

A1.1 The quantum mechanics of atoms and molecules

John F Stanton

A1.1.1 INTRODUCTION

At the turn of the 19th century, it was generally believed that the great distance between earth and the stars would forever limit what could be learned about the universe. Apart from their approximate size and distance from earth, there seemed to be no hope of determining intensive properties of stars, such as temperature and composition. While this pessimistic attitude may seem quaint from a modern perspective, it should be remembered that all knowledge gained in these areas has been obtained by exploiting a scientific technique that did not exist 200 years ago—spectroscopy.

In 1859, Kirchoff made a breakthrough discovery about the nearest star—our sun. It had been known for some time that a number of narrow dark lines are found when sunlight is bent through a prism. These absences had been studied systematically by Fraunhofer, who also noted that dark lines can be found in the spectrum of other stars; furthermore, many of these absences are found at the same wavelengths as those in the solar spectrum. By burning substances in the laboratory, Kirchoff was able to show that some of the features are due to the presence of sodium atoms in the solar atmosphere. For the first time, it had been demonstrated that an element found on our planet is not unique, but exists elsewhere in the universe. Perhaps most important, the field of modern spectroscopy was born.

Armed with the empirical knowledge that each element in the periodic table has a characteristic spectrum, and that heating materials to a sufficiently high temperature disrupts all interatomic interactions, Bunsen and Kirchoff invented the spectroscope, an instrument that atomizes substances in a flame and then records their emission spectrum. Using this instrument, the elemental composition of several compounds and minerals were deduced by measuring the wavelength of radiation that they emit. In addition, this new science led to the discovery of elements, notably caesium and rubidium.

Despite the enormous benefits of the fledgling field of spectroscopy for chemistry, the underlying physical processes were completely unknown a century ago. It was believed that the characteristic frequencies of elements were caused by (nebulously defined) vibrations of the atoms, but even a remotely satisfactory quantitative theory proved to be elusive. In 1885, the Swiss mathematician Balmer noted that wavelengths in the visible region of the hydrogen atom emission spectrum could be fitted by the empirical equation

$$\lambda = b \left(\frac{n^2}{n^2 - m^2} \right) \quad (\text{A1.1.1})$$

where $m = 2$ and n is an integer. Subsequent study showed that frequencies in other regions of the hydrogen spectrum could be fitted to this equation by assigning different integer values to m , albeit with a different value of the constant b . Ritz noted that a simple modification of Balmer's formula

$$\frac{1}{\lambda} = R_{\text{H}} \left(\frac{1}{m^2} - \frac{1}{n^2} \right) \quad (\text{A1.1.2})$$

succeeds in fitting all the line spectra corresponding to different values of m with only the single constant R_H . Although this formula provides an important clue regarding the underlying processes involved in spectroscopy, more than two decades passed before a theory of atomic structure succeeded in deriving this equation from first principles.

The origins of line spectra as well as other unexplained phenomena such as radioactivity and the intensity profile in the emission spectrum of hot objects eventually led to a realization that the physics of the day was incomplete. New ideas were clearly needed before a detailed understanding of the submicroscopic world of atoms and molecules could be gained. At the turn of the 20th century, Planck succeeded in deriving an equation that gave a correct description of the radiation emitted by an idealized isolated solid (blackbody radiation). In the derivation, Planck assumed that the energy of electromagnetic radiation emitted by the vibrating atoms of the solid cannot have just any energy, but must be an integral multiple of $h\nu$, where ν is the frequency of the radiation and h is now known as Planck's constant. The resulting formula matched the experimental blackbody spectrum perfectly.

Another phenomenon that could not be explained by classical physics involved what is now known as the photoelectric effect. When light impinges on a metal, ionization leading to ejection of electrons happens only at wavelengths ($\lambda = c/\nu$, where c is the speed of light) below a certain threshold. At shorter wavelengths (higher frequency), the kinetic energy of the photoelectrons depends linearly on the frequency of the applied radiation field and is independent of its intensity. These findings were inconsistent with conventional electromagnetic theory. A brilliant analysis of this phenomenon by Einstein convincingly demonstrated that electromagnetic energy is indeed absorbed in bundles, or quanta (now called photons), each with energy $h\nu$ where h is precisely the same quantity that appears in Planck's formula for the blackbody emission spectrum.

While the revolutionary ideas of Planck and Einstein forged the beginnings of the quantum theory, the physics governing the structure and properties of atoms and molecules remained unknown. Independent experiments by Thomson, Weichert and Kaufmann had established that atoms are not the indivisible entities postulated by Democritus 2000 years ago and assumed in Dalton's atomic theory. Rather, it had become clear that all atoms contain identical negative charges called electrons. At first, this was viewed as a rather esoteric feature of matter, the electron being an entity that 'would never be of any use to anyone'. With time, however, the importance of the electron and its role in the structure of atoms came to be understood. Perhaps the most significant advance was Rutherford's interpretation of the scattering of alpha particles from a thin gold foil in terms of atoms containing a very small, dense, positively charged core surrounded by a cloud of electrons. This picture of atoms is fundamentally correct, and is now learned each year by millions of elementary school students.

Like the photoelectric effect, the atomic model developed by Rutherford in 1911 is not consistent with the classical theory of electromagnetism. In the hydrogen atom, the force due to Coulomb attraction between the nucleus and the electron results in acceleration of the electron (Newton's first law). Classical electromagnetic theory mandates that all accelerated bodies bearing charge must emit radiation. Since emission of radiation necessarily results in a loss of energy, the electron should eventually be captured by the nucleus. But this catastrophe does not occur. Two years after Rutherford's gold-foil experiment, the first quantitatively successful theory of an atom was developed by Bohr. This model was based on a combination of purely classical ideas, use of Planck's constant h and the bold assumption that radiative loss of energy does not occur provided the electron adheres to certain special orbits, or 'stationary states'. Specifically, electrons that move in a circular path about the nucleus with a classical angular momentum mvr equal to an integral multiple of Planck's constant divided by 2π (a quantity of sufficient general use that it is designated by the simple symbol \hbar) are immune from energy loss in the Bohr model. By simply writing the classical energy of the orbiting electron in terms of its mass m , velocity v , distance r from the nucleus and charge e ,

$$E = \frac{1}{2}mv^2 - \frac{e^2}{r} \quad (\text{A1.1.3})$$

invoking the (again classical) virial theorem that relates the average kinetic ($\langle T \rangle$) and potential ($\langle V \rangle$) energy of a system governed by a potential that depends on pairwise interactions of the form r^k via

$$\langle T \rangle = \frac{k}{2} \langle V \rangle \quad (\text{A1.1.4})$$

and using Bohr's criterion for stable orbits

$$r = \frac{n\hbar}{mv} \quad (\text{A1.1.5})$$

it is relatively easy to demonstrate that energies associated with orbits having angular momentum $n\hbar$ in the hydrogen atom are given by

$$E_n = -\frac{me^4}{2n^2\hbar^2} \quad (\text{A1.1.6})$$

with corresponding radii

$$r_n = \frac{n^2\hbar^2}{me^2}. \quad (\text{A1.1.7})$$

Bohr further postulated that *quantum jumps* between the different allowed energy levels are always accompanied by absorption or emission of a photon, as required by energy conservation, *viz.*

$$\Delta E \equiv E_n - E_m = \frac{me^4}{2\hbar^2} \left(\frac{1}{m^2} - \frac{1}{n^2} \right) = h\nu_{\text{photon}} \quad (\text{A1.1.8})$$

or perhaps more illustratively

$$\frac{1}{\lambda_{\text{photon}}} = \frac{me^4}{4\pi\hbar^3c} \left(\frac{1}{m^2} - \frac{1}{n^2} \right) \quad (\text{A1.1.9})$$

precisely the form of the equation deduced by Ritz. The constant term of [equation \(A1.1.2\)](#) calculated from Bohr's equation did not exactly reproduce the experimental value at first. However, this situation was quickly remedied when it was realized that a proper treatment of the two-particle problem involved use of the reduced mass of the system $\mu \equiv mm_{\text{proton}}/(m + m_{\text{proton}})$, a minor modification that gives striking agreement with experiment.

Despite its success in reproducing the hydrogen atom spectrum, the Bohr model of the atom rapidly encountered difficulties. Advances in the resolution obtained in spectroscopic experiments had shown that the spectral features of the hydrogen atom are actually composed of several closely spaced lines; these are not accounted for by quantum jumps between Bohr's allowed orbits. However, by modifying the Bohr model to

allow for elliptical orbits and to include the special theory of relativity, Sommerfeld was able to account for some of the *fine structure* of spectral lines. More serious problems arose when the planetary model was applied to systems that contained more than one electron. Efforts to calculate the spectrum of helium were completely unsuccessful, as was a calculation of the spectrum of the hydrogen molecule ion (H_2^+) that used a generalization of the Bohr model to treat a problem involving two nuclei. This latter work formed the basis of the PhD thesis of Pauli, who was to become one of the principal players in the development of a more mature and comprehensive theory of atoms and molecules.

In retrospect, the Bohr model of the hydrogen atom contains several flaws. Perhaps most prominent among these is that the angular momentum of the hydrogen ground state ($n = 1$) given by the model is \hbar ; it is now known that the correct value is zero. Efforts to remedy the Bohr model for its insufficiencies, pursued doggedly by Sommerfeld and others, were ultimately unsuccessful. This ‘old’ quantum theory was replaced in the 1920s by a considerably more abstract framework that forms the basis for our current understanding of the detailed physics governing chemical processes. The modern quantum theory, unlike Bohr’s, does not involve classical ideas coupled with an *ad hoc* incorporation of Planck’s quantum hypothesis. It is instead founded upon a limited number of fundamental principles that cannot be proven, but must be regarded as laws of nature. While the modern theory of quantum mechanics is exceedingly complex and fraught with certain philosophical paradoxes (which will not be discussed), it has withstood the test of time; no contradiction between predictions of the theory and actual atomic or molecular phenomena has ever been observed.

The purpose of this chapter is to provide an introduction to the basic framework of quantum mechanics, with an emphasis on aspects that are most relevant for the study of atoms and molecules. After summarizing the basic principles of the subject that represent required knowledge for all students of physical chemistry, the independent-particle approximation so important in molecular quantum mechanics is introduced. A significant effort is made to describe this approach in detail and to communicate how it is used as a foundation for qualitative understanding and as a basis for more accurate treatments. Following this, the basic techniques used in accurate calculations that go beyond the independent-particle picture (variational method and perturbation theory) are described, with some attention given to how they are actually used in practical calculations.

It is clearly impossible to present a comprehensive discussion of quantum mechanics in a chapter of this length. Instead, one is forced to present cursory overviews of many topics or to limit the scope and provide a more rigorous treatment of a select group of subjects. The latter alternative has been followed here. Consequently, many areas of quantum mechanics are largely ignored. For the most part, however, the areas lightly touched upon or completely absent from this chapter are specifically dealt with elsewhere in the encyclopedia. Notable among these are the interaction between matter and radiation, spin and magnetism, techniques of quantum chemistry including the Born–Oppenheimer approximation, the Hartree–Fock method and electron correlation, scattering theory and the treatment of internal nuclear motion (rotation and vibration) in molecules.

A1.1.2 CONCEPTS OF QUANTUM MECHANICS

A1.1.2.1 BEGINNINGS AND FUNDAMENTAL POSTULATES

The modern quantum theory derives from work done independently by Heisenberg and Schrödinger in the mid-1920s. Superficially, the mathematical formalisms developed by these individuals appear very different; the quantum mechanics of Heisenberg is based on the properties of matrices, while that of Schrödinger is founded upon a differential equation that bears similarities to those used in the classical theory of waves. Schrödinger’s formulation was strongly influenced by the work of de Broglie, who made the revolutionary

hypothesis that entities previously thought to be strictly particle-like (electrons) can exhibit wavelike behaviour (such as diffraction) with particle ‘wavelength’ and momentum (p) related by the equation $\lambda = h/p$. This truly startling premise was subsequently verified independently by Davisson and Germer as well as by Thomson, who showed that electrons exhibit diffraction patterns when passed through crystals and very small circular apertures, respectively. Both the treatment of Heisenberg, which did not make use of wave theory concepts, and that of Schrödinger were successfully applied to the calculation of the hydrogen atom spectrum. It was ultimately proven by both Pauli and Schrödinger that the ‘matrix mechanics’ of Heisenberg and the ‘wave mechanics’ of Schrödinger are mathematically equivalent. Connections between the two methods were further clarified by the transformation theory of Dirac and Jordan. The importance of this new quantum theory was recognized immediately and Heisenberg, Schrödinger and Dirac shared the 1932 Nobel Prize in physics for their work.

While not unique, the Schrödinger picture of quantum mechanics is the most familiar to chemists principally because it has proven to be the simplest to use in practical calculations. Hence, the remainder of this section will focus on the Schrödinger formulation and its associated wavefunctions, operators and eigenvalues. Moreover, effects associated with the special theory of relativity (which include spin) will be ignored in this subsection. Treatments of alternative formulations of quantum mechanics and discussions of relativistic effects can be found in the reading list that accompanies this chapter.

Like the geometry of Euclid and the mechanics of Newton, quantum mechanics is an axiomatic subject. By making several assertions, or postulates, about the mathematical properties of and physical interpretation associated with solutions to the Schrödinger equation, the subject of quantum mechanics can be applied to understand behaviour in atomic and molecular systems. The first of these postulates is:

1. Corresponding to any collection of n particles, there exists a time-dependent function $\Psi(q_1, q_2, \dots, q_n; t)$ that comprises all information that can be known about the system. This function must be continuous and single valued, and have continuous first derivatives at all points where the classical force has a finite magnitude.

In classical mechanics, the state of the system may be completely specified by the set of Cartesian particle coordinates r_i and velocities dr_i/dt at any given time. These evolve according to Newton’s equations of motion. In principle, one can write down equations involving the state variables and forces acting on the particles which can be solved to give the location and velocity of each particle at any later (or earlier) time t' , provided one knows the precise state of the classical system at time t . In quantum mechanics, the state of the system at time t is instead described by a well behaved mathematical function of the particle coordinates q_i rather than a simple list of positions and velocities.

-6-

The relationship between this *wavefunction* (sometimes called *state function*) and the location of particles in the system forms the basis for a second postulate:

2. The product of $\Psi(q_1, q_2, \dots, q_n; t)$ and its complex conjugate has the following physical interpretation. The probability of finding the n particles of the system in the regions bounded by the coordinates q'_1, q'_2, \dots, q'_n and $q''_1, q''_2, \dots, q''_n$ at time t is proportional to the integral

$$\int_{q'_1}^{q''_1} \int_{q'_2}^{q''_2} \dots \int_{q'_n}^{q''_n} \Psi^*(q_1, q_2, \dots, q_n; t) \Psi(q_1, q_2, \dots, q_n; t) dq_1 dq_2 \dots dq_n. \quad (\text{A1.1.10})$$

The proportionality between the integral and the probability can be replaced by an equivalence if the wavefunction is scaled appropriately. Specifically, since the probability that the n particles will be found somewhere must be unity, the wavefunction can be scaled so that the equality

$$\int \Psi^*(q_1, q_2, \dots, q_n; t) \Psi(q_1, q_2, \dots, q_n; t) d\tau = 1 \quad (\text{A1.1.11})$$

is satisfied. The symbol $d\tau$ introduced here and used throughout the remainder of this section indicates that the integral is to be taken over the full range of all particle coordinates. Any wavefunction that satisfies equation (A1.1.11) is said to be *normalized*. The product $\Psi^*\Psi$ corresponding to a normalized wavefunction is sometimes called a probability, but this is an imprecise use of the word. It is instead a *probability density*, which must be integrated to find the chance that a given measurement will find the particles in a certain region of space. This distinction can be understood by considering the classical counterpart of $\Psi^*\Psi$ for a single particle moving on the x -axis. In classical mechanics, the probability at time t for finding the particle at the coordinate (x') obtained by propagating Newton's equations of motion from some set of initial conditions is exactly equal to one; it is zero for any other value of x . What is the corresponding probability density function, $P(x; t)$ Clearly, $P(x; t)$ vanishes at all points other than x' since its integral over any interval that does not include x' must equal zero. At x' , the value of $P(x; t)$ must be chosen so that the normalization condition

$$\int_{-\infty}^{\infty} P(x; t) dx = 1 \quad (\text{A1.1.12})$$

is satisfied. Functions such as this play a useful role in quantum mechanics. They are known as *Dirac delta functions*, and are designated by $\delta(r - r_0)$. These functions have the properties

$$\int \delta(\mathbf{r} - \mathbf{r}_0) d\tau = 1 \quad (\text{A1.1.13})$$

$$\int f(\mathbf{r}) \delta(\mathbf{r} - \mathbf{r}_0) d\tau = f(\mathbf{r}_0) \quad (\text{A1.1.14})$$

$$\delta(\mathbf{r} - \mathbf{r}_0) = 0 \quad \text{for } \mathbf{r} \neq \mathbf{r}_0. \quad (\text{A1.1.15})$$

-7-

Although a seemingly odd mathematical entity, it is not hard to appreciate that a simple one-dimensional realization of the classical $P(x; t)$ can be constructed from the familiar Gaussian distribution centred about x' by letting the standard deviation (σ) go to zero,

$$P(x; t) = \lim_{\sigma \rightarrow 0} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(x - x')^2}{2\sigma^2}\right]. \quad (\text{A1.1.16})$$

Hence, although the probability for finding the particle at x' is equal to one, the corresponding probability density function is infinitely large. In quantum mechanics, the probability density is generally nonzero for all values of the coordinates, and its magnitude can be used to determine which regions are most likely to contain particles. However, because the number of possible coordinates is infinite, the probability associated with any precisely specified choice is zero. The discussion above shows a clear distinction between classical and quantum mechanics; given a set of initial conditions, the locations of the particles are determined exactly at all future times in the former, while one generally can speak only about the probability associated with a given range of coordinates in quantum mechanics.

To extract information from the wavefunction about properties other than the probability density, additional postulates are needed. All of these rely upon the mathematical concepts of operators, eigenvalues and eigenfunctions. An extensive discussion of these important elements of the formalism of quantum mechanics is precluded by space limitations. For further details, the reader is referred to the reading list supplied at the end of this chapter. In quantum mechanics, the classical notions of position, momentum, energy etc are replaced by mathematical operators that act upon the wavefunction to provide information about the system. The third postulate relates to certain properties of these operators:

3. Associated with each system property A is a linear, Hermitian operator \hat{A} .

Although not a unique prescription, the quantum-mechanical operators \hat{A} can be obtained from their classical counterparts A by making the substitutions $x \rightarrow x$ (coordinates); $t \rightarrow t$ (time); $p_q \rightarrow -i\hbar\partial/\partial q$ (component of momentum). Hence, the quantum-mechanical operators of greatest relevance to the dynamics of an n -particle system such as an atom or molecule are:

Dynamical variable A	Classical quantity	Quantum-mechanical operator \hat{A}
Time	t	t
Position of particle i	r_i	r_i
Momentum of particle i	$m_i v_i$	$-i\hbar\nabla_i$
Angular momentum of particle i	$m_i v_i \times r_i$	$-i\hbar\nabla_i \times r_i$
Kinetic energy of particle i	$\frac{p_i \cdot p_i}{2m_i}$	$-\frac{\hbar^2}{2m_i} \nabla_i^2$
Potential energy	$V(q, t)$	$V(q, t)$

-8-

where the gradient

$$\nabla_i \equiv \frac{\partial}{\partial x_i} \mathbf{i} + \frac{\partial}{\partial y_i} \mathbf{j} + \frac{\partial}{\partial z_i} \mathbf{k} \quad (\text{A1.1.17})$$

and Laplacian

$$\nabla_i^2 \equiv \nabla \cdot \nabla = \frac{\partial^2}{\partial x_i^2} + \frac{\partial^2}{\partial y_i^2} + \frac{\partial^2}{\partial z_i^2} \quad (\text{A1.1.18})$$

operators have been introduced. Note that a potential energy which depends upon only particle coordinates and time has exactly the same form in classical and quantum mechanics. A particularly useful operator in quantum mechanics is that which corresponds to the total energy. This *Hamiltonian* operator is obtained by simply adding the potential and kinetic energy operators

$$\hat{H} \equiv \hat{T} + \hat{V} = - \sum_{\text{particles}} \frac{\hbar^2}{2m_i} \left[\frac{\partial^2}{\partial x_i^2} + \frac{\partial^2}{\partial y_i^2} + \frac{\partial^2}{\partial z_i^2} \right] + V(\mathbf{q}, t). \quad (\text{A1.1.19})$$

The relationship between the abstract quantum-mechanical operators \hat{A} and the corresponding physical quantities A is the subject of the fourth postulate, which states:

4. If the system property A is measured, the only values that can possibly be observed are those that correspond to eigenvalues of the quantum-mechanical operator \hat{A} .

An illustrative example is provided by investigating the possible momenta for a single particle travelling in the x -direction, p_x . First, one writes the equation that defines the eigenvalue condition

$$\hat{p}_x f(x) = -i\hbar \frac{df(x)}{dx} = \lambda f(x) \quad (\text{A1.1.20})$$

where λ is an eigenvalue of the momentum operator and $f(x)$ is the associated eigenfunction. It is easily verified that this differential equation has an infinite number of solutions of the form

$$f_k(x) = A \exp(ikx) \quad (\text{A1.1.21})$$

with corresponding eigenvalues

$$\lambda_k = \hbar k \quad (\text{A1.1.22})$$

-9-

in which k can assume any value. Hence, nature places no restrictions on allowed values of the linear momentum. Does this mean that a quantum-mechanical particle in a particular state $\psi(x; t)$ is allowed to have any value of p_x ? The answer to this question is 'yes', but the interpretation of its consequences rather subtle. Eventually a fifth postulate will be required to establish the connection between the quantum-mechanical wavefunction ψ and the possible outcomes associated with measuring properties of the system. It turns out that the set of possible momenta for our particle depends entirely on its wavefunction, as might be expected from the first postulate given above. The infinite set of solutions to [equation \(A1.1.20\)](#) means only that no values of the momentum are excluded, in the sense that they can be associated with a particle described by an *appropriately chosen* wavefunction. However, the choice of a *specific* function might (or might not) impose restrictions on which values of p_x are allowed.

The rather complicated issues raised in the preceding paragraph are central to the subject of quantum mechanics, and their resolution forms the basis of one of the most important postulates associated with the Schrödinger formulation of the subject. In the example above, discussion focuses entirely on the eigenvalues of the momentum operator. What significance, if any, can be attached to the eigenfunctions of quantum-mechanical operators? In the interest of simplicity, the remainder of this subsection will focus entirely on the quantum mechanics associated with operators that have a finite number of eigenvalues. These are said to have a *discrete spectrum*, in contrast to those such as the linear momentum, which have a *continuous spectrum*. Discrete spectra of eigenvalues arise whenever boundaries limit the region of space in which a system can be. Examples are particles in hard-walled boxes, or soft-walled shells and particles attached to springs. The results developed below can all be generalized to the continuous case, but at the expense of increased mathematical complexity. Readers interested in these details should consult chapter 1 of Landau and Lifschitz (see additional reading).

It can be shown that the eigenfunctions of Hermitian operators necessarily exhibit a number of useful mathematical properties. First, if all eigenvalues are distinct, the set of eigenfunctions $\{f_1, f_2 \dots f_n\}$ are *orthogonal* in the sense that the integral of the product formed from the complex conjugate of eigenfunction j (f_j^*) and eigenfunction k (f_k) vanishes unless $j = k$,

$$\int f_j^* f_k \, d\tau = 0 \text{ if } j \neq k. \quad (\text{A1.1.23})$$

If there are identical eigenvalues (a common occurrence in atomic and molecular quantum mechanics), it is permissible to form linear combinations of the eigenfunctions corresponding to these *degenerate* eigenvalues, as these must also be eigenfunctions of the operator. By making a judicious choice of the expansion coefficients, the degenerate eigenfunctions can also be made orthogonal to one another. Another useful property is that the set of eigenfunctions is said to be *complete*. This means that any function of the coordinates that appear in the operator can be written as a linear combination of its eigenfunctions, provided that the function obeys the same *boundary conditions* as the eigenfunctions and shares any fundamental symmetry property that is common to all of them. If, for example, all of the eigenfunctions vanish at some point in space, then only functions that vanish at the same point can be written as linear combinations of the eigenfunctions. Similarly, if the eigenfunctions of a particular operator in one dimension are all odd functions of the coordinate, then all linear combinations of them must also be odd. It is clearly impossible in the latter case to expand functions such as $\cos(x)$, $\exp(-x^2)$ etc in terms of odd functions. This qualification is omitted in some elementary treatments of quantum mechanics, but it is one that turns out to be important for systems containing several identical particles. Nevertheless, if these criteria are met by a suitable function g , then it is always possible to find coefficients c_k such that

-10-

$$g = \sum_k c_k f_k \quad (\text{A1.1.24})$$

where the coefficient c_j is given by

$$c_j = \frac{\int f_j^* g \, d\tau}{\int f_j^* f_j \, d\tau}. \quad (\text{A1.1.25})$$

If the eigenfunctions are normalized, this expression reduces to

$$c_j = \int f_j^* g \, d\tau. \quad (\text{A1.1.26})$$

When normalized, the eigenfunctions corresponding to a Hermitian operator are said to represent an *orthonormal set*.

The mathematical properties discussed above are central to the next postulate:

5. In any experiment, the probability of observing a particular non-degenerate value for the system property A can be determined by the following procedure. First, expand the wavefunction in terms of the complete set of normalized eigenfunctions of the quantum-mechanical operator, \hat{A} ,

$$\Psi = \sum_j c_j \phi_j. \quad (\text{A1.1.27})$$

The probability of measuring $A = \lambda_k$, where λ_k is the eigenvalue associated with the normalized eigenfunction ϕ_k , is precisely equal to $|c_k|^2$ ($\equiv c_k^* c_k$). For degenerate eigenvalues, the probability of observation is given by $\sum |c_k|^2$, where the sum is taken over all of the eigenfunctions ϕ_k that correspond to the degenerate eigenvalue λ_k .

At this point, it is appropriate to mention an elementary concept from the theory of probability. If there are n possible numerical outcomes (ξ_n) associated with a particular process, the average value ($\langle \xi \rangle$) can be calculated by summing up all of the outcomes, each weighted by its corresponding probability

$$\langle \xi \rangle = \sum_i P_i \xi_i. \quad (\text{A1.1.28})$$

-11-

As an example, the possible outcomes and associated probabilities for rolling a pair of six-sided dice are

Sum	Probability
2	1/36
3	1/18
4	1/12
5	1/9
6	5/36
7	1/6
8	5/36
9	1/9
10	1/12
11	1/18
12	1/36

The average value is therefore given by the sum

$$\frac{1}{36}(2) + \frac{1}{18}(3) + \frac{1}{12}(4) + \frac{1}{9}(5) + \frac{5}{36}(6) + \frac{1}{6}(7) + \frac{5}{36}(8) + \frac{1}{9}(9) + \frac{1}{12}(10) + \frac{1}{18}(11) + \frac{1}{36}(12) = 7.$$

What does this have to do with quantum mechanics? To establish a connection, it is necessary to first expand the wavefunction in terms of the eigenfunctions of a quantum-mechanical operator \hat{A} ,

$$\Psi = \sum_k c_k \phi_k. \quad (\text{A1.1.29})$$

We will assume that both the wavefunction and the orthogonal eigenfunctions are normalized, which implies that

$$\sum_j c_j^* c_j = \sum_j |c_j|^2 = 1. \quad (\text{A1.1.30})$$

Now, the operator \hat{A} is applied to both sides of equation (A1.1.29), which because of its *linearity*, gives

$$\hat{A}\Psi = \hat{A} \sum_k c_k \phi_k = \sum_k c_k \hat{A}\phi_k = \sum_k c_k \lambda_k \phi_k \quad (\text{A1.1.31})$$

-12-

where λ_k represents the eigenvalue associated with the eigenfunction ϕ_k . Next, both sides of the preceding equation are multiplied from the left by the complex conjugate of the wavefunction and integrated over all space

$$\int \Psi^* \hat{A}\Psi \, d\tau = \int \sum_j \sum_k c_j^* c_k \lambda_k \phi_j^* \phi_k \, d\tau \quad (\text{A1.1.32})$$

$$= \sum_j \sum_k c_j^* c_k \lambda_k \int \phi_j^* \phi_k \, d\tau \quad (\text{A1.1.33})$$

$$= \sum_k c_k^* c_k \lambda_k = \sum_k |c_k|^2 \lambda_k. \quad (\text{A1.1.34})$$

The last identity follows from the orthogonality property of eigenfunctions and the assumption of normalization. The right-hand side in the final result is simply equal to the sum over all eigenvalues of the operator (possible results of the measurement) multiplied by the respective probabilities. Hence, an important corollary to the fifth postulate is established:

$$\langle A \rangle = \int \Psi^* \hat{A}\Psi \, d\tau. \quad (\text{A1.1.35})$$

This provides a recipe for calculating the average value of the system property associated with the quantum-mechanical operator \hat{A} , for a specific but arbitrary choice of the wavefunction Ψ , notably those choices which are not eigenfunctions of \hat{A} .

The fifth postulate and its corollary are extremely important concepts. Unlike classical mechanics, where everything can in principle be known with precision, one can generally talk only about the probabilities associated with each member of a set of possible outcomes in quantum mechanics. By making a measurement of the quantity A , all that can be said with certainty is that *one* of the eigenvalues of \hat{A} will be observed, and its probability can be calculated precisely. However, if it happens that the wavefunction corresponds to one of the eigenfunctions of the operator \hat{A} , then and only then is the outcome of the experiment certain: the measured value of A will be the corresponding eigenvalue.

Up until now, little has been said about time. In classical mechanics, complete knowledge about the system at any time t suffices to predict with absolute certainty the properties of the system at any other time t' . The situation is quite different in quantum mechanics, however, as it is not possible to know everything about the system at *any* time t . Nevertheless, the temporal behavior of a quantum-mechanical system evolves in a well defined way that depends on the Hamiltonian operator and the wavefunction Ψ according to the last postulate

6. The time evolution of the wavefunction is described by the differential equation

$$i\hbar \frac{d}{dt} \Psi(q_1, q_2, \dots, q_n; t) = \hat{H} \Psi(q_1, q_2, \dots, q_n; t). \quad (\text{A1.1.36})$$

The differential equation above is known as the time-dependent Schrödinger equation. There is an interesting and

-13-

intimate connection between this equation and the classical expression for a travelling wave

$$A(x, t) = A \exp\left(2\pi i \left\{ \frac{x}{\lambda} - \nu t \right\}\right). \quad (\text{A1.1.37})$$

To convert (A1.1.37) into a quantum-mechanical form that describes the ‘matter wave’ associated with a free particle travelling through space, one might be tempted to simply make the substitutions $\nu = E/h$ (Planck’s hypothesis) and $\lambda = h/p$ (de Broglie’s hypothesis). It is relatively easy to verify that the resulting expression satisfies the time-dependent Schrödinger equation. However, it should be emphasized that this is not a derivation, as there is no compelling reason to believe that this *ad hoc* procedure should yield one of the fundamental equations of physics. Indeed, the time-dependent Schrödinger equation cannot be derived in a rigorous way and therefore must be regarded as a postulate.

The time-dependent Schrödinger equation allows the precise determination of the wavefunction at any time t from knowledge of the wavefunction at some initial time, provided that the forces acting within the system are known (these are required to construct the Hamiltonian). While this suggests that quantum mechanics has a deterministic component, it must be emphasized that it is not the observable system properties that evolve in a precisely specified way, but rather the probabilities associated with values that might be found for them in a measurement.

A1.1.2.2 STATIONARY STATES, SUPERPOSITION AND UNCERTAINTY

From the very beginning of the 20th century, the concept of energy conservation has made it abundantly clear that electromagnetic energy emitted from and absorbed by material substances must be accompanied by compensating energy changes within the material. Hence, the discrete nature of atomic line spectra suggested that only certain energies are allowed by nature for each kind of atom. The wavelengths of radiation emitted or absorbed must therefore be related to the *difference* between energy levels via Planck’s hypothesis, $\Delta E = h\nu = hc/\lambda$.

The Schrödinger picture of quantum mechanics summarized in the previous subsection allows an important deduction to be made that bears directly on the subject of energy levels and spectroscopy. Specifically, the energies of spectroscopic transitions must correspond precisely to differences between distinct eigenvalues of the Hamiltonian operator, as these correspond to the allowed energy levels of the system. Hence, the set of eigenvalues of the Hamiltonian operator are of central importance in chemistry. These can be determined by solving the so-called *time-independent Schrödinger equation*,

$$H \psi_k(q_1, q_2, \dots, q_n) = E_k \psi_k(q_1, q_2, \dots, q_n) \quad (\text{A1.1.38})$$

for the eigenvalues E_k and eigenfunctions ψ_k . It should be clear that the set of eigenfunctions and eigenvalues does not evolve with time provided the Hamiltonian operator itself is time independent. Moreover, since the

eigenfunctions of the Hamiltonian (like those of any other operator) form a complete set, it is always possible to expand the exact wavefunction of the system at any time in terms of them:

$$\Psi(\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n; t) = \sum_j c_j(t) \psi_j(\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n). \quad (\text{A1.1.39})$$

-14-

It is important to point out that this expansion is valid even if time-dependent terms are added to the Hamiltonian (as, for example, when an electric field is turned on). If there is more than one nonzero value of c_j at any time t , then the system is said to be in a *superposition* of the energy eigenstates ψ_k associated with non-vanishing expansion coefficients, c_k . If it were possible to measure energies directly, then the fifth postulate of the previous section tells us that the probability of finding energy E_k in a given measurement would be $c_k^* c_k$.

When a molecule is isolated from external fields, the Hamiltonian contains only kinetic energy operators for all of the electrons and nuclei as well as terms that account for repulsion and attraction between all distinct pairs of like and unlike charges, respectively. In such a case, the Hamiltonian is constant in time. When this condition is satisfied, the representation of the time-dependent wavefunction as a superposition of Hamiltonian eigenfunctions can be used to determine the time dependence of the expansion coefficients. If [equation \(A1.1.39\)](#) is substituted into the time-dependent Schrödinger equation

$$i\hbar \frac{d}{dt} \sum_k c_k(t) \psi_k = H \sum_k c_k(t) \psi_k \quad (\text{A1.1.40})$$

the simplification

$$i\hbar \sum_k \psi_k \frac{d}{dt} c_k(t) = \sum_k E_k c_k(t) \psi_k \quad (\text{A1.1.41})$$

can be made to the right-hand side since the restriction of a time-independent Hamiltonian means that ψ_k is *always* an eigenfunction of H . By simply equating the coefficients of the ψ_k , it is easy to show that the choice

$$c_k(t) = c_k(0) \exp\left(\frac{iE_k t}{\hbar}\right) \quad (\text{A1.1.42})$$

for the time-dependent expansion coefficients satisfies equation (A1.1.41). Like any differential equation, there are an infinite number of solutions from which a choice must be made to satisfy some set of initial conditions. The state of the quantum-mechanical system at time $t = 0$ is used to fix the arbitrary multipliers $c_k(0)$, which can always be chosen as real numbers. Hence, the wavefunction Ψ becomes

$$\Psi = \sum_k c_k(0) \exp\left(\frac{iE_k t}{\hbar}\right) \psi_k. \quad (\text{A1.1.43})$$

Suppose that the system property A is of interest, and that it corresponds to the quantum-mechanical operator \hat{A} . The average value of A obtained in a series of measurements can be calculated by exploiting the corollary to the fifth postulate

$$\langle A \rangle = \int \Psi^* \hat{A} \Psi \, d\tau = \sum_j \sum_k c_j(0) c_k(0) \int \exp\left(\frac{-iE_j t}{\hbar}\right) \psi_j^* \hat{A} \exp\left(\frac{iE_k t}{\hbar}\right) \psi_k \, d\tau. \quad (\text{A1.1.44})$$

-15-

Now consider the case where \hat{A} is itself a time-independent operator, such as that for the position, momentum or angular momentum of a particle or even the energy of the benzene molecule. In these cases, the time-dependent expansion coefficients are unaffected by application of the operator, and one obtains

$$\begin{aligned} \langle A \rangle &= \sum_j \sum_k c_j(0) c_k(0) \exp\left[\frac{i(E_k - E_j)t}{\hbar}\right] \int \psi_j^* \hat{A} \psi_k \, d\tau \\ &= \sum_j |c_j(0)|^2 \int \psi_j^* \hat{A} \psi_j + \sum_j \sum_{k \neq j} c_j(0) c_k(0) \cos\left[\frac{(E_j - E_k)t}{\hbar}\right] \int \psi_j^* \hat{A} \psi_k \, d\tau. \end{aligned} \quad (\text{A1.1.45})$$

As one might expect, the first term that contributes to the expectation value of A is simply its value at $t = 0$, while the second term exhibits an oscillatory time dependence. If the superposition initially includes large contributions from states of widely varying energy, then the oscillations in $\langle A \rangle$ will be rapid. If the states that are strongly mixed have similar energies, then the timescale for oscillation in the properties will be slower. However, there is one special class of system properties A that exhibit no time dependence whatsoever. If (and only if) every one of the states ψ_k is an eigenfunction of \hat{A} , then the property of orthogonality can be used to show that every contribution to the second term vanishes. An obvious example is the Hamiltonian operator itself; it turns out that the expectation value for the energy of a system subjected to forces that do not vary with time is a constant. Are there other operators that share the same set of eigenfunctions ψ_k with \hat{H} , and if so, how can they be recognized? It can be shown that any two operators which satisfy the property

$$\hat{A}\hat{B}f = \hat{B}\hat{A}f \Rightarrow [\hat{A}, \hat{B}]f = 0 \quad (\text{A1.1.46})$$

for all functions f share a common set of eigenfunctions, and A and B are said to *commute*. (The symbol $[\hat{A}, \hat{B}]$ meaning $\hat{A}\hat{B} - \hat{B}\hat{A}$, is called the *commutator* of the operators \hat{A} and \hat{B} .) Hence, there is no time dependence for the expectation value of any system property that corresponds to a quantum-mechanical operator that commutes with the Hamiltonian. Accordingly, these quantities are known as *constants of the motion*: their average values will not vary, provided the environment of the system does not change (as it would, for example, if an electromagnetic field were suddenly turned on). In nonrelativistic quantum mechanics, two examples of constants of the motion are the square of the total angular momentum, as well as its projection along an arbitrarily chosen axis. Other operators, such as that for the dipole moment, do not commute with the Hamiltonian and the expectation value associated with the corresponding properties can indeed oscillate with time. It is important to note that the frequency of these oscillations is given by differences between the allowed energies of the system divided by Planck's constant. These are the so-called *Bohr frequencies*, and it is perhaps not surprising that these are exactly the frequencies of electromagnetic radiation that cause transitions between the corresponding energy levels.

Close inspection of equation (A1.1.45) reveals that, under very special circumstances, the expectation value does not change with time for *any* system properties that correspond to fixed (static) operator representations. Specifically, if the spatial part of the time-dependent wavefunction is the exact eigenfunction ψ_j of the Hamiltonian, then $c_j(0) = 1$ (the zero of time can be chosen arbitrarily) and all other $c_k(0) = 0$. The second term clearly vanishes in these cases, which are known as *stationary states*. As the name implies, all observable properties of these states do not vary with time. In a stationary state, the energy of the system has a precise value (the corresponding eigenvalue of \hat{H}) as do observables that are associated with operators that commute with \hat{H} . For all other properties (such as the position and momentum),

one can speak only about average values or probabilities associated with a given measurement, but these quantities themselves do not depend on time. When an external perturbation such as an electric field is applied or a collision with another atom or molecule occurs, however, the system and its properties generally will evolve with time. The energies that can be absorbed or emitted in these processes correspond precisely to differences in the stationary state energies, so it should be clear that solving the time-independent Schrödinger equation for the stationary state wavefunctions and eigenvalues provides a wealth of spectroscopic information. The importance of stationary state solutions is so great that it is common to refer to [equation \(A1.1.38\)](#) as ‘the Schrödinger equation’, while the qualified name ‘time-dependent Schrödinger equation’ is generally used for [equation \(A1.1.36\)](#). Indeed, the subsequent subsections are devoted entirely to discussions that centre on the former and its exact and approximate solutions, and the qualifier ‘time independent’ will be omitted.

Starting with the quantum-mechanical postulate regarding a one-to-one correspondence between system properties and Hermitian operators, and the mathematical result that only operators which commute have a common set of eigenfunctions, a rather remarkable property of nature can be demonstrated. Suppose that one desires to determine the values of the two quantities A and B , and that the corresponding quantum-mechanical operators do not commute. In addition, the properties are to be measured simultaneously so that both reflect the same quantum-mechanical state of the system. If the wavefunction is neither an eigenfunction of \hat{A} nor \hat{B} , then there is necessarily some uncertainty associated with the measurement. To see this, simply expand the wavefunction ψ in terms of the eigenfunctions of the relevant operators

$$\psi = \sum_k a_k f_k^A \tag{A1.1.47}$$

$$\psi = \sum_k b_k f_k^B \tag{A1.1.48}$$

where the eigenfunctions f_k^A and f_k^B of operators \hat{A} and \hat{B} , respectively, are associated with corresponding eigenvalues λ_k^A and λ_k^B . Given that ψ is not an eigenfunction of either operator, at least two of the coefficients a_k and two of the b_k must be nonzero. Since the probability of observing a particular eigenvalue is proportional to the square of the expansion coefficient corresponding to the associated eigenfunction, there will be no less than four possible outcomes for the set of values A and B . Clearly, they both cannot be determined precisely. Indeed, under these conditions, neither of them can be!

In a more favourable case, the wavefunction ψ might indeed correspond to an eigenfunction of one of the operators. If $\psi = f_m^A$, then a measurement of A necessarily yields λ_m^A , and this is an unambiguous result. What can be said about the measurement of B in this case? It has already been said that the eigenfunctions of two commuting operators are identical, but here the pertinent issue concerns eigenfunctions of two operators that do not commute. Suppose f_m^A is an eigenfunction of \hat{A} . Then, it must be true that

$$\begin{aligned} \hat{A} f_m^A &= \lambda_m^A f_m^A \\ \hat{B} \hat{A} f_m^A &= \lambda_m^A \hat{B} f_m^A. \end{aligned} \tag{A1.1.49}$$

If f_m^A is also an eigenfunction of \hat{B} , then it follows that $\hat{A}\hat{B}f_m^A = \hat{B}\hat{A}f_m^A = \lambda_m^A\lambda_m^B f_m^A$, which contradicts the assumption that \hat{A} and \hat{B} do not commute. Hence, no nontrivial eigenfunction of \hat{A} can also be an eigenfunction of \hat{B} . Therefore, if measurement of A yields a precise result, then some uncertainty must be associated with B . That is, the expansion of ψ in terms of eigenfunctions of \hat{B} (equation (A1.1.48)) must have at least two non-vanishing coefficients; the corresponding eigenvalues therefore represent distinct possible outcomes of the experiment, each having probability $b_k^*b_k$. A physical interpretation of $\hat{A}f_m^A$ is the process of measuring the value of A for a system in a state with a unique value for this property λ_m^A . However $\hat{B}f_m^A$ represents a measurement that changes the state of the system, so that if after we measure B and then measure A , we would no longer find λ_m^A as its value: $\hat{B}\hat{A}f_m^A = \lambda_m^A\hat{B}f_m^A \neq \hat{A}\hat{B}f_m^A$.

The *Heisenberg uncertainty principle* offers a rigorous treatment of the qualitative picture sketched above. If several measurements of A and B are made for a system in a particular quantum state, then quantitative uncertainties are provided by standard deviations in the corresponding measurements. Denoting these as σ_A and σ_B , respectively, it can be shown that

$$\sigma_A\sigma_B \geq \frac{1}{2}|\langle[\hat{A}, \hat{B}]\rangle|. \quad (\text{A1.1.50})$$

One feature of this inequality warrants special attention. In the previous paragraph it was shown that the precise measurement of A made possible when ψ is an eigenfunction of \hat{A} necessarily results in some uncertainty in a simultaneous measurement of B when the operators \hat{A} and \hat{B} do not commute. However, the mathematical statement of the uncertainty principle tells us that measurement of B is in fact completely uncertain: one can say nothing at all about B apart from the fact that any and all values of B are equally probable! A specific example is provided by associating A and B with the position and momentum of a particle moving along the x -axis. It is rather easy to demonstrate that $[p_x, x] = -i\hbar$, so that $\sigma_{p_x}\sigma_x \geq \hbar/2$. If the system happens to be described by a Dirac delta function at the point x_0 (which is an eigenfunction of the position operator corresponding to eigenvalue x_0), then the probabilities associated with possible momenta can be determined by expanding $\delta(x-x_0)$ in terms of the momentum eigenfunctions $A \exp(ikx)$. Carrying out such a calculation shows that all of the infinite number of possible momenta (the momentum operator has a continuous spectrum) appear in the wavefunction expansion, all with precisely the same weight. Hence, no particular momentum or (more properly in this case) range bounded by $p_x + dp_x$ is more likely to be observed than any other.

A1.1.2.3 SOME QUALITATIVE FEATURES OF STATIONARY STATES

A great number of qualitative features associated with the stationary states that correspond to solutions of the time-independent Schrödinger can be worked out from rather general mathematical considerations and use of the postulates of quantum mechanics. Mastering these concepts and the qualifications that may apply to them is essential if one is to obtain an intuitive feeling for the subject. In general, the systems of interest to chemists are atoms and molecules, both in isolation as well as how they interact with each other or with an externally applied field. In all of these cases, the forces acting upon the particles in the system give rise to a potential energy function that varies with the positions of the particles, strength of the applied fields etc. In general, the potential is a smoothly varying function of the coordinates, either growing without bound for large values of the coordinates or tending asymptotically towards a finite value. In these cases, there is necessarily a minimum value at what is known as the global equilibrium position (there may be several global minima that are equivalent by symmetry). In many cases, there are also other minima

(meaning that the matrix of second derivatives with respect to the coordinates has only non-negative eigenvalues) that have higher energies, which are called local minima. If the potential becomes infinitely large for infinite values of the coordinates (as it does, for example, when the force on a particle varies linearly with its displacement from equilibrium) then all solutions to the Schrödinger equation are known as *bound states*; that with the smallest eigenvalue is called the *ground state* while the others are called *excited states*. In other cases, such as potential functions that represent realistic models for diatomic molecules by approaching a constant finite value at large separation (zero force on the particles, with a finite dissociation energy), there are two classes of solutions. Those associated with eigenvalues that are below the asymptotic value of the potential energy are the bound states, of which there is usually a finite number; those having higher energies are called the *scattering* (or *continuum*) states and form a continuous spectrum. The latter are dealt with in [section A3.11](#) of the encyclopedia and will be mentioned here only when necessary for mathematical reasons.

Bound state solutions to the Schrödinger equation decay to zero for infinite values of the coordinates, and are therefore integrable since they are continuous functions in accordance with the first postulate. The solutions may assume zero values elsewhere in space and these regions—which may be a point, a plane or a three- or higher-dimensional hypersurface—are known as *nodes*. From the mathematical theory of differential eigenvalue equations, it can be demonstrated that the lowest eigenvalue is always associated with an eigenfunction that has the same sign at all points in space. From this result, which can be derived from the calculus of variations, it follows that the wavefunction corresponding to the smallest eigenvalue of the Hamiltonian must have no nodes. It turns out, however, that relativistic considerations require that this statement be qualified. For systems that contain more than two identical particles of a specific type, not all solutions to the Schrödinger equation are allowed by nature. Because of this restriction, which is described in [subsection \(A1.1.3.3\)](#), it turns out that the ground states of lithium, all larger atoms and all molecules other than H_2^+ , H_2 and isoelectronic species have nodes. Nevertheless, our conceptual understanding of electronic structure as well as the basis for almost all highly accurate calculations is ultimately rooted in a single-particle approximation. The quantum mechanics of one-particle systems is therefore important in chemistry.

Shapes of the ground- and first three excited-state wavefunctions are shown in [figure A1.1.1](#) for a particle in one dimension subject to the potential $V = \frac{1}{2}kx^2$, which corresponds to the case where the force acting on the particle is proportional in magnitude and opposite in direction to its displacement from equilibrium ($f \equiv -\nabla V = -kx$). The corresponding Schrödinger equation

$$-\frac{\hbar^2}{2m} \frac{d^2}{dx^2} \psi + \frac{1}{2}kx^2 = E\psi \quad (\text{A1.1.51})$$

can be solved analytically, and this problem (probably familiar to most readers) is that of the quantum harmonic oscillator. As expected, the ground-state wavefunction has no nodes. The first excited state has a single node, the second two nodes and so on, with the number of nodes growing with increasing magnitude of the eigenvalue. From the form of the kinetic energy operator, one can infer that regions where the slope of the wavefunction is changing rapidly (large second derivatives) are associated with large kinetic energy. It is quite reasonable to accept that wavefunctions with regions of large curvature (where the function itself has appreciable magnitude) describe states with high energy, an expectation that can be made rigorous by applying a quantum-mechanical version of the virial theorem.

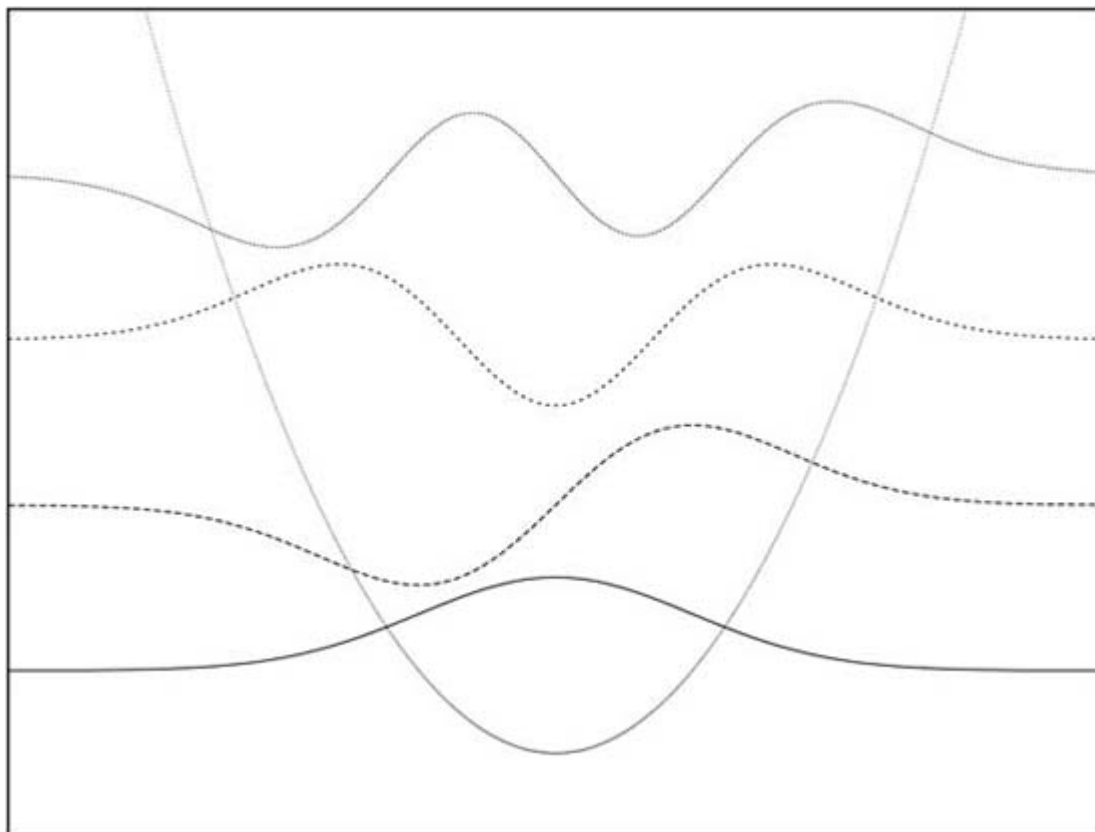


Figure A1.1.1. Wavefunctions for the four lowest states of the harmonic oscillator, ordered from the $n = 0$ ground state (at the bottom) to the $n = 3$ state (at the top). The vertical displacement of the plots is chosen so that the location of the classical turning points are those that coincide with the superimposed potential function (dotted line). Note that the number of nodes in each state corresponds to the associated quantum number.

Classically, a particle with fixed energy E described by a quadratic potential will move back and forth between the points where $V = E$, known as the *classical turning points*. Movement beyond the classical turning points is forbidden, because energy conservation implies that the particle will have a negative kinetic energy in these regions, and imaginary velocities are clearly inconsistent with the Newtonian picture of the universe. Inside the turning points, the particle will have its maximum kinetic energy as it passes through the minimum, slowing in its climb until it comes to rest and subsequently changes direction at the turning points (imagine a marble rolling in a parabola). Therefore, if a camera were to take snapshots of the particle at random intervals, most of the pictures would show the particle near the turning points (the equilibrium position is actually the least likely location for the particle). A more detailed analysis of the problem shows that the probability of seeing the classical particle in the neighbourhood of a given position x is proportional to $\frac{1}{\sqrt{E-V(x)}}$. Note that the situation found for the ground state described by quantum mechanics bears very little resemblance to the classical situation. The particle is most likely to be found at the equilibrium position and, within the classically allowed region, least likely to be seen at the turning points. However, the situation is even stranger than this: the probability of finding the particle outside the turning points is non-zero! This phenomenon, known as *tunnelling*, is not unique to the harmonic oscillator. Indeed, it occurs for bound states described by every potential

that tends asymptotically to a finite value since the wavefunction and its derivatives must approach zero in a

smooth fashion for large values of the coordinates where (by the definition of a bound state) V must exceed E . However, at large energies (see the 29th excited state probability density in figure A1.1.2, the situation is more consistent with expectations based on classical theory: the probability density has its largest value near the turning points, the general appearance is as implied by the classical formula (if one ignores the oscillations) and its magnitude in the classically forbidden region is reduced dramatically with respect to that found for the low-lying states. This merging of the quantum-mechanical picture with expectations based on classical theory always occurs for highly excited states and is the basis of the *correspondence principle*.

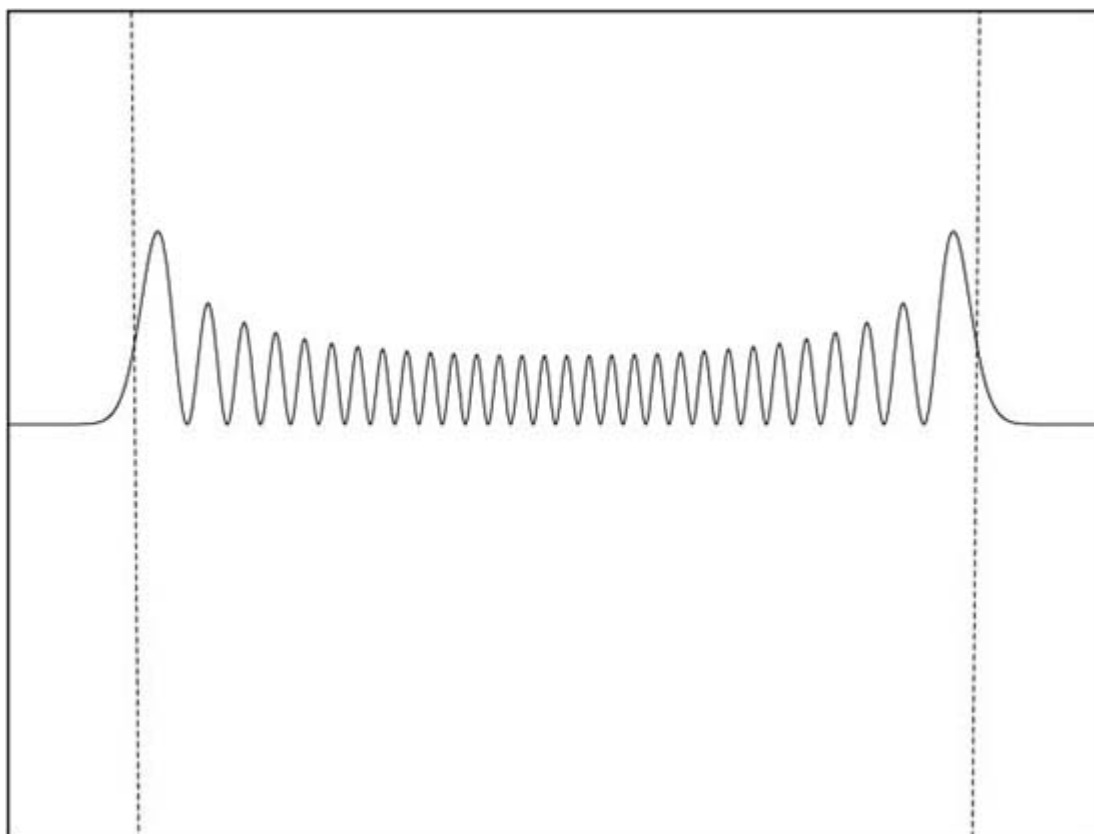


Figure A1.1.2. Probability density ($\psi^*\psi$) for the $n = 29$ state of the harmonic oscillator. The vertical state is chosen as in [figure A1.1.1](#), so that the locations of the turning points coincide with the superimposed potential function.

The energy level spectrum of the harmonic oscillator is completely regular. The ground state energy is given by $\frac{1}{2}h\nu$, where ν is the classical frequency of oscillation given by

$$\nu = \frac{1}{2\pi} \sqrt{\frac{k}{m}} \tag{A1.1.52}$$

although it must be emphasized that our inspection of the wavefunction shows that the motion of the particle cannot be literally thought of in this way. The energy of the first excited state is $h\nu$ above that of the ground state and precisely the same difference separates each excited state from those immediately above and below. A different example is provided by a particle trapped in the *Morse potential*

$$V(x) = D_e[\exp(-ax) - 1]^2, \quad (\text{A1.1.53})$$

originally suggested as a realistic model for the vibrational motion of diatomic molecules. Although the wavefunctions associated with the Morse levels exhibit largely the same qualitative features as the harmonic oscillator functions and are not shown here, the energy level structures associated with the two systems are qualitatively different. Since $V(x)$ tends to a finite value (D_e) for large x , there are only a limited number of bound state solutions, and the spacing between them decreases with increasing eigenvalue. This is another general feature; energy level spacings for states associated with potentials that tend towards asymptotic values at infinity tend to decrease with increasing quantum number.

The one-dimensional cases discussed above illustrate many of the qualitative features of quantum mechanics, and their relative simplicity makes them quite easy to study. Motion in more than one dimension and (especially) that of more than one particle is considerably more complicated, but many of the general features of these systems can be understood from simple considerations. While one relatively common feature of multidimensional problems in quantum mechanics is degeneracy, it turns out that the ground state must be non-degenerate. To prove this, simply assume the opposite to be true, i.e.

$$H\psi_1 = E_0\psi_1 \quad (\text{A1.1.54})$$

$$H\psi_2 = E_0\psi_2 \quad (\text{A1.1.55})$$

where E_0 is the ground state energy, and

$$\int \psi_1^* \psi_2 \, d\tau = 0. \quad (\text{A1.1.56})$$

In order to satisfy equation (A1.1.56), the two functions must have identical signs at some points in space and different signs elsewhere. It follows that at least one of them must have at least one node. However, this is incompatible with the nodeless property of ground-state eigenfunctions.

Having established that the ground state of a single-particle system is non-degenerate and nodeless, it is straightforward to prove that the wavefunctions associated with every excited state must contain at least one node (though they need not be degenerate!), just as seen in the example problems. It follows from the orthogonality of eigenfunctions corresponding to a Hermitian operator that

$$\int \psi_g^* \psi_x \, d\tau = 0 \quad (\text{A1.1.57})$$

for all excited states ψ_x . In order for this equality to be satisfied, it is necessary that the integrand either vanishes at all points in space (which contradicts the assumption that both ψ_g and ψ_x are nodeless) or is positive in some regions of space and negative in others. Given that the ground state has no nodes, the latter condition can be satisfied only if the excited-state wavefunction changes sign at one or more points in space. Since the first postulate states that all wavefunctions are continuous, it is therefore necessary that ψ_x has at least one node.

In classical mechanics, it is certainly possible for a system subject to dissipative forces such as friction to come to rest. For example, a marble rolling in a parabola lined with sandpaper will eventually lose its kinetic energy and come to rest at the bottom. Rather remarkably, making a measurement of E that coincides with

V_{\min} (as would be found classically for our stationary marble) is incompatible with quantum mechanics. Turning back to our example, the ground-state energy is indeed larger than the minimum value of the potential energy for the harmonic oscillator. That this property of *zero-point energy* is guaranteed in quantum mechanics can be demonstrated by straightforward application of the basic principles of the subject. Unlike nodal features of the wavefunction, the arguments developed here also hold for many-particle systems. Suppose the total energy of a stationary state is E . Since the energy is the sum of kinetic and potential energies, it must be true that expectation values of the kinetic and potential energies are related according to

$$E = \langle T \rangle + \langle V \rangle. \quad (\text{A1.1.58})$$

If the total energy associated with the state is equal to the potential energy at the equilibrium position, it follows that

$$V_{\min} - \langle V \rangle = \langle T \rangle. \quad (\text{A1.1.59})$$

Two cases must be considered. In the first, it will be assumed that the wavefunction is nonzero at one or more points for which $V > V_{\min}$ (for the physically relevant case of a smoothly varying and continuous potential, this includes all possibilities other than that in which the wavefunction is a Dirac delta function at the equilibrium position). This means that $\langle V \rangle$ must also be greater than V_{\min} thereby forcing the average kinetic energy to be negative. This is not possible. The kinetic energy operator for a quantum-mechanical particle moving in the x -direction has the (unnormalized) eigenfunctions

$$f = \exp(ikx) \quad (\text{A1.1.60})$$

where

$$k = \left(\frac{2m\alpha}{\hbar^2} \right)^{\frac{1}{2}} \quad (\text{A1.1.61})$$

and α are the corresponding eigenvalues. It can be seen that negative values of α give rise to real arguments of the exponential and correspondingly divergent eigenfunctions. Zero and non-negative values are associated with constant and oscillatory solutions in which the argument of the exponential vanishes or is imaginary, respectively. Since divergence of the actual wavefunction is incompatible with its probabilistic interpretation, no contribution from negative α eigenfunctions can appear when the wavefunction is expanded in terms of kinetic energy eigenfunctions.

It follows from the fifth postulate that the kinetic energy of each particle in the system (and therefore the total kinetic energy) is restricted to non-negative values. Therefore, the expectation value of the kinetic energy cannot be negative. The other possibility is that the wavefunction is non-vanishing only when $V = V_{\min}$. For the case of a smoothly varying, continuous potential, this corresponds to a state described by a Dirac delta function at the equilibrium position, which is the quantum-mechanical equivalent of a particle at rest. In any event, the fact that the wavefunction vanishes at all points for which $V \neq V_{\min}$ means that the expectation value of the kinetic energy operator must also vanish if there is to be no zeropoint energy. Considering the discussion above, this can occur only when the wavefunction is the same as the zero-kinetic-energy eigenfunction ($\psi = \text{constant}$). This contradicts the assumption used in this case, where the wavefunction is a

delta function. Following the general arguments used in both cases above, it is easily shown that E can only be larger than V_{\min} , which means that any measurement of E for a particle in a stationary or non-stationary state must give a result that satisfies the inequality $E > V_{\min}$.

A1.1.3 QUANTUM MECHANICS OF MANY-PARTICLE SYSTEMS

A1.1.3.1 THE HYDROGEN ATOM

It is admittedly inconsistent to begin a section on many-particle quantum mechanics by discussing a problem that can be treated as a single particle. However, the hydrogen atom and atomic ions in which only one electron remains (He^+ , Li^{2+} etc) are the only atoms for which exact analytic solutions to the Schrödinger equation can be obtained. In no cases are exact solutions possible for molecules, even after the Born–Oppenheimer approximation (see [section B3.1.1.1](#)) is made to allow for separate treatment of electrons and nuclei. Despite the limited interest of hydrogen atoms and hydrogen-like ions to chemistry, the quantum mechanics of these systems is both highly instructive and provides a basis for treatments of more complex atoms and molecules. Comprehensive discussions of one-electron atoms can be found in many textbooks; the emphasis here is on qualitative aspects of the solutions.

The Schrödinger equation for a one-electron atom with nuclear charge Z is

$$\frac{-\hbar^2}{2\mu} \nabla^2 \psi - \frac{Ze^2}{r} \psi = E\psi \quad (\text{A1.1.62})$$

where μ is the reduced mass of the electron–nucleus system and the Laplacian is most conveniently expressed in spherical polar coordinates. While not trivial, this differential equation can be solved analytically. Some of the solutions are normalizable, and others are not. The former are those that describe the bound states of one-electron atoms, and can be written in the form

$$\psi_{nlm} = N R_{nl}(r) Y_{l,m_l}(\theta, \phi) \quad (\text{A1.1.63})$$

where N is a normalization constant, and $R_{nl}(r)$ and $Y_{l,m_l}(\theta, \phi)$ are specific functions that depend on the *quantum numbers* n , l and m_l . The first of these is called the principal quantum number, while l is known as the angular momentum, or azimuthal, quantum number, and m_l the magnetic quantum number. The quantum numbers that allow for normalizable wavefunctions are limited to integers that run over the ranges

-24-

$$n = 1, 2, 3, \dots \quad (\text{A1.1.64})$$

$$l = -n + 1, -n + 2, \dots, 0, 1, 2, \dots, n - 1 \quad (\text{A1.1.65})$$

$$m_l = -l, -l + 1, \dots, l - 1, l. \quad (\text{A1.1.66})$$

The fact that there is no restriction on n apart from being a positive integer means that there are an infinite number of bound-state solutions to the hydrogen atom, a peculiarity that is due to the form of the Coulomb potential. Unlike most bound state problems, the range of the potential is infinite (it goes to zero at large r , but diverges to negative infinity at $r = 0$). The eigenvalues of the Hamiltonian depend only on the principal

quantum number and are (in attojoules (10^{-18} J))

$$E_n = -2.18 \frac{Z^2}{n^2} \quad (\text{A1.1.67})$$

where it should be noted that the zero of energy corresponds to infinite separation of the particles. For each value of n , the Schrödinger equation predicts that all states are degenerate, regardless of the choice of l and m_l . Hence, any linear combination of wavefunctions corresponding to some specific value of n is also an eigenfunction of the Hamiltonian with eigenvalue E_n . States of hydrogen are usually characterized as ns , np , nd etc where n is the principal quantum number and s is associated with $l = 0$, p with $l = 1$ and so on. The functions $R_{nl}(r)$ describe the radial part of the wavefunctions and can all be written in the form

$$R_{nl}(r) = \exp(-\rho/2) \rho^l L_{nl}(\rho) \quad (\text{A1.1.68})$$

where ρ is proportional to the electron–nucleus separation r and the atomic number Z . L_{nl} is a polynomial of order $n - l - 1$ that has zeros (where the wavefunction, and therefore the probability of finding the electron, vanishes—a *radial node*) only for positive values of ρ . The functions $Y_{l,m_l}(\theta, \phi)$ are the *spherical harmonics*. The first few members of this series are familiar to everyone who has studied physical chemistry: $Y_{0,0}$ is a constant, leading to a spherically symmetric wavefunction, while $Y_{1,0}$, and specific linear combinations of $Y_{1,1}$ and $Y_{1,-1}$, vanish (have an *angular node*) in the xy , xz and yz planes, respectively. In general, these functions exhibit l nodes, meaning that the number of overall nodes corresponding to a particular ψ_{nlm_l} is equal to $n - 1$. For example, the 4d state has two angular nodes ($l = 2$) and one radial node ($L_{nl}(\rho)$ has one zero for positive ρ). In passing, it should be noted that many of the ubiquitous qualitative features of quantum mechanics are illustrated by the wavefunctions and energy levels of the hydrogen atom. First, the system has a zero-point energy, meaning that the ground-state energy is larger than the lowest value of the potential ($-\infty$) and the spacing between the energy levels decreases with increasing energy. Second, the ground state of the system is nodeless (the electron may be found at any point in space), while the number of nodes exhibited by the excited states increases with energy. Finally, there is a finite probability that the electron is found in a classically forbidden region in all bound states. For the hydrogen atom ground state, this corresponds to all electron–proton separations larger than 105.8 pm, where the electron is found 23.8% of the time. As usual, this tunnelling phenomenon is less pronounced in excited states: the corresponding values for the 3s state are 1904 pm and 16.0%.

The Hamiltonian commutes with the angular momentum operator \hat{L}_z as well as that for the square of the angular momentum \hat{L}^2 . The wavefunctions above are also eigenfunctions of these operators, with eigenvalues $m_l \hbar$ (\hat{L}_z) and $l(l+1)\hbar^2$ (\hat{L}^2). It should be emphasized that the total angular momentum is $L = \sqrt{l(l+1)}\hbar$, and not a simple

integral multiple of \hbar as assumed in the Bohr model. In particular, the ground state of hydrogen has zero angular momentum, while the Bohr atom ground state has $L = \hbar$. The meaning associated with the m_l quantum number is more difficult to grasp. The choice of z instead of x or y seems to be (and is) arbitrary and it is illogical that a specific value of the angular momentum projection along one coordinate must be observed in any experiment, while those associated with x and y are not similarly restricted. However, the states with a given l are degenerate, and the wavefunction at any particular time will in general be some linear combination of the m_l eigenfunctions. The only way to isolate a specific ψ_{nlm_l} (and therefore ensure the result of measuring L_z) is to apply a magnetic field that lifts the degeneracy and breaks the symmetry of the problem. The z axis

then corresponds to the magnetic field direction, and it is the projection of the angular momentum vector on this axis that must be equal to $m_j\hbar$.

The quantum-mechanical treatment of hydrogen outlined above does not provide a completely satisfactory description of the atomic spectrum, even in the absence of a magnetic field. Relativistic effects cause both a scalar shifting in all energy levels as well as splittings caused by the magnetic fields associated with both motion and intrinsic properties of the charges within the atom. The features of this *fine structure* in the energy spectrum were successfully (and miraculously, given that it preceded modern quantum mechanics by a decade and was based on a two-dimensional picture of the hydrogen atom) predicted by a formula developed by Sommerfeld in 1915. These interactions, while small for hydrogen, become very large indeed for larger atoms where very strong electron–nucleus attractive potentials cause electrons to move at velocities close to the speed of light. In these cases, quantitative calculations are extremely difficult and even the separability of orbital and intrinsic angular momenta breaks down.

A1.1.3.2 THE INDEPENDENT-PARTICLE APPROXIMATION

Applications of quantum mechanics to chemistry invariably deal with systems (atoms and molecules) that contain more than one particle. Apart from the hydrogen atom, the stationary-state energies cannot be calculated exactly, and compromises must be made in order to estimate them. Perhaps the most useful and widely used approximation in chemistry is the *independent-particle approximation*, which can take several forms. Common to all of these is the assumption that the Hamiltonian operator for a system consisting of n particles is approximated by the sum

$$\hat{H}_0 = \hat{h}_1 + \hat{h}_2 + \cdots + \hat{h}_n \quad (\text{A1.1.69})$$

where the single-particle Hamiltonians \hat{h}_i consist of the kinetic energy operator plus a potential ($\hat{H}_0 = \hat{h}_1 + \hat{h}_2 + \cdots + \hat{h}_n$) that does not explicitly depend on the coordinates of the other $n - 1$ particles in the system. Of course, the simplest realization of this model is to completely neglect forces due to the other particles, but this is often too severe an approximation to be useful. In any event, the quantum mechanics of a system described by a Hamiltonian of the form given by equation (A1.1.69) is worthy of discussion simply because the independent-particle approximation is the foundation for molecular orbital theory, which is the central paradigm of descriptive chemistry.

Let the orthonormal functions $\chi_i(1), \chi_j(2), \dots, \chi_p(n)$ be selected eigenfunctions of the corresponding single-particle Hamiltonians $\hat{h}_1, \hat{h}_2, \dots, \hat{h}_n$, with eigenvalues $\lambda_i, \lambda_j, \dots, \lambda_p$. It is easily verified that the product of these *single-particle wavefunctions* (which are often called *orbitals* when the particles are electrons in atoms and molecules)

-26-

$$\phi = \chi_i(1)\chi_j(2) \cdots \chi_p(n) \quad (\text{A1.1.70})$$

satisfies the approximate Schrödinger equation for the system

$$\hat{H}_0\phi = E_0\phi \quad (\text{A1.1.71})$$

with the corresponding energy

$$E_0 = \lambda_i + \lambda_j + \dots + \lambda_p. \quad (\text{A1.1.72})$$

Hence, if the Hamiltonian can be written as a sum of terms that individually depend only on the coordinates of one of the particles in the system, then the wavefunction of the system can be written as a product of functions, each of which is an eigenfunction of one of the single-particle Hamiltonians, h_i . The corresponding eigenvalue is then given by the sum of eigenvalues associated with each single-particle wavefunction χ appearing in the product.

The approximation embodied by [equation \(A1.1.69\)](#), [equation \(A1.1.70\)](#), [equation \(A1.1.71\)](#) and [equation \(A1.1.72\)](#) presents a conceptually appealing picture of many-particle systems. The behaviour and energetics of each particle can be determined from a simple function of three coordinates and the eigenvalue of a differential equation considerably simpler than the one that explicitly accounts for all interactions. It is precisely this simplification that is invoked in qualitative interpretations of chemical phenomena such as the inert nature of noble gases and the strongly reducing property of the alkali metals. The price paid is that the model is only approximate, meaning that properties predicted from it (for example, absolute ionization potentials rather than just trends within the periodic table) are not as accurate as one might like. However, as will be demonstrated in the latter parts of this section, a carefully chosen independent-particle description of a many-particle system provides a starting point for performing more accurate calculations. It should be mentioned that even qualitative features might be predicted incorrectly by independent-particle models in extreme cases. One should always be aware of this possibility and the oft-misunderstood fact that there really is no such thing as an orbital. Fortunately, however, it turns out that qualitative errors are uncommon for electronic properties of atoms and molecules when the best independent-particle models are used.

One important feature of many-particle systems has been neglected in the preceding discussion. Identical particles in quantum mechanics must be indistinguishable, which implies that the exact wavefunctions ψ which describe them must satisfy certain symmetry properties. In particular, interchanging the coordinates of any two particles in the mathematical form of the wavefunction cannot lead to a different prediction of the system properties. Since any rearrangement of particle coordinates can be achieved by successive pairwise permutations, it is sufficient to consider the case of a single permutation in analysing the symmetry properties that wavefunctions must obey. In the following, it will be assumed that the wavefunction is real. This is not restrictive, as stationary state wavefunctions for isolated atoms and molecules can always be written in this way. If the operator P_{ij} is that which permutes the coordinates of particles i and j , then indistinguishability requires that

$$\left(\int P_{ij}\psi \right)^* \hat{A} P_{ij}\psi \, d\tau = \int \psi^* \hat{A}\psi \, d\tau \quad (\text{A1.1.73})$$

-27-

for any operator \hat{A} (including the identity) and choice of i and j . Clearly, a wavefunction that is symmetric with respect to the interchange of coordinates for any two particles

$$P_{ij}\psi = \psi \quad (\text{A1.1.74})$$

satisfies the indistinguishability criterion. However, [equation \(A1.1.73\)](#) is also satisfied if the permutation of particle coordinates results in an overall sign change of the wavefunction, i.e.

$$P_{ij}\psi = -\psi. \quad (\text{A1.1.75})$$

Without further considerations, the only acceptable real quantum-mechanical wavefunctions for an n -particle system would appear to be those for which

$$P_{ij}\psi = \pm\psi \quad (\text{A1.1.76})$$

where i and j are any pair of identical particles. For example, if the system comprises two protons, a neutron and two electrons, the relevant permutations are that which interchanges the proton coordinates and that which interchanges the electron coordinates. The other possible pairs involve distinct particles and the action of the corresponding P_{ij} operators on the wavefunction will in general result in something quite different. Since indistinguishability is a necessary property of exact wavefunctions, it is reasonable to impose the same constraint on the approximate wavefunctions ϕ formed from products of single-particle solutions. However, if two or more of the χ_i in the product are different, it is necessary to form linear combinations if the condition $P_{ij}\psi = \pm\psi$ is to be met. An additional consequence of indistinguishability is that the h_i operators corresponding to identical particles must also be identical and therefore have precisely the same eigenfunctions. It should be noted that there is nothing mysterious about this perfectly reasonable restriction placed on the mathematical form of wavefunctions.

For the sake of simplicity, consider a system of two electrons for which the corresponding single-particle states are $\chi_i, \chi_j, \chi_k, \dots, \chi_n$, with eigenvalues $\lambda_i, \lambda_j, \lambda_k, \dots, \lambda_n$. Clearly, the two-electron wavefunction $\phi = \chi_i(1)\chi_j(2)$ satisfies the indistinguishability criterion and describes a stationary state with energy $E_0 = 2\lambda_i$. However, the state $\chi_i(1)\chi_j(2)$ is not satisfactory. While it is a solution to the Schrödinger equation, it is neither symmetric nor antisymmetric with respect to particle interchange. However, two such states can be formed by taking the linear combinations

$$\phi_S = \sqrt{\frac{1}{2}}[\chi_i(1)\chi_j(2) + \chi_i(2)\chi_j(1)] \quad (\text{A1.1.77})$$

$$\phi_A = \sqrt{\frac{1}{2}}[\chi_i(1)\chi_j(2) - \chi_i(2)\chi_j(1)] \quad (\text{A1.1.78})$$

which are symmetric and antisymmetric with respect to particle interchange, respectively. Because the functions χ are orthonormal, the energies calculated from ϕ_S and ϕ_A are the same as that corresponding to the unsymmetrized product state $\chi_i(1)\chi_j(2)$, as demonstrated explicitly for ϕ_S :

$$\begin{aligned} \int \phi_S \hat{H} \phi_S \, d\tau &= \frac{1}{2} \left[\int \chi_i(1)\chi_j(2) \hat{H} \chi_i(1)\chi_j(2) \, d\tau_1 \, d\tau_2 + \int \chi_i(1)\chi_j(2) \hat{H} \chi_i(2)\chi_j(1) \, d\tau_1 \, d\tau_2 \right. \\ &\quad \left. + \int \chi_i(2)\chi_j(1) \hat{H} \chi_i(1)\chi_j(2) \, d\tau_1 \, d\tau_2 + \int \chi_i(2)\chi_j(1) \hat{H} \chi_i(2)\chi_j(1) \, d\tau_1 \, d\tau_2 \right] \\ &= \frac{1}{2}(\lambda_i + \lambda_j) \left[\int \chi_i(1)\chi_i(1) \, d\tau_1 \int \chi_j(2)\chi_j(2) \, d\tau_2 + \int \chi_i(1)\chi_j(1) \, d\tau_1 \int \chi_j(2)\chi_i(2) \, d\tau_2 \right. \\ &\quad \left. + \int \chi_j(1)\chi_i(1) \, d\tau_1 \int \chi_i(2)\chi_j(2) \, d\tau_2 + \int \chi_j(1)\chi_j(1) \, d\tau_1 \int \chi_i(2)\chi_i(2) \, d\tau_2 \right] \\ &= \frac{1}{2}(\lambda_i + \lambda_j)[1 + 0 + 0 + 1] = \lambda_i + \lambda_j. \end{aligned} \quad (\text{A1.1.79})$$

It should be mentioned that the single-particle Hamiltonians in general have an infinite number of solutions, so that an uncountable number of wavefunctions ψ can be generated from them. Very often, interest is focused on the ground state of many-particle systems. Within the independent-particle approximation, this state can be represented by simply assigning each particle to the lowest-lying energy level. If a calculation is

performed on the lithium atom in which interelectronic repulsion is ignored completely, the single-particle Schrödinger equations are precisely the same as those for the hydrogen atom, apart from the difference in nuclear charge. The following lithium atom wavefunction could then be constructed from single-particle orbitals

$$\phi = N \chi_{1s}(1) \chi_{1s}(2) \chi_{1s}(3) \quad (\text{A1.1.80})$$

a form that is obviously symmetric with respect to interchange of particle coordinates. If this wavefunction is used to calculate the expectation value of the energy using the exact Hamiltonian (which includes the explicit electron–electron repulsion terms),

$$\epsilon = \int \psi^* H \psi \, d\tau \quad (\text{A1.1.81})$$

one obtains an energy lower than the actual result, which (see [\(A1.1.4.1\)](#)) suggests that there are serious problems with this form of the wavefunction. Moreover, a relatively simple analysis shows that ionization potentials of atoms would increase monotonically—approximately linearly for small atoms and quadratically for large atoms—if the independent-particle picture discussed thus far has any validity. Using a relatively simple model that assumes that the lowest lying orbital is a simple exponential, ionization potentials of 13.6, 23.1, 33.7 and 45.5 electron volts (eV) are predicted for hydrogen, helium, lithium and beryllium, respectively. The value for hydrogen (a one-electron system) is exact and that for helium is in relatively good agreement with the experimental value of 24.8 eV. However, the other values are well above the actual ionization energies of Li and Be (5.4 and 9.3 eV, respectively), both of which are smaller than those of H and He! All freshman chemistry students learn that ionization potentials do not increase monotonically with atomic number, and that there are in fact many pronounced and more subtle decreases that appear when this property is plotted as a function of atomic number.

There is evidently a grave problem here. The wavefunction proposed above for the lithium atom contains all of the particle coordinates, adheres to the boundary conditions (it decays to zero when the particles are removed to infinity) and obeys the restrictions $P_{12}\phi = P_{13}\phi = P_{23}\phi = \pm\phi$ that govern the behaviour of the exact wavefunctions. Therefore, if no other restrictions are placed on the wavefunctions of multiparticle systems, the product wavefunction for lithium

-29-

must lie in the space spanned by the exact wavefunctions. However, it clearly does not, because it is proven in [subsection \(A1.1.4.1\)](#) that any function expressible as a linear combination of Hamiltonian eigenfunctions cannot have an energy lower than that of the exact ground state. This means that there is at least one additional symmetry obeyed by all of the exact wavefunctions that is not satisfied by the product form given for lithium in [equation \(A1.1.80\)](#).

This missing symmetry provided a great puzzle to theorists in the early part days of quantum mechanics. Taken together, ionization potentials of the first four elements in the periodic table indicate that wavefunctions which assign two electrons to the same single-particle functions such as

$$\phi = \chi_a(1) \chi_a(2) \quad (\text{A1.1.82})$$

(helium) and

$$\phi = S\chi_a(1)\chi_a(2)\chi_b(3)\chi_b(4) \quad (\text{A1.1.83})$$

(beryllium, the operator \hat{S} produces the labelled $\chi_a\chi_a\chi_b\chi_b$ product that is symmetric with respect to interchange of particle indices) are somehow acceptable but that those involving three or more electrons in one state are not! The resolution of this *zweideutigkeit* (two-valuedness) puzzle was made possible only by the discovery of electron spin, which is discussed below.

A1.1.3.3 SPIN AND THE PAULI PRINCIPLE

In the early 1920s, spectroscopic experiments on the hydrogen atom revealed a striking inconsistency with the Bohr model, as adapted by Sommerfeld to account for relativistic effects. Studies of the fine structure associated with the $n = 4 \rightarrow n = 3$ transition revealed five distinct peaks, while six were expected from arguments based on the theory of interaction between matter and radiation. The problem was ultimately reconciled by Uhlenbeck and Goudsmit, who reinterpreted one of the quantum numbers appearing in Sommerfeld's fine structure formula based on a startling assertion that the electron has an intrinsic angular momentum independent of that associated with its motion. This idea was also supported by previous experiments of Stern and Gerlach, and is now known as *electron spin*. Spin is a mysterious phenomenon with a rather unfortunate name. Electrons are fundamental particles, and it is no more appropriate to think of them as charges that resemble extremely small billiard balls than as waves. Although they exhibit behaviour characteristic of both, they are in fact neither. Elementary textbooks often depict spin in terms of spherical electrons whirling about their axis (a compelling idea in many ways, since it reinforces the Bohr model by introducing a spinning planet), but this is a purely classical perspective on electron spin that should not be taken literally.

Electrons and most other fundamental particles have two distinct spin wavefunctions that are degenerate in the absence of an external magnetic field. Associated with these are two abstract states which are eigenfunctions of the intrinsic spin angular momentum operator \hat{S}_z

$$S_z\sigma = m_s\hbar\sigma. \quad (\text{A1.1.84})$$

-30-

The allowed quantum numbers m_s are $\frac{1}{2}$ and $-\frac{1}{2}$, and the corresponding eigenfunctions are usually written as α and β , respectively. The associated eigenvalues $\frac{\hbar}{2}$ and $-\frac{\hbar}{2}$ give the projection of the intrinsic angular momentum vector along the direction of a magnetic field that can be applied to resolve the degeneracy. The overall spin angular momentum of the electron is given in terms of the quantum number s by $\sqrt{s(s+1)}\hbar$. For an electron, $s = \frac{1}{2}$. For a collection of particles, the overall spin and its projection are given in terms of the spin quantum numbers S and M_S (which are equal to the corresponding lower-case quantities for single particles) by $\sqrt{S(S+1)}\hbar$ and $S = \frac{1}{2}$, respectively. S must be positive and can assume either integral or half-integral values, and the M_S quantum numbers lie in the interval

$$M_S = -S, -S+1, -S+2, \dots, 0, \dots, S-1, S \quad (\text{A1.1.85})$$

where a correspondence to the properties of orbital angular momentum should be noted. The *multiplicity* of a state is given by $2S+1$ (the number of possible M_S values) and it is customary to associate the terms *singlet* with $S=0$, *doublet* with $S = \frac{1}{2}$, *triplet* with $S=1$ and so on.

In the non-relativistic quantum mechanics discussed in this chapter, spin does not appear naturally. Although

Dirac showed in 1928 that a fourth quantum number associated with intrinsic angular momentum appears in a relativistic treatment of the free electron, it is customary to treat spin heuristically. In general, the wavefunction of an electron is written as the product of the usual spatial part (which corresponds to a solution of the non-relativistic Schrödinger equation and involves only the Cartesian coordinates of the particle) and a *spin part* σ , where σ is either α or β . A common shorthand notation is often used, whereby

$$\psi \equiv \psi_{\text{spatial}}\alpha \quad (\text{A1.1.86})$$

$$\bar{\psi} \equiv \psi_{\text{spatial}}\beta. \quad (\text{A1.1.87})$$

In the context of electronic structure theory, the composite functions above are often referred to as *spin orbitals*. When spin is taken into account, one finds that the ground state of the hydrogen atom is actually doubly degenerate. The spatial part of the wavefunction is the Schrödinger equation solution discussed in section (A1.1.3.1), but the possibility of either spin α or β means that there are two distinct overall wavefunctions. The same may be said for any of the excited states of hydrogen (all of which are, however, already degenerate in the nonrelativistic theory), as the level of degeneracy is doubled by the introduction of spin. Spin may be thought of as a fourth coordinate associated with each particle. Unlike Cartesian coordinates, for which there is a continuous distribution of possible values, there are only two possible values of the spin coordinate available to each particle. This has important consequences for our discussion of indistinguishability and symmetry properties of the wavefunction, as the concept of coordinate permutation must be amended to include the spin variable of the particles. As an example, the independent-particle ground state of the helium atom based on hydrogenic wavefunctions

$$\chi_{1s}(1)\chi_{1s}(2) \quad (\text{A1.1.88})$$

must be replaced by the four possibilities

$$\chi_{1s}(1)\chi_{1s}(2) \quad (\text{A1.1.89})$$

-31-

$$\bar{\chi}_{1s}(1)\chi_{1s}(2) \quad (\text{A1.1.90})$$

$$\chi_{1s}(1)\bar{\chi}_{1s}(2) \quad (\text{A1.1.91})$$

$$\bar{\chi}_{1s}(1)\bar{\chi}_{1s}(2). \quad (\text{A1.1.92})$$

While the first and fourth of these are symmetric with respect to particle interchange and thereby satisfy the indistinguishability criterion, the other two are not and appropriate linear combinations must be formed. Doing so, one finds the following four wavefunctions

$$\phi_{S1} = \chi_{1s}(1)\chi_{1s}(2) \quad (\text{A1.1.93})$$

$$\phi_{S2} = \sqrt{\frac{1}{2}}[\chi_{1s}(1)\bar{\chi}_{1s}(2) + \chi_{1s}(2)\bar{\chi}_{1s}(1)] \quad (\text{A1.1.94})$$

$$\phi_{S3} = \bar{\chi}_{1s}(1)\bar{\chi}_{1s}(2) \quad (\text{A1.1.95})$$

$$\phi_A = \sqrt{\frac{1}{2}}[\chi_{1s}(1)\bar{\chi}_{1s}(2) - \chi_{1s}(2)\bar{\chi}_{1s}(1)] \quad (\text{A1.1.96})$$

where the first three are symmetric with respect to particle interchange and the last is antisymmetric. This suggests that under the influence of a magnetic field, the ground state of helium might be resolved into components that differ in terms of overall spin, but this is not observed. For the lithium example, there are

eight possible ways of assigning the spin coordinates, only two of which

$$\phi = \chi_{1s}(1)\chi_{1s}(2)\chi_{1s}(3) \quad (\text{A1.1.97})$$

$$\phi = \bar{\chi}_{1s}(1)\bar{\chi}_{1s}(2)\bar{\chi}_{1s}(3) \quad (\text{A1.1.98})$$

satisfy the criterion $P_{ij}\phi = \pm\phi$. The other six must be mixed in appropriate linear combinations. However, there is an important difference between lithium and helium. In the former case, all assignments of the spin variable to the state given by [equation \(A1.1.88\)](#) produce a product function in which the same state (in terms of both spatial and spin coordinates) appears at least twice. A little reflection shows that it is not possible to generate a linear combination of such functions that is antisymmetric with respect to all possible interchanges; only symmetric combinations such as

$$\phi = \sqrt{\frac{1}{3}}[\chi_{1s}(1)\chi_{1s}(2)\bar{\chi}_{1s}(3) + \chi_{1s}(1)\bar{\chi}_{1s}(2)\chi_{1s}(3) + \bar{\chi}_{1s}(1)\chi_{1s}(2)\chi_{1s}(3)] \quad (\text{A1.1.99})$$

can be constructed. The fact that antisymmetric combinations appear for helium (where the independent-particle ground state made up of hydrogen 1s functions is qualitatively consistent with experiment) and not for lithium (where it is not) raises the interesting possibility that the exact wavefunction satisfies a condition more restrictive than $P_{ij}\psi = \pm\psi$, namely $P_{ij}\psi = -\psi$. For reasons that are not at all obvious, or even intuitive, nature does indeed enforce this restriction, which is one statement of the *Pauli exclusion principle*. When this idea is first met with, one usually learns an equivalent but less general statement that applies only within the independent-particle approximation: *no two electrons can have the same quantum numbers*. What does this mean? Within the independent-particle picture of an atom, each single-particle wavefunction, or orbital, is described by the quantum numbers n, l, m_l and (when spin is considered) m_s .

Since it is not possible to generate antisymmetric combinations of products if the same spin orbital appears twice in each term, it follows that states which assign the same set of four quantum numbers twice cannot possibly satisfy the requirement $P_{ij}\psi = -\psi$, so this statement of the exclusion principle is consistent with the more general symmetry requirement. An even more general statement of the exclusion principle, which can be regarded as an additional postulate of quantum mechanics, is

The wavefunction of a system must be antisymmetric with respect to interchange of the coordinates of identical particles γ and δ if they are *fermions*, and symmetric with respect to interchange of γ and δ if they are *bosons*.

Electrons, protons and neutrons and all other particles that have $s = \frac{1}{2}$ are known as fermions. Other particles are restricted to $s = 0$ or 1 and are known as bosons. There are thus profound differences in the quantum-mechanical properties of fermions and bosons, which have important implications in fields ranging from statistical mechanics to spectroscopic selection rules. It can be shown that the spin quantum number S associated with an even number of fermions must be integral, while that for an odd number of them must be half-integral. The resulting composite particles behave collectively like bosons and fermions, respectively, so the wavefunction symmetry properties associated with bosons can be relevant in chemical physics. One prominent example is the treatment of nuclei, which are typically considered as composite particles rather than interacting protons and neutrons. Nuclei with even atomic number therefore behave like individual bosons and those with odd atomic number as fermions, a distinction that plays an important role in rotational spectroscopy of polyatomic molecules.

A1.1.3.4 INDEPENDENT-PARTICLE MODELS IN ELECTRONIC STRUCTURE

At this point, it is appropriate to make some comments on the construction of approximate wavefunctions for the many-electron problems associated with atoms and molecules. The Hamiltonian operator for a molecule is given by the general form

$$\hat{H} = -\frac{\hbar^2}{2} \left[\sum_{\text{nuclei } \alpha} \frac{\nabla_{\alpha}^2}{M_{\alpha}} + \sum_{\text{electrons } i} \frac{\nabla_i^2}{m_e} \right] + \sum_{\text{nuclei } \alpha < \beta} \frac{Z_{\alpha} Z_{\beta} e^2}{r_{\alpha\beta}} + \sum_{\text{electrons } i < j} \frac{e^2}{r_{ij}} - \sum_i \sum_{\text{nuclei } \alpha} \frac{e Z_{\alpha}}{r_{i\alpha}}. \quad (\text{A1.1.100})$$

It should be noted that nuclei and electrons are treated equivalently in \hat{H} , which is clearly inconsistent with the way that we tend to think about them. Our understanding of chemical processes is strongly rooted in the concept of a potential energy surface which determines the forces that act upon the nuclei. The potential energy surface governs all behaviour associated with nuclear motion, such as vibrational frequencies, mean and equilibrium internuclear separations and preferences for specific conformations in molecules as complex as proteins and nucleic acids. In addition, the potential energy surface provides the transition state and activation energy concepts that are at the heart of the theory of chemical reactions. Electronic motion, however, is never discussed in these terms. All of the important and useful ideas discussed above derive from the Born–Oppenheimer approximation, which is discussed in some detail in [section B3.1](#). Within this model, the *electronic states* are solutions to the equation

-33-

$$\left[-\frac{\hbar^2}{2m} \sum_{\text{electrons } i} \nabla_i^2 - \sum_{\text{electrons } i} \sum_{\text{nuclei } \alpha} \frac{e Z_{\alpha}}{r_{i\alpha}} + \sum_{\text{electrons } i < j} \frac{e^2}{r_{ij}} \right] \psi = \lambda \psi \quad (\text{A1.1.101})$$

where the nuclei are assumed to be stationary. The *electronic energies* are given by the eigenvalues (usually augmented by the wavefunction-independent internuclear repulsion energy) of \hat{H} . The functions obtained by plotting the electronic energy as a function of nuclear position are the potential energy surfaces described above. The latter are different for every electronic state; their shape gives the usual information about molecular structure, barrier heights, isomerism and so on. The Born–Oppenheimer separation is also made in the study of electronic structure in atoms. However, this is a rather subtle point and is not terribly important in applications since the only assumption made is that the nucleus has infinite mass.

Although a separation of electronic and nuclear motion provides an important simplification and appealing qualitative model for chemistry, the electronic Schrödinger equation is still formidable. Efforts to solve it approximately and apply these solutions to the study of spectroscopy, structure and chemical reactions form the subject of what is usually called *electronic structure theory* or *quantum chemistry*. The starting point for most calculations and the foundation of molecular orbital theory is the independent-particle approximation.

For many-electron systems such as atoms and molecules, it is obviously important that approximate wavefunctions obey the same boundary conditions and symmetry properties as the exact solutions. Therefore, they should be antisymmetric with respect to interchange of each pair of electrons. Such states can always be constructed as linear combinations of products such as

$$\chi_i(1)\chi_j(2)\chi_k(3)\dots\chi_q(n). \quad (\text{A1.1.102})$$

The χ are assumed to be spin orbitals (which include both the spatial and spin parts) and each term in the product differs in the way that the electrons are assigned to them. Of course, it does not matter how the electrons are distributed amongst the χ in equation (A1.1.102), as the necessary subsequent antisymmetrization makes all choices equivalent apart from an overall sign (which has no physical significance). Hence, the product form is usually written without assigning electrons to the individual orbitals, and the set of unlabelled χ included in the product represents an *electron configuration*. It should be noted that all of the single-particle orbitals χ in the product are distinct. A very convenient method for constructing antisymmetrized combinations corresponding to products of particular single-particle states is to form the *Slater determinant*

$$\phi = \frac{1}{\sqrt{n!}} \begin{vmatrix} \chi_i(1) & \chi_i(2) & \chi_i(3) & \cdots & \chi_i(n) \\ \chi_j(1) & \chi_j(2) & \chi_j(3) & \cdots & \chi_j(n) \\ \chi_k(1) & \chi_k(2) & \chi_k(3) & \cdots & \chi_k(n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \chi_q(1) & \chi_q(2) & \chi_q(3) & \cdots & \chi_q(n) \end{vmatrix} \quad (\text{A1.1.103})$$

-34-

where the nominal electron configuration can be determined by simply scanning along the main diagonal of the matrix. A fundamental result of linear algebra is that the determinant of a matrix changes sign when any two rows or columns are interchanged. Inspection of [equation \(A1.1.103\)](#) shows that interchanging any two columns of the Slater determinant corresponds to interchanging the labels of two electrons, so the Pauli exclusion principle is automatically incorporated into this convenient representation. Whether all orbitals in a given row are identical and all particle labels the same in each column (as above) or *vice versa* is not important, as determinants are invariant with respect to transposition. In particular, it should be noted that the Slater determinant necessarily vanishes when two of the spin orbitals are identical, reflecting the alternative statement of the Pauli principle—no two electrons can have the same quantum number. One qualification which should be stated here is that Slater determinants are not necessarily eigenfunctions of the S^2 operator, and it is often advantageous to form linear combinations of those corresponding to electron configurations that differ only in the assignment of the spin variable to the spatial orbitals. The resulting functions ϕ are sometimes known as *spin-adapted configurations*.

Within an independent-particle picture, there are a very large number of single-particle wavefunctions χ available to each particle in the system. If the single-particle Schrödinger equations can be solved exactly, then there are often an infinite number of solutions. Approximate solutions are, however, necessitated in most applications, and some subtleties must be considered in this case. The description of electrons in atoms and molecules is often based on the Hartree–Fock approximation, which is discussed in [section B3.1](#) of this encyclopedia. In the Hartree–Fock method, only briefly outlined here, the orbitals are chosen in such a way that the total energy of a state described by the Slater determinant that comprises them is minimized. There are cogent reasons for using an energy minimization strategy that are based on the variational principle discussed later in this section. The Hartree–Fock method derives from an earlier treatment of Hartree, in which indistinguishability and the Pauli principle were ignored and the wavefunction expressed as in [equation \(A1.1.102\)](#). However, that approach is not satisfactory because it can lead to pathological solutions such as that discussed earlier for lithium. In Hartree–Fock theory, the orbital optimization is achieved at the expense of introducing a very complicated single-particle potential term v_i . This potential depends on all of the other orbitals in which electrons reside, requires the evaluation of difficult integrals and necessitates a self-consistent (iterative) solution. The resulting one-electron Hamiltonian is known as the Fock operator, and it has (in principle) an infinite number of eigenfunctions; a subset of these are exactly the same as the χ that correspond to the occupied orbitals upon which it is parametrized. The resulting equations cannot be solved analytically; for atoms, exact solutions for the occupied orbitals can be determined by numerical methods, but

the infinite number of unoccupied functions are unknown apart from the fact that they must be orthogonal to the occupied ones. In molecular calculations, it is customary to assume that the orbitals χ can be written as linear combinations of a fixed set of N *basis functions*, where N is typically of the order of tens to a few hundred. Iterative solution of a set of matrix equations provides approximations for the orbitals describing the n electrons of the molecule and $N - n$ unoccupied orbitals.

The choice of basis functions is straightforward in atomic calculations. It can be demonstrated that all solutions to an independent-particle Hamiltonian have the symmetry properties of the hydrogenic wavefunctions. Each is, or can be written as, an eigenfunction of the \hat{L}_z and \hat{L}^2 operators and involves a radial part multiplied by a spherical harmonic. Atomic calculations that use basis sets (not all of them do) typically choose functions that are similar to those that solve the Schrödinger equation for hydrogen. If the complete set of hydrogenic functions is used, the solution to the basis set equations are the exact Hartree–Fock solutions. However, practical considerations require the use of finite basis sets; the corresponding solutions are therefore only approximate. Although the distinction is rarely made, it is preferable to refer to these as *self-consistent field* (SCF) solutions and energies in order to distinguish them from the exact Hartree–Fock results. As the quality of a basis is improved, the energy approaches that of the Hartree–Fock solution from above.

-35-

In molecules, things are a great deal more complicated. In principle, one can always choose a subset of all the hydrogenic wavefunctions centred at some point in space. Since the resulting basis functions include all possible electron coordinates and Slater determinants constructed from them vanish at infinity and satisfy the Pauli principle, the corresponding approximate solutions must lie in the space spanned by the exact solutions and be qualitatively acceptable. In particular, use of enough basis functions will result in convergence to the exact Hartree–Fock solution. Because of the difficulties involved with evaluating integrals involving exponential hydrogenic functions centred at more than one point in space, such *single-centre expansions* were used in the early days of quantum chemistry. The main drawback is that convergence to the exact Hartree–Fock result is extraordinarily slow. The states of the hydrogen molecule are reasonably well approximated by linear combinations of hydrogenic functions centred on each of the two nuclei. Hence, a more practical strategy is to construct a basis by choosing a set of hydrogenic functions for each atom in the molecule (the same functions are usually used for identical atoms, whether or not they are equivalent by symmetry). Linear combinations of a relatively small number of these functions are capable of describing the electronic distribution in molecules much better than is possible with a corresponding number of functions in a single-centre expansion. This approach is often called the *linear combination of atomic orbitals* (LCAO) approximation, and is used in virtually all molecular SCF calculations performed today. The problems associated with evaluation of multicentre integrals alluded to above was solved more than a quarter-century ago by the introduction of Gaussian—rather than exponential—basis functions, which permit all of the integrals appearing in the Fock operator to be calculated analytically. Although Gaussian functions are not hydrogenic functions (and are inferior basis functions), the latter can certainly be approximated well by linear combinations of the former. The ease of integral evaluation using Gaussian functions makes them the standard choice for practical calculations. The importance of selecting an appropriate basis set is of great practical importance in quantum chemistry and many other aspects of atomic and molecular quantum mechanics. An illustrative example of basis set selection and its effect on calculated energies is given in [subsection \(A1.1.4.2\)](#). While the problem studied there involves only the motion of a single particle in one dimension, an analogy with the LCAO and single-centre expansion methods should be apparent, with the desirable features of the former clearly illustrated.

Even Hartree–Fock calculations are difficult and expensive to apply to large molecules. As a result, further simplifications are often made. Parts of the Fock operator are ignored or replaced by parameters chosen by some sort of statistical procedure to account, in an average way, for the known properties of selected

compounds. While calculating properties that have already been measured experimentally is of limited interest to anyone other than theorists trying to establish the accuracy of a method, the hope of these approximate Hartree–Fock procedures (which include the well known Hückel approximation and are collectively known as *semiempirical methods*) is that the parametrization works just as well for both unmeasured properties of known molecules (such as transition state structures) and the structure and properties of transient or unknown species. No further discussion of these approaches is given here (more details are given in [section B3.1](#) and [section B3.2](#)); it should only be emphasized that all of these methods are based on the independent-particle approximation.

Regardless of how many single-particle wavefunctions χ are available, this number is overwhelmed by the number of n -particle wavefunctions ϕ (Slater determinants) that can be constructed from them. For example, if a two-electron system is treated within the Hartree–Fock approximation using 100 basis functions, both of the electrons can be assigned to any of the χ obtained in the calculation, resulting in 10,000 two-electron wavefunctions. For water, which has ten electrons, the number of electronic wavefunctions with equal numbers of α and β spin electrons that can be constructed from 100 single-particle wavefunctions is roughly 10^{15} ! The significance of these other solutions may be hard to grasp. If one is interested solely in the electronic ground state and its associated potential energy surface (the focus of investigation in most quantum chemistry studies), these solutions play no role whatsoever within the HF–SCF approximation. Moreover, one might think (correctly) that solutions obtained by putting an electron in one of the

-36-

unoccupied orbitals offers a poor treatment of excited states since only the occupied orbitals are optimized. However, there is one very important feature of the extra solutions. If the HF solution has been obtained and all (an infinite number) of virtual orbitals available, then the basic principles of quantum mechanics imply that the exact wavefunction can be written as the sum of Slater determinants

$$\psi_{\text{exact}} = \sum_k c_k \phi_k \tag{A1.1.104}$$

where the ϕ_k correspond to all possible electron configurations. The individual Slater determinants are thus seen to play a role in the representation of the exact wavefunction that is analogous to that played by the hydrogenic (or LCAO) functions in the expansion of the Hartree–Fock orbitals. The Slater determinants are sometimes said to form an *n-electron basis*, while the hydrogenic (LCAO) functions are the *one-electron basis*.

A similar expansion can be made in practical finite-basis calculations, except that limitations of the basis set preclude the possibility that the exact wavefunction lies in the space spanned by the available ϕ . However, it should be clear that the formation of linear combinations of the finite number of ϕ_k offers a way to better approximate the exact solution. In fact, it is possible to obtain by this means a wavefunction that exactly satisfies the electronic Schrödinger equation when the assumption is made that the solution must lie in the space spanned by the n -electron basis functions ϕ . However, even this is usually impossible, and only a select number of the ϕ_k are used. The general principle of writing n -electron wavefunctions as linear combinations of Slater determinants is known as *configuration interaction*, and the resultant improvement in the wavefunction is said to account for *electron correlation*. The origin of this term is easily understood. Returning to helium, an inspection of the Hartree–Fock wavefunction

$$\psi = \sqrt{\frac{1}{2}} [\chi_{1s}(1)\bar{\chi}_{1s}(2) - \chi_{1s}(2)\bar{\chi}_{1s}(1)] \tag{A1.1.105}$$

exhibits some rather unphysical behaviour: the probability of finding one electron at a particular point in space is entirely independent of where the other electron is! In particular, the probability does not vanish when the two particles are coincident, the associated singularity in the interelectronic repulsion potential notwithstanding. Of course, electrons do not behave in this way, and do indeed tend to avoid each other. Hence, their motion is correlated, and this qualitative feature is absent from the Hartree–Fock approximation when the electrons have different spins. When they are of like spin, then the implicit incorporation of the Pauli principle into the form of the Slater determinant allows for some measure of correlation (although these like-spin effects are characteristically overestimated) since the wavefunction vanishes when the coordinates of the two electrons coincide. Treatments of electron correlation and the related concept of *correlation energy* (the difference between the Hartree–Fock and exact non-relativistic results) take a number of different forms that differ in the strategies used to determine the expansion coefficients c_k and the energy (which is not always given by the expectation value of the Hamiltonian over a function of the form equation (A1.1.104)). The basic theories underlying the most popular choices are the variational principle and perturbation theory, which are discussed in a general way in the remainder of this section. Specific application of these tools in electronic structure theory is dealt with in [section B3.1](#). Before leaving this discussion, it should also be mentioned that a concept very similar to the independent-particle approximation is used in the quantum-mechanical treatment of molecular vibrations. In that case, it is always possible to solve the Schrödinger equation for nuclear motion exactly if the potential energy function is assumed to be quadratic.

-37-

The corresponding functions χ_i, χ_j etc. then define what are known as the *normal coordinates* of vibration, and the Hamiltonian can be written in terms of these in precisely the form given by [equation \(A1.1.69\)](#), with the caveat that each term refers not to the coordinates of a single particle, but rather to independent coordinates that involve the collective motion of many particles. An additional distinction is that treatment of the vibrational problem does not involve the complications of antisymmetry associated with identical fermions and the Pauli exclusion principle. Products of the normal coordinate functions nevertheless describe all vibrational states of the molecule (both ground and excited) in very much the same way that the product states of single-electron functions describe the electronic states, although it must be emphasized that one model is based on independent motion and the other on collective motion, which are qualitatively very different. Neither model faithfully represents reality, but each serves as an extremely useful conceptual model and a basis for more accurate calculations.

A1.1.4 APPROXIMATING EIGENVALUES OF THE HAMILTONIAN

Since its eigenvalues correspond to the allowed energy states of a quantum-mechanical system, the time-independent Schrödinger equation plays an important role in the theoretical foundation of atomic and molecular spectroscopy. For cases of chemical interest, the equation is always easy to write down but impossible to solve exactly. Approximation techniques are needed for the application of quantum mechanics to atoms and molecules. The purpose of this subsection is to outline two distinct procedures—the variational principle and perturbation theory—that form the theoretical basis for most methods used to approximate solutions to the Schrödinger equation. Although some tangible connections are made with ideas of quantum chemistry and the independent-particle approximation, the presentation in the next two sections (and example problem) is intended to be entirely general so that the scope of applicability of these approaches is not underestimated by the reader.

A1.1.4.1 THE VARIATIONAL PRINCIPLE

Although it may be impossible to solve the Schrödinger equation for a specific choice of the Hamiltonian, it is

always possible to guess! While randomly chosen functions are unlikely to be good approximations to the exact quantum-mechanical wavefunction, an educated guess can usually be made. For example, if one is interested in the ground state of a single particle subjected to some potential energy function, the qualitative features discussed in [subsection \(A1.1.2.3\)](#) can be used as a guide in constructing a guess. Specifically, an appropriate choice would be one that decays to zero at positive and negative infinity, has its largest values in regions where the potential is deepest, and has no nodes. For more complicated problems—especially those involving several identical particles—it is not so easy to intuit the form of the wavefunction. Nevertheless, guesses can be based on solutions to a (perhaps grossly) simplified Schrödinger equation, such as the Slater determinants associated with independent-particle models.

In general, approaches based on guessing the form of the wavefunction fall into two categories. In the first, the ground-state wavefunction is approximated by a function that contains one or more nonlinear parameters. For example, if $\exp(ax)$ is a solution to a simplified Schrödinger equation, then the function $\exp(bx)$ provides a plausible guess for the actual problem. The parameter b can then be varied to obtain the most accurate description of the exact ground state. However, there is an apparent contradiction here. If the exact ground-state wavefunction and energy are not known (and indeed impossible to obtain analytically), then how is one to determine the best choice for the parameter b ?

-38-

The answer to the question that closes the preceding paragraph is the essence of the *variational principle* in quantum mechanics. If a guessed or *trial wavefunction* ϕ is chosen, the energy ϵ obtained by taking the expectation value of the Hamiltonian (it must be emphasized that the actual Hamiltonian is used to evaluate the expectation value, rather than the approximate Hamiltonian that may have been used to generate the form of the trial function) over ϕ must be higher than the exact ground-state energy. It seems rather remarkable that the mathematics seems to know precisely where the exact eigenvalue lies, even though the problem cannot be solved exactly. However, it is not difficult to prove that this assertion is true. The property of mathematical completeness tells us that our trial function can be written as a linear combination of the exact wavefunctions (so long as our guess obeys the boundary conditions and fundamental symmetries of the problem), even when the latter cannot be obtained. Therefore one can always write

$$\phi = \sum_k c_k \psi_k \tag{A1.1.106}$$

where ψ_k is the exact Hamiltonian eigenfunction corresponding to eigenvalue λ_k , and ordered so that $\lambda_0 \leq \lambda_1 \leq \lambda_2 \dots$. Assuming normalization of both the exact wavefunctions and the trial function, the expectation value of the Hamiltonian is

$$\begin{aligned} \epsilon &= \int \phi^* H \phi \, d\tau \\ &= \int \left(\sum_j c_j^* \psi_j^* \right) H \left(\sum_k c_k \psi_k \right) d\tau = \sum_k \sum_j c_j^* c_k \int \psi_j^* H \psi_k. \end{aligned} \tag{A1.1.107}$$

Since the ψ_k represent exact eigenfunctions of the Hamiltonian, equation (A1.1.107) simplifies to

$$\epsilon = \sum_j \sum_k c_j^* c_k \lambda_k \int \psi_j^* \psi_k = \sum_k c_k^* c_k \lambda_k = \sum_k |c_k|^2 \lambda_k. \tag{A1.1.108}$$

The assumption of normalization imposed on the trial function means that

$$\sum_k |c_k|^2 = 1 \quad (\text{A1.1.109})$$

hence

$$|c_0|^2 = 1 - |c_1|^2 - |c_2|^2 - |c_3|^2 - \dots \quad (\text{A1.1.110})$$

Inserting equation (A1.1.110) into equation (A1.1.108) yields

$$\epsilon = \lambda_0 + |c_1|^2(\lambda_1 - \lambda_0) + |c_2|^2(\lambda_2 - \lambda_0) + \dots \quad (\text{A1.1.111})$$

The first term on the right-hand side of the equation for ϵ is the exact ground-state energy. All of the remaining contributions involve norms of the expansion coefficients and the differences $\lambda_k - \lambda_0$, both of which must be either positive or zero. Therefore, ϵ is equal to the ground-state energy plus a number that cannot be negative. In the case where the trial function is precisely equal to the ground-state wavefunction, then $\epsilon = \lambda_0$; otherwise $\epsilon > \lambda_0$. Hence, the expectation value of the Hamiltonian with respect to any arbitrarily chosen trial function provides an upper bound to the exact ground-state energy. The dilemma raised earlier—how to define the best value of the variational parameter b —has a rather straightforward answer, namely the choice that minimizes the value of ϵ , known as the *variational energy*.

A concrete example of the variational principle is provided by the Hartree–Fock approximation. This method asserts that the electrons can be treated independently, and that the n -electron wavefunction of the atom or molecule can be written as a Slater determinant made up of orbitals. These orbitals are defined to be those which minimize the expectation value of the energy. Since the general mathematical form of these orbitals is not known (especially in molecules), then the resulting problem is highly nonlinear and formidably difficult to solve. However, as mentioned in [subsection \(A1.1.3.2\)](#), a common approach is to assume that the orbitals can be written as linear combinations of one-electron basis functions. If the basis functions are fixed, then the optimization problem reduces to that of finding the best set of coefficients for each orbital. This tremendous simplification provided a revolutionary advance for the application of the Hartree–Fock method to molecules, and was originally proposed by Roothaan in 1951. A similar form of the trial function occurs when it is assumed that the exact (as opposed to Hartree–Fock) wavefunction can be written as a linear combination of Slater determinants (see [equation \(A1.1.104\)](#)). In the conceptually simpler latter case, the objective is to minimize an expression of the form

$$\epsilon = \int \phi^* \hat{H} \phi \, d\tau \quad (\text{A1.1.112})$$

where ϕ is parametrized as

$$\phi = \sum_{k=0}^N c_k \chi_k \quad (\text{A1.1.113})$$

and both the (fixed functions) χ_k and ϕ are assumed to be normalized.

The representation of trial functions as linear combinations of fixed basis functions is perhaps the most common approach used in variational calculations; optimization of the coefficients c_k is often said to be an application of the *linear variational principle*. Although some very accurate work on small atoms (notably helium and lithium) has been based on complicated trial functions with several nonlinear parameters, attempts to extend these calculations to larger atoms and molecules quickly runs into formidable difficulties (not the least of which is how to choose the form of the trial function). Basis set expansions like that given by equation (A1.1.113) are much simpler to design, and the procedures required to obtain the coefficients that minimize ϵ are all easily carried out by computers.

-40-

For the example discussed above, where $\exp(ax)$ is the solution to a simpler problem, a trial function using five basis functions

$$\phi = c_1 e^{(a-2)x} + c_2 e^{(a-1)x} + c_3 e^{ax} + c_4 e^{(a+1)x} + c_5 e^{(a+2)x} \quad (\text{A1.1.114})$$

could be used instead of $\exp(bx)$ if the exact function is not expected to deviate too much from $\exp(ax)$. What is gained from replacing a trial function containing a single parameter by one that contains five? To see, consider the problem of how coefficients can be chosen to minimize the variational energy ϵ ,

$$\epsilon = \frac{\int \phi^* \hat{H} \phi \, d\tau}{\int \phi^* \phi \, d\tau}. \quad (\text{A1.1.115})$$

The denominator is included in equation (A1.1.115) because it is impossible to ensure that the trial function is normalized for arbitrarily chosen coefficients c_k . In order to minimize the value of ϵ for the trial function

$$\phi = \sum_{k=0}^N c_k \chi_k \quad (\text{A1.1.116})$$

it is necessary (but not sufficient) that its first partial derivatives with respect to all expansion coefficients vanish, *viz*

$$\frac{\partial \epsilon}{\partial c_1} = \frac{\partial \epsilon}{\partial c_2} = \frac{\partial \epsilon}{\partial c_3} = \dots = 0. \quad (\text{A1.1.117})$$

It is worthwhile, albeit tedious, to work out the condition that must be satisfied in order for equation (A1.1.117) to hold true. Expanding the trial function according to [equation \(A1.1.113\)](#), assuming that the basis functions and expansion coefficients are real and making use of the technique of implicit differentiation, one finds

$$\begin{aligned} \frac{\partial \epsilon}{\partial c_k} \left[\sum_{i=0}^N \sum_{j=0}^N c_i c_j S_{ij} \right] + 2\epsilon \left[\sum_{j=0}^N c_j S_{jk} \right] &= 2 \sum_{j=0}^N c_j H_{jk} \\ \frac{\partial \epsilon}{\partial c_k} \left[\sum_{i=0}^N \sum_{j=0}^N c_i c_j S_{ij} \right] &= 2 \sum_{j=0}^N c_j (H_{jk} - \epsilon S_{jk}) \end{aligned} \quad (\text{A1.1.118})$$

where shorthand notations for the *overlap matrix elements*

$$S_{jk} \equiv \int \chi_j \chi_k \, d\tau \quad (\text{A1.1.119})$$

-41-

and *Hamiltonian matrix elements*

$$H_{jk} \equiv \int \chi_j H \chi_k \, d\tau \quad (\text{A1.1.120})$$

have been introduced. Since the term multiplying the derivative of the expansion coefficient is simply the norm of the wavefunction, the variational condition [equation \(A1.1.117\)](#) is satisfied if the term on the right-hand side of [equation \(A1.1.118\)](#) vanishes for all values of k . Specifically, the set of homogeneous linear equations corresponding to the matrix expression

$$(c_1 \ c_2 \ \cdots \ c_N) \begin{pmatrix} H_{00} - \epsilon S_{00} & H_{01} - \epsilon S_{01} & \cdots & H_{0N} - \epsilon S_{0N} \\ H_{10} - \epsilon S_{10} & H_{11} - \epsilon S_{11} & \cdots & H_{1N} - \epsilon S_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ H_{N0} - \epsilon S_{N0} & H_{N1} - \epsilon S_{N1} & \cdots & H_{NN} - \epsilon S_{NN} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad (\text{A1.1.121})$$

must be satisfied. It is a fundamental principle of linear algebra that systems of equations of this general type are satisfied only for certain choices of ϵ , namely those for which the *determinant*

$$\begin{vmatrix} H_{00} - \epsilon S_{00} & H_{01} - \epsilon S_{01} & \cdots & H_{0N} - \epsilon S_{0N} \\ H_{10} - \epsilon S_{10} & H_{11} - \epsilon S_{11} & \cdots & H_{1N} - \epsilon S_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ H_{N0} - \epsilon S_{N0} & H_{N1} - \epsilon S_{N1} & \cdots & H_{NN} - \epsilon S_{NN} \end{vmatrix} \quad (\text{A1.1.122})$$

is identically equal to zero. There are precisely N values of ϵ that satisfy this condition, some of which might be degenerate, and their determination constitutes what is known as the generalized eigenvalue problem. While this is reasonably well suited to computation, a further simplification is usually made. When suited to the problem under consideration, the basis functions are usually chosen to be members of an orthonormal set. In other cases (for example, in the LCAO treatment of molecules) where this is not possible, the original basis functions χ'_k corresponding to the overlap matrix \mathbf{S}' can be subjected to the orthonormalizing transformation

$$\chi_k = \sum_l \chi'_l X_{lk} \quad (\text{A1.1.123})$$

where \mathbf{X} is the reciprocal square root of the overlap matrix in the primed basis,

$$\mathbf{X} \equiv \mathbf{S}'^{-1/2}. \quad (\text{A1.1.124})$$

The simplest way to obtain \mathbf{X} is to diagonalize \mathbf{S}' , take the reciprocal square roots of the eigenvalues and then transform the matrix back to its original representation, i.e.

$$\mathbf{X} = \mathbf{C}' \mathbf{s}^{-1/2} \mathbf{C}'^\dagger \quad (\text{A1.1.125})$$

where \mathbf{s} is the diagonal matrix of reciprocal square roots of the eigenvalues, and \mathbf{C}' is the matrix of eigenvectors for the original \mathbf{S}' matrix. Doing this, one finds that the transformed basis functions are orthonormal. In terms of implementation, elements of the Hamiltonian are usually first evaluated in the primed basis, and the resulting *matrix representation* of \mathbf{H} is then transformed to the orthogonal basis ($\mathbf{H} = \mathbf{X}^\dagger \mathbf{H}' \mathbf{X}$).

In an orthonormal basis, $S_{kj} = 1$ if $k = j$, and vanishes otherwise. The problem of finding the variational energy of the ground state then reduces to that of determining the smallest value of ϵ that satisfies

$$\begin{vmatrix} H_{00} - \epsilon & H_{01} & \cdots & H_{0N} \\ H_{10} & H_{11} - \epsilon & \cdots & H_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ H_{N0} & H_{N1} & \cdots & H_{NN} - \epsilon \end{vmatrix} = 0 \quad (\text{A1.1.126})$$

a task that modern digital computers can perform very efficiently. Given an orthonormal basis, the variational problem can be solved by *diagonalizing* the matrix representation of the Hamiltonian, \mathbf{H} . Associated with each eigenvalue ϵ is an eigenvector ($c_0, c_1, c_2, \dots, c_N$) that tells how the basis functions are combined in the corresponding approximate wavefunction ϕ as parametrized by [equation \(A1.1.116\)](#). That the lowest eigenvalue ϵ of \mathbf{H} provides an upper bound to the exact ground-state energy has already been proven; it is also true (but will not be proved here) that the first excited state of the actual system must lie below the next largest eigenvalue λ_1 , and indeed all remaining eigenvalues provide upper bounds to the corresponding excited states. That is,

$$\epsilon_0 \geq \lambda_0, \epsilon_1 \geq \lambda_1, \epsilon_2 \geq \lambda_2, \dots, \epsilon_N \geq \lambda_N. \quad (\text{A1.1.127})$$

The equivalence between variational energies and the exact eigenvalues of the Hamiltonian is achieved only in the case where the corresponding exact wavefunctions can be written as linear combinations of the basis functions. Suppose that the Schrödinger equation for the problem of interest cannot be solved, but another simpler problem that involves precisely the same set of coordinates lends itself to an analytic solution. In practice, this can often be achieved by ignoring certain interaction terms in the Hamiltonian, as discussed earlier. Since the eigenfunctions of the simplified Hamiltonian form a complete set, they provide a conceptually useful basis since all of the eigenfunctions of the intractable Hamiltonian can be written as linear combinations of them (for example, Slater determinants for electrons or products of normal mode wavefunctions for vibrational states). In this case, diagonalization of \mathbf{H} in this basis of functions provides an exact solution to the Schrödinger equation. It is worth pausing for a moment to analyse what is meant by this rather remarkable statement. One simply needs to ignore interaction terms in the Hamiltonian that preclude an analytic determination of the stationary states and energies of the system. The corresponding Schrödinger equation can then be solved to provide a set of orthonormal basis functions, and the integrals that represent the matrix elements of \mathbf{H}

$$H_{ij} = \int \chi_i^* H \chi_j \, d\tau \quad (\text{A1.1.128})$$

computed. Diagonalization of the resulting matrix provides the sought-after solution to the quantum-mechanical problem. Although this process replaces an intractable differential equation by a problem in linear algebra, the latter offers its own insurmountable hurdle: the *dimension of the matrix* (equal to the number of rows or columns) is equal to the number of functions included in the complete set of solutions to the simplified Schrödinger equation. Regrettably, this number is usually infinite. At present, special algorithms can be used with modern computers to obtain eigenvalues of matrices with dimensions of about 100 million relatively routinely, but this still falls far short of infinity. Therefore, while it seems attractive (and much simpler) to do away with the differential equation in favour of a matrix diagonalization, it is not a magic bullet that makes exact quantum-mechanical calculations a possibility.

In order to apply the linear variational principle, it is necessary to work with a matrix sufficiently small that it can be diagonalized by a computer; such calculations are said to employ a *finite basis*. Use of a finite basis means that the eigenvalues of \mathbf{H} are not exact unless the basis chosen for the problem has the miraculous (and extraordinarily unlikely) property of being sufficiently flexible to allow one or more of the exact solutions to be written as linear combinations of them. For example, if the intractable system Hamiltonian contains only small interaction terms that are ignored in the simplified Hamiltonian used to obtain the basis functions, then χ_0 is probably a reasonably good approximation to the exact ground-state wavefunction. At the very least, one can be relatively certain that it is closer to ψ_0 than are those that correspond to the thousandth, millionth and billionth excited states of the simplified system. Hence, if the objective of a variational calculation is to determine the ground-state energy of the system, it is important to include χ_0 and other solutions to the simplified problem with relatively low lying energies, while $\chi_{1\ 000\ 000}$ and other high lying solutions can be excluded more safely.

A1.1.4.2 EXAMPLE PROBLEM: THE DOUBLE-WELL OSCILLATOR

To illustrate the use of the variational principle, results are presented here for calculations of the five lowest energy states (the ground state and the first four excited states) of a particle subject to the potential

$$V(q) = 0.05q^4 - q^2 \quad (\text{A1.1.129})$$

which is shown in [figure A1.1.3](#). The potential goes asymptotically to infinity (like that for the harmonic oscillator), but exhibits two symmetrical minima at $q = \pm\sqrt{10}$, and a maximum at the origin. This function is known as a double well, and provides a qualitative description of the potential governing a variety of quantum-mechanical processes, such as motion involved in the inversion mode of ammonia (where the minima play the role of the two equivalent pyramidal structures and the maximum that of planar NH_3). For simplicity, the potential is written in terms of the dimensionless coordinate q defined by

$$q \equiv \alpha x \equiv \left(\frac{mk}{\hbar^2} \right)^{\frac{1}{4}} x \quad (\text{A1.1.130})$$

where x is a Cartesian displacement and k is a constant with units of $(\text{mass})(\text{time})^{-2}$. The corresponding Schrödinger

equation can be written as

$$\left[-\frac{1}{2} \frac{d}{dq^2} + 0.05q^4 - q^2 \right] \psi = E\psi \quad (\text{A1.1.131})$$

where the energy is given as a multiple of $\hbar^2\alpha^2/m$. This value corresponds to $h\nu$ where ν is the frequency corresponding to a quantum harmonic oscillator with force constant k .

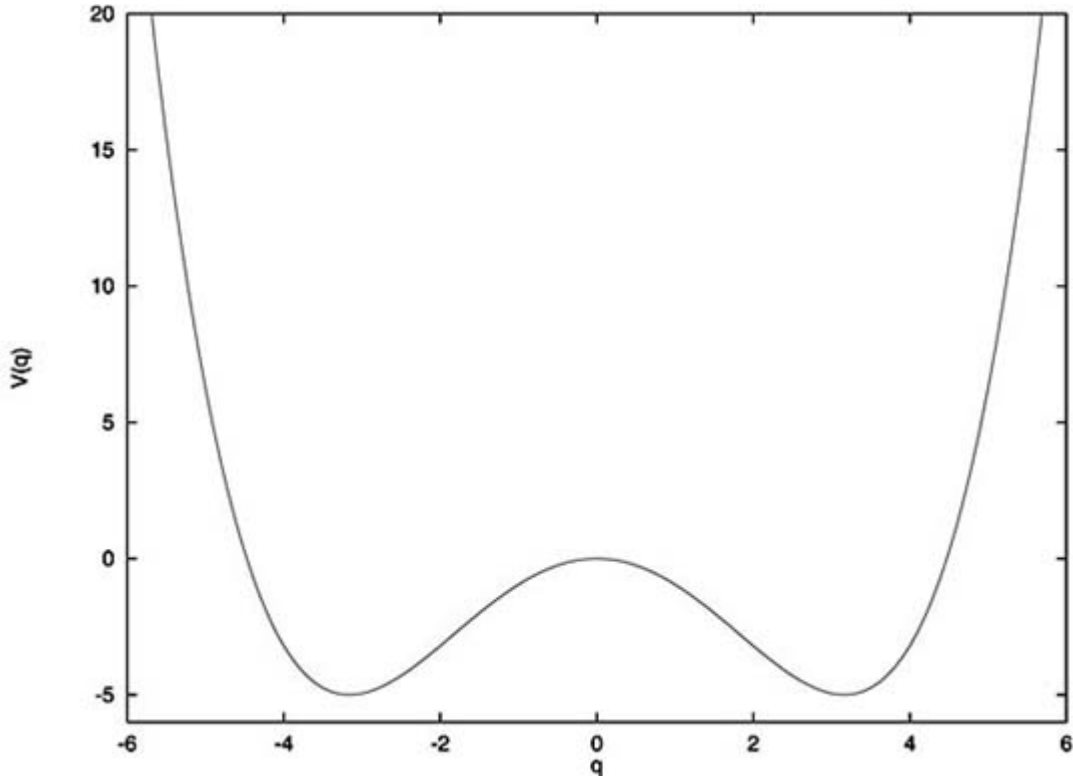


Figure A1.1.3. Potential function used in the variational calculations. Note that the energies of all states lie above the lowest point of the potential ($V = -5$), which occurs at $q = \pm\sqrt{10}$.

It is not possible to solve this equation analytically, and two different calculations based on the linear variational principle are used here to obtain the approximate energy levels for this system. In the first, eigenfunctions corresponding to the potential $V = 2q^2$ (this corresponds to the shape of the double-well potential in the vicinity of its minima) are used as a basis. It should be noted at the outset that these functions form a complete set, and it is therefore possible to write exact solutions to the double-well oscillator problem in terms of them. However, since we expect the ground-state wavefunction to have maximum amplitude in the regions around $q = \pm\sqrt{10}$, it is unlikely that the first few harmonic oscillator functions (which have maxima closer to the origin) are going to provide a good representation of the exact ground state. The first four eigenvalues of the potential are given in the table below, where N indicates

the size of the variational basis which includes the N lowest energy harmonic oscillator functions centred at the origin.

N	λ_1	λ_2	λ_3	λ_4
2	0.259 37	0.796 87	—	—
4	-0.467 37	-0.358 63	-1.989 99	-3.248 62
6	-1.414 39	-1.051 71	-0.689 35	-1.434 57
8	-2.225 97	-1.850 67	-0.097 07	-0.396 14
10	-2.891 74	-2.580 94	-0.358 57	-0.339 30
20	-4.021 22	-4.012 89	-2.162 21	-2.125 38
30	-4.026 63	-4.026 60	-2.204 11	-2.200 79
40	-4.026 63	-4.026 60	-2.204 11	-2.200 79
50	-4.026 63	-4.026 60	-2.204 11	-2.200 79

Note that the energies decrease with increasing size of the basis set, as expected from the variational principle. With 30 or more functions, the energies of the four states are well converged (to about one part in 100,000). In [figure A1.1.4](#) the wavefunctions of the ground and first excited states of the system calculated with 40 basis functions are shown. As expected, the probability density is localized near the symmetrically disposed minima on the potential. The ground state has no nodes and the first excited state has a single node. The ground-state wavefunction calculated with only eight basis functions (shown in [figure A1.1.5](#) is clearly imperfect. The rapid oscillations in the wavefunction are not real, but rather an artifact of the incomplete basis used in the calculation. A larger number of functions is required to reduce the amplitude of the oscillations.

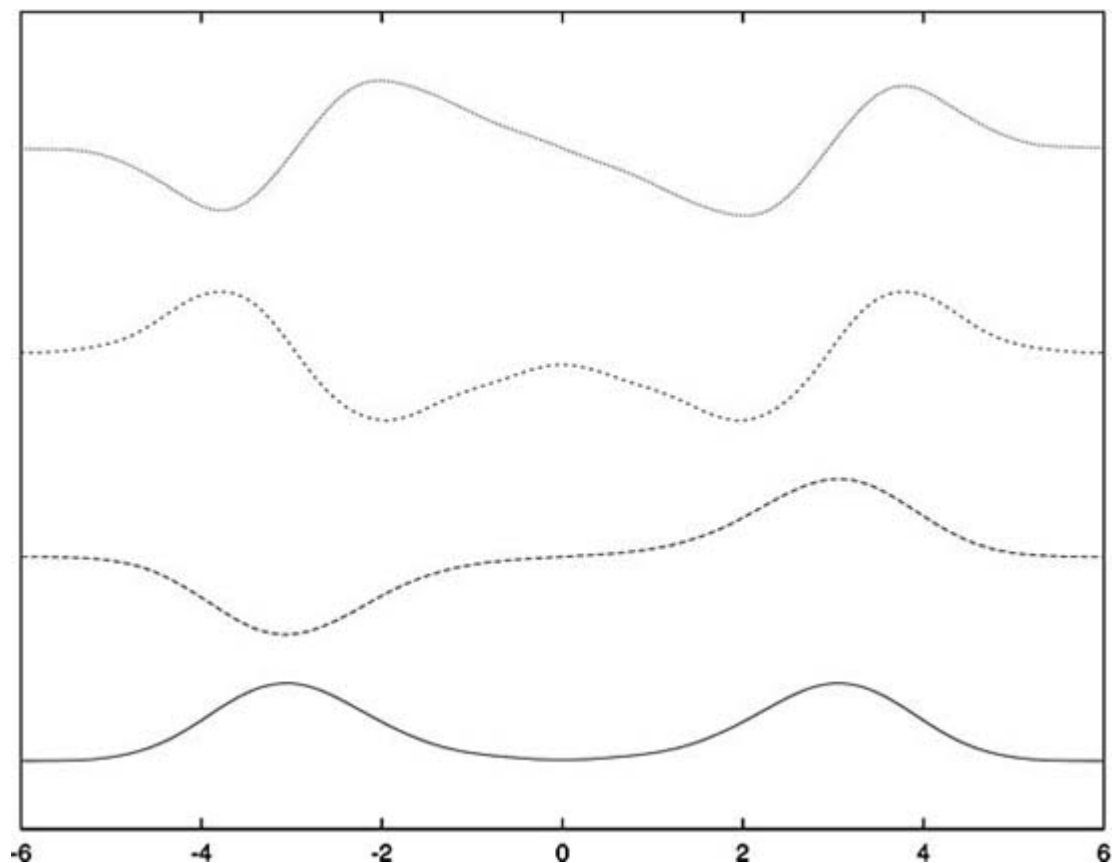


Figure A1.1.4. Wavefunctions for the four lowest states of the double-well oscillator. The ground-state wavefunction is at the bottom and the others are ordered from bottom to top in terms of increasing energy.

-47-

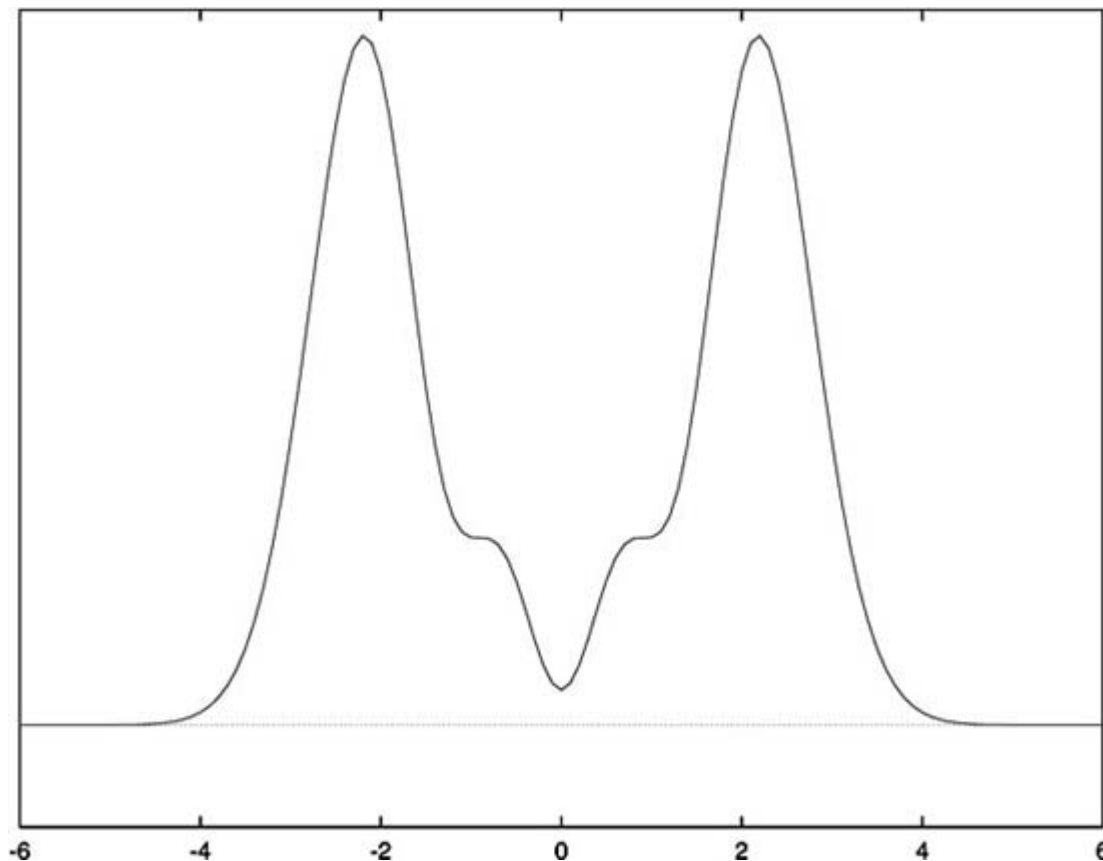


Figure A1.1.5. Ground state wavefunction of the double-well oscillator, as obtained in a variational calculation using eight basis functions centred at the origin. Note the spurious oscillatory behaviour near the origin and the location of the peak maxima, both of which are well inside the potential minima.

The form of the approximate wavefunctions suggests another choice of basis for this problem, namely one comprising some harmonic oscillator functions centred about one minimum and additional harmonic oscillator functions centred about the other minimum. The only minor difficulty in this calculation is that the basis set is not orthogonal (which should be clear simply by inspecting the overlap of the ground-state harmonic oscillator wavefunctions centred at the two points) and an orthonormalization based on [equation \(A1.1.123\)](#), [equation \(A1.1.124\)](#) and [equation \(A1.1.125\)](#) is necessary. Placing an equal number of $V = 2q^2$ harmonic oscillator functions at the position of each minimum (these correspond to solutions of the harmonic oscillator problems with $V = 2(q - \sqrt{10})^2$ and $V = 2(q + \sqrt{10})^2$, respectively) yields the eigenvalues given below for the four lowest states (in each case, there are $N/2$ functions centred at each point).

-48-

N	λ_1	λ_2	λ_3	λ_4
-----	-------------	-------------	-------------	-------------

2	-3.990 62	-3.990 62	—	—
4	-4.01787	-4.01787	-1.92588	-1.92588
6	-4.01851	-4.01851	-2.12522	-2.12521
8	-4.02523	-4.02523	-2.14247	-2.14245
10	-4.02632	-4.02632	-2.17690	-2.17680
20	-4.02663	-4.02660	-2.20290	-2.20064
30	-4.02663	-4.02660	-2.20411	-2.20079
40	-4.02663	-4.02660	-2.20411	-2.20079
50	-4.02663	-4.02660	-2.20411	-2.20079

These results may be compared to those obtained with the basis centred at $q = 0$. The rate of convergence is faster in the present case, which attests to the importance of a carefully chosen basis. It should be pointed out that there is a clear correspondence between the two approaches used here and the single-centre and LCAO expansions used in molecular orbital theory; the reader should appreciate the advantages of choosing an appropriately designed multicentre basis set in achieving rapid convergence in some calculations. Finally, in [figure A1.1.6](#) the ground-state wavefunctions calculated with a mixed basis of eight functions (four centred about each of the two minima) are displayed. Note that oscillations seen in the single-centre basis calculation using the same number of functions are completely missing in the non-orthogonal basis calculation.

-49-

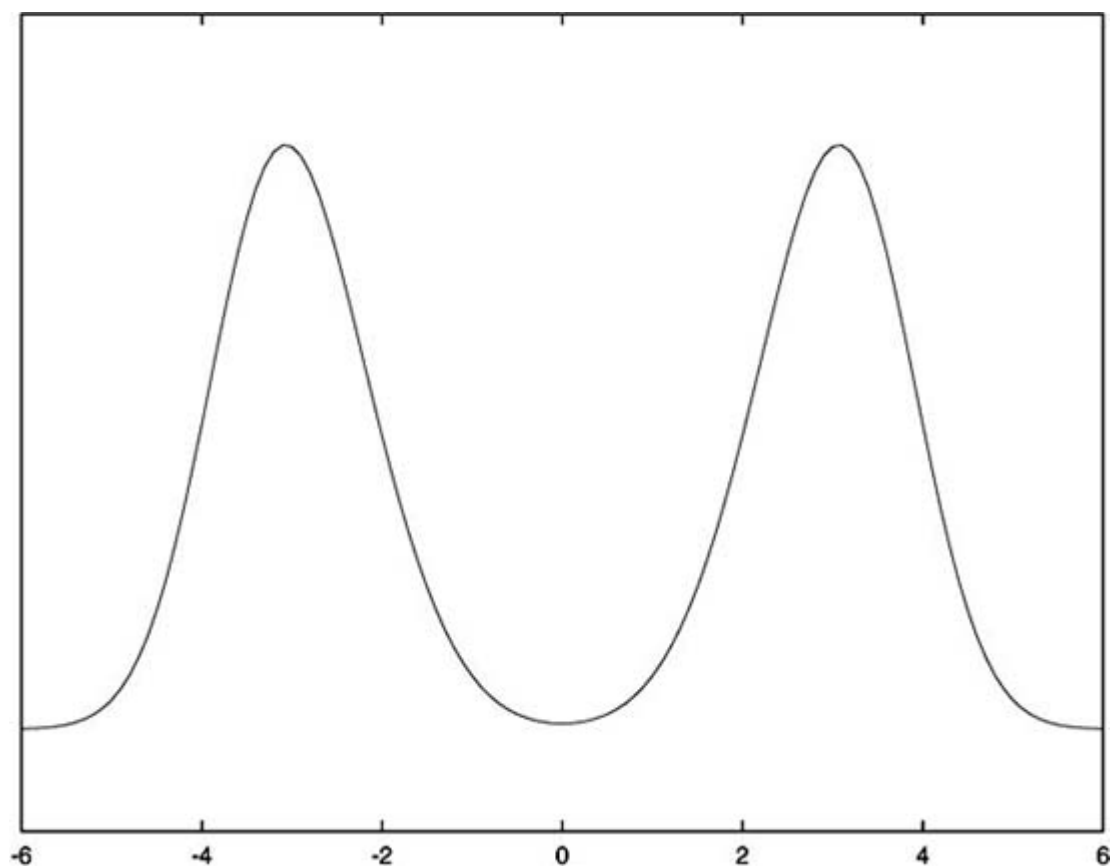


Figure A1.1.6. Ground-state wavefunction of the double-well oscillator, as obtained in a variational calculation using four basis functions centred at $q = \sqrt{10}$ and four centred at $q = -\sqrt{10}$. Note the absence of a node at the origin.

A1.1.4.3 PERTURBATION THEORY

Calculations that employ the linear variational principle can be viewed as those that obtain the exact solution to an approximate problem. The problem is approximate because the basis necessarily chosen for practical calculations is not sufficiently flexible to describe the exact states of the quantum-mechanical system. Nevertheless, within this finite basis, the problem is indeed solved exactly: the variational principle provides a recipe to obtain the best possible solution in the *space spanned by the basis functions*. In this section, a somewhat different approach is taken for obtaining approximate solutions to the Schrödinger equation. Instead of obtaining exact eigenvalues of \mathbf{H} in a finite basis, a strategy is developed for determining approximate eigenvalues of the exact matrix representation of \hat{H} . It can also be used (and almost always is in practical calculations) to obtain approximate eigenvalues to approximate (incomplete basis) Hamiltonian matrices that are nevertheless much larger in dimension than those that can be diagonalized exactly. The standard textbook presentation of this technique, which is known as perturbation theory, generally uses the Schrödinger differential equation as the starting point. However, some of the generality and usefulness of the technique can be lost in the treatment. Students may not come away with an appreciation for the role of linear algebra in perturbation theory, nor do they usually grasp the (approximate problem, exact answer)/(right—or

at least less approximate— problem/approximate answer) distinction between matrix diagonalization in the linear variational principle and the use of perturbation theory.

In perturbation theory, the Hamiltonian is divided into two parts. One of these corresponds to a Schrödinger equation that can be solved exactly

$$\hat{H}_0 \chi_k^{(0)} = \lambda_k^{(0)} \chi_k^{(0)} \tag{A1.1.132}$$

while the remainder of the Hamiltonian is designated here as \hat{V} . The orthonormal eigenfunctions $\chi_k^{(0)}$ of the *unperturbed, or zeroth-order Hamiltonian* \hat{H}_0 form a convenient basis for a matrix representation of the Hamiltonian \hat{H} . Diagonalization of \mathbf{H} gives the exact quantum-mechanical energy levels if the complete set of $\chi_k^{(0)}$ is used, and approximate solutions if the basis is truncated. Instead of focusing on the exact eigenvalues of \mathbf{H} , however, the objective of perturbation theory is to approximate them. The starting point is the matrix representation of H_0 and V , which will be designated as \mathbf{h}

$$\mathbf{h} = \begin{pmatrix} h_{00} & 0 & 0 & \cdots & 0 \\ 0 & h_{11} & 0 & \cdots & 0 \\ 0 & 0 & h_{22} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & h_{NN} \end{pmatrix} \tag{A1.1.133}$$

and \mathbf{v}

$$\mathbf{v} = \begin{pmatrix} v_{00} & v_{01} & v_{02} & \cdots & v_{0N} \\ v_{10} & v_{11} & v_{12} & \cdots & v_{1N} \\ v_{20} & v_{21} & v_{22} & \cdots & v_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ v_{N0} & v_{N1} & v_{N2} & \cdots & v_{NN} \end{pmatrix} \quad (\text{A1.1.134})$$

respectively, where the matrix elements h_{ii} and v_{ij} are given by the integrals

$$h_{ii} \equiv \int \chi_i^{(0)*} H_0 \chi_i^{(0)} d\tau = \lambda_i^{(0)} \quad (\text{A1.1.135})$$

and

$$v_{ij} \equiv \int \chi_i^{(0)*} V \chi_j^{(0)} d\tau. \quad (\text{A1.1.136})$$

-51-

Note that \mathbf{h} is simply the diagonal matrix of zeroth-order eigenvalues $\lambda_k^{(0)}$. In the following, it will be assumed that the zeroth-order eigenfunction $\chi_0^{(0)}$ is a reasonably good approximation to the exact ground-state wavefunction (meaning that $\lambda_0^{(0)} \sim \lambda_0$), and \mathbf{h} and \mathbf{v} will be written in the compact representations

$$\mathbf{h} = \begin{pmatrix} \lambda_0^{(0)} & \mathbf{0} \\ \mathbf{0} & \Lambda_q^{(0)} \end{pmatrix} \quad (\text{A1.1.137})$$

$$\mathbf{v} = \begin{pmatrix} v_{00} & \mathbf{v}_{0q} \\ \mathbf{v}_{q0} & \mathbf{v}_{qq} \end{pmatrix}. \quad (\text{A1.1.138})$$

It is important to realize that while the uppermost diagonal elements of these matrices are numbers, the other diagonal element is a matrix of dimension N . Specifically, these are the matrix representations of H_0 and V in the basis \mathbf{q} which consists of all $\chi_k^{(0)}$ in the original set, apart from $\chi_0^{(0)}$, i.e.

$$\mathbf{q} = \{\chi_1^{(0)}, \chi_2^{(0)} \cdots \chi_N^{(0)}\}. \quad (\text{A1.1.139})$$

The off-diagonal elements in this representation of \mathbf{h} and \mathbf{v} are the zero vector of length N (for \mathbf{h}) and matrix elements which *couple* the zeroth-order ground-state eigenfunction $\chi_0^{(0)}$ to members of the set \mathbf{q} (for \mathbf{v}):

$$\mathbf{v}_{q0} \ni v_{k0} \equiv \int \chi_k^{(0)*} V \chi_0^{(0)} \quad (k \neq 0). \quad (\text{A1.1.140})$$

The exact ground-state eigenvalue λ_0 and corresponding eigenvector

$$\mathbf{c} \equiv \begin{pmatrix} c_0 \\ \mathbf{c}_q \end{pmatrix} \quad (\text{A1.1.141})$$

clearly satisfy the coupled equations

$$H_{00}c_0 + \mathbf{H}_{0q}c_q = c_0\lambda_0 \quad (\text{A1.1.142})$$

$$\mathbf{H}_{q0}c_0 + \mathbf{H}_{qq}c_q = c_q\lambda_0. \quad (\text{A1.1.143})$$

The latter of these can be solved for c_q

$$c_q = [\lambda_0 \mathbf{1} - \mathbf{H}_{qq}]^{-1} \mathbf{h}_{q0} c_0 \quad (\text{A1.1.144})$$

(the N by N identity matrix is represented here and in the following by $\mathbf{1}$) and inserted into equation (A1.1.142) to yield the implicit equation

-52-

$$\lambda_0 = \{H_{00} + \mathbf{H}_{0q}[\lambda_0 \mathbf{1} - \mathbf{H}_{qq}]^{-1} \mathbf{H}_{q0}\}. \quad (\text{A1.1.145})$$

Thus, one can solve for the eigenvalue iteratively, by guessing λ_0 , evaluating the right-hand side of equation (A1.1.145), using the resulting value as the next guess and continuing in this manner until convergence is achieved. However, this is not a satisfactory method for solving the Schrödinger equation, because the problem of diagonalizing a matrix of dimension $N + 1$ is replaced by an iterative procedure in which a matrix of dimension N must be inverted for each successive improvement in the guessed eigenvalue. This is an even more computationally intensive problem than the straightforward diagonalization approach associated with the linear variational principle.

Nevertheless, equation (A1.1.145) forms the basis for the approximate diagonalization procedure provided by perturbation theory. To proceed, the exact ground-state eigenvalue and corresponding eigenvector are written as the sums

$$\mathbf{c} = \mathbf{c}^{(0)} + \mathbf{c}^{(1)} + \mathbf{c}^{(2)} + \dots \quad (\text{A1.1.146})$$

and

$$\lambda_0 = \lambda_0^{(0)} + \lambda_0^{(1)} + \lambda_0^{(2)} + \dots \quad (\text{A1.1.147})$$

where $c^{(k)}$ and $\lambda_0^{(k)}$ are said to be k th-order contributions in the *perturbation expansion*. What is meant here by order? Ultimately, the various contributions to c and λ_0 will be written as matrix products involving the unperturbed Hamiltonian matrix \mathbf{h} and the matrix representation of the perturbation \mathbf{v} . The order of a particular contribution is defined by the number of times \mathbf{v} appears in the corresponding matrix product. Roughly speaking, if $\lambda_0^{(0)} \mathbf{1} - \Lambda_q^{(0)}$ is of order unity, and the matrix elements of \mathbf{v} are an order of magnitude or two smaller, then the third-order energy contribution should be in the range 10^{-3} – 10^{-6} . Therefore, one expects the low order contributions to be most important and the expansions given by equation (A1.1.146) and equation (A1.1.147) to converge rapidly, provided the zeroth-order description of the quantum-mechanical system is reasonably accurate.

To derive equations for the order-by-order contributions to the eigenvalue λ , the implicit equation for the eigenvalue is first rewritten as

$$\begin{aligned}
\lambda_0^{(0)} + \Delta\lambda &= \{\lambda_0^{(0)} + v_{00} + \mathbf{v}_{0q}[\lambda_0^{(0)}\mathbf{1} + \Delta\lambda\mathbf{1} - \Lambda_q^{(0)} - \mathbf{v}_{qq}]^{-1}\mathbf{v}_{q0}\} \\
&= \{\lambda_0^{(0)} + v_{00} + \mathbf{v}_{0q}[\lambda_0^{(0)}\mathbf{1} - \Lambda_q^{(0)} + \Delta\lambda\mathbf{1} - \mathbf{v}_{qq}]^{-1}\mathbf{v}_{q0}\} \\
&= \{\lambda_0^{(0)} + v_{00} + \mathbf{v}_{0q}[(\lambda_0^{(0)}\mathbf{1} - \Lambda_q^{(0)})\{\mathbf{1} - (\lambda_0^{(0)}\mathbf{1} - \Lambda_q^{(0)})^{-1}(\mathbf{v}_{qq} - \Delta\lambda\mathbf{1})\}]^{-1}\mathbf{v}_{q0}\} \\
&= \{\lambda_0^{(0)} + v_{00} + \mathbf{v}_{0q}[\mathbf{1} - (\lambda_0^{(0)}\mathbf{1} - \Lambda_q^{(0)})^{-1}(\mathbf{v}_{qq} - \Delta\lambda\mathbf{1})]^{-1}(\lambda_0^{(0)}\mathbf{1} - \Lambda_q^{(0)})^{-1}\mathbf{v}_{q0}\}
\end{aligned} \tag{A1.1.148}$$

where $\Delta\lambda$ is a shorthand notation for the error in the zeroth-order eigenvalue λ

$$\Delta\lambda \equiv \lambda_0 - \lambda_0^{(0)} = \lambda_0^{(1)} + \lambda_0^{(2)} + \lambda_0^{(3)} + \dots \tag{A1.1.149}$$

-53-

There are two matrix inverses that appear on the right-hand side of these equations. One of these is trivial; the matrix $\lambda_0^{(0)}$ is diagonal. The other inverse

$$[\mathbf{1} - (\lambda_0^{(0)}\mathbf{1} - \Lambda_q^{(0)})^{-1}(\mathbf{v}_{qq} - \Delta\lambda\mathbf{1})]^{-1} \tag{A1.1.150}$$

is more involved because the matrix \mathbf{v}_{qq} is not diagonal, and direct inversion is therefore problematic. However, if the zeroth-order ground-state energy is well separated from low lying excited states, the diagonal matrix hereafter designated as \mathbf{R}_q

$$\mathbf{R}_q \equiv (\lambda_0^{(0)}\mathbf{1} - \Lambda_q^{(0)})^{-1} \tag{A1.1.151}$$

that acts in equation (A1.1.150) to scale

$$(\mathbf{v}_{qq} - \Delta\lambda\mathbf{1}) \tag{A1.1.152}$$

will consist of only small elements. Thus, the matrix to be inverted can be considered as

$$\mathbf{1} - \mathbf{X} \tag{A1.1.153}$$

where \mathbf{X} is, in the sense of matrices, small with respect to $\mathbf{1}$. It can be shown that the inverse of the matrix $\mathbf{1} - \mathbf{X}$ can be written as a series expansion

$$(\mathbf{1} - \mathbf{X})^{-1} = \mathbf{1} + \mathbf{X} + \mathbf{X}\mathbf{X} + \mathbf{X}\mathbf{X}\mathbf{X} + \mathbf{X}\mathbf{X}\mathbf{X}\mathbf{X} + \dots \tag{A1.1.154}$$

that converges if all eigenvalues of \mathbf{X} lie within the unit circle in the complex plane (complex numbers $a + bi$ such that $a^2 + b^2 < 1$). Applications of perturbation theory in quantum mechanics are predicated on the assumption that the series converges for the inverse given by equation (A1.1.150), but efforts are rarely made to verify that this is indeed the case. Use of the series representation of the inverse in [equation \(A1.1.148\)](#) gives the unwieldy formal equality

$$\begin{aligned}
\lambda_0^{(0)} + \Delta\lambda &= \lambda_0^{(0)} + v_{00} + \mathbf{v}_{0q}\mathbf{R}_q\mathbf{v}_{q0} + \mathbf{v}_{0q}\mathbf{R}_q(\mathbf{v}_{qq} - \Delta\lambda\mathbf{1})\mathbf{R}_q\mathbf{v}_{q0} \\
&\quad + \mathbf{v}_{0q}\mathbf{R}_q(\mathbf{v}_{qq} - \Delta\lambda\mathbf{1})\mathbf{R}_q(\mathbf{v}_{qq} - \Delta\lambda\mathbf{1})\mathbf{R}_q\mathbf{v}_{q0} + \dots
\end{aligned} \tag{A1.1.155}$$

from which the error in the zeroth-order energy $\Delta\lambda$ is easily seen to be

$$\begin{aligned} \lambda_0^{(1)} + \lambda_0^{(2)} + \lambda_0^{(3)} + \dots = & \mathbf{v}_{00} + \mathbf{v}_{0q} \mathbf{R}_q \mathbf{v}_{q0} + \mathbf{v}_{0q} \mathbf{R}_q (\mathbf{v}_{qq} - \Delta\lambda \mathbf{1}) \mathbf{R}_q \mathbf{v}_{q0} \\ & + \mathbf{v}_{0q} \mathbf{R}_q (\mathbf{v}_{qq} - \Delta\lambda \mathbf{1}) \mathbf{R}_q (\mathbf{v}_{qq} - \Delta\lambda \mathbf{1}) \mathbf{R}_q \mathbf{v}_{q0} + \dots \end{aligned} \quad (\text{A1.1.156})$$

-54-

Each term on the right-hand side of the equation involves matrix products that contain \mathbf{v} a specific number of times, either explicitly or implicitly (for the terms that involve $\Delta\lambda$). Recognizing that \mathbf{R}_q is a zeroth-order quantity, it is straightforward to make the associations

$$\lambda_0^{(1)} = v_{00} \quad (\text{A1.1.157})$$

$$\lambda_0^{(2)} = \mathbf{v}_{0q} \mathbf{R}_q \mathbf{v}_{q0} \quad (\text{A1.1.158})$$

$$\lambda_0^{(3)} = \mathbf{v}_{0q} \mathbf{R}_q \mathbf{v}_{qq} \mathbf{R}_q \mathbf{v}_{q0} - \lambda_0^{(1)} \mathbf{v}_{0q} \mathbf{R}_q \mathbf{R}_q \mathbf{v}_{q0} \quad (\text{A1.1.159})$$

$$\begin{aligned} \lambda_0^{(4)} = & \mathbf{v}_{0q} \mathbf{R}_q \mathbf{v}_{qq} \mathbf{R}_q \mathbf{v}_{qq} \mathbf{R}_q \mathbf{v}_{q0} - \lambda_0^{(2)} \mathbf{v}_{0q} \mathbf{R}_q \mathbf{R}_q \mathbf{v}_{q0} \\ & - 2\lambda_0^{(1)} \mathbf{v}_{0q} \mathbf{R}_q \mathbf{R}_q \mathbf{v}_{qq} \mathbf{R}_q \mathbf{v}_{q0} + \lambda_0^{(1)} \lambda_0^{(1)} \mathbf{v}_{0q} \mathbf{R}_q \mathbf{R}_q \mathbf{R}_q \mathbf{v}_{q0} \dots \end{aligned} \quad (\text{A1.1.160})$$

which provide recipes for calculating corrections to the energy to fourth order. Similar analysis of [equation \(A1.1.146\)](#) provides successively ordered corrections to the zeroth-order eigenvector ($c_0^{(0)} = 1, c_q^{(0)} = 0$), specifically

$$c_q^{(1)} = \mathbf{R}_q \mathbf{v}_{q0} \quad (\text{A1.1.161})$$

$$c_q^{(2)} = \mathbf{R}_q \mathbf{v}_{qq} \mathbf{R}_q \mathbf{v}_{q0} - \lambda_0^{(1)} \mathbf{R}_q \mathbf{R}_q \mathbf{v}_{q0} \quad (\text{A1.1.162})$$

\vdots

At this point, it is appropriate to make some general comments about perturbation theory that relate to its use in qualitative aspects of chemical physics. Very often, our understanding of complex systems is based on some specific zeroth-order approximation that is then modified to allow for the effect of a perturbation. For example, chemical bonding is usually presented as a weak interaction between atoms in which the atomic orbitals interact to form bonds. Hence, the free atoms represent the zeroth-order picture, and the perturbation is the decrease in internuclear distance that accompanies bond formation. Many rationalizations for bonding trends traditionally taught in descriptive chemistry are ultimately rooted in perturbation theory. As a specific illustration, the decreasing bond strength of carbon–halogen bonds in the sequence C–F > C–Cl > C–Br > C–I (a similar trend is found in the sequence CO, CS, CSe, CTe) can be attributed to a ‘mismatch’ of the np halogen orbitals with the $2p$ orbitals of carbon as for larger values of n . From the point of perturbation theory, it is easily understood that the interaction between the bonding electrons is maximized when the corresponding energy levels are close (small denominators, large values of \mathbf{R}_q) while large energy mismatches (such as that between the valence orbitals of iodine and carbon) allow for less interaction and correspondingly weaker bonds.

For qualitative insight based on perturbation theory, the two lowest order energy corrections and the first-order wavefunction corrections are undoubtedly the most useful. The first-order energy corresponds to averaging the effects of the perturbation over the approximate wavefunction χ_0 , and can usually be evaluated without difficulty. The sum of $\lambda_0^{(1)}$ and \mathbf{v}_{00} is precisely equal to the expectation value of the Hamiltonian over the zeroth-order description χ_0 , and is therefore the proper energy to associate with a simplified model. (It

should be pointed out that it is this energy and not the zeroth-order energy obtained by summing up orbital eigenvalues that is used as the basis for orbital optimization in Hartree–Fock theory. It is often stated that the first-order correction to the Hartree–Fock energy vanishes, but this is

-55-

misleading; the first-order energy is defined instead to be part of the Hartree–Fock energy.) The second-order correction allows for *interaction* between the zeroth-order wavefunction and all others, weighted by the reciprocal of the corresponding energy differences and the magnitude of the matrix elements $\chi_q^{(0)}$. The same interactions between $c_q^{(1)}$ and the $\chi_q^{(0)}$ determine the extent to which the latter are *mixed in* to the first-order perturbed wavefunction described by $c_q^{(1)}$. This is essentially the idea invoked in the theory of orbital hybridization. In the presence of four identical ligands approaching a carbon atom tetrahedrally, its valence s and p orbitals are mixed (through the corresponding $\chi_q^{(0)}$ elements, which vanish at infinite separation) and their first-order correction in the presence of the perturbation (the ligands) can be written as four equivalent linear combinations between the s and three p zeroth-order orbitals.

Some similarities and differences between perturbation theory and the linear variational principle need to be emphasized. First, neither approach can be used in practice to obtain exact solutions to the Schrödinger equation for intractable Hamiltonians. In either case, an infinite basis is required; neither the sums given by perturbation theory nor the matrix diagonalization of a variational calculation can be carried out. Hence, the strengths and weaknesses of the two approaches should be analysed from the point of view that the basis is necessarily truncated. Within this constraint, diagonalization of \mathbf{H} represents the best solution that is possible in the space spanned by the basis set. In variational calculations, rather severe truncation of \mathbf{H} is usually required, with the effect that its eigenvalues might be poor approximations to the exact values. The problem, of course, is that the basis is not sufficiently flexible to accurately represent the true quantum-mechanical wavefunction. In perturbation theory, one can include significantly more functions in the calculation. It turns out that the results of a low order perturbation calculation are often superior to a practical variational treatment of the same problem. Unlike variational methods, perturbation theory does not provide an upper bound to the energy (apart from a first-order treatment) and is not even guaranteed to converge. However, in chemistry, it is virtually always energy differences—and not absolute energies—that are of interest, and differences of energies obtained variationally are not themselves upper (or lower) bounds to the exact values. For example, suppose a spectroscopic transition energy between the states ψ_i and ψ_j is calculated from the difference $\lambda_i - \lambda_j$, obtained by diagonalizing \mathbf{H} in a truncated basis. There is no way of knowing whether this value is above or below the exact answer, a situation no different than that associated with taking the difference between two approximate eigenvalues obtained from two separate calculations based on perturbation theory.

In the quantum mechanics of atoms and molecules, both perturbation theory and the variational principle are widely used. For some problems, one of the two classes of approach is clearly best suited to the task, and is thus an established choice. However, in many others, the situation is less clear cut, and calculations can be done with either of the methods or a combination of both.

FURTHER READING

Berry R S, Rice S A and Ross J R 1980 *Physical Chemistry* 6th edn (New York, NY: Wiley)

The introductory treatment of quantum mechanics presented in this textbook is excellent. Particularly appealing is the effort devoted to developing a qualitative understanding of quantum-mechanical principles.

Karplus M and Porter R N 1970 *Atoms and Molecules: an Introduction for Students of Physical Chemistry* (Reading, MA: Addison-Wesley)

An excellent treatment of molecular quantum mechanics, on a level comparable to that of Szabo and Ostlund. The scope of this book is quite different, however, as it focuses mainly on the basic principles of quantum mechanics and the theoretical treatment of spectroscopy.

Levine I N 1991 *Quantum Chemistry* 4th edn (Englewood Cliffs, NJ: Wiley)

A relatively complete survey of quantum chemistry, written on a level just below that of the Szabo and Ostlund text. Levine has done an excellent job in including up-to-date material in successive editions of this text, which makes for interesting as well as informative reading.

Szabo A and Ostlund N S 1996 *Modern Quantum Chemistry* (New York: Dover)

Although it is now somewhat dated, this book provides one of the best treatments of the Hartree–Fock approximation and the basic ideas involved in evaluating the correlation energy. An especially valuable feature of this book is that much attention is given to how these methods are actually implemented.

Pauling L and Wilson E B 1935 *Introduction to Quantum Mechanics* (New York: Dover)

This venerable book was written in 1935, shortly after the birth of modern quantum mechanics. Nevertheless, it remains one of the best sources for students seeking to gain an understanding of quantum-mechanical principles that are relevant in chemistry and chemical physics. Equally outstanding jobs are done in dealing with both quantitative and qualitative aspects of the subject. More accessible to most chemists than Landau and Lifschitz.

Landau L D and Lifschitz E M 1977 *Quantum Mechanics (Nonrelativistic Theory)* (Oxford: Pergamon)

A marvellous and rigorous treatment of non-relativistic quantum mechanics. Although best suited for readers with a fair degree of mathematical sophistication and a desire to understand the subject in great depth, the book contains all of the important ideas of the subject and many of the subtle details that are often missing from less advanced treatments. Unusual for a book of its type, highly detailed solutions are given for many illustrative example problems.

Simons J and Nichols J 1997 *Quantum Mechanics in Chemistry* (New York: Oxford)

A new text that provides a relatively broad view of quantum mechanics in chemistry ranging from electron correlation to time-dependent processes and scattering.

Parr R G and Yang W 1994 *Density-Functional Theory of Atoms and Molecules* (New York: Oxford)

A comprehensive treatment of density functional theory, an idea that is currently very popular in quantum chemistry.

Albright T A, Burdett J K and Whangbo M-H 1985 *Orbital Interactions in Chemistry* (New York: Wiley)

A superb treatment of applied molecular orbital theory and its application to organic, inorganic and solid state chemistry. Perhaps the best source for appreciating the power of the independent-particle approximation and its remarkable ability to account for qualitative behaviour in chemical systems.

Salem L 1966 *Molecular Orbital Theory of Conjugated Systems* (Reading, MA: Benjamin)

A highly readable account of early efforts to apply the independent-particle approximation to problems of organic chemistry. Although more accurate computational methods have since been developed for treating all of the problems discussed in the text, its discussion of approximate Hartree–Fock (semiempirical) methods and their accuracy is still useful. Moreover, the view supplied about what was understood and what was not understood in physical organic chemistry three decades ago is

fascinating.

Pais A 1988 *Inward Bound: of Matter and Forces in the Physical World* (Oxford: Oxford University Press)

A good account of the historical development of quantum mechanics. While much of the book deals with quantum field theory and particle physics, the first third of the book focuses on the period 1850–1930 and the origins of quantum theory. An admirable job is done in placing events in a proper historical context.

-1-

A 1.2 Internal molecular motions

Michael E Kellman

A 1.2.1 INTRODUCTION

Ideas on internal molecular motions go back to the very beginnings of chemistry as a natural science, to the days of Robert Boyle and Isaac Newton [1]. Much of Boyle's interest in chemistry, apart from the 'bewitchment' he found in performing chemical experiments [2], arose from his desire to revive and transform the corpuscular philosophy favoured by some of the ancient Greeks, such as Epicurus [3]. This had lain dormant for centuries, overshadowed by the apparently better-founded Aristotelian cosmology [4], including the theory of the four elements. With the revolution in celestial mechanics that was taking place in modern Europe in the 17th century, Boyle was concerned to persuade natural philosophers that chemistry, then barely emerging from alchemy, was potentially of great value for investigating the corpuscular view, which was re-emerging as a result of the efforts of thinkers such as Francis Bacon and Descartes. This belief of Boyle's was based partly on the notion that the qualitative properties of real substances and their chemical changes could be explained by the joining together of elementary corpuscles, and the 'local motions' within these aggregates—what we now call the internal motions of molecules. Boyle influenced his younger colleague in the Royal Society, Isaac Newton. Despite immense efforts in chemical experimentation, Newton wrote only one paper in chemistry, in which he conjectured the existence of short-range forces in what we now recognize as molecules. Thus, in a true sense, with Boyle and Newton was born the science of chemical physics [1].

This was a child whose development was long delayed, however. Not until the time of John Dalton in the early 19th century, after the long interlude in which the phlogiston theory triumphed and then was overthrown in the chemistry of Lavoisier, did the nascent corpuscular view of Boyle and Newton really begin to grow into a useful atomic and molecular theory [1, 5]. It became apparent that it was necessary to think of the compound states of the elements of Lavoisier in terms of definite molecular formulae, to account for the facts that were becoming known about the physical properties of gases and the reactions of the elements, their joining into compounds and their separation again into elements.

However, it was still a long time even after Dalton before anything definite could be known about the internal motions in molecules. The reason was that the microscopic nature of atoms and molecules was a bar to any knowledge of their internal constituents. Furthermore, nothing at all was known about the physical laws that applied at the microscopic level. The first hints came in the late 19th century, with the classical Maxwell–Lorentz theory of the dynamics of charged particles interacting through the electromagnetic field. The electron was discovered by Thomson, and a little later the nuclear structure of the atom by Rutherford. This set the stage in the 20th century for a physical understanding in terms of quantum theory of the constituents of molecules, and the motions of which they partake.

This section will concentrate on the motions of atoms within molecules—'internal molecular motions'—as comprehended by the revolutionary quantum ideas of the 20th century. Necessarily, limitations of space prevent many topics from being treated in the detail they deserve. Some of these are treated in more detail in

other articles in this Encyclopedia, or in references in the Bibliography. The emphasis is on treating certain key topics in sufficient depth to build a foundation for further exploration by the reader, and for branching off into related topics that cannot be treated

-2-

in depth at all. There will not be much focus on molecules undergoing chemical reactions, except for unimolecular rearrangements, which are a rather extreme example of internal molecular motion. However, it must be emphasized that the distinctions between the internal motions of molecules, the motions of atoms in a molecule which is undergoing dissociation and the motion of atoms in two or more molecules undergoing reaction are somewhat artificial. Even the motions which are most properly called 'internal' play a central role in theories of reaction dynamics. In fact, their character in chemical reactions is one of the most important unsolved mysteries in molecular motion. Although we will not have anything directly to say about general theories of reaction [6], the internal motion of molecules undergoing isomerization and the importance of the internal motions in efforts to control reactions with sophisticated laser sources will be two of the topics considered.

A key theme of contemporary chemical physics and physical chemistry is 'ultrafast' molecular processes [7, 8 and 9], including both reaction dynamics and internal molecular motions that do not involve reaction. The probing of ultrafast processes generally is thought of in terms of very short laser pulses, through the window of the time domain. However, most of the emphasis of this section is on probing molecules through the complementary window of the frequency domain, which usually is thought of as the realm of the time-independent processes, which is to say, the 'ultraslow'. One of the key themes of this section is that encrypted within the totality of the information which can be gathered on a molecule in the frequency domain is a vast store of information on ultrafast internal motions. The decoding of this information by new theoretical techniques for analysis of experimental spectra is a leading theme of recent work.

A 1.2.2 QUANTUM THEORY OF ATOMIC AND MOLECULAR STRUCTURE AND MOTION

The understanding of molecular motions is necessarily based on quantum mechanics, the theory of microscopic physical behaviour worked out in the first quarter of the 20th century. This is because molecules are microscopic systems in which it is impossible—or at least very dangerous!—to ignore the dual wave–particle nature of matter first recognized in quantum theory by Einstein (in the case of classical waves) and de Broglie (in the case of classical particles).

The understanding of the quantum mechanics of atoms was pioneered by Bohr, in his theory of the hydrogen atom. This combined the classical ideas on planetary motion—applicable to the atom because of the formal similarity of the gravitational potential to the Coulomb potential between an electron and nucleus—with the quantum ideas that had recently been introduced by Planck and Einstein. This led eventually to the formal theory of quantum mechanics, first discovered by Heisenberg, and most conveniently expressed by Schrödinger in the wave equation that bears his name.

However, the hydrogen atom is relatively a very simple quantum mechanical system, because it contains only two constituents, the electron and the nucleus. This situation is the quantum mechanical analogue of a single planet orbiting a sun. It might be thought that an atom with more than one electron is much like a solar system with more than one planet, in which the motion of each of the planets is more or less independent and regular. However, this is not the case, because the relative strength of the interaction between the electrons is much stronger than the attraction of the planets in our solar system. The problem of the internal dynamics of atoms—the internal motion when there is more than one electron—is still very far from a complete understanding. The electrons are not really independent, nor would their motion, if it were described by

classical rather than quantum mechanics, be regular, unlike the annual orbits of the

-3-

planets. Instead, in general, it would be *chaotic*. The corresponding complexity of the quantum mechanical atom with more than one electron, or even one electron in a field, is to this day a challenge [10, 11 and 12]. (In fact, even in the solar system, despite the relative strengths of planetary attraction, there are constituents, the asteroids, with very irregular, chaotic behaviour. The issue of chaotic motion in molecules is an issue that will appear later with great salience.)

As we shall see, in molecules as well as atoms, the interplay between the quantum description of the internal motions and the corresponding classical analogue is a constant theme. However, when referring to the internal motions of molecules, we will be speaking, loosely, of the motion of the atoms in the molecule, rather than of the fundamental constituents, the electrons and nuclei. This is an extremely fundamental point to which we now turn.

A 1.2.3 THE MOLECULAR POTENTIAL ENERGY SURFACE

One of the most salient facts about the structure of molecules is that the electrons are far lighter than the nuclei, by three orders of magnitude and more. This is extremely fortunate for our ability to attain a rational understanding of the internal motion of the electrons and nuclei. In fact, without this it might well be that not much progress would have been made at all! Soon after the discovery of quantum mechanics it was realized that the vast difference in the mass scales of the electrons and nuclei means that it is possible, in the main, to separate the problem into two parts, an electronic and a nuclear part. This is known as the Born–Oppenheimer separability or approximation [13]. The underlying physical idea is that the electrons move much faster than the nuclei, so they adjust rapidly to the relatively much slower nuclear motion. Therefore, the electrons are described by a quantum mechanical ‘cloud’ obtained by solving the Schrödinger wave equation. The nuclei then move slowly within this cloud, which in turn adjusts rapidly as the nuclei move.

The result is that, to a very good approximation, as treated elsewhere in this Encyclopedia, the nuclei move in a mechanical potential created by the much more rapid motion of the electrons. The electron cloud itself is described by the quantum mechanical theory of electronic structure. Since the electronic and nuclear motion are approximately separable, the electron cloud can be described mathematically by the quantum mechanical theory of electronic structure, in a framework where the nuclei are fixed. The resulting Born–Oppenheimer potential energy surface (PES) created by the electrons is the mechanical potential in which the nuclei move. When we speak of the internal motion of molecules, we therefore mean essentially the motion of the nuclei, which contain most of the mass, on the molecular potential energy surface, with the electron cloud rapidly adjusting to the relatively slow nuclear motion.

We will now treat the internal motion on the PES in cases of progressively increasing molecular complexity. We start with the simplest case of all, the diatomic molecule, where the notions of the Born–Oppenheimer PES and internal motion are particularly simple.

The potential energy surface for a diatomic molecule can be represented as in [figure A1.2.1](#). The x -axis gives the internuclear separation R and the y -axis the potential function $V(R)$. At a given value of R , the potential $V(R)$ is determined by solving the quantum mechanical electronic structure problem in a framework with the nuclei fixed at the given value of R . (To reiterate the discussion above, it is only possible to regard the nuclei as fixed in this calculation because of the Born–Oppenheimer separability, and it is important to keep in mind that this is only an approximation.)

There can be subtle but important *non-adiabatic effects* [14, 15], due to the non-exactness of the separability of the nuclei and electrons. These are treated elsewhere in this Encyclopedia.) The potential function $V(R)$ is determined by repeatedly solving the quantum mechanical electronic problem at different values of R . Physically, the variation of $V(R)$ is due to the fact that the electronic cloud adjusts to different values of the internuclear separation R in a subtle interplay of mutual particle attractions and repulsions: electron–electron repulsions, nuclear–nuclear repulsions and electron–nuclear attractions.

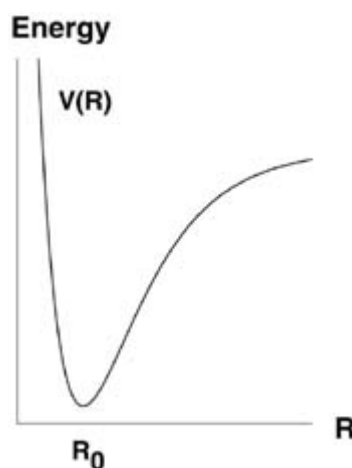


Figure A1.2.1. Potential $V(R)$ of a diatomic molecule as a function of the internuclear separation R . The equilibrium distance R_0 is at the potential minimum.

The potential function in figure A1.2.1 has several crucial characteristics. It has a minimum at a certain value R_0 of the internuclear separation. This is the equilibrium internuclear distance. Near R_0 , the function $V(R)$ rises as R increases or decreases. This means that there is an attractive mechanical force tending to restore the nuclei to R_0 . At large values of R , $V(R)$ flattens out and asymptotically approaches a value which in figure A1.2.1 is arbitrarily chosen to be zero. This means that the molecule dissociates into separated atoms at large R . The difference between the equilibrium potential $V(R_0)$ and the asymptotic energy is the dissociation, or binding, energy. At values of R less than R_0 , the potential $V(R)$ again rises, but now without limit. This represents the repulsion between nuclei as the molecule is compressed.

Classically, the nuclei vibrate in the potential $V(R)$, much like two steel balls connected by a spring which is stretched or compressed and then allowed to vibrate freely. This vibration along the nuclear coordinate R is our first example of internal molecular motion. Most of the rest of this section is concerned with different aspects of molecular vibrations in increasingly complicated situations.

Near the bottom of the potential well, $V(R)$ can be approximated by a parabola, so the function $V(R)$ is approximated as

$$V(R) = kR^2. \quad (\text{A 1.2.1})$$

This is the form of the potential for a harmonic oscillator, so near the bottom of the well, the nuclei undergo nearly

harmonic vibrations. For a harmonic oscillator with potential as in (A1.2.1), the classical frequency of

oscillation is independent of energy and is given by [16, 17 and 18]

$$\omega_0 = 2\pi\nu_0 = \sqrt{k/\mu} \quad (\text{A 1.2.2})$$

where μ is the reduced mass. Quantum mechanically, the oscillator has a series of discrete energy levels, characterized by the number of quanta n in the oscillator. This is the quantum mechanical analogue for the oscillator of the quantized energy levels of the electron in a hydrogen atom. The energy levels of the harmonic oscillator are given by

$$E_n = \omega_0(n + \frac{1}{2}) \quad (\text{A 1.2.3})$$

where \hbar , i.e. Planck's constant h divided by 2π , has been omitted as a factor on the right-hand side, as is appropriate when the customary wavenumber (cm^{-1}) units are used [18].

A 1.2.4 ANHARMONICITY

If the potential were exactly harmonic for all values of R , the vibrational motion would be extremely simple, consisting of vibrations with frequency ω_0 for any given amount of vibrational energy. The fact that this is a drastic oversimplification for a real molecule can be seen from the fact that such a molecule would never dissociate, lacking the flatness in the potential at large R that we saw in [figure A1.2.1](#). As the internuclear separation departs from the bottom of the well at R_0 , the harmonic approximation ([A1.2.1](#)) progressively becomes less accurate as a description of the potential. This is known as *anharmonicity* or *nonlinearity*. Anharmonicity introduces complications into the description of the vibrational motion. The frequency is no longer given by the simple harmonic formula (A1.2.2). Instead, it varies with the amount of energy in the oscillator. This variation of frequency with the number of quanta is the essence of the nonlinearity.

The variation of the frequency can be approximated by a series in the number of quanta, so the energy levels are given by

$$E_n = \omega_0(n + \frac{1}{2}) + \gamma_1(n + \frac{1}{2})^2 + \gamma_2(n + \frac{1}{2})^3 + \dots \quad (\text{A 1.2.4})$$

Often, it is a fair approximation to truncate the series at the quadratic term with γ_1 . The energy levels are then approximated as

$$E_n = \omega_0(n + \frac{1}{2}) + \gamma_1(n + \frac{1}{2})^2. \quad (\text{A 1.2.5})$$

The first term is known as the harmonic contribution and the second term as the quadratic anharmonic correction.

Even with these complications due to anharmonicity, the vibrating diatomic molecule is a relatively simple mechanical system. In polyatomics, the problem is fundamentally more complicated with the presence of more than two atoms. The anharmonicity leads to many extremely interesting effects in the internal molecular motion, including the possibility of chaotic dynamics.

It must be pointed out that another type of internal motion is the overall rotation of the molecule. The vibration and rotation of the molecule are shown schematically in figure A1.2.2.

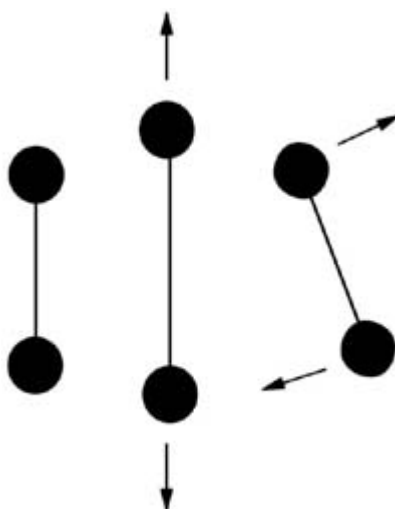


Figure A1.2.2. Internal nuclear motions of a diatomic molecule. Top: the molecule in its equilibrium configuration. Middle: vibration of the molecule. Bottom: rotation of the molecule.

A 1.2.5 POLYATOMIC MOLECULES

In polyatomic molecules there are many more degrees of freedom, or independent ways in which the atoms of the molecule can move. With n atoms, there are a total of $3n$ degrees of freedom. Three of these are the motion of the centre of mass, leaving $(3n-3)$ internal degrees of freedom [18]. Of these, except in linear polyatomics, three are rotational degrees of freedom, leaving $(3n-6)$ vibrational degrees of freedom. (In linear molecules, there are only two rotational degrees of freedom, corresponding to the two individual orthogonal axes of rotation about the molecular axis, leaving $(3n-5)$ vibrational degrees of freedom. For example, the diatomic has only one vibrational degree of freedom, the vibration along the coordinate R which we encountered above.)

Because of limitations of space, this section concentrates very little on rotational motion and its interaction with the vibrations of a molecule. However, this is an extremely important aspect of molecular dynamics of long-standing interest, and with development of new methods it is the focus of intense investigation [18, 19, 20, 21, 22 and 23]. One very interesting aspect of rotation–vibration dynamics involving *geometric phases* is addressed in [section A1.2.20](#).

The $(3n-6)$ degrees of vibrational motion again take place on a PES. This implies that the PES itself must be a function in a $(3n-6)$ dimensional space, i.e. it is a function of $(3n-6)$ *internal coordinates* $r_1 \dots r_N$, where $N = (3n-6)$, which depend on the positions of all the nuclei. The definition of the coordinates $r_1 \dots r_N$ has a great deal of flexibility. To be concrete, for H_2O one choice is the set of internal coordinates illustrated in figure A1.2.3. These are a bending coordinate, i.e. the angular bending displacement from the equilibrium geometry, and two bond displacement coordinates, i.e. the stretching displacement of each O–H bond from its equilibrium length.

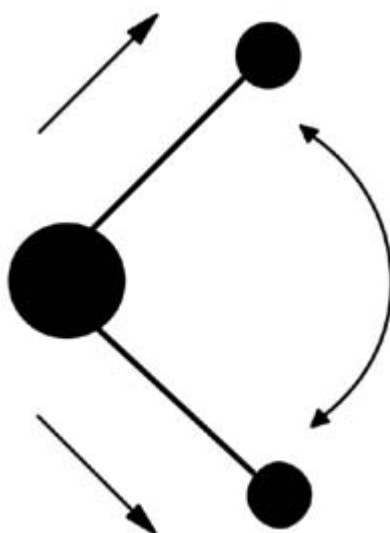


Figure A1.2.3. The internal coordinates of the H_2O molecule. There are two bond stretching coordinates and a bend coordinate.

An equilibrium configuration for the molecule is any configuration $(r_{10} \dots r_{n0})$ where the PES has a minimum, analogous to the minimum in the diatomic potential at R_0 in [figure A1.2.1](#). In general, there can be a number of local equilibrium configurations in addition to the lowest equilibrium configuration, which is called the global equilibrium or minimum. We will refer to an equilibrium configuration in speaking of any of the local equilibria, and *the* equilibrium configuration when referring to the global minimum. In the very close vicinity of the equilibrium configuration, the molecule will execute harmonic vibrations. Since there are $(3n-6)$ vibrational degrees of freedom, there must be $(3n-6)$ harmonic *modes*, or independent vibrational motions. This means that on the multi-dimensional PES, there must be $(3n-6)$ independent coordinates, along any of which the potential is harmonic, near the equilibrium configuration. We will denote these independent degrees of freedom as the *normal modes coordinates* $R_1 \dots R_N$. Each of the R_i in general is some combination of the internal coordinates $r_1 \dots r_N$ in terms of which the nuclear positions and PES were defined earlier. These are illustrated for the case of water in [figure A1.2.4](#). One of the normal modes is a bend, very much like the internal bending coordinate in [figure A1.2.3](#). The other two modes are a symmetric and antisymmetric stretch. Near the equilibrium configuration, given knowledge of the molecular potential, it is possible by the procedure of *normal mode analysis* [24] to calculate the frequencies of each of the normal modes and their exact expression in terms of the original internal coordinates $r_1 \dots r_N$.

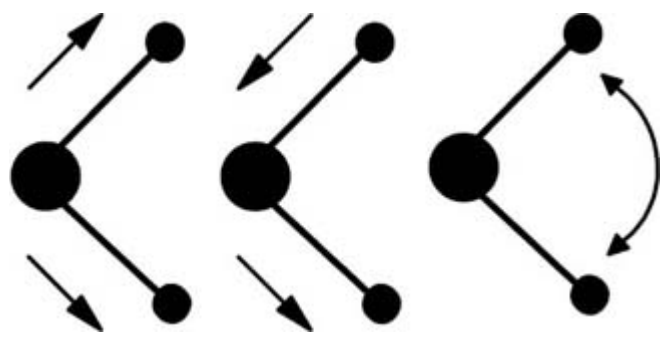


Figure A1.2.4. The normal vibrational coordinates of H_2O . Left: symmetric stretch. Middle: antisymmetric stretch. Right: bend.

It is often very useful to describe classical vibrations in terms of a *trajectory* in the space of coordinates $r_1 \dots r_N$. If the motion follows one of the normal modes, the trajectory is one in which the motion repeats itself

along a closed curve. An example is shown in figure A1.2.5 for the symmetric and antisymmetric stretch modes. The x and y coordinates r_1, r_2 are the displacements of the two O–H bonds. (For each mode i there is a family of curves, one for each value of the energy, with the amplitude of vibration along the normal modes in figure A1.2.5 increasing with energy; the figure shows the trajectory of each mode for one value of the energy.)

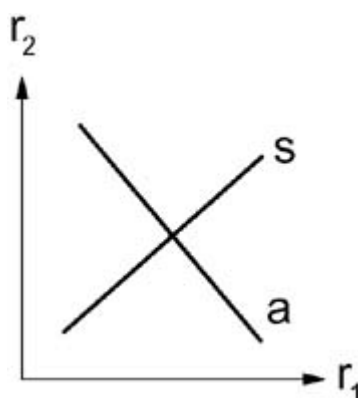


Figure A1.2.5. Harmonic stretch normal modes of a symmetric triatomic. The symmetric stretch s and antisymmetric stretch a are plotted as a function of the bond displacements r_1, r_2 .

In general, each normal mode in a molecule has its own frequency, which is determined in the normal mode analysis [24]. However, this is subject to the constraints imposed by molecular symmetry [18, 25, 26]. For example, in the methane molecule CH_4 , four of the normal modes can essentially be designated as normal stretch modes, i.e. consisting primarily of collective motions built from the four C–H bond displacements. The molecule has tetrahedral symmetry, and this constrains the stretch normal mode frequencies. One mode is the totally symmetric stretch, with its own characteristic frequency. The other three stretch normal modes are all constrained by symmetry to have the same frequency, and are referred to as being *triply-degenerate*.

-9-

The $(3n-6)$ normal modes with coordinates $R_1 \dots R_N$ are often designated $\nu_1 \dots \nu_N$. (Not to be confused with the common usage of ν to denote a frequency, as in [equation \(A1.2.2\)](#), the last such usage in this section.) Quantum mechanically, each normal mode ν_i is characterized by the number of vibrational quanta ν_i in the mode. Then the vibrational state of the molecule is designated or *assigned* by the number of quanta ν_i in each of the modes, i.e. $(n_1 \dots n_N)$. In the harmonic approximation in which each mode i is characterized by a frequency ω_i , the vibrational energy of a state assigned as $(n_1 \dots n_N)$ is given by

$$E(n_1 \dots n_N) = (n_1 + \frac{1}{2})\omega_1 + (n_2 + \frac{1}{2})\omega_2 + \dots + (n_N + \frac{1}{2})\omega_N \quad (\text{A 1.2.6})$$

A 1.2.6 ANHARMONIC NORMAL MODES

In the polyatomic molecule, just as in the diatomic, the PES must again be highly anharmonic away from the vicinity of the potential minimum, as seen from the fact that the polyatomic can dissociate; in fact in a multiplicity of ways, because in general there can be several dissociation products. In addition, the molecule can have complicated internal rearrangements in which it isomerizes. This means that motion takes place from one minimum in the PES, over a saddle, or ‘pass’, and into another minimum. We will have something to say about these internal rearrangements later. However, the fact of anharmonicity raises important questions about the normal modes even in the near vicinity of an equilibrium configuration. We saw above that anharmonicity in a diatomic means that the frequency of the vibrational motion varies with the amount of vibrational energy.

An analogous variation of frequency of the normal modes occurs in polyatomics.

However, there is a much more profound prior issue concerning anharmonic normal modes. The existence of the normal vibrational modes, involving the collective motion of all the atoms in the molecule as illustrated for H₂O in [figure A1.2.4](#) was predicated on the basis of the existence of a harmonic potential. But if the potential is not exactly harmonic, as is the case everywhere except right at the equilibrium configuration, are there still collective normal modes? And if so, since they cannot be harmonic, what is their nature and their relation to the harmonic modes?

The beginning of an answer comes from a theorem of Moser and Weinstein in mathematical nonlinear dynamics [27, 28]. This theorem states that in the vicinity of a potential minimum, a system with $(3n-6)$ vibrational degrees of freedom has $(3n-6)$ *anharmonic normal modes*. What is the difference between the harmonic normal modes and the *anharmonic* normal modes proven to exist by Moser and Weinstein? [Figure A1.2.6](#) shows anharmonic stretch normal modes. The symmetric stretch looks the same as its harmonic counterpart in [Figure A1.2.5](#); this is necessarily so because of the symmetry of the problem. The antisymmetric stretch, however, is distinctly different, having a curvilinear appearance in the zero-order bond modes. The significance of the Moser–Weinstein theorem is that it guarantees that in the vicinity of a minimum in the PES, there must be a set of $(3n-6)$ of these anharmonic modes.

-10-

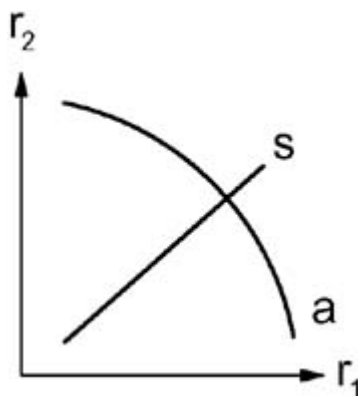


Figure A1.2.6. Anharmonic stretch normal modes of a symmetric triatomic. The plot is similar to [figure A1.2.5](#), except the normal modes are now anharmonic and can be curvilinear in the bond displacement coordinates r_1, r_2 . The antisymmetric stretch is curved, but the symmetric stretch is linear because of symmetry.

It is sometimes very useful to look at a trajectory such as the symmetric or antisymmetric stretch of [figure A1.2.5](#) and [figure A1.2.6](#) not in the physical spatial coordinates $(r_1 \dots r_N)$, but in the *phase space* of Hamiltonian mechanics [16, 29], which in addition to the coordinates $(r_1 \dots r_N)$ also has as additional coordinates the set of conjugate momenta $(p_1 \dots p_N)$. In phase space, a one-dimensional trajectory such as the antisymmetric stretch again appears as a one-dimensional curve, but now the curve closes on itself. Such a trajectory is referred to in nonlinear dynamics as a *periodic orbit* [29]. One says that the anharmonic normal modes of Moser and Weinstein are *stable* periodic orbits.

What does it mean to say the modes are stable? Suppose that one fixes the initial conditions—the initial values of the coordinates and momenta, for a given fixed value of the energy—so the trajectory does not lie entirely on one of the anharmonic modes. At any given time the position and momentum is some combination of each of the normal motions. An example of the kind of trajectory that can result is shown in [figure A1.2.7](#). The trajectory lies in a box with extensions in each of the anharmonic normal modes, filling the box in a very regular, ‘woven’ pattern. In phase space, a regular trajectory in a box is no longer a one-dimensional closed curve, or periodic orbit. Instead, in phase space a box-filling trajectory lies on a surface which has the

qualitative form, or topology, of a torus—the surface of a doughnut. The confinement of the trajectory to such a box indicates that the normal modes are stable. (Unstable modes do exist and will be of importance later.) Another quality of the trajectory in the box is its ‘woven’ pattern. Such a trajectory is called *regular*. We will consider other, *chaotic* types of trajectories later; the chaos and instability of modes are closely related. The issues of periodic orbits, stable modes and regular and chaotic motion have been studied in great depth in the theory of Hamiltonian or energy-preserving dynamical systems [29, 30]. We will return repeatedly to concepts of classical dynamical systems.

-11-

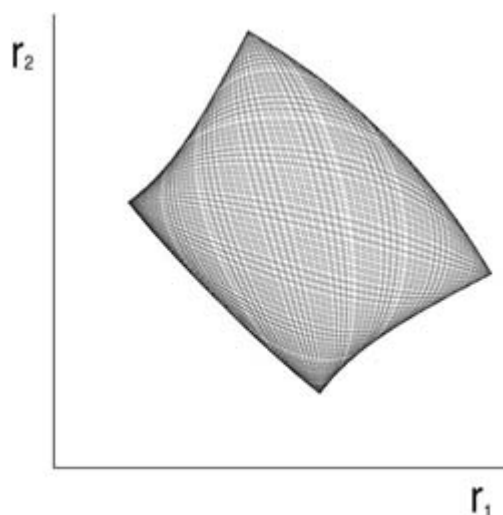


Figure A1.2.7. Trajectory of two coupled stretches, obtained by integrating Hamilton’s equations for motion on a PES for the two modes. The system has stable anharmonic symmetric and antisymmetric stretch modes, like those illustrated in [figure A1.2.6](#). In this trajectory, semiclassically there is one quantum of energy in each mode, so the trajectory corresponds to a combination state with quantum numbers $[n_s, n_a] = [1, 1]$. The ‘woven’ pattern shows that the trajectory is regular rather than chaotic, corresponding to motion in phase space on an invariant torus.

However, the reader may be wondering, what is the connection of all of these classical notions—stable normal modes, regular motion on an invariant torus—to the quantum spectrum of a molecule observed in a spectroscopic experiment? Recall that in the harmonic normal modes approximation, the quantum levels are defined by the set of quantum numbers (n_1, \dots, n_N) giving the number of quanta in each of the normal modes.

Does it make sense to associate a definite quantum number n_i to each mode i in an anharmonic system? In general, this is an extremely difficult question! But remember that so far, we are speaking of the situation in some small vicinity of a minimum on the PES, where the Moser–Weinstein theorem *guarantees* the existence of the anharmonic normal modes. This essentially guarantees that quantum levels with low enough v_i values correspond to trajectories that lie on invariant tori. Since the levels are quantized, these must be special tori, each characterized by quantized values of the classical actions $I_i = (n_i + \frac{1}{2})\hbar$, which are constants of the motion on the invariant torus. As we shall see, the possibility of assigning a set of N quantum numbers n_i to a level, one for each mode, is a very special situation that holds only near the potential minimum, where the motion is described by the N anharmonic normal modes. However, let us continue for now with the region of the spectrum where this special situation applies.

If there are n_i quanta in mode i and zero quanta in all the other modes, the state is called an *overtone* of the normal mode i . What does such a state correspond to in terms of a classical trajectory? Consider the overtone of the antisymmetric stretch, again neglecting the bend. If all the energy in the overtone were in mode i , the trajectory would look like the anharmonic mode itself in [figure A1.2.6](#). However, because of the unavoidable

quantum mechanical zero-point energy associated with the action $\hbar/2$ in each mode, an overtone state actually has a certain amount of energy in all of the normal modes. Therefore, classically, the overtone of the antisymmetric stretch corresponds to a box-like trajectory, with most of the extension along the antisymmetric stretch, but with some extension along the symmetric stretch, and corresponding to the irreducible zero-point energy.

-12-

The other kind of quantum level we considered above is one with quanta in more than one mode, i.e. $(n_1 \dots n_N)$ with more than one of the n_i not equal to zero. Such a state is called a *combination* level. This corresponds, classically, to a box-like trajectory with extension in each mode corresponding to the number of quanta; an example was seen in [figure A1.2.7](#).

What does one actually observe in the experimental spectrum, when the levels are characterized by the set of quantum numbers $(n_1 \dots n_N)$ for the normal modes? The most obvious spectral observation is simply the set of energies of the levels; another important observable quantity is the intensities. The latter depend very sensitively on the type of probe of the molecule used to obtain the spectrum; for example, the intensities in absorption spectroscopy are in general far different from those in Raman spectroscopy. From now on we will focus on the energy levels of the spectrum, although the intensities most certainly carry much additional information about the molecule, and are extremely interesting from the point of view of theoretical dynamics.

If the molecule really had harmonic normal modes, the energy formula ([A1.2.6](#)) would apply and the spectrum would be extremely simple. It is common to speak of a *progression* in a mode i ; a progression consists of the series of levels containing the fundamental, with $n_i = 1$, along with the overtone levels $n_i > 1$. Each progression of a harmonic system would consist of equally spaced levels, with the level spacing given by the frequency ω_i . It is also common to speak of sequences, in which the sum of the number of quanta in two modes is fixed. In a harmonic spectrum, the progressions and sequences would be immediately evident to the eye in a plot of the energy levels.

In a system with anharmonic normal modes, the spectral pattern is not so simple. Instead of the simple energy level formula ([A1.2.6](#)), in addition to the harmonic terms there are anharmonic terms, similar to the terms $\gamma_1 (n + \frac{1}{2})^2$, $\gamma_2 (n + \frac{1}{2})^3$, ... in ([A1.2.4](#)). For each mode i , there is a set of such terms $\gamma_{ii}(n_i + \frac{1}{2})^2$, $\gamma_{iii}(n_i + \frac{1}{2})^3$, etc, where now by common convention the i 's in the subscript refer to mode i and the order of the subscript and superscript match, for example γ_{ii} with the quadratic power $(n_i + \frac{1}{2})^2$. However, there are also *cross terms* $\gamma_{ij}(n_i + \frac{1}{2})(n_j + \frac{1}{2})$, $\gamma_{ijj}(n_i + \frac{1}{2})^2(n_j + \frac{1}{2})$, etc. As an example, the anharmonic energy level formula for just a symmetric and antisymmetric stretch is given to the second order in the quantum numbers by

$$\begin{aligned}
 E(n_s, n_a) = & \omega_s(n_s + \frac{1}{2}) + \omega_a(n_a + \frac{1}{2}) + \omega_b(n_b + \frac{1}{2}) + \gamma_{ss}(n_s + \frac{1}{2})^2 \\
 & + \gamma_{aa}(n_a + \frac{1}{2})^2 + \gamma_{bb}(n_b + \frac{1}{2})^2 + \gamma_{sa}(n_s + \frac{1}{2})(n_a + \frac{1}{2}) \\
 & + \gamma_{sb}(n_s + \frac{1}{2})(n_b + \frac{1}{2}) + \gamma_{ab}(n_a + \frac{1}{2})(n_b + \frac{1}{2}).
 \end{aligned}
 \tag{A 1.2.7}$$

An energy expression for a polyatomic in powers of the quantum numbers like ([A1.2.7](#)) is an example of an anharmonic expansion [[18](#)]. In the anharmonic spectrum, within a progression or sequence there will not be equal spacings between levels; rather, the spacings will depend on the quantum numbers of the adjacent levels. Nonetheless, the spectrum will appear very regular to the eye. Spectra that follow closely a formula such as ([A1.2.7](#)), perhaps including higher powers in the quantum numbers, are very common in the spectroscopy of polyatomic molecules at relatively low energy near the minimum of the PES. This regularity is not too surprising, when one recalls that it is associated with the existence of the good quantum numbers $(n_1 \dots n_N)$, which themselves correspond classically to regular motion of the kind shown in [figure A1.2.7](#).

A 1.2.7 SPECTRA THAT ARE NOT SO REGULAR

If this was all there is to molecular spectra they would be essentially well understood by now and their power to give information on molecules nearly exhausted. However, this cannot be the case: consider that molecules dissociate—a very irregular type of motion!—while a molecule whose spectrum strictly followed a formula such as (A1.2.7) would have quantum levels all corresponding semiclassically to motion on invariant tori that are described by the N anharmonic normal modes. Motion as simple as this is expected only near a potential minimum, where the Weinstein–Moser theorem applies. How is the greater complexity of real molecules manifested in a spectrum? The spectrum is a reflection of the physical PES, since the vibrational spectrum is determined quantum mechanically by the PES. Since the PES contains the possibility of much less regular motion than that reflected in a Dunham formula such as (A1.2.7), how can a Dunham formula be modified so as to represent a real spectrum, including portions corresponding to less regular motion? We will consider first what these modifications must look like, then pose the following question: suppose we have a generalized spectral Hamiltonian and use this to represent experimental observations, how can we use this representation to *decode* the dynamical information on the internal molecular motions that is contained in the spectrum?

A 1.2.8 RESONANCE COUPLINGS

The fact that terms in addition to those present in the energy level formula (A1.2.7) might arise in molecular spectra is already strongly suggested by one of the features already discussed; the cross-anharmonic terms such as $\gamma_{ij}(n_i + \frac{1}{2})(n_j + \frac{1}{2})$. These terms show that the anharmonicity arises not only from the normal modes themselves—the ‘self-anharmonicity’ terms like $\gamma_{ii}(n_i + \frac{1}{2})^2$ —but also from *couplings between the normal modes*. The cross-anharmonic terms depend only on the vibrational quantum numbers—the Hamiltonian so far is diagonal in the normal mode quantum numbers. However, there are also terms in the generalized Hamiltonian that are not diagonal in the quantum numbers. It is these that are responsible for profoundly greater complexity of the internal motion of a polyatomic, as compared to a diatomic.

Consider how these non-diagonal terms would arise in the analysis of an experimental spectrum. Given a set of spectral data, one would try to fit the data to a Hamiltonian of the form of (A1.2.7). The Hamiltonian then is to be regarded as a ‘phenomenological’ or ‘effective’ spectroscopic Hamiltonian, to be used to describe the results of experimental observations. The fitting consists of adjusting the parameters of the Hamiltonian, for example ω ’s, the γ ’s, etc, until the best match possible is obtained between the spectroscopic Hamiltonian and the data. If a good fit is not obtained with a given number of terms in the Dunham expansion, one could simply add terms of higher order in the quantum numbers. However, it is found in fitting the spectrum of the stretch modes of a molecule like H₂O that this does not work at all well. Instead, a large *resonance coupling* term which *exchanges quanta* between the modes is found to be necessary to obtain a good fit to the data, as was first discovered long ago by Darling and Dennison [31]. Specifically, the Darling–Dennison coupling takes two quanta out of the symmetric stretch, and places two into the antisymmetric stretch. There is also a coupling which does the reverse, taking two quanta from the antisymmetric stretch and placing them into the symmetric stretch. It is convenient to represent this coupling in terms of the raising and lowering operators [32] a_i^+ , a_i . These, respectively, have the action of placing a quantum into or removing a quantum from an oscillator which originally has n quanta:

$$a^+|n\rangle = |n+1\rangle \quad a|n\rangle = |n-1\rangle \quad (\text{A 1.2.8})$$

The raising and lowering operators originated in the algebraic theory of the quantum mechanical oscillator, essentially by the path followed by Heisenberg in formulating quantum mechanics [33]. In terms of raising and lowering operators, the Darling–Dennison coupling operator is

$$\kappa_{\text{DD}}(a_s^+ a_s^+ a_a a_a + a_s a_s a_a^+ a_a^+) \quad (\text{A 1.2.9})$$

where κ_{DD} is a parameter which defines the strength of the coupling; κ_{DD} is optimized to obtain the best possible fit between the data and the spectroscopic Hamiltonian.

Physically, why does a term like the Darling–Dennison coupling arise? We have said that the spectroscopic Hamiltonian is an abstract representation of the more concrete, physical Hamiltonian formed by letting the nuclei in the molecule move with specified initial conditions of displacement and momentum on the PES, with a given total kinetic plus potential energy. This is the sense in which the spectroscopic Hamiltonian is an ‘effective’ Hamiltonian, in the nomenclature used above. The concrete Hamiltonian that it mimics is expressed in terms of particle momenta and displacements, in the representation given by the normal coordinates. Then, in general, it may contain terms proportional to all the powers of the products of the normal coordinates $R_i^{\eta_l} R_j^{\eta_p}$. (It will also contain terms containing the momenta that arise from the kinetic energy; however, these latter kinetic energy terms are more restricted in form than the terms from the potential.) In the spectroscopic Hamiltonian, these will partly translate into expressions with terms proportional to the powers of the quantum numbers, as in (A1.2.7). However, there will also be resonance couplings, such as the Darling–Dennison coupling (A1.2.9). These arise directly from the fact that the oscillator raising and lowering operators (A1.2.8) have a close connection to the position and momentum operators of the oscillator [32], so the resonance couplings are implicit in the terms of the physical Hamiltonian such as $R_i^{\eta_l} R_j^{\eta_p}$.

Since all powers of the coordinates appear in the physical PES, and these give rise to resonance couplings, one might expect a large, in fact infinite, number of resonance couplings in the spectroscopic Hamiltonian. However, in practice, a small number of resonance couplings—and often none, especially at low energy—is sufficient to give a good fit to an experimental spectrum, so effectively the Hamiltonian has a rather simple form. To understand why a small number of resonance couplings is usually sufficient we will focus again on H₂O.

In fitting the H₂O stretch spectrum, it is found that the Darling–Dennison coupling is necessary to obtain a good fit, but *only* the Darling–Dennison and no other. (It turns out that a second coupling, between the symmetric stretch and bend, is necessary to obtain a good fit when significant numbers of bending quanta are involved; we will return to this point later.) If all resonance terms in principle are involved in the Hamiltonian, why it is that, empirically, only the Darling–Dennison coupling is important? To understand this, a very important notion, the *polyad quantum number*, is necessary.

A 1.2.9 POLYAD NUMBER

The characteristic of the Darling–Dennison coupling is that it exchanges two quanta between the symmetric and antisymmetric stretches. This means that the individual quantum numbers n_s, n_a are no longer good quantum numbers of the Hamiltonian containing V_{DD} . However, the *total* number of stretch quanta

$$n_{\text{str}} = (n_s + n_a) \quad (\text{A } 1.2.10)$$

is left unchanged by V_{DD} . Thus, while it might appear that V_{DD} has destroyed two quantum numbers, corresponding to two constants of motion, it has in fact preserved n_{str} as a good quantum number, often referred to as a *polyad* quantum number. So, the Darling–Dennison term V_{DD} couples together a set of zero-order states with common values of the polyad number n_{str} . For example, the set with $n_{\text{str}} = 4$ contains zero-order states $[n_s, n_a] = [4, 0], [3, 1], [2, 2], [1, 3], [0, 4]$. These five, zero-order states are referred to as the zero-order polyad with $n_{\text{str}} = 4$.

If only zero-order states from the same polyad are coupled together, this constitutes a fantastic simplification in the Hamiltonian. Enormous computational economies result in fitting spectra, because the spectroscopic Hamiltonian is block diagonal in the polyad number. That is, only zero-order states within blocks with the same polyad number are coupled; the resulting small matrix diagonalization problem is vastly simpler than diagonalizing a matrix with all the zero-order states coupled to each other.

However, why should such a simplification be a realistic approximation? For example, why should not a coupling of the form

$$(a_s^+ a_s^+ a_s^+ a_a + a_s a_s a_s a_a^+) \quad (\text{A } 1.2.11)$$

which would break the polyad number n_{str} , be just as important as V_{DD} ? There is no reason *a priori* why it might not have just as large a contribution as V_{DD} when the coordinate representation of the PES is expressed in terms of the raising and lowering operators a_i^+ , a_i . To see why it nonetheless is found empirically to be unimportant in the fit, and therefore is essentially negligible, consider again the molecule H_2O . A coupling like (A1.2.11), which removes three quanta from one mode but puts only one quantum in the other mode, is going to couple zero-order states with vastly different zero-order energy. For example, $[n_s, n_a] = [3, 0]$ will be coupled to $[0, 1]$, but these zero-order states are nowhere near each other in energy. By general quantum mechanical arguments of perturbation theory [32], the coupling of states which differ greatly in energy will have a correspondingly small effect on the wavefunctions and energies. In a molecule like H_2O , such a coupling can essentially be ignored in the fitting Hamiltonian.

This is why the coupling V_{DD} is often called a Darling–Dennison *resonance* coupling: it is significant precisely when it couples zero-order states that differ by a small number of quanta which are approximately degenerate with each other, which classically is to say that they are in resonance. The Darling–Dennison coupling, because it involves taking two quanta from one mode and placing two in another, is also called a 2:2 coupling. Other orders of coupling $n:m$ also arise in different situations (such as the stretch–bend coupling in H_2O), and these will be considered later.

However, if *only* the Darling–Dennison coupling is important for the coupled stretches, what is its importance telling us about the internal molecular motion? It turns out that the right kind of analysis of the spectroscopic fitting Hamiltonian reveals a vast amount about the dynamics of the molecule: it allows us to decipher the story encoded in the spectrum of what the molecule is ‘really doing’ in its internal motion. We will approach this ‘spectral cryptology’ from two complementary directions:

the spectral pattern of the Darling–Dennison spectroscopic Hamiltonian; and, less directly, the analysis of a classical Hamiltonian corresponding to the spectroscopic quantum Hamiltonian. We will see that the Darling–

Dennison coupling produces a pattern in the spectrum that is very distinctly different from the pattern of a ‘pure normal modes Hamiltonian’, without coupling, such as (A1.2.7). Then, when we look at the classical Hamiltonian corresponding to the Darling–Dennison quantum fitting Hamiltonian, we will subject it to the mathematical tool of bifurcation analysis [34]. From this, we will infer a dramatic birth in bifurcations of new ‘natural motions’ of the molecule, i.e. *local modes*. This will be directly connected with the distinctive quantum spectral pattern of the polyads. Some aspects of the pattern can be accounted for by the classical bifurcation analysis; while others give evidence of intrinsically non-classical effects in the quantum dynamics.

It should be emphasized here that while the discussion of contemporary techniques for decoding spectra for information on the internal molecular motions will largely concentrate on spectroscopic Hamiltonians and bifurcation analysis, there are distinct, but related, contemporary developments that show great promise for the future. For example approaches using advanced ‘algebraic’ techniques [35, 36] for alternative ways to build the spectroscopic Hamiltonian, and ‘hierarchical analysis’ using techniques related to general classification methods [37].

A 1.2.10 SPECTRAL PATTERN OF THE DARLING–DENNISON HAMILTONIAN

Consider the polyad $n_{\text{str}} = 6$ of the Hamiltonian (A1.2.7). This polyad contains the set of levels conventionally assigned as $[6, 0]$, $[5, 1]$, . . . , $[0, 6]$. If a Hamiltonian such as (A1.2.7) described the spectrum, the polyad would have a pattern of levels with monotonically varying spacing, like that shown in figure A1.2.8. However, suppose the fit of the experimental spectrum requires the addition of a strong Darling–Dennison term V_{DD} , as empirically is found to be the case for the stretch spectrum of a molecule like H_2O . In general, because of symmetry, only certain levels may be spectroscopically allowed; for example, in absorption spectra, only levels with odd number of quanta n_a in the antisymmetric stretch. However, diagonalization of the polyad Hamiltonian gives all the levels of the polyad. When these are plotted for the Darling–Dennison Hamiltonian, including the spectroscopically unobserved levels with even n_a , a striking pattern, shown in figure A1.2.9, is immediately evident. At the top of the polyad the level spacing pattern is like that of the anharmonic normal modes, as in figure A1.2.8, but at the bottom of the polyad the levels come in near-degenerate doublets. What is this pattern telling us about the change in the internal molecular motion resulting from inclusion of the Darling–Dennison coupling?

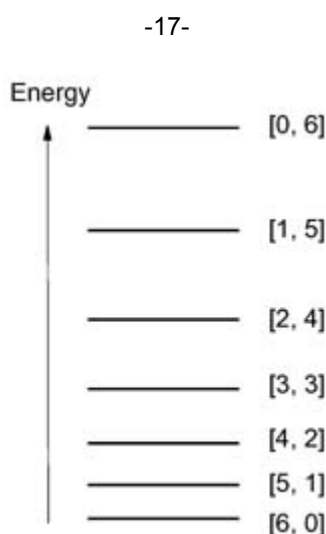


Figure A1.2.8. Typical energy level pattern of a sequence of levels with quantum numbers $[n_s, n_a]$ for the number of quanta in the symmetric and antisymmetric stretch. The bend quantum number is neglected and may be taken as fixed for the sequence. The total number of quanta ($n_s + n_a = 6$) is the polyad number, which

is the same for all levels. $[6, 0]$ and $[0, 6]$ are the overtones of the symmetric and antisymmetric stretch; the other levels are combination levels. The levels have a monotonic sequence of energy spacings from bottom to top.



Figure A1.2.9. Energy level pattern of polyad 6 of a spectroscopic Hamiltonian for coupled stretches with strong Darling–Dennison coupling. Within the polyad the transition from normal to local modes is evident. At the bottom of the polyad are two nearly degenerate pairs of levels. Semiclassically, the bottom pair derive from local mode overtone states. The levels are symmetrized mixtures of the individual local mode overtones. Semiclassically, they are exactly degenerate; quantum mechanically, a small splitting is present, due to tunnelling. The next highest pair are symmetrized local mode combination states. The tunnelling splitting is larger than in the bottom pair; above this pair, the levels have normal mode character, as evidenced by the energy level pattern.

-18-

This has been the subject of a great deal of work by many people over more than 20 years. Breakthroughs in the theoretical understanding of the basic physics began to accumulate in the early 1980s [38, 39, 40 and 41]. One approach that has a particularly close relation between experiment and theory uses bifurcation analysis of a classical analogue of the spectroscopic fitting Hamiltonian. The mathematical details are presented elsewhere [42, 43, 44 and 45]; the qualitative physical meaning is easily described.

A classical Hamiltonian is obtained from the spectroscopic fitting Hamiltonian by a method that has come to be known as the ‘Heisenberg correspondence’ [46], because it is closely related to the techniques used by Heisenberg in fabricating the form of quantum mechanics known as matrix mechanics.

Once the classical Hamiltonian has been obtained, it is subjected to bifurcation analysis. In a bifurcation, typically, a stable motion of the molecule—say, one of the Weinstein–Moser normal modes—suddenly becomes unstable; and new stable, anharmonic modes suddenly branch out from the normal mode. An illuminating example is presented in [figure A1.2.10](#) which illustrates the results of the bifurcation analysis of the classical version of the Darling–Dennison Hamiltonian. One of the normal modes—it can be either the symmetric or antisymmetric stretch depending on the specific parameters found empirically in the fitting Hamiltonian—remains stable. Suppose it is the antisymmetric stretch that remains stable. At the bifurcation, the symmetric stretch suddenly becomes unstable. This happens at some critical value of the mathematical ‘control parameters’ [34], which we may take to be some critical combination of the energy and polyad number. From the unstable symmetric stretch, there immediately emerge two new stable periodic orbits, or anharmonic modes. As the control parameter is increased, the new stable modes creep out from the symmetric

stretch—which remains in ‘fossilized’ form as an unstable periodic orbit. Eventually, the new modes point more or less along the direction of the zero-order bond displacements, but as curvilinear trajectories. We can say that in this bifurcation, anharmonic local modes have been born.

It is the ‘skeleton’ of stable and unstable modes in [figure A1.2.10\(c\)](#) that explains the spectral pattern seen in [figure A1.2.9](#). Some of the levels in the polyad, those in the upper part, have wavefunctions that are quantized in patterns that shadow the normal modes—the still-stable antisymmetric stretch and the now-unstable symmetric stretch. Other states, the lower ones in the polyad, are quantized along the local modes. These latter states, described by *local mode quantum numbers*, account for the pattern of near-degenerate doublets. First, why is the degeneracy there at all? The two classical local modes have exactly the same energy and frequency, by symmetry. In the simplest semiclassical [29] picture, there are two *exactly* degenerate local mode overtones, each pointed along one or the other of the local modes. There are also combination states possible with quanta in each of the local modes and, again, semiclassically these must come in exactly degenerate pairs.

-19-

Normal-Local Bifurcation

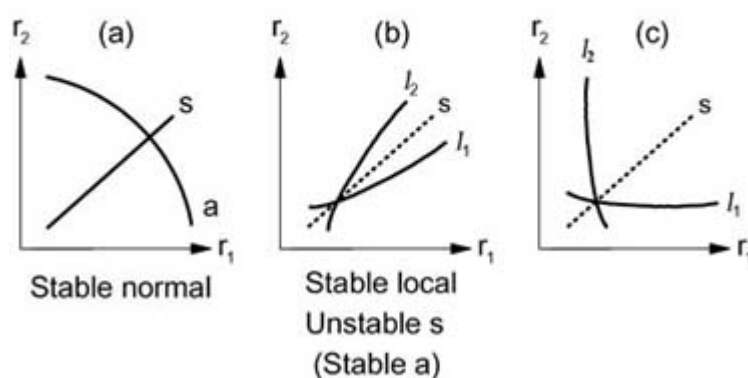


Figure A1.2.10. Birth of local modes in a bifurcation. In (a), before the bifurcation there are stable anharmonic symmetric and antisymmetric stretch modes, as in [figure A1.2.6](#). At a critical value of the energy and polyad number, one of the modes, in this example the symmetric stretch, becomes unstable and new stable local modes are born in a bifurcation; the system is shown shortly after the bifurcation in (b), where the new modes have moved away from the unstable symmetric stretch. In (c), the new modes clearly have taken the character of the anharmonic local modes.

The classical bifurcation analysis has succeeded in decoding the spectrum to reveal the existence of local and normal modes, and the local modes have accounted for the changeover from a normal mode spectral pattern to the pattern of degenerate doublets. But why the *splitting* of the near-degenerate doublets? Here, non-classical effects unique to quantum mechanics come into play. A trajectory in the box for one of the local modes is confined in phase space to an invariant torus, and classically will never leave its box. However, quantum mechanically, there is some probability for classically forbidden processes to take place in which the trajectory jumps from one box to the other! This may strike the reader as akin to the quantum mechanical phenomenon of tunnelling. In fact, this is more than an analogy. The effect has been called ‘dynamical tunnelling’ [41, 47], and it can be formulated rigorously as a mathematical tunnelling problem [40, 48]. The effect of the dynamical tunnelling on the energy levels comes through in another unique manifestation of quantum mechanics. The quantum eigenfunctions—the wavefunctions for the energy levels of the true quantum spectrum—are symmetrized combinations of the two semiclassical wavefunctions corresponding to the two classical boxes [38]. These wavefunctions come in pairs of + and – symmetry; the two levels of a near-degenerate pair are split into a +state and a –state. The amount of the splitting is directly related to the

rate of the non-classical tunnelling process [49].

A 1.2.11 FERMI RESONANCES

In the example of H₂O, we saw that the Darling–Dennison coupling between the stretches led to a profound change in the internal dynamics; the birth of local modes in a bifurcation from one of the original low-energy normal modes. The question arises of the possibility of other types of couplings, if not between two identical stretch modes, then between other kinds of modes. We have seen that, effectively, only a very small subset of possible resonance couplings between

-20-

the stretches is actually important; in the case of the H₂O stretches, only the 2:2 Darling–Dennison coupling. This great simplification came about because of the necessity to satisfy a condition of frequency resonance between the zero-order modes for the 2:2 Darling–Dennison coupling to be important. In H₂O, there is also an approximate 2:1 resonance condition satisfied between the stretch and bend frequencies. Not surprisingly, in fitting the H₂O spectrum, in particular when several bending quanta are present, it is necessary to consider a 2:1 coupling term between the symmetric stretch (*s*) and bend (*b*), of the form

$$\kappa_{sbb}(a_s^+ a_b a_b + a_s a_b^+ a_b^+). \quad (\text{A 1.2.12})$$

(The analogous coupling between the antisymmetric stretch and bend is forbidden in the H₂O Hamiltonian because of symmetry.) The 2:1 resonance is known as a ‘Fermi resonance’ after its introduction [50] in molecular spectroscopy. The 2:1 resonance is often very prominent in spectra, especially between stretch and bend modes, which often have approximate 2:1 frequency ratios. The 2:1 coupling leaves unchanged as a polyad number the sum:

$$n_{sb} = (n_s + n_b/2). \quad (\text{A 1.2.13})$$

Other resonances, of order *n:m*, are possible in various systems. Another type of resonance is a ‘multimode’ resonance. For example, in C₂H₂ the coupling

$$\kappa_{2345}(a_3^+ a_2 a_4 a_5 + a_3 a_2^+ a_4^+ a_5^+) \quad (\text{A 1.2.14})$$

that transfers one quantum from the antisymmetric stretch ν_3 to the C–C stretch ν_2 and each of the bends ν_4 and ν_5 is important [51, 52 and 53]. Situations where couplings such as the *n:m* resonance and the 2345 multimode resonance need to be invoked are often referred to as ‘Fermi resonances’, though some authors restrict this term to the 2:1 resonance and use the term ‘anharmonic resonance’ to describe the more general *n:m* or multimode cases. Here, we will use the terms ‘Fermi’ and ‘anharmonic’ resonances interchangeably.

It turns out that the language of ‘normal and local modes’ that emerged from the bifurcation analysis of the Darling–Dennison Hamiltonian is not sufficient to describe the general Fermi resonance case, because the bifurcations are qualitatively different from the normal-to-local bifurcation in [figure A1.2.10](#). For example, in 2:1 Fermi systems, one type of bifurcation is that in which ‘resonant collective modes’ are born [54]. The resonant collective modes are illustrated in [figure A1.2.11](#) their difference from the local modes of the Darling–Dennison system is evident. Other types of bifurcations are also possible in Fermi resonance systems; a detailed treatment of the 2:1 resonance can be found in [44].

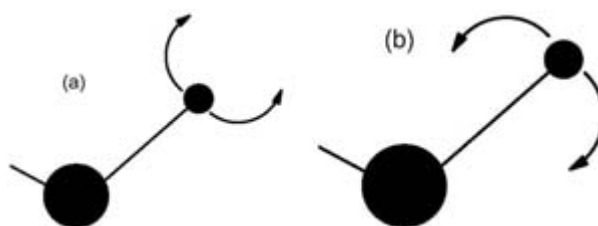


Figure A1.2.11. Resonant collective modes of the 2:1 Fermi resonance system of a coupled stretch and bend with an approximate 2:1 frequency ratio. Shown is one end of a symmetric triatomic such as H_2O . The normal stretch and bend modes are superseded by the horseshoe-shaped modes shown in (a) and (b). These two modes have different frequency, as further illustrated in [figure A1.2.12](#).

A 1.2.12 MORE SUBTLE ENERGY LEVEL PATTERNS

The Darling–Dennison Hamiltonian displayed a striking energy level pattern associated with the bifurcation to local modes: approximately degenerate local mode doublets, split by dynamical tunnelling. In general Fermi resonance systems, the spectral hallmarks of bifurcations are not nearly as obvious. However, subtle, but clearly observable spectral markers of bifurcations do exist. For example, associated with the formation of resonant collective modes in the 2:1 Fermi system there is a pattern of a minimum in the spacing of adjacent energy levels within a polyad [55], as illustrated in [figure A1.2.12](#). This pattern has been invoked [56, 57] in the analysis of ‘isomerization spectra’ of the molecule HCP, which will be discussed later. Other types of bifurcations have their own distinct, characteristic spectral patterns; for example, in 2:1 Fermi systems a second type of bifurcation has a pattern of alternating level spacings, of a ‘fan’ or a ‘zigzag’, which was predicted in [55] and subsequently s [57].

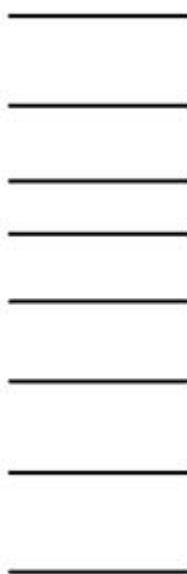


Figure A1.2.12. Energy level pattern of a polyad with resonant collective modes. The top and bottom energy levels correspond to overtone motion along the two modes shown in [figure A1.2.11](#), which have a different frequency. The spacing between adjacent levels decreases until it reaches a minimum between the third and fourth levels from the top. This minimum is the hallmark of a separatrix [29, 45] in phase space.

A 1.2.13 MULTIPLE RESONANCES IN POLYATOMICS

Implicit in the discussion of the Darling–Dennison and Fermi resonances has been the assumption that we can isolate each individual resonance, and consider its bifurcations and associated spectral patterns separately from other resonances in the system. However, strictly speaking, this cannot be the case. Consider again H_2O . The Darling–Dennison resonance couples the symmetric and antisymmetric stretches; the Fermi resonance couples the symmetric stretch and bend. Indirectly, all three modes are coupled, and the two resonances are linked. It is no longer true that the stretch polyad number ($n_s + n_a$) is conserved, because it is broken by the 2:1 Fermi coupling; nor is the Fermi polyad number ($n_s + n_b/2$) preserved, because it is broken by the Darling–Dennison coupling. However, there is still a generalized ‘total’ polyad number

$$n_{\text{total}} = (n_s + n_a + n_b/2) \quad (\text{A 1.2.15})$$

that is conserved by both couplings, as may readily be verified. All told, the Hamiltonian with both couplings has two constants of motion, the energy and the polyad number (A1.2.15). A system with fewer constants than the number of degrees of freedom, in this case two constants and three degrees of freedom, is ‘non-integrable’, in the language of classical mechanics [29]. This means that, in general, trajectories do not lie on higher-dimensional invariant tori; instead, they may be chaotic, and in fact this is often observed to be the case [58, 59] in trajectories of the semiclassical Hamiltonian for H_2O .

-23-

Nonetheless, it is still possible to perform the bifurcation analysis on the multiresonance Hamiltonian. In fact, the existence of the polyad number makes this almost as easy, despite the presence of chaos, as in the case of an isolated single Fermi or Darling–Dennison resonance. It is found [60] that most often (though not always), the same qualitative bifurcation behaviour is seen as in the single resonance case, explaining why the simplified individual resonance analysis very often is justified. The bifurcation analysis has now been performed for triatomics with two resonances [60] and for C_2H_2 with a number of resonances [61].

A 1.2.14 POTENTIAL AND EXPERIMENT: CLOSING THE CIRCLE

We have alluded to the connection between the molecular PES and the spectroscopic Hamiltonian. These are two very different representations of the molecular Hamiltonian, yet both are supposed to describe the same molecular dynamics. Furthermore, the PES often is obtained via *ab initio* quantum mechanical calculations; while the spectroscopic Hamiltonian is most often obtained by an empirical fit to an experimental spectrum. Is there a direct link between these two seemingly very different ways of apprehending the molecular Hamiltonian and dynamics? And if so, how consistent are these two distinct ways of viewing the molecule?

There has been a great deal of work [62, 63] investigating how one can use perturbation theory to obtain an effective Hamiltonian like the spectroscopic Hamiltonian, starting from a given PES. It is found that one can readily obtain an effective Hamiltonian in terms of normal mode quantum numbers and coupling. Furthermore, the actual Hamiltonians obtained very closely match those obtained via the empirical fitting of spectra! This consistency lends great confidence that both approaches are complementary, mutually consistent ways of apprehending real information on molecules and their internal dynamics.

Is it possible to approach this problem the other way, from experiment to the molecular PES? This is difficult

to answer in general, because ‘inversion’ of spectra is not a very well-posed question mathematically. Nonetheless, using spectra to gain information on potentials has been pursued with great vigor. Even for diatomics, surprising new, mathematically powerful methods are being developed [64]. For polyatomics, it has been shown [65] how the effective spectroscopic Hamiltonian is a very useful way-station on the road from experiment back to the PES. This closes the circle, because it shows that one can go from an assumed PES to the effective Hamiltonian derived via perturbation theory; or take the opposite path from the experimentally obtained effective spectroscopic Hamiltonian to the PES.

A 1.2.15 POLYAD QUANTUM NUMBERS IN LARGER SYSTEMS

We have seen that resonance couplings destroy quantum numbers as constants of the spectroscopic Hamiltonian. With both the Darling–Dennison stretch coupling and the Fermi stretch–bend coupling in H₂O, the individual quantum numbers n_s , n_a and n_b were destroyed, leaving the total polyad number $(n_s + n_a + n_b/2)$ as the only remaining quantum number. We can ask: (1) Is there also a good polyad number in larger molecules? (2) If so, how robust is this quantum number? For example, how high in energy does it persist as the molecule approaches dissociation or a barrier to isomerization? (3) Is the total polyad number the only good vibrational quantum number left over after the resonances have been taken into account, or can there be others?

-24-

It may be best to start with question (3). Given the set of resonance coupling operators found to be necessary to obtain a good fit of an experimental spectrum, it can be shown that the resonance couplings may be represented as vectors, which are not necessarily orthogonal. This leads to a simple but very powerful ‘resonance vector analysis’ [62, 66, 67]. The original vector space of the normal mode coordinates has N dimensions. The subspace spanned by the resonance vectors is the space of the vibrational quantum numbers that was destroyed; the complement of this space gives the quantities that remain as good quantum numbers. In general, there can be more than one such quantum number; we will encounter an example of this in C₂H₂, and see that it has important implications for the internal molecular dynamics. The set of good quantum numbers may contain one or more of the original individual normal mode quantum numbers; but in general, the good constants are combinations of the original quantum numbers. Examples of this are the polyad numbers that we have already encountered.

The resonance vector analysis has been used to explore all of the questions raised above on the fate of the polyad numbers in larger molecules, the most thoroughly investigated case so far probably being C₂H₂. This molecule has been very extensively probed by absorption as well as stimulated emission pumping and dispersed fluorescence techniques [52, 53, 68, 69, 70 and 71], the experimental spectra have been analysed in great detail and the fits to data have been carefully refined with each new experiment. A large number of resonance coupling operators has been found to be important, a good many more than the number of vibrational modes, which are seven in number: a symmetric C–H stretch ν_1 , antisymmetric C–H stretch ν_3 , C–C stretch ν_2 and two bends ν_4 and ν_5 , each doubly degenerate. Despite the plethora of couplings, the resonance vector analysis shows that the total polyad number

$$N_{\text{total}} = (5n_1 + 3n_2 + 5n_3 + n_4 + n_5) \quad (\text{A 1.2.16})$$

is a good quantum number up to at least about 15,000 cm⁻¹. This is at or near the barrier to the formation of the isomer vinylidene! (The coefficients 5, 3, 5, 1 and 1 in (A1.2.16) are close to the frequency ratios of the zero-order normal modes, which is to say, the polyad number satisfies a resonance condition, as in the earlier examples for H₂O.) The polyad number N_{total} has been used with great effect to identify remarkable order

[66, 67, 68, 69 and 70] in the spectrum: groups of levels can clearly be identified that belong to distinct polyads. Furthermore, there are additional ‘polyad’ constants—that is, quantum numbers that are combinations of the original quantum numbers—in addition to the total polyad number (A1.2.16). These additional constants have great significance for the molecular dynamics. They imply the existence of energy transfer pathways [67]. For example, in dispersed fluorescence spectra in which pure bending motion is excited, it has been found that with as many as 22 quanta of bend, all of the vibrational excitation remains in the bends on the time scale associated with dispersed fluorescence spectroscopy, with no energy transfer to the stretches [72].

A 1.2.16 ISOMERIZATION SPECTRA

We have spoken of the simplicity of the bifurcation analysis when the spectroscopic Hamiltonian possesses a good polyad number, and also of the persistence of the polyad number in C_2H_2 as the molecule approaches the barrier to isomerization to the species vinylidene. This suggests that it might be possible to use detailed spectra to probe the dynamics of a system undergoing an internal rearrangement. Several groups [56, 57] have been investigating the rearrangement of HCP to the configuration CPH, through analysis of the ‘isomerization spectrum’. Many of the tools described in this section, including decoding the dynamics through analysis of bifurcations and associated spectral

-25-

patterns, have come into play. The various approaches all implicate an ‘isomerization mode’ in the rearrangement process, quite distinct from any of the low-energy normal modes of the system. An explanation has been provided [57] in terms of the abrupt birth of the isomerization mode. This occurs at a bifurcation, in which the HCP molecule suddenly acquires a stable motion that takes it along the isomerization pathway, thereby altering the geometry and with it the rotational constant.

It should be emphasized that isomerization is by no means the only process involving chemical reactions in which spectroscopy plays a key role as an experimental probe. A very exciting topic of recent interest is the observation and computation [73, 74] of the spectral properties of the transition state [6]—catching a molecule ‘in the act’ as it passes the point of no return from reactants to products. Furthermore, it has been discovered from spectroscopic observation [75] that molecules can have motions that are stable for long times even *above* the barrier to reaction.

A 1.2.17 BREAKDOWN OF THE POLYAD NUMBERS

The polyad concept is evidently a very simple but powerful tool in the analysis and description of the internal dynamics of molecules. This is especially fortunate in larger molecules, where the intrinsic spectral complexity grows explosively with the number of atoms and degrees of freedom. Does the polyad number ever break down? Strictly speaking, it must: the polyad number is only an approximate property of a molecule’s dynamics and spectrum. The actual molecular Hamiltonian contains resonance couplings of all forms, and these must destroy the polyad numbers at some level. This will show up by looking at high enough resolution at a spectrum which at lower resolution has a good polyad number. Levels will be observed of small intensity, which would be rigorously zero if the polyad numbers were exact. The fine detail in the spectrum corresponds to long-time dynamics, according to the time–energy uncertainty relation [49].

One reason the polyad-breaking couplings are of interest is because they govern the long-time intramolecular energy flow, which is important for theories on reaction dynamics. These are considered elsewhere in this Encyclopedia and in monographs [6] and will not be considered further here. The long-time energy flow may

also be important for efforts of coherent control and for problems of energy flow from a molecule to a bath, such as a surrounding liquid. Both of these will be considered later.

Several questions arise on the internal dynamics associated with the breakdown of the polyad number. We can only speculate in what follows, awaiting the illumination of future research.

When the polyad number breaks down, as evidenced by the inclusion of polyad-breaking terms in the spectroscopic Hamiltonian, what is the residue left in the spectrum of the polyads as approximately conserved entities? There is already some indication [76] that the polyad organization of the spectrum will still be evident even with the inclusion of weak polyad-breaking terms. The identification of these polyad-breaking resonances will be a challenge, because each such resonance probably only couples a given polyad to a very small subset of ‘dark’ states of the molecule that lie outside those levels visible in the polyad spectrum. There will be a large number of such resonances, each of them coupling a polyad level to a small subset of dark levels.

Another question is the nature of the changes in the classical dynamics that occur with the breakdown of the polyad number. In all likelihood there are further bifurcations. Apart from the identification of the individual polyad-breaking resonances, the bifurcation analysis itself presents new challenges. This is partly because with the breakdown

-26-

of the polyad number, the great computational simplicity afforded by the block-diagonalization of the Hamiltonian is lost. Another problem is that the bifurcation analysis is exactly solvable only when a polyad number is present [45], so approximate methods will be needed.

When the polyad number breaks down, the bifurcation analysis takes on a new kind of interest. The approximate polyad number can be thought of as a type of ‘bottleneck’ to energy flow, which is restricted to the phase space of the individual polyad; the polyad breakdown leads to energy flow in the full phase space. We can think of the goal as the search for the ‘energy transfer modes’ of long-time energy flow processes in the molecule, another step beyond the current use of bifurcation analysis to find the natural anharmonic modes that emerge within the polyad approximation.

The existence of the polyad number as a bottleneck to energy flow on short time scales is potentially important for efforts to control molecular reactivity using advanced laser techniques, discussed below in [section A1.2.20](#). Efforts at control seek to intervene in the molecular dynamics to prevent the effects of widespread vibrational energy flow, the presence of which is one of the key assumptions of Rice–Ramsperger–Kassel–Marcus (RRKM) and other theories of reaction dynamics [6].

In connection with the energy transfer modes, an important question, to which we now turn, is the significance of classical chaos in the long-time energy flow process, in particular the relative importance of chaotic classical dynamics, versus classically forbidden processes involving ‘dynamical tunnelling’.

A 1.2.18 CLASSICAL VERSUS NON-CLASSICAL EFFECTS

To understand the internal molecular motions, we have placed great store in classical mechanics to obtain a picture of the dynamics of the molecule and to predict associated patterns that can be observed in quantum spectra. Of course, the classical picture is at best an imprecise image, because the molecular dynamics are intrinsically quantum mechanical. Nonetheless, the classical metaphor must surely possess a large kernel of truth. The classical structure brought out by the bifurcation analysis has accounted for real patterns seen in wavefunctions and also for patterns observed in spectra, such as the existence of local mode doublets, and the

more subtle level-spacing patterns seen in connection with Fermi resonance spectra.

However, we have also seen that some of the properties of quantum spectra are intrinsically non-classical, apart from the discreteness of quantum states and energy levels implied by the very existence of quanta. An example is the splitting of the local mode doublets, which was ascribed to dynamical tunnelling, i.e. processes which classically are forbidden. We can ask if non-classical effects are ubiquitous in spectra and, if so, are there manifestations accessible to observation other than those we have encountered so far? If there are such manifestations, it seems likely that they will constitute subtle peculiarities in spectral patterns, whose discernment and interpretation will be an important challenge.

The question of non-classical manifestations is particularly important in view of the chaos that we have seen is present in the classical dynamics of a multimode system, such as a polyatomic molecule, with more than one resonance coupling. Chaotic classical dynamics is expected to introduce its own peculiarities into quantum spectra [29, 77]. In H_2O , we noted that chaotic regions of phase space are readily seen in the classical dynamics corresponding to the spectroscopic Hamiltonian. How important are the effects of chaos in the observed spectrum, and in the wavefunctions of the molecule? In H_2O , there were some states whose wavefunctions appeared very disordered, in the region of the

-27-

phase space where the two resonances should both be manifesting their effects strongly. This is precisely where chaos should be most pronounced, and indeed this was observed to be the case [58]. However, close examination of the states in question by Keshavamurthy and Ezra [78] showed that the disorder in the quantum wavefunction was due not primarily to chaos, but to dynamical tunnelling, the non-classical effect invoked earlier to explain the splitting of local mode doublets.

This demonstrated importance of the non-classical processes in systems with intact polyad numbers prompts us to consider again the breakdown of the polyad number. Will it be associated mainly with chaotic classical diffusion, or non-classical effects? It has been suggested [47] that high-resolution structure in spectra, which we have said is one of the manifestations of the polyad breakdown, may be predominantly due to non-classical, dynamical tunnelling processes, rather than chaotic diffusion. Independent, indirect support comes from the observation that energy flow from vibrationally excited diatomic molecules in a liquid bath is predominantly due to non-classical effects, to the extent of several orders of magnitude [79]. Whether dynamical tunnelling is a far more important energy transfer mechanism within molecules than is classical chaos is an important question for the future exploration of the interface of quantum and classical dynamics.

It should be emphasized that the existence of ‘energy transfer modes’ hypothesized earlier with the polyad breakdown is completely consistent with the energy transfer being due to non-classical, dynamical tunnelling processes. This is evident from the observation above that the disorder in the H_2O spectrum is attributable to *non-classical* effects which nonetheless are accompaniments of *classical* bifurcations.

The general question of the spectral manifestations of classical chaos and of non-classical processes, and their interplay in complex quantum systems, is a profound subject worthy of great current and future interest. Molecular spectra can provide an immensely important laboratory for the exploration of these questions. Molecules provide all the necessary elements: a mixture of regular and chaotic classical motion, with ample complexity for the salient phenomena to make their presence known and yet sufficient simplicity and control in the number of degrees of freedom to yield intelligible answers. In particular, the fantastic simplification afforded by the polyad constants, together with their gradual breakdown, may well make the spectroscopic study of internal molecular motions an ideal arena for a fundamental investigation of the quantum–classical correspondence.

A 1.2.19 MOLECULES IN CONDENSED PHASE

So far we have considered internal motions mostly of isolated molecules, not interacting with an environment. This condition will be approximately met in a dilute gas. However, many of the issues raised may be of relevance in processes where the molecule is not isolated at all. An example already briefly noted is the transfer of vibrational energy from a molecule to a surrounding bath, for example a liquid. It has been found [79] that when a diatomic molecule such as O_2 is vibrationally excited in a bath of liquid oxygen, the transfer of vibrational energy is extremely slow. This is due to the extreme mismatch between the energy of an O_2 vibrational quantum, and the far lower energy of the bath's phonon modes—vibrations involving large numbers of the bath molecules oscillating together. Classically, the energy transfer is practically non-existent; semiclassical approximations, however, show that quantum effects increase the rate by orders of magnitude.

-28-

The investigation of energy transfer in polyatomic molecules immersed in a bath is just beginning. One issue has to do with energy flow *from* the molecule to the bath. Another issue is the effect of the bath on energy flow processes *within* the molecule. Recent experimental work [80] using ultrafast laser probes of ClO_2 immersed in solvents points to the importance of bifurcations within the solute triatomic for the understanding of energy flow both within and from the molecule.

For a polyatomic, there are many questions on the role of the polyad number in energy flow from the molecule to the bath. Does polyad number conservation in the isolated molecule inhibit energy flow to the bath? Is polyad number breaking a facilitator or even a *prerequisite* for energy flow? Finally, does the energy flow to the bath increase the polyad number breaking in the molecule? One can only speculate until these questions become accessible to future research.

A 1.2.20 LASER CONTROL OF MOLECULES

So far, we have talked about the internal motions of molecules which are exhibiting their 'natural' behaviour, either isolated in the gas phase or surrounded by a bath in a condensed phase. These natural motions are inferred from carefully designed spectroscopic experiments that are sufficiently mild that they simply probe what the molecule does when left to 'follow its own lights'. However, there is also a great deal of effort toward using high-intensity, carefully sculpted laser pulses which are anything but mild, in order to control the dynamics of molecules. In this quest, what role will be played by knowledge of their natural motions?

Surprisingly, a possible answer may be 'not much of a role at all'. One promising approach [81] using *coherent* light sources seeks to have the apparatus 'learn' how to control the molecule without knowing much at all about its internal properties in advance. Instead, a 'target' outcome is selected, and a large number of automated experiments performed, in which the control apparatus learns how to achieve the desired goal by rationally programmed trial and error in tailoring coherent light sources. It might not be necessary to learn much at all about the molecule's dynamics before, during or after, to make the control process work, even though the control apparatus might seem to all appearances to be following a cunning path to achieve its ends.

It can very well be objected that such a hit-or-miss approach, no matter how cleverly designed, is not likely to get very far in controlling polyatomic molecules with more than a very small number of atoms—in fact one will do much better by harnessing knowledge of the natural internal motions of molecules in tandem with the process of external control. The counter-argument can be made that in the trial and error approach, one will hit on the 'natural' way of controlling the molecule, even if one starts out with a method which at first tries nothing but brute force, even if one remains resolutely ignorant of why the molecule is responding to the evolving control procedure. Of course, if a good way is found to control the molecule, a retrospective explanation of how and why it worked almost certainly must invoke the natural motions of the molecule,

about which much will perhaps have been learned along the way in implementing the process of control.

The view of this author is that knowledge of the internal molecular motions, perhaps as outlined in this chapter, is likely to be important in achieving successful control, in approaches that make use of coherent light sources and quantum mechanical coherence. However, at this point, opinions on these issues may not be much more than speculation.

-29-

There are also approaches [82, 83 and 84] to control that have had marked success and which do not rely on quantum mechanical coherence. These approaches typically rely explicitly on a knowledge of the internal molecular dynamics, both in the design of the experiment and in the achievement of control. So far, these approaches have exploited only implicitly the very simplest types of bifurcation phenomena, such as the transition from local to normal stretch modes. If further success is achieved along these lines in larger molecules, it seems likely that deliberate knowledge and exploitation of more complicated bifurcation phenomena will be a matter of necessity.

As discussed in [section A1.2.17](#), the existence of the approximate polyad numbers, corresponding to short-time bottlenecks to energy flow, could be very important in efforts for laser control, apart from the separate question of bifurcation phenomena.

Another aspect of laser control of molecular dynamics is the use of control techniques to *probe* the internal motions of molecules. A full account of this topic is far beyond the scope of this section, but one very interesting case in point has important relations to other branches of physics and mathematics. This is the phenomenon of ‘geometric phases’, which are closely related to gauge theories. The latter were originally introduced into quantum physics from the classical theory of electromagnetism by Weyl and others (see [85]). Quantum field theories with generalizations of the electromagnetic gauge invariance were developed in the 1950s and have since come to play a paramount role in the theory of elementary particles [86, 87]. Geometric phases were shown to have directly observable effects in quantum phenomena such as the Aharonov–Bohm effect [88]. It was later recognized that these phases are a general phenomenon in quantum systems [89]. One of the first concrete examples was pointed out [90] in molecular systems involving the coupling of rotation and vibration. A very systematic exposition of geometric phases and gauge ideas in molecular systems was presented in [91]. The possibility of the direct optical observation of the effects of the geometric phases in the time domain through coherent laser excitations has recently been explored [92].

A 1.2.21 LARGER MOLECULES

This section has focused mainly on the internal dynamics of small molecules, where a coherent picture of the detailed internal motion has been emerging from intense efforts of many theoretical and experimental workers. A natural question is whether these kinds of issues will be important in the dynamics of larger molecules, and whether their investigation at the same level of detail will be profitable or tractable.

There will probably be some similarities, but also some fundamental differences. We have mainly considered small molecules with relatively rigid structures, in which the vibrational motions, although much different from the low-energy, near-harmonic normal modes, are nonetheless of relatively small amplitude and close to an equilibrium structure. (An important exception is the isomerization spectroscopy considered earlier, to which we shall return shortly.)

Molecules larger than those considered so far are formed by linking together several smaller components. A new kind of dynamics typical of these systems is already seen in a molecule such as C_2H_6 , in which there is hindered rotation of the two methyl groups. Systems with hindered internal rotation have been studied in great

depth [93], but there are still many unanswered questions. It seems likely that semiclassical techniques, using bifurcation analysis, could be brought to bear on these systems with great benefit.

The dynamics begin to take on a qualitatively different nature as the number of components, capable of mutual

-30-

hindered rotation, starts to become only a little larger than in C_2H_6 . The reason is that large-amplitude, very flexible twisting motions, such as those that start to be seen in a small polymer chain, become very important. These large scale ‘wiggly motions’ define a new class of dynamics and associated frequency scale as a characteristic internal motion of the system.

A hint that bifurcation techniques should be a powerful aid to the understanding of these problems comes from the example already considered in HCP isomerization [57]. Here the bifurcation techniques have given dramatic insights into the motions that stray very far from the equilibrium structure, in fact approaching the top of a barrier to the rearrangement to a different molecular isomer. It seems likely that similar approaches will be invaluable for molecules with internal rotors, including flexible polymer systems, but with an increase in complexity corresponding to the larger size of the systems. Probably, techniques to separate out the characteristic large-amplitude flexible motions from faster high-frequency vibrations, such as those of the individual bonds, will be necessary to unlock, along with the tools of the bifurcation analysis, the knowledge of the detailed anharmonic motions encrypted in the spectrum. This separation of time scales would be similar in some ways to the Born–Oppenheimer separability of nuclear and electronic motion.

Another class of problems in larger systems, also related to isomerization, is the question of large-amplitude motions in clusters of atoms and molecules. The phenomena of internal rearrangements, including processes akin to ‘melting’ and the seeking of minima on potential surfaces of very high dimensionality (due to the number of particles), have been extensively investigated [94]. The question of the usefulness of bifurcation techniques and the dynamical nature of large-amplitude natural motions in these systems has yet to be explored. These problems of large-amplitude motions and the seeking of potential minima in large clusters are conceptually related to the problem of protein folding, to which we now turn.

A 1.2.22 PROTEIN FOLDING

An example of a kind of extreme challenge in the complexity of internal molecular dynamics comes with very complicated biological macromolecules. One of the major classes of these is proteins, very long biopolymers consisting of large numbers of amino acid residues [95]. They are very important in biological systems because they are the output of the translation of the genetic code: the DNA codes for the sequences of amino acid residues for each individual protein produced by the organism. A good sequence, i.e. one which forms a biologically useful protein, is one which folds to a more-or-less unique ‘native’ three-dimensional *tertiary* structure. (The sequence itself is the *primary* structure; subunits within the tertiary structure, consisting of chains of residues, fold to well defined *secondary* structures, which themselves are folded into the tertiary structure.) An outstanding problem, still very far from a complete understanding, is the connection between the sequence and the specific native structure, and even the prior question whether a given sequence has a reliable native structure at all. For sequences which do fold up into a unique structure, it is not yet possible to reliably predict what the structure will be, or what it is about the sequence that makes it a good folder. A solution of the sequence–structure problem would be very important, because it would make it possible to design sequences in the laboratory to fold to a definite, predictable structure, which then could be tailored for biological activity. A related question is the kinetic mechanism by which a good protein folds to its native structure.

Both the structural and kinetic aspects of the protein-folding problem are complicated by the fact that folding takes place within a bath of water molecules. In fact, hydrophobic interactions are almost certainly crucial for both the relation of the sequence and the native structure, and the process by which a good sequence folds to its native structure.

It is presently unknown whether the kind of detailed dynamical analysis of the natural motions of molecules outlined in this section will be useful for a problem as complicated as that of protein folding. The likely applicability of such methods to systems with several internal rotors strung together, and the incipient interest in bifurcation phenomena of small molecules immersed in a bath [80], suggests that dynamical analysis might also be useful for the much larger structures in proteins. In a protein, most of the molecular motion may be essentially irrelevant, i.e. the high-frequency, small-amplitude vibrations of the backbone of the amino acid sequence, and, also, probably much of the localized large-amplitude ‘wiggly’ motion. It is likely that there is a far smaller number of relevant large-amplitude, low-frequency motions that are crucial to the folding process. It will be of great interest to discover if techniques of dynamical systems such as bifurcation analysis can be used to reveal the ‘folding modes’ of proteins. For this to work, account must be taken of the complication of the bath of water molecules in which the folding process takes place. This introduces effects such as friction, for which there is little or no experience at present in applying bifurcation techniques in molecular systems. Proteins themselves interact with other proteins and with nucleic acids in biological processes of every conceivable kind considered at the molecular level.

A 1.2.23 OUTLOOK

Knowledge of internal molecular motions became a serious quest with Boyle and Newton, at the very dawn of modern natural science. However, real progress only became possible with the advent of quantum theory in the 20th century. The study of internal molecular motion for most of the century was concerned primarily with molecules near their equilibrium configuration on the PES. This gave an enormous amount of immensely valuable information, especially on the structural properties of molecules.

In recent years, especially the past two decades, the focus has changed dramatically to the study of highly-excited states. This came about because of a conjunction of powerful influences, often in mutually productive interaction with molecular science. Perhaps the first was the advent of lasers as revolutionary light sources for the probing of molecules. Coherent light of unprecedented intensities and spectral purity became available for studies in the traditional frequency domain of spectroscopy. This allowed previously inaccessible states of molecules to be reached, with new levels of resolution and detail. Later, the development of ultrafast laser pulses opened up the window of the ultrafast time domain as a spectroscopic complement to the new richness in the frequency domain. At the same time, revolutionary information technology made it possible to apply highly-sophisticated analytical methods, including new pattern recognition techniques, to process the wealth of new experimental information. The computational revolution also made possible the accurate investigation of highly-excited regions of molecular potential surfaces by means of quantum chemistry calculations. Finally, new mathematical developments in the study of nonlinear classical dynamics came to be appreciated by molecular scientists, with applications such as the bifurcation approaches stressed in this section.

With these radical advances in experimental technology, computational ability to handle complex systems, and new theoretical ideas, the kind of information being sought about molecules has undergone an equally profound change. Formerly, spectroscopic investigation, even of vibrations and rotations, had focused primarily on structural information. Now there is a marked drive toward dynamical information, including problems of energy flow, and

internal molecular rearrangement. As emphasized in this section, a tremendous impetus to this was the recognition that other kinds of motion, such as local modes, could be just as important as the low-energy normal modes, in the understanding of the internal dynamics of highly-excited states. Ultrafast pulsed lasers have played a major role in these dynamical investigations. There is also a growing awareness of the immense potential for frequency domain spectroscopy to yield information on ultrafast processes in the time domain. This involves sophisticated measurements and data analysis of the very complex spectra of excited states; and equally sophisticated theoretical analysis to unlock the dynamical information encoded in the spectra. One of the primary tools is the bifurcation analysis of phenomenological Hamiltonians used directly to model experimental spectra. This gives information on the birth of new anharmonic motions in bifurcations of the low-energy normal modes. This kind of analysis is yielding information of startling detail about the internal molecular dynamics of high-energy molecules, including molecules undergoing isomerization. The ramifications are beginning to be explored for molecules in condensed phase. Here, ultrafast time-domain laser spectroscopy is usually necessary; but the requisite knowledge of internal molecular dynamics at the level of bifurcation analysis must be obtained from frequency-domain, gas phase experiments. Thus, a fruitful interplay is starting between gas and condensed phase experiments, and probes using sophisticated time- and frequency-domain techniques. Extension to much larger systems such as proteins is an exciting, largely unexplored future prospect. The interplay of research on internal molecular dynamics at the levels of small molecules, intermediate-size molecules, such as small polymer chains, and the hyper-complex scale of biological macromolecules is a frontier area of chemistry which surely will yield fascinating insights and discoveries for a long time to come.

REFERENCES

- [1] Brock W H 1992 *The Norton History of Chemistry* (New York: Norton)
- [2] Hall M B 1965 *Robert Boyle on Natural Philosophy* (Bloomington, IN: Indiana University Press)
- [3] Schrödinger E 1996 *Nature and the Greeks* (Cambridge: Cambridge University Press)
Schrödinger E 1996 *Science and Humanism* (Cambridge: Cambridge University Press)
- [4] Kuhn T S 1957 *The Copernican Revolution: Planetary Astronomy in the Development of Western Thought* (Cambridge, MA: Harvard University Press)
- [5] Ihde A J 1984 *The Development of Modern Chemistry* (New York: Dover)
- [6] Steinfeld J I, Francisco J S and Hase W L 1999 *Chemical Kinetics and Dynamics* (Upper Saddle River, NJ: Prentice-Hall)
- [7] Ball P 1994 *Designing the Molecular World: Chemistry at the Frontier* (Princeton, NJ: Princeton University Press)
- [8] Rosker M J, Dantus M and Zewail A H 1988 *Science* **241** 1200
- [9] Mokhtari A, Cong P, Herek J L and Zewail A H 1990 *Nature* **348** 225
- [10] Kellman M E 1994 *Phys. Rev. Lett.* **75** 2543
- [11] Blumel R and Reinhardt W P 1997 *Chaos in Atomic Physics* (Cambridge: Cambridge University Press)
- [12] Jaffe C, Farrelly D and Uzer T 2000 *Phys. Rev. A* **60** 3833

- [13] Berry R S, Rice S A and Ross J 1980 *Physical Chemistry* (New York: Wiley)
- [14] Herzberg G 1966 *Molecular Spectra and Molecular Structure III: Electronic Spectra and Electronic Structure of Polyatomic Molecules* (New York: Van Nostrand-Reinhold)
- [15] Nikitin E E 1999 *Ann. Rev. Phys. Chem.* **50** 1
- [16] Goldstein H 1980 *Classical Mechanics* (Reading, MA: Addison-Wesley)
- [17] Herzberg G 1950 *Molecular Spectra and Molecular Structure I: Spectra of Diatomic Molecules* (New York: Van Nostrand-Reinhold)
- [18] Herzberg G 1945 *Molecular Spectra and Molecular Structure II: Infrared and Raman Spectra of Polyatomic Molecules* (New York: Van Nostrand-Reinhold)
- [19] Papoušek D and Aliev M R 1982 *Molecular Vibrational–Rotational Spectra* (Amsterdam: Elsevier)
- [20] Dai H L, Field R W and Kinsey J L 1985 *J. Chem. Phys.* **82** 2161
- [21] Frederick J H and McClelland G M 1987 *J. Chem. Phys.* **84** 4347
- [22] Littlejohn R G, Mitchell K A, Reinsch M, Aquilanti V and Cavalli S 1998 *Phys. Rev. A* **58** 3718
- [23] Sarkar P, Poulin N and Carrington T Jr 1999 *J. Chem. Phys.* **110** 10, 269
- [24] Wilson E B Jr, Decius J C and Cross P C 1955 *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra* (New York: McGraw-Hill)
- [25] Bunker P R 1979 *Molecular Symmetry and Spectroscopy* (New York: Academic)
- [26] Harter W G 1993 *Principles of Symmetry, Dynamics, and Spectroscopy* (New York: Wiley)
- [27] Weinstein A 1973 Normal modes for nonlinear Hamiltonian systems *Inv. Math.* **20** 47
- [28] Moser J 1976 Periodic orbits near an equilibrium and a theorem by Alan Weinstein *Comm. Pure Appl. Math.* **29** 727
- [29] Tabor M 1989 *Chaos and Integrability in Nonlinear Dynamics: An Introduction* (New York: Wiley)
- [30] Lichtenberg A J and Leiberman M A 1983 *Regular and Stochastic Motion* (Berlin: Springer)
- [31] Darling B T and Dennison D M 1940 *Phys. Rev.* **57** 128
- [32] Messiah A 1961 *Quantum Mechanics* transl. G M Temmer (Amsterdam: North-Holland)
- [33] Heisenberg W 1925 *Z. Phys.* **33** 879 (Engl. Transl. van der Waerden B L (ed) 1967 *Sources of Quantum Mechanics* (New York: Dover)
- [34] Golubitsky M and Schaeffer D G 1985 *Singularities and Groups in Bifurcation Theory* vol 1 (New York: Springer)
- [35] Iachello F and Levine R D 1995 *Algebraic Theory of Molecules* (Oxford: Oxford University Press)
- [36] Tamsamani M A, Champion J-M and Oss S 1999 *J. Chem. Phys.* **110** 2893

- [37] Davis M J 1995 Trees from spectra: generation, analysis, and energy transfer information *Molecular Dynamics and Spectroscopy by Stimulated Emission Pumping* ed H-L Dai and R W Field (Singapore: World Scientific)

- [38] Lawton R T and Child M S 1980 *Mol. Phys.* **40** 773
- [39] Jaffe C and Brumer P 1980 *J. Chem. Phys.* **73** 5646
- [40] Sibert E L III, Hynes J T and Reinhardt W P 1982 *J. Chem. Phys.* **77** 3583
- [41] Davis M J and Heller E J 1981 *J. Chem. Phys.* **75** 246
- [42] Xiao L and Kellman M E 1989 *J. Chem. Phys.* **90** 6086
- [43] Li Z, Xiao L and Kellman M E 1990 *J. Chem. Phys.* **92** 2251
- [44] Xiao L and Kellman M E 1990 *J. Chem. Phys.* **93** 5805
- [45] Kellman M E 1995 Dynamical analysis of highly excited vibrational spectra: progress and prospects *Molecular Dynamics and Spectroscopy by Stimulated Emission Pumping* ed H-L Dai and R W Field (Singapore: World Scientific)
- [46] Clark A P, Dickinson A S and Richards D 1977 *Adv. Chem. Phys.* **36** 63
- [47] Heller E J 1995 *J. Phys. Chem.* **99** 2625
- [48] Kellman M E 1985 *J. Chem. Phys.* **83** 3843
- [49] Merzbacher E 1998 *Quantum Mechanics* 3rd edn (New York: Wiley)
- [50] Fermi E 1931 *Z. Physik* **71** 250
- [51] Pliva J 1972 *J. Mol. Spec.* **44** 165
- [52] Smith B C and Winn J S 1988 *J. Chem. Phys.* **89** 4638
- [53] Smith B C and Winn J S 1991 *J. Chem. Phys.* **94** 4120
- [54] Kellman M E and Xiao L 1990 *J. Chem. Phys.* **93** 5821
- [55] Svitak J, Li Z, Rose J and Kellman M E 1995 *J. Chem. Phys.* **102** 4340
- [56] Ishikawa H, Nagao C, Mikami N and Field R W 1998 *J. Chem. Phys.* **109** 492
- [57] Joyeux M, Sugny D, Tyng V, Kellman M E, Ishikawa H and Field R W 2000 *J. Chem. Phys.* **112** 4162
- [58] Lu Z-M and Kellman M E 1995 *Chem. Phys. Lett.* **247** 195–203
- [59] Keshavamurthy S and Ezra G S 1997 *J. Chem. Phys.* **107** 156
- [60] Lu Z-M and Kellman M E 1997 *J. Chem. Phys.* **107** 1–15
- [61] Jacobson M P, Jung C, Taylor H S and Field R W 1999 *J. Chem. Phys.* **111** 66
- [62] Fried L E and Ezra G S 1987 *J. Chem. Phys.* **86** 6270

- [63] Sibert E L 1988 *J. Chem. Phys.* **88** 4378
- [64] Herrick D R and O'Connor S 1998 *J. Chem. Phys.* **109** 2071
- [65] Sibert E L and McCoy A B 1996 *J. Chem. Phys.* **105** 469

- [66] Kellman M E 1990 *J. Chem. Phys.* **93** 6330
- [67] Kellman M E and Chen G 1991 *J. Chem. Phys.* **95** 8671
- [68] Solina S A B, O'Brien J P, Field R W and Polik W F 1996 *J. Phys. Chem.* **100** 7797
- [69] Abbouti Temsamani M and Herman M 1995 *J. Chem. Phys.* **102** 6371
- [70] El Idrissi M I, Lievin J, Campargue A and Herman M 1999 *J. Chem. Phys.* **110** 2074
- [71] Jonas D M, Solina S A B, Rajaram B, Silbey R J, Field R W, Yamanouchi K and Tsuchiya S 1993 *J. Chem. Phys.* **99** 7350
- [72] Jacobson M P, O'Brien J P, Silbey R J and Field R W 1998 *J. Chem. Phys.* **109** 121
- [73] Waller I M, Kitsopoulos T M and Neumark D M 1990 *J. Phys. Chem.* **94** 2240
- [74] Sadeghi R and Skodje R T 1996 *J. Chem. Phys.* **105** 7504
- [75] Choi Y S and Moore C B 1991 *J. Chem. Phys.* **94** 5414
- [76] Wu G 1998 *Chem. Phys. Lett.* **292** 369
- [77] Stockmann H-J 1999 *Quantum Chaos: An Introduction* (Cambridge: Cambridge University Press)
- [78] Keshavamurthy S and Ezra G S 1996 *Chem. Phys. Lett.* **259** 81
- [79] Everitt K F, Egorov S A and Skinner J L 1998 *Chem. Phys.* **235** 115
- [80] Hayes S C, Philpott M P, Mayer S G and Reid P J 1999 *J. Phys. Chem. A* **103** 5534
- [81] Rabitz H 1997 *Adv. Chem. Phys.* **101** 315
- [82] Hoffmann R 2000 *Am. Sci.* **88** 14
- [83] VanderWal R L, Scott J L and Crim F F 1990 *J. Chem. Phys.* **92** 803
- [84] Kandel S A and Zare R N 1998 *J. Chem. Phys.* **109** 9719
- [85] O'Raifeartaigh L 1997 *The Dawning of Gauge Theory* (Princeton, NJ: Princeton University Press)
- [86] Coughlan G D and Dodd J E 1991 *The Ideas of Particle Physics: An Introduction for Scientists* 2nd edn (Cambridge: Cambridge University Press)
- [87] Aitchison I J R and Hey A J G 1996 *Gauge Theories in Particle Physics: A Practical Introduction* (Bristol: Institute of Physics Publishing)
- [88] Aharonov Y and Bohm D 1959 *Phys. Rev.* **115** 485

- [89] Berry M V 1984 *Proc. R. Soc. A* **392** 45
- [90] Mead C A 1992 *Rev. Mod. Phys.* **64** 1
- [91] Littlejohn R G and Reinsch M 1997 *Rev. Mod. Phys.* **69** 213
- [92] Cina J A 2000 *J. Raman Spec.* **31** 95
- [93] Ortigoso J, Kleiner I and Hougen J T 1999 *J. Chem. Phys.* **110** 11 688

[94] Berry R S 1999 Phases and phase changes of small systems *Theory of Atomic and Molecular Clusters* ed J Jellinek (Berlin: Springer)

[95] Creighton T E 1993 *Proteins* (New York: Freeman)

FURTHER READING

Herzberg G 1950 *Molecular Spectra and Molecular Structure I: Spectra of Diatomic Molecules* (New York: Van Nostrand-Reinhold)

Herzberg G 1945 *Molecular Spectra and Molecular Structure II: Infrared and Raman Spectra of Polyatomic Molecules* (New York: Van Nostrand-Reinhold)

Herzberg G 1966 *Molecular Spectra and Molecular Structure III: Electronic Spectra and Electronic Structure of Polyatomic Molecules* (New York: Van Nostrand-Reinhold)

The above three sources are a classic and comprehensive treatment of rotation, vibration, and electronic spectra of diatomic and polyatomic molecules.

Kellman M E 1995 Dynamical analysis of highly excited vibrational spectra: progress and prospects *Molecular Dynamics and Spectroscopy by Stimulated Emission Pumping* ed H-L Dai and R W Field (Singapore: World Scientific)

This is a didactic introduction to some of the techniques of bifurcation theory discussed in this article.

Steinfeld J I, Francisco J S and Hase W L 1999 *Chemical Kinetics and Dynamics* (Upper Saddle River, NJ: Prentice-Hall)

Papoušek D and Aliev M R 1982 *Molecular Vibrational–Rotational Spectra* (Amsterdam: Elsevier)

This is a readable and fairly comprehensive treatment of rotation–vibration spectra and their interactions.

Tabor M 1989 *Chaos and Integrability in Nonlinear Dynamics: An Introduction* (New York: Wiley)

Lichtenberg A J and Leiberman M A 1983 *Regular and Stochastic Motion* (Berlin: Springer)

The above is a comprehensive, readable introduction to modern nonlinear classical dynamics, with quantum applications.

-37-

Iachello F and Levine R D 1995 *Algebraic Theory of Molecules* (Oxford: Oxford University Press)

This is a comprehensive survey of ‘algebraic’ methods for internal molecular motions.

Kellman M E 1995 Algebraic methods in spectroscopy *Ann. Rev. Phys. Chem.* **46** 395

This survey compares ‘algebraic’ methods with more standard approaches, and the bifurcation approach in this article.

Bunker P R 1979 *Molecular Symmetry and Spectroscopy* (New York: Academic)

Harter W G 1993 *Principles of Symmetry, Dynamics, and Spectroscopy* (New York: Wiley)

The above two references are comprehensive and individualistic surveys of symmetry, molecular structure and dynamics.

A1.3 Quantum mechanics of condensed phases

James R Chelikowsky

A1.3.1 INTRODUCTION

Traditionally one categorizes matter by phases such as gases, liquids and solids. Chemistry is usually concerned with matter in the gas and liquid phases, whereas physics is concerned with the solid phase. However, this distinction is not well defined: often chemists are concerned with the solid state and reactions between solid-state phases, and physicists often study atoms and molecular systems in the gas phase. The term *condensed phases* usually encompasses both the liquid state and the solid state, but not the gas state. In this section, the emphasis will be placed on the solid state with a brief discussion of liquids.

The solid phase of matter offers a very different environment to examine the chemical bond than does a gas or liquid [1, 2, 3, 4 and 5]. The obvious difference involves describing the atomic positions. In a solid state, one can often describe atomic positions by a *static* configuration, whereas for liquid and gas phases this is not possible. The properties of the liquids and gases can be characterized only by considering some time-averaged ensemble. This difference between phases offers advantages in describing the solid phase, especially for crystalline matter. Crystals are characterized by a periodic symmetry that results in a system occupying all space [6]. Periodic, or translational, symmetry of crystalline phases greatly simplifies discussions of the solid state since knowledge of the atomic structure within a fundamental ‘subunit’ of the crystal, called the *unit cell*, is sufficient to describe the entire system encompassing all space. For example, if one is interested in the spatial distribution of electrons in a crystal, it is sufficient to know what this distribution is within a unit cell.

A related advantage of studying crystalline matter is that one can have symmetry-related operations that greatly expedite the discussion of a chemical bond. For example, in an elemental crystal of diamond, all the chemical bonds are equivalent. There are no terminating bonds and the characterization of one bond is sufficient to understand the entire system. If one were to know the binding energy or polarizability associated with one bond, then properties of the diamond crystal associated with all the bonds could be extracted. In contrast, molecular systems often contain different bonds and always have atoms at the boundary between the molecule and the vacuum.

Since solids do not exist as truly infinite systems, there are issues related to their termination (i.e. surfaces). However, in most cases, the existence of a surface does not strongly affect the properties of the crystal as a whole. The number of atoms in the interior of a cluster scale as the cube of the size of the specimen while the number of surface atoms scale as the square of the size of the specimen. For a sample of macroscopic size, the number of interior atoms vastly exceeds the number of atoms at the surface. On the other hand, there are interesting properties of the surface of condensed matter systems that have no analogue in atomic or molecular systems. For example, electronic states can exist that ‘trap’ electrons at the interface between a solid and the vacuum [1].

Issues associated with order occupy a large area of study for crystalline matter [1, 7, 8]. For nearly perfect crystals, one can have systems with defects such as *point defects* and *extended defects* such as dislocations and grain

boundaries. These defects occur in the growth process or can be mechanically induced. In contrast to molecular systems that can be characterized by ‘perfect’ molecular systems, solids always have defects. Individual atoms that are missing from the ideal crystal structure, or extra atoms unneeded to characterize the ideal crystal are called point defects. The missing atoms correspond to vacancies; additional atoms are called

interstitials. Extended defects are entire planes of atoms or interfaces that do not correspond to those of the ideal crystal. For example, edge dislocations occur when an extra half-plane of atoms is inserted in a perfect crystal and grain boundaries occur when a solid possesses regions of crystalline matter that have different structural orientations. In general, if a solid has no long-range order then one considers the phase to be an *amorphous* solid. The idea of atomic order and ‘order parameters’ is not usually considered for molecular systems, although for certain systems such as long molecular chains of atoms one might invoke a similar concept.

Another issue that distinguishes solids from atomic or molecular systems is the role of controlled defects or impurities. Often a pure, elemental crystal is not of great interest for technological applications; however, crystals with controlled additions of impurities are of great interest. The alteration of electronic properties with defects can be dramatic, involving changes in electrical conductivity by orders of magnitude. As an example, the addition of one boron atom for every 10^5 silicon atoms increases the conductivity of pure silicon by factor of 10^3 at room temperature [1]. Much of the electronic materials revolution is based on capitalizing on the dramatic changes in electronic properties via the controlled addition of electronically active dopants.

Of course, condensed phases also exhibit interesting physical properties such as electronic, magnetic, and mechanical phenomena that are not observed in the gas or liquid phase. Conductivity issues are generally not studied in isolated molecular species, but are actively examined in solids. Recent work in solids has focused on dramatic conductivity changes in superconducting solids. Superconducting solids have resistivities that are identically zero below some transition temperature [1, 9, 10]. These systems cannot be characterized by interactions over a few atomic species. Rather, the phenomenon involves a collective mode characterized by a phase representative of the entire solid.

A1.3.2 MANY-BODY WAVEFUNCTIONS IN CONDENSED PHASES

One of the most significant achievements of the twentieth century is the description of the quantum mechanical laws that govern the properties of matter. It is relatively easy to write down the Hamiltonian for interacting fermions. Obtaining a solution to the problem that is sufficient to make predictions is another matter.

Let us consider N nucleons of charge Z_n at positions $\{\mathbf{R}_n\}$ for $n = 1, \dots, N$ and M electrons at positions $\{\mathbf{r}_i\}$ for $i = 1, \dots, M$. This is shown schematically in figure A1.3.1. The Hamiltonian for this system in its simplest form can be written as

$$\begin{aligned} \hat{\mathcal{H}}(\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3, \dots; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3 \dots) = & \sum_{n=1}^N \frac{-\hbar^2 \nabla_n^2}{2\mathcal{M}_n} + \frac{1}{2} \sum_{n,m=1, n \neq m}^N \frac{Z_n Z_m e^2}{|\mathbf{R}_n - \mathbf{R}_m|} + \sum_{i=1}^M \frac{-\hbar^2 \nabla_i^2}{2m} \\ & - \sum_{n=1}^N \sum_{i=1}^M \frac{Z_n e^2}{|\mathbf{R}_n - \mathbf{r}_i|} + \frac{1}{2} \sum_{i,j=1, i \neq j}^M \frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|}. \end{aligned} \quad (\text{A1.3.1})$$

\mathcal{M}_n is the mass of the nucleon, \hbar is Planck’s constant divided by 2π , m is the mass of the electron. This expression omits some terms such as those involving relativistic interactions, but captures the essential features for most condensed matter phases.

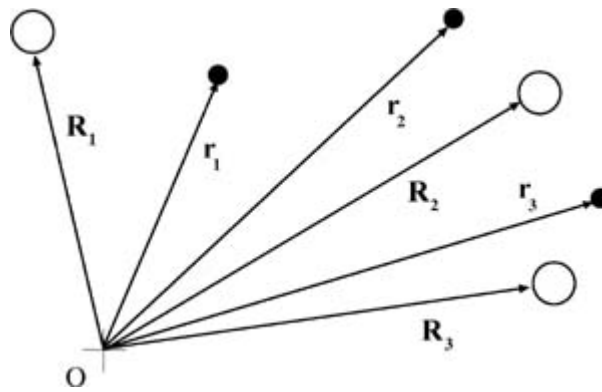


Figure A1.3.1. Atomic and electronic coordinates. The electrons are illustrated by filled circles; the nuclei by open circles.

Using the Hamiltonian in [equation A1.3.1](#), the quantum mechanical equation known as the Schrödinger equation for the electronic structure of the system can be written as

$$\hat{H}(\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3, \dots; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \dots)\Psi(\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3, \dots; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \dots) = E\Psi(\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3, \dots; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \dots) \quad (\text{A1.3.2})$$

where E is the total electronic energy of the system, and Ψ is the many-body wavefunction. In the early part of the twentieth century, it was recognized that this equation provided the means of solving for the electronic and nuclear degrees of freedom. Using the variational principle, which states that an approximate wavefunction will always have a less favourable energy than the true ground-state energy, one had an equation and a method to test the solution. One can estimate the energy from

$$E = \frac{\int \Psi^* \hat{H} \Psi \, d^3 R_1 \, d^3 R_2 \, d^3 R_3 \dots \, d^3 r_1 \, d^3 r_2 \, d^3 r_3 \dots}{\int \Psi^* \Psi \, d^3 R_1 \, d^3 R_2 \, d^3 R_3 \dots \, d^3 r_1 \, d^3 r_2 \, d^3 r_3 \dots} \quad (\text{A1.3.3})$$

Solving equation A1.3.2 for anything more complex than a few particles becomes problematic even with the most modern computers. Obtaining an approximate solution for condensed matter systems is difficult, but considerable progress has been made since the advent of digital computers. Several highly successful approximations have been made to solve for the ground-state energy. The nature of the approximations used is to remove as many degrees of freedom from the system as possible.

One common approximation is to separate the nuclear and electronic degrees of freedom. Since the nuclei are considerably more massive than the electrons, it can be assumed that the electrons will respond ‘instantaneously’ to the nuclear coordinates. This approximation is called the Born–Oppenheimer or adiabatic approximation. It allows one to treat the nuclear coordinates as classical parameters. For most condensed matter systems, this assumption is highly accurate [[11](#), [12](#)].

A1.3.2.1 THE HARTREE APPROXIMATION

Another common approximation is to construct a specific form for the many-body wavefunction. If one can obtain an accurate estimate for the wavefunction, then, via the variational principle, a more accurate estimate for the energy will emerge. The most difficult part of this exercise is to use physical intuition to define a trial wavefunction.

One can utilize some very simple cases to illustrate this approach. Suppose one considers a solution for *non-interacting electrons*: i.e. in [equation A1.3.1](#) the last term in the Hamiltonian is ignored. In this limit, it is

possible to write the many-body wavefunction as a sum of independent Hamiltonians. Using the adiabatic approximation, the *electronic* part of the Hamiltonian becomes

$$\hat{\mathcal{H}}_{\text{el}}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3 \dots) = \sum_{i=1}^M \frac{-\hbar^2 \nabla_i^2}{2m} - \sum_{n=1}^N \sum_{i=1}^M \frac{Z_n e^2}{|\mathbf{R}_n - \mathbf{r}_i|}. \quad (\text{A1.3.4})$$

Let us define a nuclear potential, V_N , which the i th electron sees as

$$V_N(\mathbf{r}_i) = - \sum_{n=1}^N \frac{Z_n e^2}{|\mathbf{R}_n - \mathbf{r}_i|}. \quad (\text{A1.3.5})$$

One can now rewrite a simplified Schrödinger equation as

$$\hat{\mathcal{H}}_{\text{el}}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3 \dots) \psi(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3 \dots) = \sum_{i=1}^M \hat{H}^i \psi(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3 \dots) \quad (\text{A1.3.6})$$

where the Hamiltonian is now defined for the i th electron as

$$\hat{H}^i = \frac{-\hbar^2 \nabla_i^2}{2m} + V_N(\mathbf{r}_i). \quad (\text{A1.3.7})$$

For this simple Hamiltonian, let us write the many-body wavefunction as

$$\psi(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3 \dots) = \phi_1(\mathbf{r}_1) \phi_2(\mathbf{r}_2) \phi_3(\mathbf{r}_3) \dots \quad (\text{A1.3.8})$$

The $\phi_i(\mathbf{r})$ orbitals can be determined from a ‘one-electron’ Hamiltonian

$$\hat{H}^i \phi_i(\mathbf{r}) = \left(\frac{-\hbar^2 \nabla^2}{2m} + V_N(\mathbf{r}) \right) \phi(\mathbf{r}) = E_i \phi_i(\mathbf{r}). \quad (\text{A1.3.9})$$

The index i for the orbital $\phi_i(\mathbf{r})$ can be taken to include the spin of the electron plus any other relevant quantum numbers. The index i runs over the number of electrons, each electron being assigned a unique set of quantum

numbers. This type of Schrödinger equation can be easily solved for fairly complex condensed matter systems. The many-body wavefunction in [equation A1.3.8](#) is known as the Hartree wavefunction. If one uses this form of the wavefunction as an approximation to solve the Hamiltonian *including* the electron–electron interactions, this is known as the Hartree approximation. By ignoring the electron–electron terms, the Hartree approximation simply reflects the electrons independently moving in the nuclear potential. The total energy of the system in this case is simply the sum of the eigenvalues, E_i .

To obtain a realistic Hamiltonian, the electron–electron interactions must be reinstated in [equation A1.3.6](#):

$$\hat{\mathcal{H}}_{\text{el}}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3 \dots) \psi(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3 \dots) = \sum_{i=1}^M \left(\hat{H}^i + \frac{1}{2} \sum_{j=1, j \neq i}^M \frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|} \right) \psi(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3 \dots). \quad (\text{A1.3.10})$$

In this case, the individual orbitals, $\phi_i(\mathbf{r})$, can be determined by minimizing the total energy as per [equation A1.3.3](#), with the constraint that the wavefunction be normalized. This minimization procedure results in the following Hartree equation:

$$\hat{H}^i \phi_i(\mathbf{r}) = \left(\frac{-\hbar^2 \nabla^2}{2m} + V_N(\mathbf{r}) + \sum_{j=1, j \neq i}^M \int \frac{e^2 |\phi_j(\mathbf{r}')|^2}{|\mathbf{r} - \mathbf{r}'|} d^3 r' \right) \phi_i(\mathbf{r}) = E_i \phi_i(\mathbf{r}). \quad (\text{A1.3.11})$$

Using the orbitals, $\phi(\mathbf{r})$, from a solution of equation A1.3.11, the Hartree many-body wavefunction can be constructed and the total energy determined from [equation A1.3.3](#).

The Hartree approximation is useful as an illustrative tool, but it is not a very accurate approximation. A significant deficiency of the Hartree wavefunction is that it does not reflect the anti-symmetric nature of the electrons as required by the Pauli principle [7]. Moreover, the Hartree equation is difficult to solve. The Hamiltonian is orbitally dependent because the summation in equation A1.3.11 does not include the i th orbital. This means that if there are M electrons, then M Hamiltonians must be considered and equation A1.3.11 solved for each orbital.

A1.3.2.2 THE HARTREE–FOCK APPROXIMATION

It is possible to write down a many-body wavefunction that will reflect the antisymmetric nature of the wavefunction. In this discussion, the spin coordinate of each electron needs to be explicitly treated. The coordinates of an electron may be specified by $\mathbf{r}_i s_i$, where s_i represents the spin coordinate. Starting with one-electron orbitals, $\phi_i(\mathbf{r} s)$, the following form can be invoked:

$$\Psi(\mathbf{r}_1 s_1, \mathbf{r}_1 s_2, \mathbf{r}_1 s_3, \dots) = \begin{vmatrix} \phi_1(\mathbf{r}_1 s_1) & \phi_1(\mathbf{r}_2 s_2) & \dots & \dots & \phi_1(\mathbf{r}_M s_M) \\ \phi_2(\mathbf{r}_1 s_1) & \phi_2(\mathbf{r}_2 s_2) & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \phi_M(\mathbf{r}_1 s_1) & \dots & \dots & \dots & \phi_M(\mathbf{r}_M s_M) \end{vmatrix}. \quad (\text{A1.3.12})$$

This form of the wavefunction is called a Slater determinant. It reflects the proper symmetry of the wavefunction and

the Pauli principle. If two electrons occupy the same orbit, two rows of the determinant will be identical and the many-body wavefunction will have zero amplitude. Likewise, the determinant will vanish if two electrons occupy the same point in generalized space (i.e. $\mathbf{r}_i s_i = \mathbf{r}_j s_j$) as two columns of the determinant will be identical. If two particles are exchanged, this corresponds to a sign change in the determinant. The Slater determinant is a convenient representation. It is probably the simplest form that incorporates the required symmetry properties for fermions, or particles with non-integer spins.

If one uses a Slater determinant to evaluate the total electronic energy and maintains the orbital normalization, then the orbitals can be obtained from the following Hartree–Fock equations:

$$\begin{aligned}
H^i \phi_i(\mathbf{r}) &= \left(\frac{-\hbar^2 \nabla^2}{2m} + V_N(\mathbf{r}) + \sum_{j=1}^M \int \frac{e^2 |\phi_j(\mathbf{r}')|^2}{|\mathbf{r} - \mathbf{r}'|} d^3 r' \right) \phi_i(\mathbf{r}) \\
&= E_i \phi_i(\mathbf{r}) - \sum_{j=1}^M \int \frac{e^2}{|\mathbf{r} - \mathbf{r}'|} \phi_j^*(\mathbf{r}') \phi_i(\mathbf{r}') d^3 r' \delta_{s_i, s_j} \phi_j(\mathbf{r})
\end{aligned}$$

It is customary to simplify this expression by defining an electronic charge density, ρ :

$$\rho(\mathbf{r}) = \sum_{j=1}^M |\phi_j(\mathbf{r})|^2 \quad (\text{A1.3.14})$$

and an orbitally dependent *exchange-charge density*, ρ_i^{HF} for the i th orbital:

$$\rho_i^{\text{HF}}(\mathbf{r}, \mathbf{r}') = \sum_{j=1}^M \frac{\phi_j^*(\mathbf{r}') \phi_i(\mathbf{r}') \phi_i^*(\mathbf{r}) \phi_j(\mathbf{r})}{\phi_i^*(\mathbf{r}) \phi_i(\mathbf{r})} \delta_{s_i, s_j}. \quad (\text{A1.3.15})$$

This ‘density’ involves a spin-dependent factor which couples only states (i, j) with the same spin coordinates (s_i, s_j) . It is not a true density in that it is dependent on \mathbf{r}, \mathbf{r}' ; it has meaning only as defined below.

With these charge densities defined, it is possible to define corresponding potentials. The Coulomb or Hartree potential, V_H , is defined by

$$V_H(\mathbf{r}) = \int \rho(\mathbf{r}') \frac{e^2}{|\mathbf{r} - \mathbf{r}'|} d^3 r' \quad (\text{A1.3.16})$$

and an *exchange potential* can be defined by

$$V_x^i(\mathbf{r}) = - \int \rho_i^{\text{HF}}(\mathbf{r}, \mathbf{r}') \frac{e^2}{|\mathbf{r} - \mathbf{r}'|} d^3 r'. \quad (\text{A1.3.17})$$

This combination results in the following Hartree–Fock equation:

$$\left(\frac{-\hbar^2 \nabla^2}{2m} + V_N(\mathbf{r}) + V_H(\mathbf{r}) + V_x^i(\mathbf{r}) \right) \phi_i(\mathbf{r}) = E_i \phi_i(\mathbf{r}). \quad (\text{A1.3.18})$$

Once the Hartree–Fock orbitals have been obtained, the total Hartree–Fock electronic energy of the system, E_{HF} , can be obtained from

$$E_{\text{HF}} = \sum_i^M E_i - \frac{1}{2} \int \rho(\mathbf{r}) V_H(\mathbf{r}) d^3 r - \frac{1}{2} \sum_i^M \int \phi_i^*(\mathbf{r}) \phi_i(\mathbf{r}) V_x^i(\mathbf{r}) d^3 r. \quad (\text{A1.3.19})$$

E_{HF} is not a sum of the Hartree–Fock orbital energies, E_i . The factor of $\frac{1}{2}$ in the electron–electron terms arises because the electron–electron interactions have been double-counted in the Coulomb and exchange potentials. The Hartree–Fock Schrödinger equation is only slightly more complex than the Hartree equation. Again, the equations are difficult to solve because the exchange potential is orbitally dependent.

There is one notable difference between the Hartree–Fock summation and the Hartree summation. The Hartree–Fock sums include the $i = j$ terms in [equation A1.3.13](#). This difference arises because the exchange term corresponding to $i = j$ cancels an equivalent term in the Coulomb summation. The $i = j$ term in both the Coulomb and exchange term is interpreted as a ‘self-screening’ of the electron. Without a cancellation between Coulomb and exchange terms a ‘self-energy’ contribution to the total energy would occur. Approximate forms of the exchange potential often do not have this property. The total energy then contains a self-energy contribution which one needs to remove to obtain a correct Hartree–Fock energy.

The Hartree–Fock wavefunctions are approximations to the true ground-state many-body wavefunctions. Terms not included in the Hartree–Fock energy are referred to as *correlation* contributions. One definition for the correlation energy, E_{corr} is to write it as the difference between the correct total energy of the system and the Hartree–Fock energies: $E_{\text{corr}} = E_{\text{exact}} - E_{\text{HF}}$. Correlation energies are sometimes included by considering Slater determinants composed of orbitals which represent excited-state contributions. This method of including unoccupied orbitals in the many-body wavefunction is referred to as *configuration interaction* or ‘CI’.

Applying Hartree–Fock wavefunctions to condensed matter systems is not routine. The resulting Hartree–Fock equations are usually too complex to be solved for extended systems. It has been argued that many-body wavefunction approaches to the condensed matter or large molecular systems do not represent a reasonable approach to the electronic structure problem of extended systems.

A1.3.3 DENSITY FUNCTIONAL APPROACHES TO QUANTUM DESCRIPTIONS OF CONDENSED PHASES

Alternative descriptions of quantum states based on a knowledge of the electronic charge density [equation A1.3.14](#) have existed since the 1920s. For example, the Thomas–Fermi description of atoms based on a knowledge of $\rho(\mathbf{r})$

was reasonably successful [[13](#), [14](#) and [15](#)]. The starting point for most discussions of condensed matter begins by considering a limiting case that may be appropriate for condensed matter systems, but not for small molecules. One often considers a free electron gas of uniform charge density. The justification for this approach comes from the observation that simple metals like aluminium and sodium have properties which appear to resemble those of a free electron gas. This model cannot be applied to systems with localized electrons such as highly covalent materials like carbon or highly ionic materials like sodium chloride. It is also not appropriate for very open structures. In these systems large variations of the electron distribution can occur.

A1.3.3.1 FREE ELECTRON GAS

Perhaps the simplest description of a condensed matter system is to imagine non-interacting electrons contained within a box of volume, Ω . The Schrödinger equation for this system is similar to [equation A1.3.9](#) with the potential set to zero:

$$\frac{-\hbar^2 \nabla^2}{2m} \phi(\mathbf{r}) = E \phi(\mathbf{r}). \quad (\text{A1.3.20})$$

Ignoring spin for the moment, the solution of equation A1.3.20 is

$$\phi(\mathbf{r}) = \frac{1}{\sqrt{\Omega}} \exp(i\mathbf{k} \cdot \mathbf{r}). \quad (\text{A1.3.21})$$

The energy is given by $E(k) = \hbar^2 k^2 / 2m$ and the charge density by $\rho = 1/\Omega$. \mathbf{k} is called a *wavevector*.

A key issue in describing condensed matter systems is to account properly for the number of states. Unlike a molecular system, the eigenvalues of condensed matter systems are closely spaced and essentially ‘infinite’ in number. For example, if one has 10^{23} electrons, then one can expect to have 10^{23} occupied states. In condensed matter systems, the number of states per energy unit is a more natural measure to describe the energy distribution of states.

It is easy to do this with *periodic boundary conditions*. Suppose one considers a one-dimensional specimen of length L . In this case the wavefunctions obey the rule $\phi(x + L) = \phi(x)$ as $x + L$ corresponds in all physical properties to x . For a free electron wavefunction, this requirement can be expressed as $\exp(ik(x + L)) = \exp(ikx)$ or as $\exp(ikL) = 1$ or $k = 2\pi n/L$ where n is an integer.

Periodic boundary conditions force k to be a discrete variable with allowed values occurring at intervals of $2\pi/L$. For very large systems, one can describe the system as continuous in the limit of $L \rightarrow \infty$. Electron states can be defined by a *density of states* defined as follows:

$$\begin{aligned} D(E) &= \lim_{\Delta E \rightarrow 0} \frac{N(E + \Delta E) - N(E)}{\Delta E} \\ &= \frac{dN}{dE} \end{aligned} \quad (\text{A1.3.22})$$

-9-

where $N(E)$ is the number of states whose energy resides below E . For the one-dimensional case, $N(k) = 2k / (2\pi/L)$ (the factor of two coming from spin) and $dN/dE = (dN/dk) \cdot (dk/dE)$. Using $E(k) = \hbar^2 k^2 / 2m$, we have $k = \sqrt{2mE}/\hbar$ and $dk/dE = \frac{1}{2} \sqrt{2m/E}/\hbar$. This results in the one-dimensional density of states as

$$D(E) = \frac{L}{\pi\hbar} \sqrt{2m/E}. \quad (\text{A1.3.23})$$

The density of states for a one-dimensional system diverges as $E \rightarrow 0$. This divergence of $D(E)$ is not a serious issue as the integral of the density of states remains finite. In three dimensions, it is straightforward to show that

$$D(E) = \frac{\Omega}{2\pi^2} \left(\frac{2m}{\hbar^2} \right)^{3/2} \sqrt{E}. \quad (\text{A1.3.24})$$

The singularity is removed, although a discontinuity in the derivative exists as $E \rightarrow 0$.

One can determine the total number of electrons in the system by integrating the density of states up to the highest occupied energy level. The energy of the highest occupied state is called the *Fermi level* or *Fermi energy*, E_F :

$$N = \frac{\Omega}{2\pi^2} \left(\frac{2m}{\hbar^2} \right)^{3/2} \int_0^{E_F} \sqrt{E} dE \quad (\text{A1.3.25})$$

and

$$E_F = \frac{\hbar^2}{2m} \left(\frac{3\pi^2 N}{\Omega} \right)^{2/3}. \quad (\text{A1.3.26})$$

By defining a *Fermi* wavevector as $k_F = (3\pi^2 n_{\text{el}})^{1/3}$ where n_{el} is the electron density, $n_{\text{el}} = N/\Omega$, of the system, one can write

$$E_F = \frac{\hbar^2 k_F^2}{2m}. \quad (\text{A1.3.27})$$

It should be noted that typical values for E_F for simple metals like sodium or potassium are of the order of several electronvolts. If one defines a temperature, T_F , where $T_F = E_F/k_B$ and k_B is the Boltzmann constant, typical values for T_F might be 10^3 – 10^4 K. Thus, at ambient temperatures one can often neglect the role of temperature in determining the Fermi energy.

A1.3.3.2 HARTREE–FOCK EXCHANGE IN A FREE ELECTRON GAS

For a free electron gas, it is possible to evaluate the Hartree–Fock exchange energy directly [3, 16]. The Slater determinant is constructed using free electron orbitals. Each orbital is labelled by a \mathbf{k} and a spin index. The Coulomb

-10-

potential for an infinite free electron gas diverges, but this divergence can be removed by imposing a compensating uniform positive charge. The resulting Hartree–Fock eigenvalues can be written as

$$E_k = \frac{\hbar^2 k^2}{2m} - \frac{1}{\Omega} \sum_{k' < k_F} \frac{4\pi e^2}{|\mathbf{k} - \mathbf{k}'|^2} \quad (\text{A1.3.28})$$

where the summation is over occupied \mathbf{k} -states. It is possible to evaluate the summation by transposing the summation to an integration. This transposition is often done for solid-state systems as the state density is so high that the system can be treated as a continuum:

$$\frac{1}{\Omega} \sum_{k' < k_F} \frac{4\pi e^2}{|\mathbf{k} - \mathbf{k}'|^2} = \frac{1}{(2\pi)^3} \int_{k' < k_F} \frac{4\pi e^2}{|\mathbf{k} - \mathbf{k}'|^2} d^3k. \quad (\text{A1.3.29})$$

This integral can be solved analytically. The resulting eigenvalues are given by

$$E_k = \frac{\hbar^2 k^2}{2m} - \frac{e^2 k_F}{\pi} \left(1 + \frac{1 - (k/k_F)^2}{2(k/k_F)} \ln \left| \frac{k + k_F}{k - k_F} \right| \right). \quad (\text{A1.3.30})$$

Using the above expression and [equation A1.3.19](#), the total electron energy, $E_{\text{HF}}^{\text{FEG}}$, for a free electron gas within the Hartree–Fock approximation is given by

$$E_{\text{HF}}^{\text{FEG}} = 2 \sum_{k < k_{\text{F}}} \frac{\hbar^2 k^2}{2m} - \frac{e^2 k_{\text{F}}}{\pi} \sum_{k < k_{\text{F}}} \left(1 + \frac{1 - (k/k_{\text{F}})^2}{2(k/k_{\text{F}})} \ln \left| \frac{k + k_{\text{F}}}{k - k_{\text{F}}} \right| \right). \quad (\text{A1.3.31})$$

The factor of 2 in the first term comes from spin. In the exchange term, there is no extra factor of 2 because one can subtract off a ‘double-counting term’ (see [equation A1.3.19](#)). The summations can be executed as per [equation A1.3.29](#) to yield

$$E_{\text{HF}}^{\text{FEG}}/N = \frac{3}{5} E_{\text{F}} - \frac{3e^2}{4\pi} k_{\text{F}}. \quad (\text{A1.3.32})$$

The first term corresponds to the average energy per electron in a free electron gas. The second term corresponds to the exchange energy per electron. The exchange energy is attractive and scales with the cube root of the average density. This form provides a clue as to what form the exchange energy might take in an interacting electron gas or non-uniform electron gas.

Slater was one of the first to propose that one replace V_{x}^i in [equation A1.3.18](#) by a term that depends only on the cube root of the charge density [[17](#), [18](#) and [19](#)]. In analogy to [equation A1.3.32](#), he suggested that V_{x}^i be replaced by

-11-

$$V_{\text{x}}^{\text{Slater}}[\rho(\mathbf{r})] = -\frac{3e^2}{2\pi} (3\pi\rho(\mathbf{r}))^{1/3}. \quad (\text{A1.3.33})$$

This expression is not orbitally dependent. As such, a solution of the Hartree–Fock equation ([equation A1.3.18](#)) is much easier to implement. Although Slater exchange was not rigorously justified for non-uniform electron gases, it was quite successful in replicating the essential features of atomic and molecular systems as determined by Hartree–Fock calculations.

A1.3.3.3 THE LOCAL DENSITY APPROXIMATION

In a number of classic papers Hohenberg, Kohn and Sham established a theoretical framework for justifying the replacement of the many-body wavefunction by one-electron orbitals [[15](#), [20](#), [21](#)]. In particular, they proposed that the charge density plays a central role in describing the electronic structure of matter. A key aspect of their work was the *local density approximation* (LDA). Within this approximation, one can express the exchange energy as

$$E_{\text{x}}[\rho(\mathbf{r})] = \int \rho(\mathbf{r}) \varepsilon_{\text{x}}[\rho(\mathbf{r})] d^3r \quad (\text{A1.3.34})$$

where $\varepsilon_{\text{x}}[\rho]$ is the exchange energy per particle of uniform gas at a density of ρ . Within this framework, the exchange potential in [equation A1.3.18](#) is replaced by a potential determined from the functional derivative of $E_{\text{x}}[\rho]$:

$$V_{\text{x}}[\rho] = \frac{\delta E_{\text{x}}[\rho]}{\delta \rho}. \quad (\text{A1.3.35})$$

One serious issue is the determination of the exchange energy per particle, ε_{x} , or the corresponding exchange potential, V_{x} . The exact expression for either of these quantities is unknown, save for special cases. If one

assumes the exchange energy is given by [equation A1.3.32](#), i.e. the Hartree–Fock expression for the exchange energy of the free electron gas, then one can write

$$E_x[\rho] = -\frac{3e^2}{4\pi}(3\pi^2)^{1/3} \int [\rho(\mathbf{r})]^{4/3} d^3r \quad (\text{A1.3.36})$$

and taking the functional derivative, one obtains

$$V_x[\rho] = -\frac{e^2}{\pi}(3\pi^2\rho(\mathbf{r}))^{1/3}. \quad (\text{A1.3.37})$$

Comparing this to the form chosen by Slater, we note that this form, known as Kohn–Sham exchange, differs by a factor of $\frac{2}{3}$: i.e. $V_x = 2V_x^{\text{Slater}}/3$. For a number of years, some controversy existed as to whether the Kohn–Sham or Slater exchange was more accurate for realistic systems [15]. Slater suggested that a parameter be introduced that would allow one to vary the exchange between the Slater and Kohn–Sham values [19]. The parameter, α , was often

-12-

placed in front of the Slater exchange: $V_{x\alpha} = \alpha V_x^{\text{Slater}}$. α was often chosen to replicate some known feature of an exact Hartree–Fock calculation such as the total energy of an atom or ion. Acceptable values of α were viewed to range from $\alpha = \frac{2}{3}$ to $\alpha = 1$. Slater’s so-called ‘ X_α ’ method was very successful in describing molecular systems [19]. Notable drawbacks of the X_α method centre on its ad hoc nature through the α parameter and the omission of an explicit treatment of correlation energies.

In contemporary theories, α is taken to be $\frac{2}{3}$, and correlation energies are explicitly included in the energy functionals [15]. Sophisticated numerical studies have been performed on uniform electron gases resulting in local density expressions of the form $V_{xc}[\rho(\mathbf{r})] = V_x[\rho(\mathbf{r})] + V_c[\rho(\mathbf{r})]$ where V_c represents contributions to the total energy beyond the Hartree–Fock limit [22]. It is also possible to describe the role of spin explicitly by considering the charge density for up and down spins: $\rho = \rho_\uparrow + \rho_\downarrow$. This approximation is called the *local spin density approximation* [15].

The Kohn–Sham equation [21] for the electronic structure of matter is given by

$$\left(\frac{-\hbar^2 \nabla^2}{2m} + V_N(\mathbf{r}) + V_H(\mathbf{r}) + V_{xc}[\rho(\mathbf{r})] \right) \phi_i(\mathbf{r}) = E_i \phi_i(\mathbf{r}). \quad (\text{A1.3.38})$$

This equation is usually solved ‘self-consistently’. An approximate charge is assumed to estimate the exchange-correlation potential and to determine the Hartree potential from [equation A1.3.16](#). These approximate potentials are inserted in the Kohn–Sham equation and the total charge density is obtained from [equation A1.3.14](#). The ‘output’ charge density is used to construct new exchange-correlation and Hartree potentials. The process is repeated until the input and output charge densities or potentials are identical to within some prescribed tolerance.

Once a solution of the Kohn–Sham equation is obtained, the total energy can be computed from

$$(\text{A1.3.39})$$

$$E_{\text{KS}} = \sum_i^M E_i - \frac{1}{2} \int \rho(\mathbf{r}) V_{\text{H}}(\mathbf{r}) d^3r + \int \rho(\mathbf{r}) (\mathcal{E}_{\text{xc}}[\rho(\mathbf{r})] - V_{\text{xc}}[\rho(\mathbf{r})]) d^3r.$$

The electronic energy, as determined from E_{KS} , must be added to the ion–ion interactions to obtain the structural energies. This is a straightforward calculation for confined systems. For extended systems such as crystals, the calculations can be done using Madelung summation techniques [2].

Owing to its ease of implementation and overall accuracy, the local density approximation is the current method of choice for describing the electronic structure of condensed matter. It is relatively easy to implement and surprisingly accurate. Moreover, recent developments have included so-called gradient corrections to the local density approximation. In this approach, the exchange–correlation energy depends on the local density the gradient of the density. This approach is called the generalized gradient approximation or GGA [23].

When first proposed, density functional theory was not widely accepted in the chemistry community. The theory is not ‘rigorous’ in the sense that it is not clear how to improve the estimates for the ground-state energies. For wavefunction-based methods, one can include more Slater determinants as in a configuration interaction approach. As the wavefunctions improve via the variational theorem, the energy is lowered. In density functional theory, there is no

-13-

analogous procedure. The Kohn–Sham equations are also variational, but need not approach the true ground-state energy. This is not a problem provided that one is interested in *relative* energies and any inherent density functional errors cancel in the difference.

In some sense, density functional theory is an *a posteriori* theory. Given the transference of the exchange–correlation energies from an electron gas, it is not surprising that errors would arise in its implementation to highly non-uniform electron gas systems as found in realistic systems. However, the degree of error cancellations is rarely known *a priori*. The reliability of density functional theory has only been established by numerous calculations for a wide variety of condensed matter systems. For example, the cohesive energies, compressibility, structural parameters and vibrational spectra of elemental solids have been calculated within the density functional theory [24]. The accuracy of the method is best for systems in which the cancellation of errors is expected to be complete. Since cohesive energies involve the difference in energies between atoms in solids and atoms in free space, error cancellations are expected to be significant. This is reflected in the fact that historically cohesive energies have presented greater challenges for density functional theory: the errors between theory and experiment are typically ~5–10%, depending on the nature of the density functional. In contrast, vibrational frequencies which involve small structural changes within a given crystalline environment are easily reproduced to within 1–2%.

A1.3.4 ELECTRONIC STATES IN PERIODIC POTENTIALS: BLOCH’S THEOREM

Crystalline matter serves as the testing ground for electronic structure methods applied to extended systems. Owing to the translational periodicity of the system, a knowledge of the charge density in part of the crystal is sufficient to understand the charge density throughout the crystal. This greatly simplifies quantum descriptions of condensed matter.

A1.3.4.1 THE STRUCTURE OF CRYSTALLINE MATTER

A key aspect in defining a crystal is the existence of a building block which, when translated by a precise

prescription an infinite number of times, replicates the structure of interest. This building block is called a *unit cell*. The numbers of atoms required to define a unit cell can vary greatly from one solid to another. For simple metals such as sodium only one atom may be needed in defining the unit cell. Complex organic crystals can require thousands of atoms to define the building block.

The unit cell can be defined in terms of three *lattice vectors*: (\mathbf{a} , \mathbf{b} , \mathbf{c}). In a periodic system, the point \mathbf{x} is equivalent to any point \mathbf{x}' , provided the two points are related as follows:

$$\mathbf{x} = \mathbf{x}' + n_1 \mathbf{a} + n_2 \mathbf{b} + n_3 \mathbf{c} \quad (\text{A1.3.40})$$

where n_1, n_2, n_3 are arbitrary integers. This requirement can be used to define the translation vectors. Equation A1.3.40 can also be written as

$$\mathbf{x} = \mathbf{x}' + \mathbf{R}_{n_1, n_2, n_3} \quad (\text{A1.3.41})$$

-14-

where $\mathbf{R}_{n_1, n_2, n_3} = n_1 \mathbf{a} + n_2 \mathbf{b} + n_3 \mathbf{c}$ is called a *translation vector*. The set of points located by $\mathbf{R}_{n_1, n_2, n_3}$ formed by all possible combinations of (n_1, n_2, n_3) is called a *lattice*.

Knowing the lattice is usually not sufficient to reconstruct the crystal structure. A knowledge of the vectors (\mathbf{a} , \mathbf{b} , \mathbf{c}) does not specify the positions of the atoms within the unit cell. The positions of the atoms within the unit cell is given by a set of vectors: $\tau_j, j = 1, 2, 3 \dots n$ where n is the number of atoms in the unit cell. The set of vectors, τ_j , is called the *basis*. For simple elemental structures, the unit cell may contain only one atom. The lattice sites in this case can be chosen to correspond to the atomic sites, and no basis exists.

The position of the i th atom in a crystal, \mathbf{r}_i , is given by

$$\mathbf{r}_i = \tau_j + \sum_{n_1, n_2, n_3} \mathbf{R}_{n_1, n_2, n_3} \quad (\text{A1.3.42})$$

where the index j refers to the j th atom in the cell and the indices n_1, n_2, n_3 refer to the cell. The construction of the unit cell, i.e. the lattice vectors $\mathbf{R}_{n_1, n_2, n_3}$ and the basis vector τ_j , is not unique. The choice of unit cell is usually dictated by convenience. The smallest possible unit cell which properly describes a crystal is called the *primitive unit cell*.

(A) FACE-CENTRED CUBIC (FCC) STRUCTURE

The FCC structure is illustrated in [figure A1.3.2](#). Metallic elements such as calcium, nickel, and copper form in the FCC structure, as well as some of the inert gases. The conventional unit cell of the FCC structure is cubic with the length of the edge given by the *lattice parameter*, a . There are four atoms in the conventional cell. In the primitive unit cell, there is only one atom. This atom coincides with the lattice points. The lattice vectors for the primitive cell are given by

$$\mathbf{a} = a(\hat{\mathbf{y}} + \hat{\mathbf{z}})/2 \quad \mathbf{b} = a(\hat{\mathbf{x}} + \hat{\mathbf{z}})/2 \quad \mathbf{c} = a(\hat{\mathbf{x}} + \hat{\mathbf{y}})/2. \quad (\text{A1.3.43})$$

This structure is called ‘close packed’ because the number of atoms per unit volume is quite large compared with other simple crystal structures.

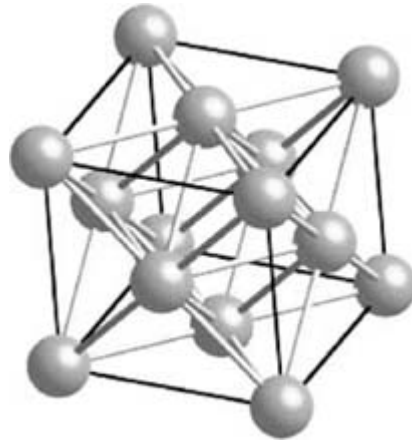


Figure A1.3.2. Structure of a FCC crystal.

(B) BODY-CENTRED CUBIC (BCC) STRUCTURE

The BCC structure is illustrated in [figure A1.3.3](#). Elements such as sodium, tungsten and iron form in the BCC structure. The conventional unit cell of the BCC structure is cubic, like FCC, with the length of the edge given by the *lattice parameter*, a . There are two atoms in the conventional cell. In the primitive unit cell, there is only one atom and the lattice vectors are given by

$$\mathbf{a} = a(-\hat{x} + \hat{y} + \hat{z})/2 \quad \mathbf{b} = a(\hat{x} - \hat{y} + \hat{z})/2 \quad \mathbf{c} = a(\hat{x} + \hat{y} - \hat{z})/2. \quad (\text{A1.3.44})$$

(C) DIAMOND STRUCTURE

The diamond structure is illustrated in [figure A1.3.4](#). Elements such as carbon, silicon and germanium form in the diamond structure. The conventional unit cell of the diamond structure is cubic with the length of the edge given by the *lattice parameter*, a . There are eight atoms in the conventional cell. The diamond structure can be constructed by considering two interpenetrating FCC crystals displaced one-fourth of the body diagonal. For the primitive unit cell, the lattice vectors are the same as for the FCC crystal; however, each lattice point has a basis associated with it. The basis can be chosen as

$$\boldsymbol{\tau}_1 = -a(1, 1, 1)/8 \quad \boldsymbol{\tau}_2 = a(1, 1, 1)/8. \quad (\text{A1.3.45})$$

(D) ROCKSALT STRUCTURE

The rocksalt structure is illustrated in [figure A1.3.5](#). This structure represents one of the simplest compound structures. Numerous ionic crystals form in the rocksalt structure, such as sodium chloride (NaCl). The conventional unit cell of the rocksalt structure is cubic. There are eight atoms in the conventional cell. For the primitive unit cell, the lattice vectors are the same as FCC. The basis consists of two atoms: one at the origin and one displaced by one-half the body diagonal of the conventional cell.

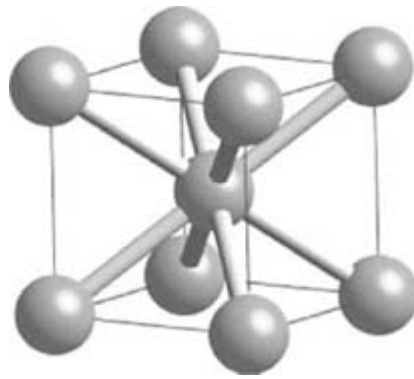


Figure A1.3.3. Structure of a BCC crystal.

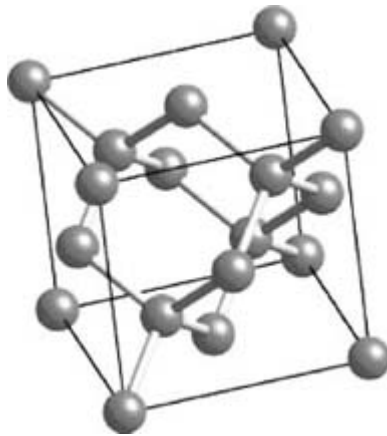


Figure A1.3.4. Structure of a diamond crystal.

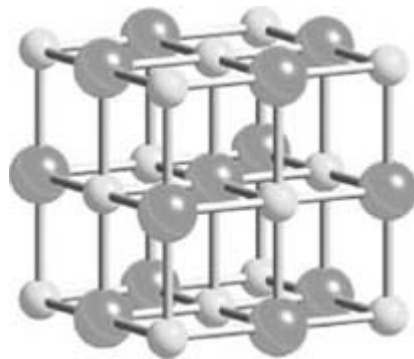


Figure A1.3.5. Structure of a rocksalt crystal.

A1.3.4.2 BLOCH'S THEOREM

The periodic nature of crystalline matter can be utilized to construct wavefunctions which reflect the translational symmetry. Wavefunctions so constructed are called *Bloch functions* [1]. These functions greatly simplify the electronic structure problem and are applicable to any periodic system.

For example, consider a simple crystal with one atom per lattice point: the total ionic potential can be written as

$$V_{\text{ion}}^{\text{xtal}}(\mathbf{r}) = \sum_{\mathbf{R}, \tau} V_{\text{ion}}^{\text{a}}(\mathbf{r} - \mathbf{R} - \tau). \quad (\text{A1.3.46})$$

This ionic potential is periodic. A translation of \mathbf{r} to $\mathbf{r} + \mathbf{R}$ can be accommodated by simply reordering the summation. Since the valence charge density is also periodic, the total potential is periodic as the Hartree and exchange-correlation potentials are functions of the charge density. In this situation, it can be shown that the wavefunctions for crystalline matter can be written as

$$\phi_{\mathbf{k}}(\mathbf{r}) = \exp(i\mathbf{k} \cdot \mathbf{r})u_{\mathbf{k}}(\mathbf{r}) \quad (\text{A1.3.47})$$

where \mathbf{k} is a wavevector and $u_{\mathbf{k}}(\mathbf{r})$ is a periodic function, $u_{\mathbf{k}}(\mathbf{r} + \mathbf{R}) = u_{\mathbf{k}}(\mathbf{r})$. This is known as *Bloch's theorem*. In the limit of a free electron, \mathbf{k} can be identified with the momentum of the electron and $u_{\mathbf{k}} = 1$.

The wavevector is a *good* quantum number: e.g., the orbitals of the Kohn–Sham equations [21] can be rigorously labelled by \mathbf{k} and spin. In three dimensions, four quantum numbers are required to characterize an eigenstate. In spherically symmetric atoms, the numbers correspond to n, l, m, s , the principal, angular momentum, azimuthal and spin quantum numbers, respectively. Bloch's theorem states that the equivalent quantum numbers in a crystal are k_x, k_y, k_z and spin. The spin index is usually dropped for non-magnetic materials.

By taking the $\phi_{\mathbf{k}}$ orbitals to be of the Bloch form, the Kohn–Sham equations can be written as

$$\left(\frac{(\mathbf{p} + \hbar\mathbf{k})^2}{2m} + V_{\text{N}}(\mathbf{r}) + V_{\text{H}}(\mathbf{r}) + V_{\text{xc}}[\rho(\mathbf{r})] \right) u_{\mathbf{k}}(\mathbf{r}) = E(\mathbf{k})u_{\mathbf{k}}(\mathbf{r}). \quad (\text{A1.3.48})$$

Knowing the energy distributions of electrons, $E(\mathbf{k})$, and the spatial distribution of electrons, $\rho(\mathbf{r})$, is important in obtaining the structural and electronic properties of condensed matter systems.

A1.3.5 ENERGY BANDS FOR CRYSTALLINE SOLIDS

A1.3.5.1 KRONIG–PENNEY MODEL

One of the first models to describe electronic states in a periodic potential was the *Kronig–Penney* model [1]. This model is commonly used to illustrate the fundamental features of Bloch's theorem and solutions of the Schrödinger

equation for a periodic system.

This model considers the solution of wavefunctions for a one-dimensional Schrödinger equation:

$$\left[\frac{-\hbar^2 \nabla^2}{2m} + V(x) \right] \psi(x) = E \psi(x). \quad (\text{A1.3.49})$$

This Schrödinger equation has a particularly simple solution for a finite energy well: $V(x) = -V_0$ for $0 < x < a$ (region I) and $V(x) = 0$ elsewhere (region II) as indicated in [figure A1.3.6](#). This is a standard problem in

elementary quantum mechanics. For a bound state ($E < 0$) the wavefunctions have solutions in region I: $\psi_I(x) = B \exp(iKx) + C \exp(-iKx)$ and in region II: $\psi_{II}(x) = A \exp(-Q|x|)$. The wavefunctions are required to be continuous: $\psi_I(0) = \psi_{II}(0)$ and $\psi_I(a) = \psi_{II}(a)$ and have continuous first derivatives: $\psi_I'(0) = \psi_{II}'(0)$ and $\psi_I'(a) = \psi_{II}'(a)$. With these conditions imposed at $x = 0$

$$B/C = -(1 + iK/Q)^2 / (1 + K^2/Q^2) \quad (\text{A1.3.50})$$

and at $x = a$

$$B/C = -(1 - iK/Q)^2 \exp(-2iKa) / (1 + K^2/Q^2). \quad (\text{A1.3.51})$$

A nontrivial solution will exist only if

$$(1 + iK/Q)^2 = (1 - iK/Q)^2 \exp(-2iKa) \quad (\text{A1.3.52})$$

or

$$Q^2 - 2QK \cot(Ka) - K^2 = 0. \quad (\text{A1.3.53})$$

This results in two solutions:

$$Q = -K \cot(Ka/2) \quad \text{and} \quad Q = K \tan(Ka/2). \quad (\text{A1.3.54})$$

If ψ_I and ψ_{II} are inserted into the one-dimensional Schrödinger equation, one finds $E = \hbar^2 K^2 / 2m - V_0$ or $K = \sqrt{2m(E + V_0) / \hbar^2}$ and $E = -\hbar^2 Q^2 / 2m$. In the limit $V_0 \rightarrow \infty$, or $K \rightarrow \infty$, equation A1.3.53 can result in a finite value for Q only if $\tan(Ka/2) \rightarrow 0$, or $\cot(Ka/2) \rightarrow 0$ (i.e. $Ka = n\pi$ where n is an integer). The energy levels in this limit correspond to the standard 'particle in a box' eigenvalues:

$$E_n = \frac{\hbar^2 (2n\pi/a)^2}{2m}.$$

In the Kronig–Penney model, a periodic array of such potentials is considered as illustrated in figure A1.3.6. The width of the wells is a and the wells are separated by a distance b . Space can be divided to distinct regions: region I ($-b < x < 0$), region II ($0 < x < a$) and region III ($a < x < a + b$). In region I, the wavefunction can be taken as

$$\psi_I(x) = C \exp(Qx) + D \exp(-Qx). \quad (\text{A1.3.55})$$

In region II, the wavefunction is

$$\psi_{II}(x) = A \exp(iKx) + B \exp(-iKx). \quad (\text{A1.3.56})$$

Unlike an isolated well, there is no restriction on the sign on the exponentials, i.e. both $\exp(+Qx)$ and $\exp(-Qx)$ are allowed. For an isolated well, the sign was restricted so that the exponential vanished as $|x| \rightarrow \infty$. Either sign is allowed for the periodic array as the extent of the wavefunction within each region is finite.

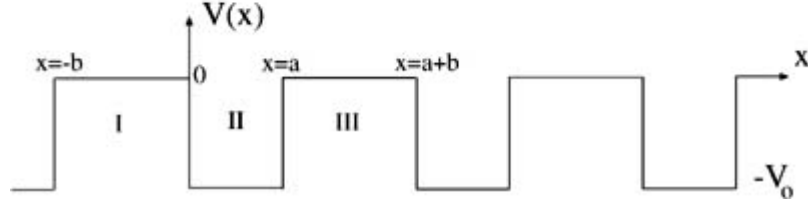
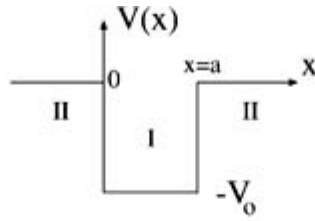


Figure A1.3.6. An isolated square well (top). A periodic array of square wells (bottom). This model is used in the Kronig–Penney description of energy bands in solids.

Because our system is periodic, one need only consider the wavefunctions in I and II and apply the periodic boundary conditions for other regions of space. Bloch's theorem can be used in this case: $\psi(x + a) = \exp(ika) \psi(x)$ or $\psi(x + (a + b)) = \exp(ik(a + b)) \psi(x)$. This relates ψ_{III} and ψ_I :

$$\psi_{III}(x) = \exp(ik(a + b)) \psi_I(x) \quad (\text{A1.3.57})$$

or

$$\psi_{III}(x) = \exp(ik(a + b))(C \exp(Q(x - a - b)) + D \exp(-Q(x - a - b))). \quad (\text{A1.3.58})$$

-20-

k now serves to label the states in the same sense n serves to label states for a square well.

As in the case of the isolated well, one can impose continuity of the wavefunctions and the first derivatives of the wavefunctions at $x = 0$ and $x = a$. At $x = 0$,

$$A + B = C + D \quad iK(A - B) = Q(C - D) \quad (\text{A1.3.59})$$

and at $x = a$

$$A \exp(iKa) + B \exp(-iKa) = \exp(ik(a + b))(C \exp(-Qb) + D \exp(Qb)) \quad (\text{A1.3.60})$$

$$iKa(A \exp(iKa) - B \exp(-iKa)) = Q \exp(ik(a + b))(C \exp(-Qb) + D \exp(Qb)). \quad (\text{A1.3.61})$$

This results in four equations and four unknowns. Since the equations are homogeneous, a nontrivial solution exists only if the determinant formed by the coefficients of A , B , C and D vanishes. The solution to this equation is

$$\frac{(Q^2 - K^2)}{2QK} \sinh(Qb) \sin(Ka) + \cosh(Qb) \cos(Ka) = \cos(k(a + b)). \quad (\text{A1.3.62})$$

Equation A1.3.62 provides a relationship between the wavevector, k , and the energy, E , which is implicit in Q and K .

Before this result is explored in more detail, consider the limit where $b \rightarrow \infty$. In this limit, the wells become isolated and k has no meaning. As $b \rightarrow \infty$, $\sinh(Qb) \rightarrow \exp(Qb)/2$ and $\cosh(Qb) \rightarrow \exp(Qb)/2$. One can rewrite equation A1.3.62 as

$$(\exp(Qb)/2) \left(\frac{(Q^2 - K^2)}{2QK} \sin(Ka) + \cos(Ka) \right) = \cos(k(a+b)). \quad (\text{A1.3.63})$$

As $\exp(Qb)/2 \rightarrow \infty$, this equation can be valid if

$$\frac{(Q^2 - K^2)}{2QK} \sin(Ka) + \cos(Ka) \rightarrow 0 \quad (\text{A1.3.64})$$

otherwise the rhs of equation A1.3.63 would diverge. In this limit, equation A1.3.64 reduces to the isolated well solution (equation A1.3.53):

$$Q^2 - 2QK \cot(Ka) - K^2 = 0. \quad (\text{A1.3.65})$$

Since k does not appear in equation A1.3.65 in this limit, it is undefined.

One can illustrate how the energy states evolve from discrete levels in an isolated well to states appropriate for periodic systems by varying the separation between wells. In figure A1.3.7 solutions for E versus k are shown for isolated wells and for strongly interacting wells. It is important to note that k is not defined except within a factor of $2\pi m/(a+b)$ where m is an integer as $\cos((k + 2\pi m/(a+b))(a+b)) = \cos(k(a+b))$. The E versus k plot need be displayed only for k between 0 and $\pi/(a+b)$ as larger values of k can be mapped into this interval by subtracting off values of $2\pi/(a+b)$.

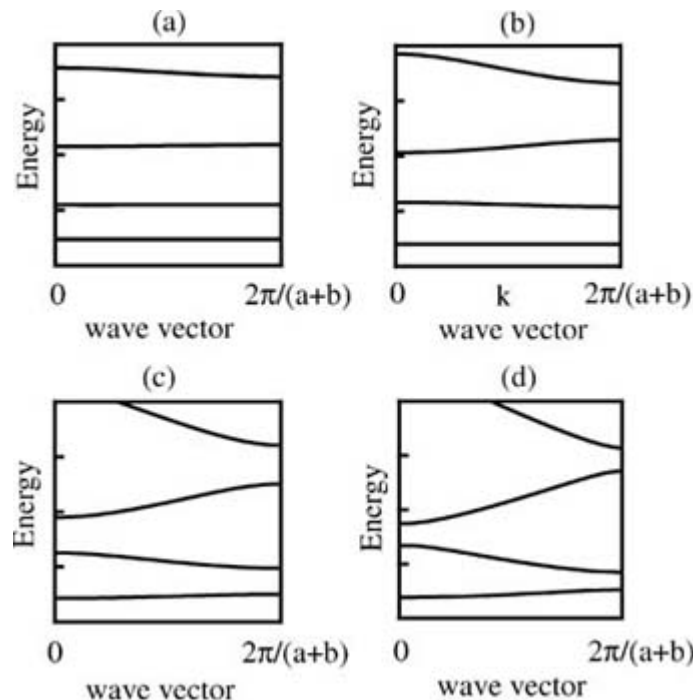


Figure A1.3.7. Evolution of energy bands in the Kronig–Penney model as the separation between wells, b (figure A1.3.6) is decreased from (a) to (d). In (a) the wells are separated by a large distance (large value of b) and the energy bands resemble discrete levels of an isolated well. In (d) the wells are quite close together (small value of b) and the energy bands are free-electron-like.

In the case where the wells are far apart, the resulting energy levels are close to the isolated well. However, an interesting phenomenon occurs as the atoms are brought closer together. The energy levels cease being constant as a function of the wavevector, k . There are regions of allowed solutions and regions where no energy state occurs. The region of allowed energy values is called an *energy band*. The range of energies within the band is called the *band width*. As the width of the band increases, it is said that the band has greater *dispersion*.

The Kronig–Penney solution illustrates that, for *periodic* systems, gaps can exist between bands of energy states. As for the case of a free electron gas, each band can hold $2N$ electrons where N is the number of wells present. In one dimension, this implies that if a well contains an *odd* number, one will have *partially occupied* bands. If one has an *even* number of electrons per well, one will have *fully occupied energy* bands. This distinction between odd and even numbers of electrons per cell is of fundamental importance. The Kronig–Penney model implies that crystals with an odd number of electrons per unit cell are *always* metallic whereas an even number of electrons per unit cell implies an

insulating state. This simple rule is valid for more realistic potentials and need be only slightly modified in three dimensions. In three dimensions, an even number of electrons per unit cells is a necessary condition for an insulating state, but not a sufficient condition.

One of the major successes of *energy band* theory is that it can be used to predict whether a crystal exists as a metal or insulator. If a band is filled, the Pauli principle prevents electrons from changing their momentum in response to the electric field as all possible momentum states are occupied. In a metal this constraint is not present as an electron can change its momentum state by moving from a filled to an occupied state within a given band. The distinct types of energy bands for insulators, metals, semiconductors and semimetals are schematically illustrated in figure A1.3.8. In an insulator, energy bands are either completely empty or completely filled. The band gap between the highest occupied band and lowest empty band is large, e.g. above 5 eV. In a semiconductor, the bands are also completely filled or empty, but the gap is smaller, e.g. below 3 eV. In metals bands are not completely occupied and no gap between filled and empty states occurs. Semimetals are a special case. No gap exists, but one band is almost completely occupied; it overlaps with a band that is almost completely empty.

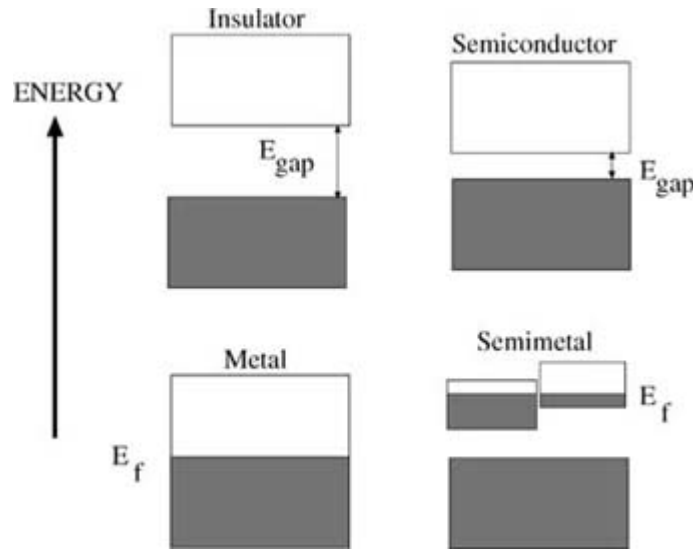


Figure A1.3.8. Schematic energy bands illustrating an insulator (large band gap), a semiconductor (small band gap), a metal (no gap) and a semimetal. In a semimetal, one band is almost filled and another band is almost empty.

A1.3.5.2 RECIPROCAL SPACE

Expressing $E(\mathbf{k})$ is complicated by the fact that \mathbf{k} is not unique. In the Kronig–Penney model, if one replaced k by $k + 2\pi/(a + b)$, the energy remained unchanged. In three dimensions \mathbf{k} is known only to within a *reciprocal lattice vector*, \mathbf{G} . One can define a set of reciprocal vectors, given by

$$\mathbf{G} = m_1\mathbf{A} + m_2\mathbf{B} + m_3\mathbf{C} \quad (\text{A1.3.66})$$

where the set $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ define a lattice in reciprocal space. These vectors can be defined by

-23-

$$\mathbf{A} = \frac{2\pi}{\Omega}\mathbf{b} \times \mathbf{c} \quad \mathbf{B} = \frac{2\pi}{\Omega}\mathbf{c} \times \mathbf{a} \quad \mathbf{C} = \frac{2\pi}{\Omega}\mathbf{a} \times \mathbf{b} \quad (\text{A1.3.67})$$

where Ω is defined as the unit cell volume. Note that $\Omega = |\mathbf{a} \cdot \mathbf{b} \times \mathbf{c}|$ from elementary vector analysis. It is easy to show that

$$\begin{array}{lll} \mathbf{A} \cdot \mathbf{a} = 2\pi & \mathbf{A} \cdot \mathbf{b} = 0 & \mathbf{A} \cdot \mathbf{c} = 0 \\ \mathbf{B} \cdot \mathbf{a} = 0 & \mathbf{B} \cdot \mathbf{b} = 2\pi & \mathbf{B} \cdot \mathbf{c} = 0 \\ \mathbf{C} \cdot \mathbf{a} = 0 & \mathbf{C} \cdot \mathbf{b} = 0 & \mathbf{C} \cdot \mathbf{c} = 2\pi. \end{array} \quad (\text{A1.3.68})$$

It is apparent that

$$\mathbf{G} \cdot \mathbf{R} = 2\pi(n_1m_1 + n_2m_2 + n_3m_3). \quad (\text{A1.3.69})$$

Reciprocal lattice vectors are useful in defining periodic functions. For example, the valence charge density, $\rho(\mathbf{r})$, can be expressed as

$$\rho(\mathbf{r}) = \sum_{\mathbf{G}} \rho(\mathbf{G}) \exp(i\mathbf{G} \cdot \mathbf{r}). \quad (\text{A1.3.70})$$

It is clear that $\rho(\mathbf{r} + \mathbf{R}) = \rho(\mathbf{r})$ from equation A1.3.69. The Fourier coefficients, $\rho(\mathbf{G})$, can be determined from

$$\rho(\mathbf{G}) = \frac{1}{\Omega} \int \rho(\mathbf{r}) \exp(-i\mathbf{G} \cdot \mathbf{r}) d^3r. \quad (\text{A1.3.71})$$

Because $E(\mathbf{k}) = E(\mathbf{k} + \mathbf{G})$, a knowledge of $E(\mathbf{k})$ within a given volume called the *Brillouin zone* is sufficient to determine $E(\mathbf{k})$ for all \mathbf{k} . In one dimension, $G = 2\pi n/d$ where d is the lattice spacing between atoms. In this case, $E(k)$ is known once k is determined for $-\pi/d < k < \pi/d$. (For example, in the Kronig–Penney model (figure A1.3.6), $d = a + b$ and k was defined only to within a vector $2\pi/(a + b)$.) In three dimensions, this subspace can result in complex polyhedrons for the Brillouin zone.

-24-

In figure A1.3.9 the Brillouin zone for a FCC and a BCC crystal are illustrated. It is a common practice to label high-symmetry point and directions by letters or symbols. For example, the $\mathbf{k} = 0$ point is called the Γ point. For cubic crystals, there exist 48 symmetry operations and this symmetry is maintained in the energy bands: e.g., $E(k_x, k_y, k_z)$ is invariant under sign permutations of (x, y, z) . As such, one need only have knowledge of $E(\mathbf{k})$ in $\frac{1}{48}$ of the zone to determine the energy band throughout the zone. The part of the zone which cannot be reduced by symmetry is called the *irreducible Brillouin zone*.

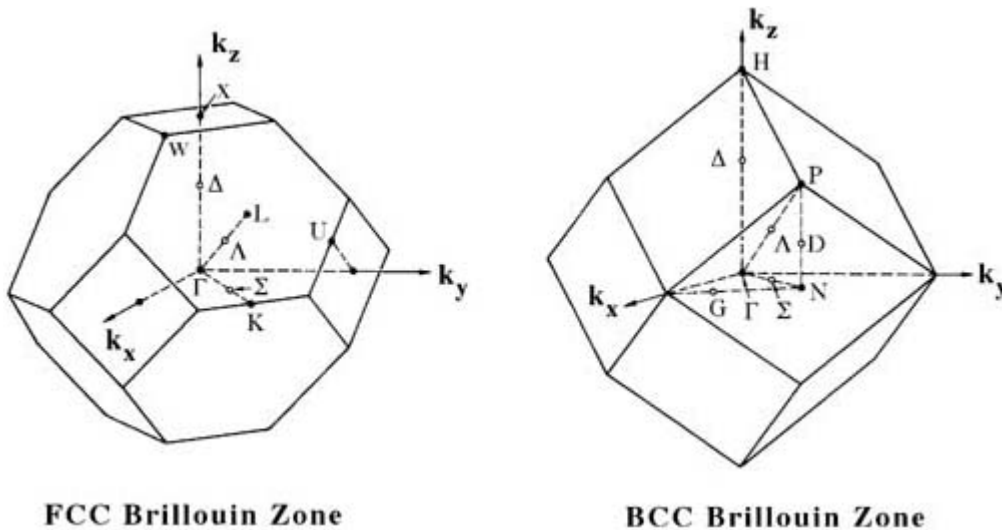


Figure A1.3.9. Brillouin zones for the FCC and BCC crystal structures.

A1.3.5.3 REALISTIC ENERGY BANDS

Since the electronic structure of a solid can be determined from a knowledge of the spatial and energetic distribution of electrons (i.e. from the charge density, $\rho(\mathbf{r})$, and the electronic density of states, $D(E)$), it is highly desirable to have the ability to determine the quantum states of crystal. The first successful electronic structure calculations for energy bands of crystalline matter were not performed from ‘first principles’. Although elements of density functional theory were understood by the mid-1960s, it was not clear how reliable these methods were. Often, two seemingly identical calculations would yield very different results for simple issues such as whether a solid was a metal or an insulator. Consequently, some of the first reliable energy bands were constructed using *empirical pseudopotentials* [25]. These potentials were extracted from experimental data and not determined from first principles.

A1.3.5.4 EMPIRICAL PSEUDOPOTENTIALS

The first reliable energy band theories were based on a powerful approximation, call the *pseudopotential approximation*. Within this approximation, the *all-electron potential* corresponding to interaction of a valence electron with the inner, core electrons and the nucleus is replaced by a pseudopotential. The pseudopotential reproduces only the properties of the outer electrons. There are rigorous theorems such as the *Phillips–Kleinman* cancellation theorem that can be used to justify the pseudopotential model [2, 3, 26]. The Phillips–Kleinman cancellation theorem states that the *orthogonality* requirement of the valence states to the core states can be described by an effective repulsive

-25-

potential. This repulsive potential cancels the strong Coulombic potential within the core region. The cancellation theorem explains, in part, why valence electrons feel a less attractive potential than would be expected on the basis of the Coulombic part of the potential. For example, in alkali metals an ‘empty’ core pseudopotential approximation is often made. In this model pseudopotential, the valence electrons experience no Coulomb potential within the core region.

Since the pseudopotential does not bind the core states, it is a very weak potential. Simple basis functions can be used to describe the pseudo-wavefunctions. For example, a simple grid or plane wave basis will yield a converged solution [25]. The simplicity of the basis is important as it results in an unbiased, flexible description of the charge density. Also, since the nodal structure of the pseudo-wavefunctions has been removed, the charge density varies slowly in the core region. A schematic model of the pseudopotential model is illustrated in figure A1.3.10. The pseudopotential model describes a solid as a sea of valence electrons moving in a periodic background of cores (composed of nuclei and inert core electrons). In this model many of the complexities of all-electron calculations, calculations that include the core and valence electrons on an equal footing, are avoided. A group IV solid such as C with 6 electrons per atom is treated in a similar fashion to Sn with 50 electrons per atom since both have 4 valence electrons per atom. In addition, the focus of the calculation is only on the accuracy of the valence electron wavefunction in the spatial region away from the chemically inert core.

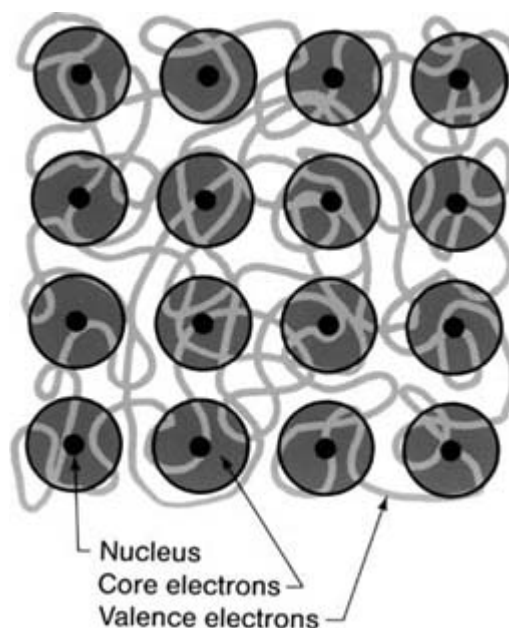


Figure A1.3.10. Pseudopotential model. The outer electrons (valence electrons) move in a fixed arrangement of chemically inert ion cores. The ion cores are composed of the nucleus and core electrons.

One can quantify the pseudopotential by writing the total crystalline potential for an elemental solid as

$$V_p(\mathbf{r}) = \sum_{\mathbf{G}} S(\mathbf{G}) V_p^a(\mathbf{G}) \exp(i\mathbf{G} \cdot \mathbf{r}). \quad (\text{A1.3.72})$$

-26-

$S(\mathbf{G})$ is the *structure factor* given by

$$S(\mathbf{G}) = \frac{1}{N_a} \sum_{\tau} \exp(i\mathbf{G} \cdot \tau) \quad (\text{A1.3.73})$$

where N_a is the number of atoms in the unit cell and τ is a basis vector. $\frac{2\pi}{a}(\mathbf{G})$ is the *form factor* given by

$$V_p^a(\mathbf{G}) = \frac{1}{\Omega_a} \int V_p^a(\mathbf{r}) \exp(i\mathbf{G} \cdot \mathbf{r}) d^3r \quad (\text{A1.3.74})$$

where Ω_a is the volume per atom and $\frac{2\pi}{a}(\mathbf{r})$ is a pseudopotential associated with an atom. Often this potential is assumed to be spherically symmetry. In this case, the form factor depends only on the magnitude of \mathbf{G} : $\frac{2\pi}{a}(\mathbf{G}) = \frac{2\pi}{a}(|\mathbf{G}|)$. A schematic pseudopotential is illustrated in figure A1.3.11. Outside the core region the pseudopotential is commensurate with the all-electron potential. When this potential is transformed into Fourier space, it is often sufficient to keep just a few unique form factors to characterize the potential. These form factors are then treated as adjustable parameters which can be fitted to experimental data. This is illustrated in [figure A1.3.12](#).

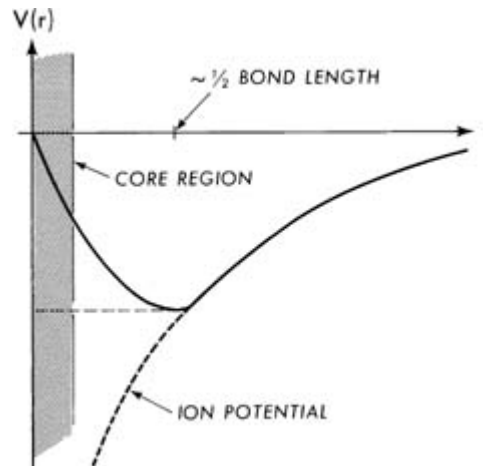


Figure A1.3.11. Schematic pseudopotential in real space.

-27-

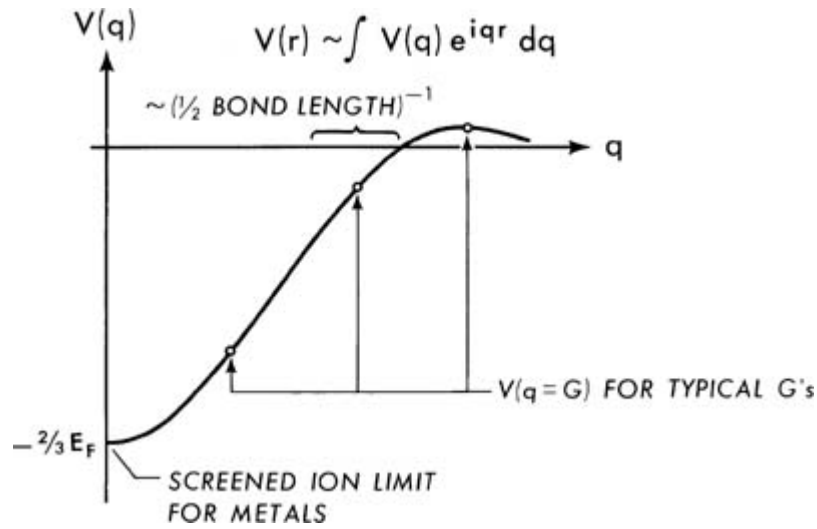


Figure A1.3.12. Schematic pseudopotential in reciprocal space.

The empirical pseudopotential method can be illustrated by considering a specific semiconductor such as silicon. The crystal structure of Si is diamond. The structure is shown in [figure A1.3.4](#). The lattice vectors and basis for a primitive cell have been defined in the section on crystal structures ([A1.3.4.1](#)). In Cartesian coordinates, one can write \mathbf{G} for the diamond structure as

$$\mathbf{G} = \frac{2\pi}{a}(n, l, m) \quad (\text{A1.3.75})$$

where the indices (n, l, m) must be either all odd or all even: e.g., $\mathbf{G} = \frac{2\pi}{a}(1, 0, 0)$ is not allowed, but $\mathbf{G} = \frac{2\pi}{a}(2, 0, 0)$ is permitted. It is convenient to organize \mathbf{G} -vectors by their magnitude squared in units of $(2\pi/a)^2$. In this scheme: $G^2 = 0, 3, 4, 8, 11, 12, \dots$. The structure factor for the diamond structure is $S(\mathbf{G}) = \cos(\mathbf{G} \cdot \boldsymbol{\tau})$. For some values of G , this structure factor vanishes: e.g., if $\mathbf{G} = (2\pi/a)(2, 0, 0)$, then $\mathbf{G} \cdot \boldsymbol{\tau} = \pi/2$ and $S(\mathbf{G}) = 0$. If the structure factor vanishes, the corresponding form factor is irrelevant as it is multiplied by a zero structure factor. In the case of diamond structure, this eliminates the $G^2 = 4, 12$ form factors. Also, the $G^2 = 0$ factor is not important for spectroscopy as it corresponds to the average potential and serves to shift the energy bands by a constant. The rapid convergence of the pseudopotential in Fourier space coupled with the vanishing of the structure factor for certain \mathbf{G} means that only three form factors are required to fix the energy bands for diamond semiconductors like Si and Ge: $\frac{2\pi}{a}(G^2 = 3)$, $\frac{2\pi}{a}(G^2 = 8)$ and $\frac{2\pi}{a}(G^2 = 11)$. These form factors can be fixed by comparisons to reflectivity measurements or photoemission [25].

A1.3.5.5 DENSITY FUNCTIONAL PSEUDOPOTENTIALS

Another realistic approach is to construct pseudopotentials using density functional theory. The implementation of the Kohn–Sham equations to condensed matter phases without the pseudopotential approximation is not easy owing to the dramatic span in length scales of the wavefunction and the energy range of the eigenvalues. The pseudopotential eliminates this problem by removing the core electrons from the problem and results in a much simpler problem [27].

In the pseudopotential construction, the atomic wavefunctions for the valence electrons are taken to be nodeless. The pseudo-wavefunction is taken to be identical to the appropriate all-electron wavefunction in the regions of interest for solid-state effects. For the core region, the wavefunction is extrapolated back to the

origin in a manner consistent with the normalization condition. This type of construction was first introduced by Fermi to account for the shift in the wavefunctions of high-lying states of alkali atoms subject to perturbations from foreign atoms. In this remarkable paper, Fermi introduced the conceptual basis for both the pseudopotential and the scattering length [28].

With the density functional theory, the first step in the construction of a pseudopotential is to consider the solution for an isolated atom [27]. If the atomic wavefunctions are known, the pseudo-wavefunction can be constructed by removing the nodal structure of the wavefunction. For example, if one considers a valence wavefunction for the isolated atom, $\psi_v(r)$, then a pseudo-wavefunction, $\phi_p(r)$, might have the properties

$$\begin{aligned} \phi_p(r) &= r^l \exp(-\alpha r^4 - \beta r^3 - \gamma r^2 - \delta) & r < r_c \\ &= \psi_v(r) & r > r_c. \end{aligned} \tag{A1.3.76}$$

The pseudo-wavefunction within this frame work is guaranteed to be nodeless. The parameters $(\alpha, \beta, \gamma, \delta)$ are fixed so that (1) ϕ_v and ϕ_p have the same eigenvalue, \mathcal{E}_v , and the same norm:

$$\int_0^{r_c} |\psi_v(r)|^2 r^2 dr = \int_0^{r_c} |\phi_p(r)|^2 r^2 dr. \tag{A1.3.77}$$

This ensures that $\phi_p(r) = \psi_v(r)$ for $r > r_c$ after the wavefunctions have been normalized. (2) The pseudo-wavefunction should be continuous and have continuous first and second derivatives at r_c . An example of a pseudo-wavefunction is given in [figure A1.3.13](#).

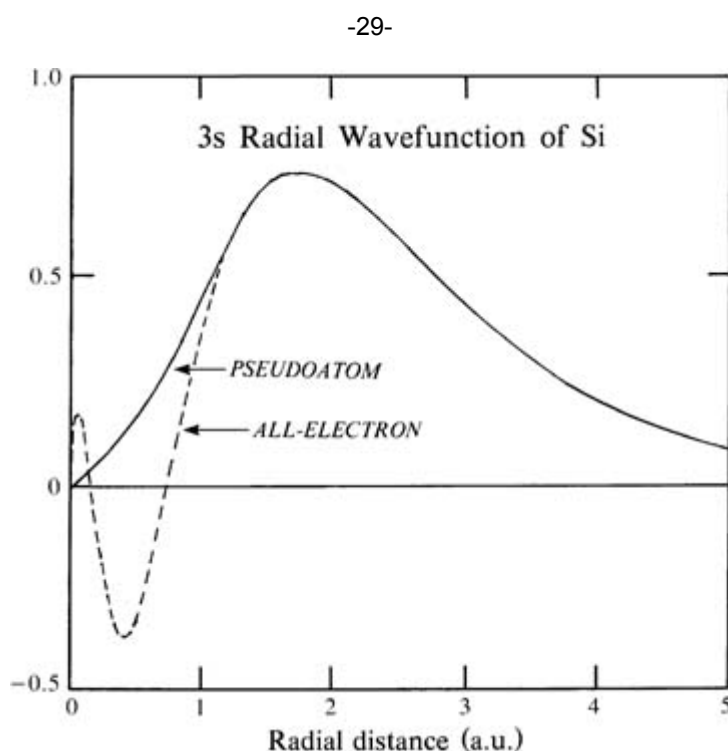


Figure A1.3.13. All-electron and pseudopotential wavefunction for the 3s state in silicon. The all-electron 3s state has nodes which arise because of an orthogonality requirement to the 1s and 2s core states.

Once the eigenvalue and pseudo-wavefunction are known for the atom, the Kohn–Sham equation can be inverted to yield the ionic pseudopotential:

$$V_{\text{ion}}^{\text{p}}(\mathbf{r}) = -\mathcal{E}_{\text{v}} - V_{\text{H}}(\mathbf{r}) - V_{\text{xc}}(\mathbf{r}) + \frac{\hbar^2 \nabla^2 \phi_{\text{p}}}{2m\phi_{\text{p}}}. \quad (\text{A1.3.78})$$

Since V_{H} and V_{xc} depend only on the valence charge densities, they can be determined once the valence pseudo-wavefunctions are known. Because the pseudo-wavefunctions are nodeless, the resulting pseudopotential is well defined despite the last term in equation A1.3.78. Once the pseudopotential has been constructed from the atom, it can be transferred to the condensed matter system of interest. For example, the ionic pseudopotential defined by equation A1.3.78 from an *atomistic* calculation can be transferred to *condensed matter phases* without any significant loss of accuracy.

There are complicating issues in defining pseudopotentials, e.g. the pseudopotential in equation A1.3.78 is state dependent, orbitally dependent and the energy and spatial separations between valence and core electrons are sometimes not transparent. These are not insurmountable issues. The state dependence is usually weak and can be ignored. The orbital dependence requires different potentials for different angular momentum components. This can be incorporated via *non-local* operators. The distinction between valence and core states can be addressed by incorporating the core level in question as part of the valence shell. For example, in Zn one can treat the $3d^{10}$ shell as a valence shell. In this case, the valency of Zn is 12, not 2. There are also very reliable approximate methods for treating the outer core states without explicitly incorporating them in the valence shell.

-30-

A1.3.5.6 OTHER APPROACHES

There are a variety of other approaches to understanding the electronic structure of crystals. Most of them rely on a density functional approach, with or without the pseudopotential, and use different bases. For example, instead of a plane wave basis, one might write a basis composed of atomic-like orbitals:

$$\psi_{\mathbf{k}}(\mathbf{r}) = \sum_{i, \mathbf{R}} \alpha_i(\mathbf{k}) \exp(i\mathbf{k} \cdot \mathbf{R}) \phi_i(\mathbf{r} - \mathbf{R}) \quad (\text{A1.3.79})$$

where the $\exp(i\mathbf{k} \cdot \mathbf{R})$ is explicitly written to illustrate the Bloch form of this wavefunction: i.e. $\psi_{\mathbf{k}}(\mathbf{r} + \mathbf{R}) = \exp(i\mathbf{k} \cdot \mathbf{R}) \psi_{\mathbf{k}}(\mathbf{r})$. The orbitals ϕ_i can be taken from atomic structure solutions where i is a general index such as $lmns$, or ϕ_i can be taken to be a some localized function such as an exponential, called a Slater-type orbital, or a Gaussian orbital. Provided the basis functions are appropriately chosen, this approach works quite well for a wide variety of solids. This approach is called the *tight binding method* [2, 7].

An approach closely related to the pseudopotential is the *orthogonalized plane wave method* [29]. In this method, the basis is taken to be as follows:

$$\phi_{\mathbf{k}}^{\text{OPW}}(\mathbf{r}) = \exp(i\mathbf{k} \cdot \mathbf{r}) - \sum_i \beta_i \chi_{i, \mathbf{k}}(\mathbf{r}) \quad (\text{A1.3.80})$$

and

$$\chi_{i, \mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{R}} \exp(i\mathbf{k} \cdot \mathbf{R}) a_i(\mathbf{r} - \mathbf{R}) \quad (\text{A1.3.81})$$

where $\chi_{i, \mathbf{k}}$ is a tight binding wavefunction composed of atomic core functions, a_i . As an example, one would take (a_{1s}, a_{2s}, a_{2p}) atomic orbitals for the core states of silicon. The form for $\phi_{\mathbf{k}}(\mathbf{r})$ is motivated by several factors. In the interstitial regions of a crystal, the potential should be weak and slowly varying. The

wavefunction should look like a plane wave in this region. Near the nucleus, the wavefunction should look atomic-like. The basis reflects these different regimes by combining plane waves with atomic orbitals. Another important attribute of the wavefunction is an *orthogonality* condition. This condition arises from the form of the Schrödinger equation; higher-energy eigenvalues must have wavefunctions which are orthogonal to more tightly bound states of the same symmetry: e.g., the 2s wavefunction of an atom must be orthogonal to the 1s state. It is possible to choose β_i so that

$$\int \phi_{\mathbf{k}}^*(\mathbf{r}) \chi_{i,\mathbf{k}}(\mathbf{r}) d^3r = 0. \quad (\text{A1.3.82})$$

The orthogonality condition assures one that the lowest energy state will not converge to core-like states, but valence states. The wavefunction for the solid can be written as

$$\psi_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}} \alpha(\mathbf{k}, \mathbf{G}) \phi_{\mathbf{k}}^{\text{OPW}}(\mathbf{r}). \quad (\text{A1.3.83})$$

-31-

As with any basis method the $\alpha(\mathbf{k}, \mathbf{G})$ coefficients are determined by solving a secular equation.

Other methods for determining the energy band structure include cellular methods, Green function approaches and augmented plane waves [2, 3]. The choice of which method to use is often dictated by the particular system of interest. Details in applying these methods to condensed matter phases can be found elsewhere (see [section B3.2](#)).

A1.3.6 EXAMPLES FOR THE ELECTRONIC STRUCTURE AND ENERGY BANDS OF CRYSTALS

Many phenomena in solid-state physics can be understood by resort to energy band calculations. Conductivity trends, photoemission spectra, and optical properties can all be understood by examining the quantum states or energy bands of solids. In addition, electronic structure methods can be used to extract a wide variety of properties such as structural energies, mechanical properties and thermodynamic properties.

A1.3.6.1 SEMICONDUCTORS

A prototypical semiconducting crystal is silicon. Historically, silicon has been the testing ground for quantum theories of condensed matter. This is not surprising given the importance of silicon for technological applications. The energy bands for Si are shown in [figure A1.3.14](#). Each band can hold two electrons per unit cell. There are four electrons per silicon atom and two atoms in the unit cell. This would lead to four filled bands. It is customary to show the filled bands and the lowest few empty bands. In the case of silicon the bands are separated by a gap of approximately 1 eV. Semiconductors have *band gaps* that are less than a few electronvolts. Displaying the energy bands is not a routine matter as $E(\mathbf{k})$ is often a complex function. The bands are typically displayed only along high-symmetry directions in the Brillouin zone (see [figure A1.3.9](#)). For example, one might plot the energy bands along the (100) direction (the Δ direction).

-32-

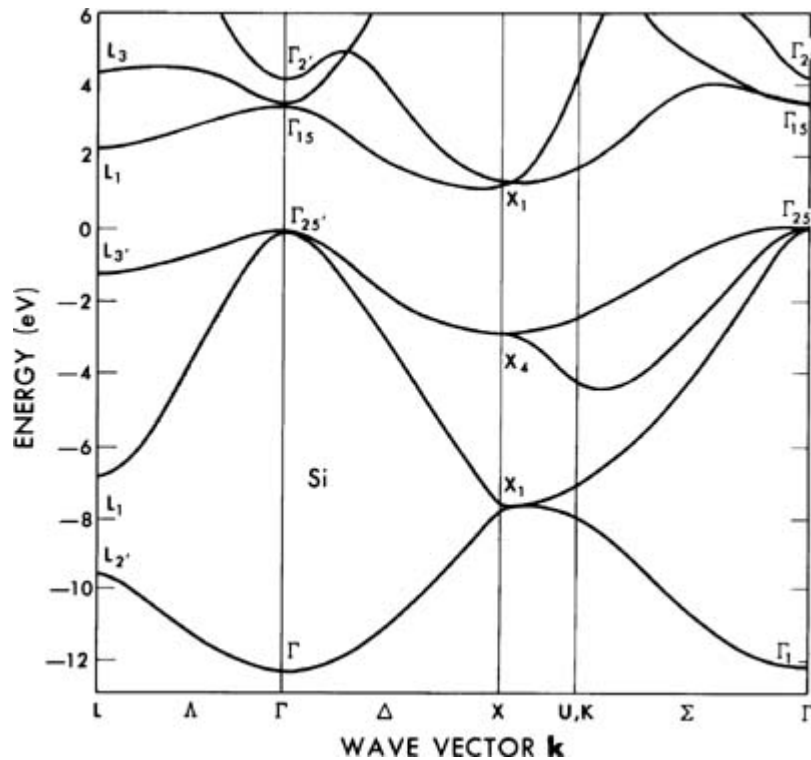


Figure A1.3.14. Band structure for silicon as calculated from empirical pseudopotentials [25].

The occupied bands are called *valence* bands; the empty bands are called *conduction* bands. The top of the valence band is usually taken as energy zero. The lowest conduction band has a minimum along the Δ direction; the highest occupied valence band has a maximum at Γ . Semiconductors which have the highest occupied \mathbf{k}_v -state and lowest empty state \mathbf{k}_c at different points are called *indirect* gap semiconductors. If $\mathbf{k}_v = \mathbf{k}_c$, the semiconductor is called *direct* gap semiconductor. Germanium is also an indirect gap semiconductor whereas GaAs has a direct gap. It is not easy to predict whether a given semiconductor will have a direct gap or not.

Electronic and optical excitations usually occur between the upper valence bands and lowest conduction band. In optical excitations, electrons are transferred from the valence band to the conduction band. This process leaves an empty state in the valence band. These empty states are called *holes*. Conservation of wavevectors must be obeyed in these transitions: $\mathbf{k}_{\text{photon}} + \mathbf{k}_v = \mathbf{k}_c$ where $\mathbf{k}_{\text{photon}}$ is the wavevector of the photon, \mathbf{k}_v is the wavevector of the electron in the initial valence band state and \mathbf{k}_c is the wavevector of the electron in the final conduction band state. For optical excitations, $\mathbf{k}_{\text{photon}} \approx 0$. This implies that the excitation must be *direct*: $\mathbf{k}_v \approx \mathbf{k}_c$. Because of this conservation rule, direct optical excitations are stronger than indirect excitations.

Semiconductors are poor conductors of electricity at low temperatures. Since the valence band is completely occupied, an applied electric field cannot change the total momentum of the valence electrons. This is a reflection of the Pauli principle. This would not be true for an electron that is excited into the conduction band. However, for a band gap of 1 eV or more, few electrons can be thermally excited into the conduction band at ambient temperatures. Conversely, the electronic properties of semiconductors at ambient temperatures can be profoundly altered by the

addition of impurities. In silicon, each atom has four covalent bonds, one to each neighbouring atom. All the valence electrons are consumed in saturating these bonds. If a silicon atom is removed and replaced by an atom with a different number of valence electrons, there will be a mismatch between the number of electrons and the number of covalent bonds. For example, if one replaces a silicon atom by a phosphorous atom, then

there will be an extra electron that cannot be accommodated as phosphorous possesses five instead of four valence electrons. This extra electron is only loosely bound to the phosphorous atom and can be easily excited into the conduction band. Impurities with an ‘extra’ electron are called *donors*. Under the influence of an electric field, this donor electron can contribute to the electrical conductivity of silicon. If one were to replace a silicon atom by a boron atom, the opposite situation would occur. Boron has only three valence electrons and does not possess a sufficient number of electrons to saturate the bonds. In this case, an electron in the valence band can readily move into the unsaturated bond. Under the influence of an electric field, this unsaturated bond can propagate and contribute to the electrical conductivity as if it were a positively charged particle. The unsaturated bond corresponds to a *hole* excitation. Impurity atoms that have less than the number of valence electrons to saturate all the covalent bonds are called *acceptors*.

Several factors determine how efficient impurity atoms will be in altering the electronic properties of a semiconductor. For example, the size of the band gap, the shape of the energy bands near the gap and the ability of the valence electrons to screen the impurity atom are all important. The process of adding controlled impurity atoms to semiconductors is called *doping*. The ability to produce well defined doping levels in semiconductors is one reason for the revolutionary developments in the construction of solid-state electronic devices.

Another useful quantity in defining the electronic structure of a solid is the electronic *density of states*. In general the density of states can be defined as

$$D(E) = \frac{\Omega}{(2\pi)^3} \sum_n \int_{\text{BZ}} \delta(E - E_n(\mathbf{k})) d^3k. \quad (\text{A1.3.84})$$

Unlike the density of states defined in [equation A1.3.24](#), which was specific for the free electron gas, equation A1.3.84 is a general expression. The sum in equation A1.3.84 is over all energy bands and the integral is over all \mathbf{k} -points in the Brillouin zone. The density of states is an extensive function that scales with the size of the sample. It is usually normalized to the number of electronic states per atom. In the case of silicon, the number of states contained by integrating $D(E)$ up to the highest occupied states is four states per atom. Since each state can hold two electrons with different spin coordinates, eight electrons can be accommodated within the valence bands. This corresponds to the number of electrons within the unit cell with the resulting valence bands being fully occupied.

The density of states for crystalline silicon is shown in [figure A1.3.15](#). The density of states is a more general representation of the energetic distribution of electrons than the energy band structure. The distribution of states can be given without regard to the \mathbf{k} wavevector. It is possible to compare the density of states from the energy band structure directly to experimental probes such as those obtained in photoemission. *Photoemission measurements* can be used to measure the distribution of binding electrons within a solid. In these measurements, a photon with a well defined energy impinges on the sample. If the photon carries sufficient energy, an electron can be excited from the valence state to a free electron state. By knowing the energy of the absorbed photon and the emitted electron, it is possible to determine the energy of the electron in the valence state. The number of electrons emitted is proportional to the number of electrons in the initial valence states; the density of states gives a measure of the number of photoemitted electrons for a given binding energy. In realistic calculations of the photoemission spectra, the probability of making a transition from the valence band to the vacuum must be included, but often the transition probabilities are

similar over the entire valence band. This is illustrated in [figure A1.3.15](#). Empty states cannot be measured using photoemission so these contributions are not observed.

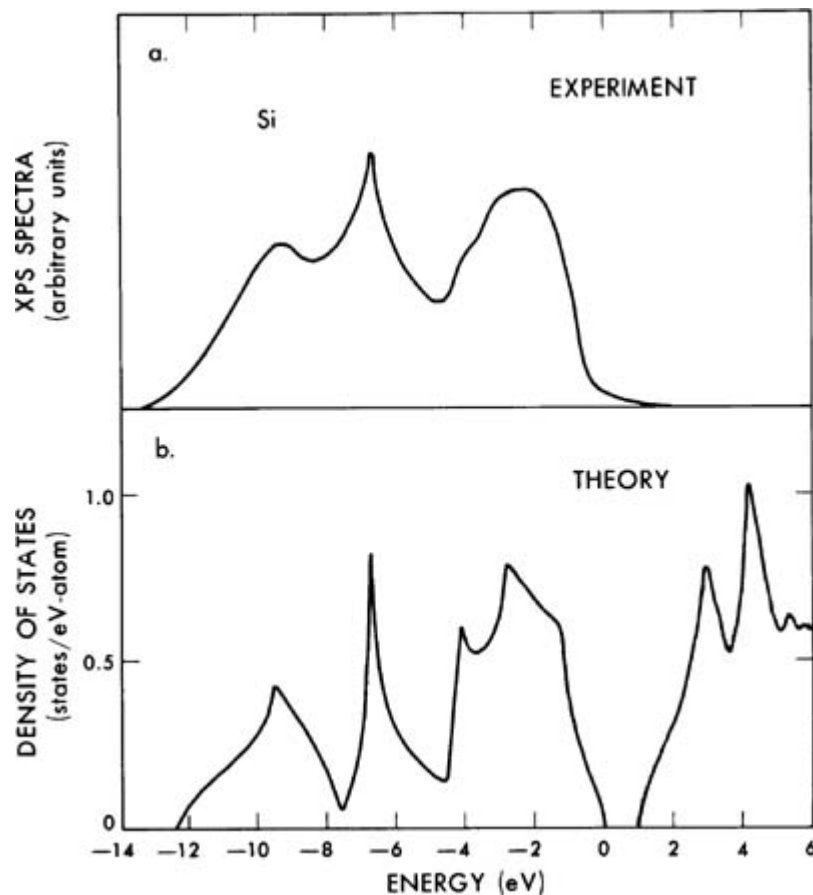


Figure A1.3.15. Density of states for silicon (bottom panel) as calculated from empirical pseudopotential [25]. The top panel represents the photoemission spectra as measured by x-ray photoemission spectroscopy [30]. The density of states is a measure of the photoemission spectra.

By examining the spatial character of the wavefunctions, it is possible to attribute atomic characteristics to the density of states spectrum. For example, the lowest states, 8 to 12 eV below the top of the valence band, are s-like and arise from the atomic 3s states. From 4 to 6 eV below the top of the valence band are states that are also s-like, but change character very rapidly toward the valence band maximum. The states residing within 4 eV of the top of the valence band are p and arise from the 3p states.

A major achievement of the quantum theory of matter has been to explain the interaction of light and matter. For example, the first application of quantum theory, the Bohr model of the atom, accurately predicted the electronic excitations in the hydrogen atom. In atomic systems, the absorption and emission of light is characterized by sharp lines. Predicting the exact frequencies for atomic absorption and emission lines provides a great challenge and testing ground for any theory. This is in apparent contrast to the spectra of solids. The continuum of states in solids, i.e.

energy bands, allows many possible transitions. A photon with energy well above the band gap can excite a number of different states corresponding to different bands and \mathbf{k} -points. The resulting spectra correspond to broad excitation spectra without the sharp structures present in atomic transitions. This is illustrated in figure A1.3.16. The spectrum consists of three broad peaks with the central peak at about 4.5 eV.

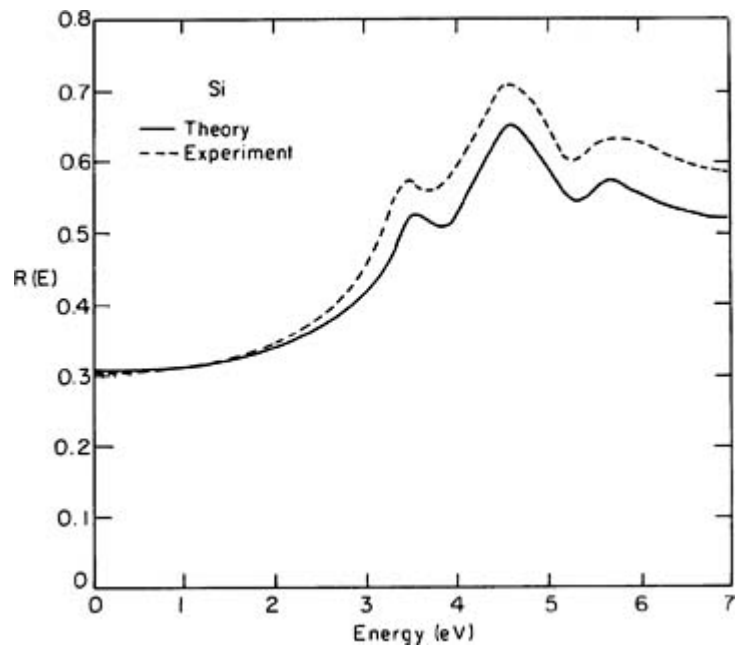


Figure A1.3.16. Reflectivity of silicon. The theoretical curve is from an empirical pseudopotential method calculation [25]. The experimental curve is from [31].

The interpretation of solid-state spectra as featureless and lacking the information content of atomic spectra is misleading. If one modulates the reflectivity spectra of solids, the spectra are quite rich in structure. This is especially the case at low temperatures where vibrational motions of the atoms are reduced. In [figure A1.3.17](#) the spectra of silicon is differentiated. The process of measuring a differentiated spectra is called *modulation spectroscopy*. In modulated reflectivity spectra, broad undulating features are suppressed and sharp features are enhanced. It is possible to modulate the reflectivity spectrum in a variety of ways. For example, one can mechanically vibrate the crystal, apply an alternating electric field or modulate the temperature of the sample. One of the most popular methods is to measure the reflectivity directly and then numerically differentiate the reflectivity data. This procedure has the advantage of being easily interpreted[25].

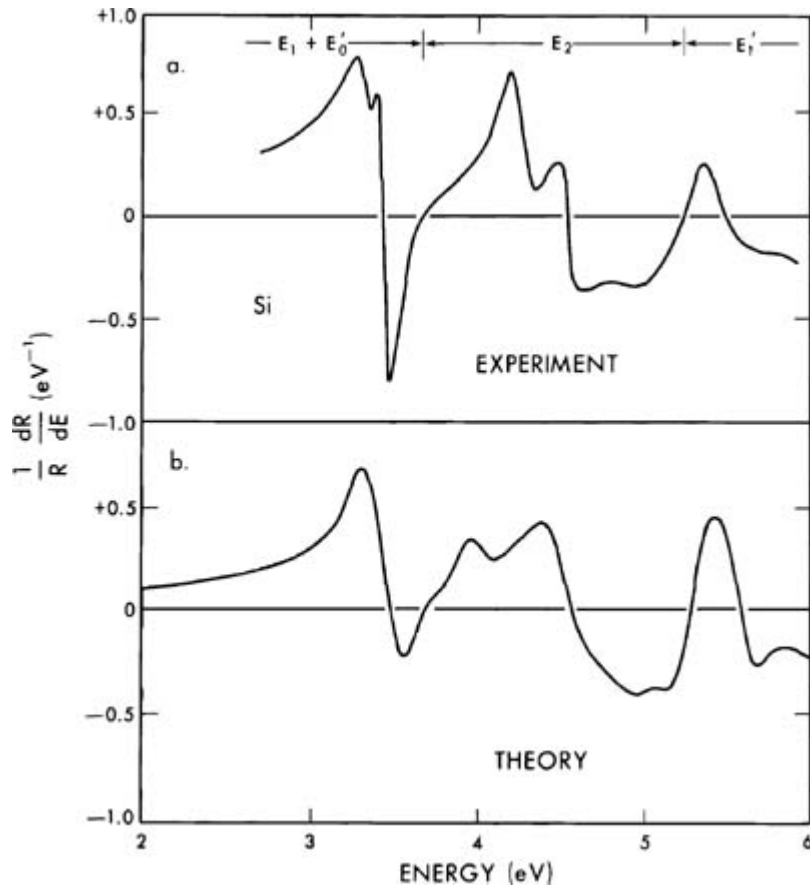


Figure A1.3.17. Modulated reflectivity spectrum of silicon. The theoretical curve is obtained from an empirical pseudopotential calculation [25]. The experimental curve is from a wavelength modulation experiment from [32].

The structure in the reflectivity can be understood in terms of band structure features: i.e. from the quantum states of the crystal. The normal incident reflectivity from matter is given by

$$R = \left(\frac{I}{I_0} \right)^2 = \left| \frac{N - 1}{N + 1} \right|^2 \quad (\text{A1.3.85})$$

where I_0 is the incident intensity of the light and I is the reflected intensity. N is the complex index of refraction. The complex index of refraction, N , can be related to the dielectric function of matter by

$$N^2 = \epsilon_1 + i\epsilon_2 \quad (\text{A1.3.86})$$

where ϵ_1 is the real part of the dielectric function and ϵ_2 is the imaginary part of the dielectric function.

It is possible to make a connection between the quantum states of a solid and the resulting optical properties of a solid.

In contrast to metals, most studies have concentrated on insulators and semiconductors where the optical structure readily lends itself to a straightforward interpretation. Within certain approximations, the imaginary part of the dielectric function for semiconducting or insulating crystals is given by

$$\epsilon_2(\omega) = \frac{4\pi e^2 \hbar}{3m^2 \omega^2} \sum_{vc} \frac{2}{(2\pi)^3} \int_{\text{BZ}} \delta(\omega_{vc}(\mathbf{k}) - \omega) |M_{vc}(\mathbf{k})|^2 d^3k. \quad (\text{A1.3.87})$$

The matrix elements are given by

$$M_{vc}(\mathbf{k}) = \int u_{\mathbf{k},v}^*(\mathbf{r}) \nabla u_{\mathbf{k},c}(\mathbf{r}) d^3r \quad (\text{A1.3.88})$$

where $u_{\mathbf{k},c}$ is the periodic part of the Bloch wavefunction. The summation in equation A1.3.87 is over all occupied to empty state transitions from valence (v) to conduction bands (c). The energy difference between occupied and empty states is given by $E_c(\mathbf{k}) - E_v(\mathbf{k})$ which can be defined as a frequency; $\omega_{vc}(\mathbf{k}) = (E_c(\mathbf{k}) - E_v(\mathbf{k}))/\hbar$. The delta function term, $\delta(\omega_{vc}(\mathbf{k}) - \omega)$, ensures conservation of energy. The matrix elements, M_{vc} , control the oscillator strength. As an example, suppose that the $v \rightarrow c$ transition couples states which have similar parity. The matrix elements will be small because the momentum operator is odd. Although angular momentum is not a good quantum number in condensed matter phases, atomic selection rules remain approximately true.

This expression for ϵ_2 neglects the spatial variation of the perturbing electric field. The wavelength of light for optical excitations is between 4000–7000 Å and greatly exceeds a typical bond length of 1–2 Å. Thus, the assumption of a uniform field is usually a good approximation. Other effects ignored include many-body contributions such as correlation and electron–hole interactions.

Once the imaginary part of the dielectric function is known, the real part can be obtained from the Kramers–Kronig relation:

$$\epsilon_1(\omega) = 1 + \frac{2}{\pi} P \int_0^\infty \frac{\omega' \epsilon_2(\omega')}{\omega'^2 - \omega^2} d\omega'. \quad (\text{A1.3.89})$$

The principal part of the integral is taken and the integration must be done over all frequencies. In practice, the integration is often terminated outside of the frequency range of interest. Once the full dielectric function is known, the reflectivity of the solid can be computed.

It is possible to understand the fine structure in the reflectivity spectrum by examining the contributions to the imaginary part of the dielectric function. If one considers transitions from two bands ($v \rightarrow c$), equation A1.3.87 can be written as

$$\epsilon_2(\omega)_{vc} = \frac{4\pi e^2 \hbar}{3m^2 \omega^2} \frac{2}{(2\pi)^3} |M_{vc}| \int_{\text{BZ}} \delta(\omega_{vc}(\mathbf{k}) - \omega) d^3k. \quad (\text{A1.3.90})$$

Under the assumption that the matrix elements can be treated as constants, they can be factored out of the integral. This is a good approximation for most crystals. By comparison with [equation A1.3.84](#), it is possible to define a function similar to the density of states. In this case, since both valence and conduction band states are included, the function is called the *joint density of states*:

$$J_{vc}(\omega) = \frac{2}{(2\pi)^3} \int_{\text{BZ}} \delta(\omega_{vc}(\mathbf{k}) - \omega) d^3k. \quad (\text{A1.3.91})$$

With this definition, one can write

$$\epsilon_2(\omega)_{vc} = \frac{4\pi e^2 \hbar}{3m^2 \omega^2} |M_{vc}| J_{vc}(\omega). \quad (\text{A1.3.92})$$

Within this approximation, the structure in $\epsilon_2(\omega)_{vc}$ can be related to structure in the joint density of states. The joint density of states can be written as a surface integral [1]:

$$J_{vc}(\omega) = \frac{2}{(2\pi)^3} \int_{\omega_{vc}=\omega} \frac{ds}{|\nabla_{\mathbf{k}} \omega_{vc}(\mathbf{k})|}. \quad (\text{A1.3.93})$$

ds is a surface element defined by $\omega_{vc}(\mathbf{k}) = \omega$. The sharp structure in the joint density of states arises from zeros in the dominator. This occurs at *critical points* where

$$\nabla_{\mathbf{k}} \omega_{vc}(\mathbf{k}) = 0 \quad (\text{A1.3.94})$$

or

$$\nabla_{\mathbf{k}} E_v(\mathbf{k}) = \nabla_{\mathbf{k}} E_c(\mathbf{k}) \quad (\text{A1.3.95})$$

when the slopes of the valence band and conduction band are equal. The group velocity of an electron or hole is defined as $\mathbf{v}_g = \nabla_{\mathbf{k}} E(\mathbf{k})$. Thus, the critical points occur when the hole and electrons have the same group velocity.

The band energy difference or $\omega_{vc}(\mathbf{k})$ can be expanded around a critical point \mathbf{k}_{cp} as

$$\omega_{vc}(\mathbf{k}) = \omega_{vc}(\mathbf{k}_{cp}) + \sum_{n=1}^3 \alpha_n (\mathbf{k} - \mathbf{k}_{cp})_n^2 + \dots \quad (\text{A1.3.96})$$

The expansion is done around the principal axes so only three terms occur in the summation. The nature of the critical point is determined by the signs of the α_n . If $\alpha_n > 0$ for all n , then the critical point corresponds to a local minimum. If $\alpha_n < 0$ for all n , then the critical point corresponds to a local maximum. Otherwise, the critical points correspond to saddle points.

The types of critical points can be labelled by the number of α_n less than zero. Specifically, the critical points are labelled by M_i where i is the number of α_n which are negative: i.e. a local minimum critical point would be labelled by M_0 , a local maximum by M_3 and the saddle points by (M_1, M_2) . Each critical point has a characteristic line shape. For example, the M_0 critical point has a joint density of state which behaves as $J_{vc} = \text{constant} \times \sqrt{\omega - \omega_0}$ for $\omega > \omega_0$ and zero otherwise, where ω_0 corresponds to the M_0 critical point energy. At $\omega = \omega_0$, J_{vc} has a discontinuity in the first derivative. In figure A1.3.18 the characteristic structure of the joint density of states is presented for each type of critical point.

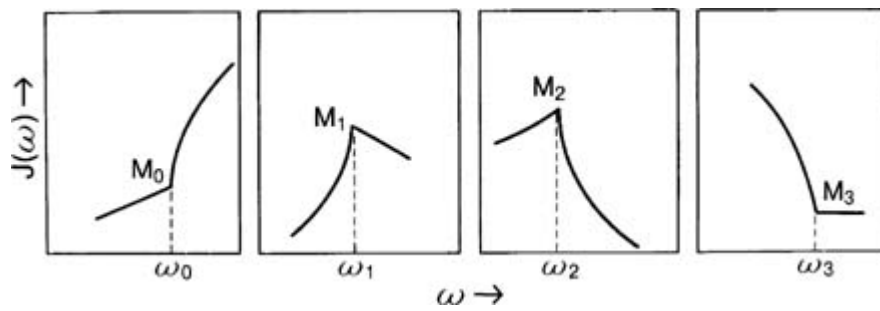


Figure A1.3.18. Typical critical point structure in the joint density of states.

For a given pair of valence and conduction bands, there must be at least one M_0 and one M_3 critical points and at least three M_1 and three M_2 critical points. However, it is possible for the saddle critical points to be degenerate. In the simplest possible configuration of critical points, the joint density of states appears as in figure A1.3.19.

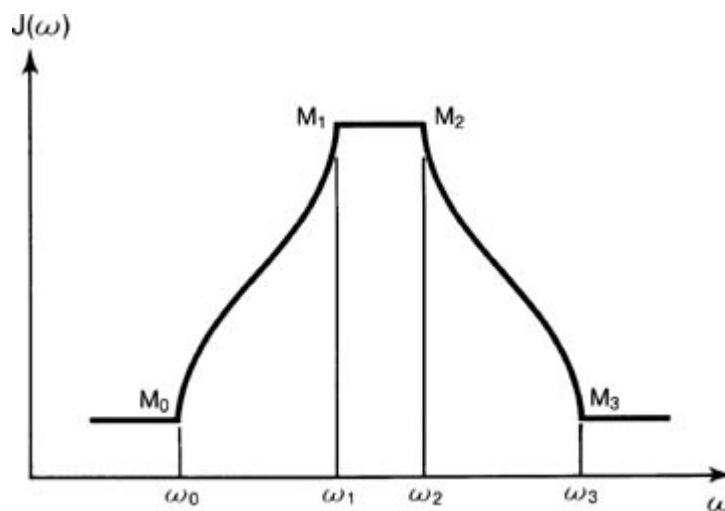


Figure A1.3.19. Simplest possible critical point structure in the joint density of states for a given energy band.

It is possible to identify particular spectral features in the modulated reflectivity spectra to band structure features. For example, in a direct band gap the joint density of states must resemble that of a M_0 critical point. One of the first applications of the *empirical pseudopotential method* was to calculate reflectivity spectra for a given energy band. Differences between the calculated and measured reflectivity spectra could be assigned to errors in the energy band

structure. Such errors usually involve the incorrect placement or energy of a critical point feature. By making small adjustments in the pseudopotential, it is almost always possible to extract an energy band structure consistent with the measure reflectivity.

The critical point analysis performed for the joint density of states can also be applied to the density of states. By examining the photoemission spectrum compared with the calculated density of states, it is also possible to assess the quality of the energy band structure. Photoemission spectra are superior to reflectivity spectra in the sense of giving the band structure energies relative to a fixed energy reference, such as the vacuum level. Reflectivity measurements only give relative energy differences between energy bands.

In figure A1.3.20 and [figure A1.3.21](#) the real and imaginary parts of the dielectric function are illustrated for

silicon. There are some noticeable differences in the line shapes between theory and experiment. These differences can be attributed to issues outside of elementary band theory such as the interactions of electrons and holes. This issue will be discussed further in the following section on insulators. Qualitatively, the real part of the dielectric function appears as a simple harmonic oscillator with a resonance at about 4.5 eV. This energy corresponds approximately to the cohesive energy per atom of silicon.

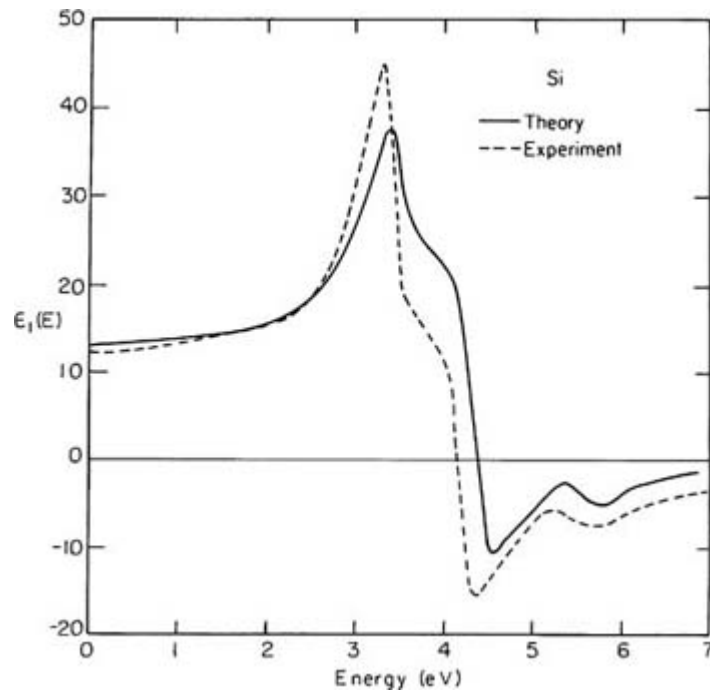


Figure A1.3.20. Real part of the dielectric function for silicon. The experimental work is from [31]. The theoretical work is from an empirical pseudopotential calculation [25].

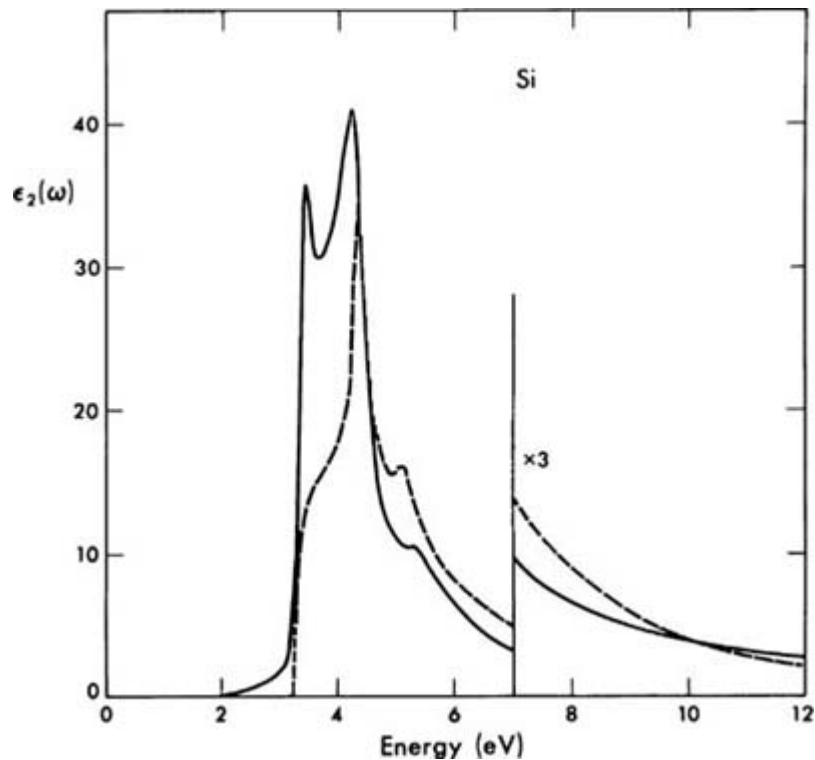


Figure A1.3.21. Imaginary part of the dielectric function for silicon. The experimental work is from [31]. The theoretical work is from an empirical pseudopotential calculation [25].

It is possible to determine the spatial distributions, or charge densities, of electrons from a knowledge of the wavefunctions. The arrangement of the charge density is very useful in characterizing the bond in the solid. For example, if the charge is highly localized between neighbouring atoms, then the bond corresponds to a covalent bond. The classical picture of the covalent bond is the sharing of electrons between two atoms. This picture is supported by quantum calculations. In [figure A1.3.22](#) the electronic distribution charge is illustrated for crystalline carbon and silicon in the diamond structure. In carbon the midpoint between neighbouring atoms is a saddle point: this is typical of the covalent bond in organics, but not in silicon where the midpoint corresponds to a maximum of the charge of the density. X-ray measurements also support the existence of the covalent bonding charge as determined from quantum calculations [33].

-42-

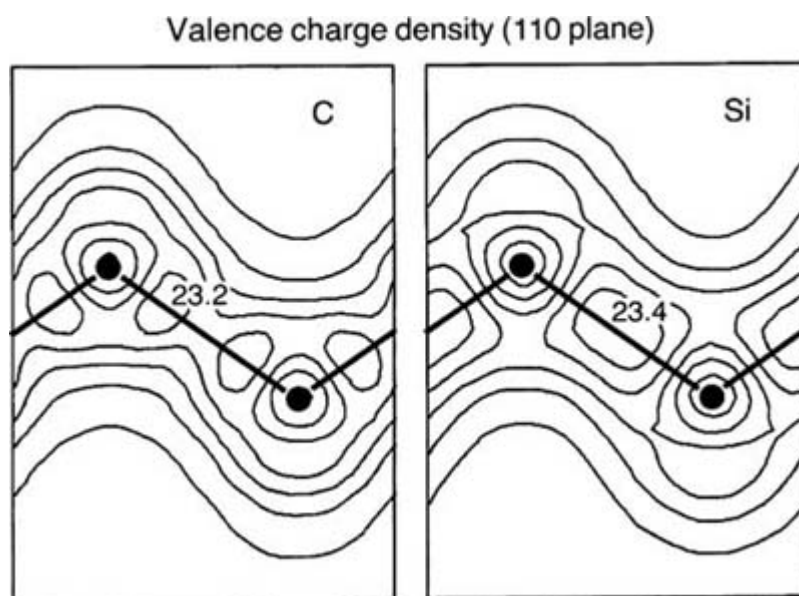


Figure A1.3.22. Spatial distributions or charge densities for carbon and silicon crystals in the diamond structure. The density is only for the valence electrons; the core electrons are omitted. This charge density is from an *ab initio* pseudopotential calculation [27].

Although empirical pseudopotentials present a reliable picture of the electronic structure of semiconductors, these potentials are not applicable for understanding structural properties. However, density-functional-derived pseudopotentials can be used for examining the structural properties of matter. Once a self-consistent field solution of the Kohn–Sham equations has been achieved, the total electronic energy of the system can be determined from [equation A1.3.39](#). One of the first applications of this method was to forms of crystalline silicon. Various structural forms of silicon were considered: diamond, hexagonal diamond, β -Sn, simple cubic, FCC, BCC and so on. For a given volume, the lattice parameters and any internal parameters can be optimized to achieve a ground-state energy. In [figure A1.3.23](#) the total structural energy of the system is plotted for eight different forms of silicon. The lowest energy form of silicon is correctly predicted to be the diamond structure. By examining the change in the structural energy with respect to volume, it is possible to determine the equation of state for each form. It is possible to determine which phase is lowest in energy for a specified volume and to determine transition pressures between different phases. As an example, one can predict from this phase diagram the transition pressure to transform silicon in the diamond structure to the white tin (β -Sn) structure. This pressure is predicted to be approximately 90 MPa; the measured pressure is about 120 MPa [34]. The role of temperature has been neglected in the calculation of the structural energies. For most applications, this is not a serious issue as the role of temperature is often less than the inherent errors

within density functional theory.

-43-

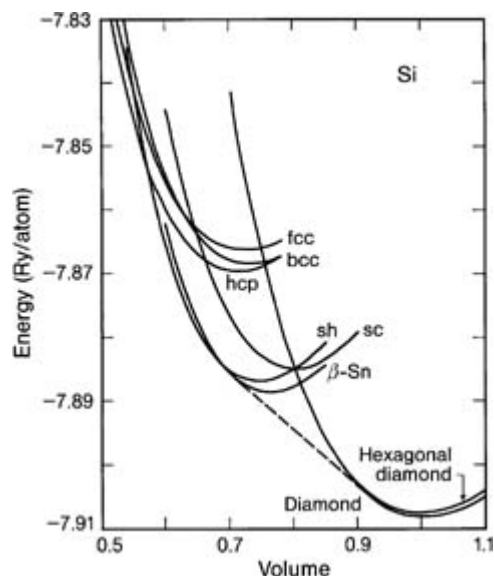


Figure A1.3.23. Phase diagram of silicon in various polymorphs from an *ab initio* pseudopotential calculation [34]. The volume is normalized to the experimental volume. The binding energy is the total electronic energy of the valence electrons. The slope of the dashed curve gives the pressure to transform silicon in the diamond structure to the β -Sn structure. Other polymorphs listed include face-centred cubic (fcc), body-centred cubic (bcc), simple hexagonal (sh), simple cubic (sc) and hexagonal close-packed (hcp) structures.

One notable consequence of the phase diagram in figure A1.3.23 was the prediction that high-pressure forms of silicon might be superconducting [35, 36]. This prediction was based on the observation that some high-pressure forms of silicon are metallic, but retain strong covalent-like bonds. It was later verified by high-pressure measurements that the predicted phase was a superconductor [36]. This success of the structural phase diagram of silicon helped verify the utility of the pseudopotential density functional method and has resulted in its widespread applicability to condensed phases.

A1.3.6.2 INSULATORS

Insulating solids have band gaps which are notably larger than semiconductors. It is not unusual for an alkali halide to have a band gap of ~ 10 eV or more. Electronic states in insulators are often highly localized around the atomic sites in insulating materials. In most cases, this arises from a large transfer of electrons from one site to another. Exceptions are insulating materials like sulfur and carbon where the covalent bonds are so strong as to strongly localize charge between neighbouring atoms.

As an example of the energy band structures for an insulator, the energy bands for lithium fluoride are presented in figure A1.3.24. LiF is a highly ionic material which forms in the rocksalt structure (figure A1.3.25)). The bonding in this crystal can be understood by transferring an electron from the highly electropositive Li to the electronegative F atoms: i.e. one can view crystalline LiF as consisting of Li^+F^- constituents. The highly localized nature of the electronic charge density results in very narrow, almost atomic-like, energy bands.

-44-

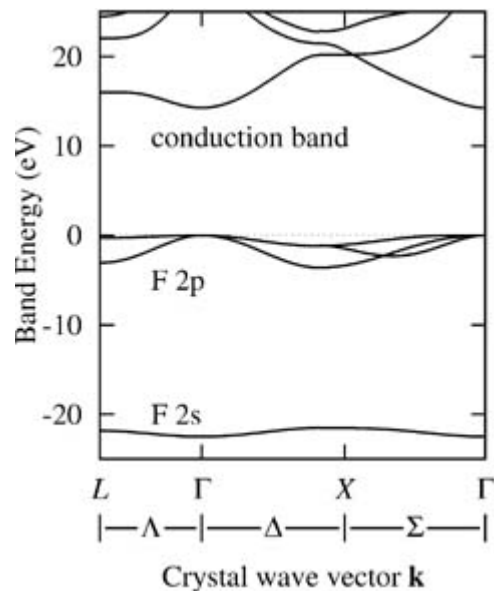


Figure A1.3.24. Band structure of LiF from *ab initio* pseudopotentials [39].

One challenge of modern electronic structure calculations has been to reproduce excited-state properties. Density functional theory is a ground-state theory. The eigenvalues for empty states do not have physical meaning in terms of giving excitation energies. If one were to estimate the band gap from density functional theory by taking the eigenvalue differences between the highest occupied and lowest empty states, the energy difference would badly underestimate the band gap. Contemporary approaches [37, 38] have resolved this issue by correctly including spatial variations in the electron–electron interactions and including self-energy terms (see [section A1.3.2.2](#)).

Because of the highly localized nature of electronic and hole states in insulators, it is difficult to describe the optical excitations. The excited electron is strongly affected by the presence of the hole state. One failure of the energy band picture concerns the interaction between the electron and hole. The excited electron and the hole can form a hydrogen atomic-like interaction resulting in the formation of an *exciton*, or a bound electron–hole pair. The exciton binding energy reduces the energy for an excitation below that of the conduction band and results in strong, discrete optical lines. This is illustrated in [figure A1.3.25](#).

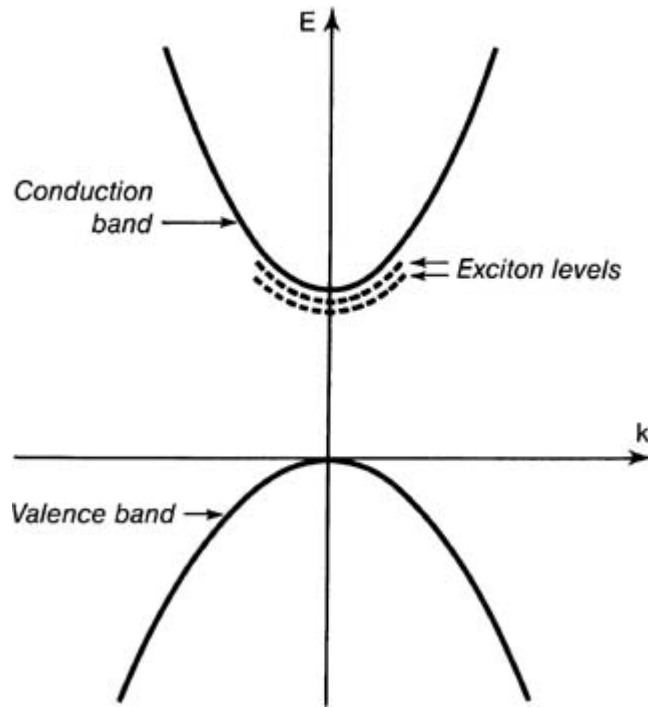


Figure A1.3.25. Schematic illustration of exciton binding energies in an insulator or semiconductor.

A simple model for the exciton is to assume a screened interaction between the electron and hole using a static dielectric function. In addition, it is common to treat the many-body interactions in a crystal by replacing the true mass of the electron and hole by a dynamical or *effective mass*. Unlike a hydrogen atom, where the proton mass exceeds that of the electron by three orders of magnitude, the masses of the interacting electron and hole are almost equivalent. Using the reduced mass for this system, we have $1/\mu = 1/m_e + 1/m_h$. Within this model, the binding energy of the exciton can be found from

$$\left[\frac{-\hbar^2 \nabla^2}{2\mu} - \frac{e^2}{\epsilon r} \right] \psi(\mathbf{r}) = E_b \psi(\mathbf{r}) \quad (\text{A1.3.97})$$

where ϵ is the static dielectric function for the insulator of interest. The binding energy from this hydrogenic Schrödinger equation is given by

$$E_b = -\frac{\mu e^4}{2\hbar^2 \epsilon^2 n^2} \quad (\text{A1.3.98})$$

where $n = 1, 2, 3, \dots$. Typical values for a semiconductor are μ and ϵ are $\mu = 0.1 m$ and $\epsilon = 10$. This results in a binding energy of about 0.01 eV for the ground state, $n = 1$. For an insulator, the binding energy is much larger. For a material like silicon dioxide, one might have $\mu = 0.5 m$ and $\epsilon = 3$ or a binding energy of roughly 1 eV. This estimate suggests that reflectivity spectra in insulators might be strongly altered by exciton interactions.

Even in semiconductors, where it might appear that the exciton binding energies would be of interest only for low temperature regimes, excitonic effects can strongly alter the line shape of excitations away from the band gap.

The size of the electron-hole pair can be estimated from the Bohr radius for this system:

$$r_0 = -\frac{\epsilon\hbar^2}{\mu e^2}. \quad (\text{A1.3.99})$$

The size of the exciton is approximately 50 Å in a material like silicon, whereas for an insulator the size would be much smaller: for example, using our numbers above for silicon dioxide, one would obtain a radius of only ~3 Å or less. For excitons of this size, it becomes problematic to incorporate a static dielectric constant based on macroscopic crystalline values.

The reflectivity of LiF is illustrated in figure A1.3.26. The first large peak corresponds to an excitonic transition.

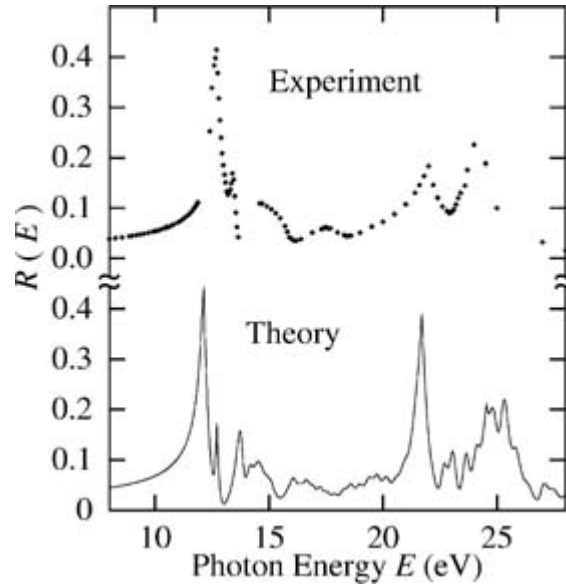


Figure A1.3.26. Reflectivity of LiF from *ab initio* pseudopotentials. (Courtesy of E L Shirley, see [39] and references therein.)

A1.3.6.3 METALS

Metals are fundamentally different from insulators as they possess no gap in the excitation spectra. Under the influence of an external field, electrons can respond by readily changing from one \mathbf{k} state to another. The ease by which the ground-state configuration is changed accounts for the high conductivity of metals.

Arguments based on a free electron model can be made to explain the conductivity of a metal. It can be shown that the \mathbf{k} will evolve following a Newtonian law [1]:

$$\hbar \frac{d\mathbf{k}}{dt} = -e\mathcal{E}. \quad (\text{A1.3.100})$$

This can be integrated to yield

$$(\mathbf{k} - \mathbf{k}_0) = -e\mathcal{E}(t - t_0)/\hbar. \quad (\text{A1.3.101})$$

After some typical time, τ , the electron will scatter off a lattice imperfection. This imperfection might be a lattice vibration or an impurity atom. If one assumes that no memory of the event resides after the scattering event, then on average one has $\Delta \mathbf{k} = -e\mathcal{E}\tau/\hbar$. In this picture, the conductivity of the metal, σ , can be extracted from Ohm's law: $\sigma = J/\mathcal{E}$ where J is the current density. The current density is given by

$$\mathbf{J} = -ne\Delta\mathbf{v} = -ne(-e\mathcal{E}\tau/m) = ne^2\tau\mathcal{E}/m \quad (\text{A1.3.102})$$

or

$$\sigma = ne^2\tau/m. \quad (\text{A1.3.103})$$

This expression for the conductivity is consistent with experimental trends.

Another important accomplishment of the free electron model concerns the heat capacity of a metal. At low temperatures, the heat capacity of a metal goes linearly with the temperature and vanishes at absolute zero. This behaviour is in contrast with classical statistical mechanics. According to classical theories, the equipartition theory predicts that a free particle should have a heat capacity of $\frac{3}{2}k_B$ where k_B is the Boltzmann constant. An ideal gas has a heat capacity consistent with this value. The electrical conductivity of a metal suggests that the conduction electrons behave like 'free particles' and might also have a heat capacity of $\frac{3}{2}k_B$, which would be strongly at variance with the observed behaviour and in violation of the third law of thermodynamics.

The resolution of this issue is based on the application of the Pauli exclusion principle and Fermi–Dirac statistics. From the free electron model, the total electronic energy, U , can be written as

$$U(T) = \int_0^\infty \epsilon f(\epsilon, T) D(\epsilon) d\epsilon \quad (\text{A1.3.104})$$

where $f(\epsilon, T)$ is the Fermi–Dirac distribution function and $D(\epsilon)$ is the density of states. The Fermi–Dirac function gives the probability that a given orbital will be occupied:

$$f(\epsilon, T) = \frac{1}{\exp((\epsilon - E_F)/kT) + 1}. \quad (\text{A1.3.105})$$

-48-

The value of E_F at zero temperature can be estimated from the electron density (equation A1.3.26). Typical values of the Fermi energy range from about 1.6 eV for Cs to 14.1 eV for Be. In terms of temperature ($T_F = E_F/k$), the range is approximately 2000–16,000 K. As a consequence, the Fermi energy is a very weak function of temperature under ambient conditions. The electronic contribution to the heat capacity, C , can be determined from

$$C = \frac{dU}{dT} = \int_0^\infty \epsilon D(\epsilon) \frac{df(\epsilon, T)}{dT} d\epsilon. \quad (\text{A1.3.106})$$

The integral can be approximated by noting that the derivative of the Fermi function is highly localized around E_F . To a very good approximation, the heat capacity is

$$C = \frac{\pi^2}{3} D(\epsilon_F) k^2 T. \quad (\text{A1.3.107})$$

The linear dependence of C with temperature agrees well with experiment, but the pre-factor can differ by a factor of two or more from the free electron value. The origin of the difference is thought to arise from several factors: the electrons are not truly free, they interact with each other and with the crystal lattice, and the dynamical behaviour the electrons interacting with the lattice results in an *effective mass* which differs from the free electron mass. For example, as the electron moves through the lattice, the lattice can distort and exert a dragging force.

Simple metals like alkalis, or ones with only s and p valence electrons, can often be described by a free electron gas model, whereas transition metals and rare earth metals which have d and f valence electrons cannot. Transition metal and rare earth metals do not have energy band structures which resemble free electron models. The formed bonds from d and f states often have some strong covalent character. This character strongly modulates the free-electron-like bands.

An example of metal with significant d-bonding is copper. The atomic configuration of copper is $1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^1$. If the 3d states were truly core states, then one might expect copper to resemble potassium as its atomic configuration is $1s^2 2s^2 2p^6 3s^2 3p^6 4s^1$. The strong differences between copper and potassium in terms of their chemical properties suggest that the 3d states interact strongly with the valence electrons. This is reflected in the energy band structure of copper (figure A1.3.27).

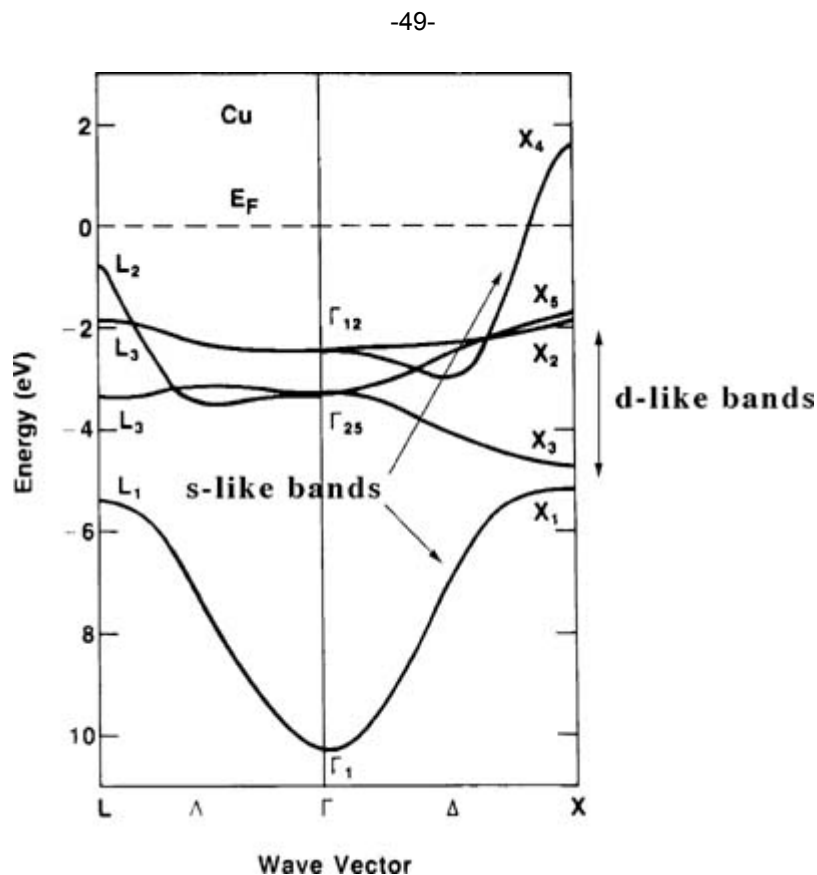


Figure A1.3.27. Energy bands of copper from *ab initio* pseudopotential calculations [40].

Copper has a FCC structure with one atom in the primitive unit cell. From simple orbital counting, one might expect the ten d electrons to occupy five d-like bands and the one s electron to occupy one s-like band. This is apparent in the figure, although the interpretation is not straightforward. The lowest band (L_1 to Γ_1 to X_1) is

s-like, but it mixes strongly with the d-like bands (at Γ_{25} and Γ_{12}), these bands are triply and doubly degenerate at Γ . Were it not for the d-mixing, the s-like band would be continuous from Γ_1 to X_4 . The d-mixing 'splits' the s bands. The Fermi level cuts the s-like band along the Δ direction, reflecting the partial occupation of the s levels.

A1.3.7 NON-CRYSTALLINE MATTER

A1.3.7.1 AMORPHOUS SOLIDS

Crystalline matter can be characterized by *long-range order*. For a perfect crystal, a prescription can be used to generate the positions of atoms arbitrarily far away from a specified origin. However, 'real crystals' always contain imperfections. They contain defects which can be characterized as *point defects* localized to an atomic site or *extended defects* spread over a number of sites. Vacancies on the lattice site or atoms of impurities are examples of point defects. Grain boundaries or dislocations are examples of extended defects. One might imagine starting from an ideal

-50-

crystal and gradually introducing defects such as vacancies. At some point the number of defects will be so large as to ruin the long-range order of the crystal. Solid materials that lack long-range order are called *amorphous* solids or *glasses*. The precise definition of an amorphous material is somewhat problematic. Usually, any material which does not display a sharp x-ray pattern is considered to be 'amorphous'. Some text books [1] define amorphous solids as 'not crystalline on any significant scale'.

Glassy materials are usually characterized by an additional criterion. It is often possible to cool a liquid below the thermodynamic melting point (i.e. to supercool the liquid). In glasses, as one cools the liquid state significantly below the melting point, it is observed that at a temperature well below the melting point of the solid the viscosity of the supercooled liquid increases dramatically. This temperature is called the *glass transition* temperature, and labelled as T_g . This increase of viscosity delineates the supercooled liquid state from the glass state. Unlike thermodynamic transitions between the liquid and solid state, the liquid \rightarrow glass transition is not well defined. Most amorphous materials such as tetrahedrally coordinated semiconductors like silicon and germanium do not exhibit a glass transformation.

Defining order in an amorphous solid is problematic at best. There are several 'qualitative concepts' that can be used to describe disorder [7]. In figure A1.3.28 a perfect crystal is illustrated. A simple form of disorder involves crystals containing more than one type of atom. Suppose one considers an alloy consisting of two different atoms (A and B). In an ordered crystal one might consider each A surrounded by B and *vice versa*. In a random alloy, one might consider the lattice sites to remain unaltered but randomly place A and B atoms. This type of disorder is called *compositional disorder*. Other forms of disorder may involve minor distortions of the lattice that destroy the long-range order of the solid, but retain the chemical ordering and short-range order of the solid. For example, in short-range ordered solids, the coordination number of each atom might be preserved. In a highly disordered solid, no short-range order is retained: the chemical ordering is random with a number of over- and under-coordinated species.

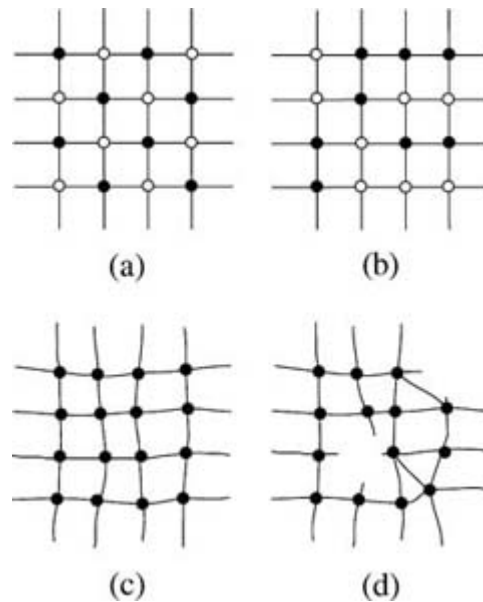


Figure A1.3.28. Examples of disorder: (a) perfect crystal, (b) compositional disorder, (c) positional disorder which retains the short-range order and (d) no long-range or short-range order.

-51-

In general, it is difficult to quantify structural properties of disordered matter via experimental probes as with x-ray or neutron scattering. Such probes measure statistically averaged properties like the *pair-correlation function*, also called the *radial distribution function*. The pair-correlation function measures the average distribution of atoms from a particular site.

Several models have been proposed to describe amorphous solids, in particular glasses. The structure of glasses often focus on two archetypes [1]: the continuous random network and microcrystallite models. In the continuous random network model, short-range order is preserved. For example, in forms of crystalline silica each silicon atom is surrounded by four oxygen atoms. The SiO_4 tetrahedra are linked together in a regular way which establishes the crystal structure. In a continuous random network each SiO_4 tetrahedral unit is preserved, but the relative arrangement between tetrahedral units is random. In another model, the so-called microcrystallite model, small ‘crystallites’ of the perfect structure exist, but these crystallites are randomly arranged. The difference between the random network model and the crystallite model cannot be experimentally determined unless the crystallites are sufficiently large to be detected; this is usually not the situation.

Amorphous materials exhibit special quantum properties with respect to their electronic states. The loss of periodicity renders Bloch’s theorem invalid; \mathbf{k} is no longer a good quantum number. In crystals, structural features in the reflectivity can be associated with critical points in the joint density of states. Since amorphous materials cannot be described by \mathbf{k} -states, selection rules associated with \mathbf{k} are no longer appropriate. Reflectivity spectra and associated spectra are often featureless, or they may correspond to highly smoothed versions of the crystalline spectra.

One might suppose that optical gaps would not exist in amorphous solids, as the structural disorder would result in allowed energy states throughout the solid. However, this is not the case, as disordered insulating solids such as silica are quite transparent. This situation reflects the importance of local order in determining gaps in the excitation spectra. It is still possible to have gaps in the joint density of states without resort to a description of energy *versus* wavevector. For example, in silica the large energy gap arises from the existence of SiO_4 units. Disordering these units can cause states near the top of the occupied states and near the bottom of the empty states to tail into the gap region, but not remove the gap itself.

Disorder plays an important role in determining the extent of electronic states. In crystalline matter one can view states as existing throughout the crystal. For disordered matter, this is not the case: electronic states become localized near band edges. The effect of localization has profound effects on transport properties. Electrons and holes can still carry current in amorphous semiconductors, but the carriers can be strongly scattered by the disordered structure. For the localized states near the band edges, electrons can be propagated only by a thermally activated hopping process.

A1.3.7.2 LIQUIDS

Unlike the solid state, the liquid state cannot be characterized by a static description. In a liquid, bonds break and reform continuously as a function of time. The quantum states in the liquid are similar to those in amorphous solids in the sense that the system is also disordered. The liquid state can be quantified only by considering some ensemble averaging and using statistical measures. For example, consider an elemental liquid. Just as for amorphous solids, one can ask what is the distribution of atoms at a given distance from a reference atom on average, i.e. the *radial distribution function* or the *pair correlation function* can also be defined for a liquid. In scattering experiments on liquids, a *structure factor* is measured. The radial distribution function, $g(r)$, is related to the structure factor, $S(q)$, by

-52-

$$S(q) = 1 + \rho_0 \int [g(r) - 1] \exp(i\mathbf{q} \cdot \mathbf{r}) d^3r \quad (\text{A1.3.108})$$

where ρ_0 is the average concentration density of the liquid. By taking the Fourier transform of the structure, it is possible to determine the radial distribution function of the liquid.

Typical results for a semiconducting liquid are illustrated in figure A1.3.29 where the experimental pair correlation and structure factors for silicon are presented. The radial distribution function shows a sharp first peak followed by oscillations. The structure in the radial distribution function reflects some local ordering. The nature and degree of this order depends on the chemical nature of the liquid state. For example, semiconductor liquids are especially interesting in this sense as they are believed to retain covalent bonding characteristics even in the melt.

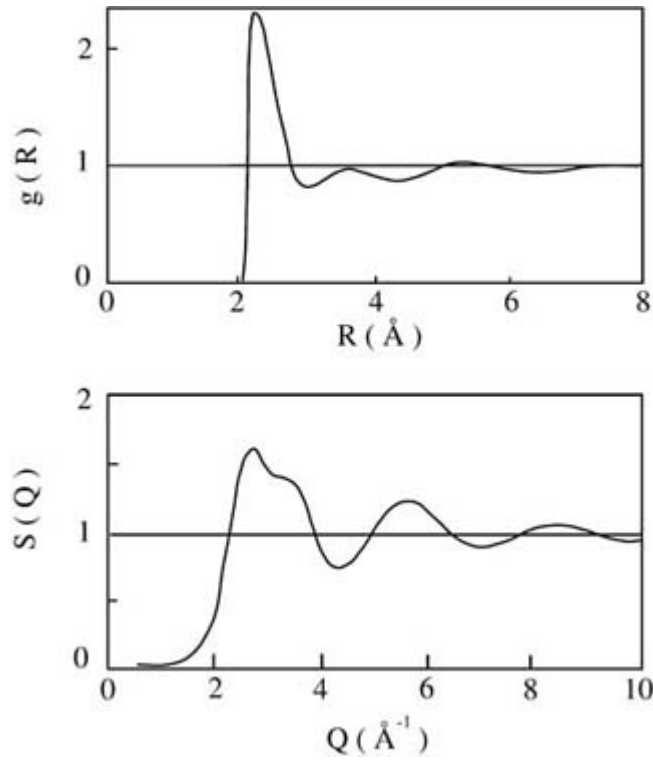


Figure A1.3.29. Pair correlation and structure factor for liquid silicon from experiment [41].

One simple measure of the liquid structure is the average coordination number of an atom. For example, the average coordination of a silicon atom is four in the *solid phase* at ambient pressure and increases to six in high pressure forms of silicon. In the *liquid state*, the average coordination of silicon is six. The average coordination of the liquid can be determined from the radial distribution function. One common prescription is to integrate the area under the first peak of the radial distribution function. The integration is terminated at the first local minimum after the first peak. For a crystalline case, this procedure gives the exact number of nearest neighbours. In general, coordination numbers greater than four correspond to metallic states of silicon. As such, the radial distribution function suggests that silicon is a metal in the liquid state. This is consistent with experimental values of the conductivity. Most tetrahedrally coordinated

-53-

semiconductors, e.g. Ge, GaAs, InP and so on, become metallic upon melting.

It is possible to use the quantum states to predict the electronic properties of the melt. A typical procedure is to implement molecular dynamics simulations for the liquid, which permit the wavefunctions to be determined at each time step of the simulation. As an example, one can use the eigenpairs for a given atomic configuration to calculate the optical conductivity. The real part of the conductivity can be expressed as

$$\sigma_r(\omega) = \frac{2\pi e^2}{3m^2\omega\Omega} \sum_{n,m} \sum_{\alpha=x,y,z} |\langle \psi_m | p_\alpha | \psi_n \rangle|^2 \times \delta(E_n - E_m - \hbar\omega) \quad (\text{A1.3.109})$$

where E_i and ψ_i are eigenvalues and eigenfunctions, and Ω is the volume of the supercell. The dipole transition elements, $\langle \psi_m | p_\alpha | \psi_n \rangle$, reflect the spatial resolution of the initial and final wavefunctions. If the initial and final states were to have an even parity, then the electromagnetic field would not couple to these states.

The conductivity can be calculated for each time step in a simulation and averaged over a long simulation time. This procedure can be used to distinguish the metallic and semiconducting behaviour of the liquid state. As an example, the calculated frequency dependence of the electrical conductivity of gallium arsenide and cadmium telluride are illustrated in figure A1.3.30. In the melt, gallium arsenide is a metal. As the temperature of the liquid is increased, its DC conductivity decreases. For cadmium telluride, the situation is reversed. As the temperature of the liquid is increased, the DC conductivity increases. This is similar to the behaviour of a semiconducting solid. As the temperature of the solid is increased, more carriers are thermally excited into the conduction bands and the conductivity increases. The relative conductivity of GaAs *versus* CdTe as determined via theoretical calculations agrees well with experiment.

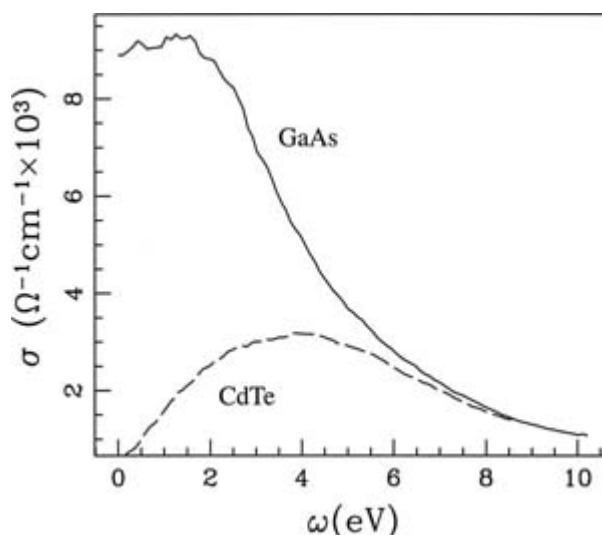


Figure A1.3.30. Theoretical frequency-dependent conductivity for GaAs and CdTe liquids from *ab initio* molecular dynamics simulations [42].

REFERENCES

- [1] Kittel C 1996 *Introduction to Solid State Physics* 7th edn (New York: Wiley)
- [2] Ziman J M 1986 *Principles of the Theory of Solids* 2nd edn (Cambridge: Cambridge University Press)
- [3] Kittel C 1987 *Quantum Theory of Solids* 2nd revn (New York: Wiley)
- [4] Callaway J 1974 *Quantum Theory of the Solid State* (Boston: Academic)
- [5] Yu P and Cardona M 1996 *Fundamentals of Semiconductors* (New York: Springer)
- [6] Wells A F 1984 *Structural Inorganic Chemistry* 5th edn (Oxford: Clarendon)
- [7] Madelung O 1996 *Introduction to Solid State Theory* (New York: Springer)
- [8] Tauc J (ed) 1974 *Amorphous and Liquid Semiconductors* (New York: Plenum)
- [9] Phillips J C 1989 *Physics of High- T_C Superconductors* (Boston: Academic)
- [10] Anderson P W 1997 *The Theory of Superconductivity in the High- T_C Cuprates (Princeton Series in Physics)* (Princeton: Princeton University Press)

- [11] Ziman J M 1960 *Electrons and Phonons* (Oxford: Oxford University Press)
 - [12] Haug A 1972 *Theoretical Solid State Physics* (New York: Pergamon)
 - [13] Thomas L H 1926 *Proc. Camb. Phil. Soc.* **23** 542
 - [14] Fermi E 1928 *Z. Phys.* **48** 73
 - [15] Lundqvist S and March N H (eds) 1983 *Theory of the Inhomogeneous Electron Gas* (New York: Plenum)
 - [16] Ashcroft N W and Mermin N D 1976 *Solid State Physics* (New York: Holt, Rinehart and Winston)
 - [17] Slater J C 1951 *Phys. Rev.* **81** 385
 - [18] Slater J C 1964–74 *Quantum Theory of Molecules and Solids* vols 1–4 (New York: McGraw-Hill)
 - [19] Slater J C 1968 *Quantum Theory of Matter* (New York: McGraw-Hill)
 - [20] Hohenberg P and Kohn W 1964 *Phys. Rev. B* **136** 864
 - [21] Kohn W and Sham L 1965 *Phys. Rev. A* **140** 1133
 - [22] Ceperley D M and Alder B J 1980 *Phys. Rev. Lett.* **45** 566
 - [23] Perdew J P, Burke K and Wang Y 1996 *Phys. Rev. B* **54** 16 533 and references therein
-

-55-

- [24] Chelikowsky J R and Louie S G (eds) 1996 *Quantum Theory of Real Materials* (Boston: Kluwer)
- [25] Cohen M L and Chelikowsky J R 1989 *Electronic Structure and Optical Properties of Semiconductors* 2nd edn (Springer)
- [26] Phillips J C and Kleinman L 1959 *Phys. Rev.* **116** 287
- [27] Chelikowsky J R and Cohen M L 1992 *Ab initio* pseudopotentials for semiconductors *Handbook on Semiconductors* vol 1, ed P Landsberg (Amsterdam: Elsevier) p 59
- [28] Fermi E 1934 *Nuovo Cimento* **11** 157
- [29] Herring C 1940 *Phys. Rev.* **57** 1169
- [30] Ley L, Kowalczyk S P, Pollack R A and Shirley D A 1972 *Phys. Rev. Lett.* **29** 1088
- [31] Philipp H R and Ehrenreich H 1963 *Phys. Rev. Lett.* **127** 1550
- [32] Zucca R R L and Shen Y R 1970 *Phys. Rev. B* **1** 2668
- [33] Yang L W and Coppens P 1974 *Solid State Commun.* **15** 1555
- [34] Yin M T and Cohen M L 1980 *Phys. Rev. Lett.* **45** 1004
- [35] Chang K J, Dacorogna M M, Cohen M L, Mignot J M, Chouteau G and Martinez G 1985 *Phys. Rev. Lett.* **54** 2375
- [36] Dacorogna M M, Chang K J and Cohen M L 1985 *Phys. Rev. B* **32** 1853
- [37] Hybertsen M and Louie S G 1985 *Phys. Rev. Lett.* **55** 1418
- [38] Hybertsen M and Louie S G 1986 *Phys. Rev. B* **34** 5390
- [39] Benedict L X and Shirley E L 1999 *Phys. Rev. B* **59** 5441

- [40] Chelikowsky J R and Chou M Y 1988 *Phys. Rev. B* **38** 7966
- [41] Waseda Y 1980 *The Structure of Non-Crystalline Materials* (New York: McGraw-Hill)
- [42] Godlevsky V, Derby J and Chelikowsky J R 1998 *Phys. Rev. Lett.* **81** 4959
-

FURTHER READING

- Anderson P W 1963 *Concepts in Solids* (New York: Benjamin)
- Cox P A 1987 *The Electronic Structure and Chemistry of Solids* (Oxford: Oxford University Press)
- Harrison W A 1999 *Elementary Electronic Structure* (River Edge: World Scientific)
- Harrison W A 1989 *Electronic Structure and the Properties of Solids: The Physics of the Chemical Bond* (New York: Dover)
-

-56-

- Hummel R 1985 *Electronic Properties of Materials* (New York: Springer)
- Jones W and March N 1973 *Theoretical Solid State Physics* (New York: Wiley)
- Lerner R G and Trigg G L (eds) 1983 *Concise Encyclopedia of Solid State Physics* (Reading, MA: Addison-Wesley)
- Myers H P 1997 *Introductory Solid State Physics* (London: Taylor and Francis)
- Patterson J D 1971 *Introduction to the Theory of Solid State Physics* (Reading, MA: Addison-Wesley)
- Peierls R 1955 *Quantum Theory of Solids* (Oxford: Clarendon)
- Phillips J C 1973 *Bands and Bonds in Semiconductors* (New York: Academic)
- Pines D 1963 *Elementary Excitations in Solids* (New York: Benjamin)
- Seitz F 1948 *Modern Theory of Solids* (New York: McGraw-Hill)
-

-1-

A1.4 The symmetry of molecules

Per Jensen and P R Bunker

A1.4.1 INTRODUCTION

Unlike most words in a glossary of terms associated with the theoretical description of molecules, the word ‘symmetry’ has a meaning in every-day life. Many objects look exactly like their mirror image, and we say that they are symmetrical or, more precisely, that they have *reflection* symmetry. In addition to having reflection symmetry, a pencil (for example) is such that if we rotate it through any angle about its long axis it

will look the same. We say it has *rotational* symmetry. The concepts of rotation and reflection symmetry are familiar to us all.

The ball-and-stick models used in elementary chemistry education to visualize molecular structure are frequently symmetrical in the sense discussed above. Reflections in certain planes, rotations by certain angles about certain axes, or more complicated symmetry operations involving both reflection and rotation, will leave them looking the same. One might initially think that this is ‘the symmetry of molecules’ discussed in the present chapter, but it is not. Ball-and-stick models represent molecules fixed at their equilibrium configuration, that is, at the minimum (or at one of the minima) of the potential energy function for the electronic state under consideration. A real molecule is not static and generally it does not possess the rotation–reflection symmetry of its equilibrium configuration. Anyway, the use we make of molecular symmetry in understanding molecules, their spectra and their dynamics, has its basis in considerations other than the appearance of the molecule at equilibrium.

The true basis for understanding molecular symmetry involves studying the operations that leave the energy of a molecule unchanged, rather than studying the rotations or reflections that leave a molecule in its equilibrium configuration looking the same. Symmetry is a general concept. Not only does it apply to molecules, but it also applies, for example, to atoms, to atomic nuclei and to the particles that make up atomic nuclei. Also, the concept of symmetry applies to *nonrigid* molecules such as ammonia NH_3 , ethane C_2H_6 , the hydrogen dimer $(\text{H}_2)_2$, the water trimer $(\text{H}_2\text{O})_3$ and so on, that easily contort through structures that differ in the nature of their rotational and reflection symmetry. For a hydrogen molecule that is translating, rotating and vibrating in space, with the electrons orbiting, it is clear that the total energy of the molecule is unchanged if we interchange the coordinates and momenta of the two protons; the total kinetic energy is unchanged (since the two protons have the same mass), and the total electrostatic potential energy is unchanged (since the two protons have the same charge). However, the interchange of an electron and a proton will almost certainly not leave the molecular energy unchanged. Thus the permutation of identical particles is a symmetry operation and we will introduce others. In quantum mechanics the possible molecular energies are the eigenvalues of the molecular Hamiltonian and if the Hamiltonian is invariant to a particular operation (or, equivalently, if the Hamiltonian commutes with a particular operation) then that operation is a symmetry operation.

We collect symmetry operations into various ‘symmetry groups’, and this chapter is about the definition and use of such symmetry operations and symmetry groups. Symmetry groups are used to label molecular states and this labelling makes the states, and their possible interactions, much easier to understand. One important symmetry group that we describe is called *the molecular symmetry group* and the symmetry operations it contains are permutations of identical nuclei with and without the inversion of the molecule at its centre of mass. One fascinating outcome is that indeed for

-2-

rigid molecules (i.e., molecules that do not undergo large amplitude contortions to become *nonrigid* as discussed above) we can obtain a group of rotation and reflection operations that describes the rotation and reflection symmetry of the equilibrium molecular structure from the molecular symmetry group. However, by following the energy-invariance route we can understand the generality of the concept of symmetry and can readily deduce the symmetry groups that are appropriate for nonrigid molecules as well.

This introductory section continues with a subsection that presents the general motivation for using symmetry and ends with a short subsection that lists the various types of molecular symmetry.

A1.4.1.1 MOTIVATION: ROTATIONAL SYMMETRY AS AN EXAMPLE

Rotational symmetry is used here as an example to explain the motivation for using symmetry in molecular physics; it will be discussed in more detail in [section A1.4.3.2](#).

We consider an isolated molecule in field-free space with Hamiltonian \hat{H} . We let \hat{F} be the total angular momentum operator of the molecule, that is

$$\hat{F} = \hat{N} + \hat{S} + \hat{I} \quad (\text{A1.4.1})$$

where \hat{N} is the operator for the rovibronic angular momentum that results from the rotational motion of the nuclei and the orbital motion of the electrons, \hat{S} is the total electron spin angular momentum operator and \hat{I} is the total nuclear spin angular momentum operator. We introduce a Cartesian axis system (X, Y, Z) . The orientation of the (X, Y, Z) axis system is fixed in space (i.e., it is independent of the orientation in space of the molecule), but the origin is tied to the molecular centre of mass. It is well known that the molecular Hamiltonian \hat{H} commutes with the operators

$$\hat{F}^2 = \hat{F}_X^2 + \hat{F}_Y^2 + \hat{F}_Z^2 \quad (\text{A1.4.2})$$

and \hat{F}_Z where this is the component of \hat{F} along the Z axis, i.e.,

$$[\hat{F}^2, \hat{H}] = \hat{F}^2 \hat{H} - \hat{H} \hat{F}^2 = 0 \quad (\text{A1.4.3})$$

and

$$[\hat{F}_Z, \hat{H}] = 0. \quad (\text{A1.4.4})$$

It is also well known that \hat{F}^2 and \hat{F}_Z have simultaneous eigenfunctions $|F, m_F\rangle$ and that

$$\hat{F}^2 |F, m_F\rangle = F(F+1)\hbar^2 |F, m_F\rangle \quad (\text{A1.4.5})$$

-3-

and

$$\hat{F}_Z |F, m_F\rangle = m_F \hbar |F, m_F\rangle \quad (\text{A1.4.6})$$

where, for a given molecule, F assumes non-negative values that are either integral ($=0, 1, 2, 3, \dots$) or half-integral ($=1/2, 3/2, 5/2, \dots$) and, for a given F value, m_F has the $2F+1$ values $-F, -F+1, \dots, F-1, F$.

We can solve the molecular Schrödinger equation

$$\hat{H}\Psi_j = E_j\Psi_j \quad (\text{A1.4.7})$$

by representing the unknown wavefunction Ψ_j (where j is an index labelling the solutions) as a linear combination of known basis functions Ψ_n^0 ,

$$\Psi_j = \sum_n C_{jn} \Psi_n^0 \quad (\text{A1.4.8})$$

where the C_{jn} are expansion coefficients and n is an index labelling the basis functions. As described, for example, in section 6.6 of Bunker and Jensen [1], the eigenvalues E_j and expansion coefficients C_{jn} can be determined from the ‘Hamiltonian matrix’ by solving the secular equation

$$|H_{mn} - \delta_{mn}E| = 0 \quad (\text{A1.4.9})$$

where the Kronecker delta δ_{mn} has the value 1 for $m = n$ and the value 0 for $m \neq n$, and the Hamiltonian matrix elements H_{mn} are given by

$$H_{mn} = \int \Psi_m^0 * \hat{H} \Psi_n^0 d\tau \quad (\text{A1.4.10})$$

with integration carried out over the configuration space of the molecule. This process is said to involve ‘diagonalizing the Hamiltonian matrix’.

We now show what happens if we set up the Hamiltonian matrix using basis functions Ψ_n^0 that are eigenfunctions of \hat{F} and \hat{F}_Z with eigenvalues given by (equation A1.4.5) and (equation A1.4.6). We denote this particular choice of basis functions as Ψ_{n,F,m_F}^0 . From (equation A1.4.3), (equation A1.4.5) and the fact that \hat{F}^2 is a Hermitian operator, we derive

-4-

$$\begin{aligned} \langle \Psi_{m,F',m_F'}^0 | [\hat{F}^2, \hat{H}] | \Psi_{n,F'',m_F''}^0 \rangle &= \langle \hat{F}^2 \Psi_{m,F',m_F'}^0 | \hat{H} | \Psi_{n,F'',m_F''}^0 \rangle \\ &\quad - \langle \Psi_{m,F',m_F'}^0 | \hat{H} | \hat{F}^2 \Psi_{n,F'',m_F''}^0 \rangle \\ &= (F'(F'+1) - F''(F''+1)) \hbar^2 \langle \Psi_{m,F',m_F'}^0 | \hat{H} | \Psi_{n,F'',m_F''}^0 \rangle = 0 \end{aligned} \quad (\text{A1.4.11})$$

from which it follows that the matrix element $\langle \Psi_{m,F',m_F'}^0 | \hat{H} | \Psi_{n,F'',m_F''}^0 \rangle$ must vanish if $F' \neq F''$. From (equation A1.4.4) it follows in a similar manner that the matrix element must also vanish if $m_F' \neq m_F''$. That is, in the basis Ψ_{n,F,m_F}^0 the Hamiltonian matrix is block diagonal in F and m_F , and we can rewrite (equation A1.4.8) as

$$\Psi_j^{(F,m_F)} = \sum_n C_{jn}^{(F,m_F)} \Psi_{n,F,m_F}^0 \quad (\text{A1.4.12})$$

the eigenfunctions of \hat{H} are also eigenfunctions of \hat{F}^2 and \hat{F}_Z . We can further show that since m_F quantizes the molecular angular momentum along the arbitrarily chosen, space-fixed Z axis, the energy (i.e., the eigenvalue of \hat{H} associated with the function $\Psi_j^{(F,m_F)}$) is independent of m_F . That is, the $2F + 1$ states with common values of j and F and $m_F = -F, -F + 1, \dots, F$, are degenerate.

In order to solve (equation A1.4.7) we do not have to choose the basis functions to be eigenfunctions of \hat{F}^2 and \hat{F}_Z , but there are obvious advantages in doing so:

- The Hamiltonian matrix factorizes into blocks for basis functions having common values of F and m_F . This reduces the numerical work involved in diagonalizing the matrix.

- The solutions can be labelled by their values of F and m_F . We say that F and m_F are *good quantum numbers*. With this labelling, it is easier to keep track of the solutions and we can use the good quantum numbers to express selection rules for molecular interactions and transitions. In field-free space only states having the same values of F and m_F can interact, and an electric dipole transition between states with $F = F'$ and F'' will take place if and only if

$$|F' - F''| \leq 1 \text{ and } F' + F'' \geq 1. \quad (\text{A1.4.13})$$

At this point the reader may feel that we have done little in the way of explaining molecular symmetry. All we have done is to state basic results, normally treated in introductory courses on quantum mechanics, connected with the fact that it is possible to find a complete set of simultaneous eigenfunctions for two or more commuting operators. However, as we shall see in [section A1.4.3.2](#), the fact that the molecular Hamiltonian \hat{H} commutes with \hat{F}^2 and \hat{F}_Z is intimately connected to the fact that \hat{H} commutes with (or, equivalently, is invariant to) any rotation of the molecule about a space-fixed axis passing through the centre of mass of the molecule. As stated above, an operation that leaves the Hamiltonian invariant is a symmetry operation of the Hamiltonian. The infinite set of all possible rotations of the

-5-

molecule about all possible axes that pass through the molecular centre of mass can be collected together to form a *group* (see below). Following the notation of Bunker and Jensen [1] we call this group \mathbf{K} (spatial). Since all elements of \mathbf{K} (spatial) are symmetry operations of \hat{H} , we say that \mathbf{K} (spatial) is a *symmetry group* of \hat{H} . Any group has a set of *irreducible representations* and they define the way coordinates, wavefunctions and operators have to transform under the operations in the group; it so happens that the irreducible representations of \mathbf{K} (spatial), $D^{(F)}$, are labelled by the angular momentum quantum number F . The $2F + 1$ functions $|F, m_F\rangle$ (or Ψ_{n, F, m_F}^0 or $\Psi_j^{(F, m_F)}$) with a common value of F (and n or j) and $m_F = -F, -F + 1, \dots, F$ transform according to the irreducible representation $D^{(F)}$ of \mathbf{K} (spatial). As a result, we can reformulate our procedure for solving the Schrödinger equation of a molecule as follows:

- For the Hamiltonian \hat{H} we identify a symmetry group, and this is a group of symmetry operations of \hat{H} a symmetry operation being defined as an operation that leaves \hat{H} invariant (i.e., that commutes with \hat{H}). In our example, the symmetry group is \mathbf{K} (spatial).
- Having done this we solve the Schrödinger equation for the molecule by diagonalizing the Hamiltonian matrix in a complete set of known basis functions. We choose the basis functions so that they transform according to the irreducible representations of the symmetry group.
- The Hamiltonian matrix will be block diagonal in this basis set. There will be one block for each irreducible representation of the symmetry group.
- As a result the eigenstates of \hat{H} can be labelled by the irreducible representations of the symmetry group and these irreducible representations can be used as ‘good quantum numbers’ for understanding interactions and transitions.

We have described here one particular type of molecular symmetry, *rotational symmetry*. On one hand, this example is complicated because the appropriate symmetry group, \mathbf{K} (spatial), has infinitely many elements. On the other hand, it is simple because each irreducible representation of \mathbf{K} (spatial) corresponds to a particular value of the quantum number F which is associated with a physically observable quantity, the angular momentum. Below we describe other types of molecular symmetry, some of which give rise to finite symmetry groups.

A1.4.1.2 A LIST OF THE VARIOUS TYPES OF MOLECULAR SYMMETRY

The possible types of symmetry for the Hamiltonian of an isolated molecule in field-free space (all of them are discussed in more detail later on in the article) can be listed as follows:

- (i) *Translational symmetry*. A translational symmetry operation displaces all nuclei and electrons in the molecule uniformly in space (i.e., all particles are moved in the same direction and by the same distance). This symmetry is a consequence of the uniformity of space.
 - (ii) *Rotational symmetry*. A rotational symmetry operation rotates all nuclei and electrons by the same angle about a space-fixed axis that passes through the molecular centre of mass. This symmetry is a consequence of the isotropy of space.
-

-6-

- (iii) *Inversion symmetry*. The Hamiltonian that we customarily use to describe a molecule involves only the electromagnetic forces between the particles (nuclei and electrons) and these forces are invariant to the 'inversion operation' E^* which inverts all particle positions through the centre of mass of the molecule. Thus such a Hamiltonian commutes with E^* ; the use of this operation leads (as we see in [section A1.4.2.5](#)) to the concept of *parity*, and parity can be + or -. This symmetry results from the fact that the electromagnetic force is invariant to inversion. It is not a property of space.
- (iv) *Identical particle permutation symmetry*. The corresponding symmetry operations permute identical particles in a molecule. These particles can be electrons, or they can be identical nuclei. This symmetry results from the indistinguishability of identical particles.
- (v) *Time reversal symmetry*. The time reversal symmetry operation T or $\hat{\theta}$ reverses the direction of motion in a molecule by reversing the sign of all linear and angular momenta. This symmetry results from the properties of the Schrödinger equation of a system of particles moving under the influence of electromagnetic forces. It is not a property of space-time.

We hope that by now the reader has it firmly in mind that the way molecular symmetry is defined and used is based on energy invariance and not on considerations of the geometry of molecular equilibrium structures. Symmetry defined in this way leads to the idea of *conservation*. For example, the total angular momentum of an isolated molecule in field-free space is a conserved quantity (like the total energy) since there are no terms in the Hamiltonian that can mix states having different values of F . This point is discussed further in [section A1.4.3.1](#) and [section A1.4.3.2](#).

A1.4.2 GROUP THEORY

The use of symmetry involves the mathematical apparatus of *group theory*, and in this section we summarize the basics. We first define the concept of a *group* by considering the permutations of the protons in the phosphine molecule PH_3 (figure A1.4.1) as an example. This leads to the definition of the nuclear permutation group for PH_3 . We briefly discuss point groups and then introduce *representations* of a group; in particular we define *irreducible representations*. We then go on to show how wavefunctions are transformed by symmetry operations, and how this enables molecular states to be labelled according to the irreducible representations of the applicable symmetry group. The final subsection explains *the vanishing integral rule* which is of major use in applying molecular symmetry in order to determine which transitions and interactions can and cannot occur.

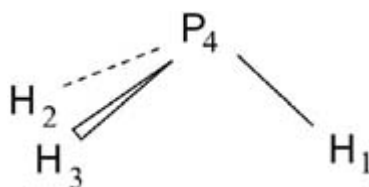


Figure A1.4.1. A PH_3 molecule at equilibrium. The protons are labelled 1, 2 and 3, respectively, and the phosphorus nucleus is labelled 4.

-7-

A1.4.2.1 NUCLEAR PERMUTATION GROUPS

The three protons in PH_3 are identical and indistinguishable. Therefore the molecular Hamiltonian will commute with any operation that permutes them, where such a permutation interchanges the space and spin coordinates of the protons. Although this is a rather obvious symmetry, and a proof is hardly necessary, it can be proved by formal algebra as done in chapter 6 of [1].

How many distinct ways of permuting the three protons are there? For example, we can interchange protons 1 and 2. The corresponding symmetry operation is denoted (12) (pronounced ‘one–two’) and it is to be understood quite literally: protons 1 and 2 interchange their positions in space. There are obviously two further distinct operations of this type: (23) and (31)¹. A permutation operation that interchanges just two nuclei is called a *transposition*. A more complicated symmetry operation is (123). Here, nucleus 1 is replaced by nucleus 2, nucleus 2 by nucleus 3 and nucleus 3 by nucleus 1. Thus, after (123) nucleus 2 ends up at the position in space initially occupied by nucleus 1, nucleus 3 ends up at the position in space initially occupied by nucleus 2 and nucleus 1 ends up at the position in space initially occupied by nucleus 3. Such an operation, which involves more than two nuclei, is called a *cycle*. A moment’s thought will show that in the present case, there exists one other distinct cycle, namely (132). We could write further cycles like (231), (321) etc, but we discover that each of them has the same effect as (123) or (132). There are thus five distinct ways of permuting three protons: (123), (132), (12), (23) and (31).

We can apply permutations successively. For example, we can first apply (12), and then (123); the net effect of doing this is to interchange protons 1 and 3. Thus we have

$$(123)(12) = (31). \quad (\text{A1.4.14})$$

When we apply permutations (or other symmetry operations) successively (this is commonly referred to as *multiplying* the operations so that (31) is the *product* of (123) and (12)), we write the operation to be applied first to the right in the manner done for general quantum mechanical operators. Permutations do not necessarily commute. For example,

$$(12)(123) = (23). \quad (\text{A1.4.15})$$

If we apply the operation (12) twice, or the operation (123) three times, we obviously get back to the starting point. We write this as

$$(\text{A1.4.16})$$

$$(12)(12) = (123)(123)(123) = E$$

where the *identity operation* E leaves the molecule unchanged by definition. Having defined E , we define the *reciprocal* (or *inverse*) R^{-1} of a symmetry operation R (which, in our present example, could be (123), (132), (12), (23) or (31)) by the equation

$$RR^{-1} = R^{-1}R = E. \tag{A1.4.17}$$

-8-

It is easy to verify that for example

$$(12)^{-1} = (12) \text{ and } (123)^{-1} = (132). \tag{A1.4.18}$$

The six operations

$$\mathcal{S}_3 = \{E, (123), (132), (12), (23), (31)\} \tag{A1.4.19}$$

are said to form a *group* because they satisfy the following *group axioms*:

- (i) We can multiply (i.e., successively apply) the operations together in pairs and the result is a member of the group.
- (ii) One of the operations in the group is the identity operation E .
- (iii) The reciprocal of each operation is a member of the group.
- (iv) Multiplication of the operations is associative; that is, in a multiple product the answer is independent of how the operations are associated in pairs, e.g.,

$$(12)(123)(23) = (12) \underbrace{[(123)(23)]}_{(12)} = \underbrace{[(12)(123)]}_{(23)}(23) = E. \tag{A1.4.20}$$

The fact that the group axioms (i), (ii), (iii) and (iv) are satisfied by the set in (equation A1.4.19) can be verified by inspecting the *multiplication table* of the group \mathcal{S}_3 given in table A1.4.1; this table lists all products R_1R_2 where R_1 and R_2 are members of \mathcal{S}_3 . The group \mathcal{S}_3 is the permutation group (or symmetric group) of degree 3, and it consists of all permutations of three objects. There are six elements in \mathcal{S}_3 and the group is said to have *order* six. In general, the permutation group \mathcal{S}_n (all permutations of n objects) has order $n!$.

Table A1.4.1 The multiplication table of the \mathcal{S}_3 group.

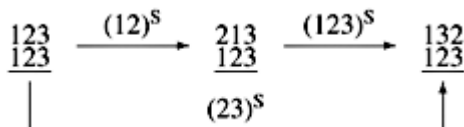
E (123)(132) (12) (23) (31)

E	E	(123)	(132)	(12)	(23)	(31)
(123)	(123)	(132)	E	(31)	(12)	(23)
(132)	(132)	E	(123)	(23)	(31)	(12)
(12)	(12)	(23)	(31)	E	(123)	(132)
(23)	(23)	(31)	(12)	(132)	E	(123)
(31)	(31)	(12)	(23)	(123)	(132)	E

Each entry is the product of first applying the permutation at the top of the column and then applying the permutation at the left end of the row.

-9-

There is another way of developing the algebra of permutation multiplication, and we briefly explain it. In this approach for PH_3 three positions in space are introduced and labelled $\underline{1}$, $\underline{2}$ and $\underline{3}$; the three protons are labelled H_1 , H_2 and H_3 . The permutation $(12)^S$ (where S denotes space-fixed position labels) is defined in this approach as permuting the nuclei that are in positions $\underline{1}$ and $\underline{2}$, and the permutation $(123)^S$ as replacing the proton in position $\underline{1}$ by the proton in position $\underline{2}$ etc. With this definition the effect of first doing $(12)^S$ and then doing $(123)^S$ can be drawn as



and we see that

$$(123)^S(12)^S = (23)^S. \quad (\text{A1.4.21})$$

This is not the same as (equation A1.4.14). In fact, in this convention, which we can call the S -convention, the multiplication table is the transpose of that given in table A1.4.1. The convention we use and which leads to the multiplication table given in table A1.4.1, will be called the N -convention (where N denotes nuclear-fixed labels).

A1.4.2.2 POINT GROUPS

Having defined the concept of a group in section A1.4.2.1, we discuss in the present section a particular type of group that most readers will have heard about: the *point group*. We do this with some reluctance since point group operations do not commute with the complete molecular Hamiltonian and thus they are not true symmetry operations of the kind discussed in section A1.4.1.2. Also the actual effect that the operations have on molecular coordinates is not straightforward to explain. From a pedagogical and logical point of view it would be better to bring them into the discussion of molecular symmetry only after groups consisting of the true symmetry operations enumerated in section A1.4.1.2 have been thoroughly explained. However, because of their historical importance we have decided to introduce them early on. As explained in section A1.4.4 the operations of a molecular point group involve the rotation and/or reflection of vibrational displacement coordinates and electronic coordinates, within the molecular-fixed coordinate system which itself remains fixed in space. Thus the rotational variables (called Euler angles) that define the orientation of a molecule in space are not transformed and in particular the molecule is not subjected to an overall rotation by the operations that are called 'rotations' in the molecular point group. It turns out that the molecular point group is a symmetry group of use in the understanding of the vibrational and electronic states of molecules. However,

because of centrifugal and Coriolis forces the vibrational and electronic motion is not completely separable from the rotational motion and, as we explain in [section A1.4.5](#), the molecular point group is only a *near symmetry group* of the complete molecular Hamiltonian appropriate for the hypothetical, non-rotating molecule.

In general, a point group symmetry operation is defined as a rotation or reflection of a macroscopic object such that, after the operation has been carried out, the object looks the same as it did originally. The macroscopic objects we consider here are models of molecules in their equilibrium configuration; we could also consider idealized objects such as cubes, pyramids, spheres, cones, tetrahedra etc. in order to define the various possible point groups.

-10-

As an example, we again consider the PH_3 molecule. In its pyramidal equilibrium configuration PH_3 has all three P–H distances equal and all three bond angles $\angle(\text{HPH})$ equal. This object has the point group symmetry C_{3v} where the operations of the group are

$$C_{3v} = \{E, C_3, C_3^2, \sigma_1, \sigma_2, \sigma_3\}. \quad (\text{A1.4.22})$$

The operations in the group can be understood by referring to figure A1.4.2 In this figure the right-handed Cartesian (p, q, r) axis system has origin at the molecular centre of mass, the P nucleus is above the pq plane (the plane of the page), and the three H nuclei are below the pq plane². The operations C_3 and C_3^2 in (equation A1.4.22) are right-handed rotations of 120° and 240° , respectively, about the r axis. In general, we use the notation C_n for a rotation of $2\pi/n$ radians about an axis³. Somewhat unfortunately, it is customary to use the symbol C_n to denote not only the rotation operation, but also the rotation axis. That is, we say that the r axis in figure A1.4.2 is a C_3 axis. The operation σ_1 is a reflection in the pr plane (which, with the same unfortunate lack of distinction used in the case of the C_3 operation and the C_3 axis, we call the σ_1 plane), and σ_2 and σ_3 are reflections in the σ_2 and σ_3 planes; these planes are obtained by rotating by 120° and 240° , respectively, about the r axis from the pr plane. As shown in figure A1.4.2, each of the H nuclei in the PH_3 molecule lies in a σ_k plane ($k = 1, 2, 3$) and the P nucleus lies on the C_3 axis. It is clear that the operations of C_{3v} as defined here leave the PH_3 molecule in its static equilibrium configuration looking unchanged. It is important to realize that when we apply the point group operations we do not move the (p, q, r) axes (we call this the ‘space-fixed’ axis convention) and we will now show how this aspect of the way point group operations are defined affects the construction of the group multiplication table.

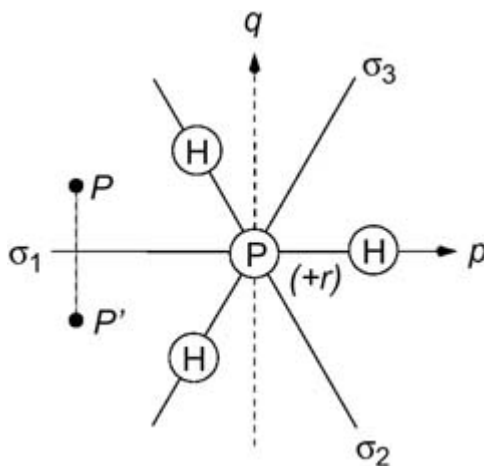


Figure A1.4.2. The PH_3 molecule at equilibrium. The symbol $(+r)$ indicates that the r axis points up, out of the plane of the page.

Formally, we can say that the operations in C_{3v} act on points in space. For example, we show in figure A1.4.2 how a point P in the pq plane is transformed into another point P' by the operation σ_1 ; we can say that $P' = \sigma_1 P$. The reader can now show by geometrical considerations that if we first reflect a point P in the σ_1 plane to obtain $P' = \sigma_1 P$, and we then reflect P' in the σ_2 plane to obtain $P'' = \sigma_2 P' = \sigma_2 \sigma_1 P$, then P'' can be obtained directly from P by a 240° anticlockwise rotation about the r axis. Thus $P'' = C_3^2 P$ or generally

-11-

$$C_3^2 = \sigma_2 \sigma_1. \quad (\text{A1.4.23})$$

We can also show that

$$C_3 = \sigma_3 \sigma_1. \quad (\text{A1.4.24})$$

The complete multiplication table of the C_{3v} point group, worked out using arguments similar to those leading to (equation A1.4.23) and (equation A1.4.24), is given in table A1.4.2. It is left as an exercise for the reader to use this table to show that the elements of C_{3v} satisfy the group axioms given in [section A1.4.2.1](#).

Table A1.4.2 The multiplication table of the C_{3v} point group using the space-fixed axis convention (see text).

$E \ C_3 \ C_3^2 \ \sigma_1 \ \sigma_2 \ \sigma_3$

E	E	C_3	C_3^2	σ_1	σ_2	σ_3
C_3	C_3	C_3^2	E	σ_3	σ_1	σ_2
C_3^2	C_3^2	E	C_3	σ_2	σ_3	σ_1
σ_1	σ_1	σ_2	σ_3	E	C_3	C_3^2
σ_2	σ_2	σ_3	σ_1	C_3^2	E	C_3
σ_3	σ_3	σ_1	σ_2	C_3	C_3^2	E

Each entry is the product of first applying the operation at the top of the column and then applying the operation at the left end of the row.

If we were to define the operations of the point group as also rotating and reflecting the (p,q,r) axis system (in which case the axes would be 'tied' to the positions of the nuclei), we would obtain a different multiplication table. We could call this the 'nuclear-fixed axis convention.' To implement this the protons in the σ_1 , σ_2 and σ_3 planes in [figure A1.4.2](#) would be numbered H_1 , H_2 and H_3 respectively. With this convention the C_3 operation would move the σ_1 plane to the position in space originally occupied by the σ_2 plane. If we follow such a C_3 operation by the σ_1 reflection (in the plane containing H_1) we find that, in the nuclear-fixed axis convention:

$$\sigma_1 C_3 = \sigma_3. \quad (\text{A1.4.25})$$

Similarly, with the nuclear-fixed axis convention, we determine that

$$C_3 = \sigma_1 \sigma_3 \quad (\text{A1.4.26})$$

and this result also follows by multiplying (equation A1.4.25) on the left by σ_1 . The multiplication table obtained using the nuclear-fixed axis convention is the transpose of the multiplication table obtained using the space-fixed axis convention (compare (equation A1.4.24) and (equation A1.4.26)). In dealing with point groups we will use the space-fixed axis convention. For defining the effect of permutation operations the S-convention (see [\(equation A1.4.21\)](#))

-12-

is related to the N-convention (see [\(equation A1.4.14\)](#)) in the same way that the space-fixed and nuclear-fixed axis conventions for point groups are related.

The operations in a point group are associated with so-called *symmetry elements*. Symmetry elements can be rotation axes (such as the C_3 axis that gives rise to the C_3 and operations in C_{3v}) or reflection planes (such as the planes $\sigma_1, \sigma_2, \sigma_3$; each of which gives rise to a reflection operation in C_{3v}). A third type of symmetry element not present in C_{3v} is the *rotation-reflection axis* or *improper axis*. For example, an allene molecule H_2CCCH_2 in its equilibrium configuration will be unchanged in appearance by a rotation of 90° about the CCC axis combined with a reflection in a plane perpendicular to this axis and containing the 'middle' C nucleus. This operation (a *rotation-reflection* or an *improper rotation*) is called S_4 ; it is an element of the point group of allene, D_{2d} . Allene is said to have as a symmetry element the rotation-reflection axis or improper axis S_4 . It should be noted that neither the rotation of 90° about the CCC axis nor the reflection in the plane perpendicular to it are themselves in D_{2d} . For an arbitrary point group, all symmetry elements will intersect at the centre of mass of the object; this point is left unchanged by the group operations and hence the name point group. In order to determine the appropriate point group for a given static arrangement of nuclei, one first identifies the symmetry elements present. Cotton [2] gives in his section 3.14 a systematic procedure to select the appropriate point group from the symmetry elements found. The labels customarily used for point groups (such as C_{3v} and D_{2d}) are named *Schönflies symbols* after their inventor. The most important point groups (defined by their symmetry elements) are

- C_n one n -fold rotation axis,
- C_{nv} one n -fold rotation axis and n reflection planes containing this axis,
- C_{nh} one n -fold rotation axis and one reflection plane perpendicular to this axis,
- D_n one n -fold rotation axis and n twofold rotation axes perpendicular to it,
- D_{nd} those of D_n plus n reflection planes containing the n -fold rotation axis and bisecting the angles between the n twofold rotation axes,
- D_{nh} those of D_n plus a reflection plane perpendicular to the n -fold rotation axis,
- S_n one alternating axis of symmetry (about which rotation by $2\pi/n$ radians followed by reflection in a plane perpendicular to the axis is a symmetry operation).

The point groups T_d , O_h and I_h consist of all rotation, reflection and rotation-reflection symmetry operations of a regular tetrahedron, cube and icosahedron, respectively.

Point groups are discussed briefly in sections 4.3 and 4.4 of [1] and very extensively in chapter 3 of Cotton

[2]. We refer the reader to these literature sources for more details.

A1.4.2.3 IRREDUCIBLE REPRESENTATIONS AND CHARACTER TABLES

If we have two groups A and B , of the same order h :

$$\tag{A1.4.27}$$

-13-

$$B = \{B_1, B_2, B_3, \dots, B_h\} \tag{A1.4.28}$$

where $A_1 = B_1 = E$, the identity operation and if there is a one-to-one correspondence between the elements of A and B , $A_k \leftrightarrow B_k$, $k = 1, 2, 3, \dots, h$, so that if

$$A_i A_j = A_m \tag{A1.4.29}$$

it can be inferred that

$$B_i B_j = B_m \tag{A1.4.30}$$

for all $i \leq h$ and $j \leq h$, then the two groups A and B are said to be *isomorphic*.

As an example we consider the group S_3 introduced in (equation A1.4.19) and the point group C_{3v} given in (equation A1.4.22). Inspection shows that the multiplication table of C_{3v} in table A1.4.2 can be obtained from the multiplication table of the group S_3 (table A1.4.1) by the following mapping:

$$\begin{array}{l} S_3 : \quad E \quad (123) \quad (132) \quad (12) \quad (23) \quad (31) \\ C_{3v} : \quad E \quad C_3 \quad C_3^2 \quad \sigma_3 \quad \sigma_1 \quad \sigma_2. \end{array} \tag{A1.4.31}$$

Thus, C_{3v} and S_3 are isomorphic.

Homomorphism is analogous to isomorphism. Where an isomorphism is a one-to-one correspondence between elements of groups of the same order, homomorphism is a many-to-one correspondence between elements of groups having different orders. The larger group is said to be homomorphic onto the smaller group. For example, the point group C_{3v} is homomorphic onto $S_2 = \{E, (12)\}$ with the following correspondences:

$$\begin{array}{l} C_{3v} : \quad E \quad C_3 \quad C_3^2 \quad \sigma_1 \quad \sigma_2 \quad \sigma_3 \\ S_2 : \quad \quad \underbrace{E} \quad \underbrace{(12)} \quad . \end{array} \tag{A1.4.32}$$

The multiplication table of S_2 has the entries $EE = E$, $E(12) = (12)E = (12)$ and $(12)(12) = E$. If, in the multiplication table of C_{3v} (table A1.4.2), the elements E , C_3 and C_3^2 are each replaced by E (of S_2) and σ_1 , σ_2 and σ_3 each by (12) , we obtain the multiplication table of S_2 nine times over.

We are particularly concerned with isomorphisms and homomorphisms, in which one of the groups involved is a *matrix group*. In this circumstance the matrix group is said to be a *representation* of the other group. The elements of a matrix group are square matrices, all of the same dimension. The ‘successive application’ of two

matrix group elements (in the sense of group axiom (i) in [section A1.4.2.1](#)) is matrix multiplication. Thus, the identity operation E of a matrix group is the unit matrix of the appropriate dimension, and the inverse element of a matrix is its inverse matrix. Matrices and matrix groups are discussed in more detail in section 5.1 of [1].

-14-

For the group A in ([equation A1.4.27](#)) to be isomorphic to, or homomorphic onto, a matrix group containing matrices of dimension ℓ , say, each element A_k of A is mapped onto an $\ell \times \ell$ matrix \mathbf{M}_k , $k = 1, 2, 3, 4, \dots, h$, and ([equation A1.4.29](#)) and ([equation A1.4.30](#)) can be rewritten in the form

$$\text{if } A_i A_j = A_m \text{ then } \mathbf{M}_i \mathbf{M}_j = \mathbf{M}_m \quad (\text{A1.4.33})$$

for all $i \leq h$ and $j \leq h$. The latter part of this equation says that the $\ell \times \ell$ matrix \mathbf{M}_m is the product of the two $\ell \times \ell$ matrices \mathbf{M}_i and \mathbf{M}_j .

If we have found one representation of ℓ -dimensional matrices $\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3, \dots$ of the group A , then, at least for $\ell > 1$, we can define infinitely many other *equivalent representations* consisting of the matrices

$$\mathbf{M}'_k = \mathbf{V}^{-1} \mathbf{M}_k \mathbf{V} \quad k = 1, 2, 3, \dots, h \quad (\text{A1.4.34})$$

where \mathbf{V} is an $\ell \times \ell$ matrix. The determinant of \mathbf{V} must be nonvanishing, so that \mathbf{V}^{-1} exists, but otherwise \mathbf{V} is arbitrary. We say that \mathbf{M}'_k is obtained from \mathbf{M}_k by a *similarity transformation*. It is straightforward to show that the matrices \mathbf{M}'_k , $k = 1, 2, 3, \dots, h$ form a representation of A since they satisfy an equation analogous to ([equation A1.4.33](#)).

It is well known that the *trace* of a square matrix (i.e., the sum of its diagonal elements) is unchanged by a similarity transformation. If we define the traces

$$\chi'_k = \sum_{p=1}^{\ell} (\mathbf{M}'_k)_{pp} \text{ and } \chi_k = \sum_{p=1}^{\ell} (\mathbf{M}_k)_{pp} \quad (\text{A1.4.35})$$

we have

$$\chi'_k = \chi_k. \quad (\text{A1.4.36})$$

The traces of the representation matrices are called the *characters* of the representation, and (equation A1.4.36) shows that all equivalent representations have the same characters. Thus, the characters serve to distinguish inequivalent representations.

If we select an element of \mathcal{A} , A_j say, and determine the set of elements S given by forming all products

$$S = R^{-1} A_j R \quad (\text{A1.4.37})$$

-15-

where R runs over all elements of \mathcal{A} , then the set of distinct elements obtained, which will include A_j (since for $R = R^{-1} = E$ we have $S = A_j$), is said to form a *class* of \mathcal{A} . For any group the identity operation E is always in a class of its own since for all R we have $S = R^{-1} E R = R^{-1} R = E$. The reader can use the multiplication table (table A1.4.1) to determine the classes of the group \mathcal{S}_3 (equation (A1.4.19)); there are three classes $[E]$, $[(123),(132)]$ and $[(12),(23),(31)]$. Since the groups \mathcal{S}_3 and \mathcal{C}_{3v} ((equation A1.4.22)) are isomorphic, the class structure of \mathcal{C}_{3v} can be immediately inferred from the class structure of \mathcal{S}_3 together with (equation A1.4.31). \mathcal{C}_{3v} has the classes $[E]$, $[C_3, C_3^2]$ and $[\sigma_1, \sigma_2, \sigma_3]$.

If two elements of \mathcal{A} , A_i and A_j say, are in the same class, then there exists a third element of \mathcal{A} , R , such that

$$A_i = R^{-1} A_j R. \quad (\text{A1.4.38})$$

Then by (equation A1.4.33)

$$\mathbf{M}_i = \mathbf{M}_R^{-1} \mathbf{M}_j \mathbf{M}_R \quad (\text{A1.4.39})$$

where \mathbf{M}_i , \mathbf{M}_j and \mathbf{M}_R are the representation matrices associated with A_i , A_j and R , respectively. That is, \mathbf{M}_i is obtained from \mathbf{M}_j in a similarity transformation, and

these two matrices thus have the same trace or character. Consequently, all the elements in a given class of a group are represented by matrices with the same character.

If we start with an ℓ -dimensional representation of A consisting of the matrices $\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3, \dots$, it may be that we can find a matrix \mathbf{V} such that when it is used with (equation A1.4.34) it produces an equivalent representation $\mathbf{M}'_1, \mathbf{M}'_2, \mathbf{M}'_3, \dots$ each of whose matrices is in the same *block diagonal form*. For example, the nonvanishing elements of each of the matrices \mathbf{M}'_k could form an upper-left-corner $\ell_1 \times \ell_1$ block and a lower-right-corner $\ell_2 \times \ell_2$ block, where $\ell_1 + \ell_2 = \ell$. In this situation, a few moments' consideration of the rules of matrix multiplication shows that all the upper-left-corner $\ell_1 \times \ell_1$ blocks, taken on their own, form an ℓ_1 -dimensional representation of A and all the lower-right-corner $\ell_2 \times \ell_2$ blocks, taken on their own, form an ℓ_2 -dimensional representation. In these circumstances the original representation Γ consisting of $\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3, \dots$ is *reducible* and we have *reduced* it to the sum of the two representations, Γ_1 and Γ_2 say, of dimensions ℓ_1 and ℓ_2 , respectively. We write this reduction as

$$\Gamma = \Gamma_1 \oplus \Gamma_2. \quad (\text{A1.4.40})$$

Clearly, a one-dimensional representation (also called a *non-degenerate* representation) is of necessity *irreducible* in that it cannot be reduced to representations of lower dimensions. *Degenerate representations* (i.e., groups of matrices with dimension higher than 1) can also be irreducible, which means that there is no matrix that by a similarity transformation will bring all the matrices of the representation into the same block diagonal form. It can be shown that the number of irreducible representations of a given group is equal to the number of classes in the group. We have seen that the group \mathcal{S}_3 has three classes $[E], [(123), (132)]$ and $[(23), (31), (12)]$ and therefore it has three irreducible representations. For a general group with n irreducible representations with dimensions $\ell_1, \ell_2, \ell_3, \dots, \ell_n$, it can also be shown that

-16-

$$\sum_{i=1}^n \ell_i^2 = h \quad (\text{A1.4.41})$$

where h is the order of the group. For \mathcal{S}_3 this equation yields

$$\ell_1^2 + \ell_2^2 + \ell_3^2 = 6 \quad (\text{A1.4.42})$$

and, since the ℓ_i have to be positive integers, we obtain $\ell_1 = \ell_2 = 1$ and $\ell_3 = 2$. When developing general formulae we label the irreducible representations of a group as $\Gamma_1, \Gamma_2, \dots, \Gamma_n$ and denote the characters associated with Γ_i as $\chi^{\Gamma_i}[R]$, where R is an element of the group under study. However, the irreducible representations of symmetry groups are denoted by various other special symbols such as A_2, Σ^- and $D^{(4)}$. The characters of the irreducible representations of a symmetry group are collected together into a *character table* and the character table of the group S_3 is given in table A1.4.3. The construction of character tables for finite groups is treated in section 4.4 of [2] and section 3-4 of [3].

Table A1.4.3 The character table of the S_3 group.

	<i>E</i>	(123)	(12)
S_3	1	2	3
A_1	1	1	1
A_2	1	1	-1
<i>E</i>	2	-1	0

One representative element in each class is given, and the number written below each element is the number of elements in the class.

For any Γ_i we have $\chi^{\Gamma_i}[E] = \ell_i$, the dimension of Γ_i . This is because the identity operation E is always represented by an $\ell_i \times \ell_i$ unit matrix whose trace obviously is ℓ_i . For any group there will be one irreducible representation (called the *totally symmetric representation* $\Gamma^{(s)}$) which has all $\chi^{\Gamma_i}[R] = 1$. Such a representation exists because any group is homomorphic onto the one-member matrix group $\{1\}$ (where the ‘1’ is interpreted as a 1×1 matrix). The irreducible characters $\chi^{\Gamma_i}[R]$ satisfy several equations (see, for example, section 4.3 of [2] and section 3-3 of [3]), for example

$$\sum_R \chi^{\Gamma_i}[R]^* \chi^{\Gamma_j}[R] = h \delta_{ij} \tag{A1.4.43}$$

where the sum runs over all elements R of the group.

In applications of group theory we often obtain a reducible representation, and we then need to reduce it to its irreducible components. The way that a given representation of a group is reduced to its irreducible components depends only on the characters of the matrices in the representation and on the characters of the matrices in the irreducible representations of the group. Suppose that the reducible representation is Γ and that the group involved

has irreducible representations that we label $\Gamma_1, \Gamma_2, \Gamma_3, \dots$. What we mean by ‘reducing’ Γ is finding the integral coefficients a_i in the expression

$$\tag{A1.4.44}$$

where

(A1.4.45)

with the sum running over all the irreducible representations of the group. Multiplying (equation A1.4.45) on the right by $[R]^*$ and summing over R it follows from the character orthogonality relation (equation (A1.4.43)) that the required a_i are given by

(A1.4.46)

where h is the order of the group and R runs over all the elements of the group.

A1.4.2.4 THE EFFECTS OF SYMMETRY OPERATIONS

For the PH_3 molecule, which we continue using as an example, we consider that proton i ($= 1, 2$ or 3) initially has the coordinates (X_i, Y_i, Z_i) in the (X, Y, Z) axis system, and the phosphorus nucleus has the coordinates (X_4, Y_4, Z_4) . After applying the permutation operation (12) to the PH_3 molecule, nucleus 1 is where nucleus 2 was before. Consequently, nucleus 1 now has the coordinates (X_2, Y_2, Z_2) . Nucleus 2 is where nucleus 1 was before and has the coordinates (X_1, Y_1, Z_1) . Thus we can write

(A1.4.47)

where X_i, Y_i, Z_i are the X, Y and Z coordinates of nucleus i after applying the permutation (12). By convention we always give first the (X, Y, Z) coordinates of nucleus 1, then those of nucleus 2, then those of nucleus 3 etc.

Similarly, after applying the operation (123) to the PH_3 molecule, nucleus 2 is where nucleus 1 was before and has the coordinates (X_1, Y_1, Z_1) . Nucleus 3 is where nucleus 2 was before and has the coordinates (X_2, Y_2, Z_2) and, finally, nucleus 1 is where nucleus 3 was before and has the coordinates (X_3, Y_3, Z_3) . So

-18-

$$\begin{aligned} (123) [X_1, Y_1, Z_1, X_2, Y_2, Z_2, X_3, Y_3, Z_3, X_4, Y_4, Z_4] \\ = [X'_1, Y'_1, Z'_1, X'_2, Y'_2, Z'_2, X'_3, Y'_3, Z'_3, X_4, Y_4, Z_4] \\ = [X_3, Y_3, Z_3, X_1, Y_1, Z_1, X_2, Y_2, Z_2, X_4, Y_4, Z_4] \end{aligned} \quad (\text{A1.4.48})$$

where here X'_i, Y'_i and Z'_i are the X, Y and Z coordinates of nucleus i after applying the permutation (123).

The procedure exemplified by (equation A1.4.47) and (equation A1.4.48) can be trivially generalized to define the effect of any symmetry operation, R say, on the coordinates (X_i, Y_i, Z_i) of any nucleus or electron i in any molecule by writing

$$R[X_i, Y_i, Z_i] = [RX_i, RY_i, RZ_i] = [X'_i, Y'_i, Z'_i]. \quad (\text{A1.4.49})$$

We can also write

$$R^{-1}[X'_i, Y'_i, Z'_i] = [R^{-1}X'_i, R^{-1}Y'_i, R^{-1}Z'_i] = [X_i, Y_i, Z_i]. \quad (\text{A1.4.50})$$

We use the nuclear permutation operations (123) and (12) to show what happens when we apply two operations in succession. We write the successive effect of these two permutations as (remember that we are using the N-convention; see (equation A1.4.14))

$$\begin{aligned}
(123)(12) [X_1, Y_1, Z_1, X_2, Y_2, Z_2, X_3, Y_3, Z_3, X_4, Y_4, Z_4] \\
= (123) [X'_1, Y'_1, Z'_1, X'_2, Y'_2, Z'_2, X'_3, Y'_3, Z'_3, X'_4, Y'_4, Z'_4] \\
= [X'_3, Y'_3, Z'_3, X'_1, Y'_1, Z'_1, X'_2, Y'_2, Z'_2, X'_4, Y'_4, Z'_4] \\
= [X_3, Y_3, Z_3, X_2, Y_2, Z_2, X_1, Y_1, Z_1, X_4, Y_4, Z_4] \\
= (31) [X_1, Y_1, Z_1, X_2, Y_2, Z_2, X_3, Y_3, Z_3, X_4, Y_4, Z_4]
\end{aligned} \tag{A1.4.51}$$

where X'_i, Y'_i, Z'_i are the coordinates of the nuclei after applying the operation (12). The result in (equation A1.4.51) is in accord with (equation A1.4.14).

Molecular wavefunctions are functions of the coordinates of the nuclei and electrons in a molecule, and we are now going to consider how such functions can be transformed by the general symmetry operation R as defined in (equation A1.4.49). To do this we introduce three functions of the coordinates, $f(X_p, Y_p, Z_p)$, $f_N^R(X_p, Y_p, Z_p)$ and $f_S^R(X_p, Y_p, Z_p)$. The functions f_N^R and f_S^R are such that their values at any point in configuration space are each related to the value of the function f at another point in configuration space, where the coordinates of this 'other' point are defined by the effect of R as follows:

$$f_N^R(X_i, Y_i, Z_i) = f(X'_i, Y'_i, Z'_i) \tag{A1.4.52}$$

-19-

and

$$f_S^R(X'_i, Y'_i, Z'_i) = f(X_i, Y_i, Z_i) \tag{A1.4.53}$$

or equivalently,

$$f_N^R(X_i, Y_i, Z_i) = f(RX_i, RY_i, RZ_i) \tag{A1.4.54}$$

and

$$f_S^R(X_i, Y_i, Z_i) = f(R^{-1}X_i, R^{-1}Y_i, R^{-1}Z_i). \tag{A1.4.55}$$

This means that f_N^R is such that its value at any point (X_p, Y_p, Z_p) is the same as the value of f at the point (RX_p, RY_p, RZ_p) , and that f_S^R is such that its value at any point (X_p, Y_p, Z_p) is the same as the value of f at the point $(R^{-1}X_p, R^{-1}Y_p, R^{-1}Z_p)$. Alternatively, for the latter we can say that f_S^R is such that its value at (RX_p, RY_p, RZ_p) is the same as the value of f at the point (X_p, Y_p, Z_p) .

We define the effect of a symmetry operation on a wavefunction in two different ways depending on whether the symmetry operation concerned uses a moving or fixed 'reference frame' (see [4]). Either we define its effect using the equation

$$Rf(X_i, Y_i, Z_i) = f_N^R(X_i, Y_i, Z_i) = f(RX_i, RY_i, RZ_i), \tag{A1.4.56}$$

or we define its effect using

$$Rf(X_i, Y_i, Z_i) = f_S^R(X_i, Y_i, Z_i) = f(R^{-1}X_i, R^{-1}Y_i, R^{-1}Z_i). \quad (\text{A1.4.57})$$

Nuclear permutations in the N-convention (which convention we always use for nuclear permutations) and rotation operations relative to a nuclear-fixed or molecule-fixed reference frame, are defined to transform wavefunctions according to (equation A1.4.56). These symmetry operations involve a moving reference frame. Nuclear permutations in the S-convention, point group operations in the space-fixed axis convention (which is the convention that is always used for point group operations; see [section A1.4.2.2](#) and rotation operations relative to a space-fixed frame are defined to transform wavefunctions according to (equation A1.4.57). These operations involve a fixed reference frame.

Another distinction we make concerning symmetry operations involves the *active* and *passive* pictures. Below we consider translational and rotational symmetry operations. We describe these operations in a space-fixed axis system (X, Y, Z) with axes parallel to the (X, Y, Z) axes, but with the origin fixed in space. In the active picture, which we adopt here, a translational symmetry operation displaces all nuclei and electrons in the molecule along a vector \mathbf{A} , say,

-20-

and leaves the (X, Y, Z) axis system unaffected. In the passive picture, the molecule is left unaffected but the (X, Y, Z) axis system is displaced by $-\mathbf{A}$. Similarly, in the active picture a rotational symmetry operation physically rotates the molecule, leaving the axis system unaffected, whereas in the passive picture the axis system is rotated and the molecule is unaffected. If we think about symmetry operations in the passive picture, it is immediately obvious that they must leave the Hamiltonian invariant (i.e., commute with it). The energy of an isolated molecule in field-free space is obviously unaffected if we translate or rotate the (X, Y, Z) axis system.

A1.4.2.5 THE LABELLING OF MOLECULAR ENERGY LEVELS

The irreducible representations of a symmetry group of a molecule are used to label its energy levels. The way we label the energy levels follows from an examination of the effect of a symmetry operation on the molecular Schrödinger equation.

$$\hat{H}\Psi_n(X_i, Y_i, Z_i) = E_n\Psi_n(X_i, Y_i, Z_i) \quad (\text{A1.4.58})$$

where $\Psi_n(X_i, Y_i, Z_i)$ is a molecular eigenfunction having eigenvalue E_n .

By definition, a symmetry operation R commutes with the molecular Hamiltonian \hat{H} and so we can write the operator equation:

$$\hat{H}R = R\hat{H}. \quad (\text{A1.4.59})$$

If we act with each side of this equation on an eigenfunction $\Psi_n(X_i, Y_i, Z_i)$ from (equation A1.4.58) we derive

$$\begin{aligned} \hat{H}R\Psi_n(X_i, Y_i, Z_i) &= R\hat{H}\Psi_n(X_i, Y_i, Z_i) = RE_n\Psi_n(X_i, Y_i, Z_i) \\ &= E_nR\Psi_n(X_i, Y_i, Z_i). \end{aligned} \quad (\text{A1.4.60})$$

The second equality follows from (equation A1.4.58)⁴, and the third equality from the fact that E_n is a number and numbers are not affected by symmetry operations. We can rewrite the result of (equation A1.4.60) as

$$\widehat{H}[R\Psi_n(X_i, Y_i, Z_i)] = E_n[R\Psi_n(X_i, Y_i, Z_i)]. \quad (\text{A1.4.61})$$

Thus

$$R\Psi_n(X_i, Y_i, Z_i) = \Psi_n^R(X_i, Y_i, Z_i) \quad (\text{A1.4.62})$$

is an eigenfunction having the same eigenvalue as $\Psi_n(X_i, Y_i, Z_i)$. If E_n is a nondegenerate eigenvalue then Ψ_n^R cannot be linearly independent of Ψ_n , which means that we can only have

-21-

$$R\Psi_n(X_i, Y_i, Z_i) = c\Psi_n(X_i, Y_i, Z_i) \quad (\text{A1.4.63})$$

where c is a constant. An arbitrary symmetry operation R is such that $R^m = E$ the identity, where m is an integer. From (equation A1.4.63) we deduce that

$$R^m\Psi_n(X_i, Y_i, Z_i) = c^m\Psi_n(X_i, Y_i, Z_i). \quad (\text{A1.4.64})$$

Since $R^m = E$ we must have $c^m = 1$ in (equation A1.4.64), which gives

$$c = \sqrt[m]{1}. \quad (\text{A1.4.65})$$

Thus, for example, for the PH_3 molecule any nondegenerate eigenfunction can only be multiplied by $+1$, $\omega = \exp(2\pi i/3)$, or $\omega^2 = \exp(4\pi i/3)$ by the symmetry operation (123) since $(123)^3 = E$ (so that $m = 3$ in (equation A1.4.65)). In addition, such a function can only be multiplied by $+1$ or -1 by the symmetry operations (12), (23) or (31) since each of these operations is self-reciprocal (so that $m = 2$ in (equation A1.4.65)).

We will apply this result to the H_2 molecule as a way of introducing the fact that nondegenerate molecular energy levels can be labelled according to the one-dimensional irreducible representations of a symmetry group of the molecular Hamiltonian. The Hamiltonian for the H_2 molecule commutes with E^* and with the operation (12) that permutes the protons. Thus, the eigenfunction of any nondegenerate molecular energy level is either invariant, or changed in sign, by the inversion operation E^* since $(E^*)^2 = E$ (i.e., $m = 2$ for $R = E^*$ in (equation A1.4.65)); invariant states are said to have positive parity (+) and states that are changed in sign by E^* to have negative parity (-). Similarly, any nondegenerate energy level will be invariant or changed in sign by the proton permutation operation (12); states that are invariant are said to be symmetric (s) with respect to (12) and states that are changed in sign are said to be antisymmetric (a). This enables us to label nondegenerate energy levels of the H_2 molecule as being (+ s), (- s), (+ a) or (- a) according to the effect of the operations E^* and (12). For the H_2 molecule we can form a symmetry group using these elements: $\{E, (12), E^*, (12)^*\}$, where

$$(12)^* = (12)E^* = E^*(12) \quad (\text{A1.4.66})$$

and the character table of the group is given in [table A1.4.4](#). The effect of the operation $(12)^*$ on a wavefunction is simply the product of the effects of (12) and E^* . The labelling of the states as (+ s), (- s), (+ a)

or $(-a)$ is thus according to the irreducible representations of the symmetry group and the nondegenerate energy levels of the H_2 molecule are of four different symmetry types in this group.

Table A1.4.4 The character table of a symmetry group for the H_2 molecule.

	E	(12)	E^*	$(12)^*$
$+s$	1	1	1	1
$-s$	1	1	-1	-1
$+a$	1	-1	1	-1
$-a$	1	-1	-1	1

The energy level of an l -fold degenerate eigenstate can be labelled according to an l -fold degenerate irreducible representation of the symmetry group, as we now show.

Suppose the l orthonormal⁵ eigenfunctions $\Psi_{n1}, \Psi_{n2}, \dots, \Psi_{nl}$ all have the same eigenvalue E_n of the molecular Hamiltonian. If we apply a symmetry operation R to one of these functions the resulting function will also be an eigenfunction of the Hamiltonian with eigenvalue E_n (see (equation A1.4.61) and the sentence after it) and the most general function of this type is a linear combination of the l functions Ψ_{ni} given above. Thus, using matrix notation, we can write the effect of R as⁶

$$R\Psi_{ni} = \sum_{j=1}^l D[R]_{ij} \Psi_{nj} \quad (\text{A1.4.67})$$

where $i = 1, 2, \dots, l$. For example, choosing $i = 1$, we have the effect of R on Ψ_{n1} as:

$$R\Psi_{n1} = D[R]_{11}\Psi_{n1} + D[R]_{12}\Psi_{n2} + \dots + D[R]_{1l}\Psi_{nl}. \quad (\text{A1.4.68})$$

The $D[R]_{ij}$ are numbers and $D[R]$ is a matrix of these numbers; the matrix $D[R]$ is *generated* by the effect of R on the l functions Ψ_{ni} . We can visualize (equation A1.4.67) as the effect of R acting on a column matrix Ψ_n being equal to the product of a square matrix $D[R]$ and a column matrix Ψ_n , i.e.,

$$R[\Psi_n] = [D[R]][\Psi_n]. \quad (\text{A1.4.69})$$

Each operation in a symmetry group of the Hamiltonian will generate such an $l \times l$ matrix, and it can be shown (see, for example, appendix 6-1 of [1]) that if three operations of the group P_1, P_2 and P_{12} are related by

$$P_1 P_2 = P_{12} \quad (\text{A1.4.70})$$

then the matrices generated by application of them to the Ψ_{ni} (as described by (equation A1.4.67)) will satisfy

$$D[P_1]D[P_2] = D[P_{12}]. \quad (\text{A1.4.71})$$

Thus, the matrices will have a multiplication table with the same structure as the multiplication table of the symmetry group and hence will form an l -dimensional representation of the group.

-23-

A given l -fold degenerate state can generate a reducible or an irreducible l -dimensional representation of the symmetry group considered. If the representation is irreducible then the degeneracy is said to be *necessary*, i.e., imposed symmetry of the Hamiltonian. However, if the representation is reducible then the degeneracy between the different states is said to be accidental and it is not imposed by the symmetry of the Hamiltonian. The occurrence of accidental degeneracy can indicate that some other symmetry operation has been forgotten, or paradoxically it can indicate many symmetry operations (called *unfeasible* symmetry operations in [section A1.4.4](#)) have been introduced.

These considerations mean that, for example, using the symmetry group S_3 for the PH_3 molecule (see [table A1](#)) energy levels are determined to be of symmetry type A_1 , A_2 or E . In molecular physics the labelling of molecular energy levels according to the irreducible representations of a symmetry group is mainly what we use symmetry for. Once we have labelled the energy levels of a molecule, we can use the labels to determine which of the levels can interact with each other as the result of adding a term \hat{H}' to the molecular Hamiltonian. This term could be the result of applying an external perturbation such as an electric or magnetic field, it could be the result of including a previously unaccounted term from the Hamiltonian, or this term could result from the effect of shining electromagnetic radiation through the molecules. In this latter case the symmetry labels enable us to determine the selection rules for allowed transitions in the spectrum of the molecule. All this becomes possible by making use of the *vanishing integral rule*.

A1.4.2.6 THE VANISHING INTEGRAL RULE

To explain the vanishing integral rule we first have to explain how we determine the symmetry of a product. Given an s -fold degenerate state of energy E_n and symmetry Γ_n , with eigenfunctions $\Phi_{n1}, \Phi_{n2}, \dots, \Phi_{ns}$, and an r -fold degenerate state of energy E_m and symmetry Γ_m , with eigenfunctions $\Phi_{m1}, \Phi_{m2}, \dots, \Phi_{mr}$, we wish to determine the symmetry of the set of functions $\Psi_{ij} = \Phi_{ni} \Phi_{mj}$, where $i = 1, 2, \dots, s$ and $j = 1, 2, \dots, r$. There will be $s \times r$ functions of the type Ψ_{ij} . The matrices D^{Γ_n} and D^{Γ_m} in the representations Γ_n and Γ_m , respectively, are obtained from (see [\(equation A1.4.67\)](#))

$$R\Phi_{ni} = \sum_{k=1}^s D^{\Gamma_n}[R]_{ik} \Phi_{nk}$$

and

$$R\Phi_{mj} = \sum_{l=1}^r D^{\Gamma_m}[R]_{jl} \Phi_{ml}$$

where R is an operation of the symmetry group. To obtain the matrices in the representation Γ_{nm} we write

$$R[\Phi_{ni} \Phi_{mj}] = \sum_{k=1}^s \sum_{l=1}^r D^{\Gamma_n}[R]_{ik} D^{\Gamma_m}[R]_{jl} \Phi_{nk} \Phi_{ml}$$

and we can write this as

$$R\Psi_{ij} = \sum_{k=1}^s \sum_{l=1}^r D^{\Gamma_{nm}}[R]_{ij,kl} \Psi_{kl}.$$

From this we see that the $s \times r$ dimensional representation Γ_{nm} generated by the $s \times r$ functions Ψ_{ij} has matrix elements given by

$$D^{\Gamma_{nm}}[R]_{ij,kl} = D^{\Gamma_n}[R]_{ik} D^{\Gamma_m}[R]_{jl}$$

where each element of $D^{\Gamma_{nm}}$ is indexed by a row label ij and a column label kl , each of which runs over $s \times r$ values. The diagonal element is given by

$$D^{\Gamma_{nm}}[R]_{ij,ij} = D^{\Gamma_n}[R]_{ii} D^{\Gamma_m}[R]_{jj}$$

and the character of the matrix is given by

$$\begin{aligned} \chi^{\Gamma_{nm}}[R] &= \sum_{k=1}^s \sum_{l=1}^r D^{\Gamma_{nm}}[R]_{ij,ij} = \sum_{k=1}^s \sum_{l=1}^r D^{\Gamma_n}[R]_{ii} D^{\Gamma_m}[R]_{jj} \\ &= \chi^{\Gamma_n}[R] \chi^{\Gamma_m}[R]. \end{aligned}$$

We can therefore calculate the character, under a symmetry operation R , in the representation generated by the product of two sets of functions, by multiplying together the characters under R in the representations generated by each set of functions. We write Γ_{nm} symbolically as

$$\Gamma_{nm} = \Gamma_n \otimes \Gamma_m$$

where the characters satisfy (equation A1.4.78) in which usual algebraic multiplication is used. Knowing the character in Γ_{nm} from (equation A1.4.78) we can then reduce the representation to its irreducible components using (equation A1.4.47). Suppose Γ_{nm} can be reduced to irreducible representations Γ_1, Γ_2 and Γ_3 according to

$$\Gamma_n \otimes \Gamma_m = 3\Gamma_1 \oplus \Gamma_2 \oplus 2\Gamma_3.$$

In this circumstance we say that Γ_{nm} contains Γ_1, Γ_2 and Γ_3 ; since $\Gamma_n \otimes \Gamma_m$ contains Γ_1 , for example, we write

$$\Gamma_n \otimes \Gamma_m \supset \Gamma_1.$$

Suppose that we can write the total Hamiltonian as $\hat{H} = \hat{H}^0 + \hat{H}'$, where \hat{H}' is a perturbation. Let us further suppose that the Hamiltonian \hat{H}^0 (\hat{H}' having been neglected) has normalized eigenfunctions Ψ_m^0 and Ψ_n^0 , with eigenvalues E_m^0 and E_n^0 , respectively, and that \hat{H}^0 commutes with the group of symmetry operations $G = \{R_1, R_2, \dots, R_h\}$. \hat{H}^0 will transform as the totally symmetric representation $\Gamma^{(s)}$ of G , and we let Ψ_m^0, Ψ_n^0 and \hat{H}'

generate the representations Γ_m , Γ_n and Γ' of G , respectively. The complete set of eigenfunctions of \hat{H}^0 forms a basis set for determining the eigenfunctions and eigenvalues of the Hamiltonian $\hat{H} = \hat{H}^0 + \hat{H}'$ and the Hamiltonian matrix H in this basis set is a matrix with elements H_{mn} given by the integrals

$$H_{mn} = \int \Psi_m^{0*} (\hat{H}^0 + \hat{H}') \Psi_n^0 d\tau = \delta_{mn} E_n^0 + H'_{mn} \quad (\text{A1.4.82})$$

where

$$H'_{mn} = \int \Psi_m^{0*} \hat{H}' \Psi_n^0 d\tau. \quad (\text{A1.4.83})$$

The eigenvalues E of \hat{H} can be determined from the Hamiltonian matrix by solving the secular equation

$$|H_{mn} - \delta_{mn} E| = 0. \quad (\text{A1.4.84})$$

In solving the secular equation it is important to know which of the off-diagonal matrix elements H'_{mn} vanish since this will enable us to simplify the equation.

We can use the symmetry labels Γ_m and Γ_n on the levels E_m^0 and E_n^0 , together with the symmetry Γ' of \hat{H}' , to determine which H'_{mn} elements must vanish. The function $\Psi_m^{0*} \hat{H}' \Psi_n^0$ generates the product representation $\Gamma_m^* \otimes \Gamma' \otimes \Gamma_n = \Gamma'_{mn}$ (Ψ_m^{0*} has symmetry Γ_m^*). We can now state *the vanishing integral rule*⁷: the matrix element

$$\int \Psi_m^{0*} \hat{H}' \Psi_n^0 d\tau = 0 \quad (\text{A1.4.85})$$

if

$$\Gamma_m^* \otimes \Gamma' \otimes \Gamma_n \not\supset \Gamma^{(s)}$$

(A1.4.86)

-26-

where $\Gamma^{(s)}$ is the totally symmetric representation. If \hat{H}' is totally symmetric in G then H'_{mn} will vanish if

$$\Gamma_m^* \otimes \Gamma_n \not\supset \Gamma^{(s)} \quad (\text{A1.4.87})$$

i.e., if

$$\Gamma_m \neq \Gamma_n. \quad (\text{A1.4.88})$$

It would be an accident if H'_{mn} vanished even though $\Gamma'_{mn} \supset \Gamma^{(s)}$, but if this were the case it might well indicate that there is some extra unconsidered symmetry present.

The value of the vanishing integral rule is that it allows the matrix H to be block diagonalized. This occurs if

we order the eigenfunctions Ψ_n^0 according to their symmetry when we set up H . Let us initially consider the case when $\Gamma' = \Gamma^{(s)}$. In this case all off-diagonal matrix elements between Ψ_n^0 basis functions of different symmetry will vanish, and the Hamiltonian matrix will block diagonalize with there being one block for each symmetry type of Ψ_n^0 function. Each eigenfunction of \hat{H} will only be a linear combination of Ψ_n^0 functions having the same symmetry in G (G being the symmetry group of \hat{H}^0). Thus the symmetry of each eigenfunction Ψ_j of \hat{H} in the group G will be the same as the symmetry of the Ψ_n^0 basis functions that make it up (G is a symmetry group of \hat{H} when $\Gamma' = \Gamma^{(s)}$) and each block of a block diagonal matrix can be diagonalized separately, which is a great simplification. The symmetry of the Ψ_j functions can be obtained from the symmetry of the Ψ_n^0 functions without worrying about the details of \hat{H} and this is frequently very useful. When $\Gamma' \neq \Gamma^{(s)}$ all off-diagonal matrix elements between Ψ^0 functions of symmetry Γ_m and Γ_n will vanish if (equation A1.4.87) is satisfied, and there will also be a block diagonalization of H (it will be necessary to rearrange the rows or columns of H , i.e., to rearrange the order of the Ψ_n^0 functions, to obtain H in block diagonal form). However, now nonvanishing matrix elements occur in H that connect Ψ_n^0 functions of different symmetry in G and as a result the eigenfunctions of \hat{H} may not contain only functions of one symmetry type of G ; when $\Gamma' \neq \Gamma^{(s)}$ the group G is not a symmetry group of \hat{H} and its eigenfunctions Ψ_j cannot be classified in G . However, the classification of the basis functions Ψ_n^0 in G will still allow a simplification of the Hamiltonian matrix.

The vanishing integral rule is not only useful in determining the nonvanishing elements of the Hamiltonian matrix H . Another important application is the derivation of *selection rules* for transitions between molecular states. For example, the intensity of an electric dipole transition from a state with wavefunction $\Psi_{j'}^{(F', m'_{F'})}$ to a state with wavefunction $\Psi_{j''}^{(F'', m''_{F''})}$ (see (equation A1.4.12)) is proportional to the quantity

$$|T|^2 = \left| \int \Psi_{j'}^{(F', m'_{F'})*} \mu_A \Psi_{j''}^{(F'', m''_{F''})} d\tau \right|^2 \quad (\text{A1.4.89})$$

-27-

where μ_A , $A = X, Y, Z$, is the component of the molecular dipole moment operator along the A axis. If $\Psi_{j'}^{(F', m'_{F'})}$ and $\Psi_{j''}^{(F'', m''_{F''})}$ belong to the irreducible representations $\Gamma_{j'}^{(F', m'_{F'})}$ and $\Gamma_{j''}^{(F'', m''_{F''})}$, respectively and μ_A has the symmetry $\Gamma(\mu_A)$, then $|T|^2$, and thus the intensity of the transition, vanishes unless

$$\Gamma_{j'}^{(F', m'_{F'})*} \otimes \Gamma(\mu_A) \otimes \Gamma_{j''}^{(F'', m''_{F''})} \supset \Gamma^{(s)}. \quad (\text{A1.4.90})$$

In the rotational symmetry group K (spatial) discussed in section A1.4.1.1, we have $\Gamma_{j'}^{(F', m'_{F'})} = D^{(F')}$, $\Gamma_{j''}^{(F'', m''_{F''})} = D^{(F'')}$ and $\Gamma(\mu_A) = D^{(1)}$. In this case the application of the vanishing integral rule leads to the selection rule given in (equation A1.4.13) (see section 7.3.2, in particular equation (7-47), of [1]).

A1.4.3 SYMMETRY OPERATIONS AND SYMMETRY GROUPS

The various types of symmetry enumerated in section A1.4.1.2 are discussed in detail here and the symmetry groups containing such symmetry operations are presented.

A1.4.3.1 TRANSLATIONAL SYMMETRY

In the active picture adopted here the (X, Y, Z) axis system remains fixed in space and a translational symmetry operation changes the (X, Y, Z) coordinates of all nuclei and electrons in the molecule by constant amounts, $(\Delta X, \Delta Y, \Delta Z)$ say,

$$(X_i, Y_i, Z_i) \rightarrow (X_i + \Delta X, Y_i + \Delta Y, Z_i + \Delta Z). \quad (\text{A1.4.91})$$

We obtain a coordinate set more suitable for describing translational symmetry by introducing the centre of mass coordinates

$$(X_0, Y_0, Z_0) = \left(\frac{1}{M} \sum_{i=1}^l m_i X_i, \frac{1}{M} \sum_{i=1}^l m_i Y_i, \frac{1}{M} \sum_{i=1}^l m_i Z_i \right) \quad (\text{A1.4.92})$$

together with

$$(X_i, Y_i, Z_i) = (X_i - X_0, Y_i - Y_0, Z_i - Z_0) \quad (\text{A1.4.93})$$

for each particle i , where there are l particles in the molecule (N nuclei and $l - N$ electrons), m_i is the mass of particle i and $M = \sum_{i=1}^l m_i$ is the total mass of the molecule. In this manner we have introduced a new axis system (X, Y, Z)

-28-

with axes parallel to the (X, Y, Z) axes but with origin at the molecular centre of mass. The molecule is described by the $3l$ coordinates

$$X_0, Y_0, Z_0, X_2, Y_2, Z_2, X_3, Y_3, Z_3, \dots, X_l, Y_l, Z_l.$$

The coordinates (X_1, Y_1, Z_1) are redundant since they can be determined from the condition that the (X, Y, Z) axis system has origin at the molecular centre of mass. Obviously, the translational symmetry operation discussed above has the effect of changing the centre of mass coordinates

$$(X_0, Y_0, Z_0) \rightarrow (X_0 + \Delta X, Y_0 + \Delta Y, Z_0 + \Delta Z) \quad (\text{A1.4.94})$$

whereas the coordinates $X_2, Y_2, Z_2, X_3, Y_3, Z_3, \dots, X_l, Y_l, Z_l$ are unchanged by this operation.

We now define the effect of a translational symmetry operation on a function. [Figure A1.4.3](#) shows how a PH_3 molecule is displaced a distance ΔX along the X axis by the translational symmetry operation that changes X_0 to $X'_0 = X_0 + \Delta X$. Together with the molecule, we have drawn a sine wave symbolizing the molecular wavefunction, Ψ_j , say. We have marked one wavecrest to better keep track of the way the function is displaced by the symmetry operation. For the physical situation to be unchanged by the symmetry operation, the marked wavecrest and thus the entire wavefunction, is displaced by ΔX along the X axis as shown in [Figure A1.4.3](#). Thus, an operator $R_T^{(\Delta X, \Delta Y, \Delta Z)}$, which describes the effect of the translational symmetry operation on a wavefunction, is defined according to the S-convention (see [equation A1.4.57](#))

(A1.4.95)

$$R_T^{(\Delta X, \Delta Y, \Delta Z)} \Psi_j(X_0, Y_0, Z_0, X_2, Y_2, Z_2, X_3, Y_3, Z_3, \dots, X_l, Y_l, Z_l) \\ = \Psi_j(X_0 - \Delta X, Y_0 - \Delta Y, Z_0 - \Delta Z, X_2, Y_2, Z_2, X_3, Y_3, Z_3, \dots, X_l, Y_l, Z_l).$$

This definition causes the wavefunction to ‘move with the molecule’ as shown for the X direction in [figure A1.4.3](#). The set of all translation symmetry operations $R_T^{(\Delta X, \Delta Y, \Delta Z)}$ constitutes a group which we call the translational group G_T . Because of the uniformity of space, G_T is a symmetry group of the molecular Hamiltonian \hat{H} in that all its elements commute with \hat{H} :

$$[R_T^{(\Delta X, \Delta Y, \Delta Z)}, \hat{H}] = 0. \quad (\text{A1.4.96})$$

We could stop here in the discussion of the translational group. However, for the purpose of understanding the relation between translational symmetry and the conservation of linear momentum, we now show how the operator $R_T^{(\Delta X, \Delta Y, \Delta Z)}$ can be expressed in terms of the quantum mechanical operators representing the translational linear momentum of the molecule; these operators are defined as

$$(\hat{P}_X, \hat{P}_Y, \hat{P}_Z) = \left(-i\hbar \frac{\partial}{\partial X_0}, -i\hbar \frac{\partial}{\partial Y_0}, -i\hbar \frac{\partial}{\partial Z_0} \right). \quad (\text{A1.4.97})$$

-29-

The translational linear momentum is conserved for an isolated molecule in field free space and, as we see below, this is closely related to the fact that the molecular Hamiltonian commutes with $R_T^{(\Delta X, \Delta Y, \Delta Z)}$ for all values of $(\Delta X, \Delta Y, \Delta Z)$. The conservation of linear momentum and translational symmetry are directly related.

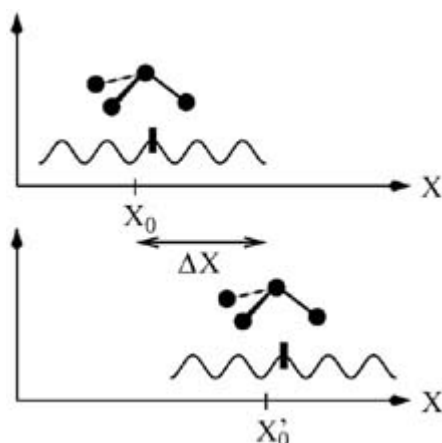


Figure A1.4.3. A PH_3 molecule and its wavefunction, symbolized by a sine wave, before (top) and after (bottom) a translational symmetry operation.

In order to determine the relationship between $R_T^{(\Delta X, \Delta Y, \Delta Z)}$ and the $(\hat{P}_X, \hat{P}_Y, \hat{P}_Z)$ operators, we consider a translation $R_T^{(\delta X, 0, 0)}$ where ΔX is infinitesimally small. In this case we can approximate the right hand side of (equation A1.4.95) by a first-order Taylor expansion:

$$\begin{aligned}
R_T^{(\delta X, 0, 0)} \Psi_j(X_0, Y_0, Z_0, X_2, Y_2, Z_2, \dots, X_l, Y_l, Z_l) \\
&= \Psi_j(X_0 - \delta X, Y_0, Z_0, X_2, Y_2, Z_2, \dots, X_l, Y_l, Z_l) \\
&= \Psi_j(X_0, Y_0, Z_0, X_2, Y_2, Z_2, \dots, X_l, Y_l, Z_l) - \frac{\partial \Psi_j}{\partial X_0} \delta X.
\end{aligned} \tag{A1.4.98}$$

From the definition of the translational linear momentum operator \hat{P}_X (in [equation A1.4.97](#)) we see that

$$\frac{\partial \Psi_j}{\partial X_0} = \frac{i}{\hbar} \hat{P}_X \Psi_j \tag{A1.4.99}$$

and by introducing this identity in [equation A1.4.98](#) we obtain

$$R_T^{(\delta X, 0, 0)} \Psi_j = \Psi_j - \frac{i}{\hbar} \delta X \hat{P}_X \Psi_j \tag{A1.4.100}$$

-30-

where we have omitted the coordinate arguments for brevity. Since the function Ψ_j in [equation A1.4.100](#) is arbitrary, it follows that we can write the symmetry operation as

$$R_T^{(\delta X, 0, 0)} = 1 - \frac{i}{\hbar} \delta X \hat{P}_X. \tag{A1.4.101}$$

The operation $R_T^{(\Delta X, 0, 0)}$, for which ΔX is an arbitrary finite length, obviously has the same effect on a wavefunction as the operation $R_T^{(\delta X, 0, 0)}$ applied to the wavefunction $\Delta X/\delta X$ times. We simply divide the translation by ΔX into $\Delta X/\delta X$ steps, each step of length δX . This remains true in the limit of $\delta X \rightarrow 0$. Thus

$$R_T^{(\Delta X, 0, 0)} = \lim_{\delta X \rightarrow 0} (R_T^{(\delta X, 0, 0)})^{\frac{\Delta X}{\delta X}} = \lim_{\delta X \rightarrow 0} \left(1 - \frac{i}{\hbar} \delta X \hat{P}_X \right)^{\frac{\Delta X}{\delta X}} = \exp \left(-\frac{i}{\hbar} \Delta X \hat{P}_X \right) \tag{A1.4.102}$$

where we have used the general identity

$$\lim_{x \rightarrow 0} (1 + ax)^{y/x} = \exp(ay). \tag{A1.4.103}$$

We can derive expressions analogous to [equation A1.4.102](#) for $R_T^{(0, \Delta Y, 0)}$ and $R_T^{(0, 0, \Delta Z)}$ and we can resolve a general translation $R_T^{(\Delta X, \Delta Y, \Delta Z)}$ as

$$R_T^{(\Delta X, \Delta Y, \Delta Z)} = R_T^{(\Delta X, 0, 0)} R_T^{(0, \Delta Y, 0)} R_T^{(0, 0, \Delta Z)}. \tag{A1.4.104}$$

Consequently,

$$\tag{A1.4.105}$$

$$R_T^{(\Delta X, \Delta Y, \Delta Z)} = \exp \left[-\frac{i}{\hbar} (\Delta X \hat{P}_X + \Delta Y \hat{P}_Y + \Delta Z \hat{P}_Z) \right].$$

We deal with the exponentials in (equation A1.4.102) and (equation A1.4.105) whose arguments are operators by using their Taylor expansion

$$\exp(i\hat{O}) = 1 + i\hat{O} + \frac{1}{2!}(i\hat{O})^2 + \dots \quad (\text{A1.4.106})$$

where \hat{O} is a Hermitian operator.

-31-

It follows from (equation A1.4.105) that (equation A1.4.96) is satisfied for arbitrary $\Delta X, \Delta Y, \Delta Z$ if and only if

$$[\hat{H}, \hat{P}_X] = [\hat{H}, \hat{P}_Y] = [\hat{H}, \hat{P}_Z] = 0. \quad (\text{A1.4.107})$$

From the fact that \hat{H} commutes with the operators ($\hat{P}_X, \hat{P}_Y, \hat{P}_Z$) it is possible to show that the linear momentum of a molecule in free space must be conserved. First we note that the time-dependent wavefunction $\Psi(t)$ of a molecule fulfills the time-dependent Schrödinger equation

$$i\hbar \frac{\partial \Psi(t)}{\partial t} = \hat{H} \Psi(t). \quad (\text{A1.4.108})$$

For $A = X, Y, \text{ or } Z$, we use this identity to derive an expression for

$$\begin{aligned} \frac{\partial}{\partial t} \langle \Psi(t) | \hat{P}_A | \Psi(t) \rangle &= \left\langle \frac{\partial \Psi(t)}{\partial t} | \hat{P}_A | \Psi(t) \right\rangle + \left\langle \Psi(t) | \frac{\partial}{\partial t} (\hat{P}_A \Psi(t)) \right\rangle \\ &= \left\langle \frac{\partial \Psi(t)}{\partial t} | \hat{P}_A | \Psi(t) \right\rangle + \left\langle \Psi(t) | \hat{P}_A | \frac{\partial \Psi(t)}{\partial t} \right\rangle \end{aligned} \quad (\text{A1.4.109})$$

where, in the last equality, we have used the fact that \hat{P}_A does not depend explicitly on t . We obtain $\partial \Psi(t) / \partial t$ from (equation A1.4.108) and insert the resulting expression in (equation A1.4.109); this yields

$$\begin{aligned} \frac{\partial}{\partial t} \langle \Psi(t) | \hat{P}_A | \Psi(t) \rangle &= \frac{i}{\hbar} (\langle \hat{H} \Psi(t) | \hat{P}_A | \Psi(t) \rangle - \langle \Psi(t) | \hat{P}_A | \hat{H} \Psi(t) \rangle) \\ &= \frac{i}{\hbar} \langle \Psi(t) | [\hat{H}, \hat{P}_A] | \Psi(t) \rangle = 0 \end{aligned} \quad (\text{A1.4.110})$$

where we have used (equation A1.4.107) in conjunction with the fact that \hat{H} is Hermitian. (Equation A1.4.110) shows that the expectation value of each linear momentum operator is conserved in time and thus the conservation of linear momentum directly follows from the translational invariance of the molecular Hamiltonian ((equation A1.4.96)).

Because of (equation A1.4.107) and because of the fact that \hat{P}_X , \hat{P}_Y and \hat{P}_Z commute with each other, we know that there exists a complete set of simultaneous eigenfunctions of \hat{P}_X , \hat{P}_Y , \hat{P}_Z and \hat{H} . An eigenfunction of \hat{P}_X , \hat{P}_Y and \hat{P}_Z has the form

$$\Psi_T(X_0, Y_0, Z_0) = \exp[i(k_X X_0 + k_Y Y_0 + k_Z Z_0)] \quad (\text{A1.4.111})$$

-32-

where

$$\hat{P}_A \Psi_T(X_0, Y_0, Z_0) = \hbar k_A \Psi_T(X_0, Y_0, Z_0) \quad (\text{A1.4.112})$$

with A = X, Y or Z, so that ((equation A1.4.105))

$$R_T^{(\Delta X, \Delta Y, \Delta Z)} \Psi_T(X_0, Y_0, Z_0) = \exp[-i(\Delta X k_X + \Delta Y k_Y + \Delta Z k_Z)] \Psi_T(X_0, Y_0, Z_0). \quad (\text{A1.4.113})$$

That is, the effect of a translational operation is determined solely by the vector with components (k_X, k_Y, k_Z) which defines the linear momentum.

For a molecular wavefunction $\Psi_j(X_0, Y_0, Z_0, X_2, Y_2, Z_2, \dots, X_l, Y_l, Z_l)$ to be a simultaneous eigenfunction of \hat{P}_X , \hat{P}_Y , \hat{P}_Z and \hat{H} it must have the form

$$\Psi_j(X_0, Y_0, Z_0, X_2, Y_2, Z_2, \dots, X_l, Y_l, Z_l) = \Psi_T(X_0, Y_0, Z_0) \Psi_{\text{int}}(X_2, Y_2, Z_2, \dots, X_l, Y_l, Z_l) \quad (\text{A1.4.114})$$

where Ψ_{int} describes the *internal* motion of the molecule (see also section 7.3.1 of [1]).

We can describe the conservation of linear momentum by noting the analogy between the time-dependent Schrödinger equation, (equation A1.4.108), and (equation A1.4.99). For an isolated molecule, \hat{H} does not depend explicitly on t and we can repeat the arguments expressed in (equation A1.4.98), (equation A1.4.99), (equation A1.4.100), (equation A1.4.101) and (equation A1.4.102) with X replaced by t and \hat{P}_X replaced by $-\hat{H}$ to show that

$$\Psi(t) = \exp\left(\frac{i}{\hbar} t \hat{H}\right) \Psi(t=0). \quad (\text{A1.4.115})$$

If the wavefunction at $t=0$, $\Psi(t=0)$, is an eigenfunction of \hat{P}_X , \hat{P}_Y , \hat{P}_Z and \hat{H} so that it can be expressed as given in (equation A1.4.114), it follows from (equation A1.4.115) that at any other time t ,

$$\Psi(t) = \exp\left(\frac{i}{\hbar} t E\right) \Psi(t=0) \quad (\text{A1.4.116})$$

where E is the energy (i.e., the eigenvalue of \hat{H} associated with the eigenfunction $\Psi(t=0)$). It is straightforward to show that this function is an eigenfunction of \hat{P}_X , \hat{P}_Y , \hat{P}_Z and \hat{H} with the same eigenvalues as $\Psi(t=0)$. This is another way of proving that linear momentum and energy are conserved in time.

A1.4.3.2 ROTATIONAL SYMMETRY

In order to discuss rotational symmetry, we must first introduce the rotational and vibrational coordinates customarily used in molecular theory. We define a set of (x, y, z) axes with an orientation relative to the (X, Y, Z) axes discussed

-33-

above that is defined by the positions of the nuclei. These axes are called ‘molecule fixed’ axes; their orientation is determined by the coordinates of the nuclei only and the coordinates of the electrons are not involved. The (x, y, z) and (X, Y, Z) axis systems are always chosen to be right handed. For any placement of the N nuclei in space (i.e., any set of values for the $3N - 3$ independent coordinates X_i, Y_i and Z_i of the nuclei) there is an unambiguous way of specifying the orientation of the (x, y, z) axes with respect to the (X, Y, Z) axes. Three equations are required to define the three Euler angles (θ, ϕ, χ) (see figure A1.4.4 that specify this orientation and the equations used are the Eckart (equation A1.4.5). The Eckart equations minimize the angular momentum in the (x, y, z) axis system and so they optimize the separation of the rotational and vibrational degrees of freedom in the rotation–vibration Schrödinger equation. It is described in detail in chapter 10 of [1] how, by introducing the Eckart equations, we can define the (x, y, z) axis system and thus the Euler angles (θ, ϕ, χ) . Suffice it to say that we describe the internal motion of a nonlinear molecule⁸ by $3l - 3$ coordinates, where the first three are the Euler angles (θ, ϕ, χ) describing rotation, the next $3N - 6$ are *normal coordinates* $Q_1, Q_2, Q_3, \dots, Q_{3N-6}$ describing the vibration of the nuclei and the remaining $3(l - N)$ are electronic coordinates $x_{N+1}, y_{N+1}, z_{N+1}, x_{N+2}, y_{N+2}, z_{N+2}, \dots, x_l, y_l, z_l$ simply chosen as the Cartesian coordinates of the electrons in the (x, y, z) axis system.

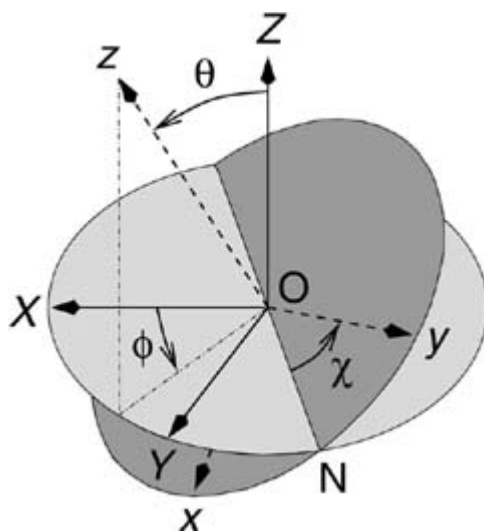


Figure A1.4.4. The definition of the Euler angles (θ, ϕ, χ) that relate the orientation of the molecule fixed (x, y, z) axes to the (X, Y, Z) axes. The origin of both axis systems is at the nuclear centre of mass O , and the node line ON is directed so that a right handed screw is driven along ON in its positive direction by twisting it from Z to z through θ where $0 \leq \theta \leq \pi$. ϕ and χ have the ranges 0 to 2π . χ is measured from the node line.

We consider rotations of the molecule about space-fixed axes in the active picture. Such a rotation causes the (x, y, z) axis system to rotate so that the Euler angles change

$$(\theta, \phi, \chi) \rightarrow (\theta + \Delta\theta, \phi + \Delta\phi, \chi + \Delta\chi). \quad (\text{A1.4.117})$$

The normal coordinates $Q_r, r = 1, 2, \dots, 3N - 6$, and the electronic coordinates $x_i, y_i, z_i, i = N + 1, N + 2, \dots, l$, all describe motion relative to the (x, y, z) axis system and are invariant to rotations.

Initially, we neglect terms depending on the electron spin \hat{S} and the nuclear spin \hat{I} in the molecular Hamiltonian \hat{H} . In this approximation, we can take the total angular momentum to be \hat{N} (see (equation A1.4.1)) which results from the rotational motion of the nuclei and the orbital motion of the electrons. The components of \hat{N} in the (X, Y, Z) axis system are given by:

$$\hat{N}_X = -i\hbar \left(-\sin \phi \frac{\partial}{\partial \theta} + \operatorname{cosec} \theta \cos \phi \frac{\partial}{\partial \chi} - \cot \theta \cos \phi \frac{\partial}{\partial \phi} \right) \quad (\text{A1.4.118})$$

$$\hat{N}_Y = -i\hbar \left(\cos \phi \frac{\partial}{\partial \theta} + \operatorname{cosec} \theta \sin \phi \frac{\partial}{\partial \chi} - \cot \theta \sin \phi \frac{\partial}{\partial \phi} \right) \quad (\text{A1.4.119})$$

and

$$\hat{N}_Z = -i\hbar \frac{\partial}{\partial \phi}. \quad (\text{A1.4.120})$$

By analogy with our treatment of translation symmetry, we aim to derive an operator $R_R^{(\Delta\theta, \Delta\phi, \Delta\chi)}$ which, when applied to a wavefunction, describes the effect of a general symmetry operation that causes the change in the Euler angles given in (equation A1.4.117). Because of the analogy between (equation A1.4.120) and the definition of \hat{P}_X in (equation A1.4.97), we can repeat the arguments expressed in (equation A1.4.98), (equation A1.4.99), (equation A1.4.100), (equation A1.4.101) and (equation A1.4.102) with X replaced by ϕ to show that

$$R_R^{(0, \Delta\phi, 0)} = \exp \left(-\frac{i}{\hbar} \Delta\phi \hat{N}_Z \right). \quad (\text{A1.4.121})$$

A more involved derivation (see, for example, section 3.2 of Zare [6]) shows that for a general rotation

$$R_R^{(\Delta\theta, \Delta\phi, \Delta\chi)} = \exp \left(-\frac{i}{\hbar} \Delta\phi \hat{N}_Z \right) \exp \left(-\frac{i}{\hbar} \Delta\theta \hat{N}_Y \right) \exp \left(-\frac{i}{\hbar} \Delta\chi \hat{N}_Z \right). \quad (\text{A1.4.122})$$

The operators \hat{N}_Y and \hat{N}_Z in (equation A1.4.122) do not commute and we have (see equation (10-90) of [1])

$$[\hat{N}_Y, \hat{N}_Z] = i\hbar \hat{N}_X. \quad (\text{A1.4.123})$$

The commutators $[\hat{N}_X, \hat{N}_Y]$ and $[\hat{N}_Z, \hat{N}_X]$ are obtained by replacing XYZ by ZXY and YZX, respectively, in (equation A1.4.123). It is, therefore, important in using (equation A1.4.122) that the exponential factors be applied in the correct order.

The set of all rotation operations $R_R^{(\Delta\theta, \Delta\phi, \Delta\chi)}$ forms a group which we call the rotational group \mathbf{K} (spatial). Since space is isotropic, \mathbf{K} (spatial) is a symmetry group of the molecular Hamiltonian \hat{H} in that all its elements commute with \hat{H} :

$$[R_R^{(\Delta\theta, \Delta\phi, \Delta\chi)} \hat{H}] = 0. \quad (\text{A1.4.124})$$

It follows from (equation A1.4.122) that (equation A1.4.124) is satisfied for arbitrary $(\Delta\theta, \Delta\phi, \Delta\chi)$ if and only if \hat{H} commutes with \hat{N}_Y and \hat{N}_Z . But then \hat{H} also commutes with \hat{N}_X because of (equation A1.4.123). That is

$$[\hat{H}, \hat{N}_X] = [\hat{H}, \hat{N}_Y] = [\hat{H}, \hat{N}_Z] = 0 \quad (\text{A1.4.125})$$

this equation is analogous to (equation A1.4.107). We discussed above (in connection with (equation A1.4.108), (equation A1.4.109) and (equation A1.4.110)) how the invariance of the molecular Hamiltonian to translation is related to the conservation of linear momentum. We now see that, in a completely analogous manner, the invariance of the molecular Hamiltonian to rotation is related to the conservation of angular momentum.

The (X, Y, Z) components of \hat{N} do not commute and so we cannot find simultaneous eigenfunctions of all the four operators occurring in (equation A1.4.125). It is straightforwardly shown from the commutation relations in (equation A1.4.123) that the operator

$$\hat{N}^2 = \hat{N}_X^2 + \hat{N}_Y^2 + \hat{N}_Z^2 \quad (\text{A1.4.126})$$

commutes with \hat{N}_X , \hat{N}_Y , and \hat{N}_Z . Because of (equation A1.4.125), this operator also commutes with \hat{H} . As a consequence, we can find simultaneous eigenfunctions of \hat{H} , \hat{N}^2 and one component of \hat{N} , customarily chosen as \hat{N}_Z . We can use this result to simplify the diagonalization of the matrix representation of the molecular Hamiltonian. We choose the basis functions as $R_R^{(\Delta\theta, \Delta\phi, \Delta\chi)}$. They are eigenfunctions of \hat{N}^2 (with eigenvalues $N(N+1)\hbar^2$, $N = 0, 1, 2, 3, 4, \dots$) and \hat{N}_Z (with eigenvalues $m\hbar$, $m = -N, -N+1, \dots, N-1, N$). The functions $\Psi_{n,N,m}^0$, $m = -N, -N+1, \dots, N-1, N$, transform according to the irreducible representation $D^{(N)}$ of K (spatial) (see section A1.4.1.1). With these basis functions, the matrix representation of the molecular Hamiltonian will be block diagonal in N and m in the manner described for the quantum numbers F and m_F in section A1.4.1.1.

If we allow for the terms in the molecular Hamiltonian depending on the electron spin \hat{S} (see chapter 7 of [1]), the resulting Hamiltonian no longer commutes with the components of \hat{N} as given in (equation A1.4.125), but with the components of

$$\hat{J} = \hat{N} + \hat{S}. \quad (\text{A1.4.127})$$

In this case, we choose the basis functions Ψ_{n,J,m_J}^0 , that is, the eigenfunctions of \hat{J}^2 (with eigenvalues $J(J+1)\hbar^2$, $J = |N-S|, |N-S|+1, \dots, N+S-1, N+S$) and \hat{N}_Z (with eigenvalues $m_J\hbar$, $m_J = -J, -J+1, \dots, J-1, J$). These functions are linear combinations of products $\Psi_{n,N,m}^0 \Psi_{e,S,m_S}^0$, where the function $\Psi_{n,N,m}^0$ is an eigenfunction of \hat{N}^2 and \hat{N}_Z as described above, and Ψ_{e,S,m_S}^0 is an eigenfunction of \hat{S}^2 (with eigenvalues $S(S+1)\hbar^2$, $S = 0, 1/2, 1, 3/2, 2, 5/2, 3, \dots$) and \hat{S}_Z (with eigenvalues $m_S\hbar$, $m_S = -S, -S+1, \dots, S-1, S$). In this

basis, the matrix representation of the molecular Hamiltonian is block diagonal in J and m_J . The functions $\Psi_{e,S,m_S}^0 = -S, -S+1, \dots, S-1, S$, transform according to the irreducible representation $D^{(S)}$ of K (spatial) and the functions $\Psi_{n,J,m_J}^0 = -J, -J+1, \dots, J-1, J$, have $D^{(J)}$ symmetry in K (spatial). Singlet states have $S = 0$ and for them $\hat{J} = \hat{N}$, $J = N$ and $m_J = m$.

Finally, we consider the complete molecular Hamiltonian which contains not only terms depending on the electron spin, but also terms depending on the nuclear spin \hat{F} (see chapter 7 of [1]). This Hamiltonian commutes with the components of \hat{F} given in (equation A1.4.1). The diagonalization of the matrix representation of the complete molecular Hamiltonian proceeds as described in section A1.4.1.1. The theory of rotational symmetry is an extensive subject and we have only scratched the surface here. A relatively new book, which is concerned with molecules, is by Zare [6] (see [7] for the solutions to all the problems in [6] and a list of the errors). This book describes, for example, the method for obtaining the functions Ψ_{n,J,m_J}^0 from $\Psi_{n,N,m}^0$ and Ψ_{e,S,m_S}^0 , and for obtaining the functions Ψ_{n,F,m_F}^0 (section A1.4.1.1) from the Ψ_{n,J,m_J}^0 combined with eigenfunctions of \hat{F}^2 and \hat{F}_Z .

A1.4.3.3 INVERSION SYMMETRY

We have already discussed inversion symmetry and how it leads to the parity label in section A1.4.1.2 and section A1.4.2.5. For any molecule in field-free space, if we neglect terms arising from the weak interaction force (see the next paragraph), the molecular Hamiltonian commutes with the inversion operation E^* and thus for such a Hamiltonian the *inversion group* $\mathcal{E} = \{E, E^*\}$ is a symmetry group. The character table of the inversion group is given in table A1.4.5 and the irreducible representations are labelled + and - to give the parity.

Table A1.4.5 The character table of the inversion group \mathcal{E}

	E	E^*
+	1	1
-	1	-1

Often molecular energy levels occur in closely spaced doublets having opposite parity. This is of particular interest when there are symmetrically equivalent minima, separated by a barrier, in the potential energy function of the electronic state under investigation. This happens in the PH_3 molecule and such pairs of levels are called ‘inversion doublets’; the splitting between such parity doublet levels depends on the extent of the quantum mechanical tunnelling through the barrier that separates the two minima. This is discussed further in section A1.4.4.

The Hamiltonian considered above, which commutes with E^* , involves the electromagnetic forces between the nuclei and electrons. However, there is another force between particles, the weak interaction force, that is not invariant to inversion. The weak charged current interaction force is responsible for the beta decay of nuclei, and the related weak neutral current interaction force has an effect in atomic and molecular systems. If we include this force between the nuclei and electrons in the molecular Hamiltonian (as we should because of electroweak unification) then the Hamiltonian will not commute with E^* , and states of opposite parity will be mixed. However, the effect of the weak neutral current interaction force is incredibly small (and it is a very short range force), although its effect has been detected in extremely precise experiments on atoms (see, for

example, Wood *et al* [8], who detect that a small part ($\sim 10^{-11}$) of a P state of caesium is mixed into an S state by this force). Its effect has not been detected in a molecule and, thus, for practical purposes we can neglect it and consider E^* to be a symmetry operation. Note that inversion symmetry is not a universal symmetry like translational or rotational symmetry and it does not derive from a general property of space. In the theoretical physics community, when dealing with particle symmetry, the inversion operation is called the ‘parity operator’ P .

An optically active molecule is a particular type of molecule in which there are two equivalent minima separated by an insuperable barrier in the potential energy surface and for which the molecular structures at these two minima are not identical (as they are in PH_3) but are mirror images of one another. The two forms of the molecule are called the dextrorotatory (D) and laevorotatory (L) forms and they can be separated. The D and L wavefunctions are not eigenfunctions of E^* and E^* interconverts them. In the general case eigenstates of the Hamiltonian are eigenstates of E^* and they have a definite parity. In the laboratory, when one makes an optically active molecule one obtains a racemic 50/50 mixture of the D and L forms, but in living organisms use is made of only one isomer; natural proteins, for example, are composed exclusively of L-amino acids, whereas nucleic acids contain only D-sugars. This fact is unexplained but it has been pointed out (see [9] and references therein) that in the molecular Hamiltonian the weak neutral current interaction term \hat{H}_{WI} would give rise to a small energy difference between the energy levels of the D and L forms, and this small energy difference could have acted to select one isomer over the long time of prebiotic evolution. The experimental determination of the energy difference between the D and L forms of any optically active molecule has yet to be achieved. However, see Daussy C, Marrel T, Amy-Klein A, Nguyen C T, Bordé C J and Chardonnet C 1999 *Phys. Rev. Lett.* **83** 1554 for a recent determination of an upper bound of 13 Hz on the energy difference between CHFCIBr enantiomers.

A very recent paper concerning the search for a parity-violating energy difference between enantiomers of a chiral molecule is by Lahamer A S, Mahurin S M, Compton R N, House D, Laerdahl J K, Lein M and Schwerdtfeger P 2000 *Phys. Rev. Lett.* **85** 4470. The importance of the parity-violating energy difference in leading to prebiotic asymmetric synthesis is discussed in Frank P, Bonner W A and Zare R N 2000 On one hand but not the other: the challenge of the origin and survival of homochirality in prebiotic chemistry *Chemistry for the 21st Century* ed E Keinan and I Schechter (Weinheim: Wiley-VCH) pp 175–208.

A1.4.3.4 IDENTICAL PARTICLE PERMUTATION SYMMETRY

If there are n electrons in a molecule there are $n!$ ways of permuting them and we can form the permutation group (or symmetric group) $S_n^{(e)}$ of degree n and order $n!$ that contains all the electron permutations. The molecular Hamiltonian is invariant to the elements of this group. Similarly, there can be sets of identical nuclei in a molecule and the Hamiltonian is invariant to the relevant identical-nucleus permutation groups. For example, the ethanol molecule

-38-

$\text{CH}_3\text{CH}_2\text{OH}$ consists of 26 electrons, a set of six identical hydrogen nuclei, a set of two identical carbon nuclei and a lone oxygen nucleus. The molecular Hamiltonian of ethanol is therefore invariant to the $26!$ ($\sim 4 \times 10^{26}$) elements of the electron permutation group $S_{26}^{(e)}$, the $6!$ = 720 possible permutations of the hydrogen nuclei in the group $S_6^{(H)}$ and the two possible permutations of the C nuclei (E and their exchange) in the group $S_2^{(C)}$. The group of all possible permutations of identical nuclei in a molecule is called the *complete nuclear permutation* (CNP) group of the molecule G^{CNP} . For ethanol G^{CNP} consists of all $6!$ elements of $S_6^{(H)}$ and of all these elements taken in combination with the exchange of the two C nuclei; $2 \times 6!$ elements in all. This CNP group is called the *direct product* of the groups $S_6^{(H)}$ and $S_2^{(C)}$ and is written

(A1.4.128)

$$G^{\text{CNP}} = S_6^{(\text{H})} \otimes S_2^{(\text{C})}.$$

The CNP group of a molecule containing l identical nuclei of one type, m of another, n of another and so on is the direct product group

$$G^{\text{CNP}} = S_l \otimes S_m \otimes S_n \dots \quad (\text{A1.4.129})$$

and the order of the group is $l! \times m! \times n! \dots$. It would seem that we have a very rich set of irreducible representation labels with which we can label the molecular energy levels of a molecule using the electron permutation group and the CNP group. But this is not the case for internal states described by Ψ_{int} (see [equation A1.4.114](#)) because there is fundamentally no observable difference between states that differ merely in the permutation of identical particles. The environment of a molecule (e.g. an external electric or magnetic field, or the effect of a neighbouring molecule) affects whether the Hamiltonian of that molecule is invariant to a rotation operation or the inversion operation; states having different symmetry labels from the rotation or inversion groups can be mixed and transitions can occur between such differently labelled states. However, the Hamiltonian of a molecule *regardless of the environment of the molecule* is invariant to any identical particle permutation. Two Ψ_{int} states that differ only in the permutation of identical particles are observationally indistinguishable and there is only one state. Since there is only one state it can only transform as one set of irreducible representations of the various identical particle permutation groups that apply for the particular molecule under investigation. It is an experimental fact that particles with half integral spin (called *fermions*), such as electrons and protons, transform as that one-dimensional irreducible representation of their permutation group that has character +1 for all even permutations⁹ and character -1 for all odd permutations. Nuclei that have integral spin (called *bosons*), such as ¹²C nuclei and deuterons, transform as the totally symmetric representation of their permutation group (having character +1 for all permutations). Thus fermion wavefunctions are changed in sign by an odd permutation but boson wavefunctions are invariant. This simple experimental observation has defied simple theoretical proof but there is a complicated proof [10] that we cannot recommend any reader of the present article to look at.

The fact that allowed fermion states have to be antisymmetric, i.e., changed in sign by any odd permutation of the fermions, leads to an interesting result concerning the allowed states. Let us write a state wavefunction for a system of n noninteracting fermions as

$$|X\rangle = |a_1\rangle|b_2\rangle|c_3\rangle \dots |q_n\rangle \quad (\text{A1.4.130})$$

-39-

where this indicates that particle 1 is in state a , particle 2 in state b and so on. Clearly this does not correspond to an allowed (i.e., antisymmetric) state since making an odd permutation of the indices, such as (12), does not give -1 times $|X\rangle$. But we can get an antisymmetric function by making all permutations of the indices in $|X\rangle$ and adding the results with the coefficient -1 for those functions obtained by making an odd permutation, i.e.,

$$|F\rangle = \sum_P \pm P |a_1\rangle|b_2\rangle|c_3\rangle \dots |q_n\rangle \quad (\text{A1.4.131})$$

where the sum over all permutations involves a + or - sign as the permutation P is even or odd respectively. We can write (equation A1.4.131) as the determinant

$$|F\rangle = \begin{vmatrix} |a_1\rangle & |b_1\rangle & |c_1\rangle & \dots & |q_1\rangle \\ |a_2\rangle & |b_2\rangle & |c_2\rangle & \dots & |q_2\rangle \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ |a_n\rangle & |b_n\rangle & |c_n\rangle & \dots & |q_n\rangle \end{vmatrix}. \quad (\text{A1.4.132})$$

The state $|F\rangle$ is such that the particle states a, b, c, \dots, q are occupied and each particle is equally likely to be in any one of the particle states. However, if two of the particle states a, b, c, \dots, q are the same then $|F\rangle$ vanishes; it does not correspond to an allowed state of the assembly. This is a characteristic of antisymmetric states and it is called ‘the Pauli exclusion principle’: no two identical fermions can be in the same particle state. The general function for an assembly of bosons is

$$|B\rangle = \sum_P P|a_1\rangle|b_2\rangle|c_3\rangle \dots |q_n\rangle \quad (\text{A1.4.133})$$

where the sum over all permutations involves just ‘+’ signs. In such a state it is possible for two or more of the particles to be in the same particle state.

It would appear that identical particle permutation groups are not of help in providing distinguishing symmetry labels on molecular energy levels as are the other groups we have considered. However, they do provide very useful restrictions on the way we can build up the complete molecular wavefunction from basis functions. Molecular wavefunctions are usually built up from basis functions that are products of electronic and nuclear parts. Each of these parts is further built up from products of separate ‘uncoupled’ coordinate (or orbital) and spin basis functions. When we combine these separate functions, the final overall product states must conform to the permutation symmetry rules that we stated above. This leads to restrictions in the way that we can combine the uncoupled basis functions.

We explain this by considering the H_2 molecule. For the H_2 molecule we label the electrons a and b , and the hydrogen nuclei 1 and 2. The electron permutation group is $\mathbf{S}_2^{(e)} = \{E, (ab)\}$, and the CNP group $\mathbf{G}^{\text{CNP}} = \{E, (12)\}$. The character tables of these groups are given in [table A1.4.6](#) and [table A1.4.7](#). If there were no restriction on permutation symmetry we might think that the energy levels of the H_2 molecule could be of any one of the following four symmetry

-40-

types using these two groups: $(\Gamma_1^{(e)}, \Gamma_1^{(\text{CNP})})$, $(\Gamma_1^{(e)}, \Gamma_2^{(\text{CNP})})$, $(\Gamma_2^{(e)}, \Gamma_1^{(\text{CNP})})$ and $(\Gamma_2^{(e)}, \Gamma_2^{(\text{CNP})})$. However, both electrons and protons are fermions (having a spin of 1/2) and so, from the above rules, the wavefunctions of the H_2 molecule must be multiplied by -1 by both (ab) and (12) . Thus the energy levels of the H_2 molecule can only be of symmetry $(\Gamma_2^{(e)}, \Gamma_2^{(\text{CNP})})$.

Table A1.4.6 The character table of the group $\mathbf{S}_2^{(e)}$.

	$E \langle /b \langle (ab)$	
$\Gamma_1^{(e)}$	1	1
$\Gamma_2^{(e)}$	1	-1

These limitations lead to electron spin multiplicity restrictions and to differing nuclear spin statistical weights for the rotational levels. Writing the electronic wavefunction as the product of an orbital function Ψ_e and a spin function Ψ_{es} , there are restrictions on how these functions can be combined. The restrictions are imposed by the fact that the complete function $\Psi_e\Psi_{es}$ has to be of symmetry $\Gamma_2^{(e)}$ in the group $S_2^{(e)}$. The orbital function Ψ_e can be of symmetry $\Gamma_1^{(e)}$ or $\Gamma_2^{(e)}$ and, for example, Ψ_e for the ground electronic state of H_2 has symmetry $\Gamma_1^{(e)}$. For a two electron system there are four possible electron spin functions¹⁰: $\alpha\alpha$, $\alpha\beta$, $\beta\alpha$ and $\beta\beta$, where α is a ‘spin-up’ function having $m_s = +1/2$ and β is a ‘spin-down’ function having $m_s = -1/2$. The functions $\Psi_{es}^{(1)} = \alpha\alpha$ and $\Psi_{es}^{(2)} = \beta\beta$ are invariant to the operation (ab) and therefore have symmetry $\Gamma_1^{(e)}$. The functions $\alpha\beta$ and $\beta\alpha$ are interchanged by (ab) and do not transform irreducibly, but it is easy to see that their sum and difference, $\Psi_{es}^{(3)} = (\alpha\beta + \beta\alpha)/\sqrt{2}$ and $\Psi_{es}^{(4)} = (\alpha\beta - \beta\alpha)/\sqrt{2}$, transform as $\Gamma_1^{(e)}$ and $\Gamma_2^{(e)}$ respectively. The three functions $\Psi_{es}^{(1)}$, $\Psi_{es}^{(2)}$ and $\Psi_{es}^{(3)}$, each of symmetry $\Gamma_1^{(e)}$, form a triplet electron spin state (with $m_s = 1, -1$ and 0 , for $S = 1$) and the function $\Psi_{es}^{(4)}$, having symmetry $\Gamma_2^{(e)}$ is a singlet state (with $S = 0$). The ground electronic state cannot be a triplet state since if it were then the symmetry of both Ψ_e and Ψ_{es} would be $\Gamma_1^{(e)}$ and the product would therefore be of symmetry $\Gamma_1^{(e)}$ which is not allowed. Hence the ground electronic state of H_2 has to be a singlet electronic state.

The way we combine the nuclear spin basis functions Ψ_{ns} with the rotation–vibration–electronic basis functions Ψ_{rve} in H_2 follows the same type of argument using the nuclear permutation group G^{CNP} . Rovibronic states of symmetry $\Gamma_1^{(CNP)}$ can only be combined with Ψ_{ns} of species $\Gamma_2^{(CNP)}$ (of which there is one with $I = 0$), and rovibronic states of symmetry $\Gamma_2^{(CNP)}$ can only be combined with Ψ_{ns} of species $\Gamma_1^{(CNP)}$ (of which there are three with $I = 1$). Thus rovibronic states of symmetry $\Gamma_1^{(CNP)}$ have a *nuclear spin statistical weight* of 1, and rovibronic states of symmetry $\Gamma_2^{(CNP)}$ have a nuclear spin statistical weight of 3. An interesting result of these considerations follows for the $^{16}O_2$ molecule by using the G^{CNP} group. Labelling the O nuclei 1 and 2 this group is as in [table A1.4.7](#). The spin of ^{16}O nuclei is 0 and so the nuclear spin wavefunction is of species $\Gamma_1^{(CNP)}$. There is no nuclear spin wavefunction of species $\Gamma_2^{(CNP)}$ in $^{16}O_2$. Since ^{16}O nuclei are bosons the complete wavefunction must be of symmetry and thus

-41-

rovibronic states of species $\Gamma_2^{(CNP)}$ (which can only be combined with a nuclear spin wavefunction of species $\Gamma_2^{(CNP)}$) have no nuclear spin partner with which to combine. Thus these states cannot occur and are ‘missing.’ This means that half the rotational levels of every vibronic state are missing in this molecule. Missing levels arise in other molecules and can also involve nuclei with nonzero spin; they arise for the ammonia molecule NH_3 .

Table A1.4.7 The character table of the group $G^{(CNP)}$ for H_2 .

E (12)

$\Gamma^{(\text{CNP})}$	1	1
$\Gamma_1^{(\text{CNP})}$	1	-1
Γ_2		

The Pauli exclusion principle follows from the indistinguishability of electrons and the rules of fermion permutation. It prevents the occurrence of states that have two or more electrons in the same particle state. As a result of the indistinguishability of nuclei, and the rules of fermion and boson permutation, there are missing levels. Both of these results can be tested experimentally. A negative result from trying to put an electron into the 1S state of Cu (this state already having two electrons of opposite spin in it) was reported by Ramberg and Snow [11] and by analysing their results they determined an upper limit for the violation of the Pauli exclusion principle of 1.7×10^{-26} ; this means that at this level the electrons are indistinguishable. Attempts to observe spectral lines that would arise from transitions between ‘missing’ levels have been made in order to see whether the levels are truly missing. Such missing levels would arise if the nuclei involved are not completely identical. Such a situation is conceivable. Three negative attempts at a sensitivity level of only about 10^{-6} have been reported [12, 13, 14].

A1.4.3.5 TIME REVERSAL SYMMETRY

The time reversal symmetry operation $\hat{\theta}$ (or T) is the operation of reversing the direction of motion; it reverses all momenta, including spin angular momenta, but not the coordinates (see [15] for a good general account of this symmetry operation). As with the inversion operation E^* the weak interaction force is not invariant to time reversal and we discuss this further in the next subsection. However, for all practical purposes in molecular physics we can take this to be a symmetry operation. This symmetry operation has the property, unlike the other symmetry operations discussed here, of being *antiunitary*. Also, time reversal invariance does not lead to any conservation law and molecular states are not eigenstates of $\hat{\theta}$. However, this symmetry operator constrains the form of the Hamiltonian, an example being that no term in the Hamiltonian can contain the product of an odd number of momenta. Also, it is sometimes a useful tool in determining whether certain matrix elements vanish (see, for example, [16]) and it can be responsible for extra degeneracies. In particular, if a symmetry group has a pair of irreducible representations, Γ and Γ^* say, whose characters are the complex conjugates of each other, then energy levels of symmetry Γ and Γ^* will always coincide in pairs and be degenerate because of time reversal symmetry. Such a pair of irreducible representations of a symmetry group are called ‘separably degenerate’. The irreducible representations E_+ and E_- of the point group C_3 (see table A1.4.8) are separably degenerate. Such a character table can be condensed by adding the characters of the separably degenerate irreducible representations and this is done for the C_3 group in table A1.4.9. In the condensed character table the separably degenerate representations are marked ‘sep’.

Table A1.4.8 The character table of the point group C_3 .

	E	C_3	
A_1	1	1	1
E_+	1	ω	ω^2
E_-	1	ω^2	ω

$$\omega = \exp(2\pi i/3).$$

Table A1.4.9 The condensed character table of the C_3 group.

E	C_3	
A_1	1	1
E	2	-1 sep

Apart from the degeneracy of separably degenerate states, time reversal symmetry leads to *Kramers' degeneracy* or *Kramers' theorem*: all energy levels of a system containing an odd number of particles with half-integral spin (i.e., fermions) must be at least doubly degenerate. One generally only considers systems having an odd number of electrons, but if nuclei with half integral spin cause the degeneracy then one must resolve the nuclear hyperfine structure for the degeneracy to be revealed.

A1.4.3.6 CONCLUDING REMARKS ABOUT SYMMETRIES

In the above we have discussed several different symmetry groups: the translation group G_T , the rotation group K (spatial), the inversion group, the electron permutation group and the complete nuclear permutation group G^{CNP} . We have also discussed the time reversal symmetry operation. The translational states Φ_{CM} can be classified according to their linear momentum using G_T , but we rarely worry about the translational state of a molecule. The internal states Φ_{int} can be labelled with their angular momentum (F, m_F) using K (spatial), and their parity (\pm) using. The symmetry in the group leads to restrictions on the electron spin multiplicities (the Pauli exclusion principle) and the symmetry in G^{CNP} leads to nuclear spin statistical weights. One might think that we should form a 'full' symmetry group of the molecular Hamiltonian, G_{FULL} say, describing all symmetry types simultaneously and symmetry classify our basis functions and eigenfunctions in this group. If we neglect time reversal symmetry (which requires special consideration because the operator is antiunitary), we have

$$(A1.4.134)$$

that is, the full symmetry group for an isolated molecule in field-free space is the direct product of the groups describing the individual symmetry types. However, it can be shown that it is completely equivalent and easier, to treat each type of symmetry and each symmetry group, separately. In order to transform irreducibly in G_{FULL} ,

a wavefunction must transform irreducibly in each of the groups G_T , K (spatial), \mathcal{E} , S_n^e and G^{CNP} . This is discussed in section 7.3 of [1]. Watson [17] has shown that for a molecule in an external electric field the full symmetry group cannot be factorized in the simple manner of (equation A1.4.134). In this case, instead of the three separate groups K (spatial), \mathcal{E} and G^{CNP} , it is necessary to consider a more complicated group containing selected elements of their direct product group. In the following section we show how the direct product of the groups \mathcal{E} and G^{CNP} , called the complete nuclear permutation–inversion (CNPI) group G^{CNPI} , is used in molecular physics; it leads to the definition of the molecular symmetry (MS) group. In the final section we show how the molecular point group emerges from the molecular symmetry group as a near symmetry group

of the molecular Hamiltonian.

As a postscript to this section we consider the operation of charge conjugation symmetry. This operation is not used in molecular physics but it is an important symmetry in nature, and it does lead to an important implication about the probable breakdown of time reversal symmetry. Classical electrodynamic forces are invariant if we change the signs of the charges. In elementary particle physics the ‘charge conjugation operation’ C is introduced as a generalization of this changing-the-sign-of-the-charge operation: it is the operation of changing every particle (including uncharged particles like the neutron) into its antiparticle. Weak interactions are not invariant to the operation C just as they are not invariant to the inversion operation P . One might hope to preserve the exact ‘mirror symmetry’ of nature if invariance to the product CP were a fact. Unfortunately, CP symmetry is not universal [18], although its violation is a small effect that has never been observed outside the neutral K meson (kaon) system and the extent of its violation cannot be calculated (unlike the situation with parity violation, which by comparison is a big effect). CP violation permits unequal treatment of particles and antiparticles and it may be responsible for the domination of matter over antimatter in the universe [19]. Very recent considerations concerning CP violation are summarized in [20]; in particular, this reference points out that the study of CP violation in neutral B mesons will probe the physics behind the ‘standard model’, which does not predict sufficient CP violation to account, by itself, for the predominance of matter over antimatter in the universe. In the light of the fact that C was introduced as a generalization of the changing-the-sign-of-the-charge operation, it is appropriate that CP violation provides an unambiguous ‘convention-free’ definition of positive charge: *it is the charge carried by the lepton preferentially produced in the decay of the long-lived neutral K meson*[21]. Although CP violation is a fact there is one invariance in nature involving C that is believed to be universal (based on quantum field theory) and that is invariance under the triple operation $TC P$, which also involves the time reversal operation T . $TC P$ symmetry implies that every particle has the same mass and lifetime as its antiparticle. However, now, if $TC P$ symmetry is true the observation of CP violation in experiments on neutral K mesons must mean that there is a compensating violation of time reversal symmetry at the same time. A direct experimental measure of the violation of time reversal symmetry has not been made, mainly because the degree of violation is very small.

A1.4.4 THE MOLECULAR SYMMETRY GROUP

The complete nuclear permutation inversion (CNPI) group of the PH_3 molecule is the direct product of the complete nuclear permutation (CNP) group \mathcal{S}_3 (see (equation A1.4.19)) and the inversion group $\mathcal{E} = \{E, E^*\}$. This is a group of 12 elements that we call \mathcal{G}_{12} :

$$\mathcal{G}_{12} = \{E, (123), (132), (12), (23), (31), E^*, (123)^*, (132)^*, (12)^*, (23)^*, (31)^*\}. \quad (\text{A1.4.135})$$

-44-

The rotation–vibration–electronic energy levels of the PH_3 molecule (neglecting nuclear spin) can be labelled with the irreducible representation labels of the group \mathcal{G}_{12} . The character table of this group is given in table A1.4.10.

Table A1.4.10 The character table of the CNPI group \mathcal{G}_{12} .

E	(123)	(12)	E^*	(123)*	(12)*
	(132)	(23)		(132)*	(23)
		(31)			(31)*

A_1^+	1	1	1	1	1	1
A_1^-	1	1	1	-1	-1	-1
A_2^+	1	1	-1	1	1	-1
A_2^-	1	1	-1	-1	-1	1
E^+	2	-1	0	2	-1	0
E^-	2	-1	0	-2	1	0

Before we consider the results of this symmetry labelling, we should consider the effect of the inversion motion in PH_3 . In [figure A1.4.5](#) we depict the two *versions* (see [\[22\]](#) for a discussion of this term) of the numbered equilibrium structure of the molecule and call them a and b. The inversion coordinate ρ is also indicated in this figure. In [figure A1.4.6](#) we schematically indicate the cross-section in the potential energy surface of the PH_3 molecule that contains the two minima and the barrier between them. In this figure we also indicate several vibrational energy levels of the molecule. The barrier to inversion is so high ($\approx 11\,300\text{ cm}^-1$; see [\[23\]](#)) that there is no observable inversion tunnelling splitting. Thus, the energy levels can be calculated by just considering the motion in one of the two minima and we do not need to consider both minima. The ‘single minimum’ calculation is represented in [figure A1.4.7](#) each minimum has a duplicate set of energy levels.

-45-

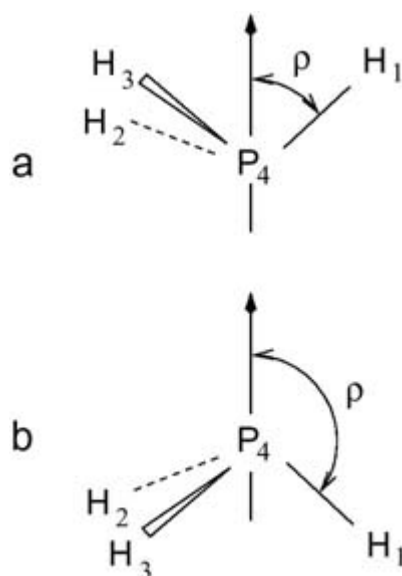


Figure A1.4.5. PH_3 inversion.

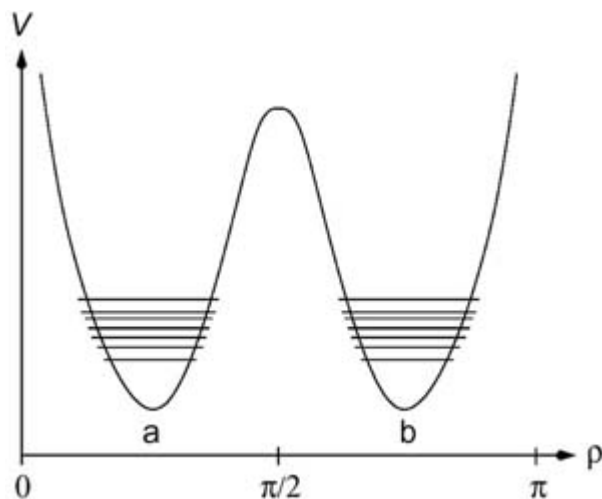


Figure A1.4.6. A cross-section of the potential energy surface of PH_3 . The coordinate ρ is defined in figure A1.4.5.

-46-

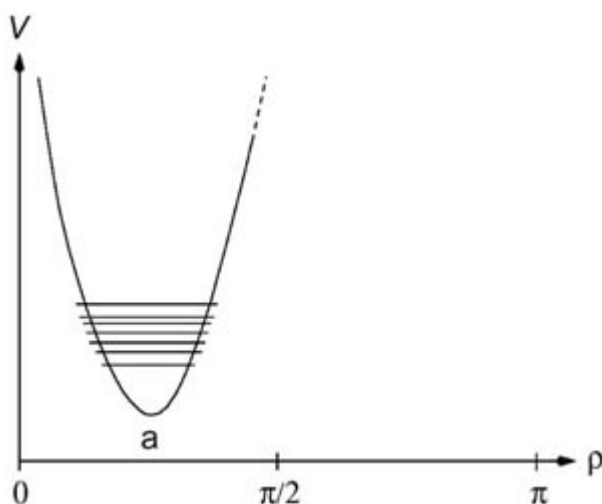


Figure A1.4.7. A cross-section of the potential energy surface of PH_3 obtained by ignoring the version b (see figure A1.4.6). The coordinate ρ is defined in figure A1.4.5.

If we were to calculate the vibrational energy levels using the double minimum potential energy surface, we would find that well below the barrier, every energy level would be doubly degenerate to within measurement accuracy for PH_3 . If we symmetry classified the levels using the group G_{12} we would find that there were three types of energy level: $A_1^+ + A_1^-$, $A_2^+ + A_2^-$ or $E^+ + E^-$. This double degeneracy would be resolved by inversion tunnelling and it is an accidental degeneracy not forced by the symmetry group G_{12} . If the inversion tunnelling is not resolved we have actually done too much work here. There are only three distinct types of level and yet we have used a symmetry group with six irreducible representations. However, Longuet-Higgins [24] showed how to obtain the appropriate subgroup of G_{12} that avoids the unnecessary double labels. This is achieved by just using the elements of G_{12} that are appropriate for a single minimum; we delete elements such as E^* and (12) that interconvert the a and b forms. Longuet-Higgins termed the deleted elements 'unfeasible.' The group obtained is 'the molecular symmetry (MS) group'. In the case of PH_3 , we obtain the particular MS group

$$C_{3v}(\text{M}) = \{E, (123), (132), (12)^*, (23)^*, (31)^*\}; \quad (\text{A1.4.136})$$

its character table (with the class structure indicated) is given in [table A1.4.11](#). Using this group, we achieve a sufficient symmetry labelling of the levels as being either A_1 , A_2 or E . All possible interactions can be understood using this group (apart from the effect of inversion tunnelling).

-47-

Table A1.4.11 The character table of the molecular symmetry group $C_{3v}(M)$.

	E	(123) (132)	$(12)^*$ $(23)^*$ $(31)^*$
A_1	1	1	1
A_2	1	1	-1
E	2	-1	0

For PH_3 the labour saved by using the MS group rather than the CNPI group is not very great, but for larger molecules, such as the water trimer for example, a great saving is achieved if all unfeasible elements of the CNPI group are eliminated from consideration. An unfeasible element of the CNPI group is one that takes the molecule between versions that are separated by an insuperable energy barrier in the potential energy function. For the water trimer the CNPI group has $6! \times 3! \times 2 = 8640$ elements. The MS group that is used to interpret the spectrum has 48 elements [25].

Ammonia (NH_3) is pyramidal like PH_3 and in its electronic ground state there are two versions of the numbered equilibrium structure exactly as shown for PH_3 in [figure A1.4.5](#). The potential barrier between the two versions, however, is around 2000 cm^{-1} for ammonia [26] and thus much lower than in PH_3 . This barrier is so low that the molecule will tunnel through it on the time scale of a typical spectroscopic experiment, and the tunnelling motion gives rise to energy level splittings that can be resolved experimentally (see, for example, figure 15-3 of [1]). Thus, for NH_3 , all elements of the group G_{12} are feasible, and the molecular symmetry group of NH_3 in its electronic ground state is G_{12} . This group is isomorphic to the point group D_{3h} and in the literature it is customarily called $D_{3h}(M)$.

A1.4.5 THE MOLECULAR POINT GROUP

The MS group is introduced by deleting unfeasible elements from the CNPI group. It can be applied to symmetry label the rotational, vibrational, electronic and spin wavefunctions of a molecule, regardless of whether the molecule is rigid or nonrigid. It is a *true* symmetry group and no terms in the Hamiltonian can violate the symmetry labels obtained (with the exception of the as yet undetected effect of the weak neutral current interaction). The MS group can be used to determine nuclear spin statistical weights, to determine which states can and cannot interact as a result of considering previously neglected higher order terms in the Hamiltonian, or the effect of externally applied magnetic or electric fields and it can be used to determine the selection rules for allowed electric and magnetic dipole transitions. What then of the molecular point group?

For a molecule that has no observable tunnelling between minima on the potential energy surface (i.e., for a rigid molecule) and for which the equilibrium structure is nonlinear¹¹, it turns out that the MS group is isomorphic to the point group of the equilibrium structure. For example, PH_3 has the molecular symmetry

group $C_{3v}(M)$ given in (equation A1.4.136) and its equilibrium structure has the point group C_{3v} given in (equation A1.4.22). It is easy to show from (equation A1.4.31) (using the fact that $E^*E^* = E$) that these two groups are isomorphic with the following mapping:

-48-

(A1.4.137)

Obviously, we have chosen the name $C_{3v}(M)$ for the molecular symmetry group of PH_3 because this group is isomorphic to C_{3v} .

Quite remarkably, if we neglect the effect of the MS group elements on the rotational variables (the Euler angles θ , ϕ and χ) then each element of the MS group rotates and/or reflects the vibrational displacements and electronic coordinates in the manner described by its partner in the point group. In fact, for the purpose of classifying vibrational and electronic wavefunctions this *defines* what the elements of the molecular point group actually do to the molecular coordinates for a rigid nonlinear molecule. By starting with the fundamental definition of symmetry in terms of energy invariance, by considering the operations of inversion and identical nuclei permutation and, finally, by deleting unfeasible elements of the CNPI group, we recover the simple description of molecular symmetry in terms of rotations and reflections, but the rotations and reflections are of the vibrational displacements and the electronic coordinates—not of the entire molecule at its equilibrium configuration. Such operations are not symmetry operations of the full Hamiltonian (unlike the elements of the MS group) since the transformation of the rotational variables is neglected. This means that such effects as Coriolis coupling for example, which involve a coupling of rotation and vibration, will mix vibrational states of different point group symmetry. The molecular point group is a *near* symmetry group of the full Hamiltonian. However, the molecular point group is a symmetry group of the vibration–electronic Hamiltonian of a rigid molecule and in practice it is always used for labelling the vibration–electronic states of such molecules. Its use enables one, for example, to classify the normal vibration coordinates and to study the transformation properties of the electronic wavefunction without having to bother about molecular rotation. This is a useful simplification, but the reader must be aware that the rotation and/or reflection operations of the molecular point group do not rotate and/or reflect the molecule in space; they rotate and/or reflect the vibrational displacements and electronic coordinates¹². To study the effect of molecular rotation (as one needs to do if one is interested in understanding high resolution rotationally resolved molecular spectra), or to study nonrigid molecules such as the water trimer, the point group is of no use and one must employ the appropriate MS group.

ACKNOWLEDGMENTS

We thank Antonio Fernandez-Ramos for critically reading the manuscript. Part of this paper was written while PJ worked as a guest at the Steacie Institute for Molecular Sciences.

¹ Clearly, the operation (21) has the same effect as (12), (13) has the same effect as (31) etc.

² The axis labels (p, q, r) are chosen in order not to confuse this axis system with other systems, such as the molecule fixed axes (x, y, z) discussed below, used to describe molecular motion.

³ For an observer viewing the pq plane from a point that has a positive r coordinate (figure A1.4.2), the positive right-handed direction of the C_3 and rotations is anticlockwise.

⁴ Equivalently, it follows if we apply R to both sides of (equation A1.4.58) and then use (equation A1.4.59) on the left hand side.

⁵ Two functions Ψ_{n_i} and Ψ_{n_j} are orthogonal if the product $\Psi_{n_i}^* \Psi_{n_j}$ integrated over all configuration space, vanishes. A function Ψ is normalized if the product $\Psi^* \Psi$ integrated over all configuration space is unity. An orthonormal set contains functions that are normalized and orthogonal to each other.

⁶ Note the order of the subscripts on $D[R]$ which follows from the fact that we use the N-convention of (equation A1.4.56) to define the effect of a permutation on a function.

⁷ Proved, for example, in section 6.5 of [1]}

⁸ That is, a molecule for which the minimum of the Born–Oppenheimer potential energy function corresponds to a nonlinear geometry. The theory of linear molecules is explained in chapter 17 of [1].

⁹ An even (odd) permutation is one that when expressed as the product of pair exchanges involves an even (odd) number of such exchanges. Thus (123)=(12)(23) and (12345)=(12)(23)(34)(45) are even permutations, whereas (12), (1234)=(12)(23)(34) and (123456)=(12)(23)(34)(45)(56) are odd permutations.

¹⁰ We give the spin of electron a first and of electron b second.

¹¹ Rigid linear molecules are a special case in which an *extended* MS group, rather than the MS group, is isomorphic to the point group of the equilibrium structure; see chapter 17 of [1].

¹² A detailed discussion of the relation between MS group operations and point group operations is given in section 4.5 of [1].

REFERENCES

- [1] Bunker P R and Jensen P 1998 *Molecular Symmetry and Spectroscopy* 2nd edn (Ottawa: NRC)
 - [2] Cotton F A 1990 *Chemical Applications of Group Theory* 3rd edn (New York: Wiley)
 - [3] Tinkham M 1964 *Group Theory and Quantum Mechanics* (New York: McGraw-Hill)
 - [4] Bunker P R and Howard B J 1983 *Symmetries and Properties of Non-Rigid Molecules: a Comprehensive Survey (Studies in Physical and Theoretical Chemistry 23)* ed J Maruani and J Serre (Amsterdam: Elsevier) p 29
 - [5] Eckart C 1935 *Phys. Rev.* **47** 552
 - [6] Zare R N 1988 *Angular Momentum* (New York: Wiley)
 - [7] Kleiman V, Gordon R J, Park H and Zare R N 1998 *Companion to Angular Momentum* (New York: Wiley)
 - [8] Wood C S, Bennett S C, Cho D, Masterson B P, Roberts J L, Tanner C E and Wieman C E 1997 Measurement of parity nonconservation and an anapole moment in cesium *Science* **275** 1759–63
-

[9] Hegstrom R A, Rein D W and Sandars P G H 1980 *J. Chem. Phys.* **73** 2329

[10] Pauli W 1940 *Phys. Rev.* **58** 716

- [11] Ramberg E and Snow G A 1990 *Phys. Lett. B* **238** 438
- [12] de Angelis M, Gagliardi G, Gianfrani L and Tino G M 1996 Test of the symmetrization postulate for spin-0 particles *Phys. Rev. Lett.* **76** 2840
- [13] Hilborn R C and Yuca C L 1996 Spectroscopic test of the symmetrization postulate for spin-0 nuclei *Phys. Rev. Lett.* **76** 2844
- [14] Naus H, de Lange A and Ubachs W 1997 *Phys. Rev. A* **56** 4755
- [15] Overseth O E 1969 *Sci. Am.* October
- [16] Watson J K G 1974 *J. Mol. Spectrosc.* **50** 281
- [17] Watson J K G 1975 *Can. J. Phys.* **53** 2210
- [18] Christenson J H, Cronin J W, Fitch V L and Turlay R 1964 *Phys. Rev. Lett.* **13** 138
- [19] Wilczek F 1980 *Sci. Am.* December
- [20] Schwarzschild B 1999 *Phys. Today* January 22
- [21] Griffiths D 1987 *Introduction to Elementary Particles* (New York: Harper and Row)
- [22] Bone R G A, Rowlands T W, Handy N C and Stone A J 1991 *Mol. Phys.* **72** 33
- [23] Špirko V, Civiš S, Ebert M and Danielis V 1986 *J. Mol. Spectrosc.* **119** 426
- [24] Longuet-Higgins H C 1963 *Mol. Phys.* **6** 445
- [25] van der Avoird A, Olthof E H T and Wormer P E S 1996 *J. Chem. Phys.* **105** 8034
- [26] Špirko V and Kraemer W P 1989 *J. Mol. Spectrosc.* **133** 331
-

FURTHER READING

- Bunker P R and Jensen P 1998 *Molecular Symmetry and Spectroscopy* 2nd edn (Ottawa: NRC)
- Cotton F A 1990 *Chemical Applications of Group Theory* 3rd edn (New York: Wiley)
- Griffiths D 1987 *Introduction to Elementary Particles* (New York: Harper and Row)
- Kleiman V, Gordon R J, Park H and Zare R N 1998 *Companion to Angular Momentum* (New York: Wiley)
- Tinkham M 1964 *Group Theory and Quantum Mechanics* (New York: McGraw-Hill)
- Zare R N 1988 *Angular Momentum* (New York: Wiley)

-1-

A 1.5 Intermolecular interactions

Ajit J Thakkar

A1.5.1 INTRODUCTION

The existence of intermolecular interactions is apparent from elementary experimental observations. There must be attractive forces because otherwise condensed phases would not form, gases would not liquefy, and liquids would not solidify. There must be short-range repulsive interactions because otherwise solids and liquids could be compressed to much smaller volumes with ease. The kernel of these notions was formulated in the late eighteenth century, and Clausius made a clear statement along the lines of this paragraph as early as 1857 [1].

Since the interaction energy V between a pair of molecules must have an attractive region at large intermolecular separations r and a steeply repulsive region at short distances, it is evident that $V(r)$ must have the schematic form illustrated in figure A1.5.1. It is conventional to denote the distance at which the interaction energy is a minimum by either r_m or r_e and to refer to this distance as the equilibrium distance. Similarly it is common to denote the shorter distance at which the interaction energy is zero by σ and refer to it as the slow collision diameter. The net potential energy of attraction at the minimum is $V(r_m) = -\varepsilon$, and ε is called the well depth.

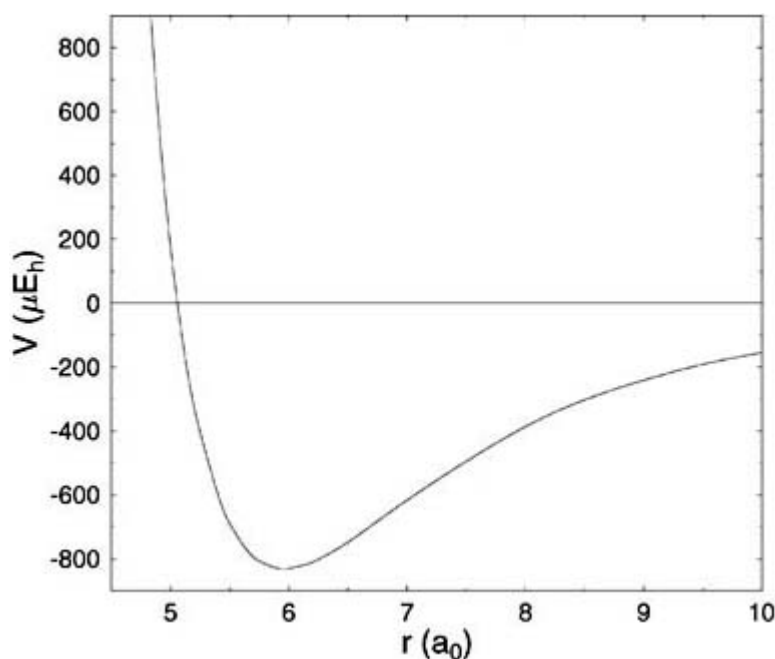


Figure A1.5.1 Potential energy curve for NeF^- based on *ab initio* calculations of Archibong *et al* [88].

-2-

In 1873, van der Waals [2] first used these ideas to account for the deviation of real gases from the ideal gas law $P\bar{V} = RT$ in which P , \bar{V} and T are the pressure, molar volume and temperature of the gas and R is the gas constant. He argued that the incompressible molecules occupied a volume b leaving only the volume $\bar{V} - b$ free for the molecules to move in. He further argued that the attractive forces between the molecules reduced the pressure they exerted on the container by a/\bar{V}^2 ; thus the pressure appropriate for the gas law is $P + a/\bar{V}^2$ rather than P . These ideas led him to the van der Waals equation of state:

$$(P + a/\bar{V}^2)(\bar{V} - b) = RT. \quad (\text{A 1.5.1})$$

The importance of the van der Waals equation is that, unlike the ideal gas equation, it predicts a gas-liquid transition and a critical point for a pure substance. Even though this simple equation has been superseded, its

remarkable success led to the custom of referring to the attractive and repulsive forces between molecules as van der Waals forces.

The feature that distinguishes intermolecular interaction potentials from intramolecular ones is their relative strength. Most typical single bonds have a dissociation energy in the 150–500 kJ mol⁻¹ range but the strength of the interactions between small molecules, as characterized by the well depth, is in the 1–25 kJ mol⁻¹ range.

A1.5.1.1 MANY-BODY EXPANSION

The total energy of an assembly of molecules can be written as

$$E = \sum_i E_i + \sum_{i>j} V_{ij} + \sum_{i>j>k} V_{ijk} + \dots \quad (\text{A 1.5.2})$$

in which E_i is the energy of isolated molecule i , V_{ij} is the energy of interaction between molecules i and j in the absence of any others, V_{ijk} is the *non-additive* energy of interaction among the three molecules i, j and k in the absence of any others, and so on. The interaction energy is then

$$V = E - \sum_i E_i = \sum_{i>j} V_{ij} + \sum_{i>j>k} V_{ijk} + \dots \quad (\text{A 1.5.3})$$

For example, if there are three molecules A, B and C, then equation (A1.5.3) can be written as

$$V = V_{AB} + V_{BC} + V_{CA} + V_{ABC}. \quad (\text{A 1.5.4})$$

V_{AB} is the interaction energy of molecules A and B in the *absence* of molecule C. The interaction between molecules A and B will be different in the presence of molecule C, and so on. The non-additive, three-body term V_{ABC} is the total correction for these errors in the three pair interactions. When there are four molecules, a three-body correction is included for each distinct triplet of molecules and the remaining error is corrected by the non-additive four-body term.

-3-

In many cases, it is reasonable to expect that the sum of two-body interactions will be much greater than the sum of the three-body terms which in turn will be greater than the sum of the four-body terms and so on. Retaining only the two-body terms in [equation \(A1.5.3\)](#) is called the *pairwise additivity* approximation. This approximation is quite good so the bulk of our attention can be focused on describing the two-body interactions. However, it is now known that the many-body terms cannot be neglected altogether, and they are considered briefly in [section A1.5.2.6](#) and [section A1.5.3.5](#).

A1.5.1.2 TYPES OF INTERMOLECULAR INTERACTIONS

It is useful to classify various contributions to intermolecular forces on the basis of the physical phenomena that give rise to them. The first level of classification is into long-range forces that vary as inverse powers of the distance r^{-n} , and short-range forces that decrease exponentially with distance as in $\exp(-\alpha r)$.

There are three important varieties of long-range forces: *electrostatic*, *induction* and *dispersion*. Electrostatic forces are due to classical Coulombic interactions between the static charge distributions of the two molecules. They are strictly pairwise additive, highly anisotropic, and can be either repulsive or attractive.

The distortions of a molecule's charge distribution induced by the electric field of all the other molecules leads to induction forces that are always attractive and highly non-additive. Dispersion forces are always present, always attractive, nearly pairwise additive, and arise from the instantaneous fluctuations of the electron distributions of the interacting molecules. If the molecules are in closed-shell ground states, then there are no other important long-range interactions. However, if one or more of the molecules are in degenerate states, then non-additive, resonance interactions of either sign can arise. Long-range forces are discussed in greater detail in [section A1.5.2](#).

The most important short-range forces are exchange and repulsion; they are very often taken together and referred to simply as exchange–repulsion. They are both non-additive and of opposing sign, but the repulsion dominates at short distances. The overlap between the electron densities of molecules when they are close to one another leads to modifications of the long-range terms and thence to short-range penetration, charge transfer and damping effects. All these effects are discussed in greater detail in [section A1.5.3](#).

A1.5.1.3 POTENTIAL ENERGY SURFACES

Only the interactions between a pair of atoms can be described as a simple function $V(r)$ of the distance between them. For nonlinear molecules, several coordinates are required to describe the relative orientation of the interacting species. Thus it is necessary to think of the interaction energy as a 'potential energy surface' (PES) that depends on many variables. There are usually several points of minimum energy on this surface; many of these will be 'local minima' and at least one will be the 'global minimum'. The interaction energy at a local minimum is lower than at any point in its neighbourhood but there can be lower energy minima further away. If there is more than one global minimum, then these are located at symmetry equivalent points on the surface, corresponding to the same minimum energy.

For the interaction between a nonlinear molecule and an atom, one can place the coordinate system at the centre of mass of the molecule so that the PES is a function of the three spherical polar coordinates r, θ, ϕ needed to specify the location of the atom. If the molecule is linear, V does not depend on ϕ and the PES is a function of only two variables. In the general case of two nonlinear molecules, the interaction energy depends on the distance between the centres of mass, and five of the six Euler angles needed to specify the relative orientation of the molecular axes with respect to the global or 'space-fixed' coordinate axes.

A1.5.2 LONG-RANGE FORCES

A1.5.2.1 LONG-RANGE PERTURBATION THEORY

Perturbation theory is a natural tool for the description of intermolecular forces because they are relatively weak. If the interacting molecules (A and B) are far enough apart, then the theory becomes relatively simple because the overlap between the wavefunctions of the two molecules can be neglected. This is called the polarization approximation. Such a theory was first formulated by London [[3](#), [4](#)], and then reformulated by several others [[5](#), [6](#) and [7](#)].

Each electron in the system is assigned to either molecule A or B, and Hamiltonian operators \mathcal{H}^A and \mathcal{H}^B for each molecule defined in terms of its assigned electrons. The unperturbed Hamiltonian for the system is then $\mathcal{H}^0 = \mathcal{H}^A + \mathcal{H}^B$, and the perturbation $\lambda\mathcal{H}'$ consists of the Coulomb interactions between the nuclei and electrons of A and those of B. The unperturbed states, eigenfunctions of \mathcal{H}^0 , are simple product functions $\Psi_{\mu}^A \Psi_{\nu}^B$. For closed-shell molecules, non-degenerate, Rayleigh–Schrödinger, perturbation theory gives the energy of the ground state of the interacting system. The first-order interaction energy is the electrostatic

energy, and the second-order energy is partitioned into induction and dispersion energies. The induction energy consists of all terms that involve excited states of only one molecule at a time, whereas the dispersion energy includes all the remaining terms that involve excited states of both molecules simultaneously.

Long-range forces are most conveniently expressed as a power series in $1/r$, the reciprocal of the intermolecular distance. This series is called the multipole expansion. It is so common to use the multipole expansion that the electrostatic, induction and dispersion energies are referred to as ‘non-expanded’ if the expansion is not used. In early work it was noted that the multipole expansion did not converge in a conventional way and doubt was cast upon its use in the description of long-range electrostatic, induction and dispersion interactions. However, it is now established [8, 9, 10, 11, 12 and 13] that the series is asymptotic in Poincaré’s sense. The interaction energy can be written as

$$V(r) = \sum_{n=0}^N V_n/r^n + O(1/r^{N+1}) \quad (\text{A 1.5.5})$$

with the assurance that the remainder left upon truncation after some chosen term in r^{-N} tends to zero in the limit as $r \rightarrow \infty$. In other words, the multipole expansion can be made as accurate as one desires for large enough intermolecular separations, even though it cannot be demonstrated to converge at any given value of r and, in some cases, diverges for all r !

Some electric properties of molecules are described in [section A1.5.2.2](#) because the coefficients of the powers of $1/r$ turn out to be related to them. The electrostatic, induction and dispersion energies are considered in turn in [section A1.5.2.3](#), [section A1.5.2.4](#) and [section A1.5.2.5](#), respectively.

A1.5.2.2 MULTIPOLE MOMENTS AND POLARIZABILITIES

The long-range interactions between a pair of molecules are determined by electric multipole moments and polarizabilities of the individual molecules. *Multipole moments* are measures that describe the non-sphericity of the charge distribution of a molecule. The zeroth-order moment is the total charge of the molecule: $Q = \sum_i q_i$ where q_i is the charge of particle i and the sum is over all electrons and nuclei in the molecule. The first-order moment is the dipole moment vector with Cartesian components given by

$$\mu_\alpha = \int \rho(\mathbf{r}) r_\alpha d^3\mathbf{r} \quad \alpha \in \{x, y, z\} \quad (\text{A 1.5.6})$$

in which $\rho(\mathbf{r})$ is the total (electronic plus nuclear) charge density of the molecule. The direction of the dipole moment is from negative to positive. Dipole moments have been measured for a vast variety of molecules [14, 15 and 16].

Next in order is the quadrupole moment tensor Θ with components:

$$\Theta_{\alpha\beta} = \frac{1}{2} \int \rho(\mathbf{r}) (3r_\alpha r_\beta - r^2 \delta_{\alpha\beta}) d^3\mathbf{r} \quad \alpha, \beta \in \{x, y, z\} \quad (\text{A 1.5.7})$$

where the ‘Kronecker delta’ $\delta_{\alpha\beta} = 1$ for $\alpha = \beta$ and $\delta_{\alpha\beta} = 0$ for $\alpha \neq \beta$. The quadrupole moment is a symmetric

$(\Theta_{\alpha\beta} = \Theta_{\beta\alpha})$ second-rank tensor. Moreover, it is traceless:

$$\Theta_{xx} + \Theta_{yy} + \Theta_{zz} = 0. \quad (\text{A 1.5.8})$$

Therefore, it has at most five independent components, and fewer if the molecule has some symmetry. Symmetric top molecules have only one independent component of Θ , and, in such cases, the axial component is often referred to as the quadrupole moment. A quadrupolar distribution can be created from four charges of the same magnitude, two positive and two negative, by arranging them in the form of two dipole moments parallel to each other but pointing in opposite directions. Centro-symmetric molecules, like CO_2 , have a zero dipole moment but a non-zero quadrupole moment.

The multipole moment of rank n is sometimes called the 2^n -pole moment. The first non-zero multipole moment of a molecule is origin independent but the higher-order ones depend on the choice of origin. Quadrupole moments are difficult to measure and experimental data are scarce [17, 18 and 19]. The octopole and hexadecapole moments have been measured only for a few highly symmetric molecules whose lower multipole moments vanish. *Ab initio* calculations are probably the most reliable way to obtain quadrupole and higher multipole moments [20, 21 and 22].

The charge redistribution that occurs when a molecule is exposed to an electric field is characterized by a set of constants called *polarizabilities*. In a uniform electric field \mathbf{F} , a component of the dipole moment is

$$\mu_\alpha = \mu_\alpha^0 + \alpha_{\alpha\beta} F_\beta + \frac{1}{2} \beta_{\alpha\beta\gamma} F_\beta F_\gamma + \frac{1}{3!} \Gamma_{\alpha\beta\gamma\delta} F_\beta F_\gamma F_\delta + \dots \quad (\text{A 1.5.9})$$

in which $\alpha_{\alpha\beta}$, $\beta_{\alpha\beta\gamma}$ and $\Gamma_{\alpha\beta\gamma\delta}$, respectively, are components of the dipole polarizability, *hyperpolarizability* and second hyperpolarizability tensors, and a summation is implied over repeated subscripts.

The dipole polarizability tensor characterizes the lowest-order dipole moment *induced* by a uniform field. The α tensor is symmetric and has no more than six independent components, less if the molecule has some symmetry. The scalar or mean dipole polarizability

$$\bar{\alpha} = \frac{1}{3} \text{Tr } \alpha = \frac{1}{3} \sum_i \alpha_{ii} \quad (\text{A 1.5.10})$$

is invariant to the choice of coordinate system and is often referred to simply as ‘the polarizability’. It is related to many important bulk properties of an ensemble of molecules including the dielectric constant, the refractive index, the extinction coefficient, and the electric susceptibility. The polarizability is a measure of the softness of the molecule’s electron density, and correlates directly with molecular size, and inversely with the ionization potential and HOMO–LUMO gap. Another scalar polarizability invariant commonly encountered is the polarizability anisotropy:

$$(\Delta\alpha)^2 = \frac{1}{2} [3\text{Tr } \alpha^2 - (\text{Tr } \alpha)^2]. \quad (\text{A 1.5.11})$$

In linear, spherical and symmetric tops the components of α along and perpendicular to the principal axis of symmetry are often denoted by α_{\parallel} and α_{\perp} , respectively. In such cases, the anisotropy is simply $\Delta\alpha = \alpha_{\parallel} - \alpha_{\perp}$. If the applied field is oscillating at a frequency ω , then the dipole polarizability is frequency dependent as well $\alpha(\omega)$. The zero frequency limit of the ‘dynamic’ polarizability $\alpha(\omega)$ is the static polarizability described above.

There are higher multipole polarizabilities that describe higher-order multipole moments induced by non-uniform fields. For example, the quadrupole polarizability is a fourth-rank tensor \mathbf{C} that characterizes the lowest-order quadrupole moment induced by an applied field gradient. There are also mixed polarizabilities such as the third-rank dipole–quadrupole polarizability tensor \mathbf{A} that describes the lowest-order response of the dipole moment to a field gradient and of the quadrupole moment to a dipolar field. All polarizabilities of order higher than dipole depend on the choice of origin. Experimental values are basically restricted to the dipole polarizability and hyperpolarizability [23, 24 and 25]. *Ab initio* calculations are an important source of both dipole and higher polarizabilities [20]; some recent examples include [26, 27].

A1.5.2.3 ELECTROSTATIC INTERACTIONS

The electrostatic potential generated by a molecule A at a distant point B can be expanded in inverse powers of the distance r between B and the centre of mass (CM) of A. This series is called the *multipole expansion* because the coefficients can be expressed in terms of the multipole moments of the molecule. With this expansion in hand, it is

-7-

straightforward to write the electrostatic interaction between molecule A and another molecule with its CM at B as a multipole expansion. The formal expression [7, 28] for this electrostatic interaction, in terms of ‘T tensors’, is intimidating to all but the experts. However, explicit expressions for individual terms in this expansion are easily understood.

Consider the case of two neutral, linear, dipolar molecules, such as HCN and KCl, in a coordinate system with its origin at the CM of molecule A and the z -axis aligned with the intermolecular vector \mathbf{r} pointing from the CM of A to the CM of B. The relative orientation of the two molecules is uniquely specified by their spherical polar angles θ_A, θ_B and the difference $\phi = \phi_A - \phi_B$ between their azimuthal angles. The leading term in the multipole expansion of the electrostatic interaction energy is the dipole–dipole term

$$V_{dd}(r, \theta_A, \theta_B, \phi) = -\frac{\mu_A \mu_B}{4\pi \epsilon_0 r^3} (2 \cos \theta_A \cos \theta_B - \sin \theta_A \sin \theta_B \cos \phi) \quad (\text{A 1.5.12})$$

in which ϵ_0 is the vacuum permittivity, and μ_A and μ_B are the magnitudes of the dipole moments of A and B. This expression is also applicable to the dipole–dipole interaction between any pair of neutral molecules provided that the angles are taken to specify the relative orientation of the dipole moment vectors of the molecules.

The leading term in the electrostatic interaction between a pair of linear, quadrupolar molecules, such as HCCH and CO_2 is

$$V_{qq} = \frac{3Q_A Q_B}{16\pi \epsilon_0 r^5} [1 - 5 \cos^2 \theta_A - 5 \cos^2 \theta_B - 15 \cos^2 \theta_A \cos^2 \theta_B + 2(4 \cos \theta_A \cos \theta_B - \sin \theta_A \sin \theta_B \cos \phi)^2] \quad (\text{A 1.5.13})$$

in which Q_A and Q_B are the axial quadrupole moments of A and B. This expression is also applicable to the quadrupole–quadrupole interaction between any pair of spherical or symmetric top molecules provided that

the angles are taken to specify the relative orientation of the axial component of the quadrupole moment tensors of the molecules.

The leading term in the electrostatic interaction between the dipole moment of molecule A and the axial quadrupole moment of a linear, spherical or symmetric top B is

$$V_{dq} = \frac{3\mu_A\Theta_B}{8\pi\epsilon_0r^4}[\cos\theta_A(3\cos^2\theta_B - 1) - \sin\theta_A\sin 2\theta_B\cos\phi]. \quad (\text{A 1.5.14})$$

Note the r dependence of these three terms: the dipole–dipole interaction varies as r^{-3} , the dipole–quadrupole as r^{-4} and the quadrupole–quadrupole as r^{-5} . In general, the interaction between a 2^ℓ -pole moment and a 2^L -pole moment varies as $r^{-(\ell+L+1)}$. Thus, the dipole–octopole interaction also varies as r^{-5} . At large enough r , only the term involving the lowest-rank, non-vanishing, multipole moment is important. Higher terms begin to play a role as r decreases. The angular variation of the electrostatic interaction is much greater than that of the induction and dispersion. Hence, electrostatic forces often determine the geometry of a van der Waals complex even when they do not constitute the dominant contribution to the overall interaction.

-8-

At a fixed distance r , the angular factor in [equation \(A1.5.12\)](#) leads to the greatest attraction when the dipoles are lined up in a linear head-to-tail arrangement, $\theta_A = \theta_B = 0$, whereas the linear tail-to-tail geometry, $\theta_A = \pi, \theta_B = 0$, is the most repulsive. A head-to-tail, parallel arrangement, $\theta_A = \theta_B = \pi/2, \phi = \pi$, is attractive but less so than the linear head-to-tail geometry. Nevertheless, if the molecules are linear, the head-to-tail, parallel geometry may be more stable because it allows the molecules to get closer and thus increases the r^{-3} factor. For example, the HCN dimer takes the linear head-to-tail geometry in the gas phase [29], but the crystal structure shows a parallel, head-to-tail packing [30].

For interactions between two quadrupolar molecules which have Θ_A and Θ_B of the same sign, at a fixed separation r , the angular factor in [equation \(A1.5.13\)](#) leads to a planar, T-shaped structure, $\theta_A = 0, \theta_B = \pi/2, \phi = 0$, being preferred. This geometry is often seen for nearly spherical quadrupolar molecules. There are other planar ($\phi = 0$) configurations with $\theta_A = \pi/2 - \theta_B$ that are also attractive. A planar, ‘slipped parallel’ structure, $\theta_A = \theta_B \approx \pi/4, \phi = 0$ is often preferred by planar molecules, and long and narrow molecules because it allows them to approach closer thereby increasing the radial factor. For example, benzene, naphthalene and many other planar quadrupolar molecules have crystal structures consisting of stacks of tilted parallel molecules.

For interactions between two quadrupolar molecules which have θ_A and θ_B of the opposite sign, at a fixed separation r , the angular factor in [equation \(A1.5.13\)](#) leads to a linear structure, $\theta_A = \theta_B = 0$, being the most attractive. Linear molecules may also prefer a C_{2v} rectangular or non-planar ‘cross’ arrangement with $\theta_A = \theta_B = \pi/2$, which allows them to approach closer and increase the radial factor.

Although such structural arguments based purely on electrostatic arguments are greatly appealing, they are also grossly over-simplified because all other interactions, such as exchange–repulsion and dispersion, are neglected, and there are serious shortcomings of the multipole expansion at smaller intermolecular separations.

A1.5.2.4 INDUCTION INTERACTIONS

If the long-range interaction between a pair of molecules is treated by quantum mechanical perturbation theory, then the electrostatic interactions considered in [section A1.5.2.3](#) arise in first order, whereas induction and dispersion effects appear in second order. The multipole expansion of the induction energy in its full generality [7, 28] is quite complex. Here we consider only explicit expressions for individual terms in the

multipole expansion that can be understood readily.

Consider the interaction of a neutral, dipolar molecule A with a neutral, S-state atom B. There are no electrostatic interactions because all the multipole moments of the atom are zero. However, the electric field of A distorts the charge distribution of B and induces multipole moments in B. The leading induction term is the interaction between the permanent dipole moment of A and the dipole moment induced in B. The latter can be expressed in terms of the polarizability of B, see [equation \(A1.5.9\)](#), and the dipole–induced-dipole interaction is given by

$$V_{\text{did}} = -\frac{\mu_{\text{A}}^2 \alpha_{\text{B}}}{2(4\pi \epsilon_0)^2 r^6} (3 \cos^2 \theta_{\text{A}} + 1) \quad (\text{A 1.5.15})$$

in which θ_{A} is the angle between the dipole moment vector of A and the intermolecular vector, and α_{B} is the mean dipole polarizability of B. Since B is a spherical atom, its polarizability tensor is diagonal with the three diagonal

-9-

components equal to one another and to the mean.

If molecule A is a linear, spherical or symmetric top that has a zero dipole moment like benzene, then the leading induction term is the quadrupole–induced-dipole interaction

$$V_{\text{qid}} = -\frac{9\Theta_{\text{A}}^2 \alpha_{\text{B}}}{8(4\pi \epsilon_0)^2 r^8} (4 \cos^4 \theta_{\text{A}} + \sin^4 \theta_{\text{A}}) \quad (\text{A 1.5.16})$$

in which θ_{A} is the angle between the axial component of the quadrupole moment tensor of A and the intermolecular vector.

If the molecule is an ion bearing a charge Q_{A} , then the leading induction term is the isotropic, charge–induced-dipole interaction

$$V_{\text{cid}} = -\frac{Q_{\text{A}}^2 \alpha_{\text{B}}}{2(4\pi \epsilon_0)^2 r^4} \quad (\text{A 1.5.17})$$

For example, this is the dominant long-range interaction between a neon atom and a fluoride anion F^- .

Note the r dependence of these terms: the charge–induced-dipole interaction varies as r^{-4} , the dipole–induced-dipole as r^{-6} and the quadrupole–induced-dipole as r^{-8} . In general, the interaction between a permanent 2^{ℓ} -pole moment and an induced 2^{L} -pole moment varies as $r^{-2(\ell + \text{L} + 1)}$. At large enough r , only the leading term is important, with higher terms increasing in importance as r decreases. The induction forces are clearly non-additive because a third molecule will induce another set of multipole moments in the first two, and these will then interact. Induction forces are almost never dominant since dispersion is usually more important.

A1.5.2.5 DISPERSION INTERACTIONS

The most important second-order forces are dispersion forces. London [3, 31, 32] showed that they are caused by a correlation of the electron distribution in one molecule with that in the other, and pointed out that the

electrons contributing most strongly to these forces are the same as those responsible for the dispersion of light. Since then, these forces have been called London or dispersion forces. Dispersion interactions are always present, even between S-state atoms such as neon and krypton, although there are no electrostatic or induction interaction terms since all the multipole moments of both species are zero.

Dispersion forces cannot be explained classically but a semiclassical description is possible. Consider the electronic charge cloud of an atom to be the time average of the motion of its electrons around the nucleus. The average cloud is spherically symmetric with respect to the nucleus, but at any instant of time there may be a polarization of charge giving rise to an instantaneous dipole moment. This instantaneous dipole induces a corresponding instantaneous dipole in the other atom and there is an interaction between the instantaneous dipoles. The dipole of either atom averages to zero over time, but the interaction energy does not because the instantaneous and induced dipoles are correlated and

-10-

they stay in phase. The average interaction energy falls off as r^{-6} just as the dipole-induced-dipole energy of equation (A1.5.15). Higher-order instantaneous multipole moments are also involved, giving rise to higher-order dispersion terms. This picture is visually appealing but it should not be taken too literally. The actual effect is not time dependent in the sense of classical fluctuations taking place.

The multipole expansion of the dispersion interaction can be written as

$$V(r) = -C_6/r^6 - C_8/r^8 - C_{10}/r^{10} - \dots \quad (\text{A 1.5.18})$$

where the dispersion coefficients C_6 , C_8 and C_{10} are positive, and depend on the electronic properties of the interacting species. The first term is the interaction between the induced-dipole moments on the atoms, the second is the induced-dipole-induced-quadrupole term and the third consists of the induced-dipole-induced-octopole term as well as the interaction between induced quadrupoles. In general, the interaction between an induced 2^ℓ -pole moment and an induced 2^L -pole moment varies as $r^{-2(\ell+L+1)}$. The dispersion coefficients are constants for atoms but, for non-spherical molecules, they depend upon the five angles describing the relative orientation of the molecules. For example, the dispersion coefficients for the interactions between an S-state atom and a Σ_g^+ -state diatomic molecule can be expressed as

$$C_{2n}(\theta) = \sum_{L=0}^{n-2} C_{2n}^{2L} P_{2L}(\cos \theta) \quad (\text{A 1.5.19})$$

where the C_{2n}^{2L} are dispersion constants, the $P_{2L}(\cos \theta)$ are Legendre polynomials, and θ is the angle between the symmetry axis of the diatomic and the intermolecular vector. Note that C_{2n}^0 is the spherical average of $C_{2n}(\theta)$ and is the appropriate quantity to use in equation (A1.5.18) if the orientation dependence is being neglected. Purely anisotropic dispersion terms varying as r^{-7}, r^{-9}, \dots arise if at least one of the interacting species lacks inversion symmetry.

Perturbation theory yields a sum-over-states formula for each of the dispersion coefficients. For example, the isotropic C_6^{AB} coefficient for the interaction between molecules A and B is given by

$$C_6^{AB} = \frac{3e^4 \hbar^4}{2m_e^2 (4\pi \epsilon_0)^2} \sum_{m,n \neq 0} \frac{f_{Am} f_{Bn}}{\Delta E_{Am} \Delta E_{Bn} (\Delta E_{Am} + \Delta E_{Bn})} \quad (\text{A 1.5.20})$$

in which \hbar is the Planck–Dirac constant, $\Delta E_{Am} = E_{Am} - E_{A0}$ is the excitation energy from the ground state $m = 0$ to state m for molecule A and f_{Am} is the corresponding dipole oscillator strength averaged over degenerate final states. Similarly, the sum-over-states formula for the mean, frequency-dependent, polarizability can be written as

-11-

$$\bar{\alpha}(\omega) = \frac{e^2}{m_e} \sum_{m \neq 0} \frac{f_{Am}}{\omega_{Am}^2 - \omega^2} \quad (\text{A 1.5.21})$$

where $\omega_{Am} = \Delta E_{Am}/\hbar$ is the m th excitation frequency. An important advance consisted in the realization [33, 34 and 35] that use of the Feynman identity

$$[ab(a+b)]^{-1} = (2/\pi) \int_0^\infty \frac{du}{(a^2+u^2)(b^2+u^2)} \quad \text{for } a > 0, b > 0 \quad (\text{A 1.5.22})$$

together with [equation \(A1.5.20\)](#) and [equation \(A1.5.21\)](#) leads to

$$C_6^{AB} = \frac{3\hbar}{\pi(4\pi\epsilon_0)^2} \int_0^\infty \bar{\alpha}_A(i\omega)\bar{\alpha}_B(i\omega) d\omega \quad (\text{A 1.5.23})$$

where $\bar{\alpha}_A(i\omega)$ is the analytic continuation of the dynamic dipole polarizability to the imaginary axis. The significance of [equation \(A1.5.23\)](#) is that it expresses an interaction coefficient in terms of properties of the individual, interacting molecules. The anisotropic components of C_6 can be written as similar integrals involving $\Delta\alpha(i\omega)$, and the higher dispersion coefficients as integrals involving components of the higher-order, dynamic polarizability tensors at imaginary frequency.

Many methods for the evaluation of C_6 from [equation \(A1.5.20\)](#) use moments of the dipole oscillator strength distribution (DOSD) defined, for molecule A, by

$$S_A(k) = (a_0/e^2)^k \sum_{m \neq 0} f_{Am} \Delta E_{Am}^k \quad \text{for } k = 2, 1, 0, -1, -2, \dots \quad (\text{A 1.5.24})$$

These moments are related to many physical properties. The Thomas–Kuhn–Reiche sum rule says that $S(0)$ equals the number of electrons in the molecule. Other sum rules [36] relate $S(2)$, $S(1)$ and $S(-1)$ to ground state expectation values. The mean static dipole polarizability is $\bar{\alpha}(0) = e^2 S(-2)/m_e$. The Cauchy expansion of the refractive index n at low frequencies ω is given by

$$n^2 - 1 = K_0[S(-2) + \omega^2 K_1 S(-4) + \omega^4 K_2 S(-6) + \dots] \quad (\text{A 1.5.25})$$

where the K_n are known constants. One approach is to use experimental photoabsorption, refractive index and Verdet constant data, together with known sum rules to construct a constrained DOSD from which dipole properties including C_6 can be calculated. This approach was pioneered by Margenau [5, 37], extended by Dalgarno and

coworkers [38, 39], and refined and exploited by Meath and coworkers [40, 41 and 42] who also generalized it to anisotropic properties [43, 44]. Many methods for bounding C_6 in terms of a few DOSD moments have been explored, and the best of these have been identified by an extensive comparative study [45]. *Ab initio* calculations are the only route to the higher-order dispersion coefficients, and Wormer and his colleagues [46, 47, 48, 49 and 50] have led the field in this area. The dimensionless ratio $C_{10}C_6/C_8^2$ is predicted to be a constant for all interactions by simple models [51], and this ratio still serves as a useful check on *ab initio* computations [48]. Dispersion coefficients of even higher order can be estimated from simple models as well [52, 53].

The dispersion coefficient for interactions C_6^{AB} between molecules A and B can be estimated to an average accuracy of 0.5% [45] from those of the A–A and B–B interactions using the Moelwyn-Hughes [54] combining rule:

$$C_6^{AB} = \frac{2C_6^{AA}C_6^{BB}\alpha_A\alpha_B}{C_6^{AA}\alpha_B^2 + C_6^{BB}\alpha_A^2} \quad (\text{A 1.5.26})$$

where α_A and α_B are the static dipole polarizabilities of A and B, respectively. This rule has a sound theoretical basis [55, 56].

A1.5.2.6 MANY-BODY LONG-RANGE FORCES

The induction energy is inherently non-additive. In fact, the non-additivity is displayed elegantly in a distributed polarizability approach [28]. Non-additive induction energies have been found to stabilize what appear to be highly improbable crystal structures of the alkaline earth halides [57].

In the third order of long-range perturbation theory for a system of three atoms A, B and C, the leading non-additive dispersion term is the Axilrod–Teller–Mutō triple–dipole interaction [58, 59]

$$V_{\text{ddd}} = C_9 \frac{(1 + 3 \cos \theta_A \cos \theta_B \cos \theta_C)}{(r_{AB}r_{BC}r_{CA})^3} \quad (\text{A 1.5.27})$$

where r_{AB} , r_{BC} and r_{CA} are the sides of the triangle formed by the atoms, and θ_A , θ_B and θ_C are its internal angles, and the C_9 coefficient can be written [60] in terms of the dynamic polarizabilities of the monomers as

$$C_9^{ABC} = \frac{3\hbar}{\pi(4\pi\epsilon_0)^2} \int_0^\infty \bar{\alpha}_A(i\omega)\bar{\alpha}_B(i\omega)\bar{\alpha}_C(i\omega) d\omega. \quad (\text{A 1.5.28})$$

Hence, the same techniques used to calculate C_6 are also used for C_9 . Note that equation (A1.5.28) has a geometrical factor whose sign depends upon the geometry, and that, unlike the case of the two-body dispersion interaction, the triple–dipole dispersion energy has no minus sign in front of the positive coefficient C_9 . For example, for an equilateral triangle configuration the triple–dipole dispersion is repulsive and varies as $+(11/8)C_9r^{-9}$. There are strongly

anisotropic, non-additive dispersion interactions arising from higher-order polarizabilities as well [61], and the relevant coefficients for rare gas atoms have been calculated *ab initio* [48].

A1.5.3 SHORT- AND INTERMEDIATE-RANGE FORCES

A1.5.3.1 EXCHANGE PERTURBATION THEORIES

The perturbation theory described in [section A1.5.2.1](#) fails completely at short range. One reason for the failure is that the multipole expansion breaks down, but this is not a fundamental limitation because it is feasible to construct a ‘non-expanded’, long-range, perturbation theory which does not use the multipole expansion [6]. A more profound reason for the failure is that the polarization approximation of zero overlap is no longer valid at short range.

When the overlap between the wavefunctions of the interacting molecules cannot be neglected, the zeroth-order wavefunction must be anti-symmetrized with respect to all the electrons. The requirement of anti-symmetrization brings with it some difficult problems. If electrons have been assigned to individual molecules in order to partition the Hamiltonian into an unperturbed part \mathcal{H}^0 and a perturbation $\lambda\mathcal{H}'$, as described in [section A1.5.2.1](#), then these parts do not commute with the antisymmetrization operator \mathcal{A}^{AB} for the full system

$$[\mathcal{A}^{AB}, \mathcal{H}^0] \neq 0, \quad [\mathcal{A}^{AB}, \lambda\mathcal{H}'] \neq 0. \quad (\text{A 1.5.29})$$

On the other hand, the system Hamiltonian $\mathcal{H}^{AB} = \mathcal{H}^0 + \lambda\mathcal{H}'$ is symmetric with respect to all the electrons and commutes with \mathcal{A}^{AB}

$$[\mathcal{A}^{AB}, \mathcal{H}^0 + \lambda\mathcal{H}'] = 0. \quad (\text{A 1.5.30})$$

Combining these commutation relations, we find

$$[\mathcal{A}^{AB}, \mathcal{H}^0] = -[\mathcal{A}^{AB}, \lambda\mathcal{H}'] \neq 0 \quad (\text{A 1.5.31})$$

which indicates that a zeroth-order quantity is equal to a non-zero, first-order quantity. Unfortunately, this means that there will be no unique definition of the order of a term in our perturbation expansion. Moreover, antisymmetrized products of the wavefunctions of A and B will be non-orthogonal, and therefore they will not be eigenfunctions of any Hermitian, zeroth-order Hamiltonian.

Given these difficulties, it is natural to ask whether we really need to antisymmetrize the zeroth-order wavefunction. If we start with the product function, can we reasonably expect that the system wavefunction obtained by perturbation theory will converge to a properly antisymmetric one? Unfortunately, in that case, the series barely converges [62, 63]. Moreover, there are an infinite number of non-physical states with bosonic character that lie below the physical ground state [64] for most systems of interest—all those containing at least one atom with atomic number greater than two

unphysical states.

Clearly, standard Rayleigh–Schrödinger perturbation theory is not applicable and other perturbation methods have to be devised. Excellent surveys of the large and confusing variety of methods, usually called ‘exchange perturbation theories’, that have been developed are available [28, 65]. Here it is sufficient to note that the methods can be classified as either ‘symmetric’ or ‘symmetry-adapted’. Symmetric methods start with antisymmetrized product functions in zeroth order and deal with the non-orthogonality problem in various ways. Symmetry-adapted methods start with non-antisymmetrized product functions and deal with the antisymmetry problem in some other way, such as antisymmetrization at each order of perturbation theory.

A further difficulty arises because the exact wavefunctions of the isolated molecules are not known, except for one-electron systems. A common starting point is the Hartree–Fock wavefunctions of the individual molecules. It is then necessary to include the effects of intramolecular electron correlation by considering them as additional perturbations. Jeziorski and coworkers [66] have developed and computationally implemented a triple perturbation theory of the symmetry-adapted type. They have applied their method, dubbed SAPT, to many interactions with more success than might have been expected given the fundamental doubts [67] raised about the method. SAPT is currently both useful and practical. A recent application [68] to the CO₂ dimer is illustrative of what can be achieved with SAPT, and a rich source of references to previous SAPT work.

A1.5.3.2 FIRST-ORDER INTERACTIONS

In all methods, the first-order interaction energy is just the difference between the expectation value of the system Hamiltonian for the antisymmetrized product function and the zeroth-order energy

$$E^{(1)} = \frac{\langle \mathcal{A}^{AB} \Psi_0^A \Psi_0^B | \mathcal{H}^{AB} | \Psi_0^A \Psi_0^B \rangle}{\langle \mathcal{A}^{AB} \Psi_0^A \Psi_0^B | \Psi_0^A \Psi_0^B \rangle} - (E_0^A + E_0^B) \quad (\text{A 1.5.32})$$

in which E_0^A and E_0^B are the ground-state energies of isolated molecules A and B. An electrostatic part is usually separated out from the first-order energy, also called the Heitler–London energy, and the remainder is called the exchange–repulsion part:

$$E^{(1)} = E_c^{(1)} + E_{\text{xr}}^{(1)}. \quad (\text{A 1.5.33})$$

The ‘non-expanded’ form of the electrostatic or ‘Coulomb’ energy is

$$E_c^{(1)} = \iint \frac{\rho_A(\mathbf{r}_1) \rho_B(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d^3r_1 d^3r_2 \quad (\text{A 1.5.34})$$

where ρ_A and ρ_B are the total (nuclear plus electronic) charge densities of A and B, respectively. A multipole expansion of [equation \(A1.5.34\)](#) leads to the long-range electrostatic energy discussed in [section A1.5.2.3](#).

The difference between the converged multipole expansion of the electrostatic energy and $E_c^{(1)}$ is sometimes called the first-order penetration energy. The exchange–repulsion is often simply called the exchange energy. For Hartree–Fock monomer wavefunctions, $E_{\text{xr}}^{(1)}$ can be divided cleanly [69] into attractive exchange and

dominant repulsion parts. The exchange part arises because the electrons of one molecule can extend over the entire system, whereas the repulsion arises because the Pauli principle does not allow electrons of the same spin to be in the same place.

Figure A1.5.2 shows $E_c^{(1)}$ and $E_{\text{nr}}^{(1)}$ for the He–He interaction computed from accurate monomer wavefunctions [70]. Figure A1.5.3 shows that, as in interactions between other species, the first-order energy $E^{(1)}$ for He–He decays exponentially with interatomic distance. It can be fitted [70] within 0.6% by a function of the form

$$E^{(1)} = (A/r) e^{-br-cr^2} \quad (\text{A 1.5.35})$$

where A, b, c are fitted parameters.

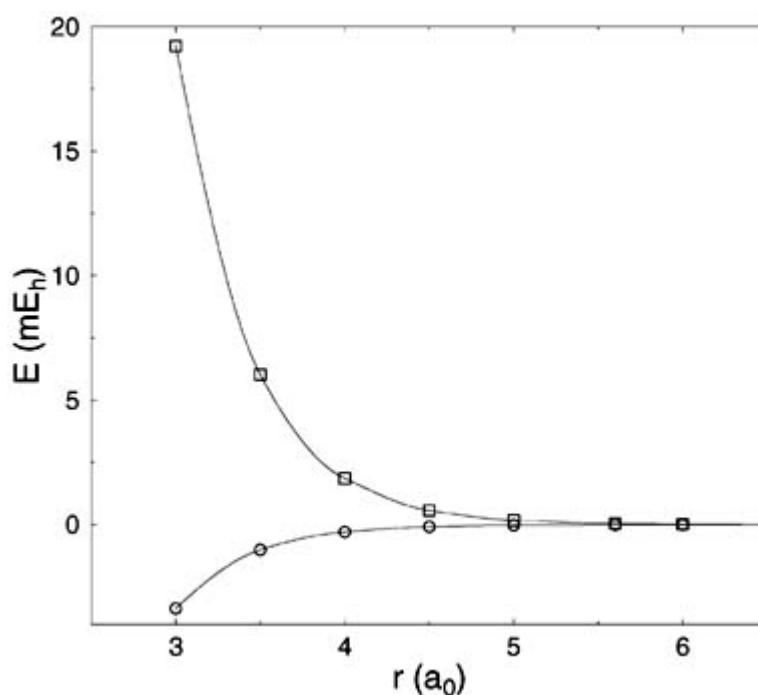


Figure A1.5.2 First-order Coulomb (○) and exchange-repulsion (□) energies for He–He. Based on data from Komasa and Thakkar [70].

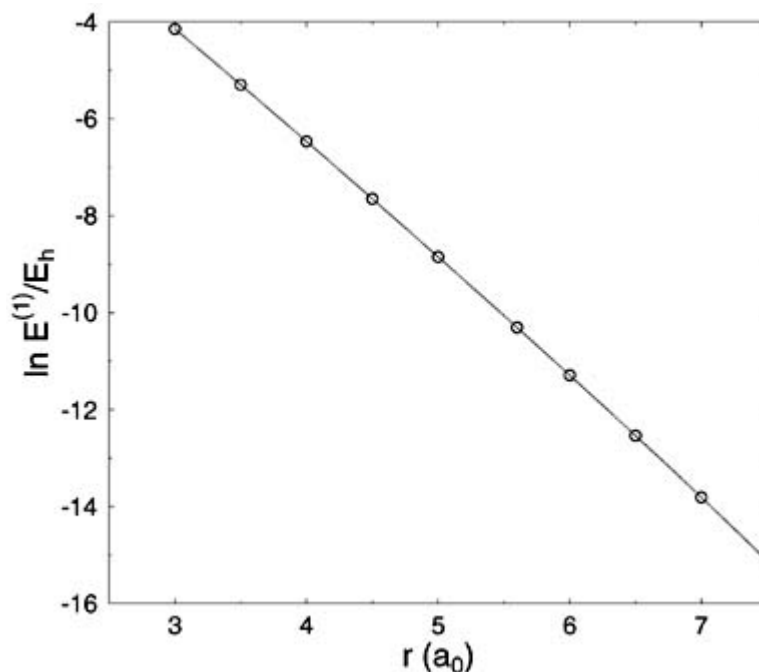


Figure A1.5.3 First-order interaction energy for He–He. Based on data from Komasa and Thakkar [70].

The exchange–repulsion energy is approximately proportional to the overlap of the charge densities of the interacting molecules [71, 72 and 73]

$$E_{\text{xr}}^{(1)} \approx k \left[\int \rho_{\text{A}}(\mathbf{r}) \rho_{\text{B}}(\mathbf{r}) d^3r \right]^n \quad (\text{A 1.5.36})$$

where $n \approx 1$.

A1.5.3.3 SECOND-ORDER INTERACTIONS

The details of the second-order energy depend on the form of exchange perturbation theory used. Most known results are numerical. However, there are some common features that can be described qualitatively. The short-range induction and dispersion energies appear in a non-expanded form and the differences between these and their multipole expansion counterparts are called penetration terms.

The non-expanded dispersion energy can be written as

$$V_{\text{disp}}(r) = -f_6(r)C_6/r^6 - f_8(r)C_8/r^8 - f_{10}(r)C_{10}/r^{10} - \dots \quad (\text{A 1.5.37})$$

where the $f_6(r), f_8(r), \dots$ are ‘damping’ functions. The damping functions tend to unity as $r \rightarrow \infty$ so that the long-range form of equation (A1.5.18) is recovered. As $r \rightarrow 0$, the damping functions tend to zero as r^n so that

they suppress the spurious r^{-n} singularity of the undamped dispersion, [equation \(A1.5.18\)](#). Meath and coworkers [[74](#), [75](#), [76](#), [77](#), [78](#) and [79](#)] have performed *ab initio* calculations of these damping functions for interactions between small species. The general form is shown in [figure A1.5.4](#). Observe that the distance at which the damping functions begin to decrease significantly below unity increases with n . The orientation dependence of the damping functions is not known. Similar damping functions also arise for the induction energy [[74](#), [76](#), [79](#)].

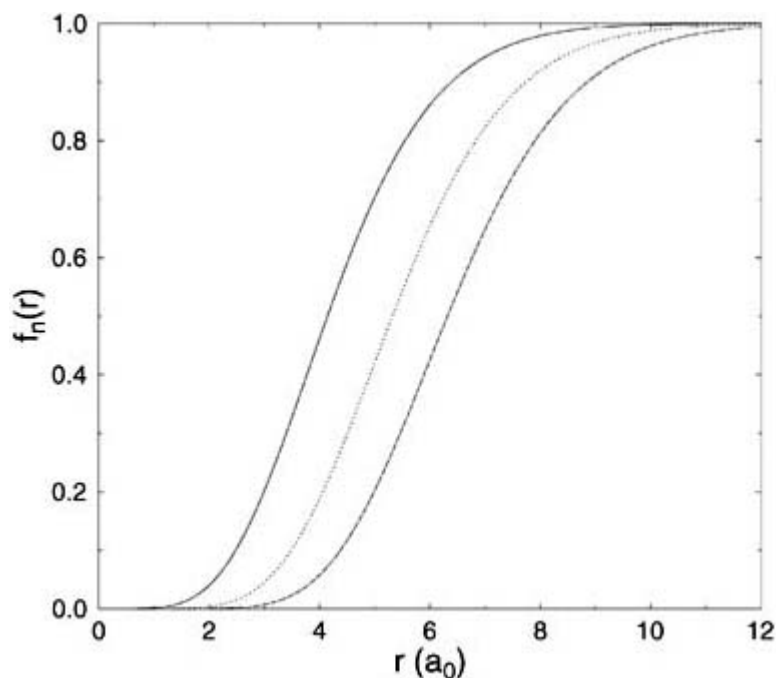


Figure A1.5.4 Dispersion damping functions, f_6 :—, f_8 : and f_{10} :---- for H–H based on data from [[74](#)].

A ‘charge transfer’ contribution is often identified in perturbative descriptions of intermolecular forces. This, however, is not a new effect but a part of the short-range induction energy. It is possible to separate the charge transfer part from the rest of the induction energy [[80](#)]. It turns out to be relatively small and often negligible. Stone [[28](#)] has explained clearly how charge transfer has often been a source of confusion and error.

A1.5.3.4 SUPERMOLECULE CALCULATIONS

The conceptually simplest way to calculate potential energy surfaces for weakly interacting species is to treat the interacting system AB as a ‘supermolecule’, use the Schrödinger equation to compute its energy as a function of the relative coordinates of the interacting molecules, and subtract off similarly computed energies of the isolated molecules. This scheme permits one to use any available method for solving the Schrödinger equation.

Unfortunately, the supermolecule approach [[81](#), [82](#)] is full of technical difficulties, which stem chiefly from the very small magnitude of the interaction energy relative to the energy of the supermolecule. Even today, a novice would be ill-advised to attempt such a computation using one of the ‘black-box’ computer programs available for performing *ab initio* calculations.

That said, the remarkable advances in computer hardware have made *ab initio* calculations feasible for small systems, provided that various technical details are carefully treated. A few examples of recent computations

include potential energy surfaces for He–He [83], Ne–Ne and Ar–Ar [84], Ar–H₂, Ar–HF and Ar–NH₃ [85], N₂–He [86, 87], He–F[−] and Ne–F[−] [88]. Density-functional theory [89] is currently unsuitable for the calculation of van der Waals interactions [90], but the situation could change.

A1.5.3.5 MANY-BODY SHORT-RANGE FORCES

A few *ab initio* calculations are the main source of our current, very meagre knowledge of non-additive contributions to the short-range energy [91]. It is unclear whether the short-range non-additivity is more or less important than the long-range, dispersion non-additivity in the rare-gas solids [28, 92].

A1.5.4 EXPERIMENTAL INFORMATION

Despite the recent successes of *ab initio* calculations, many of the most accurate potential energy surfaces for van der Waals interactions have been obtained by fitting to a combination of experimental and theoretical data. The future is likely to see many more potential energy surfaces obtained by starting with an *ab initio* surface, fitting it to a functional form and then allowing it to vary by small amounts so as to obtain a good fit to many experimental properties simultaneously; see, for example, a recent study on ‘morphing’ an *ab initio* potential energy surface for Ne–HF [93].

This section discusses how spectroscopy, molecular beam scattering, pressure virial coefficients, measurements on transport phenomena and even condensed phase data can help determine a potential energy surface.

A1.5.4.1 SPECTROSCOPY

Spectroscopy is the most important experimental source of information on intermolecular interactions. A wide range of spectroscopic techniques is being brought to bear on the problem of weakly bound or ‘van der Waals’ complexes [94, 95]. Molecular beam microwave spectroscopy, pioneered by Klemperer and refined by Flygare, has been used to determine the microwave spectra of a large number of weakly bound complexes and obtain structural information

-19-

averaged over the vibrational ground state. With the development of tunable far-infrared lasers and sophisticated detectors, far-infrared ‘vibration–rotation–tunnelling’ spectroscopy has enabled Saykally and others to measure data that probes portions of the potential energy surface further from the minimum. Other techniques including vacuum ultraviolet spectroscopy and conventional gas-phase absorption spectroscopy with very long path lengths have also been used.

Spectroscopic data for a complex formed from two atoms can be inverted by the Rydberg–Klein–Rees procedure to determine the interatomic potential in the region probed by the data. The classical turning points r_L and r_R corresponding to a specific energy level $E(\nu, J)$ with vibrational and rotational quantum numbers ν and J can be determined from a knowledge of all the vibrational and rotational energy level spacings between the bottom of the well and the given energy level. The standard equations are [96]

$$f(v, J) = r_R - r_L = \frac{2\hbar}{\sqrt{2\mu}} \int_{-1/2}^v \frac{dv'}{[E(v, J) - E(v', J)]^{1/2}}$$

and

$$g(v, J) = 1/r_L - 1/r_R = \frac{2\sqrt{2\mu}}{\hbar} \int_{-1/2}^v \frac{B(v', J) dv'}{[E(v, J) - E(v', J)]^{1/2}} \quad (\text{A 1.5.39})$$

where $B(v, J) = (2J + 1)^{-1} (\partial E / \partial J)_v$ is a generalized rotational constant. If the rotational structure has not been resolved, then the vibrational spacings alone can be used to determine the well-width function $f(v, 0)$. Similar methods have been developed which enable a spherically averaged potential function to be obtained by inversion of rotational levels, measured precisely enough to yield information on centrifugal distortion, for a single vibrational state. However, most van der Waals complexes are too floppy for a radial potential energy function to be a useful representation of the full PES.

Determination of a PES from spectroscopic data generally requires fitting a parameterized surface to the observed energy levels together with theoretical and other experimental data. This is a difficult process because it is not easy to devise realistic functional representations of a PES with parameters that are not strongly correlated, and because calculation of the vibrational and rotational energy levels from a PES is not straightforward and is an area of current research. The former issue will be discussed further in [section A1.5.5.3](#). The approaches available for the latter currently include numerical integration of a truncated set of ‘close-coupled’ equations, methods based on the discrete variable representation and diffusion Monte Carlo techniques [28]. Some early and fine examples of potential energy surfaces determined in this manner include the H_2 -rare gas surfaces of LeRoy and coworkers [97, 98 and 99], and the hydrogen halide-rare gas potential energy surfaces of Hutson [100, 101 and 102]. More recent work is reviewed by van der Avoird *et al* [103].

-20-

A1.5.4.2 MOLECULAR BEAM SCATTERING

One direct way to study molecular interactions is to cross two molecular beams, one for each of the interacting species, and to study how the molecules scatter after elastic collisions at the crossing point of the two beams. A collision of two atoms depends upon the relative kinetic energy E of collision and the impact parameter b , which is the distance by which the centres of mass would miss each other in the absence of interatomic interaction. Collimated beams with well defined initial velocities can be used, and the scattering measured as a function of deflection angle χ . However, it is not possible to restrict the collisions to a single impact parameter, and results are therefore reported in the form of differential cross sections $\sigma(\chi, E)$ which are measures of the observed scattering intensity. The integral cross section

$$Q(E) = \int \sigma(\chi, E) d\Omega \quad (\text{A 1.5.40})$$

is simply the integral of the differential cross section over all solid angles.

The situation is much the same as with spectroscopic measurements. In the case of interactions between

monatomic species, if all the oscillations in the measured differential cross sections are fully resolved, then an inversion procedure can be applied to obtain the interatomic potential [104, 105]. No formal inversion procedures exist for the determination of a PES from measured cross sections for polyatomic molecules, and it is necessary to fit a parametrized surface to the observed cross sections.

A1.5.4.3 GAS IMPERFECTIONS

The virial equation of state, first advocated by Kamerlingh Onnes in 1901, expresses the compressibility factor of a gas as a power series in the number density:

$$P\bar{V}/RT = 1 + B(T)/\bar{V} + C(T)/\bar{V}^2 + \dots \quad (\text{A 1.5.41})$$

in which $B(T), C(T), \dots$ are called the second, third, \dots virial coefficients. The importance of this equation in the study of intermolecular forces stems from the statistical mechanical proof that the second virial coefficient depends only on the pair potential, even if the total interaction contains significant many-body contributions. For spherically symmetric interactions the relationship between $B(T)$ and $V(r)$ was well established by 1908, and first Keesom in 1912, and then Jones (later known as Lennard-Jones) in the 1920s exploited it as a tool for the determination of intermolecular potentials from experiment [106, 107]. The relationship is simply [108]:

$$B(T) = -2\pi N_A \int_0^\infty [\exp(-V(r)/kT) - 1] r^2 dr. \quad (\text{A 1.5.42})$$

-21-

In the repulsive region ($r < \sigma$) there is a one-to-one correspondence between the interaction energy and the intermolecular distance. Hence it is possible, in principle at least, to obtain $V(r)$ for $r < \sigma$ by inverting $B(T)$. However, in the region of the potential well ($r > \sigma$), both the inner and outer turning points of the classical motion correspond to the same V and hence it is impossible to obtain $V(r)$ uniquely by inverting $B(T)$. In fact [109, 110], inversion of $B(T)$ can only yield the width of the well as a function of its depth. For light species, equation (A1.5.42) is the first term in a semi-classical expansion, and the following terms are called the quantum corrections [106, 107, 111]. For nonlinear molecules, the classical relationship is analogous to equation (A1.5.42) except that the integral is six dimensional since five angles are required to specify the relative orientation of the molecules. In such cases, inversion of $B(T)$ is a hopeless task. Nevertheless, virial coefficient data provide an important test of a proposed potential function.

The third virial coefficient $C(T)$ depends upon three-body interactions, both additive and non-additive. The relationship is well understood [106, 107, 111]. If the pair potential is known precisely, then $C(T)$ ought to serve as a good probe of the non-additive, three-body interaction energy. The importance of the non-additive contribution has been confirmed by $C(T)$ measurements. Unfortunately, large experimental uncertainties in $C(T)$ have precluded unequivocal tests of details of the non-additive, three-body interaction.

A1.5.4.4 TRANSPORT PROPERTIES

The viscosity, thermal conductivity and diffusion coefficient of a monatomic gas at low pressure depend only on the pair potential but through a more involved sequence of integrations than the second virial coefficient. The transport properties can be expressed in terms of ‘collision integrals’ defined [111] by

$$\bar{\Omega}^{(\ell,s)}(T) = [(s+1)!(kT)^{s+2}]^{-1} \int_0^\infty Q^{(\ell)}(E) e^{-E/kT} E^{s+1} dE \quad (\text{A 1.5.43})$$

where k is the Boltzmann constant and E is the relative kinetic energy of the collision. The collision integral is a thermal average of the transport cross section

$$Q^{(\ell)}(E) = 2\pi \left[1 - \frac{1 + (-1)^\ell}{2(1+\ell)} \right]^{-1} \int_0^\infty (1 - \cos^\ell \chi) b db \quad (\text{A 1.5.44})$$

in which b is the impact parameter of the collision, and χ is the deflection angle given by

$$\chi(E, b) = \pi - 2b \int_{r_0}^\infty \frac{dr}{r^2(1 - b^2/r^2 - V(r)/E)^{1/2}} \quad (\text{A 1.5.45})$$

where r_0 , the distance of closest approach in the collision, is the outermost classical turning point of the effective potential. The latter is the sum of the true potential and the centrifugal potential so that $V_{\text{eff}}(L, r) = V(r) + L^2/(2\mu r^2) = V(r) + Eb^2/r^2$ in which L is the angular momentum and μ the reduced mass. Hence r_0 is the outermost solution of $E = V_{\text{eff}}(L, r_0)$.

-22-

The Chapman–Enskog solution of the Boltzmann equation [112] leads to the following expressions for the transport coefficients. The viscosity of a pure, monatomic gas can be written as

$$\eta(T) = \frac{5(m\pi kT)^{1/2}}{16\bar{\Omega}^{(2,2)}(T)} f_\eta \quad (\text{A 1.5.46})$$

and the thermal conductivity as

$$\lambda(T) = \frac{75}{64} \left(\frac{\pi k^3 T}{m} \right)^{1/2} \frac{1}{\bar{\Omega}^{(2,2)}(T)} f_\lambda \quad (\text{A 1.5.47})$$

where m is the molecular mass. f_η and f_λ are higher-order correction factors that differ from unity by only 1 or 2% over a wide temperature range, and can be expressed in terms of collision integrals with different values of ℓ and s . Expression (A1.5.46) and Expression (A1.5.47) imply that

$$\frac{\lambda(T)}{\eta(T)} = \frac{15k f_\lambda}{4m f_\eta} \quad (\text{A 1.5.48})$$

and this is borne out experimentally [111] with the ratio of correction factors being a gentle function of temperature: $f_\lambda/f_\eta \approx 1 + 0.0042(1 - e^{0.33(1-T^*)})$ for $1 < T^* < 90$ with $T^* = kT/\varepsilon$. The self-diffusion coefficient

can be written in a similar fashion:

$$D(T) = \frac{3}{8n} \left(\frac{\pi kT}{m} \right)^{1/2} \frac{1}{\bar{\Omega}^{(1,1)}(T)} f_D \quad (\text{A 1.5.49})$$

where n is the number density. The higher-order correction factor f_D differs from unity by only a few per cent and can also be expressed in terms of other collision integrals.

Despite the complexity of these expressions, it is possible to invert transport coefficients to obtain information about the intermolecular potential by an iterative procedure [111] that converges rapidly, provided that the initial guess for $V(r)$ has the right well depth.

The theory connecting transport coefficients with the intermolecular potential is much more complicated for polyatomic molecules because the internal states of the molecules must be accounted for. Both quantum mechanical and semi-classical theories have been developed. McCourt and his coworkers [113, 114] have brought these theories to computational fruition and transport properties now constitute a valuable test of proposed potential energy surfaces that

-23-

can be performed routinely. Electric and magnetic field effects on transport properties [113, 114] depend primarily on the non-spherical part of the interaction, and serve as stringent checks on the anisotropy of potential energy surfaces.

A1.5.5 MODEL INTERACTION POTENTIALS

There are many large molecules whose interactions we have little hope of determining in detail. In these cases we turn to models based on simple mathematical representations of the interaction potential with empirically determined parameters. Even for smaller molecules where a detailed interaction potential has been obtained by an *ab initio* calculation or by a numerical inversion of experimental data, it is useful to fit the calculated points to a functional form which then serves as a computationally inexpensive interpolation and extrapolation tool for use in further work such as molecular simulation studies or predictive scattering computations. There are a very large number of such models in use, and only a small sample is considered here. The most frequently used simple spherical models are described in section A1.5.5.1 and some of the more common elaborate models are discussed in [section A1.5.5.2](#), [section A1.5.5.3](#) and [section A1.5.5.4](#).

A1.5.5.1 SIMPLE SPHERICAL MODELS

The hard sphere model considers each molecule to be an impenetrable sphere of diameter σ so that

$$V(r) = \begin{cases} \infty & r \leq \sigma \\ 0 & r > \sigma. \end{cases} \quad (\text{A 1.5.50})$$

This simple model is adequate for some properties of rare gas fluids. When it is combined with an accurate description of the electrostatic interactions, it can rationalize the structures of a large variety of van der Waals

complexes [115, 116 and 117].

The venerable bireciprocal potential consists of a repulsive term A/r^n and an attractive term $-B/r^m$ with $n > m$. This potential function was introduced by Mie [118] but is usually named after Lennard-Jones who used it extensively. Almost invariably, $m = 6$ is chosen so that the attractive term represents the leading dispersion term. Many different choices of n have been used, but the most common is $n = 12$ because of its computational convenience. The ‘Lennard-Jones (12,6)’ potential can be written in terms of the well depth (ϵ) and either the minimum position (r_m) or the zero potential location (σ) as

$$V(r) = 4\epsilon[(\sigma/r)^{12} - (\sigma/r)^6] = \epsilon[(r_m/r)^{12} - 2(r_m/r)^6] \quad (\text{A 1.5.51})$$

in which the relationship $\sigma = 2^{-1/6}r_m$ is a consequence of having only two parameters. Fitted values of the coefficient $4\epsilon\sigma^6$ of the r^{-6} term are often twice as large as the true C_6 value because the attractive term has to compensate for the absence of the higher-order dispersion terms. It is remarkable that this simple model continues to be used almost a century after its introduction.

-24-

Morse [119] introduced a potential energy model for the vibrations of bound molecules

$$V(r) = \epsilon[e^{-2(c/\sigma)(r-r_m)} - 2e^{-(c/\sigma)(r-r_m)}] \quad (\text{A 1.5.52})$$

where c is a dimensionless parameter related to the curvature of the potential at its minimum. This function has a more realistic repulsion than the Lennard-Jones potential, but has incorrect long-range behaviour. It has the merit that its vibrational and rotational energy levels are known analytically [119, 120].

The ‘exp-6’ potential replaces the inverse power repulsion in the Lennard-Jones (12, 6) function by a more realistic exponential form:

$$V(r) = \begin{cases} \epsilon(1 - 6/a)^{-1}[(6/a)e^{a(1-r/r_m)} - (r_m/r)^6] & r > r_{\max} \\ \infty & r \leq r_{\max} \end{cases} \quad (\text{A 1.5.53})$$

The potential has a spurious maximum at r_{\max} where the r^{-6} term again starts to dominate. The dimensionless parameter a is a measure of the steepness of the repulsion and is often assigned a value of 14 or 15. The ideas of an exponential repulsion and of its combination with an r^{-6} attraction were introduced by Slater and Kirkwood [121], and the cut-off at r_{\max} by Buckingham [122]. An exponential repulsion, Ae^{-br} , is commonly referred to as a Born–Mayer form, perhaps because their work [123] is better known than that of Slater and Kirkwood.

The parameters in simple potential models for interactions between unlike molecules A and B are often deduced from the corresponding parameters for the A–A and B–B interactions using ‘combination rules’. For example, the σ and ϵ parameters are often estimated from the ‘Lorentz–Berthelot’ rules:

$$\sigma_{AB} = (\sigma_A + \sigma_B)/2 \quad (\text{A 1.5.54})$$

$$\varepsilon_{AB} = (\varepsilon_A \varepsilon_B)^{1/2}. \quad (\text{A } 1.5.55)$$

The former is useful but the latter tends to overestimate the well depth. A harmonic mean rule

$$\varepsilon_{AB} = 2\varepsilon_A \varepsilon_B / (\varepsilon_A + \varepsilon_B) \quad (\text{A } 1.5.56)$$

proposed by Fender and Halsey [124] is generally better than the geometric mean of equation (A1.5.55). Combination rules for the steepness parameter in the exp-6 model include the arithmetic mean

$$a_{AB} = (a_A + a_B)/2 \quad (\text{A } 1.5.57)$$

-25-

and the somewhat more accurate harmonic mean

$$a_{AB} = 2a_A a_B / (a_A + a_B). \quad (\text{A } 1.5.58)$$

Many other rules, some of which are rather more elaborate, have been proposed [111], but these rules have insubstantial theoretical underpinnings and they continue to be used only because there is often no better way to proceed.

A1.5.5.2 ELABORATE SPHERICAL MODELS

The potential functions for the interactions between pairs of rare-gas atoms are known to a high degree of accuracy [125]. However, many of them use *ad hoc* functional forms parametrized to give the best possible fit to a wide range of experimental data. They will not be considered because it is more instructive to consider representations that are more firmly rooted in theory and could be used for a wide range of interactions with confidence.

Slater and Kirkwood's idea [121] of an exponential repulsion plus dispersion needs only one concept, damping functions, see section A1.5.3.3, to lead to a working template for contemporary work. Buckingham and Corner [126] suggested such a potential with an empirical damping function more than 50 years ago:

$$V(r) = A e^{-br} - (C_6/r^6 + C_8/r^8) f(r) \quad (\text{A } 1.5.59)$$

where the damping function is

$$f(r) = \begin{cases} \exp[4(1 - r/r_m)^3] & r < r_m \\ 1 & r \geq r_m. \end{cases} \quad (\text{A } 1.5.60)$$

Modern versions of this approach use a more elaborate exponential function for the repulsion, more dispersion terms, induction terms if necessary, and individual damping functions for each of the dispersion, and sometimes induction, terms as in equation (A1.5.37).

Functional forms used for the repulsion include the simple exponential multiplied by a linear combination of powers (possibly non-integer) of r , a generalized exponential function $\exp(-b(r))$, where $b(r)$ is typically a polynomial in r , and a combination of these two ideas.

Parametrized representations of individual damping dispersion functions were first obtained [127] by fitting *ab initio* damping functions [74] for H–H interactions. The one-parameter damping functions of Douketis *et al* are [127]:

$$f_n(r) = [1 - \exp(-2.1s/n - 0.109s^2/\sqrt{n})]^n \quad (\text{A 1.5.61})$$

-26-

where $s = \rho r$, and ρ is a scale parameter (defined to be $\rho=1/a_0$ for H–H) that enables the damping functions to be used for any interaction. Meath and coworkers [78, 128] prefer the more elaborate form

$$f_n(r) = [1 - \exp(-a_n s - b_n s^2 - d_n s^3)]^n \quad (\text{A 1.5.62})$$

in which the a_n, b_n, d_n ($n = 6, 8, \dots, 20$) are parameters obtained by fitting to *ab initio* damping functions for H–H. A one-parameter damping function of the incomplete gamma form, based on asymptotic arguments and the H–H interaction, is advocated by Tang and Toennies [129]:

$$f_n(r) = 1 - \exp(-br) \sum_{k=0}^n (br)^k / k! \quad (\text{A 1.5.63})$$

where b is a scale parameter which is often set equal to the corresponding steepness parameter in the Born–Mayer repulsion.

Functional forms based on the above ideas are used in the HFD [127] and Tang–Toennies models [129], where the repulsion term is obtained by fitting to Hartree–Fock calculations, and in the XC model [92] where the repulsion is modelled by an *ab initio* Coulomb term $E_c^{(1)}$ and a semi-empirical exchange–repulsion term $E_M^{(1)}$. Current versions of all these models employ an individually damped dispersion series for the attractive term.

An example of a potential energy function based on all these ideas is provided by the 10-parameter function used [88] as a representation of *ab initio* potential energy curves for He–F[−] and Ne–F[−]

$$V(r) = A \exp[-b(r)] - \sum_{n=2}^5 f_{2n}(r) C_{2n} / r^{2n} \quad (\text{A 1.5.64})$$

where $b(r) = (b_0 + b_1 z + b_2 z^2)r$ with $z = (r - r_s)/(r + r_s)$, the damping functions $f_n(r)$ are those of equation (A1.5.61), the r^{-4} term is a pure induction term, and the higher r^{-2n} terms contain both dispersion and induction. Note that this representation implicitly assumes that the dispersion damping functions are applicable to induction without change.

A1.5.5.3 MODEL NON-SPHERICAL INTERMOLECULAR POTENTIALS

The complete intermolecular potential energy surface depends upon the intermolecular distance and up to five angles, as discussed in [section A1.5.1.3](#).

The interaction energy can be written as an expansion employing Wigner rotation matrices and spherical harmonics of the angles [28, 130]. As a simple example, the interaction between an atom and a diatomic molecule can be expanded in Legendre polynomials as

-27-

$$V(r, \theta) = \sum_{L=0}^N V_L(r) P_L(\cos \theta). \quad (\text{A } 1.5.65)$$

This Legendre expansion converges rapidly only for weakly anisotropic potentials. Nonetheless, truncated expansions of this sort are used more often than justified because of their computational advantages.

A more natural way to account for the anisotropy is to treat the parameters in an interatomic potential, such as [equation \(A1.5.64\)](#), as functions of the relative orientation of the interacting molecules. Corner [131] was perhaps the first to use such an approach. Pack [132] pointed out that Legendre expansions of the well depth ϵ and equilibrium location r_m of the interaction potential converge more rapidly than Legendre expansions of the potential itself.

As an illustration, consider the function used to fit an *ab initio* surface for N₂-He [86, 87]. It includes a repulsive term of the form

$$V_{\text{rep}}(r, \theta) = \exp[A(\theta) - b(\theta)R + \gamma(\theta) \ln r] \quad (\text{A } 1.5.66)$$

in which

$$A(\theta) = A_0 + A_2 P_2(\cos \theta) + A_4 P_4(\cos \theta) \quad (\text{A } 1.5.67)$$

and similar three-term Legendre expansions are used for $b(\theta)$ and $\gamma(\theta)$. The same surface includes an anisotropic attractive term consisting of damped dispersion and induction terms:

$$V_{\text{att}}(r, \theta) = - \sum_{n=3}^5 f_{2n}(r, \theta) C_{2n}(\theta) / r^{2n} \quad (\text{A } 1.5.68)$$

in which the combined dispersion and induction coefficients $C_{2n}(\theta)$ are given by Legendre series as in [equation \(A1.5.19\)](#), and the damping functions are given by a version of [equation \(A1.5.61\)](#) modified so that the scale factor has a weak angle dependence

$$\rho(\theta) = \rho_0 + \rho_2 P_2(\cos \theta). \quad (\text{A } 1.5.69)$$

To improve the description of the short-range anisotropy, the surface also includes a repulsive ‘site–site’ term

$$V_{\text{ssr}} = \sqrt{r_A} e^{Z-\zeta r_A} + \sqrt{r_B} e^{Z-\zeta r_B} \quad (\text{A 1.5.70})$$

-28-

where r_A and r_B are distances between the nitrogen atoms and the helium atom.

A1.5.5.4 SITE-SITE INTERMOLECULAR POTENTIALS

The approach described in [section A1.5.5.3](#) is best suited for accurate representations of the PES for interactions between small molecules. Interactions between large molecules are usually handled with an atom–atom or site–site approach. For example, an atom–atom, exp-6 potential for the interaction between molecules A and B can be written as

$$V_{\text{ss}} = \sum_{a \in A} \sum_{b \in B} [A_{ab} \exp(-b_{ab} r_{ab}) - C_6^{ab} / r_{ab}^6] \quad (\text{A 1.5.71})$$

where the sums are over the atoms of each molecule, and there are three parameters A_{ab} , b_{ab} and C_6^{ab} for each distinct type of atom pair. A set of parameters was developed by Filippini and Gavezzotti [[133](#), [134](#)] for describing crystal structures and another set for hydrogen bonding.

A more accurate approach is to begin with a model of the charge distribution for each of the molecules. Various prescriptions for obtaining point charge models, such as fitting to the electrostatic potential of the molecule [[135](#), [136](#)], are currently in use. Unfortunately, these point charge models are insufficiently accurate if only atom-centred charges are used [[137](#)]. Hence, additional charges are sometimes placed at off-atom sites. This increases the accuracy of the point charge model at the expense of arbitrariness in the choice of off-atom sites and an added computational burden. A less popular but sounder procedure is to use a distributed multipole model [[28](#), [138](#), [139](#)] instead of a point charge model.

Once the models for the charge distributions are in hand, the electrostatic interaction is computed as the interaction between the sets of point charges or distributed multipoles, and added to an atom–atom, exp-6 form that represents the repulsion and dispersion interactions. Different exp-6 parameters, often from [[140](#), [141](#) and [142](#)], are used in this case. The induction interaction is frequently omitted because it is small, or it is modelled by a single site polarizability on each molecule interacting with the point charges or distributed multipoles on the other.

A further refinement [[143](#), [144](#)] is to treat the atoms as being non-spherical by rewriting the repulsive part of the atom–atom exp-6 model, equation (A1.5.71), as

$$V_{\text{rep}} = V_{\text{ref}} \sum_{a \in A} \sum_{b \in B} \exp[-b_{ab}(\Omega_{ab})(r_{ab} - \rho_{ab}(\Omega_{ab}))] \quad (\text{A 1.5.72})$$

where Ω_{ab} is used as a generic designation for all the angles required to specify the relative orientation of the molecules, and V_{ref} is an energy unit. The $\rho_{ab}(\Omega_{ab})$ functions describe the shape of the contour on which the repulsion energy between atoms a and b equals V_{ref} . The spherical harmonic expansions used to represent the angular variation of the steepness $b_{ab}(\Omega_{ab})$ and shape $\rho_{ab}(\Omega_{ab})$ functions are quite rapidly convergent.

REFERENCES

- [1] Clausius R 1857 Über die Art von Bewegung, die wir Wärme nennen *Ann. Phys. Chem.* **100** 353
- [2] van der Waals J D 1873 Over de Continuïteit van den Gas- en Vloeistofoestand *PhD Thesis* Leiden
- [3] London F 1930 Zur theorie und systematik der molekularkräfte *Z. Phys.* **63** 245
- [4] London F 1937 The general theory of molecular forces *Trans. Faraday Soc.* **33** 8
- [5] Margenau H 1939 van der Waals forces *Rev. Mod. Phys.* **11** 1
- [6] Longuet-Higgins H C 1956 The electronic states of composite systems *Proc. R. Soc. A* **235** 537
- [7] Buckingham A D 1967 Permanent and induced molecular moments and long-range intermolecular forces *Adv. Chem. Phys.* **12** 107
- [8] Brooks F C 1952 Convergence of intermolecular force series *Phys. Rev.* **86** 92
- [9] Roe G M 1952 Convergence of intermolecular force series *Phys. Rev.* **88** 659
- [10] Dalgarno A and Lewis J T 1956 The representation of long-range forces by series expansions. I. The divergence of the series *Proc. Phys. Soc. A* **69** 57
- [11] Dalgarno A and Lewis J T 1956 The representation of long-range forces by series expansions. II. The complete perturbation calculation of long-range forces *Proc. Phys. Soc. A* **69** 59
- [12] Ahlrichs R 1976 Convergence properties of the intermolecular force series ($1/r$ expansion) *Theor. Chim. Acta* **41** 7
- [13] Morgan J D III and Simon B 1980 Behavior of molecular potential energy curves for large nuclear separations *Int. J. Quantum Chem.* **17** 1143
- [14] McClellan A L 1963 *Tables of Experimental Dipole Moments* vol 1 (New York: Freeman)
- [15] McClellan A L 1974 *Tables of Experimental Dipole Moments* vol 2 (El Cerrito, CA: Raha Enterprises)
- [16] McClellan A L 1989 *Tables of Experimental Dipole Moments* vol 3 (El Cerrito, CA: Raha Enterprises)
- [17] Sutter D H and Flygare W H 1976 The molecular Zeeman effect *Topics Curr. Chem.* **63** 89
- [18] Gray C G and Gubbins K E 1984 *Theory of Molecular Fluids. 1. Fundamentals* (Oxford: Clarendon)
- [19] Spackman M A 1992 Molecular electric moments from X-ray diffraction data *Chem. Rev.* **92** 1769
- [20] Dykstra C E 1988 *Ab initio Calculation of the Structures and Properties of Molecules* (Amsterdam: Elsevier)

- [21] Bündgen P, Grein F and Thakkar A J 1995 Dipole and quadrupole moments of small molecules. An *ab initio* study using perturbatively corrected, multi-reference, configuration interaction wavefunctions *J. Mol. Struct. (Theochem)* **334** 7

- [22] Doerksen R J and Thakkar A J 1999 Quadrupole and octopole moments of heteroaromatic rings *J. Phys. Chem. A* **103** 10 009
- [23] Miller T M and Bederson B 1988 Electric dipole polarizability measurements *Adv. At. Mol. Phys.* **25** 37
- [24] Shelton D P and Rice J E 1994 Measurements and calculations of the hyperpolarizabilities of atoms and small molecules in the gas phase *Chem. Rev.* **94** 3
- [25] Bonin K D and Kresin V V 1997 *Electric-dipole Polarizabilities of Atoms, Molecules and Clusters* (Singapore: World Scientific)
- [26] Doerksen R J and Thakkar A J 1999 Structures, vibrational frequencies and polarizabilities of diazaborinines, triazadiborinines, azaboroles and oxazaboroles *J. Phys. Chem. A* **103** 2141
- [27] Maroulis G 1999 On the accurate theoretical determination of the static hyperpolarizability of trans-butadiene *J. Chem. Phys.* **111** 583
- [28] Stone A J 1996 *The Theory of Intermolecular Forces* (New York: Oxford)
- [29] Legon A C, Millen D J and Mj6berg P J 1977 The hydrogen cyanide dimer: identification and structure from microwave spectroscopy *Chem. Phys. Lett.* **47** 589
- [30] Dulmage W J and Lipscomb W N 1951 The crystal structures of hydrogen cyanide, HCN *Acta Crystallogr.* **4** 330
- [31] Eisenschitz R and London F 1930 6ber das verh6ltnis der van der Waalschen kr6ften zu density hom6opolaren bindungskr6ften *Z. Phys.* **60** 491
- [32] London F 1930 6ber einige eigenschaften und anwendungen der molekularkr6fte *Z. Phys. Chem. B* **11** 222
- [33] Casimir H B G and Polder D 1948 The influence of retardation on the London-van der Waals forces *Phys. Rev.* **73** 360
- [34] Mavroyannis C and Stephen M J 1962 Dispersion forces *Mol. Phys.* **5** 629
- [35] McLachlan A D 1963 Retarded dispersion forces between molecules *Proc. R. Soc. A* **271** 387
- [36] Bethe H A and Salpeter E E 1957 *Quantum Mechanics of One- and Two-electron Atoms* (Berlin: Springer)
- [37] Margenau H 1931 Note on the calculation of van der Waals forces *Phys. Rev.* **37** 1425
- [38] Dalgarno A and Lynn N 1957 Properties of the helium atom *Proc. Phys. Soc. London* **70** 802
- [39] Dalgarno A and Kingston A E 1961 van der Waals forces for hydrogen and the inert gases *Proc. Phys. Soc. London* **78** 607

- [40] Zeiss G D, Meath W J, MacDonald J C F and Dawson D J 1977 Dipole oscillator strength distributions, sums, and some related properties for Li, N, O, H₂, N₂, O₂, NH₃, H₂O, NO and N₂O *J. Phys.* **55** 2080
- [41] Kumar A, Fairley G R G and Meath W J 1985 Dipole properties, dispersion energy coefficients and integrated oscillator strengths for SF₆ *J. Chem. Phys.* **83** 70
- [42] Kumar A and Meath W J 1992 Dipole oscillator strength properties and dispersion energies for acetylene and benzene *Mol. Phys.* **75** 311

- [43] Meath W J and Kumar A 1990 Reliable isotropic and anisotropic dipole dispersion energies, evaluated using constrained dipole oscillator strength techniques, with application to interactions involving H₂, N₂ and the rare gases *Int. J. Quantum Chem. Symp.* **24** 501
- [44] Kumar A, Meath W J, Bündgen P and Thakkar A J 1996 Reliable anisotropic dipole properties and dispersion energy coefficients for O₂, evaluated using constrained dipole oscillator strength techniques *J. Chem. Phys.* **105** 4927
- [45] Thakkar A J 1984 Bounding and estimation of van der Waals coefficients *J. Chem. Phys.* **81** 1919
- [46] Rijks W and Wormer P E S 1989 Correlated van der Waals coefficients. II. Dimers consisting of CO, HF, H₂O and NH₃ *J. Chem. Phys.* **90** 6507
- [47] Rijks W and Wormer P E S 1990 *Erratum*: correlated van der Waals coefficients. II. Dimers consisting of CO, HF, H₂O and NH₃ *J. Chem. Phys.* **92** 5754
- [48] Thakkar A J, Hettema H and Wormer P E S 1992 *Ab initio* dispersion coefficients for interactions involving rare-gas atoms *J. Chem. Phys.* **97** 3252
- [49] Wormer P E S and Hettema H 1992 Many-body perturbation theory of frequency-dependent polarizabilities and van der Waals coefficients: application to H₂O . . .H₂O and Ar. . .NH₃ *J. Chem. Phys.* **97** 5592
- [50] Hettema H, Wormer P E S and Thakkar A J 1993 Intramolecular bond length dependence of the anisotropic dispersion coefficients for interactions of rare gas atoms with N₂, CO, Cl₂, HCl and HBr *Mol. Phys.* **80** 533
- [51] Thakkar A J and Smith V H Jr 1974 On a representation of the long range interatomic interaction potential *J. Phys. B: At. Mol. Phys.* **7** L321
- [52] Tang K T and Toennies J P 1978 A simple model of the van der Waals potential at intermediate distances. II. Anisotropic potential of He. . .H₂ and Ne. . .H₂ *J. Chem. Phys.* **68** 5501
- [53] Thakkar A J 1988 Higher dispersion coefficients: accurate values for hydrogen atoms and simple estimates for other systems *J. Chem. Phys.* **89** 2092
- [54] Moelwyn-Hughes E A 1957 *Physical Chemistry* (New York: Pergamon) p 332
- [55] Tang K T 1969 Dynamic polarizabilities and van der Waals coefficients *Phys. Rev.* **177** 108

- [56] Kutzelnigg W and Maeder F 1978 Natural states of interacting systems and their use for the calculation of intermolecular forces. III. One-term approximations of oscillator strength sums and dynamic polarizabilities *Chem. Phys.* **35** 397
- [57] Wilson M and Madden P A 1994 Anion polarization and the stability of layered structures in MX₂ systems *J. Phys.: Condens. Matter* **6** 159
- [58] Axilrod P M and Teller E 1943 Interaction of the van der Waals type between three atoms *J. Chem. Phys.* **11** 299
- [59] Mutō Y 1943 Force between non-polar molecules *J. Phys. Math. Soc. Japan* **17** 629
- [60] McLachlan A D 1963 Three-body dispersion forces *Mol. Phys.* **6** 423
- [61] Bell R J 1970 Multipolar expansion for the non-additive third-order interaction energy of three atoms *J.*

- [62] Kutzelnigg W 1992 Does the polarization approximation converge for large- r to a primitive or a symmetry-adapted wavefunction? *Chem. Phys. Lett.* **195** 77
- [63] Cwiok T, Jeziorski B, Kolos W, Moszynski R, Rychlewski J and Szalewicz K 1992 Convergence properties and large-order behavior of the polarization expansion for the interaction energy of hydrogen atoms *Chem. Phys. Lett.* **195** 67
- [64] Claverie P 1971 Theory of intermolecular forces. I. On the inadequacy of the usual Rayleigh–Schrödinger perturbation method for the treatment of intermolecular forces *Int. J. Quantum Chem.* **5** 273
- [65] Claverie P 1978 Elaboration of approximate formulas for the interactions between large molecules: applications in organic chemistry *Intermolecular Interactions: From Diatomics to Biopolymers* ed B Pullman (New York: Wiley) p 69
- [66] Jeziorski B, Moszynski R and Szalewicz K 1994 Perturbation theory approach to intermolecular potential energy surfaces of van der Waals complexes *Chem. Rev.* **94** 1887
- [67] Adams W H 1994 The polarization approximation and the Amos–Musher intermolecular perturbation theories compared to infinite order at finite separation *Chem. Phys. Lett.* **229** 472
- [68] Bukowski R, Sadlej J, Jeziorski B, Jankowski P, Szalewicz K, Kucharski S A, Williams H L and Rice B M 1999 Intermolecular potential of carbon dioxide dimer from symmetry-adapted perturbation theory *J. Chem. Phys.* **110** 3785
- [69] Hayes I C and Stone A J 1984 An intermolecular perturbation theory for the region of moderate overlap *Mol. Phys.* **53** 83
- [70] Komasa J and Thakkar A J 1995 Accurate Heitler–London interaction energy for He₂ *J. Mol. Struct. (Theochem)* **343** 43
- [71] Kita S, Noda K and Inouye H 1976 Repulsion potentials for Cl⁻-R and Br⁻-R (R = He, Ne and Ar) derived from beam experiments *J. Chem. Phys.* **64** 3446

- [72] Kim Y S, Kim S K and Lee W D 1981 Dependence of the closed-shell repulsive interaction on the overlap of the electron densities *Chem. Phys. Lett.* **80** 574
- [73] Wheatley R J and Price S L 1990 An overlap model for estimating the anisotropy of repulsion *Mol. Phys.* **69** 507
- [74] Kreek H and Meath W J 1969 Charge-overlap effects. Dispersion and induction forces *J. Chem. Phys.* **50** 2289
- [75] Knowles P J and Meath W J 1986 Non-expanded dispersion energies and damping functions for Ar₂ and Li₂ *Chem. Phys. Lett.* **124** 164
- [76] Knowles P J and Meath W J 1986 Non-expanded dispersion and induction energies, and damping functions, for molecular interactions with application to HF...He *Mol. Phys.* **59** 965
- [77] Knowles P J and Meath W J 1987 A separable method for the calculation of dispersion and induction energy damping functions with applications to the dimers arising from He, Ne and HF *Mol. Phys.* **60** 1143
- [78] Wheatley R J and Meath W J 1993 Dispersion energy damping functions, and their relative scale with interatomic separation, for (H,He,Li)–(H,He,Li) interactions *Mol. Phys.* **80** 25
- [79] Wheatley R J and Meath W J 1994 Induction and dispersion damping functions, and their relative scale

with interspecies distance, for $(\text{H}^+, \text{He}^+, \text{Li}^+) - (\text{H}, \text{He}, \text{Li})$ interactions *Chem. Phys.* **179** 341

- [80] Stone A J 1993 Computation of charge-transfer energies by perturbation theory *Chem. Phys. Lett.* **211** 101
- [81] van Lenthe J H, van Duijneveldt-van de Rijdt J G C M and van Duijneveldt F B 1987 Weakly bonded systems. *Adv. Chem. Phys.* **69** 521
- [82] van Duijneveldt F B, van Duijneveldt-van de Rijdt J G C M and van Lenthe J H 1994 State of the art in counterpoise theory *Chem. Rev.* **94** 1873
- [83] Anderson J B, Traynor C A and Boghosian B M 1993 An exact quantum Monte-Carlo calculation of the helium–helium intermolecular potential *J. Chem. Phys.* **99** 345
- [84] Woon D E 1994 Benchmark calculations with correlated molecular wavefunctions. 5. The determination of accurate *ab initio* intermolecular potentials for He_2 , Ne_2 , and Ar_2 *J. Chem. Phys.* **100** 2838
- [85] Tao F M and Klemperer W 1994 Accurate *ab initio* potential energy surfaces of Ar-HF , $\text{Ar-H}_2\text{O}$, and Ar-NH_3 *J. Chem. Phys.* **101** 1129–45
- [86] Hu C H and Thakkar A J 1996 Potential energy surface for interactions between N_2 and He: *ab initio* calculations, analytic fits, and second virial coefficients *J. Chem. Phys.* **104** 2541
- [87] Reid J P, Thakkar A J, Barnes P W, Archibong E F, Quiney H M and Simpson C J S M 1997 Vibrational deactivation of N_2 ($\nu = 1$) by inelastic collisions with ^3He and ^4He : an experimental and theoretical study *J. Chem. Phys.* **107** 2329
- [88] Archibong E F, Hu C H and Thakkar A J 1998 Interaction potentials for He-F^- and Ne-F^- *J. Chem. Phys.* **109** 3072
-

- [89] Parr R G and Yang W 1989 *Density-functional Theory of Atoms and Molecules* (Oxford: Clarendon)
- [90] Pérez-Jordy J M and Becke A D 1995 A density functional study of van der Waals forces: rare gas diatomics *Chem. Phys. Lett.* **233** 134
- [91] Elrod M J and Saykally R J 1994 Many-body effects in intermolecular forces *Chem. Rev.* **94** 1975
- [92] Meath W J and Koulis M 1991 On the construction and use of reliable two- and many-body interatomic and intermolecular potentials *J. Mol. Struct. (Theochem)* **226** 1
- [93] Meuwly M and Hutson J M 1999 Morphing *ab initio* potentials: a systematic study of Ne-HF *J. Chem. Phys.* **110** 8338
- [94] 1994 van der Waals molecules *Chem. Rev.* **94**
- [95] 1994 Structure and dynamics of van der Waals complexes *Faraday Disc.* **97**
- [96] Child M S 1991 *Semiclassical Mechanics with Molecular Applications* (Oxford: Clarendon)
- [97] LeRoy R J and van Kranendonk J 1974 Anisotropic intermolecular potentials from an analysis of spectra of H_2 - and D_2 -inert gas complexes *J. Chem. Phys.* **61** 4750
- [98] LeRoy R J and Carley J S 1980 Spectroscopy and potential energy surfaces of van der Waals molecules *Adv. Chem. Phys.* **42** 353

- [99] LeRoy R J and Hutson J M 1987 Improved potential energy surfaces for the interaction of H₂ with Ar, Kr and Xe *J. Chem. Phys.* **86** 837
- [100] Hutson J M and Howard B J 1980 Spectroscopic properties and potential surfaces for atom–diatom van der Waals molecules *Mol. Phys.* **41** 1123
- [101] Hutson J M 1989 The intermolecular potential of Ne–HCl: determination from high-resolution spectroscopy *J. Chem. Phys.* **91** 4448
- [102] Hutson J M 1990 Intermolecular forces from the spectroscopy of van der Waals molecules *Ann. Rev. Phys. Chem.* **41** 123
- [103] van der Avoird A, Wormer P E S and Moszynski R 1994 From intermolecular potentials to the spectra of van der Waals molecules and vice versa *Chem. Rev.* **94** 1931
- [104] Buck U 1974 Inversion of molecular scattering data *Rev. Mod. Phys.* **46** 369
- [105] Buck U 1975 Elastic scattering *Adv. Chem. Phys.* **30** 313
- [106] Hirschfelder J O, Curtiss C F and Bird R B 1954 *Molecular Theory of Gases and Liquids* (New York: Wiley)
- [107] Mason E A and Spurling T H 1969 *The Virial Equation of State* (Oxford: Pergamon)
- [108] McQuarrie D A 1973 *Statistical Thermodynamics* (Mill Valley, CA: University Science Books)

-35-

- [109] Keller J B and Zumino B 1959 Determination of intermolecular potentials from thermodynamic data and the law of corresponding states *J. Chem. Phys.* **30** 1351
- [110] Frisch H L and Helfand E 1960 Conditions imposed by gross properties on the intermolecular potential *J. Chem. Phys.* **32** 269
- [111] Maitland G C, Rigby M, Smith E B and Wakeham W A 1981 *Intermolecular Forces: Their Origin and Determination* (Oxford: Clarendon)
- [112] Chapman S and Cowling T G 1970 *The Mathematical Theory of Non-uniform Gases* 3rd edn (London: Cambridge University Press)
- [113] McCourt F R, Beenakker J, Köhler W E and Kúscer I 1990 *Nonequilibrium Phenomena in Polyatomic Gases. 1. Dilute Gases* (Oxford: Clarendon)
- [114] McCourt F R, Beenakker J, Köhler W E and Kúscer I 1991 *Nonequilibrium Phenomena in Polyatomic Gases. 2. Cross-sections, Scattering and Rarefied Gases* (Oxford: Clarendon)
- [115] Buckingham A D and Fowler P W 1983 Do electrostatic interactions predict structures of van der Waals molecules? *J. Chem. Phys.* **79** 6426
- [116] Buckingham A D and Fowler P W 1985 A model for the geometries of van der Waals complexes *Can. J. Chem.* **63** 2018
- [117] Buckingham A D, Fowler P W and Stone A J 1986 Electrostatic predictions of shapes and properties of van der Waals molecules *Int. Rev. Phys. Chem.* **5** 107
- [118] Mie G 1903 Zur kinetischen theorie der einatomigen körper *Ann. Phys., Lpz* **11** 657
- [119] Morse P M 1929 Diatomic molecules according to the wave mechanics: II. Vibrational levels *Phys. Rev.* **34** 57

- [120] Pekeris C L 1934 The rotation–vibration coupling in diatomic molecules *Phys. Rev.* **45** 98
- [121] Slater J C and Kirkwood J G 1931 The van der Waals forces in gases *Phys. Rev.* **37** 682
- [122] Buckingham R A 1938 The classical equation of state of gaseous helium, neon and argon *Proc. R. Soc. A* **168** 264
- [123] Born M and Mayer J E 1932 Zur gittertheorie der ionenkristalle *Z. Phys.* **75** 1
- [124] Fender B E F and Halsey G D Jr 1962 Second virial coefficients of argon, krypton and argon–krypton mixtures at low temperatures *J. Chem. Phys.* **36** 1881
- [125] Aziz R A 1984 Interatomic potentials for rare-gases: pure and mixed interactions *Inert Gases: Potentials, Dynamics and Energy Transfer in Doped Crystals* ed M L Klein (Berlin: Springer) ch 2, pp 5–86
- [126] Buckingham R A and Corner J 1947 Tables of second virial and low-pressure Joule–Thompson coefficients for intermolecular potentials with exponential repulsion *Proc. R. Soc. A* **189** 118
-

-36-

- [127] Douketis C, Scoles G, Marchetti S, Zen M and Thakkar A J 1982 Intermolecular forces via hybrid Hartree–Fock SCF plus damped dispersion (HFD) energy calculations. An improved spherical model *J. Chem. Phys.* **76** 3057
- [128] Koide A, Meath W J and Allnatt A R 1981 Second-order charge overlap effects and damping functions for isotropic atomic and molecular interactions *Chem. Phys.* **58** 105
- [129] Tang K T and Toennies J P 1984 An improved simple model for the van der Waals potential based on universal damping functions for the dispersion coefficients *J. Chem. Phys.* **80** 3726
- [130] van der Avoird A, Wormer P E S, Mulder F and Berns R M 1980 *Ab initio* studies of the interactions in van der Waals molecules *Topics Curr. Chem.* **93** 1
- [131] Corner J 1948 The second virial coefficient of a gas of non-spherical molecules *Proc. R. Soc. A* **192** 275
- [132] Pack R T 1978 Anisotropic potentials and the damping of rainbow and diffraction oscillations in differential cross-sections *Chem. Phys. Lett.* **55** 197
- [133] Filippini G and Gavezzotti A 1993 Empirical intermolecular potentials for organic crystals: the 6-exp approximation revisited *Acta Crystallogr. B* **49** 868
- [134] Gavezzotti A and Filippini G 1994 Geometry of the intermolecular XH...Y (X,Y = N,O) hydrogen bond and the calibration of empirical hydrogen-bond potentials *J. Phys. Chem.* **98** 4831
- [135] Momany F A 1978 Determination of partial atomic charges from *ab initio* molecular electrostatic potentials. Application to formamide, methanol and formic acid *J. Phys. Chem.* **82** 592
- [136] Singh U C and Kollman P A 1984 An approach to computing electrostatic charges for molecules *J. Comput. Chem.* **5** 129
- [137] Wiberg K B and Rablen P R 1993 Comparison of atomic charges by different procedures *J. Comput. Chem.* **14** 1504
- [138] Stone A J 1981 Distributed multipole analysis; or how to describe a molecular charge distribution *Chem. Phys. Lett.* **83** 233
- [139] Stone A J and Alderton M 1985 Distributed multipole analysis—methods and applications *Mol. Phys.* **56** 1047
- [140] Williams D E 1965 Non-bonded potential parameters derived from crystalline aromatic hydrocarbons *J. Chem. Phys.* **45** 3770

- [141] Williams D E 1967 Non-bonded potential parameters derived from crystalline hydrocarbons *J. Chem. Phys.* **47** 4680
- [142] Mirsky K 1978 The determination of the intermolecular interaction energy by empirical methods *Computing in Crystallography* ed R Schenk *et al* (Delft, The Netherlands: Delft University) p 169
- [143] Stone A J 1979 Intermolecular forces *The Molecular Physics of Liquid Crystals* ed G R Luckhurst and G W Gray (New York: Academic) pp 31–50
- [144] Price S L and Stone A J 1980 Evaluation of anisotropic model intermolecular pair potentials using an *ab initio* SCF-CI surface *Mol. Phys.* **40** 805
-

FURTHER READING

Stone A J 1996 *The Theory of Intermolecular Forces* (New York: Oxford)

A fine text suitable for both graduate students and researchers. Emphasizes theory of long-range forces.

Maitland G C, Rigby M, Smith E B and Wakeham W A 1981 *Intermolecular Forces: Their Origin and Determination* (Oxford: Clarendon)

A thorough reference work with emphasis on the determination of intermolecular potentials from experimental data.

Scheiner S 1997 *Hydrogen Bonding: A Theoretical Perspective* (New York: Oxford)

A survey of research on hydrogen bonding with emphasis on theoretical calculations.

1994 van der Waals molecules *Chem. Rev.* **94** 1721

A special issue devoted to review articles on various aspects of van der Waals molecules.

Müller-Dethlefs K and Hobza P 2000 Noncovalent interactions: a challenge for experiment and theory *Chem. Rev.* **100** 143

A survey of challenges that have yet to be met.

Pykkö P 1997 Strong closed-shell interactions in inorganic chemistry *Chem. Rev.* **97** 597

A review of fertile ground for further research.

Margenau H and Kestner N R 1971 *Theory of Intermolecular Forces* 2nd edn (New York: Pergamon)

An older treatment that contains a wealth of references to the earlier literature, and an interesting history of the subject beginning with the work of Clairault in the mid-eighteenth century.

Pyykkö P 1997 Strong closed-shell interactions in inorganic chemistry *Chem. Rev.* **97** 597

A review of fertile ground for further research.

Margenau H and Kestner N R 1971 *Theory of Intermolecular Forces* 2nd edn (New York: Pergamon)

An older treatment that contains a wealth of references to the earlier literature, and an interesting history of the subject beginning with the work of Clairault in the mid-eighteenth century.

A1.6 Interaction of light with matter: a coherent perspective

David J Tannor

A1.6.1 THE BASIC MATTER-FIELD INTERACTION

There has been phenomenal expansion in the range of experiments connected with light-molecule interactions. If one thinks of light as an electromagnetic (EM) wave, like any wave it has an amplitude, a frequency and a phase. The advent of the laser in 1960 completely revolutionized the control over all three of these factors. The amplitude of the EM wave is related to its intensity; current laser capabilities allow intensities up to about 10^{20} W cm⁻², fifteen orders of magnitude larger than pre-laser technology allowed. Laser beams can be made extremely monochromatic. Finally, it is increasingly possible to control the absolute phase of the laser light. There have also been remarkable advances in the ability to construct ultrashort pulses. Currently it is possible to construct pulses of the order of 10^{-15} s (several femtoseconds), a time scale short compared with typical vibrational periods of molecules. These short pulses consist of a *coherent* superposition of many frequencies of the light; the word coherent implies a precise phase relationship between the different frequency components. When these coherent ultrashort pulses interact with a molecule they excite coherently many frequency components in the molecule. Such coherent excitation, whether it is with short pulses or with monochromatic light, introduces new concepts in thinking about the light-matter interaction. These new concepts can be used passively, to learn about molecular properties via new coherent spectroscopies, or actively, to control chemical reactions using light, or to use light to cool atoms and molecules to temperatures orders of magnitude lower than 1 K.

A theme which will run through this section is the complementarity of light and the molecule with which it interacts. The simplest example is energy: when a photon of energy $E = \hbar\omega$ is absorbed by a molecule it disappears, transferring the identical quantity of energy $E = \hbar(\omega_f - \omega_i)$ to the molecule. But this is only one of a complete set of such complementary relations: the amplitude of the EM field determines the amplitude of the excitation; the phase of the EM phase determines the phase of the excitation; and the time of the interaction with the photon determines the time of excitation of the molecule. Moreover, both the magnitude and direction of the momentum of the photon are imparted to the molecules, an observation which plays a crucial role in translational cooling. Finally, because of the conservation or increase in entropy in the universe, any entropy change in the system has to be compensated for by an entropy change in the light; specifically, coherent light has zero or low entropy while incoherent light has high entropy. Entropy exchange between the system and the light plays a fundamental role in laser cooling, where entropy from the system is carried off by the light via incoherent, spontaneous emission, as well in lasing itself where entropy from incoherent light

must be transferred to the system.

This section begins with a brief description of the basic light–molecule interaction. As already indicated, coherent light pulses excite coherent superpositions of molecular eigenstates, known as ‘wavepackets’, and we will give a description of their motion, their coherence properties, and their interplay with the light. Then we will turn to linear and nonlinear spectroscopy, and, finally, to a brief account of coherent control of molecular motion.

-2-

A1.6.1.1 ELECTROMAGNETIC FIELDS

The material in this section can be found in many textbooks and monographs. Our treatment follows that in [1, 2 and 3].

(A) MAXWELL'S EQUATIONS AND ELECTROMAGNETIC POTENTIALS

The central equations of electromagnetic theory are elegantly written in the form of four coupled equations for the electric and magnetic fields. These are known as Maxwell's equations. In free space, these equations take the form:

$$\nabla \times \mathbf{E} = -\frac{1}{c} \frac{d\mathbf{B}}{dt} \quad (\text{A1.6.1})$$

(A1.6.2)

$$\nabla \cdot \mathbf{E} = 4\pi\rho \quad (\text{A1.6.3})$$

$$\nabla \cdot \mathbf{B} = 0 \quad (\text{A1.6.4})$$

where \mathbf{E} is the electric field vector, \mathbf{B} is the magnetic field vector, \mathbf{J} is the current density, ρ is the charge density and c is the speed of light. It is convenient to define two potentials, a scalar potential ϕ and a vector potential \mathbf{A} , such that the electric and magnetic fields are defined in terms of derivatives of these potentials. The four Maxwell equations are then replaced by two equations which define the fields in terms of the potentials,

$$\mathbf{E} = -\nabla\phi - \frac{1}{c} \frac{d\mathbf{A}}{dt} \quad (\text{A1.6.5})$$

$$\mathbf{B} = \nabla \times \mathbf{A} \quad (\text{A1.6.6})$$

together with two equations for the vector and scalar fields themselves. Note that there is a certain amount of flexibility in the choice of \mathbf{A} and ϕ , such that the same values for \mathbf{E} and \mathbf{B} are obtained (called gauge invariance). We will adopt below the Coulomb gauge, in which $\nabla \cdot \mathbf{A} = 0$.

In free space ($\rho = 0$, $\mathbf{J} = 0$, $\phi = \text{constant}$), the equations for the potentials decouple and take the following simple form:

$$\nabla^2\phi = 0 \quad (\text{A1.6.7})$$

$$\nabla^2 \mathbf{A} = \frac{1}{c^2} \frac{d^2 \mathbf{A}}{dt^2}. \quad (\text{A1.6.8})$$

-3-

Equation (A1.6.8), along with the definitions (A1.6.5) and (A1.6.6) constitute the central equation for the propagation of electromagnetic waves in free space. The form of section A1.6.4 admits harmonic solutions of the form

$$\mathbf{A} \equiv \mathbf{A}_0 \cos(\mathbf{k} \cdot \mathbf{r} - \omega t) \quad (\text{A1.6.9})$$

from which it follows that

$$\mathbf{E} = -\frac{\omega}{c} \mathbf{A}_0 \sin(\mathbf{k} \cdot \mathbf{r} - \omega t) \equiv E_0 \mathbf{e} \sin(kr - \omega t) \quad (\text{A1.6.10})$$

$$\mathbf{B} = -\mathbf{A}_0 (\mathbf{k} \times \mathbf{e}) \sin(\mathbf{k} \cdot \mathbf{r} - \omega t) \quad (\text{A1.6.11})$$

(\mathbf{e} is a unit vector in the direction of \mathbf{E} and $\omega = kc$).

(B) ENERGY AND PHOTON NUMBER DENSITY

In what follows it will be convenient to convert between field strength and numbers of photons in the field. According to classical electromagnetism, the energy E in the field is given by

$$E = \int d^3r \frac{\mathbf{E}^2 + \mathbf{B}^2}{8\pi}. \quad (\text{A1.6.12})$$

If we assume a single angular frequency of the field, ω , and a constant magnitude of the vector potential, A_0 , in the volume V , we obtain, using equation (A1.6.10) and equation (A1.6.11), and noting that the average value of $\sin^2(x) = 1/2$,

$$E = V \frac{E_0^2}{8\pi}. \quad (\text{A1.6.13})$$

But by the Einstein relation we know that the energy of a single photon on frequency ω is given by $\hbar\omega$, and hence the total energy in the field is

$$E = N\hbar\omega \quad (\text{A1.6.14})$$

where N is the number of photons. Combining equation (A1.6.13) and equation (A1.6.14) we find that

$$E_0 = \left(\frac{8\pi N\hbar\omega}{V} \right)^{1/2}. \quad (\text{A1.6.15})$$

-4-

Equation (A1.6.15) provides the desired relationship between field strength and the number of photons.

A1.6.1.2 INTERACTION BETWEEN FIELD AND MATTER

(A) CLASSICAL THEORY

To this point, we have considered only the radiation field. We now turn to the interaction between the matter and the field. According to classical electromagnetic theory, the force on a particle with charge e due to the electric and magnetic fields is

$$\mathbf{F} = e \left(\mathbf{E} + \frac{\mathbf{v} \times \mathbf{B}}{c} \right). \quad (\text{A1.6.16})$$

This interaction can also be expressed in terms of a Hamiltonian:

$$H(\mathbf{p}, \mathbf{A}) = \frac{1}{2m} \left(\mathbf{p} - \frac{e}{c} \mathbf{A} \right)^2 \quad (\text{A1.6.17})$$

where $\mathbf{A} = \mathbf{A}(x)$ and where \mathbf{p} and x are the conjugate variables that obey the canonical Hamilton equations. (Verifying that equation (A1.6.17) reduces to equation (A1.6.16) is non-trivial (cf [3])). Throughout the remainder of this section the radiation field will be treated using classical electromagnetic theory, while the matter will be treated quantum mechanically, that is, a ‘semiclassical’ treatment. The Hamiltonian form for the interaction, equation (A1.6.17), provides a convenient starting point for this semiclassical treatment.

(B) QUANTUM HAMILTONIAN FOR A PARTICLE IN AN ELECTROMAGNETIC FIELD

To convert the Hamiltonian for the material from a classical to a quantum form, we simply replace \mathbf{p} with $-i\hbar \nabla$. This gives:

$$H = \frac{1}{2m} \left[-i\hbar \nabla - \frac{e}{c} \mathbf{A} \right]^2 + V_s \quad (\text{A1.6.18})$$

$$= \underbrace{-\frac{\hbar^2 \nabla^2}{2m} + V_s}_{H_0} + \underbrace{\frac{i\hbar e}{2mc} (\nabla \cdot \mathbf{A} + \mathbf{A} \cdot \nabla) + \frac{e^2}{2mc} \mathbf{A} \cdot \mathbf{A}}_V \quad (\text{A1.6.19})$$

$$= H_0 + V \quad (\text{A1.6.20})$$

-5-

where H_0 is the Hamiltonian of the bare system and V is the part of the Hamiltonian that comes from the radiation field and the radiation–matter interaction. Note that an additional term, V_s , has been included in the system Hamiltonian, to allow for internal potential energy of the system. V_s contains all the interesting features that make different atoms and molecules distinct from one another, and will play a significant role in later sections.

We now make the following observations.

- (i) For many charged particles

$$V = \sum_i \frac{i\hbar e}{2mc} (\nabla \cdot \mathbf{A} + \mathbf{A} \cdot \nabla) + \frac{e^2}{2mc} \mathbf{A} \cdot \mathbf{A}.$$

- (ii) In the Coulomb gauge $\nabla \cdot \mathbf{A} = 0$. This implies that $\nabla \cdot (\mathbf{A}\psi) = \mathbf{A} \cdot \nabla\psi$ for any ψ , and hence the terms linear in \mathbf{A} can be combined:

$$\nabla \cdot \mathbf{A} + \mathbf{A} \cdot \nabla = 2\mathbf{A} \cdot \nabla.$$

- (iii) The quadratic term in \mathbf{A} ,

$$\frac{e^2}{2m} c^2 \mathbf{A} \cdot \mathbf{A}$$

can be neglected except for very strong fields, on the order of 10^{15} W cm⁻² [4].

- (iv) For isolated molecules, it is generally the case that the wavelength of light is much larger than the molecular dimensions. In this case it is a good approximation to make the replacement $e^{ik \cdot r} \approx 1$, which allows the replacement [3]

$$V = -\frac{e}{mc} \mathbf{A} \cdot \hat{\mathbf{p}} = -\mathbf{E} \cdot e\hat{\mathbf{r}}.$$

For many electrons and nuclei, V takes the following form:

$$V = -\mathbf{E} \cdot \sum_i Z_i e\hat{\mathbf{r}}_i = -\mathbf{E} \cdot \tilde{\boldsymbol{\mu}} \quad (\text{A1.6.21})$$

where we have defined the dipole operator, $\tilde{\boldsymbol{\mu}} \equiv \sum_i Z_i e\hat{\mathbf{r}}_i$. The dipole moment is seen to be a product of charge and distance, and has the physical interpretation of the degree of charge separation in the atom or molecule. Note that for not-too-intense fields, equation (A1.6.21) is the dominant term in the radiation–matter interaction; this is the dipole approximation.

A1.6.1.3 ABSORPTION, STIMULATED EMISSION AND SPONTANEOUS EMISSION OF LIGHT

Consider a quantum system with two levels, a and b , with energy levels E_a and E_b . Furthermore, let the perturbation

-6-

between these levels be of the form equation (A1.6.21), with monochromatic light, that is, $\mathbf{E} = \mathbf{E}_0 \cos(\omega t)$ resonant to the transition frequency between the levels, so $\omega = (E_b - E_a)/\hbar \equiv E_{ba}/\hbar$. The perturbation matrix element between a and b is then given by

$$V_{ba} = \mathbf{E}_0 \cos(\omega t) \cdot \langle b | \tilde{\boldsymbol{\mu}} | a \rangle = \frac{\mathbf{E}}{2} (\mathbf{e}^{i\omega t} + \mathbf{e}^{-i\omega t}) \cdot \tilde{\boldsymbol{\mu}}_{ba} \quad (\text{A1.6.22})$$

where

$$\tilde{\boldsymbol{\mu}}_{ab} \equiv \langle b | \tilde{\boldsymbol{\mu}} | a \rangle = \tilde{\boldsymbol{\mu}}_{ba}$$

is the dipole matrix element. There are three fundamental possible kinds of transitions connected by the dipole interaction: absorption ($a \rightarrow b$), corresponding to the second term in equation (A1.6.22); stimulated emission ($b \rightarrow a$) governed by the first term in equation (A1.6.22); and spontaneous emission (also $b \rightarrow a$), for which there is no term in the classical radiation field. For a microscopic description of the latter, a quantum

mechanical treatment of the radiation field is required. Nevertheless, there is a simple prescription for taking spontaneous emission into account, which was derived by Einstein during the period of the old quantum theory on the basis of considerations of thermal equilibrium between the matter and the radiation. Although for most of the remainder of this section the assumption of thermal equilibrium will not be satisfied, it is convenient to invoke it here to quantify spontaneous emission.

Fermi's Golden Rule expresses the rate of transitions between b and a as

$$W = \frac{2\pi}{\hbar} |V_{ba}|^2 \rho(E_{ba}) \quad (\text{A1.6.23})$$

where $\rho(E_{ba})$ is the density of final states for both the system and the light. As described above, we will consider the special case of both the matter and light at thermal equilibrium. The system final state is by assumption non-degenerate, but there is a frequency dependent degeneracy factor for thermal light, $\rho(E) dE$, where

$$\rho(E) = \frac{V}{(2\pi c)^3} \frac{\omega^2}{\hbar} d\Omega \quad (\text{A1.6.24})$$

and V is the volume of the 'box' and Ω is an element of solid angle.

The thermal light induces transitions from $a \rightarrow b$ and from $b \rightarrow a$ in proportion to the number of photons present. The number of transitions per second induced by absorption is

$$W_{\text{abs}}(a \rightarrow b) = \frac{2\pi}{\hbar} |V_{ab}|^2 \rho(E_b - \hbar\omega) \quad (\text{A1.6.25})$$

-7-

$$= \frac{2\pi}{\hbar} \frac{E_0^2}{4} |\mathbf{e} \cdot \bar{\mu}_{ab}|^2 \frac{V}{(2\pi c)^3} \frac{\omega^2}{\hbar} d\Omega. \quad (\text{A1.6.26})$$

Integrating over all solid angles and using [equation \(A1.6.15\)](#) and [equation \(A1.6.10\)](#) we find

$$W_{\text{abs}}(a \rightarrow b) = \frac{4}{3\hbar} N \frac{\omega^3}{c^3} |\mu_{ab}|^2. \quad (\text{A1.6.27})$$

For thermal light, the number of transitions per second induced by stimulated emission integrated over solid angles, W_{stim} , is equal to W_{abs} . The total emission, which is the sum of the stimulated and spontaneous emission, may be obtained by letting $N \rightarrow N + 1$ in the expression for stimulated emission, giving

$$W_{\text{em}}(b \rightarrow a) = \frac{4}{3\hbar} (N + 1) \frac{\omega^3}{c^3} |\mu_{ab}|^2. \quad (\text{A1.6.28})$$

Einstein's original treatment [5] used a somewhat different notation, which is still in common use:

$$W_{\text{stim}}(b \rightarrow a) \equiv B(b \rightarrow a) \bar{\rho} = W_{\text{abs}}(a \rightarrow b)$$

where

$$\bar{\rho} = \frac{2hN\hbar\omega}{V}\rho(E) = \frac{2N\omega^3\hbar}{\pi c^3}$$

is the energy in the field per unit volume between frequencies ν and $\nu + d\nu$ (the ‘radiation density’) (the factor of 2 comes from the two polarizations of the light and the factor h from the scaling between energy and frequency). Comparing with equation (A1.6.27) leads to the identification

$$B(b \rightarrow a) = B(a \rightarrow b) = \frac{2\pi}{3\hbar^2} |\mu_{ab}|^2.$$

Moreover, in Einstein’s treatment

$$W_{\text{spont}}(b \rightarrow a) \equiv A(b \rightarrow a) = \frac{4}{3\hbar} \left(\frac{\omega}{c}\right)^3 |\mu_{ab}|^2$$

leading to the following ratio of the Einstein A and B coefficients:

$$\frac{A(b \rightarrow a)}{B(b \rightarrow a)} = \frac{2\hbar}{\pi} \left(\frac{\omega}{c}\right)^3. \quad (\text{A1.6.29})$$

-8-

The argument is sometimes given that [equation \(A1.6.29\)](#) implies that the ratio of spontaneous to stimulated emission goes as the cube of the emitted photon frequency. This argument must be used with some care: recall that for light at thermal equilibrium, W_{stim} goes as $B\bar{\rho}$, and hence the rate of stimulated emission has a factor of $(\omega/c)^3$ coming from $\bar{\rho}$. The ratio of the spontaneous to the stimulated emission rates is therefore frequency independent! However, for non-thermal light sources (e.g. lasers), only a small number of energetically accessible states of the field are occupied, and the $\bar{\rho}$ factor is on the order of unity. The rate of spontaneous emission still goes as ω^3 , but the rate of stimulated emission goes as ω , and hence the ratio of spontaneous to stimulated emission goes as ω^2 . Thus, for typical light sources, spontaneous emission dominates at frequencies in the UV region and above, while stimulated emission dominates at frequencies in the far-IR region and below, with both processes participating at intermediate frequencies.

A1.6.1.4 INTERACTION BETWEEN MATTER AND FIELD

In the previous sections we have described the interaction of the electromagnetic field with matter, that is, the way the material is affected by the presence of the field. But there is a second, reciprocal perspective: the excitation of the material by the electromagnetic field generates a dipole (polarization) where none existed previously. Over a sample of finite size this dipole is macroscopic, and serves as a new source term in Maxwell’s equations. For weak fields, the source term, \mathbf{P} , is linear in the field strength. Thus,

$$\mathbf{P} = \chi \mathbf{E} \quad (\text{A1.6.30})$$

where the proportionality constant χ , called the (linear) susceptibility, is generally frequency dependent and complex. As we shall see below, the imaginary part of the linear susceptibility determines the absorption spectrum while the real part determines the dispersion, or refractive index of the material. There is a universal relationship between the real part and the imaginary part of the linear susceptibility, known as the Kramers–Kronig relation, which establishes a relationship between the absorption spectrum and the frequency-dependent refractive index. With the addition of the source term \mathbf{P} , Maxwell’s equations still have wavelike

solutions, but the relation between frequency and wavevector in [equation \(A1.6.10\)](#) must be generalized as follows:

$$\left(\frac{kc}{\omega}\right)^2 = 1 \rightarrow \left(\frac{kc}{\omega}\right)^2 = 1 + \chi. \quad (\text{A1.6.31})$$

The quantity $1 + \chi$ is known as the dielectric constant, ϵ ; it is constant only in the sense of being independent of E , but is generally dependent on the frequency of E . Since χ is generally complex so is the wavevector k . It is customary to write

$$\frac{kc}{\omega} = \eta + i\kappa \quad (\text{A1.6.32})$$

where η and κ are the refractive index and extinction coefficient, respectively. The travelling wave solutions to Maxwell's equations, propagating in the z -direction now take the form

-9-

$$\exp(i(kz - \omega t)) = \exp\left[i\omega\left(\frac{\eta z}{c} - t\right) - \left(\frac{\omega\kappa z}{c}\right)\right]. \quad (\text{A1.6.33})$$

In this form it is clear that κ leads to an attenuation of the electric field amplitude with distance (i.e. absorption).

For stronger fields the relationship between the macroscopic polarization and the incident field is non-linear. The general relation between P and E is written as

$$P = \chi^{(1)}E + \chi^{(2)} : E^2 + \chi^{(3)} : E^3 + \dots \equiv P^{(1)} + P^{(2)} + P^{(3)} + \dots. \quad (\text{A1.6.34})$$

The microscopic origin of χ and hence of P is the non-uniformity of the charge distribution in the medium. To lowest order this is given by the dipole moment, which in turn can be related to the dipole moments of the component molecules in the sample. Thus, on a microscopic quantum mechanical level we have the relation

$$P = \langle \psi | \bar{\mu} | \psi \rangle. \quad (\text{A1.6.35})$$

Assuming that the material has no permanent dipole moment, P originates from changes in the wavefunction ψ that are induced by the field; this will be our starting point in [section A1.6.4](#).

A1.6.2 COHERENCE PROPERTIES OF LIGHT AND MATTER

In the previous section we discussed light and matter at equilibrium in a two-level quantum system. For the remainder of this section we will be interested in light and matter which are not at equilibrium. In particular, laser light is completely different from the thermal radiation described at the end of the previous section. In the first place, only one, or a small number of states of the field are occupied, in contrast with the Planck distribution of occupation numbers in thermal radiation. Second, the field state can have a precise *phase*; in thermal radiation this phase is assumed to be random. If multiple field states are occupied in a laser they can have a precise phase relationship, something which is achieved in lasers by a technique called 'mode-locking'. Multiple frequencies with a precise phase relation give rise to laser pulses in time. Nanosecond experiments

have been very useful in probing, for example, radiationless transitions, intramolecular dynamics and radiative lifetimes of single vibronic levels in molecules. Picosecond experiments have been useful in probing, for example, collisional relaxation times and rotational reorientation times in solutions. Femtosecond experiments have been useful in observing the real time breaking and formation of chemical bonds; such experiments will be described in the next section. Any time that the phase is precisely correlated in time over the duration of an experiment, or there is a superposition of frequencies with well-defined relative phases, the process is called coherent. Single frequency coherent processes will be the major subject of section A1.6.2, while multifrequency coherent processes will be the focus for the remainder of the section.

-10-

A1.6.2.1 WAVEPACKETS: SOLUTIONS OF THE TIME-DEPENDENT SCHRÖDINGER EQUATION

The central equation of (non-relativistic) quantum mechanics, governing an isolated atom or molecule, is the time-dependent Schrödinger equation (TDSE):

$$i\hbar \frac{\partial \psi(x, t)}{\partial t} = H \psi(x, t). \quad (\text{A1.6.36})$$

In this equation H is the Hamiltonian (developed in the previous section) which consists of the bare system Hamiltonian and a term coming from the interaction between the system and the light. That is,

$$H = -\frac{\hbar^2}{2m} \nabla^2 + V_s(x) - E(t)\mu. \quad (\text{A1.6.37})$$

Since we are now interested in the possibility of coherent light, we have taken the interaction between the radiation and matter to be some general time-dependent interaction, $V = -E(t)\mu$, which could in principle contain many frequency components. At the same time, for simplicity, we neglect the vector character of the electric field in what follows. The vector character will be reintroduced in [section A1.6.4](#), in the context of nonlinear spectroscopy.

Real molecules in general have many quantum levels, and the TDSE can exhibit complicated behaviour even in the absence of a field. To simplify matters, it is worthwhile discussing some properties of the solutions of the TDSE in the absence of a field and then reintroducing the field. First let us consider

$$H = -\frac{\hbar^2}{2m} \nabla^2 + V_s(x). \quad (\text{A1.6.38})$$

Since in this case the Hamiltonian is time independent, the general solution can be written as

$$\Psi(x, t) = \sum_{n=1}^{\infty} a_n \psi_n(x) e^{-(i/\hbar)E_n t}. \quad (\text{A1.6.39})$$

(This expression assumes a system with a discrete level structure; for systems with both a discrete and a continuous portion to their spectrum the expression consists of a sum over the discrete states and an integral over the continuous states.) Here, $\psi_n(x)$ is a solution of the time-independent Schrödinger equation,

$$H \psi_n(x) = E_n \psi_n(x),$$

with eigenvalue E_n . The coefficients, a_n , satisfy the normalization condition $\sum_n |a_n|^2 = 1$, and are time

independent in this case. Equation (A1.6.39) describes a moving *wavepacket*, that is, a state whose average values in coordinate and momentum change with time. To see this, note that according to quantum mechanics $|\psi(x,t)|^2 dx$ is the probability to find the particle between x and $x + dx$ at time t . Using equation (A1.6.39) we see that

-11-

$$|\Psi(x, t)|^2 = \sum_{m,n=1}^{\infty} a_m^* a_n \psi_m^*(x) \psi_n(x) e^{-(i/\hbar)(E_n - E_m)t}$$

in other words, the probability density has a non-vanishing time dependence so long as there are components of two or more different energy eigenstates.

One of the remarkable features of time evolution of wavepackets is the close connection they exhibit with the motion of a classical particle. Specifically, Ehrenfest's theorem indicates that for potentials up to quadratic, the average value of position and momentum of the quantum wavepacket as a function of time is exactly the same as that of a classical particle on the same potential that begins with the corresponding initial conditions in position and momentum. This classical-like behaviour is illustrated in figure A1.6.1 for a displaced Gaussian wavepacket in a harmonic potential. For the case shown, the initial width is the same as the ground-state width, a 'coherent state', and hence the Gaussian moves without spreading. By way of contrast, if the initial Gaussian has a different width parameter, the centre of the Gaussian still satisfies the classical equations of motion; however, the width will spread and contract periodically in time, twice per period.

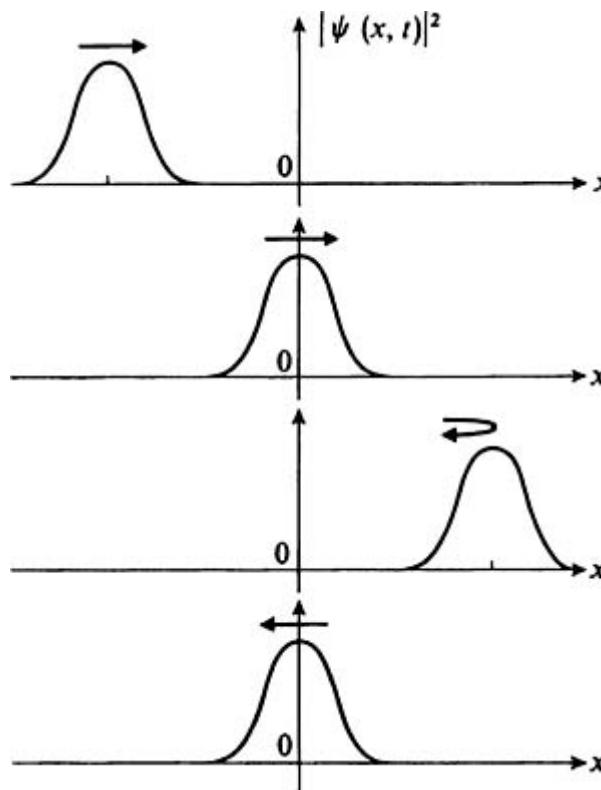


Figure A1.6.1. Gaussian wavepacket in a harmonic oscillator. Note that the average position and momentum change according to the classical equations of motion (adapted from [6]).

A1.6.2.2 COHERENCE IN A TWO-LEVEL SYSTEM: THE RABI SOLUTION

We now add the field back into the Hamiltonian, and examine the simplest case of a two-level system coupled to coherent, monochromatic radiation. This material is included in many textbooks (e.g. [6, 7, 8, 9, 10 and 11]). The system is described by a Hamiltonian H_0 having only two eigenstates, ψ_a and ψ_b , with energies $E_a = \hbar\omega_a$ and $E_b = \hbar\omega_b$. Define $\omega_0 = \omega_b - \omega_a$. The most general wavefunction for this system may be written as

$$\Psi(t) = a(t)e^{-i\omega_a t} \psi_a + b(t) e^{-i\omega_b t} \psi_b. \quad (\text{A1.6.40})$$

The coefficients $a(t)$ and $b(t)$ are subject to the constraint that $|a(t)|^2 + |b(t)|^2 = 1$. If we couple this system to a light field, represented as $V = -\mu_{ab}E \cos(\omega t)$, then we may write the TDSE in matrix form as

$$i\hbar \frac{d}{dt} \begin{pmatrix} a(t)e^{-i\omega_a t} \\ b(t)e^{-i\omega_b t} \end{pmatrix} = \begin{pmatrix} E_a & -\mu_{ab}E \cos(\omega t) \\ -\mu_{ab}E \cos(\omega t) & E_b \end{pmatrix} \begin{pmatrix} a(t)e^{-i\omega_a t} \\ b(t)e^{-i\omega_b t} \end{pmatrix}. \quad (\text{A1.6.41})$$

To continue we define a detuning parameter, $\Delta \equiv \omega - \omega_0$. If $\Delta \ll \omega_0$ then $\exp(-i(\omega - \omega_0)t)$ is slowly varying while $\exp(-i(\omega + \omega_0)t)$ is rapidly varying and cannot transfer much population from state A to state B. We therefore ignore the latter term; this is known as the ‘rotating wave approximation’. If we choose as initial conditions $|a(0)|^2 = 1$ and $|b(0)|^2 = 0$ then the solution of equation (A1.6.41) is

$$a(t) = e^{+\frac{i}{2}\Delta t} \left(\cos\left(\frac{1}{2}\Omega t\right) - i\frac{\Delta}{\Omega} \sin\left(\frac{1}{2}\Omega t\right) \right) \quad (\text{A1.6.42})$$

$$b(t) = e^{-\frac{i}{2}\Delta t} \left(\frac{\mu E}{2\hbar\Omega} \right) \left(2i \sin\left(\frac{1}{2}\Omega t\right) \right). \quad (\text{A1.6.43})$$

where the Rabi frequency, Ω , is defined as

$$\Omega = \sqrt{\Delta^2 + \left(\frac{\mu E}{\hbar}\right)^2}. \quad (\text{A1.6.44})$$

The populations as functions of time are then

$$|a(t)|^2 = \left(\frac{\Delta}{\Omega}\right)^2 + \left(\frac{\mu E}{\hbar\Omega}\right)^2 \cos^2\left(\frac{1}{2}\Omega t\right) \quad (\text{A1.6.45})$$

$$|b(t)|^2 = \left(\frac{\mu E}{\hbar\Omega}\right)^2 \sin^2\left(\frac{1}{2}\Omega t\right). \quad (\text{A1.6.46})$$

The population in the upper state as a function of time is shown in figure A1.6.2. There are several important things to note. At early times, resonant and non-resonant excitation produce the same population in the upper state because, for short times, the population in the upper state is independent of the Rabi frequency:

$$|b(t)|^2 = \left(\frac{\mu E}{\hbar\Omega}\right)^2 \sin^2\left(\frac{1}{2}\Omega t\right) \xrightarrow{t \text{ small}} \left(\frac{\mu E}{2\hbar}\right)^2 t^2. \quad (\text{A1.6.47})$$

One should also notice that resonant excitation completely cycles the population between the lower and upper state with a period of $2\pi/\Omega$. Non-resonant excitation also cycles population between the states but never completely depopulates the lower state. Finally, one should notice that non-resonant excitation cycles population between the two states at a faster rate than resonant excitation.

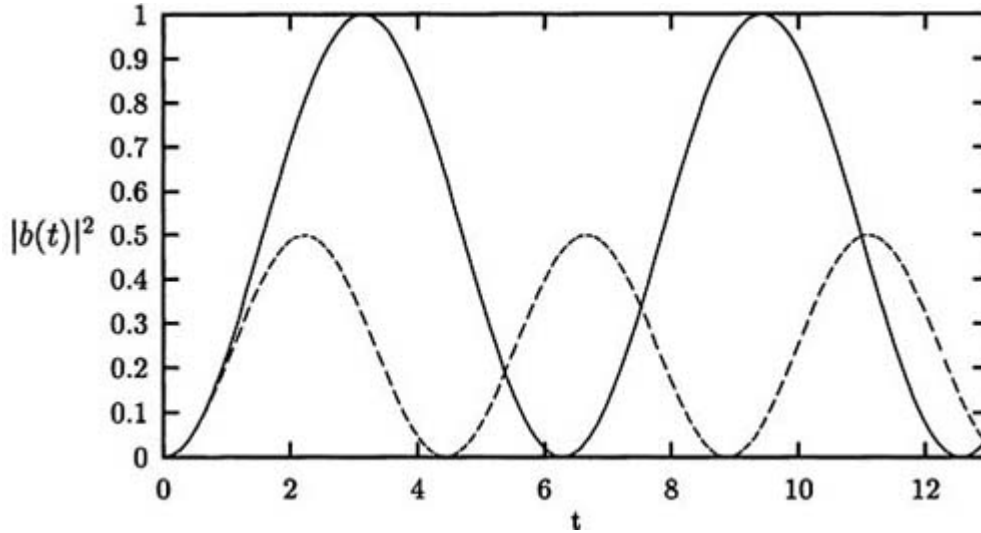


Figure A1.6.2. The population in the upper state as a function of time for resonant excitation (full curve) and for non-resonant excitation (dashed curve).

A1.6.2.3 GEOMETRICAL REPRESENTATION OF THE EVOLUTION OF A TWO-LEVEL SYSTEM

A more intuitive, and more general, approach to the study of two-level systems is provided by the Feynman–Vernon–Hellwarth geometrical picture. To understand this approach we need to first introduce the density matrix.

In the Rabi solution of the previous section we considered a wavefunction $\psi(t)$ of the form

$$\Psi(t) = a(t)e^{-i\omega_a t} \psi_a + b(t)e^{-i\omega_b t} \psi_b. \quad (\text{A1.6.48})$$

-14-

We saw that the time-dependent populations in each of the two levels is given by $P_a = |a(t)|^2$ and $P_b = |b(t)|^2$. So long as the field is on, these populations continue to change; however, once the external field is turned off, these populations remain constant (discounting relaxation processes, which will be introduced below). Yet the *amplitudes* in the states ψ_a and ψ_b do continue to change with time, due to the accumulation of time-dependent phase factors during the field-free evolution. We can obtain a convenient separation of the time-dependent and the time-independent quantities by defining a density matrix, ρ . For the case of the wavefunction $|\psi\rangle$, ρ is given as the ‘outer product’ of $|\psi\rangle$ with itself,

$$\rho \equiv |\Psi\rangle\langle\Psi|. \quad (\text{A1.6.49})$$

This outer product gives four terms, which may be arranged in matrix form as

$$\rho = \begin{pmatrix} |a|^2 & a^*b e^{-i(\omega_b - \omega_a)t} \\ ab^* e^{i(\omega_b - \omega_a)t/\hbar} & |b|^2 \end{pmatrix}. \quad (\text{A1.6.50})$$

Note that the diagonal elements of the matrix, $|a|^2$ and $|b|^2$, correspond to the *populations* in the energy levels, a and b , and contain no time dependence, while the off-diagonal elements, called the *coherences*, contain all the time dependence.

A differential equation for the time evolution of the density operator may be derived by taking the time derivative of equation (A1.6.49) and using the TDSE to replace the time derivative of the wavefunction with the Hamiltonian operating on the wavefunction. The result is called the Liouville equation, that is,

$$i\hbar \frac{\partial \rho}{\partial t} = [H, \rho]. \quad (\text{A1.6.51})$$

The strategy for representing this differential equation geometrically is to expand both H and ρ in terms of the three Pauli spin matrices, σ_1 , σ_2 and σ_3 and then view the coefficients of these matrices as time-dependent vectors in three-dimensional space. We begin by writing the two-level system Hamiltonian in the following general form,

$$H = \begin{pmatrix} E_b & V_{ba} \\ V_{ab} & E_a \end{pmatrix} \quad (\text{A1.6.52})$$

where we take the radiation–matter interaction to be of the dipole form, but allow for arbitrary time-dependent electric fields:

$$V_{ba} = -\mu_{ba} E(t). \quad (\text{A1.6.53})$$

Moreover, we will write the density matrix for the system as

$$\rho = \begin{pmatrix} bb^* & a^*b \\ ab^* & aa^* \end{pmatrix} \quad (\text{A1.6.54})$$

-15-

where a and b now contain the bare system evolution phase factors. We proceed to express both the Hamiltonian and the density matrix in terms of the standard Pauli spin matrices:

$$H = \overbrace{(V_{ab} + V_{ba})}^{E_1} \sigma_1 + i \overbrace{(V_{ba} - V_{ab})}^{E_2} \sigma_2 + \overbrace{(E_b - E_a)}^{E_3} \sigma_3$$

$$\rho = \underbrace{(ab^* + a^*b)}_{r_1} \sigma_1 + i \underbrace{(a^*b - ab^*)}_{r_2} \sigma_2 + \underbrace{(bb^* - aa^*)}_{r_3} \sigma_3.$$

We now define the three-dimensional vectors, \vec{F} and $\vec{\Omega}$, consisting of the coefficients of the Pauli matrices in the expansion of ρ and H , respectively:

$$\vec{F} = (r_1, r_2, r_3) \quad (\text{A1.6.55})$$

$$\bar{\Omega} = \frac{1}{\hbar}(E_1, E_2, E_3). \quad (\text{A1.6.56})$$

Using these vectors, we can rewrite the Liouville equation for the two-level system as

$$\frac{d}{dt}\bar{r} = \bar{\Omega} \times \bar{r}. \quad (\text{A1.6.57})$$

Note that r_3 is the population difference between the upper and lower states: having all the population in the lower state corresponds to $r_3 = -1$ while having a completely inverted population (i.e. no population in the lower state) corresponds to $r_3 = +1$.

This representation is slightly inconvenient since E_1 and E_2 in equation (A1.6.56) are explicitly time-dependent. For a monochromatic light field of frequency ω , we can transform to a frame of reference rotating at the frequency of the light field so that the vector $\bar{\Omega}$ is a constant. To completely remove the time dependence we make the rotating wave approximation (RWA) as before: $E \cos(\omega t) = \frac{1}{2}(E e^{-i\omega t} + E e^{i\omega t}) \rightarrow \frac{1}{2}E e^{-i\omega t}$. In the rotating frame, the Liouville equation for the system is

$$\frac{d}{dt}\bar{r}' = \bar{\Omega}' \times \bar{r}' \quad (\text{A1.6.58})$$

where $\bar{\Omega}'$ is now time independent. The geometrical interpretation of this equation is that the pseudospin vector, r' , precesses around the field vector, Ω' , in exactly the same way that the angular momentum vector precesses around a body fixed axis of a rigid object in classical mechanics. This representation of the two-level system is called the Feynman–Vernon–Hellwarth, or FVH representation; it gives a unified, pictorial view with which one can understand the effect of a wide variety of optical pulse effects in two-level systems. For example, the geometrical picture of Rabi

-16-

cycling within the FVH picture is shown in figure A1.6.3. Assuming that at $t = 0$ all the population is in the ground-state then the initial position of the \bar{r}' vector is $(0,0,-1)$, and so \bar{r}' points along the negative z -axis. For a resonant field, $\omega_0 - \omega = 0$ and so the $\bar{\Omega}'$ vector points along the x -axis. Equation (A1.6.58) then says that the population vector simply precesses about the x -axis. It then periodically points along the positive z -axis, which corresponds to having all the population in the upper state. If the field is non-resonant, then $\bar{\Omega}'$ no longer points along the x -axis but along some other direction in the xz -plane. The population vector still precesses about the field vector, but now at some angle to the z -axis. Thus, the projection onto the z -axis of \bar{r}' never equals one and so there is never a complete population inversion.

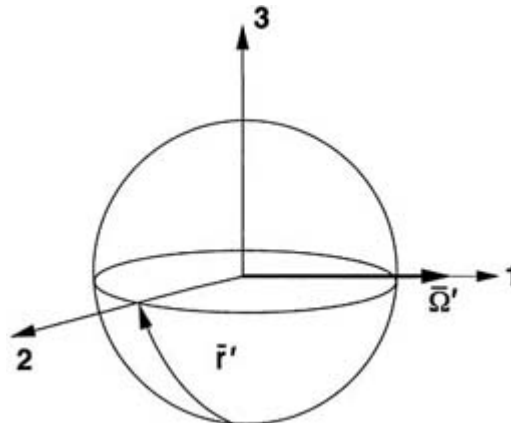


Figure A1.6.3. FVH diagram, exploiting the isomorphism between the two-level system and a pseudospin vector precessing on the unit sphere. The pseudospin vector, \vec{F}' , precesses around the field vector, $\vec{\Omega}'$, according to the equation $d\vec{F}'/dt = \vec{\Omega}' \times \vec{F}'$. The z-component of the \vec{F}' vector is the population difference between the two levels, while the x- and y-components refer to the polarization, that is, the real and imaginary parts of the coherence between the amplitude in the two levels. In the frame of reference rotating at the carrier frequency, the z-component of the $\vec{\Omega}'$ vector is the detuning of the field from resonance, while the x- and y-components indicate the field amplitude. In the rotating frame, the y-component of $\vec{\Omega}'$ may be set equal to zero (since the overall phase of the field is irrelevant, assuming no coherence of the levels at $t = 0$), unless there is non-uniform change in phase in the field during the process.

The FVH representation allows us to visualize the results of more complicated laser pulse sequences. A laser pulse which takes \vec{F}' from $(0,0,-1)$ to $(0,0,1)$ is called a π -pulse since the \vec{F}' vector precesses π radians about the field vector. Similarly, a pulse which takes \vec{F}' from $(0,0,-1)$ to $(+1,0,0)$ is called a $\pi/2$ -pulse. The state represented by the vector $(+1,0,0)$ is a coherent superposition of the upper and lower states of the system.

One interesting experiment is to apply a $\pi/2$ -pulse followed by a $\pi/2$ phase shift of the field. This phase shift will bring $\vec{\Omega}'$ parallel to \vec{F}' . Since now $\vec{\Omega}' \times \vec{F}' = 0$, the population is fixed in time in a coherent superposition between the ground and excited states. This is called photon locking.

A second interesting experiment is to begin with a pulse which is far below resonance and slowly and continuously sweep the frequency until the pulse is far above resonance. At $t = -\infty$ the field vector is pointing nearly along the $-z$ -axis, and is therefore almost parallel to the state vector. As the field vector slowly moves from $z = -1$ to $z = +1$

-17-

the state vector adiabatically follows it, precessing about the instantaneous direction of the field vector (figure A1.6.4). When, at $t \rightarrow +\infty$, the field vector is directed nearly along the $+z$ -axis, the state vector is directed there as well, signifying complete population inversion. The remarkable feature of ‘adiabatic following’, as this effect is known, is its robustness—there is almost no sensitivity to either the field strength or the exact schedule of changing the frequency, provided the conditions for adiabaticity are met.

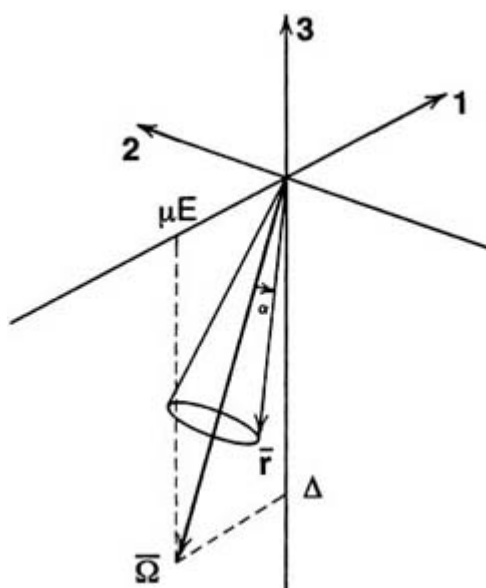


Figure A1.6.4. FVH diagram, showing the concept of adiabatic following. The Bloch vector, \vec{F}' , precesses in a narrow cone about the rotating frame torque vector, $\vec{\Omega}'$. As the detuning, Δ , changes from negative to positive, the field vector, $\vec{\Omega}'$, becomes inverted. If the change in $\vec{\Omega}'$ is adiabatic the Bloch vector follows the

field vector in this inversion process, corresponding to complete population transfer to the excited state.

A1.6.2.4 RELAXATION OF THE DENSITY OPERATOR TO EQUILIBRIUM

In real physical systems, the populations $|a(t)|^2$ and $|b(t)|^2$ are not truly constant in time, even in the absence of a field, because of relaxation processes. These relaxation processes lead, at sufficiently long times, to thermal equilibrium, characterized by the populations $P_a = e^{-\beta E_a}/Q$, $P_b = e^{-\beta E_b}/Q$, where Q is the canonical partition function which serves as a normalization factor and $\beta = 1/kT$, where k is the Boltzmann's constant and T is the temperature. The thermal equilibrium state for a two-level system, written as a density matrix, takes the following form:

$$\rho = \begin{pmatrix} e^{-\beta E_a}/Q & 0 \\ 0 & e^{-\beta E_b}/Q \end{pmatrix}. \quad (\text{A 1.6.59})$$

The populations, $e^{-\beta E_n}/Q$, appear on the diagonal as expected, but note that there are no off-diagonal elements—no coherences; this is reasonable since we expect the equilibrium state to be time-independent, and we have associated the coherences with time.

-18-

It follows that there are two kinds of processes required for an arbitrary initial state to relax to an equilibrium state: the diagonal elements must redistribute to a Boltzmann distribution and the off-diagonal elements must decay to zero. The first of these processes is called population decay; in two-level systems this time scale is called T_1 . The second of these processes is called dephasing, or coherence decay; in two-level systems there is a single time scale for this process called T_2 . There is a well-known relationship in two level systems, valid for weak system–bath coupling, that

$$\frac{1}{T_2} = \frac{1}{2T_1} + \frac{1}{T_2^*} \quad (\text{A 1.6.60})$$

where T_2^* is the time scale for so-called pure dephasing. Equation (A1.6.60) has the following significance: even without pure dephasing there is still a minimal dephasing rate that accompanies population relaxation.

In the presence of some form of relaxation the equations of motion must be supplemented by a term involving a relaxation *superoperator*—superoperator because it maps one operator into another operator. The literature on the correct form of such a superoperator is large, contradictory and incomplete. In brief, the extant theories can be divided into two kinds, those without memory relaxation (Markovian) $\Gamma\rho$ and those with memory relaxation (non-Markovian) $\int_{-\infty}^t \Gamma(t-t')\rho(t') dt'$. The Markovian theories can be further subdivided into those that preserve positivity of the density matrix (all $p_n > 0$ in [equation \(A1.6.66\)](#) for all admissible ρ) and those that do not. For example, the following widely used Markovian equation of motion is guaranteed to preserve positivity of the density operator for any choice of $\{V_i\}$:

$$\frac{\partial \rho}{\partial t} \equiv \left[\frac{H}{i\hbar}, \rho \right] + \Gamma\rho = \left[\frac{H}{i\hbar}, \rho \right] + \sum_i V_i \rho V_i^\dagger - \frac{1}{2} \sum_i [V_i^\dagger V_i \rho + \rho V_i^\dagger V_i]. \quad (\text{A 1.6.61})$$

As an example, consider the two-level system, with relaxation that arises from spontaneous emission. In this case there is just a single V_i :

$$V = \gamma^{1/2} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \quad V^\dagger = \gamma^{1/2} \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}. \quad (\text{A 1.6.62})$$

It is easy to verify that the dissipative contribution is given by

$$\Gamma\rho = \gamma \begin{pmatrix} \rho_{22} & -\rho_{12}/2 \\ -\rho_{21}/2 & -\rho_{22} \end{pmatrix}. \quad (\text{A 1.6.63})$$

We now make two connections with topics discussed earlier. First, at the beginning of this section we defined $1/T_1$ as the rate constant for population decay and $1/T_2$ as the rate constant for coherence decay. Equation (A1.6.63) shows that for spontaneous emission $1/T_1 = \gamma$, while $1/T_2 = \gamma/2$; comparing with equation (A1.6.60) we see that for spontaneous emission, $1/T_2^* = 0$. Second, note that γ is the rate constant for population transfer due to spontaneous emission; it is identical to the Einstein A coefficient which we defined in [equation \(A1.6.3\)](#).

-19-

For the two-level system, the evolution equation for ρ may also be expressed, as before, in terms of the three-vector \vec{r} :

$$\frac{d}{dt}\vec{r} = \vec{\Omega} \times \vec{r} - \vec{\Gamma} \cdot \vec{r} \quad (\text{A 1.6.64})$$

where

$$\vec{\Gamma} = \gamma \left(\frac{1}{2}, \frac{1}{2}, 1 \right) \equiv \left(\frac{1}{T_2}, \frac{1}{T_2}, \frac{1}{T_1} \right). \quad (\text{A 1.6.65})$$

Equation (A1.6.64) describes the relaxation to equilibrium of a two-level system in terms of a vector equation. It is the analogue of the Bloch equation, originally developed for magnetic resonance, in the optical regime and hence is called the optical Bloch equation.

In the above discussion of relaxation to equilibrium, the density matrix was implicitly cast in the energy representation. However, the density operator can be cast in a variety of representations other than the energy representation. Two of the most commonly used are the coordinate representation and the Wigner phase space representation. In addition, there is the diagonal representation of the density operator; in this representation, the most general form of ρ takes the form

$$\rho = \sum_i p_i |\Psi_i\rangle \langle \Psi_i| \quad (\text{A 1.6.66})$$

where the p_i are real numbers, $0 \leq p_i \leq 1$ and $\sum p_i = 1$. This equation expresses ρ as an *incoherent* superposition of fundamental density operators, $|\Psi_i\rangle \langle \Psi_i|$, where $|\Psi_i\rangle$ is a wavefunction but not necessarily an eigenstate. In equation (A1.6.66), the p_i are the *probabilities* (not amplitudes) of finding the system in state $|\Psi_i\rangle$. Note that in addition to the usual probabilistic interpretation for finding the particle described by a particular wavefunction at a specified location, there is now a probability distribution for being in different eigenstates! If one of the $p_i = 1$ and all the others are zero, the density operator takes the form [equation \(A1.6.49\)](#) and corresponds to a single wavefunction; we say the system is in a *pure state*. If more than one of the $p_i > 0$ we say the system is in a *mixed state*.

A measure of the purity or coherence of a system is given by $\sum_i p_i^2$: $\sum_i p_i^2 = 1$ for a pure state and $\sum_i p_i^2 \leq 1$ for a mixed state; the greater the degree of mixture the lower will be the purity. A general expression for the purity, which reduces to the above definition but is representation free, is given by $\text{Tr}(\rho^2)$: $\text{Tr}(\rho^2) < 1$ for a mixed state and $\text{Tr}(\rho^2) = 1$ for a pure state. Note that in the absence of dissipation, the purity of the system, as measured by $\text{Tr}(\rho^2)$, is conserved in time. To see this, take the equation of motion for ρ to be purely Hamiltonian, that is,

$$\dot{\rho} = -\frac{i}{\hbar}[H, \rho]. \quad (\text{A 1.6.67})$$

-20-

Then:

$$\frac{d}{dt}\text{Tr}(\rho^2) = 2\text{Tr}\rho\dot{\rho} = \frac{2}{i\hbar}\text{Tr}(\rho[H, \rho]) = \frac{2}{i\hbar}\text{Tr}(\rho(H\rho - \rho H)) = 0 \quad (\text{A 1.6.68})$$

where in the last step we have used the cyclic invariance of the trace. This invariance of the purity to Hamiltonian manipulations is essentially equivalent to the invariance of phase space density, or entropy, to Hamiltonian manipulations. Including the dissipative part to the equations of motion gives

$$\dot{\rho} = -\frac{i}{\hbar}[H, \rho] + \Gamma\rho \quad \text{and} \quad \frac{d}{dt}\text{Tr}(\rho^2) = 2\text{Tr}(\rho\Gamma\rho). \quad (\text{A 1.6.69})$$

In concluding this section, we note the complementarity of the light and matter, this time in terms of coherence properties (i.e. phase relations). The FVH geometrical picture shows explicitly how the *phase* of the field is inseparably intertwined with the phase change in the matter; in the next section, in the context of short pulses, we shall see how the *time* of interaction with the pulse is similarly intertwined with the time of the response of the molecule, although in general an integration over all such times must be performed. But both these forms of complementarity are on the level of the Hamiltonian portion of the evolution only. The complementarity of the *dissipation* will appear at the end of this section, in the context of laser cooling.

A1.6.3 THE FIELD TRANSFERS ITS COHERENCE TO THE MATTER

Much of the previous section dealt with two-level systems. Real molecules, however, are not two-level systems: for many purposes there are only two electronic states that participate, but each of these electronic states has many states corresponding to different quantum levels for vibration and rotation. A coherent femtosecond pulse has a bandwidth which may span many vibrational levels; when the pulse impinges on the molecule it excites a coherent superposition of all these vibrational states—a vibrational wavepacket. In this section we deal with excitation by one or two femtosecond optical pulses, as well as continuous wave excitation; in [section A1.6.4](#) we will use the concepts developed here to understand nonlinear molecular electronic spectroscopy.

The pioneering use of wavepackets for describing absorption, photodissociation and resonance Raman spectra is due to Heller [[12](#), [13](#), [14](#), [15](#) and [16](#)]. The application to pulsed excitation, coherent control and nonlinear spectroscopy was initiated by Tannor and Rice ([\[17\]](#) and references therein).

A1.6.3.1 FIRST-ORDER AMPLITUDE: WAVEPACKET INTERFEROMETRY

Consider a system governed by Hamiltonian $H \equiv H_0 + H_1$, where H_0 is the bare molecular Hamiltonian and

H_1 is the perturbation, taken to be the $-\mu E(t)$ as we have seen earlier. Adopting the Born–Oppenheimer (BO) approximation and specializing to two BO states, H_0 can be written as

-21-

$$H_0 = \begin{pmatrix} H_a & 0 \\ 0 & H_b \end{pmatrix} \quad (\text{A 1.6.70})$$

and H_1 as

$$H_1 = \begin{pmatrix} 0 & -\mu_{ab}E^*(t) \\ -\mu_{ba}E(t) & 0 \end{pmatrix}. \quad (\text{A 1.6.71})$$

The TDSE in matrix form reads:

$$i\hbar \frac{\partial}{\partial t} \begin{pmatrix} \psi_a(t) \\ \psi_b(t) \end{pmatrix} = \begin{pmatrix} H_a & -\mu_{ab}E^*(t) \\ -\mu_{ba}E(t) & H_b \end{pmatrix} \begin{pmatrix} \psi_a(t) \\ \psi_b(t) \end{pmatrix}. \quad (\text{A 1.6.72})$$

Note the structural similarity between equation (A1.6.72) and [equation \(A1.6.41\)](#), with E_a and E_b being replaced by H_a and H_b , the BO Hamiltonians governing the quantum mechanical evolution in electronic states a and b , respectively. These Hamiltonians consist of a nuclear kinetic energy part and a potential energy part which derives from nuclear–electron attraction and nuclear–nuclear repulsion, which differs in the two electronic states.

If H_1 is small compared with H_0 we may treat H_1 by perturbation theory. The first-order perturbation theory formula takes the form [[18](#), [19](#), [20](#) and [21](#)]:

$$\psi^{(1)}(x, t) = \frac{1}{i\hbar} \int_0^t e^{-(i/\hbar)H_b(t-t')} \{-\mu_{ba}E(t')\} e^{-(i/\hbar)H_a t'} \psi_a(x, 0) dt' \quad (\text{A 1.6.73})$$

where we have assumed that all the amplitude starts on the ground electronic state. This formula has a very appealing physical interpretation. At $t = 0$ the wavefunction is in, say, $v = 0$ of the ground electronic state. The wavefunction evolves from $t = 0$ until time t' under the ground electronic state, Hamiltonian, H_a . If we assume that the initial state is a vibrational eigenstate of H_a , ($H_a \psi_v = E_v \psi_v$), there is no spatial evolution, just the accumulation of an overall phase factor; that is the action of $e^{-(i/\hbar)H_a t'} \psi_a(x, 0)$ can be replaced by $e^{-(i/\hbar)E_v t'} \psi_v(x, 0) dt'$. For concreteness, in what follows we will take $v = 0$, which is the most common case of interest. At $t = t'$ the electric field, of amplitude $E(t')$, interacts with the transition dipole moment, promoting amplitude to the excited electronic state. This amplitude evolves under the influence of H_b from time t' until time t . The integral dt' indicates that one must take into account all instants in time t' at which the interaction with the field could have taken place. In general, if the field has some envelope of finite duration in time, the promotion to the excited state can take place at any instant under this envelope, and there will be interference from portions of the amplitude that are excited at one instant and portions that are excited at another. The various steps in the process may be visualized schematically with the use of Feynman diagrams. The Feynman diagram for the process just described is shown in [figure A1.6.5](#).

-22-

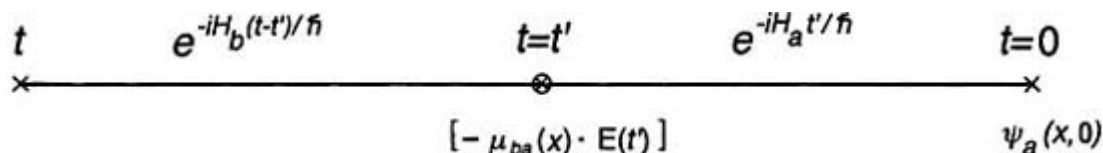


Figure A1.6.5 Feynman diagram for the first-order process described in the text.

We will now proceed to work through some applications of this formula to different pulse sequences. Perhaps the simplest is to consider the case of a δ -function excitation by light. That is,

$$E(t') = \delta(t' - t_1). \quad (\text{A 1.6.74})$$

In this case, the first-order amplitude reduces to

$$\psi^{(1)}(x, t) = \frac{1}{i\hbar} e^{-(i/\hbar)H_b(t-t_1)} \{-\mu_{ba}\} e^{-(i/\hbar)E_0 t_1} \psi_0(x, 0). \quad (\text{A 1.6.75})$$

Within the Condon approximation (μ_{ba} independent of x), the first-order amplitude is simply a constant times the initial vibrational state, propagated on the excited-state potential energy surface! This process can be visualized by drawing the ground-state vibrational wavefunction displaced vertically to the excited-state potential. The region of the excited-state potential which is accessed by the vertical transition is called the Franck–Condon region, and the vertical displacement is the Franck–Condon energy. Although the initial vibrational state was an eigenstate of H_a , in general it is not an eigenstate of H_b , and starts to evolve as a coherent wavepacket. For example, if the excited-state potential energy surface is repulsive, the wavepacket will evolve away from the Franck–Condon region toward the asymptotic region of the potential, corresponding to separated atomic or molecular fragments (see [figure A1.6.6](#)). If the excited-state potential is bound, the wavepacket will leave the Franck–Condon region, but after half a period reach a classical turning point and return to the Franck–Condon region for a complete or partial revival.

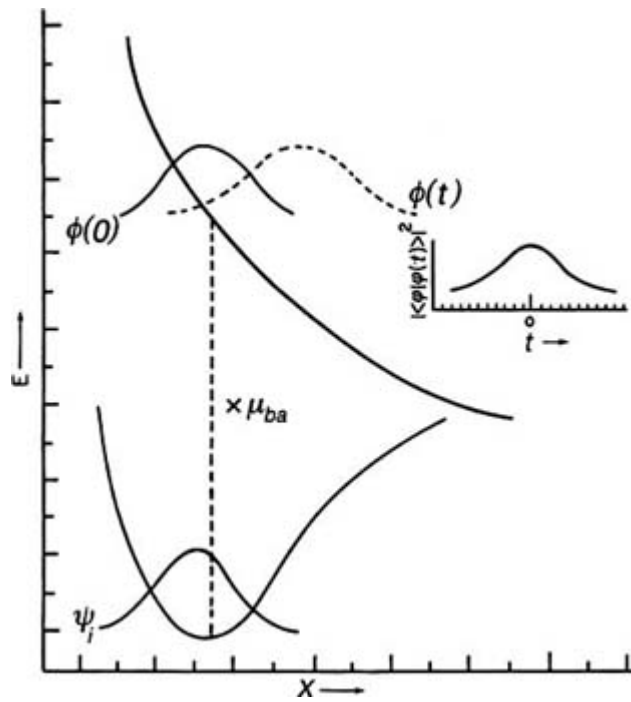


Figure A1.6.6 The wavepacket picture corresponding to the first-order process described in the text. The wavepacket propagates on the ground-state surface until time t_1 , but since it is an eigenstate of this surface it only develops a phase factor. At time t_1 a photon impinges and promotes the initial vibrational state to an excited electronic state, for which it is not an eigenstate. The state is now a *wavepacket* and begins to move according to the TDSE. Often the ensuing motion is very classical-like, the initial motion being along the gradient of the excited-state potential, with recurrences at multiples of the excited-state vibrational period (adapted from [32]).

An alternative perspective is as follows. A δ -function pulse in time has an infinitely broad frequency range. Thus, the pulse promotes transitions to all the excited-state vibrational eigenstates having good overlap (Franck–Condon factors) with the initial vibrational state. The pulse, by virtue of its coherence, in fact prepares a coherent superposition of all these excited-state vibrational eigenstates. From the earlier sections, we know that each of these eigenstates evolves with a different time-dependent phase factor, leading to coherent spatial translation of the wavepacket.

The δ -function excitation is not only the simplest case to consider; it is the fundamental building block, in the sense that the more complicated pulse sequences can be interpreted as superpositions of δ -functions, giving rise to *superpositions of wavepackets* which can in principle interfere.

The simplest case of this interference is the case of two δ -function pulses [22, 23 and 24]:

$$E(t') = \delta(t' - t_1)e^{-i\omega_L t_1} + \delta(t' - t_2)e^{-i\omega_L t_2} e^{i\phi}. \quad (\text{A 1.6.76})$$

We will explore the effect of three parameters: $t_2 - t_1$, ω_L and ϕ , that is, the time delay between the pulses, the tuning or detuning of the carrier frequency from resonance with an excited-state vibrational transition and the relative phase of the two pulses. We follow closely the development of [22]. Using equation (A1.6.73),

$$\begin{aligned}\psi^{(1)}(x, t) &= \frac{1}{i\hbar} e^{-(i/\hbar)H_b(t-t_1)} \{-\mu_{ba} e^{-i\omega_L t_1}\} e^{-(i/\hbar)E_0 t_1} \psi(x, 0) \\ &\quad + \frac{1}{i\hbar} e^{-(i/\hbar)H_b(t-t_2)} \{-\mu_{ba} e^{-i\omega_L t_2} e^{i\phi}\} e^{-(i/\hbar)E_0 t_2} \psi(x, 0)\end{aligned}\tag{A 1.6.77}$$

$$= \frac{1}{i\hbar} (e^{(i/\hbar)H_b(t_2-t_1)} e^{-i\omega_L(t_2-t_1)} e^{-i\omega_0(t_2-t_1)} e^{i\phi} + 1) e^{-(i/\hbar)H_b(t-t_1)} e^{-(i/\hbar)E_0 t_1} \{-\mu_{ba} e^{-i\omega_L t_1}\} \psi(x, 0).\tag{A 1.6.78}$$

To simplify the notation, we define $\tilde{H}_b = H_b - E_{00} - E_{0b}$, $\tilde{\omega}_L = \omega_L + \omega_0 - E_{00}/\hbar - E_{0b}/\hbar$, where E_{00} is the vertical displacement between the minimum of the ground electronic state and the minimum of the excited electronic state, E_{0b} is the zero point energy on the excited-state surface and $\omega_0 = E_0/\hbar$. Specializing to the harmonic oscillator, $e^{-(i/\hbar)\tilde{H}_b\tau} \psi(x, 0) = \psi(x, 0)$, where $\tau = 2\pi/\omega$ is the excited-state vibrational period, that is, any wavefunction in the harmonic oscillator returns exactly to its original spatial distribution after one period. To highlight the effect of detuning we write $\omega_L = n\omega + \Delta$, where Δ is the detuning from an excited-state vibrational eigenstate, and we examine time delays equal to the vibrational period $t_2 - t_1 = \tau$. We obtain:

$$\psi^{(1)}(x, t) = \frac{1}{i\hbar} (e^{(i/\hbar)\tilde{H}_b\tau} e^{-i(n\omega + \Delta)\tau} e^{i\phi} + 1) e^{-(i/\hbar)\tilde{H}_b(t-t_1)} \{-\mu_{ba} e^{-i\tilde{\omega}_L t_1}\} \psi(x, 0)\tag{A 1.6.79}$$

$$= \frac{1}{i\hbar} (e^{-i(\Delta\tau - \phi)} + 1) e^{-(i/\hbar)\tilde{H}_b(t-t_1)} \{-\mu_{ba} e^{-i\tilde{\omega}_L t_1}\} \psi(x, 0).\tag{A 1.6.80}$$

To illustrate the dependence on detuning, Δ , time delay, τ , and phase difference, ϕ , we consider some special cases. (i) If $\Delta = \phi = 0$ then the term in parentheses gives $1 + 1 = 2$. In this case, the two pulses create two wavepackets which add constructively, giving two units of amplitude or four units of excited-state population. (ii) If $\Delta = 0$ and $\phi = \pm\pi$ then the term in parentheses gives $-1 + 1 = 0$. In this case, the two pulses create two wavepackets which add destructively, giving no excited-state population! Viewed from the point of view of the light, this is stimulated emission. Emission against absorption is therefore controlled by the relative phase of the second pulse relative to the first. (iii) If $\Delta = 0$ and $\phi = \pm(\pi/2)$ then the term in parentheses gives $\pm i + 1$. In this case, the excited-state population, $\langle \psi^{(1)} | \psi^{(1)} \rangle$, is governed by the factor $(-i + 1)(i + 1) = 2$. The amplitude created by the two pulses overlap, but have no net interference contribution. This result is related to the phenomenon of ‘photon locking’, which was discussed in [section A1.6.2](#). (iv) If $\Delta = \omega/2$ and $\phi = 0$ then the term in parentheses gives $-1 + 1 = 0$. This is the ultrashort excitation counterpart of tuning the excitation frequency between vibrational resonances in a single frequency excitation: no net excited-state population is produced. As in the case above, of the two pulses π out of phase, the two wavepackets destructively interfere. In this case, the destructive interference comes from the offset of the carrier frequency from resonance, leading to a phase factor of $(\omega/2)\tau = \pi$. For time delays that are significantly different from τ the first wavepacket is not in the Franck–Condon region when the second packet is promoted to the excited state, and the packets do not interfere; two units of population are prepared on the excited state, as in the case of a $\pm(\pi/2)$ phase shift. These different cases are summarized in [figure A1.6.7](#).

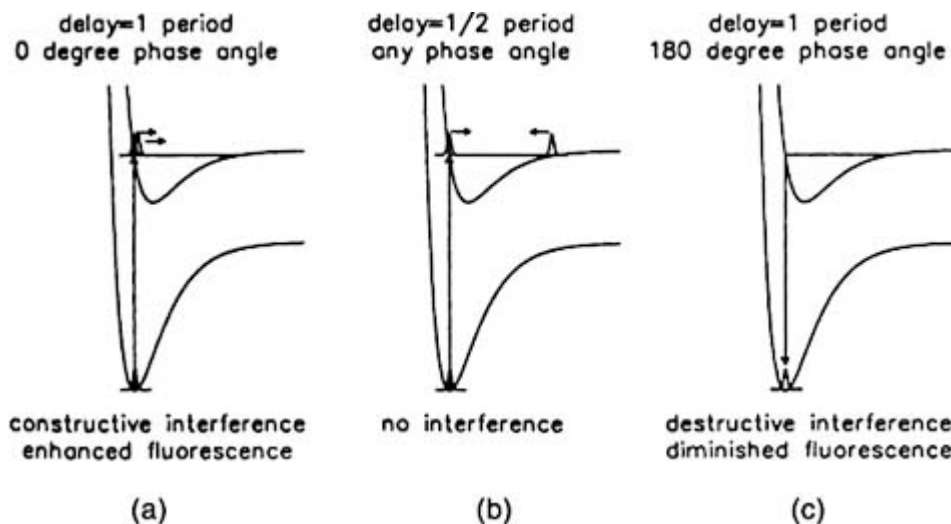


Figure A1.6.7. Schematic diagram illustrating the different possibilities of interference between a pair of wavepackets, as described in the text. The diagram illustrates the role of phase ((a) and (c)), as well as the role of time delay (b). These cases provide the interpretation for the experimental results shown in figure A1.6.8. Reprinted from [22].

Figure A1.6.8 shows the experimental results of Scherer *et al* of excitation of I_2 using pairs of phase locked pulses. By the use of heterodyne detection, those authors were able to measure just the interference contribution to the total excited-state fluorescence (i.e. the difference in excited-state population from the two units of population which would be prepared if there were no interference). The basic qualitative dependence on time delay and phase is the same as that predicted by the harmonic model: significant interference is observed only at multiples of the excited-state vibrational frequency, and the relative phase of the two pulses determines whether that interference is constructive or destructive.

There is a good analogy between the effects of pulse pairs and pulse shapes, and Fresnel and Fraunhofer diffraction in optics. Fresnel diffraction refers to the interference pattern obtained by light passing through two slits; interference from the wavefronts passing through the two slits is the spatial analogue of the interference from the two pulses in time discussed above. Fraunhofer diffraction refers to interference arising from the finite width of a single slit. The different subportions of a single slit can be thought of as independent slits that happen to adjoin; wavefronts passing through each of these subslits will interfere. This is the analogue of a single pulse with finite duration: there is interference from excitation coming from different subportions of the pulse, which may be insignificant if the pulse is short but can be important for longer pulse durations.

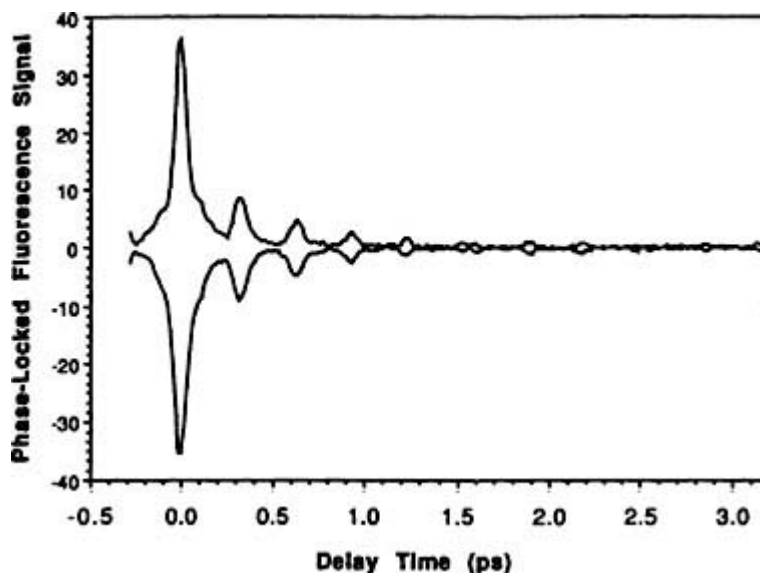


Figure A1.6.8. Wavepacket interferometry. The *interference contribution* to the excited-state fluorescence of I_2 as a function of the time delay between a pair of ultrashort pulses. The interference contribution is isolated by heterodyne detection. Note that the structure in the interferogram occurs only at multiples of 300 fs, the excited-state vibrational period of I_2 : it is only at these times that the wavepacket promoted by the first pulse is back in the Franck–Condon region. For a phase shift of 0 between the pulses the returning wavepacket and the newly promoted wavepacket are in phase, leading to constructive interference (upper trace), while for a phase shift of π the two wavepackets are out of phase, and interfere destructively (lower trace). Reprinted from Scherer N F *et al* 1991 *J. Chem. Phys.* **95** 1487.

There is an alternative, and equally instructive, way of viewing the effect of different pulse sequences, by Fourier transforming the pulse train to the frequency domain. In the time domain, the wavefunction produced is the convolution of the pulse sequence with the excited-state dynamics; in frequency it is simply the product of the frequency envelope with the Franck–Condon spectrum (the latter is simply the spectrum of overlap factors between the initial vibrational state and each of the excited vibrational states). The Fourier transform of δ -function excitation is simply a constant excitation in frequency, which excites the entire Franck–Condon spectrum. The Fourier transform of a sequence of two δ -functions in time with spacing τ is a spectrum having peaks with a spacing of $2\pi/\tau$. If the carrier frequency of the pulses is resonant and the relative phase between the pulses is zero, the frequency spectrum of the pulses will lie on top of the Franck–Condon spectrum and the product will be non-zero; if, on the other hand, the carrier frequency is between resonances, or the relative phase is π , the frequency spectrum of the pulses will lie in between the features of the Franck–Condon spectrum, signifying zero net absorption. Similarly, a single pulse of finite duration may have a frequency envelope which is smaller than that of the entire Franck–Condon spectrum. The absorption process will depend on the overlap of the frequency spectrum with the Franck–Condon spectrum, and hence on both pulse shape and carrier frequency.

A1.6.3.2 SECOND-ORDER AMPLITUDE: CLOCKING CHEMICAL REACTIONS

We now turn to the second-order amplitude. This quantity is given by [18, 19, 20 and 21]

$$\psi^{(2)}(x, t) = \left(\frac{1}{i\hbar}\right)^2 \int_0^t \int_0^{t'} e^{-(i/\hbar)H_c(t-t')} \{-\mu_{cb}E(t')\} e^{-(i/\hbar)H_b(t'-t'')} \{-\mu_{ba}E(t'')\} e^{-(i/\hbar)H_a t''} \psi(x, 0) dt' dt'' \quad (\text{A 1.6.81})$$

This expression may be interpreted in a very similar spirit to that given above for one-photon processes. Now there is a second interaction with the electric field and the subsequent evolution is taken to be on a third surface, with Hamiltonian H_c . In general, there is also a second-order interaction with the electric field through μ_{ab} which returns a portion of the excited-state amplitude to surface a , with subsequent evolution on surface a . The Feynman diagram for this second-order interaction is shown in figure A1.6.9.

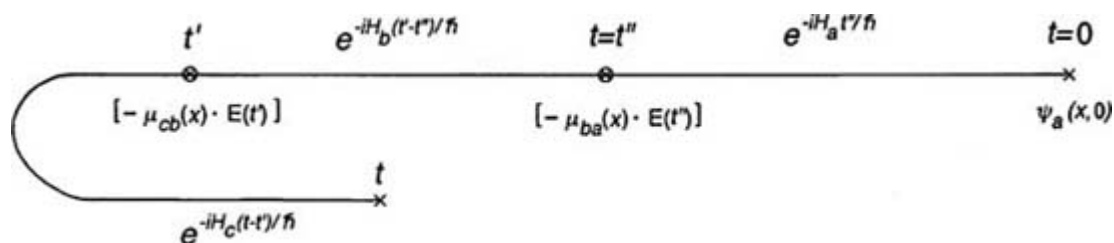


Figure A1.6.9. Feynman diagram for the second-order process described in the text.

Second-order effects include experiments designed to ‘clock’ chemical reactions, pioneered by Zewail and co-workers [25]. The experiments are shown schematically in figure A1.6.10. An initial 100–150 fs pulse moves population from the bound ground state to the dissociative first excited state in ICN. A second pulse, time delayed from the first then moves population from the first excited state to the second excited state, which is also dissociative. By noting the frequency of light absorbed from the second pulse, Zewail can estimate the distance between the two excited-state surfaces and thus infer the motion of the initially prepared wavepacket on the first excited state (figure A1.6.10).

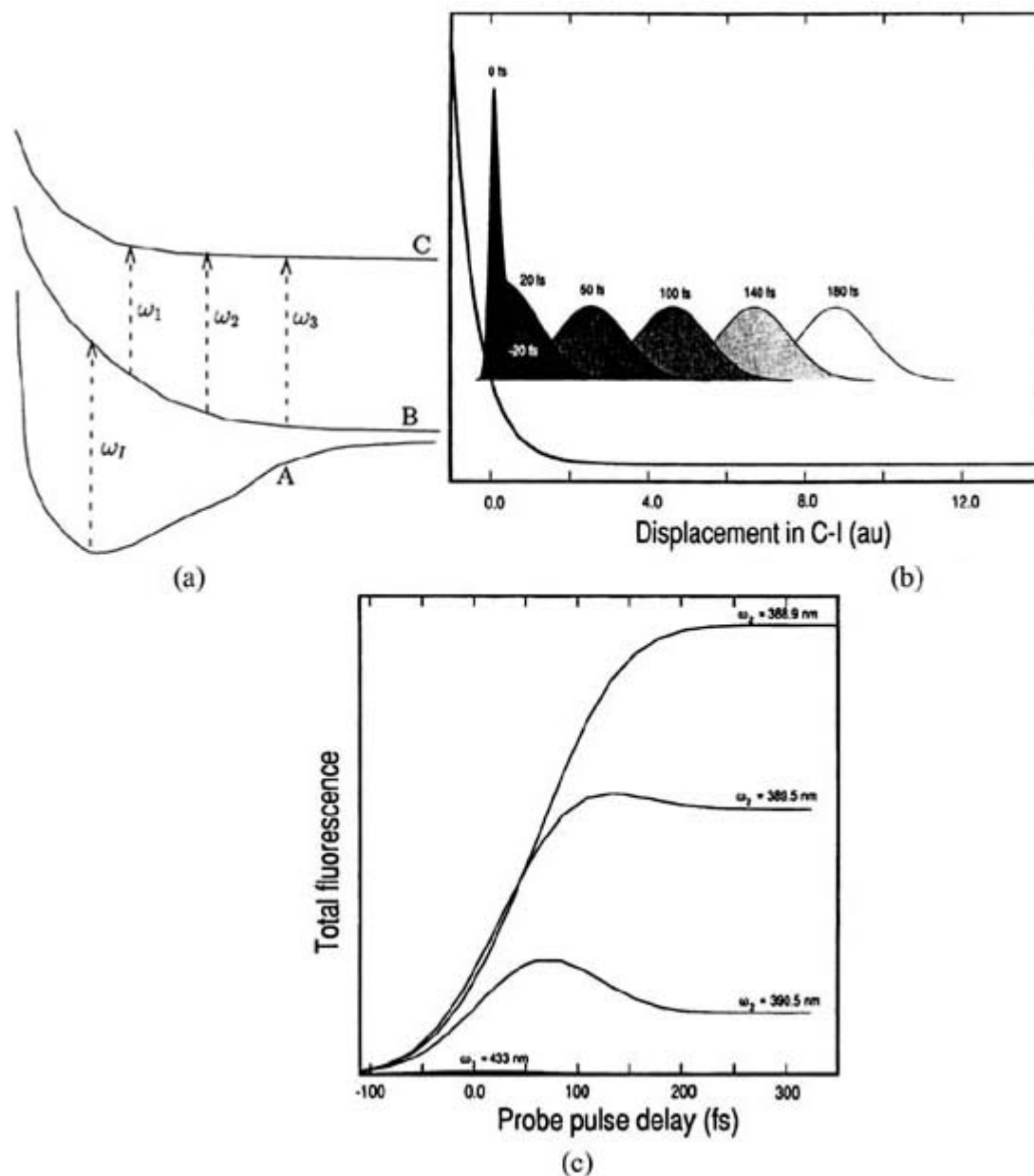


Figure A1.6.10. (a) Schematic representation of the three potential energy surfaces of ICN in the Zewail experiments. (b) Theoretical quantum mechanical simulations for the reaction $\text{ICN} \rightarrow \text{ICN}^* \rightarrow [\text{I} \cdots \text{CN}]^{\ddagger*} \rightarrow \text{I} + \text{CN}$. Wavepacket moves and spreads in time, with its centre evolving about 5 \AA in 200 fs. Wavepacket dynamics refers to motion on the intermediate potential energy surface B. Reprinted from Williams S O and Imre D G 1988 *J. Phys. Chem.* **92** 6648. (c) Calculated FTS signal (total fluorescence from state C) as a function of the time delay between the first excitation pulse ($\text{A} \rightarrow \text{B}$) and the second excitation pulse ($\text{B} \rightarrow \text{C}$). Reprinted from Williams S O and Imre D G, as above.

A dramatic set of experiments by Zewail involves the use of femtosecond pulse pairs to probe the wavepacket dynamics at the crossing between covalent and ionic states of NaI [25]. A first pulse promotes wavepacket amplitude from the ionic to the covalent potential curve. The packet begins to move out, but most of the amplitude is reflected back from the crossing between the covalent and ionic curves, that is, the adiabatic potential changes character to ionic at large distances, and this curve is bound, leading to wavepacket reflection back to the FC region. The result is a long progression of wavepacket revivals, with a slow overall decay coming from amplitude which dissociates on the diabatic curve every period.

Femtosecond pump–probe experiments have burgeoned in the last ten years, and this field is now commonly referred to as ‘laser femtochemistry’ [26, 27, 28 and 29].

A1.6.3.3 SPECTROSCOPY AS THE RATE OF ABSORPTION OF MONOCHROMATIC RADIATION

In this section we will discuss more conventional spectroscopies: absorption, emission and resonance Raman scattering. These spectroscopies are generally measured under single frequency conditions, and therefore our formulation will be tailored accordingly: we will insert monochromatic perturbations of the form $e^{i\omega\tau}$ into the perturbation theory formulae used earlier in the section. We will then *define* the spectrum as the time rate of change of the population in the final level. The same formulae apply with only minor modifications to electronic absorption, emission, photoionization and photodetachment/transition state spectroscopy. If the CW perturbation is inserted into the second-order perturbation theory one obtains the formulae for resonance Raman scattering, two-photon absorption and dispersed fluorescence spectroscopy. The spectroscopies of this section are to be contrasted with coherent nonlinear spectroscopies, such as coherent anti-Stokes Raman spectroscopy (CARS) or photon echoes, in which the signal is directional, which will be described in [section A1.6.4](#).

(A) ELECTRONIC ABSORPTION AND EMISSION SPECTROSCOPY

Consider the radiation–matter Hamiltonian, [equation \(A1.6.73\)](#), with the interaction term of the form:

$$H_1(t) = -\mu E(t) = \begin{cases} \frac{-\mu E_0}{2} e^{-i\omega_1 t} & \text{absorption} \\ \frac{-\mu E_0}{2} e^{+i\omega_1 t} & \text{emission} \end{cases} \quad (\text{A 1.6.82})$$

where the incident (scattered) light has frequency $\omega_1(\omega_S)$ and μ is the (possibly coordinate-dependent) transition dipole moment for going from the lower state to the upper state. This form for the matter–radiation interaction Hamiltonian represents a light field that is ‘on’ all the time from $-\infty$ to ∞ . This interaction will continuously move population from the lower state to the upper state. The propagating packets on the upper states will interfere with one another: constructively, if the incident light is resonant with a transition from an eigenstate of the lower surface to an eigenstate of the upper surface, destructively if not. Since for a one-photon process we have two potential energy surfaces we have, in effect, two different H_0 ’s: one for before excitation (call it H_a) and one for after (call it H_b). With this in mind, we can use the results of [section A1.6.2](#) to write down the first-order correction to the unperturbed wavefunction. If $|\psi_i(-\infty)\rangle$ is an eigenstate of the ground-state Hamiltonian, H_a , then

-30-

$$|\psi^{(1)}(t)\rangle = \frac{-1}{2i\hbar} \int_{-\infty}^t e^{-(i/\hbar)H_b(t-t')} \mu E_0 e^{-i\omega_1 t'} e^{-(i/\hbar)E_i t'} |\psi_i(-\infty)\rangle dt'. \quad (\text{A 1.6.83})$$

Defining $\tilde{\omega} = E_j/\hbar + \omega_j$, replacing $\psi(-\infty)$ by $\psi(0)$, since the difference is only a phase factor, which exactly cancels in the bra and ket, and assuming that the electric field vector is time independent, we find

$$\frac{d}{dt} \langle \psi^{(1)}(t) | \psi^{(1)}(t) \rangle = \frac{1}{4\hbar^2} \int_{-\infty}^{\infty} \langle \psi_i(0) | \mu E_0 e^{-(i/\hbar)H_b t} E_0 \mu | \psi_i(0) \rangle e^{i\tilde{\omega} t} dt. \quad (\text{A 1.6.84})$$

The absorption spectrum, $\sigma(\omega)$, is the ratio of transition probability per unit time/incident photon flux. The incident photon flux is the number of photons per unit area per unit time passing a particular location, and is

given by

$$\frac{Nc}{V} = \frac{E_0^2 c}{8\pi\hbar\omega}$$

where we have used [equation \(A1.6.15\)](#). Finally, we obtain [[12](#), [13](#)]:

$$\sigma(\omega) = \frac{2\pi\hbar\omega}{E_0^2 c} \frac{d}{dt} \langle \psi^{(1)}(t) | \psi^{(1)}(t) \rangle \quad (\text{A 1.6.85})$$

$$= \frac{2\pi\omega}{\hbar c} \int_{-\infty}^{\infty} \langle \psi_i(0) | \mu e^{-iHt/\hbar} \mu | \psi_i(0) \rangle e^{i\omega t} dt. \quad (\text{A 1.6.86})$$

Rotational averaging yields

$$\sigma(\omega) = \frac{2\pi\omega}{3\hbar c} \int_{-\infty}^{\infty} \langle \phi_i(0) | \phi_i(t) \rangle e^{i\omega t} dt \quad (\text{A 1.6.87})$$

where in the last equation we have defined $|\phi_i(0)\rangle \equiv \mu |\psi_i(0)\rangle$ and $|\phi_i(t)\rangle = e^{-iHt/\hbar} |\phi_i(0)\rangle$.

Since the absorption spectrum is a ratio it is amenable to other interpretations. One such interpretation is that the absorption spectrum is the ratio of energy absorbed to energy incident. From this perspective, the quantity $\hbar\omega(d/dt)\langle \psi^{(1)}(t) | \psi^{(1)}(t) \rangle$ is interpreted as the rate of energy absorption (per unit volume), since $dE/dt = \hbar\omega(dN/dt)$ while the quantity $E_0^2 c / \hbar\omega$ is interpreted as the incident energy flux, which depends only on the field intensity and is independent of frequency.

[Equation A1.6.87](#) expresses the absorption spectrum as the Fourier transform of a wavepacket correlation function. This is a result of central importance. The Fourier transform relationship between the wavepacket autocorrelation function and the absorption spectrum provides a powerful tool for interpreting absorption spectra in terms of the underlying nuclear wavepacket dynamics that follows the optically induced transition. The relevant correlation function is that of the moving wavepacket on the excited-state potential energy surface (or more generally, on the potential energy surface accessed by interaction with the light) with the stationary wavepacket on the ground-state surface (more generally, the initial wavepacket on the potential surface of origin), and thus the spectrum is a probe of excited-state dynamics, particularly in the Franck–Condon region (i.e. the region accessed by the packet undergoing a vertical transition at $t = 0$). Since often only short or intermediate dynamics enter in the spectrum (e.g. because of photodissociation or radiationless transitions to other electronic states) computation of the time correlation function can be much simpler than evaluation of the spectrum in terms of Franck–Condon overlaps, which formally can involve millions of eigenstates for an intermediate sized molecule.

We now proceed to some examples of this Fourier transform view of optical spectroscopy. Consider, for example, the UV absorption spectrum of CO_2 , shown in [figure A1.6.11](#). The spectrum is seen to have a long progression of vibrational features, each with fairly uniform shape and width. What is the physical interpretation of this vibrational progression and what is the origin of the width of the features? The goal is to come up with a dynamical model that leads to a wavepacket autocorrelation function whose Fourier transform

agrees with the spectrum in [figure A1.6.11](#). [figure A1.6.12](#) gives a plausible dynamical model leading to such an autocorrelation function. In (a), equipotential contours of the excited-state potential energy surface of CO_2 are shown, as a function of the two bond lengths, R_1 and R_2 , or, equivalently, as a function of the symmetric and antisymmetric stretch coordinates, v and u (the latter are linear combinations of the former). Along the axis $u = 0$ the potential has a minimum; along the axis $v = 0$ (the local ‘reaction path’) the potential has a maximum. Thus, the potential in the region $u = 0, v = 0$ has a ‘saddle-point’. There are two symmetrically related exit channels, for large values of R_1 and R_2 , respectively, corresponding to the formation of $\text{OC} + \text{O}$ versus $\text{O} + \text{CO}$. [figure A1.6.12](#) (a) also shows the initial wavepacket, which is approximately a two-dimensional Gaussian. Its centre is displaced from the minimum in the symmetric stretch coordinate. [figure A1.6.12\(b\)–\(f\)](#) show the subsequent dynamics of the wavepacket. It moves downhill along the v coordinate, while at the same time spreading. After one vibrational period in the v coordinate the centre of the wavepacket comes back to its starting point in v , but has spread in u ([figure A1.6.12\(e\)](#)). The resulting wavepacket autocorrelation function is shown in [figure A1.6.12\(right\) \(a\)](#). At $t = 0$ the autocorrelation function is 1. On a time scale τ_b the correlation function has decayed to nearly 0, reflecting the fact that the wavepacket has moved away from its initial Franck–Condon location ([figure A1.6.12\(b\)](#)). At time τ_e the wavepacket has come back to the Franck–Condon region in the v coordinate, and the autocorrelation function has a recurrence. However, the magnitude of the recurrence is much smaller than the initial value, since there is irreversible spreading of the wavepacket in the u coordinate. Note there are further, smaller recurrences at multiples of τ_e .

-32-

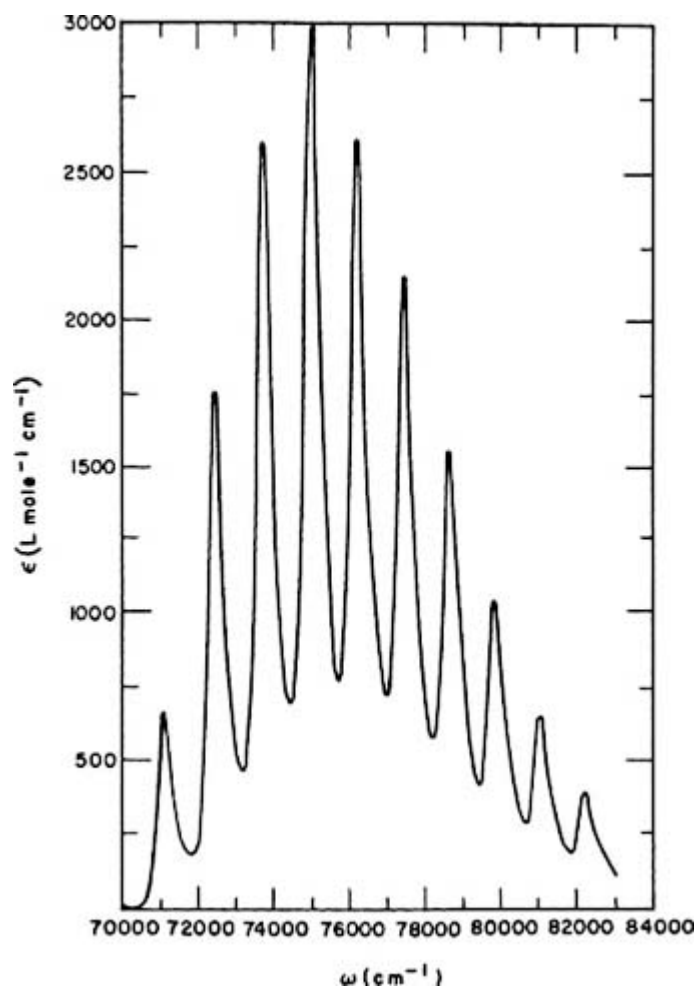


Figure A1.6.11. Idealized UV absorption spectrum of CO_2 . Note the regular progression of intermediate resolution vibrational progression. In the frequency regime this structure is interpreted as a Franck–Condon

progression in the symmetric stretch, with broadening of each of the lines due to predissociation. Reprinted from [31].

The spectrum obtained by Fourier transform of figure A1.6.12 (right) (a) is shown in figure A1.6.12 (right) (b). Qualitatively, it has all the features of the spectrum in figure A1.6.11 : a broad envelope with resolved vibrational structure underneath, but with an ultimate, unresolvable linewidth. Note that the shortest time decay, δ , determines the overall envelope in frequency, $1/\delta$; the recurrence time, T , determines the vibrational frequency spacing, $2\pi/T$; the overall decay time determines the width of the vibrational features. Moreover, note that decays in time correspond to widths in frequency, while recurrences in time correspond to spacings in frequency.

-33-

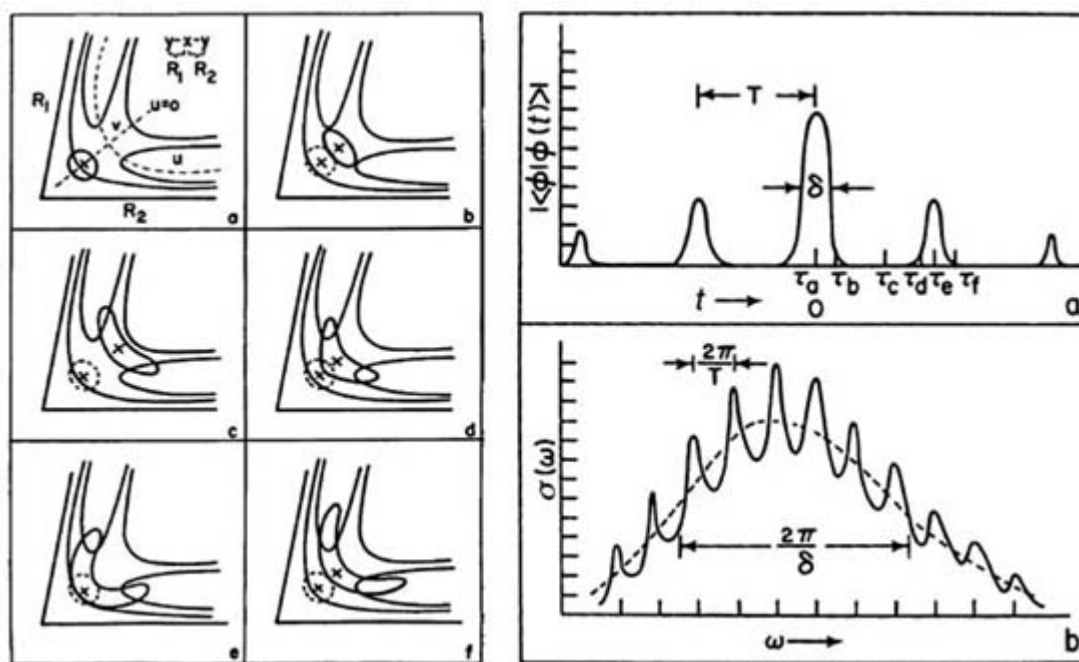


Figure A1.6.12. Left: A qualitative diagram showing evolution of $\phi(t)$ on the upper potential surface. Note the oscillation along the v (symmetric stretch) coordinate, and the spreading along the u (antisymmetric stretch) coordinate. Reprinted from [32]. Right: (a) The absolute value of the correlation function, $|\langle \phi | \phi(t) \rangle|$ versus t for the dynamical situation shown in figure A1.6.12. (b) The Fourier transform of $\langle \phi | \phi(t) \rangle$, giving the absorption spectrum. Note that the central lobe in the correlation function, with decay constant δ , gives rise to the overall width of the absorption spectrum, on the order of $2\pi/\delta$. Furthermore, the recurrences in the correlation on the time scale T give rise to the oscillations in the spectrum on the time scale $2\pi/T$. Reprinted from [32].

Perhaps the more conventional approach to electronic absorption spectroscopy is cast in the energy, rather than in the time domain. It is straightforward to show that equation (A1.6.87) can be rewritten as

$$\sigma(\omega) = \frac{4\pi^2\omega}{3c\hbar} \sum_n |\langle \psi_n | \mu | \psi_i \rangle|^2 \delta(\tilde{\omega} - \omega_n). \quad (\text{A 1.6.88})$$

Note that if we identify the sum over δ -functions with the density of states, then equation (A1.6.88) is just Fermi's Golden Rule, which we employed in section A1.6.1. This is consistent with the interpretation of the absorption spectrum as the transition rate from state i to state n .

The coefficients of the δ -function in the sum are called Franck–Condon factors, and reflect the overlap of the initial state with the excited-state ψ_n at energy $E_n = \hbar\omega_n$ (see figure A1.6.13). Formally, [equation \(A1.6.88\)](#) gives a ‘stick’ spectrum of the type shown in figure A1.6.13(b); generally, however, the experimental absorption spectrum is diffuse, as in [figure A1.6.11](#). This highlights one of the advantages of the time domain approach: that the broadening of the stick spectrum need not be introduced artificially, but arises naturally from the decay of the wavepacket correlation function, as we have seen in [figure A1.6.11](#).

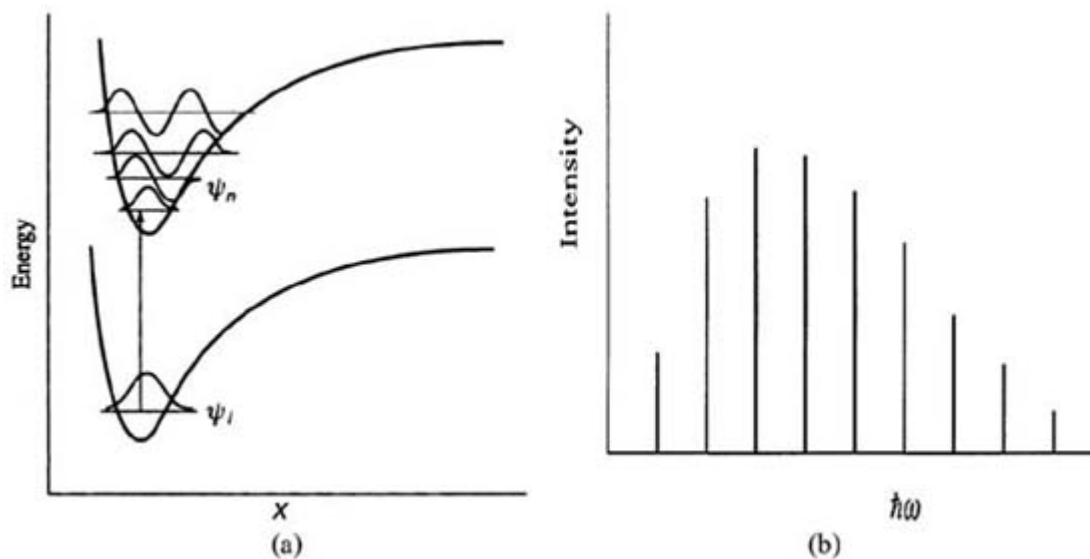


Figure A1.6.13. (a) Potential energy curves for two electronic states. The vibrational wavefunctions of the excited electronic state and for the lowest level of the ground electronic state are shown superimposed. (b) Stick spectrum representing the Franck–Condon factors (the square of overlap integral) between the vibrational wavefunction of the ground electronic state and the vibrational wavefunctions of the excited electronic state (adapted from [3]).

The above formulae for the absorption spectrum can be applied, with minor modifications, to other one-photon spectroscopies, for example, emission spectroscopy, photoionization spectroscopy and photodetachment spectroscopy (photoionization of a negative ion). For stimulated emission spectroscopy, the factor of ω_1 is simply replaced by ω_S , the stimulated light frequency; however, for spontaneous emission spectroscopy, the prefactor ω_1 is replaced by the prefactor ω_S^3 . The extra factor of ω_S^2 is due to the density of states of vacuum field states which induce the spontaneous emission, which increase quadratically with frequency. Note that in emission spectroscopy the roles of the ground- and excited-state potential energy surfaces are reversed: the initial wavepacket starts from the vibrational ground state of the *excited* electronic state and its spectrum has information on the vibrational eigenstates and potential energy surface of the *ground* electronic state.

(B) RESONANCE RAMAN SPECTROSCOPY

We will now look at two-photon processes. We will concentrate on Raman scattering although two-photon absorption can be handled using the same approach. In Raman scattering, absorption of an incident photon of frequency ω_1 carries

the initial wavefunction, ψ_i , from the lower potential to the upper. The emission of a photon of frequency ω_S returns the system to the lower potential, to state ψ_f . If $\omega_S = \omega_I$ then the scattering is elastic and the process is called Rayleigh scattering. Raman scattering occurs when $\omega_S \neq \omega_I$ and in that case $\psi_f \neq \psi_i$. The measured quantity is the Raman intensity, $I(\omega_I; \omega_S)$. The amplitudes of the incident and emitted fields are taken as E_I and E_S ; for simplicity, we begin with the case of stimulated Raman scattering, and then discuss the modifications for spontaneous Raman scattering at the end.

We start from the expression for the second-order wavefunction:

$$|\psi^{(2)}(t)\rangle = -\frac{1}{\hbar^2} \int_{-\infty}^t dt' \int_{-\infty}^{t'} dt'' e^{-(i/\hbar)H_a(t-t')} E_S \mu e^{+i\omega_S t'} \} e^{-(i/\hbar)H_b(t'-t'')} e^{-\frac{\gamma}{2}(t'-t'')} E_I \mu e^{-i\tilde{\omega}_I t''} \} \times |\psi_i(-\infty)\rangle + NRT \quad (A 1.6.89)$$

where H_a (H_b) is the Hamiltonian for the lower (upper) potential energy surface and, as before, $\tilde{\omega}_I = \omega_I + \omega_i$. In words, equation (A1.6.89) is saying that the second-order wavefunction is obtained by propagating the initial wavefunction on the ground-state surface until time t'' , at which time it is excited up to the excited state, upon which it evolves until it is returned to the ground state at time t' , where it propagates until time t . *NRT* stands for non-resonant term: it is obtained by $E_I \leftrightarrow E_S$ and $\omega_I \leftrightarrow -\omega_S$, and its physical interpretation is the physically counterintuitive possibility that the emitted photon precedes the incident photon. γ is the spontaneous emission rate.

If we define $\tilde{\omega}_S = \omega_S - \tilde{\omega}_I$, then we can follow the same approach as in the one-photon case. We now take the time derivative of the norm of $|\psi^{(2)}(t)\rangle$, with the result:

$$\omega_I \omega_S \frac{d}{dt} \langle \psi^{(2)}(t) | \psi^{(2)}(t) \rangle / |E|^2 = \omega_I \omega_S \frac{1}{\hbar^4} \sum_j |\alpha_{fi}(\omega_1)|^2 \delta(\omega_f + \omega_S - (\omega_I + \omega_i)). \quad (A 1.6.90)$$

where

$$\alpha_{fi}(\omega_1) = \int_0^\infty \langle \psi_f | \mu e^{-(i/\hbar)H_b t} \mu | \psi_i \rangle e^{-\frac{\gamma}{2}t} e^{i\tilde{\omega}_I t} dt + NRT. \quad (A 1.6.91)$$

Again, *NRT* is obtained from the first term by the replacement $\omega_I \rightarrow -\omega_S$. If we define $|\phi_f\rangle = \mu |\psi_f\rangle$ and $|\phi_i(t)\rangle = e^{-(i/\hbar)H_b t} \mu |\psi_i\rangle$, then we see that the frequency-dependent polarizability, $\alpha_{fi}(\omega_1)$, can be written in the following compact form [14]:

$$\alpha_{fi}(\omega_1) = \int_0^\infty \langle \phi_f | \phi_i(t) \rangle e^{-\frac{\gamma}{2}t} e^{i\tilde{\omega}_I t} dt + NRT. \quad (A 1.6.92)$$

The only modification of equation (A1.6.90) for spontaneous Raman scattering is the multiplication by the density of states of the cavity, equation (A1.6.24), leading to a prefactor of the form $\omega_I \omega_S^3$.

Equation (A1.6.92) has a simple physical interpretation. At $t = 0$ the initial state, ψ_i is multiplied by μ (which may be thought of as approximately constant in many cases, the Condon approximation). This product, denoted ϕ_i , constitutes an initial wavepacket which begins to propagate on the excited-state potential energy surface (figure A1.6.14). Initially, the wavepacket will have overlap only with ψ_i , and will be orthogonal to all other ψ_f on the ground-state surface. As the wavepacket begins to move on the excited state, however, it will develop overlap with ground vibrational states of ever-increasing quantum number. Eventually, the wavepacket will reach a turning point and begin moving back towards the Franck–Condon region of the excited-state surface, now overlapping ground vibrational states in decreasing order of their quantum number. These time-dependent overlaps determine the Raman intensities via equation (A1.6.92) and equation (A1.6.90). If the excited state is dissociative, then the wavepacket never returns to the Franck–Condon region and the Raman spectrum has a monotonically decaying envelope. If the wavepacket bifurcates on the excited state due to a bistable potential, then it will only have non-zero overlaps with ground vibrational states which are of even parity; the Raman spectrum will then have ‘missing’ lines. In multidimensional systems, there are ground vibrational states corresponding to each mode of vibration. The Raman intensities then contain information about the extent to which different coordinates participate in the wavepacket motion, to what extent, and even in what sequence [15, 33]. Clearly, resonance Raman intensities can be a sensitive probe of wavepacket dynamics on the excited-state potential.

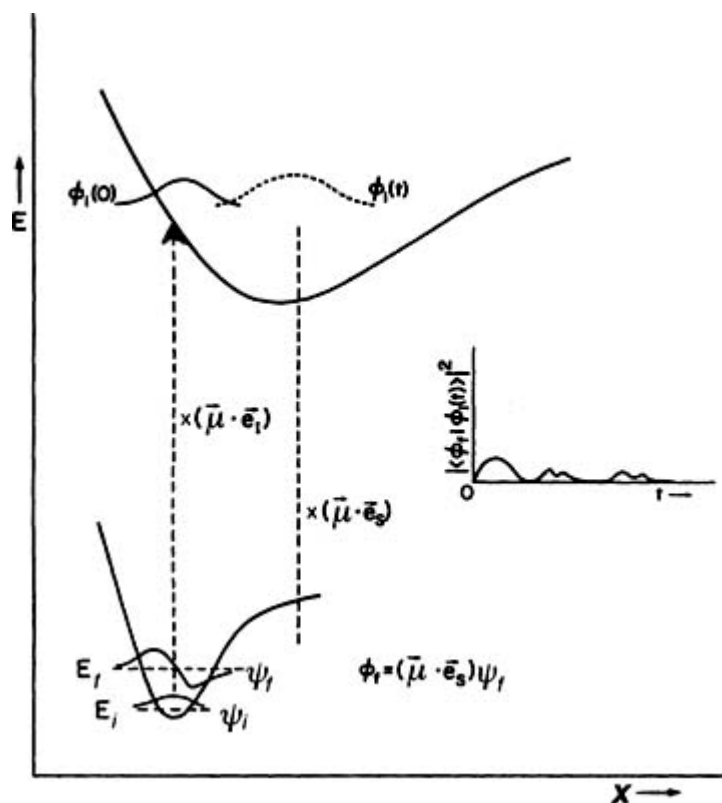


Figure A1.6.14. Schematic diagram showing the promotion of the initial wavepacket to the excited electronic state, followed by free evolution. Cross-correlation functions with the excited vibrational states of the ground-state surface (shown in the inset) determine the resonance Raman amplitude to those final states (adapted from [14]).

One of the most interesting features of the Raman spectrum is its dependence on the incident light frequency, ω_1 . When ω_1 is on resonance with the excited electronic state, the scattering process closely resembles a process of absorption followed by emission. However, as ω_1 is detuned from resonance there are no longer

any nearby eigenstates, and thus no absorption: the transition from the initial state i to the final state f is a ‘scattering’ process. In the older literature the non-existent intermediate state was called a ‘virtual’ state.

There can be no completely rigorous separation between absorption–emission and Raman scattering. This is clear from the time-domain expression, [equation \(A1.6.92\)](#), in which the physical meaning of the variable t is the time *interval* between incident and emitted photons. If the second photon is emitted long after the first photon was incident the process is called absorption/emission. If the second photon is emitted almost immediately after the first photon is incident the process is called scattering. The limits on the integral in [\(A1.6.92\)](#) imply that the Raman amplitude has contributions from all values of this interval ranging from 0 (scattering) to ∞ (absorption/emission). However, the regions that contribute most depend on the incident light frequency. In particular, as the incident frequency is detuned from resonance there can be no absorption and the transition becomes dominated by scattering. This implies that as the detuning is increased, the relative contribution to the integral from small values of t is greater.

Mathematically, the above observation suggests a time–energy uncertainty principle [15]. If the incident frequency is detuned by an amount $\Delta\omega$ from resonance with the excited electronic state, the wavepacket can ‘live’ on the excited state only for a time $\tau \approx 1/\Delta\omega$ (see [figure A1.6.15](#)). This follows from inspection of the integral in [equation \(A1.6.92\)](#): if the incident light frequency is mismatched from the intrinsic frequencies of the evolution operator, there will be a rapidly oscillating phase to the integrand. Normally, such a rapidly oscillating phase would kill the integral completely, but there is a special effect that comes into play here, since the lower bound of the integral is 0 and not $-\infty$. The absence of contributions from negative t leads to an incomplete cancellation of the portions of the integral around $t = 0$. The size of the region around $t = 0$ is inversely proportional to the mismatch in frequencies, $\Delta\omega$. Since the physical significance of t is time delay between incident and scattered photons, and this time delay is the effective wavepacket lifetime in the excited state, we are led to conclude that the effective lifetime decreases as the incident frequency is detuned from resonance.

Because of the two frequencies, ω_I and ω_S , that enter into the Raman spectrum, Raman spectroscopy may be thought of as a ‘two-dimensional’ form of spectroscopy. Normally, one fixes ω_I and looks at the intensity as a function of ω_S ; however, one may vary ω_I and probe the intensity as a function of $\omega_I - \omega_S$. This is called a Raman excitation profile.

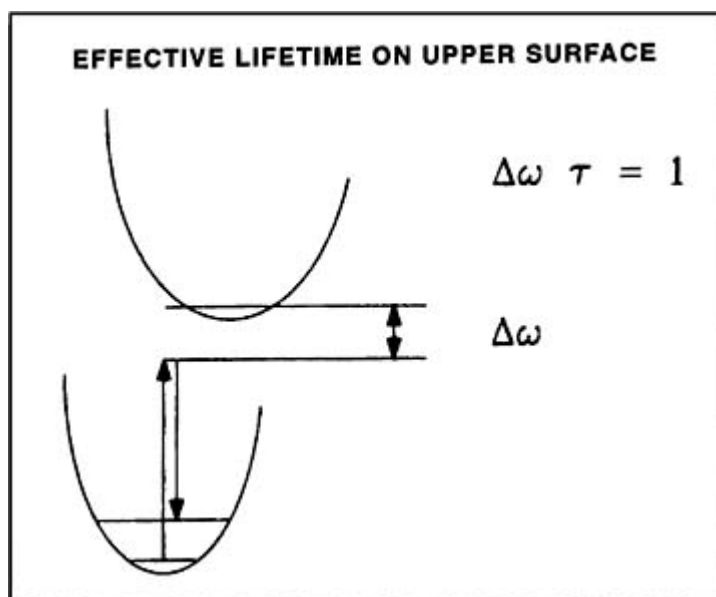


Figure A1.6.15. Schematic diagram, showing the time–energy uncertainty principle operative in resonance Raman scattering. If the incident light is detuned from resonance by an amount $\Delta\omega$, the effective lifetime on the excited-state is $\tau \approx 1/\Delta\omega$ (adapted from [15]).

The more conventional, energy domain formula for resonance Raman scattering is the expression by Kramers–Heisenberg–Dirac (KHD). The differential cross section for Raman scattering into a solid angle $d\Omega$ can be written in the form

$$\frac{d\sigma_{fi}(\omega_l)}{d\Omega} = \frac{\omega_l \omega_S^3}{c^4} \langle |(\alpha_{fi})_{\rho\lambda}(\omega_l)|^2 \rangle \quad (\text{A 1.6.93})$$

where

$$(\alpha_{fi})_{\rho\lambda}(\omega_l) = \sum_j \frac{\langle \psi_f | \mathbf{e}_S \cdot \boldsymbol{\mu}_o | \psi_j \rangle \langle \psi_j | \mathbf{e}_I \cdot \boldsymbol{\mu}_\lambda | \psi_i \rangle}{E_i + \hbar\omega_l - E_j - i\gamma/2} + \frac{\langle \psi_f | \mathbf{e}_S \cdot \boldsymbol{\mu}_o | \psi_j \rangle \langle \psi_j | \mathbf{e}_I \cdot \boldsymbol{\mu}_\lambda | \psi_i \rangle}{E_i - \hbar\omega_S - E_j - i\gamma/2} \quad (\text{A 1.6.94})$$

and the angular brackets indicate orientational averaging. The labels $\mathbf{e}_{I,S}$ refer to the direction of polarization of the incident and scattered light, respectively, while the subscripts ρ and λ refer to x , y and z components of the vector $\vec{\mu}$. Integrated over all directions and polarizations one obtains [33, 34]:

$$\sigma_{fi}(\omega_l) = \frac{8\pi \omega_l \omega_S^3}{9c^4} \sum_{\rho\lambda} |\alpha_{\rho\lambda}|^2. \quad (\text{A 1.6.95})$$

Equation (A1.6.94) is called the KHD expression for the polarizability, α . Inspection of the denominators indicates that the first term is the resonant term and the second term is the non-resonant term. Note the product of Franck–Condon factors in the numerator: one corresponding to the amplitude for excitation and the other to the amplitude for emission. The KHD formula is sometimes called the ‘sum-over-states’ formula, since formally it requires a sum over all intermediate states j , each intermediate state participating according to how far it is from resonance and the size of the matrix elements that connect it to the states ψ_i and ψ_f . The KHD formula is fully equivalent to the time domain formula, equation (A1.6.92), and can be derived from the latter in a straightforward way. However, the time domain formula can be much more convenient, particularly as one detunes from resonance, since one can exploit the fact that the effective dynamic becomes shorter and shorter as the detuning is increased.

A1.6.4 COHERENT NONLINEAR SPECTROSCOPY

As described at the end of section A1.6.1, in nonlinear spectroscopy a polarization is created in the material which depends in a nonlinear way on the strength of the electric field. As we shall now see, the microscopic description of this nonlinear polarization involves multiple interactions of the material with the electric field. The multiple interactions in principle contain information on both the ground electronic state and excited electronic state dynamics, and for a molecule in the presence of solvent, information on the molecule–solvent interactions. Excellent general introductions to nonlinear spectroscopy may be found in [35, 36 and 37]. Raman spectroscopy, described at the end of the previous section, is also a nonlinear spectroscopy, in the sense that it involves more than one interaction of light with the material, but it is a pathological example since the second interaction is through spontaneous emission and therefore not proportional to a driving field

and not directional; at the end of this section we will connect the present formulation with Raman spectroscopy [38].

What information is contained in nonlinear spectroscopy? For gas-phase experiments, that is, experiments in which the state of the system undergoes little or no dissipation, the goal of nonlinear spectroscopy is generally as in linear spectroscopy, that is, revealing the quantum energy level structure of the molecule, both in the ground and the excited electronic state(s). For example, two-photon spectroscopy allows transitions that are forbidden due to symmetry with one photon; thus the two-photon spectrum allows the spectroscopic study of many systems that are otherwise dark. Moreover, nonlinear spectroscopy allows one to access highly excited vibrational levels that cannot be accessed by ordinary spectroscopies, as in the example of time-dependent CARS spectroscopy below. Moreover, nonlinear spectroscopy has emerged as a powerful probe of molecules in anisotropic environments, for example, molecules at interfaces, where there is a $P^{(2)}$ signal which is absent for molecules in an isotropic environment.

A feature of nonlinear spectroscopy which is perhaps unique is the ability to probe not only energy levels and their populations, but to probe directly coherences, be they electronic or vibrational, via specially designed pulse sequences. For an isolated molecule this is generally uninteresting, since in the absence of relaxation the coherences are completely determined by the populations; however, for a molecule in solution the decay of the coherence is an indicator of molecule–solvent interactions. One normally distinguishes two sources of decay of the coherence: inhomogeneous decay, which represents static differences in the environment of different molecules; and homogeneous decay, which represents the dynamics interaction with the surroundings and is the same for all molecules. Both these sources of decay contribute to the *linewidth* of spectral lines; in many cases the inhomogeneous decay is faster than the homogeneous decay, masking the latter. In echo spectroscopies, which are related to a particular subset of diagrams in $P^{(3)}$, one can at least partially discriminate between homogeneous and inhomogeneous decay.

-40-

From the experimental point of view, nonlinear spectroscopy has the attractive feature of giving a directional signal (in a direction other than that of any of the incident beams), and hence a background free signal (figure A1.6.16). A significant amount of attention is given in the literature on nonlinear spectroscopy to the directionality of the signals that are emitted in different directions, and their dynamical interpretation. As we shall see, many dynamical pathways can contribute to the signal in each direction, and the dynamical interpretation of the signal depends on sorting out these contributions or designing an experiment which selects for just a single dynamical pathway.

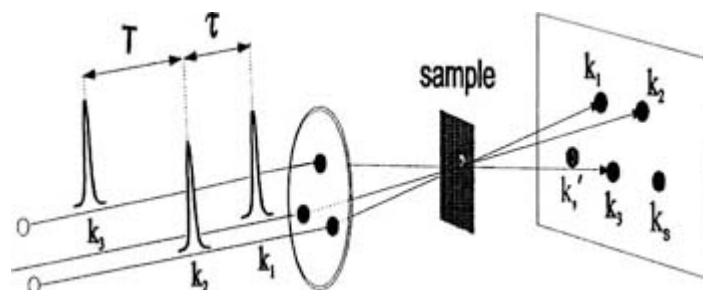


Figure A1.6.16. Diagram showing the directionality of the signal in coherent spectroscopy. Associated with the carrier frequency of each interaction with the light is a wavevector, \mathbf{k} . The output signal in coherent spectroscopies is determined from the direction of each of the input signals via momentum conservation (after [48a]).

A1.6.4.1 GENERAL DEVELOPMENT

As discussed in [section A1.6.1](#), on a microscopic quantum mechanical level, within the dipole approximation, the polarization, $P(t)$, is given by

$$P(t) \equiv \langle \psi | \mu | \psi \rangle. \quad (\text{A 1.6.96})$$

Assuming that the system has no permanent dipole moment, the existence of $P(t)$ depends on a non-stationary ψ induced by an external electric field. For weak fields, we may expand the polarization in orders of the perturbation,

$$P(t) \equiv \langle \psi | \mu | \psi \rangle = P^{(0)}(t) + P^{(1)}(t) + P^{(2)}(t) + P^{(3)}(t) + \dots, \quad (\text{A 1.6.97})$$

We can then identify each term in the expansion with one or more terms in the perturbative expansion of

$$P^{(0)}(t) \equiv \langle \psi^{(0)}(t) | \mu | \psi^{(0)}(t) \rangle \quad (\text{A 1.6.98})$$

$$P^{(1)}(t) \equiv \langle \psi^{(0)}(t) | \mu | \psi^{(1)}(t) \rangle + \text{cc} \quad (\text{A 1.6.99})$$

$$P^{(2)}(t) = \langle \psi^{(0)}(t) | \mu | \psi^{(2)}(t) \rangle + \text{cc} + \langle \psi^{(1)}(t) | \mu | \psi^{(1)}(t) \rangle \quad (\text{A 1.6.100})$$

and

$$P^{(3)}(t) \equiv \langle \psi^{(0)}(t) | \mu | \psi^{(3)}(t) \rangle + \text{cc} + \langle \psi^{(1)}(t) | \mu | \psi^{(2)}(t) \rangle + \text{cc} \quad (\text{A 1.6.101})$$

etc. Note that for an isotropic medium, terms of the form $P^{(2n)}(t)$ ($P^{(0)}(t)$, $P^{(2)}(t)$, etc) do not survive orientational averaging. For example, the first term, $\langle \psi^{(0)} | \mu | \psi^{(0)} \rangle$, is the permanent dipole moment, which gives zero when averaged over an isotropic medium. At an interface, however (e.g. between air and water), these even orders of $P(t)$ do not vanish, and in fact are sensitive probes of interface structure and dynamics.

The central dynamical object that enters into the polarization are the coherences of the form $\langle \psi^{(0)}(t) | \mu | \psi^{(1)}(t) \rangle$ and $\langle \psi^{(1)}(t) | \mu | \psi^{(2)}(t) \rangle$, etc. These quantities are overlaps between wavepackets moving on different potential energy surfaces [40, 41 and 42, 52]: the instantaneous overlap of the wavepackets creates a non-vanishing transition dipole moment which interacts with the light. This view is appropriate both in the regime of weak fields, where perturbation theory is valid, and for strong fields, where perturbation theory is no longer valid. Note that in the previous sections we saw that the absorption and Raman spectra were related to $\frac{d}{dt} \langle \psi^{(1)}(t) | \psi^{(1)}(t) \rangle$ and $\frac{d}{dt} \langle \psi^{(2)}(t) | \psi^{(2)}(t) \rangle$. The coherences that appear in [equation \(A1.6.99\)](#) and [equation \(A1.6.101\)](#) are precisely equivalent to these derivatives: the rate of change of a population is proportional to the instantaneous coherence, a relationship which can be observed already in the vector precession model of the two-level system ([section A1.6.2.3](#)).

The coherences can be written compactly using the language of density matrices. The total polarization is given by

$$P = \text{Tr}(\rho \mu) = P^{(0)}(t) + P^{(1)}(t) + P^{(2)}(t) + P^{(3)}(t) + \dots, \quad (\text{A 1.6.102})$$

where the different terms in the perturbative expansion of P are accordingly as follows:

$$P^{(1)} = \text{Tr}(\rho^{(1)}\mu) \quad P^{(2)} = \text{Tr}(\rho^{(2)}\mu) \quad P^{(3)} = \text{Tr}(\rho^{(3)}\mu) \text{ etc.} \quad (\text{A 1.6.103})$$

In the absence of dissipation and pure state initial conditions, equation (A1.6.102) and equation (A1.6.103) are equivalent to equation (A1.6.97), (A1.6.98), (A1.6.99), (A1.6.100) and (A1.6.101). But equation (A1.6.102) and equation (A1.6.103) are more general, allowing for the possibility of dissipation, and hence for describing nonlinear spectroscopy in the presence of an environment. There is an important caveat however. In the presence of an environment, it is customary to define a reduced density matrix which describes the system, in which the environment degrees of freedom have been traced out. The tracing out of the environment should be performed only at the end, after all the interactions of the system environment with the field, otherwise important parts of the nonlinear spectrum (e.g. phonon sidebands) will be missing. The tracing of the environment at the end can be done analytically if the system is a two-level system and the environment is harmonic, the so-called spin-boson or Brownian oscillator model. However, in general the dynamics in the full system-environment degrees of freedom must be calculated, which essentially entails a return to a wavefunction description, equation (A1.6.97), equation (A1.6.98), equation (A1.6.99), equation (A1.6.100) and equation (A1.6.101), but in a larger space.

The total of three interactions of the material with the field can be distributed in several different ways between the ket and the bra (or more generally, between left and right interactions of the field with the density operator). For example, the first term in equation (A1.6.101) corresponds to all three interactions being with the ket, while the second term corresponds to two interactions with the ket and one with the bra. The second term can be further subdivided into three possibilities: that the single interaction with the bra is before, between or after the two interactions with the ket (or correspondingly, left/right interactions of the field with the density operator) [37]. These different contributions to $P^{(3)}$ (or, equivalently, to $\rho^{(3)}$) are represented conveniently using double-sided Feynman diagrams, a generalization of the single-sided Feynman diagrams introduced in section A1.6.3, as shown in figure A1.6.17.

-42-

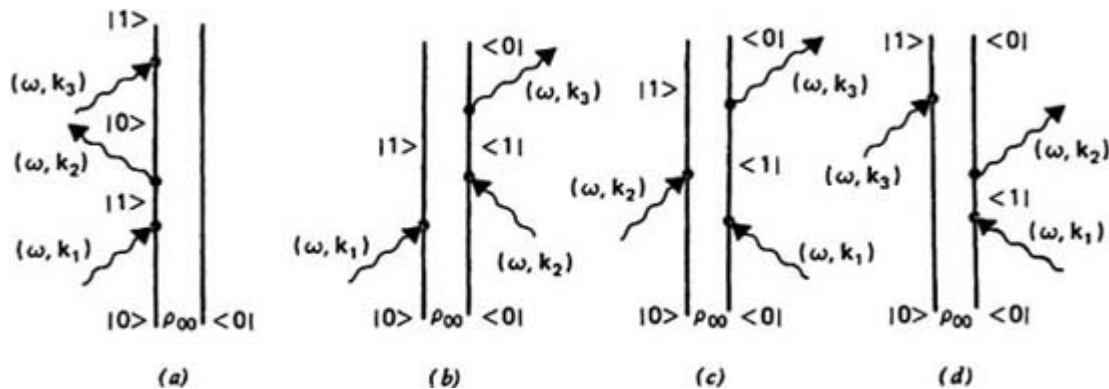


Figure A1.6.17. Double-sided Feynman diagrams, showing the interaction time with the ket (left) and the bra (right). Time moves forward from down to up (adapted from [36]).

The subdivision of the second term into three possibilities has an interesting physical interpretation. The ordering of the interactions determines whether diagonal vs off-diagonal elements of the density matrix are produced: populations versus coherences. In the presence of relaxation processes (dephasing and population relaxation) the order of the interactions and the duration between them determines the duration for which population versus coherence relaxation mechanisms are in effect. This can be shown schematically using a Liouville space diagram, figure A1.6.18 [37]. The different pathways in Liouville space are drawn on a lattice, where ket interactions are horizontal steps and bra interactions are vertical. The diagonal vertices represent populations and the off-diagonal vertices are coherences. The three different time orderings for contributions to $|\psi^{(2)}\rangle\langle\psi^{(1)}|$ correspond to the three Liouville pathways shown in figure A1.6.18. From such a diagram one

sees at a glance which pathways pass through intermediate populations (i.e. diagonal vertices) and hence are governed by population decay processes, and which pathways do not.

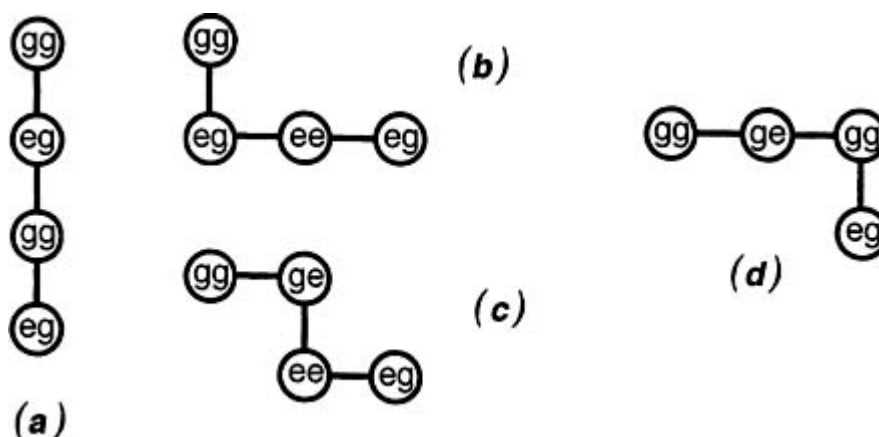


Figure A1.6.18. Liouville space lattice representation in one-to-one correspondence with the diagrams in figure A1.6.17. Interactions of the density matrix with the field from the left (right) is signified by a vertical (horizontal) step. The advantage to the Liouville lattice representation is that populations are clearly identified as diagonal lattice points, while coherences are off-diagonal points. This allows immediate identification of the processes subject to population decay processes (adapted from [37]).

As a first application of the lattice representation of Liouville pathways, it is interesting to re-examine the process of electromagnetic spontaneous light emission, discussed in the previous section. Note that formally, diagrams A1.6.18 (all contribute to the Kramers–Heisenberg–Dirac formula for resonance Raman scattering. However, diagrams (b) and (c) produce an excited electronic state population (both the bra and ket are excited in the first two interactions) and hence are subject to excited-state vibrational population relaxation processes, while diagram (d) does not. Typically, in the condensed phase, the fluorescence spectrum consists of sharp lines against a broad background. Qualitatively speaking, the sharp lines are associated with diagram (d), and are called the resonance Raman spectrum, while the broad background is associated with diagrams (a) and (b), and is called the resonance fluorescence spectrum [38]. Indeed, the emission frequency of the sharp lines changes with the excitation frequency, indicating no excited electronic state population relaxation, while the broad background is independent of excitation frequency, indicating vibrationally relaxed fluorescence.

There is an aspect of nonlinear spectroscopy which we have so far neglected, namely the spatial dependence of the signal. In general, three incident beams, described by \mathbf{k} -vectors \mathbf{k}_1 , \mathbf{k}_2 and \mathbf{k}_3 will produce an outgoing beam at each of the directions:

$$\mathbf{k}_{\text{out}} = \pm \mathbf{k}_1 \pm \mathbf{k}_2 \pm \mathbf{k}_3. \quad (\text{A } 1.6.104)$$

Figure A1.6.19 shows eight out of the 48 Feynman diagrams that contribute to an outgoing \mathbf{k} -vector at $-\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3$. The spatial dependence is represented by the wavevector \mathbf{k} on each of the arrows in figure A1.6.19. Absorption (emission) by the ket corresponds to a plus (minus) sign of \mathbf{k} ; absorption (emission) by the bra corresponds to a minus (plus) sign of \mathbf{k} . The eight diagrams shown dominate under conditions of electronic resonance; the other 40 diagrams correspond to non-resonant contributions, involving emission before absorption. The reason there are eight resonant diagrams now, instead of the four in figure A1.6.17, is a result of the fact that the introduction of the \mathbf{k} -dependence makes the order of the interactions distinguishable. At the same time, the \mathbf{k} -dependence of the detection eliminates many additional processes that might otherwise

contribute; for example, detection at $-k_1 + k_2 + k_3$ eliminates processes in which k_1 and k_2 are interchanged, as well as processes representing two or more interactions with a single beam. Under conditions in which the interactions have a well-defined temporal sequence, just two diagrams dominate, while two of the diagrams in figure A1.6.17 are eliminated since they emit to $k_1 - k_2 + k_3$. Below we will see that in resonant CARS, where in addition to the electronic resonance there is a vibrational resonance after the second interaction, there is only a single resonant diagram. All else being equal, the existence of multiple diagrams complicates the interpretation of the signal, and experiments that can isolate the contribution of individual diagrams have a better chance for complete interpretation and should be applauded.

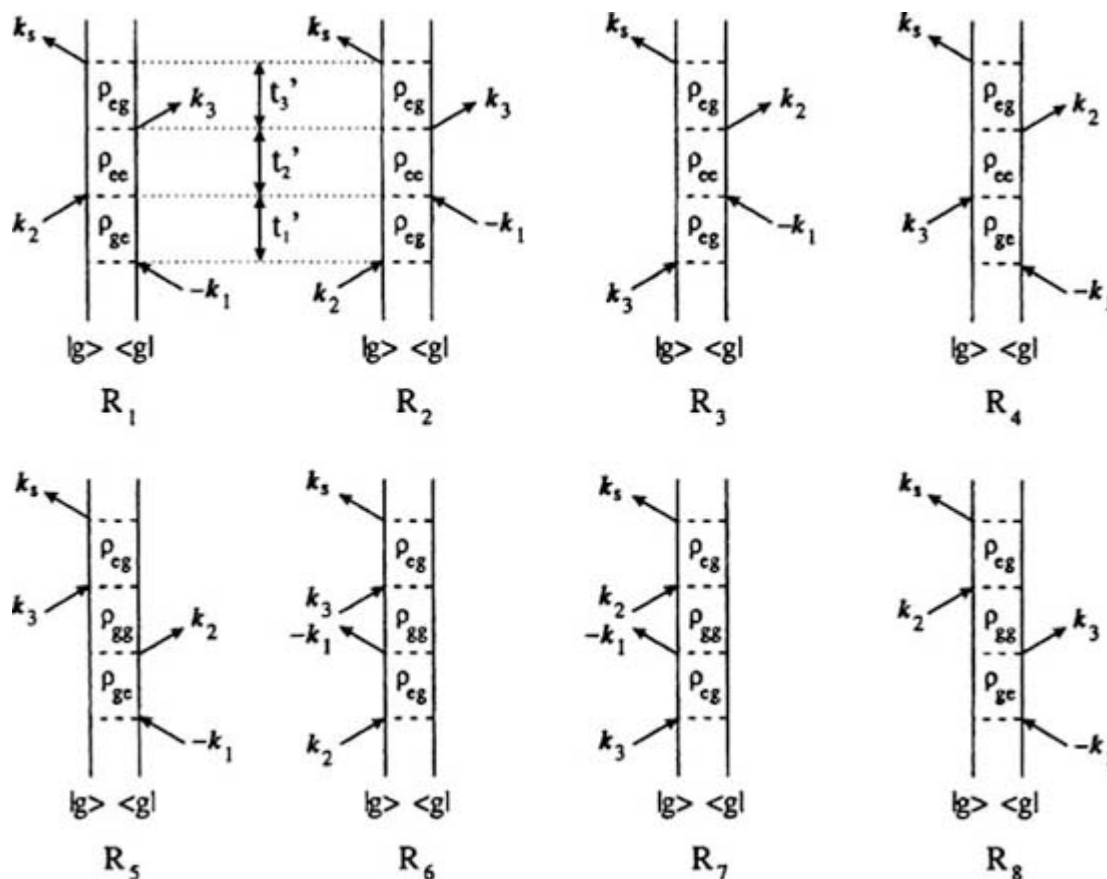


Figure A1.6.19. Eight double-sided Feynman diagrams corresponding to electronic resonance and emission at $-k_1 + k_2 + k_3$. Absorption is shown by an incoming arrow, while emission is indicated by an outgoing arrow. Note that if an arrow is moved from (to) the ket to (from) the bra while changing from (to) absorption to (from) emission, the slope of the arrow and therefore its k -vector will be unchanged. The eight diagrams are for arbitrary time ordering of the interactions; with a fixed time ordering of the interactions, as in the case of non-overlapping pulses, only two of the diagrams survive (adapted from [48]).

A1.6.4.2 LINEAR RESPONSE

We now proceed to the spectrum, or frequency-dependent response [41, 42]. The power, or rate of energy absorption, is given by

$$P = \frac{dE}{dt} = \frac{d\langle\psi|H|\psi\rangle}{dt} = -2\text{Re}\{\langle\psi_a|\mu|\psi_b\rangle\dot{E}^*(t)\}. \tag{A 1.6.105}$$

(In the second step we have used [equation \(A1.6.72\)](#) and noted that the terms involving $\partial\psi/\partial t$ cancel.) To lowest order, this gives

$$P = -2\text{Re}\{P_{01}^{(1)}\dot{E}^*\}. \quad (\text{A 1.6.106})$$

The total energy absorbed, ΔE , is the integral of the power over time. Keeping just the lowest order terms we find

$$\Delta E = \int_{-\infty}^{\infty} P dt = -2\text{Re} \int_{-\infty}^{\infty} P_{01}^{(1)}(t)\dot{E}^*(t) dt = 2\text{Im} \int_{-\infty}^{\infty} \omega \tilde{P}_{01}^{(1)}(\omega)\tilde{E}^*(\omega) d\omega \quad (\text{A 1.6.107})$$

where

$$\tilde{P}_{01}^{(1)}(\omega) \equiv \int_{-\infty}^{\infty} P_{01}^{(1)}(t)e^{i\omega t} dt \quad (\text{A 1.6.108})$$

and

$$\tilde{E}(\omega) \equiv \int_{-\infty}^{\infty} E(t)e^{i\omega t} dt. \quad (\text{A 1.6.109})$$

The last relation in [equation \(A1.6.107\)](#) follows from the Fourier convolution theorem and the property of the Fourier transform of a derivative; we have also assumed that $E(\omega) = E(-\omega)$. The absorption spectrum is defined as the total energy absorbed at frequency ω , normalized by the energy of the incident field at that frequency. Identifying the integrand on the right-hand side of [equation \(A1.6.107\)](#) with the total energy absorbed at frequency ω , we have

$$\sigma(\omega) = \frac{|\tilde{E}'(\omega)|^2}{|\tilde{E}(\omega)|^2} = \frac{4\pi\omega}{3c\hbar} \frac{\text{Im}(\tilde{P}_{01}^{(1)}(\omega)\tilde{E}^*(\omega))}{|\tilde{E}(\omega)|^2}. \quad (\text{A 1.6.110})$$

Note the presence of the ω prefactor in the absorption spectrum, as in [equation \(A1.6.87\)](#); again its origin is essentially the faster rate of the change of the phase of higher frequency light, which in turn is related to a higher rate of energy absorption. The equivalence between the other factors in [equation \(A1.6.110\)](#) and [equation \(A1.6.87\)](#) under linear response will now be established.

In the perturbative regime one may decompose these coherences into the contribution from the field and a part which is intrinsic to the matter, the *response* function. For example, note that the expression $P_{01}^{(1)}(t) = \langle \psi^{(0)}(t) | \mu | \psi^{(1)}(t) \rangle$ is not simply an intrinsic function of the molecule: it depends on the functional form of the field, since $\psi^{(1)}(t)$ does. However, since the dependence on the field is linear it is possible to write $P_{01}^{(1)}$ as a *convolution* of the field with a response function which depends on the material. Using the definition of $\psi^{(1)}$,

$$\psi^{(1)} = \frac{1}{i\hbar} \int_{-\infty}^t e^{-iH_b(t-t')} \{-\mu E(t')\} e^{-iE_0 t'} \psi^{(0)} dt' \quad (\text{A 1.6.111})$$

we find that

$$P_{01}^{(1)}(t) = \frac{1}{i\hbar} \int_{-\infty}^t \langle \psi^{(0)}(t) | \mu e^{-iH_b(t-t')} \{-\mu E(t')\} e^{-iE_0 t'} | \psi^{(0)} \rangle dt' \quad (\text{A 1.6.112})$$

$$= \frac{i}{\hbar} \int_0^{\infty} \langle \psi^{(0)}(t) | \mu e^{-iH_b \tau} \mu | \psi^{(0)} \rangle E(t - \tau) d\tau \quad (\text{A 1.6.113})$$

$$= \frac{i}{\hbar} \{E(t) \otimes S_{00}(t)\} \quad (\text{A 1.6.114})$$

where $S_{00}(t)$ is the half or *causal* form of the autocorrelation function:

$$S_{00}(t) = \begin{cases} C_{00}(t) & t > 0 \\ 0 & t < 0 \end{cases} \quad (\text{A 1.6.115})$$

$$(\text{A 1.6.116})$$

and \otimes signifies convolution. We have defined the wavepacket autocorrelation function

$$C_{00}(t) \equiv \langle \psi^{(0)} | \mu e^{-iH_b t/\hbar} \mu | \psi^{(0)} \rangle. \quad (\text{A 1.6.117})$$

where $C_{00}(t)$ is just the wavepacket autocorrelation function we encountered in [section A1.6.3.3](#). There we saw that the Fourier transform of $C_{00}(t)$ is proportional to the linear absorption spectrum. The same result appears here but with a different interpretation. There, the correlation function governed the rate of excited-state population change. Here, the expectation value of the dipole moment operator with the correlation function is viewed as the *response* function of the molecule.

By the Fourier convolution theorem

$$P_{01}^{(1)}(\omega) \equiv \int_{-\infty}^{\infty} P_{01}^{(1)}(t) e^{i\omega t} dt = \left\{ \frac{1}{i\hbar} \tilde{E}(\omega) \tilde{S}_{00}(\omega) \right\}. \quad (\text{A 1.6.118})$$

Using the definition of the susceptibility, χ ([equation \(A1.6.30\)](#)) we see that

$$\chi^{(1)}(\omega) = \left\{ \frac{1}{i\hbar} \tilde{S}_{00}(\omega) \right\}. \quad (\text{A 1.6.119})$$

Substituting $P_{01}^{(1)}(\omega)$ into [equation \(A1.6.110\)](#) we find that the linear absorption spectrum is given by

$$\sigma(\omega) = \frac{4\pi\omega}{3c\hbar} \text{Re}\{\tilde{S}_{00}(\omega)\} \quad (\text{A 1.6.120})$$

$$= \frac{2\pi\omega}{3c\hbar} \int_{-\infty}^{\infty} C_{00}(t) e^{i\omega t} dt \quad (\text{A 1.6.121})$$

in agreement with [equation \(A1.6.87\)](#). We also find that

$$\sigma(\omega) = \frac{4\pi\omega}{3c\hbar} \text{Im} \{ \chi^{(1)}(\omega) \} \quad (\text{A 1.6.122})$$

establishing the result in section A1.6.1.4 that the absorption spectrum is related to the imaginary part of the susceptibility χ at frequency ω .

A1.6.4.3 NONLINEAR RESPONSE: ISOLATED SYSTEMS

As discussed above, the nonlinear material response, $P^{(3)}(t)$ is the most commonly encountered nonlinear term since $P^{(2)}$ vanishes in an isotropic medium. Because of the special importance of $P^{(3)}$ we will discuss it in some detail. We will now focus on a few examples of $P^{(3)}$ spectroscopy where just one or two of the 48 double-sided Feynman diagrams are important, and will stress the dynamical interpretation of the signal. A pictorial interpretation of all the different resonant diagrams in terms of wavepacket dynamics is given in [41].

COHERENT ANTI-STOKES RAMAN SPECTROSCOPY (CARS)

Our first example of a $P^{(3)}$ signal is coherent anti-Stokes Raman spectroscopy, or CARS. Formally, the emission signal into direction $k = k_1 - k_2 + k_3$ has 48 Feynman diagrams that contribute. However, if the frequency ω_1 is resonant with the electronic transition from the ground to the excited electronic state, and the mismatch between frequencies ω_1 and ω_2 is resonant with a ground-state vibrational transition or transitions, only one diagram is resonant, namely, the one corresponding to R_6 in figure A1.6.19 (with the interchange of labels k_1 and k_2).

To arrive at a dynamical interpretation of this diagram it is instructive to write the formula for the dominant term in $P^{(3)}$ explicitly:

$$\begin{aligned} P^{(3)}(t) &= \langle \psi^{(0)}(t) | \mu | \psi^{(3)}(t) \rangle \\ &= \frac{(-)^3}{(i\hbar)^3} \int_{-\infty}^{t_4} dt_3 \int_{-\infty}^{t_3} dt_2 \int_{-\infty}^{t_2} dt_1 \langle \psi^{(0)}(t') | \{ \mu \} e^{-iH_b(t-t_3)/\hbar} \\ &\quad \{ \mu E_3(t_3) \} e^{-iH_a(t_3-t_2)/\hbar} \end{aligned} \quad (\text{A 1.6.123})$$

$$\times \{ \mu E_2(t_2) \} e^{-iH_b(t_2-t_1)/\hbar} \{ \mu E_1(t_1) \} e^{-iH_a t_1} | \psi^{(0)} \rangle \quad (\text{A 1.6.124})$$

where in the second line we have substituted explicitly for the third-order wavefunction, $\psi^{(3)}(t)$. This formula, although slightly longer than the formulae for the first- and second-order amplitude discussed in the previous section, has the same type of simple dynamical interpretation. The initial wavepacket, $\psi^{(0)}$ interacts with the field at time t_1 and propagates on surface b for time $t_2 - t_1$; at time t_2 it interacts a second time with the field and propagates on the ground surface a for time $t_3 - t_2$; at time t_3 it interacts a third time with the field and propagates on surface b until variable time t . The third-order wavepacket on surface b is projected onto the initial wavepacket on the ground state; this overlap

is a measure of the coherence which determines both the magnitude and phase of the CARS signal. Formally, the expression involves an integral over three time variables, reflecting the coherent contribution of all possible instants at which the interaction with the light took place, for each of the three interactions. However, if the interaction is with pulses that are short compared with a vibrational period, as we saw in equation (A1.6.76), one can approximate the pulses by δ -functions in time, eliminate the three integrals and the simple dynamical interpretation above becomes precise.

Qualitatively, the delay between interaction 1 and 2 is a probe of excited-state dynamics, while the delay between interaction 2 and 3 reflects ground-state dynamics. If pulses 1 and 2 are coincident, the combination of the first two pulses prepares a vibrationally excited wavepacket on the ground-state potential energy surface; the time delay between pulses 2 and 3 then determines the time interval for which the wavepacket evolves on the ground-state potential, and is thus a probe of ground-state dynamics [43, 45, 52]. If a second delay, the delay between pulses 1 and 2, is introduced this allows large wavepacket excursions on the excited state before coming back to the ground state. The delay between pulses 1 and 2 can be used in a very precise way to tune the level of ground-state vibrational excitation, and can prepare ground vibrational wavepackets with extremely high energy content [44]. The sequence of pulses involving one against two time delays is shown in figure A1.6.20 (a) and figure A1.6.20(b). The control over the vibrational energy content in the ground electronic state via the delay between pulses 1 and 2 is illustrated in figure A1.6.20 (right).

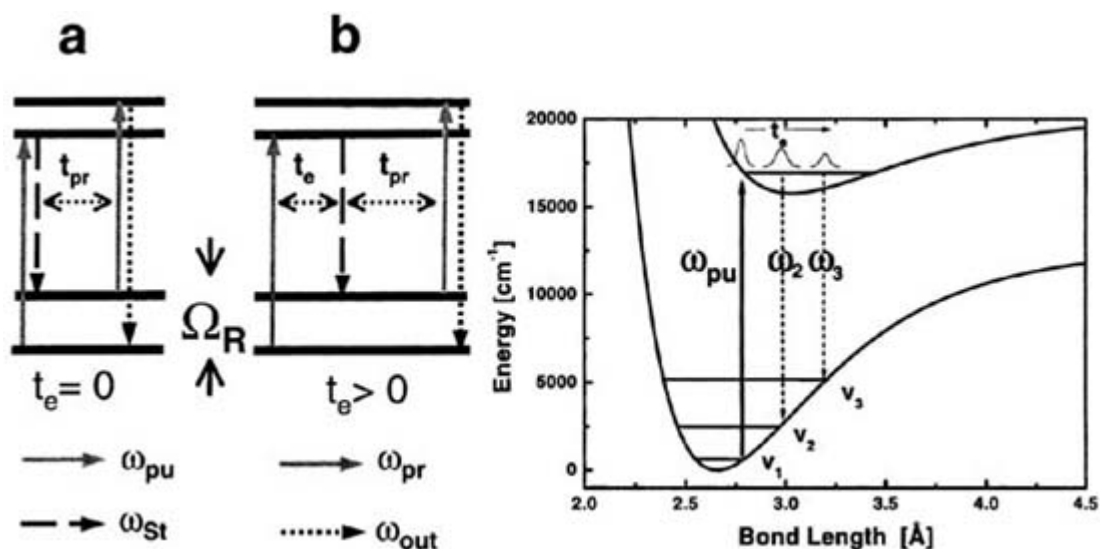


Figure A1.6.20. (Left) Level scheme and nomenclature used in (a) single time-delay CARS. (b) Two-time delay CARS ((TD)²CARS). The wavepacket is excited by ω_{pu} , then transferred back to the ground state by ω_{st} with Raman shift ω_R . Its evolution is then monitored by ω_{pr} (after [44]). (Right) Relevant potential energy surfaces for the iodine molecule. The creation of the wavepacket in the excited state is done by ω_{pu} . The transfer to the final state is shown by the dashed arrows according to the state one wants to populate (after [44]).

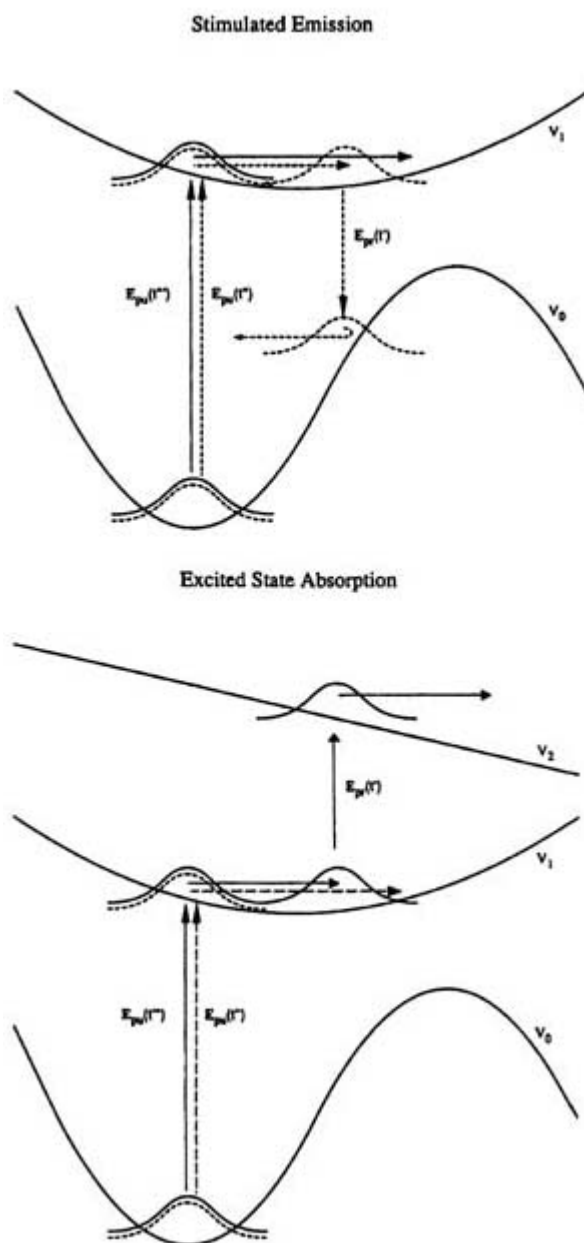


Figure A1.6.21. Bra and ket wavepacket dynamics which determine the coherence overlap, $\langle \phi^{(1)} | \phi^{(2)} \rangle$. Vertical arrows mark the transitions between electronic states and horizontal arrows indicate free propagation on the potential surface. Full curves are used for the ket wavepacket, while dashed curves indicate the bra wavepacket. (a) Stimulated emission. (b) Excited state (transient) absorption (from [41]).

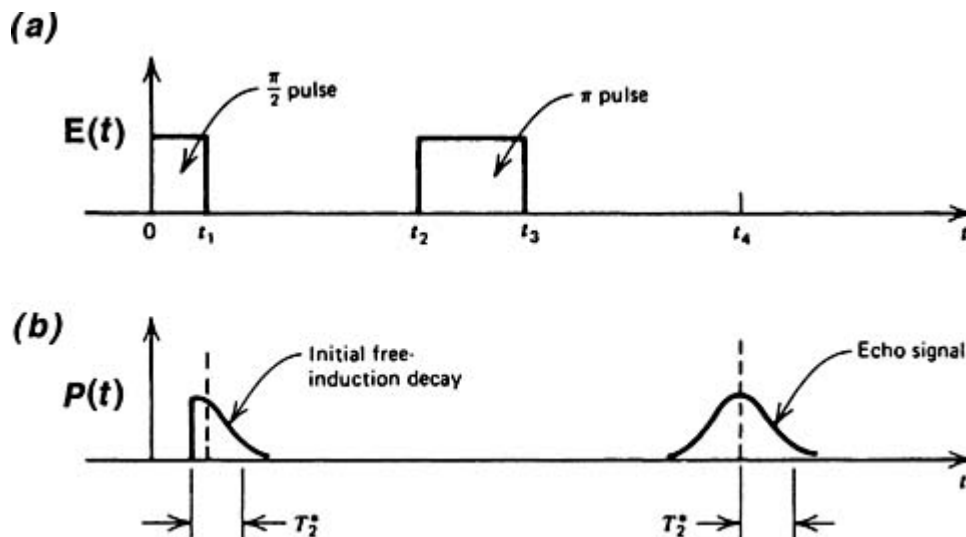


Figure A1.6.22 (a) Sequence of pulses in the canonical echo experiment. (b) Polarization versus time for the pulse sequence in (a), showing an echo at a time delay equal to the delay between the excitation pulses.

(B) STIMULATED RAMAN AND DYNAMIC ABSORPTION SPECTROSCOPY

In CARS spectroscopy, $\omega_1 = \omega_3$, and ω_2 is generally different and of lower frequency. If $\omega_1 = \omega_2 = \omega_3$ the process is called degenerate four-wave mixing (DFWM). Now, instead of a single diagram dominating, two diagrams participate if the pulses are non-overlapping, four dominate if two pulses overlap and all eight resonant diagrams contribute if all three pulses overlap (e.g., in continuous wave excitation) [43, 46]. The additional diagrams correspond to terms of the form $\langle \psi^{(1)}(t) | \mu | \psi^{(2)}(t) \rangle$ discussed above; this is the overlap of a second-order wavepacket on the ground-state surface with a first-order wavepacket on the excited-state surface. These new diagrams come in for two reasons. First, even if the pulses are non-overlapping, the degeneracy of the first two interactions allows the second interaction to produce an absorption, not just emission. If the pulses are overlapping there is the additional flexibility of interchanging the order of pulses 1 and 2 (at the same time exchanging their role in producing absorption versus emission). The contribution of these additional diagrams to the $P^{(3)}$ signal is not simply additive, but there are interference terms among all the contributions, considerably complicating the interpretation of the signal. Diagrams R_1 – R_4 are commonly referred to as stimulated Raman scattering: the first two interactions produce an excited-state population while the last interaction produces stimulated emission back to the ground electronic state.

A process which is related diagrammatically to stimulated Raman scattering is transient absorption spectroscopy. In an ordinary absorption spectrum, the initial state is typically the ground vibrational eigenstate of the ground electronic state. Dynamic absorption spectroscopy refers to the excitation of a vibrational wavepacket to an electronic state b via a first pulse, and then the measurement of the spectrum of that moving wavepacket on a third electronic state c as function of time delay between the pump and the probe. The time delay controls the instantaneous wavepacket on state b whose spectrum is being measured with the second pulse; in an ideal situation, one may obtain ‘snapshots’ of the wavepacket on electronic b as a function of time, by observing its shadow onto surface c . This form of spectroscopy is very similar in spirit to the pump–probe experiments of Zewail *et al* [25], described in section A1.6.3.2, but there are two differences. First, the signal in a dynamic absorption spectrum is a coherent signal in the direction of the probe

pulse (pulse 3), as opposed to measuring fluorescence from state c , which is non-directional. Second, field intensity in the direction of the probe pulse can be frequency resolved to give simultaneous time and

frequency resolution of the transient absorption. Although in principle the fluorescence from state c can also be frequency resolved, this fluorescence takes place over a time which is orders of magnitude longer than the vibrational dynamics of interest and the signal contains a complicated combination of all excited- and ground-state frequency differences.

The dynamic absorption signal, $P^{(3)}$, can be written in a form which looks analogous to the linear absorption signal $P^{(1)}$ (see [equation \(A1.6.113\)](#)),

$$P^{(3)}(t) = \frac{i}{\hbar} \int_{-\infty}^{\infty} \langle \psi^{(1)}(t') | \mu e^{-iH_c(t-t')/\hbar} \mu | \psi^{(1)}(t') \rangle E(t') dt'. \quad (\text{A 1.6.125})$$

However, because of the t' dependence in $\psi^{(1)}(t')$ one cannot write that $P^{(3)} = E(t) \otimes S_{11}(t)$. For the latter to hold, it is necessary to go to the limit of a probe pulse which is short compared with the dynamics on surface 1. In this case, $\psi^{(1)}(t')$ is essentially frozen and we can write $\psi^{(1)} \approx \psi_{\tau}^{(1)}$, where we have indicated explicitly the parametric dependence on the pump–probe delay time, τ . In this case, [equation \(A1.6.125\)](#) is isomorphic with [equation \(A1.6.113\)](#), indicating that under conditions of impulsive excitation, dynamic absorption spectroscopy is just first-order spectroscopy on the frozen state, $\phi_{\tau}^{(1)}$, on surface c . Note the residual dependence of the frozen state on τ , the pump–probe delay, and thus variation of the variables (ω, τ) generates a two-dimensional dynamic absorption spectrum. Note that the pair of variables (ω, τ) are not limited by some form of time–energy uncertainty principle. This is because, although the absorption is finished when the probe pulse is finished, the spectral analysis of which frequency components were absorbed depends on the full time evolution of the system, beyond its interaction with the probe pulse. Thus, the dynamic absorption signal can give high resolution both in time (i.e. time delay between pump and probe pulses) and frequency, simultaneously.

A1.6.4.4 NONLINEAR RESPONSE: SYSTEMS COUPLED TO AN ENVIRONMENT

(A) ECHO SPECTROSCOPY

In discussing spectroscopy in condensed phase environments, one normally distinguishes two sources of decay of the coherence: inhomogeneous decay, which represents static differences in the environment of different molecules, and homogeneous decay, which represents the dynamics interaction with the surroundings and is the same for all molecules. Both these sources of decay contribute to the *linewidth* of spectral lines; in many cases the inhomogeneous decay is faster than the homogeneous decay, masking the latter. In echo spectroscopies, which are related to a particular subset of diagrams in $P^{(3)}$, one can at least partially discriminate between homogeneous and inhomogeneous decay.

Historically, photon echoes grew up as optical analogues of spin echoes in NMR. Thus, the earliest photon echo experiments were based on a sequence of two excitation pulses, a $\pi/2$ pulse followed by a π pulse, analogous to the pulse sequence used in NMR. Conceptually, the $\pi/2$ pulse prepares an optical coherence, which will proceed to dephase due to both homogeneous and inhomogeneous mechanisms. After a delay time τ , the π pulse reverses the role of the excited and ground electronic states, which causes the inhomogeneous contribution to the dephasing to reverse itself but does not affect the homogeneous decay. The reversal of phases generated by the π -pulse has been described in many colourful ways over the years (see [figure A1.6.23](#)).

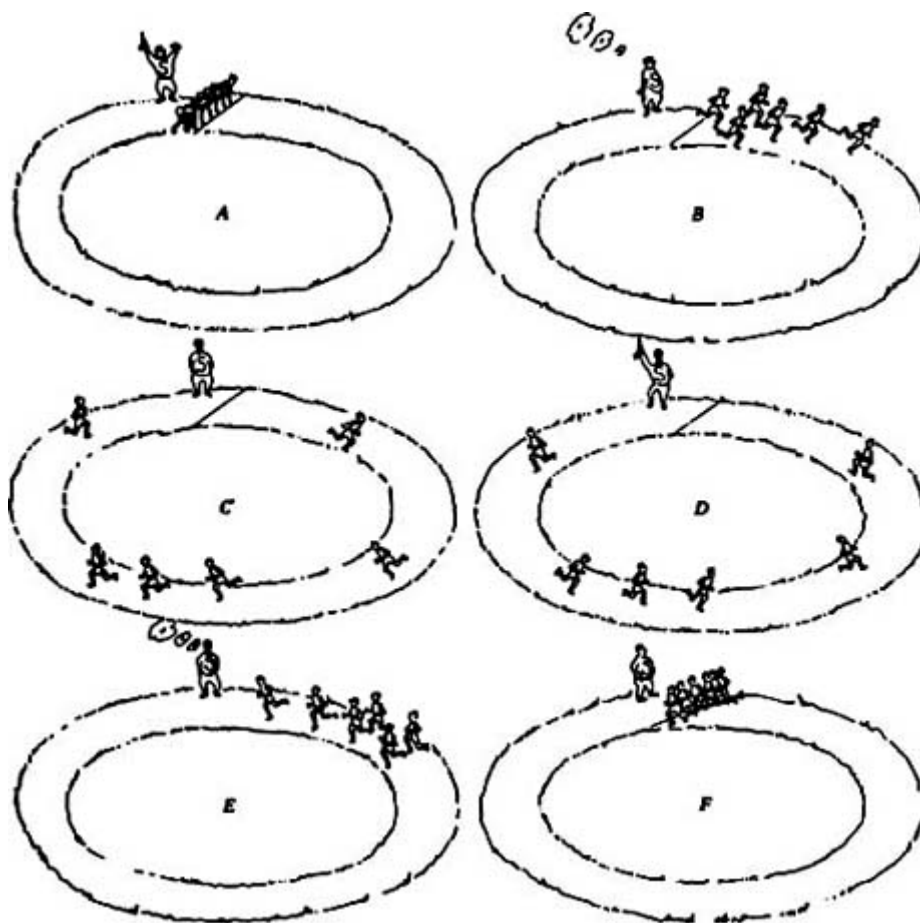


Figure A1.6.23. Schematic representation of dephasing and reversal on a race track, leading to coherent rephasing and an 'echo' of the starting configuration. From *Phys. Today*, (Nov. 1953), front cover.

Fundamentally, the above description of photon echoes is based on a two-level description of the system. As we have seen throughout this article, much of molecular electronic spectroscopy is described using two electronic states, albeit with a vibrational manifold in each of these electronic states. This suggests that photon echoes can be generalized to include these vibrational manifolds, provided that the echo signal is now defined in terms of a wavepacket overlap (or density matrix coherence) involving the coherent superposition of all the participating vibrational levels. This is shown schematically in [figure A1.6.24](#). The $\pi/2$ pulse transfers 50% of the wavepacket amplitude to the excited electronic state. This creates a non-stationary vibrational wavepacket in the excited electronic state (and generally, the remaining amplitude in the ground electronic state is non-stationary as well). After a time delay τ a π pulse comes in, exchanging the wavepackets on the ground and excited electronic states. The wavepackets continue to evolve on their new respective surfaces. At some later time, when the wavepackets overlap, an echo will be observed. This sequence is shown in [figure A1.6.24](#). Note that this description refers only to the isolated molecule; if there are dephasing mechanisms due to the environment as well, the echo requires the rephasing in both the intramolecular and the environmental degrees of freedom.

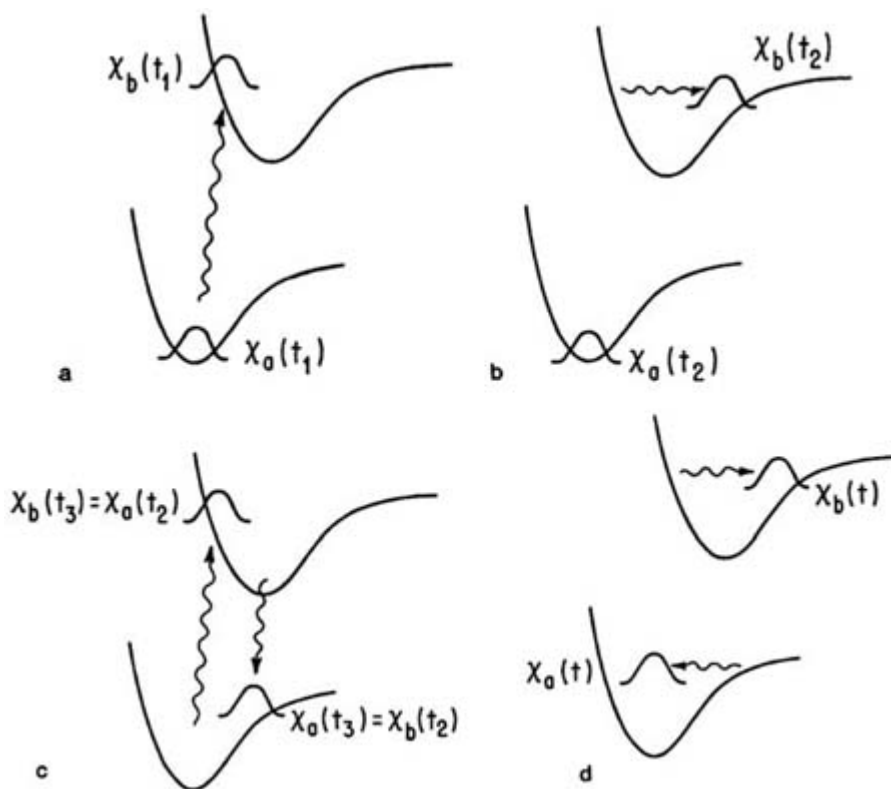


Figure A1.6.24. Schematic representation of a photon echo in an isolated, multilevel molecule. (a) The initial pulse prepares a superposition of ground- and excited-state amplitude. (b) The subsequent motion on the ground and excited electronic states. The ground-state amplitude is shown as stationary (which in general it will not be for strong pulses), while the excited-state amplitude is non-stationary. (c) The second pulse exchanges ground- and excited-state amplitude. (d) Subsequent evolution of the wavepackets on the ground and excited electronic states. When they overlap, an echo occurs (after [40]).

Although the early photon echo experiments were cast in terms of $\pi/2$ and π pulses, these precise inversions of the population are by no means necessary [36]. In fact echoes can be observed using sequences of weak pulses, and can be described within the perturbative $P^{(3)}$ formalism which we have used throughout [section A1.6.4](#). Specifically, the diagrams R_1 , R_4 , R_5 and R_8 in [figure A1.6.19](#) correspond to echo diagrams, while the diagrams R_2 , R_3 , R_6 and R_7 do not. In the widely used Brownian oscillator model for the relaxation of the system [37, 48], the central dynamical object is the electronic frequency correlation,

$$M(t) = \frac{\langle \Delta\omega(0)\Delta\omega(t) \rangle}{\langle \Delta\omega^2 \rangle} \quad (\text{A 1.6.126})$$

where $\Delta\omega(t) = \langle \omega_{eg} \rangle - \omega(t)$. Here $\langle \omega_{eg} \rangle$ is the average transition frequency, $\omega(t)$ is the transition frequency at time t , and the brackets denote an ensemble average. It can be shown that as long as $M(t)$ is a monotonically decaying function, the diagrams R_1 , R_4 , R_5 and R_8 can cause rephasing of $P^{(3)}$ while the diagrams R_2 , R_3 , R_6 and R_7 cannot (see [figure A1.6.25](#)).

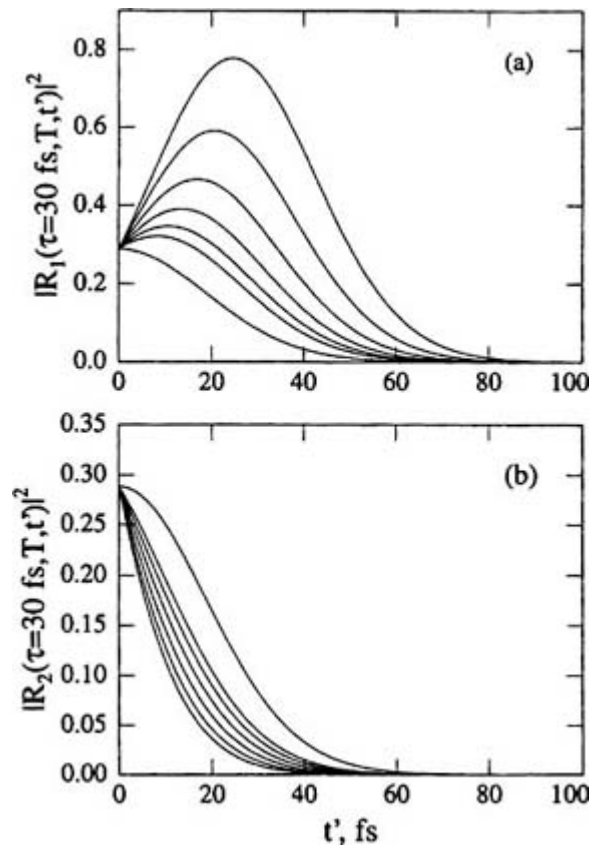


Figure A1.6.25. Modulus squared of the rephasing, $|R_1|^2$, (a), and non-rephasing, $|R_2|^2$, (b), response functions versus final time t for a near-critically overdamped Brownian oscillator model $M(t)$. The time delay between the second and third pulse, T , is varied as follows: (a) from top to bottom, $T = 0, 20, 40, 60, 80, 100, \infty$ fs; (b) from bottom to top, $T = 0, 20, 40, 60, 80, 100, \infty$ fs. Note that $|R_1|^2$ and $|R_2|^2$ are identical at $T = \infty$. After [48].

It is instructive to contrast echo spectroscopy with single time-delayed CARS spectroscopy, discussed above. Schematically, TD-CARS spectroscopy involves the interaction between pulses 1 and 2 being close in time, creating a ground-state coherence, and then varying the delay before interaction 3 to study ground-state dynamics. In contrast, echo spectroscopy involves an isolated interaction 1 creating an electronic coherence between the ground and the excited electronic state, followed by a pair of interactions 2 and 3, one of which operates on the bra and the other on the ket. The pair of interactions 2,3 essentially reverses the role of the ground and the excited electronic states. If there is any inhomogeneous broadening, or more generally any bath motions that are slow compared with the time intervals between the pulses, these modes will show up as echo signal after the third pulse is turned off [47].

We close with three comments. First, there is preliminary work on retrieving not only the amplitude but also the phase of photon echoes [49]. This appears to be a promising avenue to acquire complete 2-dimensional time and frequency information on the dynamics, analogous to methods that have been used in NMR. Second, we note that there is a growing literature on non-perturbative, numerical simulation of nonlinear spectroscopies. In these methods, the consistency of the order of interaction with the field and the appropriate relaxation process is achieved automatically,

and thus these methods may become a useful alternative to the perturbative formalism [50, 51]. Third, there is

a growing field of single molecule spectroscopy. If the optical response from individual molecules in a condensed phase environment is detected, then one has a more direct approach than echo spectroscopy for removing the effect of environmental inhomogeneity. Moreover, the spectral change of individual molecules can be followed in time, giving data that are masked in even the best echo spectrum.

A1.6.5 COHERENT CONTROL OF MOLECULAR DYNAMICS

Not only has there been great progress in making femtosecond pulses in recent years, but also progress has been made in the shaping of these pulses, that is, giving each component frequency any desired amplitude and phase. Given the great experimental progress in shaping and sequencing femtosecond pulses, the inexorable question is: How is it possible to take advantage of this wide possible range of coherent excitations to bring about selective and energetically efficient photochemical reactions? Many intuitive approaches to laser selective chemistry have been tried since 1980. Most of these approaches have focused on depositing energy in a sustained manner, using monochromatic radiation, into a particular state or mode of the molecule. Virtually all such schemes have failed, due to rapid intramolecular energy redistribution.

The design of pulse sequences to selectively control chemical bond breaking is naturally formulated as a problem in the calculus of variations [17, 52]. This is the mathematical apparatus for finding the best shape, subject to certain constraints. For example, the shape which encloses the maximum area for a given perimeter; the minimum distance between two points on a sphere subject to the constraint that the connecting path be on the sphere; the shape of a cable of fixed length and fixed endpoints which minimizes the potential energy; the trajectory of least time; the path of least action; all these are searches for the best shape, and are problems in the classical calculus of variations. In our case, we are searching for the best shape of laser pulse intensity against time. If we admit complex pulses this involves an optimization over the real and imaginary parts of the pulse shape. We may be interested in the optimal pulse subject to some constraints, for example for a fixed total energy in the pulse.

It turns out that there is another branch of mathematics, closely related to the calculus of variations, although historically the two fields grew up somewhat separately, known as optimal control theory (OCT). Although the boundary between these two fields is somewhat blurred, in practice one may view optimal control theory as the application of the calculus of variations to problems with differential equation constraints. OCT is used in chemical, electrical, and aeronautical engineering; where the differential equation constraints may be chemical kinetic equations, electrical circuit equations, the Navier–Stokes equations for air flow, or Newton’s equations. In our case, the differential equation constraint is the TDSE in the presence of the control, which is the electric field interacting with the dipole (permanent or transition dipole moment) of the molecule [53, 54, 55 and 56]. From the point of view of control theory, this application presents many new features relative to conventional applications; perhaps most interesting mathematically is the admission of a complex state variable and a complex control; conceptually, the application of control techniques to steer the *microscopic* equations of motion is both a novel and potentially very important new direction.

A very exciting approach adopted more recently involves letting the laser learn to design its own optimal pulse shape in the laboratory [59, 60, 61, 62 and 63]. This is achieved by having a feedback loop, such that the increase or decrease in yield from making a change in the pulse is fed back to the pulse shaper, guiding the design of the next trial pulse. A particular implementation of this approach is the ‘genetic algorithm’, in which large set of initial pulses are generated; those giving the highest yield are used as ‘parents’ to produce a new ‘generation’ of pulses, by allowing segments of the parent pulses to combine in random new combinations.

The various approaches to laser control of chemical reactions have been discussed in detail in several recent reviews [64, 65].

A1.6.5.1 INTUITIVE CONTROL CONCEPTS

Consider the ground electronic state potential energy surface in figure A1.6.26. This potential energy surface, corresponding to collinear ABC, has a region of stable ABC and two exit channels, one corresponding to A + BC and one to AB + C. This system is the simplest paradigm for control of chemical product formation: a two degree of freedom system is the minimum that can display two distinct chemical products. The objective is, starting out in a well-defined initial state ($v = 0$ for the ABC molecule) to design an electric field as a function of time which will steer the wavepacket out of channel 1, with no amplitude going out of channel 2, and *vice versa* [19, 52].

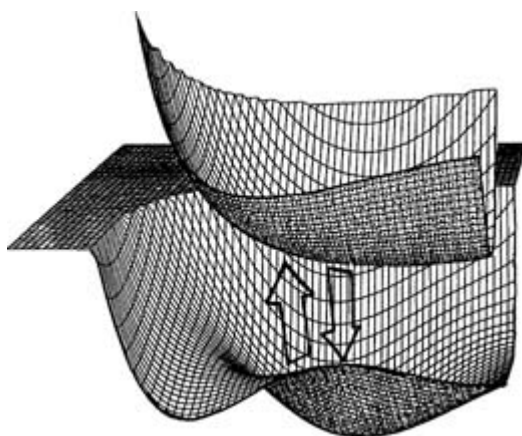


Figure A1.6.26. Stereoscopic view of ground- and excited-state potential energy surfaces for a model collinear ABC system with the masses of HHD. The ground-state surface has a minimum, corresponding to the stable ABC molecule. This minimum is separated by saddle points from two distinct exit channels, one leading to AB + C the other to A + BC. The object is to use optical excitation and stimulated emission between the two surfaces to ‘steer’ the wavepacket selectively out of one of the exit channels (reprinted from [54]).

-57-

We introduce a single excited electronic state surface at this point. The motivation is severalfold. (i) Transition dipole moments are generally much stronger than permanent dipole moments. (ii) The difference in functional form of the excited and ground potential energy surface will be our dynamical kernel; with a single surface one must make use of the (generally weak) coordinate dependence of the dipole. Moreover, the use of excited electronic states facilitates large changes in force on the molecule, effectively instantaneously, without necessarily using strong fields. (iii) The technology for amplitude and phase control of optical pulses is significantly ahead of the corresponding technology in the infrared.

The object now will be to steer the wavefunction out of a specific exit channel on the ground electronic state, using the excited electronic state as an intermediate. Insofar as the control is achieved by transferring amplitude between two electronic states, all the concepts regarding the central quantity μ_{eg} introduced above will now come into play.

(A) PUMP-DUMP SCHEME

Consider the following intuitive scheme, in which the timing between a pair of pulses is used to control the identity of products [52]. The scheme is based on the close correspondence between the centre of a wavepacket in time and that of a classical trajectory (Ehrenfest's theorem). The first pulse produces an excited electronic state wavepacket. The time delay between the pulses controls the time that the wavepacket evolves on the excited electronic state. The second pulse stimulates emission. By the Franck–Condon principle, the second step prepares a wavepacket on the ground electronic state with the same position and momentum, instantaneously, as the excited-state wavepacket. By controlling the position and momentum of the wavepacket produced on the ground state through the second step, one can gain some measure of control over product formation on the ground state. This ‘pump–dump’ scheme is illustrated classically in [figure A1.6.27](#). The trajectory originates at the ground-state surface minimum (the equilibrium geometry). At $t = 0$ it is promoted to the excited-state potential surface (a two-dimensional harmonic oscillator in this model) where it originates at the Condon point, that is, vertically above the ground-state minimum. Since this position is displaced from equilibrium on the excited state, the trajectory begins to evolve, executing a two-dimensional Lissajous motion. After some time delay, the trajectory is brought down vertically to the ground state (keeping both the instantaneous position and momentum it had on the excited state) and allowed to continue to evolve on the ground-state. [figure A1.6.27](#) shows that for one choice of time delay it will exit into channel 1, for a second choice of time delay it will exit into channel 2. Note how the position and momentum of the trajectory on the ground state, immediately after it comes down from the excited state, are both consistent with the values it had when it left the excited state, and at the same time are ideally suited for exiting out their respective channels.

-58-

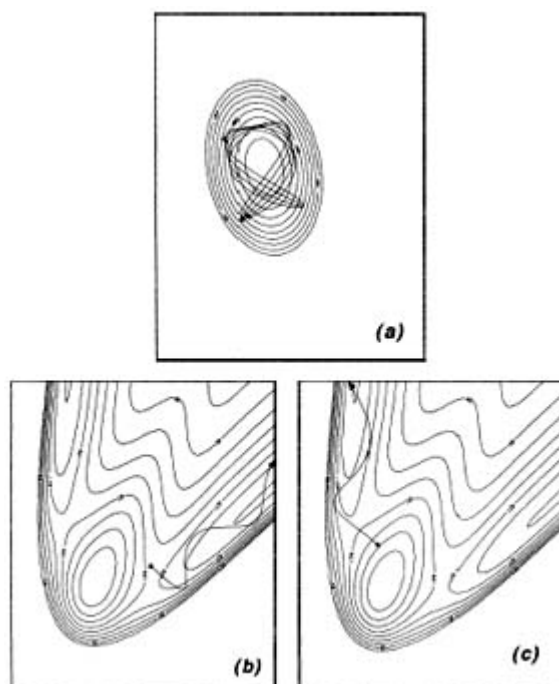


Figure A1.6.27. Equipotential contour plots of (a) the excited- and (b), (c) ground-state potential energy surfaces. (Here a harmonic excited state is used because that is the way the first calculations were performed.) (a) The classical trajectory that originates from rest on the ground-state surface makes a vertical transition to the excited state, and subsequently undergoes Lissajous motion, which is shown superimposed. (b) Assuming a vertical transition down at time t_1 (position and momentum conserved) the trajectory continues to evolve on the ground-state surface and exits from channel 1. (c) If the transition down is at time t_2 the classical trajectory exits from channel 2 (reprinted from [52]).

A full quantum mechanical calculation based on these classical ideas is shown in [figure A1.6.28](#) and [figure A1.6.29](#) [19]. The dynamics of the two-electronic-state model was solved, starting in the lowest vibrational eigenstate of the ground electronic state, in the presence of a pair of femtosecond pulses that couple the states. Because the pulses were taken to be much shorter than a vibrational period, the effect of the pulses is to prepare a wavepacket on the excited/ground state which is almost an exact replica of the instantaneous wavefunction on the other surface. Thus, the first pulse prepares an initial wavepacket which is almost a perfect Gaussian, and which begins to evolve on the excited-state surface. The second pulse transfers the instantaneous wavepacket at the arrival time of the pulse back to the ground state, where it continues to evolve on the ground-state surface, given its position and momentum at the time of arrival from the excited state. For one choice of time delay the exit out of channel 1 is almost completely selective ([figure A1.6.28](#)), while for a second choice of time delay the exit out of channel 2 is almost completely selective ([A1.6.29](#)). Note the close correspondence with the classical model: the wavepacket on the excited state is executing a Lissajous motion almost identical with that of the classical trajectory (the wavepacket is a nearly Gaussian wavepacket on a two-dimensional harmonic oscillator). On the groundstate, the wavepacket becomes spatially extended but its exit channel, as well as the partitioning of energy into translation and vibration (i.e. parallel and perpendicular to the exit direction) are seen to be in close agreement with the corresponding classical trajectory.

-59-

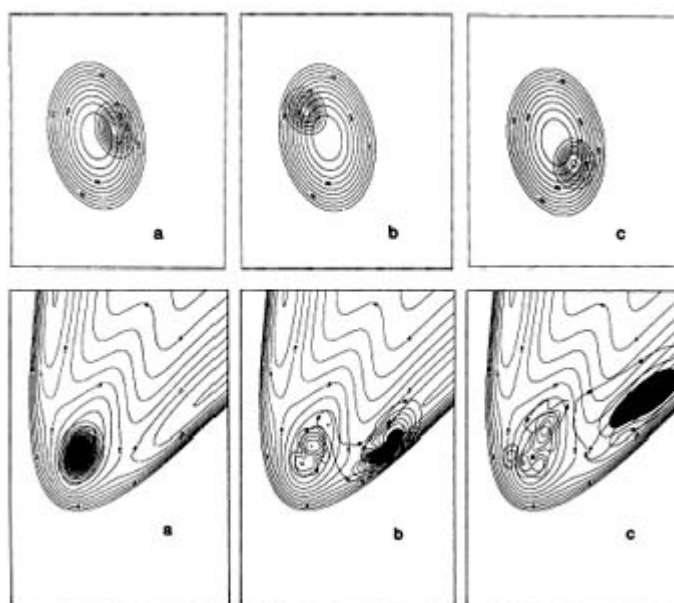


Figure A1.6.28. Magnitude of the excited-state wavefunction for a pulse sequence of two Gaussians with time delay of 610 a.u. = 15 fs. (a) $t = 200$ a.u., (b) $t = 400$ a.u., (c) $t = 600$ a.u. Note the close correspondence with the results obtained for the classical trajectory (figure A1.6.27(a) and (b)). Magnitude of the ground-state wavefunction for the same pulse sequence, at (a) $t = 0$, (b) $t = 800$ a.u., (c) $t = 1000$ a.u. Note the close correspondence with the classical trajectory of figure A1.6.27(c)). Although some of the amplitude remains in the bound region, that which does exit does so exclusively from channel 1 (reprinted from [52]).

-60-

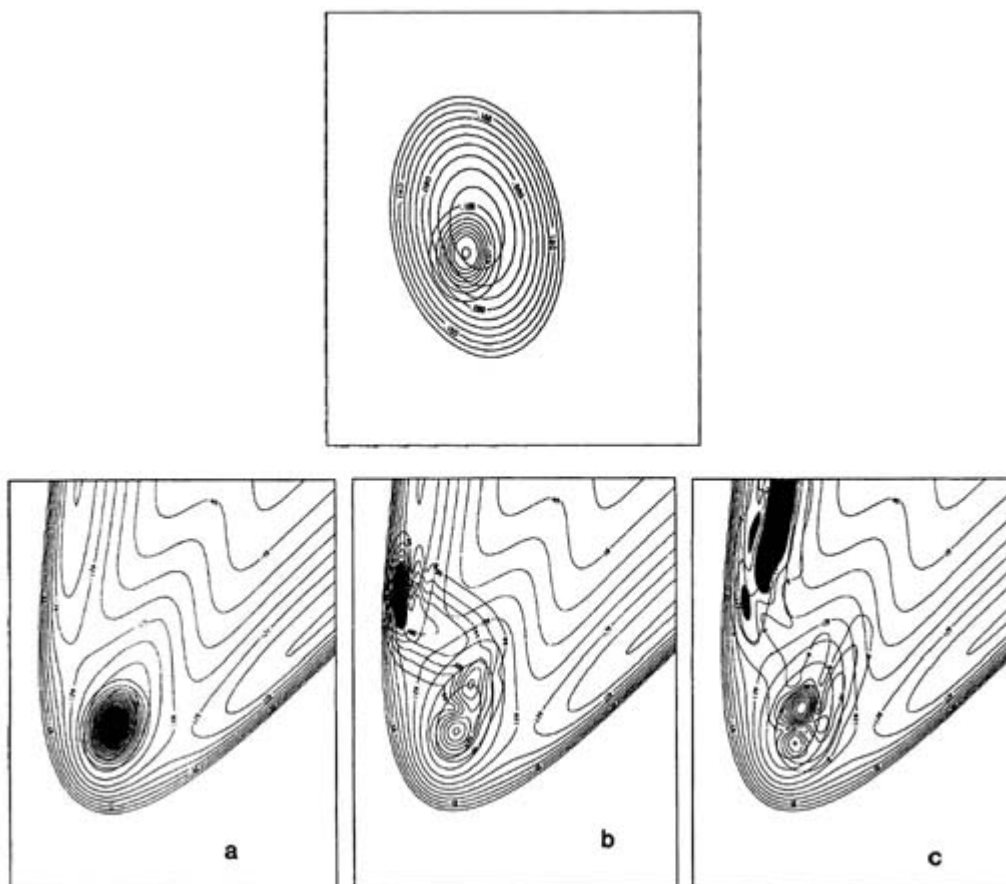


Figure A1.6.29. Magnitude of the ground- and excited-state wavefunctions for a sequence of two Gaussian pulses with time delay of 810 a.u. (upper diagram) excited-state wavefunction at 800 a.u., before the second pulse. (a) Ground-state wavefunction at 0 a.u. (b) Ground-state wavefunction at 1000 a.u. (c) Ground-state wavefunction at 1200 a.u. That amplitude which does exit does so exclusively from channel 2. Note the close correspondence with the classical trajectory of figure A1.6.27(c) (reprinted from [52]).

This scheme is significant for three reasons: (i) it shows that control is possible, (ii) it gives a starting point for the design of optimal pulse shapes, and (iii) it gives a framework for interpreting the action of two pulse and more complicated pulse sequences. Nevertheless, the approach is limited: in general with the best choice of time delay and central frequency of the pulses one may achieve only partial selectivity. Perhaps most importantly, this scheme does not exploit the phase of the light. Intuition breaks down for more complicated processes and classical pictures cannot adequately describe the role of the phase of the light and the wavefunction. Hence, attempts were made to develop a systematic procedure for improving an initial pulse sequence.

Before turning to these more systematic procedures for designing shaped pulses, we point out an interesting alternative perspective on pump–dump control. A central tenet of Feynman’s approach to quantum mechanics was to think of quantum interference as arising from multiple dynamical paths that lead to the same final state. The simple example of this interference involves an initial state, two intermediate states and a single final state, although if the objective is to control some branching ratio at the final energy then at least two final states are necessary. By controlling the phase with which each of the two intermediate states contributes to the final state, one may control constructive versus destructive interference in the final states. This is the basis of the Brumer–Shapiro approach to coherent control [57, 58]. It is interesting to note that pump–dump control

can be viewed entirely from this perspective. Now, however, instead of two intermediate states there are many, corresponding to the vibrational levels of the excited electronic state (see [figure A1.6.31](#)). The control of the phase which determines how each of these intermediate levels contributes to the final state is achieved via the time delay between the excitation and the stimulated emission pulse. This ‘interfering pathways’ interpretation of pump–dump control is shown in [figure A1.6.30](#).

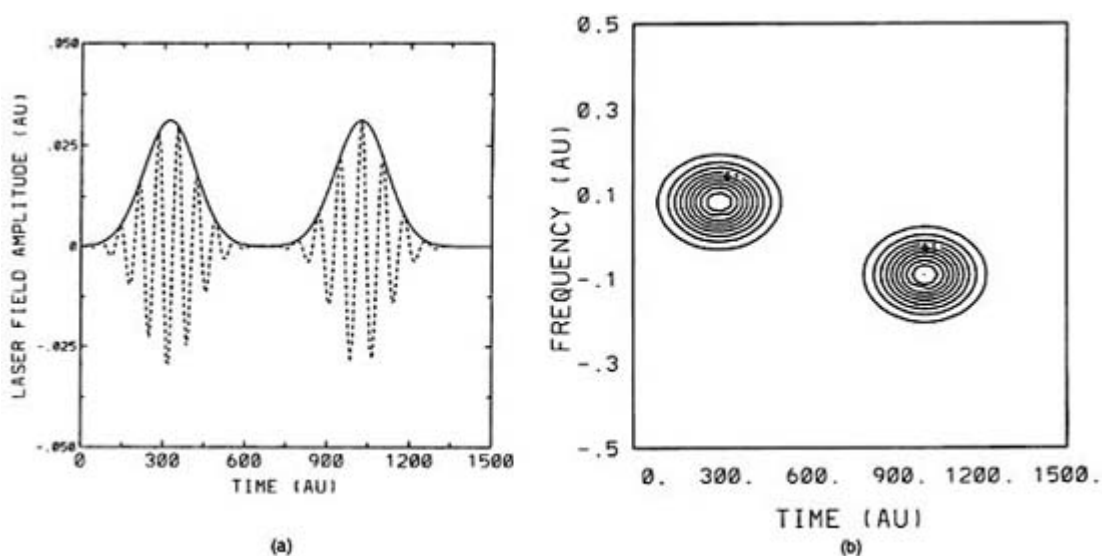


Figure A1.6.30. (a) Two pulse sequence used in the Tannor–Rice pump–dump scheme. (b) The Husimi time–frequency distribution corresponding to the two pump sequence in (a), constructed by taking the overlap of the pulse sequence with a two-parameter family of Gaussians, characterized by different centres in time and carrier frequency, and plotting the overlap as a function of these two parameters. Note that the Husimi distribution allows one to visualize both the time delay and the frequency offset of pump and dump simultaneously (after [52a]).

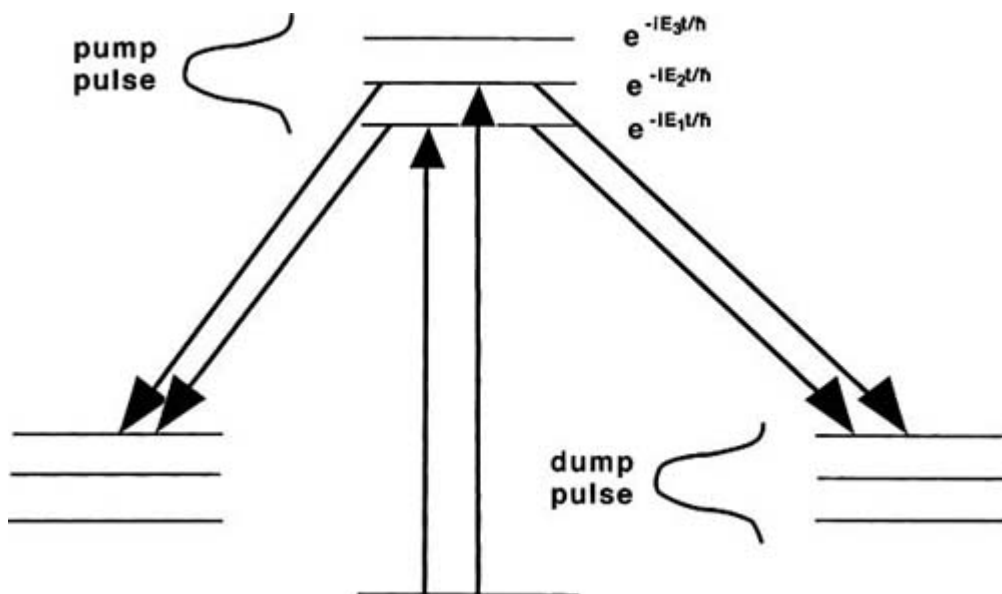


Figure A1.6.31. Multiple pathway interference interpretation of pump–dump control. Since each of the pair of pulses contains many frequency components, there are an infinite number of combination frequencies which lead to the same final energy state, which generally interfere. The time delay between the pump and

dump pulses controls the relative phase among these pathways, and hence determines whether the interference is constructive or destructive. The frequency domain interpretation highlights two important features of coherent control. First, if final products are to be controlled there must be degeneracy in the dissociative continuum. Second, a single interaction with the light, no matter how it is shaped, cannot produce control of final products: at least two interactions with the field are needed to obtain interfering pathways.

A1.6.5.2 VARIATIONAL FORMULATION OF CONTROL OF PRODUCT FORMATION

The next step, therefore, is to address the question: how is it possible to take advantage of the many additional available parameters: pulse shaping, multiple pulse sequences, etc—in general an $E(t)$ with arbitrary complexity—to maximize and perhaps obtain perfect selectivity? Posing the problem mathematically, one seeks to maximize

$$J \equiv \lim_{T \rightarrow \infty} \langle \psi(T) | P_\alpha | \psi(T) \rangle \quad (\text{A 1.6.127})$$

where P_α is a projection operator for chemical channel α (here, α takes on two values, referring to arrangement channels $A + BC$ and $AB + C$; in general, in a triatomic molecule ABC , α takes on three values, 1,2,3, referring to arrangement channels $A + BC$, $AB + C$ and $AC + B$). The time T is understood to be longer than the duration of the pulse sequence, $E(t)$; the yield, J , is defined as $T \rightarrow \infty$, that is, after the wavepacket amplitude has time to reach its asymptotic arrangement. The key observation is that the quantity J is a *functional* of $E(t)$, that is, J is a function of a function, because $\psi(T)$ depends on the whole history of $E(t)$. To make this dependence on $E(T)$ explicit we may write

-63-

$$J[E(t)] \equiv \lim_{T \rightarrow \infty} \langle \psi[E(t)](T) | P_\alpha | \psi[E(t)](T) \rangle \quad (\text{A 1.6.128})$$

where square brackets are used to indicate functional dependence. The problem of maximizing a function of a function has a rich history in mathematical physics, and falls into the class of problems belonging to the calculus of variations.

In the OCT formulation, the TDSE written as a 2×2 matrix in a BO basis set, [equation \(A1.6.72\)](#), is introduced into the objective functional with a Lagrange multiplier, $\chi(x, t)$ [54]. The modified objective functional may now be written as

$$\bar{J} \equiv \lim_{T \rightarrow \infty} \langle \psi(T) | P_\alpha | \psi(T) \rangle + 2\text{Re} \int_0^T dt \left(\chi(t) \left| \frac{\partial}{\partial t} - \frac{H}{i\hbar} \right| \psi(t) \right) - \lambda \int_0^T dt |E(t)|^2 \quad (\text{A 1.6.129})$$

where a constraint (or penalty) on the time integral of the energy in the electric field has also been added. It is clear that as long as ψ satisfies the TDSE the new term in \bar{J} will vanish for any $\chi(x, t)$. The function of the new term is to make the variations of \bar{J} with respect to E and with respect to ψ independent, to first-order in δE (i.e. to ‘deconstrain’ ψ and E).

The requirement that $\delta \bar{J} / \delta \psi = 0$ leads to the following equations:

$$i\hbar \frac{\partial \chi}{\partial t} = H \chi \quad (\text{A 1.6.130})$$

$$\chi(x, T) = P_\alpha \psi(x, T) \quad (\text{A 1.6.131})$$

that is, the Lagrange multiplier must obey the TDSE, subject to the boundary condition at the *final* time T that χ be equal to the projection operator operating on the Schrödinger wavefunction. These conditions ‘conspire’, so that a change in E , which would ordinarily change \bar{J} through the dependence of $\psi(T)$ on E , does not do so to first-order in the field. For a physically meaningful solution it is required that

$$i\hbar \frac{\partial \psi}{\partial t} = H \psi \quad (\text{A 1.6.132})$$

$$\psi(x, 0) = \psi_0(x). \quad (\text{A 1.6.133})$$

Finally, the optimal $E(t)$ is given by the condition that $\delta \bar{J} / \delta E = 0$ which leads to the equation

$$E(t) = \frac{-i}{\hbar \lambda} [\langle \chi_a | \mu | \psi_b \rangle - \langle \psi_a | \mu | \chi_b \rangle]. \quad (\text{A 1.6.134})$$

The interested reader is referred to [54] for the details of the derivation.

Equation (A1.6.129), equation (A1.6.130), equation (A1.6.131), equation (A1.6.132) and equation (A1.6.133) form the basis for a double-ended boundary value problem. ψ is known at $t = 0$, while χ is known at $t = T$. Taking a guess for $E(t)$ one can propagate ψ forward in time to obtain $\psi(t)$; at time T the projection operator P_α may be applied to obtain $\chi(t)$, which may be propagated backwards in time to obtain $\chi(t)$. Note, however, that the above description is not self-consistent: the guess of $E(t)$ used to propagate $\psi(t)$ forward in time and to propagate $\chi(t)$ backwards in time is not, in general, equal to the value of $E(t)$ given by equation (A1.6.133). Thus, in general, one has to solve these equations iteratively until self-consistency is achieved. Optimal control theory has become a widely used tool for designing laser pulses with specific objectives. The interested reader can consult the review in [65] for further examples.

A1.6.5.3 OPTIMAL CONTROL AND LASER COOLING OF MOLECULES

The use of lasers to cool atomic translational motion has been one of the most exciting developments in atomic physics in the last 15 years. For excellent reviews, see [66, 67]. Here we give a non-orthodox presentation, based on [68].

(A) CALIBRATION OF COOLING: THE ZEROth LAW

Consider, figure A1.6.32 in which a system is initially populated with an incoherent distribution of populations with Boltzmann probabilities, P_n , $\sum_n P_n = 1$. The simple-minded definition of cooling is to manipulate all the population into the lowest energy quantum state, i.e. to make $P_0 = 1$ and all the other $P_n = 0$. Cooling can then be measured by the quantity $\sum_n P_n^2$: for the initial, incoherent distribution $\sum_n P_n^2 < 1$ while for the final distribution $\sum_n P_n^2 = 1$. However, adoption of this definition of cooling implies that if all the population is put into *any* single quantum state, not necessarily the lowest energy state, the degree of cooling is identical. Although this seems surprising at first, it is in fact quite an appealing definition of cooling. It highlights the fact that the essence of cooling is the creation of a pure state starting from a mixed state; once the state is pure then coherent manipulations, which are relatively straightforward, can transfer this

population to the ground state. As described in [section A1.6.2.4](#), the conventional measure of the degree of purity of a system in quantum mechanics is $\text{Tr}(\rho^2)$, where ρ is the system's density matrix, and thus we have here defined cooling as the process of bringing $\text{Tr}(\rho^2)$ from its initial value less than 1 to unity. The definition of cooling in terms of $\text{Tr}(\rho^2)$ leads to an additional surprise, namely, that the single quantum state need not even be an eigenstate: it can, in principle, be a superposition consisting of a coherent superposition of many eigenstates. So long as the state is pure (i.e. can be described by a single Schrödinger wavefunction) it can be manipulated into the lowest energy state by a unitary transformation, and in a very real sense is already cold! [figure A1.6.32](#) gives a geometrical interpretation of cooling. The density matrix is represented as a point on a generalized Bloch sphere of radius $R = \text{Tr}(\rho^2)$. For an initially thermal state the radius $R < 1$, while for a pure state $R = 1$. Thus, the object of cooling, that is, increasing the purity of the density matrix, corresponds to manipulating the density matrix onto spheres of increasingly larger radius.

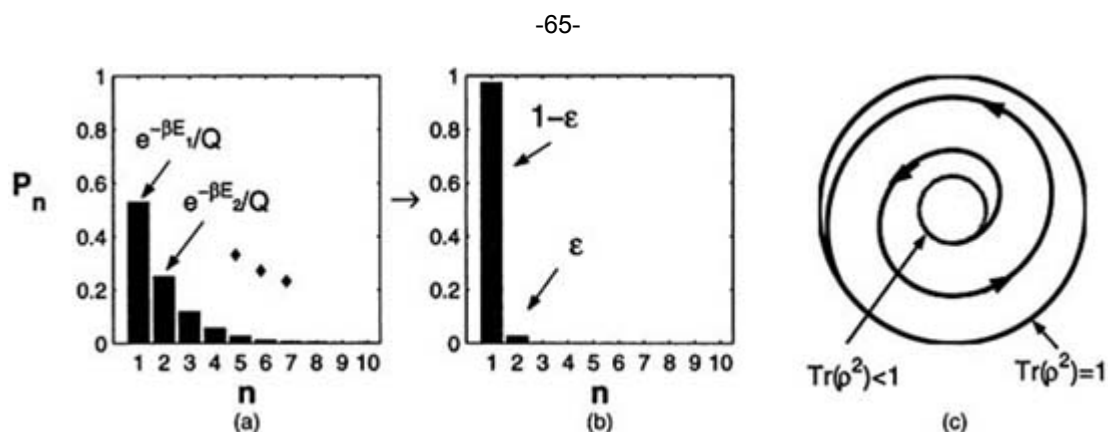


Figure A1.6.32. (a) Initial and (b) final population distributions corresponding to cooling. (c) Geometrical interpretation of cooling. The density matrix is represented as a point on generalized Bloch sphere of radius $R = \text{Tr}(\rho^2)$. For an initially thermal state the radius $R < 1$, while for a pure state $R = 1$. The object of cooling is to manipulate the density matrix onto spheres of increasingly larger radius.

We have seen in [section A1.6.2.4](#) that external fields alone cannot change the value of $\text{Tr}(\rho^2)$! Changes in the purity can arise only from the spontaneous emission, which is inherently uncontrollable. Where then is the control?

A first glimmer of the resolution to the paradox of how control fields can control purity content is obtained by noting that the second derivative, $\text{Tr}(\dot{\rho}^2)$, does depend on the external field. Loosely speaking, the independence of the first derivative and the dependence of the second derivative on the control field indicates that the control of cooling is achieved only in two-stages: preparation of the initial state by the control field, followed by spontaneous emission into that recipient state. This two-stage interpretation will now be quantified.

To find the boundary between heating and cooling we set $\text{Tr}(\dot{\rho}^2) = 0$. [Figure A1.6.33](#) shows isocontours of $\text{Tr}(\dot{\rho}^2)$ as a function of the parameters $\rho_{22}(z)$ and $|\rho_{12}|(x)$. The dark region corresponds to $\frac{d}{dt}\text{Tr}(\rho^2) < 0$; that is, cooling, while the light region corresponds to $\frac{d}{dt}\text{Tr}(\rho^2) > 0$, (i.e. heating). Note that the cooling region fills part, but not all of the lower hemisphere. For fixed z , the maximum occurs along the line $x = 0$, with the global maximum at $z = 1/4, x = 0$.

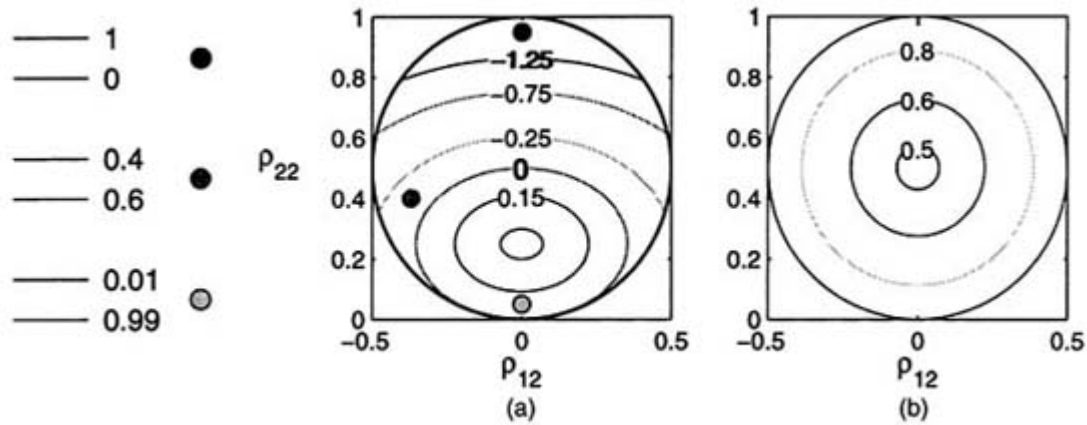


Figure A1.6.33. (a) Contour map of $\frac{d}{dt}\text{Tr}(\rho^2)$ as a function of the parameters $\rho_{22}(z)$ and $|\rho_{12}|(x)$. The dark region corresponds to $\frac{d}{dt}\text{Tr}(\rho^2) < 0$, i.e. cooling while the light region corresponds to $\frac{d}{dt}\text{Tr}(\rho^2) > 0$, i.e. heating. For fixed z , the maximum occurs along the line $x = 0$. (b) Isopurity, or isocoherence contours (contours of fixed $\text{Tr}(\rho^2)$) as a function of $\rho_{22}(z)$ and $|\rho_{12}|(x)$ for the two-level system. The contour takes its maximum value of 1, corresponding to a pure state, along the outermost circle, while the function takes its minimum value of 1/2, representing the most impure state, at the centre.

To gain a qualitative understanding for the heating and cooling regions we consider three representative points (top to bottom, figure A1.6.33(a)). (i) Spontaneous emission will lead from 1:99 to 0:100 and hence purity increase. (ii) Spontaneous emission will lead from 100:0 to 99:1 and hence purity decrease. (iii) Spontaneous emission will lead from 40:60 to 30:70 which suggests a purity increase; however, if there is purity stored in the coherences ρ_{12} , spontaneous emission will force these to decay at a rate $\Gamma_2 = 1/2\Gamma_1$; this leads to a decrease in purity which is greater than the increase in purity brought about by the population transfer.

The manipulations allowed by the external field are those that move the system along a contour of constant value of $\text{Tr}(\rho^2)$, an isocoherence contour; it is clear from figure A1.6.33 that the location on this contour has a profound affect on $\text{Tr}(\dot{\rho}^2)$. This gives a second perspective on how the external field cannot directly change $\text{Tr}(\rho^2)$, but can still affect the rate of change of $\text{Tr}(\rho^2)$. If we imagine that at every instant in time the external field moves the system along the instantaneous isocoherence contour until it intersects the curve of maximum $\text{Tr}(\dot{\rho}^2)$, that would provide an optimal cooling strategy. This last observation is the crux of our cooling theory and puts into sharp perspective the role played by the external field: while the external field cannot itself change the purity of the system it can perform purity-preserving transformations which subsequently affect the rate of change of purity.

To summarize, we have the following chain of dependence: $(\rho_{22}, |\rho_{12}|) \rightarrow \text{Tr}(\rho^2) \rightarrow (\bar{\rho}_{22}, |\bar{\rho}_{12}|) \rightarrow \text{Tr}(\dot{\rho}^2)$. This chain of dependence gives $\text{Tr}(\dot{\rho}^2)$ as a function of $\text{Tr}(\rho^2)$, which is a *differential equation for the optimal trajectory* $\text{Tr}(\rho^2)(t)$. By studying the rate of approach of the optimal trajectory to absolute zero (i.e. to a pure state) we will have found an inviolable limitation on cooling rate with the status of a third law of thermodynamics.

Note that the differential equation obtained from this approach will never agree perfectly with the results of a simulation. The above formulation is essentially an adiabatic formulation of the process: the spontaneous emission is considered to be slow compared with the time scale for the purity-preserving transformations generated by the external field, which is what allows us to assume in the theory that the external field

manipulation along the isocoherence contour is instantaneous. If the external field is sufficiently intense, the population transfer may become nearly instantaneous relative to the spontaneous emission, and the adiabatic approximation will be excellent.

(B) COOLING AND LASING AS COMPLEMENTARY PROCESSES

It is interesting to consider the regions of heating, that is, regions where $\text{Tr}(\dot{\rho}^2) < 0$. We conjecture that these regions correspond to regions where lasing can occur. The conjecture is based on the following considerations:

- (i) Note that for the two-level system with no coherence ($\rho_{12} = 0$), the region where $\text{Tr}(\dot{\rho}^2) < 0$ corresponds to $\rho_{22} > \frac{1}{2}$. This corresponds to the conventional population inversion criterion for lasing: that population in the excited state be larger than in the ground state.
- (ii) The fact that in this region the system coherence is decreasing, leaves open the possibility that coherence elsewhere can increase. In particular, excitation with incoherent light can lead to emission of coherent light. This is precisely the reverse situation as with laser cooling, where coherent light is transformed to incoherent light (spontaneous emission), increasing the level of coherence of the system.
- (iii) The regions with $\text{Tr}(\dot{\rho}^2) < 0$ and $d < \frac{1}{2}$ necessarily imply $\gamma > 0$, that is, coherences between the ground and excited state. This may correspond to lasing without population inversion, an effect which has attracted a great deal of attention in recent years, and is made possible by coherences between the ground and excited states. Indeed, in the three-level λ system the boundary between heating and cooling is in exact agreement with the boundary between lasing and non-lasing.

Fundamentally, the conditions for lasing are determined unambiguously once the populations and coherences of the system density matrix are known. Yet, we have been unable to find in the literature any simple criterion for lasing in multilevel systems in terms of the system density matrix alone. Our conjecture is that entropy, as expressed by the purity content $\text{Tr}(\rho^2)$, is the unifying condition; the fact that such a simple criterion could have escaped previous observation may be understood, given the absence of thermodynamic considerations in conventional descriptions of lasing.

REFERENCES

- [1] Jackson J D 1975 *Classical Electrodynamics* (New York: Wiley)
- [2] Loudon R 1983 *The Quantum Theory of Light* (Oxford: Oxford University Press)
- [3] Schatz G C and Ratner M A 1993 *Quantum Mechanics in Chemistry* (Englewood Cliffs, NJ: Prentice-Hall) ch 5

- [4] We follow [3] here. However, see van Kranendonk J and Sipe J E 1976 *Can. J. Phys.* **54** 471
- [5] Einstein A 1917 On the quantum theory of radiation *Phys. Z.* **18** 121 Reprinted ter Haar D 1967 *The Old Quantum Theory* (New York: Pergamon)

- [6] Cohen-Tannoudji C, Diu B and Laloë F 1977 *Quantum Mechanics* vol 1 (New York: Wiley) ch 4
- [7] Allen L and Eberly J H 1987 *Optical Resonance and Two-Level Atoms* (New York: Dover)
- [8] Steinfeld J I 1986 *Molecules and Radiation* (Cambridge, MA: MIT)
- [9] Siegman A E 1986 *Lasers* (Mill Valley, CA: University Science Books)
- [10] Sargent III M, Scully M O and Lamb W E Jr 1974 *Laser Physics* (Reading, MA: Addison-Wesley)
- [11] Cohen-Tannoudji C, Dupont-Roc J and Grynberg G 1992 *Atom-Photon Interaction* (New York: Wiley)
- [12] Heller E J 1978 Quantum corrections to classical photodissociation models *J. Chem. Phys.* **68** 2066
- [13] Kulander K C and Heller E J 1978 Time-dependent formulation of polyatomic photofragmentation: application to H_3^+ *J. Chem. Phys.* **69** 2439
- [14] Lee S-Y and Heller E J 1979 Time-dependent theory of Raman scattering *J. Chem. Phys.* **71** 4777
- [15] Heller E J, Sundberg R L and Tannor D J 1982 Simple aspects of Raman scattering *J. Phys. Chem.* **86** 1822–33
- [16] Heller E J 1981 The semiclassical way to molecular spectroscopy *Acc. Chem. Res.* **14** 368
- [17] Tannor D J and Rice S A 1988 Coherent pulse sequence control of product formation in chemical reactions *Adv. Chem. Phys.* **70** 441–524
- [18] Lee S-Y and Heller E J 1979 *op. cit.* [14] equations (2.9) and (2.10)
- [19] Tannor D J, Kosloff R and Rice S A 1986 Coherent pulse sequence induced control of selectivity of reactions: exact quantum mechanical calculations *J. Chem. Phys.* **85** 5805–20, equations (1)–(6)
- [20] Tannor D J and Rice S A 1988 Coherent pulse sequence control of product formation in chemical reactions *Adv. Chem. Phys.* **70** 441–524, equations (5.1)–(5.6)
- [21] Tannor D J 2001 *Introduction to Quantum Mechanics: A Time Dependent Perspective* (Mill Valley, CA: University Science Books)
- [22] Scherer N F, Carlson R J, Matro A, Du M, Ruggiero A J, Romero-Rochin V, Cina J A, Fleming G R and Rice S A 1991 Fluorescence-detected wave packet interferometry: time resolved molecular spectroscopy with sequences of femtosecond phase-locked pulses *J. Chem. Phys.* **95** 1487
- [23] Scherer N F, Matro A, Ziegler L D, Du M, Cina J A and Fleming G R 1992 Fluorescence-detected wave packet interferometry. 2. Role of rotations and determination of the susceptibility *J. Chem. Phys.* **96** 4180
- [24] Engel V and Metiu H 1994 2-Photon wave-packet interferometry *J. Chem. Phys.* **100** 5448
- [25] Zewail A H 1988 Laser femtochemistry *Science* **242** 1645

- [26] Zewail A H (ed) 1994 *Femtochemistry* vols 1 and 2 (Singapore: World Scientific)
- [27] Manz J and Wöste L (eds) 1995 *Femtosecond Chemistry* (Heidelberg: VCH)
- [28] Baumert T, Engel V, Meier Ch and Gerber G 1992 High laser field effects in multiphoton ionization of Na_2 – experiment and quantum calculations *Chem. Phys. Lett.* **200** 488
- [29] Wang Q, Schoenlein R W, Peteanu L A, Mathies R A and Shank C V 1994 Vibrationally coherent photochemistry in the femtosecond primary event of vision *Science* **266** 422
- [30] Pugliano N, Gnanakaran S and Hochstrasser R M 1996 The dynamics of photodissociation reactions in solution *J. Photochem. and Photobiol. A—Chemistry* **102** 21–8

- [31] Pack R T 1976 Simple theory of diffuse vibrational structure in continuous UV spectra of polyatomic molecules. I. Collinear photodissociation of symmetric triatomics *J. Chem. Phys.* **65** 4765
- [32] Heller E J 1978 Photofragmentation of symmetric triatomic molecules: Time dependent picture *J. Chem. Phys.* **68** 3891
- [33] Myers A B and Mathies R A 1987 Resonance Raman intensities: A probe of excited-state structure and dynamics *Biological Applications of Raman Spectroscopy* vol 2, ed T G Spiro (New York: Wiley-Interscience) pp 1–58
- [34] Albrecht A C 1961 On the theory of Raman intensities *J. Chem. Phys.* **34** 1476
- [35] Bloembergen N 1965 *Nonlinear Optics* (Reading, MA: Benjamin-Cummings)
- [36] Shen Y R 1984 *The Principles of Nonlinear Optics* (New York: Wiley)
- [37] Mukamel S 1995 *Principles of Non-linear Optical Spectroscopy* (New York: Oxford University Press)
- [38] Lee D and Albrecht A C 1985 *Advances in Infrared and Raman Spectroscopy* **12** 179
- [39] Tannor D J, Rice S A and Weber P M 1985 Picosecond CARS as a probe of ground electronic state intramolecular vibrational redistribution *J. Chem. Phys.* **83** 6158
- [40] Tannor D J and Rice S A 1987 Photon echoes in multilevel systems *Understanding Molecular Properties* ed J Avery *et al* (Dordrecht: Reidel) p 205
- [41] Pollard W T, Lee S-Y and Mathies R A 1990 Wavepacket theory of dynamic absorption spectra in femtosecond pump–probe experiments *J. Chem. Phys.* **92** 4012
- [42] Lee S-Y 1995 Wave-packet model of dynamic dispersed and integrated pump–probe signals in femtosecond transition state spectroscopy *Femtosecond Chemistry* ed J Manz and L Wöste (Heidelberg: VCH)
- [43] Meyer S and Engel V 2000 Femtosecond time-resolved CARS and DFWM spectroscopy on gas-phase I₂: a wave-packet description *J. Raman Spectrosc.* **31** 33
- [44] Knopp G, Pinkas I and Prior Y 2000 Two-dimensional time-delayed coherent anti-Stokes Raman spectroscopy and wavepacket dynamics of high ground-state vibrations *J. Raman Spectrosc.* **31** 51
- [45] Pausch R, Heid M, Chen T, Schwoerer H and Kiefer W 2000 Quantum control by stimulated Raman scattering *J. Raman Spectrosc.* **31** 7
- [46] Pastirk I, Brown E J, Grimberg B I, Lozovoy V V and Dantus M 1999 Sequences for controlling laser excitation with femtosecond three-pulse four-wave mixing *Faraday Discuss.* **113** 401

- [47] Shen Y R 1984 *The Principles of Nonlinear Optics* (New York: Wiley) ch 21 for a clear discussion of the connection between the perturbative and nonperturbative treatment of photon echoes
- [48] Joo T, Jia Y, Yu J-Y, Lang M J and Fleming G R 1996 Third-order nonlinear time domain probes of solvation dynamics *J. Chem. Phys.* **104** 6089
- [48a] Passino S A, Nagasawa Y, Joo T and Fleming G R 1997 Three pulse photon echoes *J. Phys. Chem. A* **101** 725
- [49] Gallagher Faeder S M and Jonas D 1999 Two-dimensional electronic correlation and relaxation spectra: theory and model calculations *J. Phys. Chem. A* **102** 10 489–505
- [50] Seidner L, Stock G and Domcke W 1995 Nonperturbative approach to femtosecond spectroscopy – general theory and application to multidimensional nonadiabatic photoisomerization processes *J. Chem. Phys.* **103** 4002
- [51] Ashkenazi G, Banin U, Bartana A, Ruhman S and Kosloff R 1997 Quantum description of the impulsive photodissociation dynamics of I₃⁻ in solution *Adv. Chem. Phys.* **100** 229
- [52] Tannor D J and Rice S A 1985 Control of selectivity of chemical reaction via control of wave packet evolution *J.*

- [52a] Tannor D J 1994 Design of femtosecond optical pulses to control photochemical products *Molecules in Laser Fields* ed A Bandrauk (New York: Dekker) p 403
- [53] Peirce A P, Dahleh M A and Rabitz H 1988 Optimal control of quantum mechanical systems – Existence, numerical approximations and applications *Phys. Rev. A* **37** 4950
- [54] Kosloff R, Rice S A, Gaspard P, Tersigni S and Tannor D J 1989 Wavepacket dancing: achieving chemical selectivity by shaping light pulses *Chem. Phys.* **139** 201–20
- [55] Warren W S, Rabitz H and Dahleh M 1993 Coherent control of quantum dynamics: the dream is alive *Science* **259** 1581
- [56] Yan Y J, Gillilan R E, Whitnell R M, Wilson K R and Mukamel S 1993 Optimal control of molecular dynamics – Liouville space theory *J. Chem. Phys.* **97** 2320
- [57] Shapiro M and Brumer P 1986 Laser control of product quantum state populations in unimolecular reactions *J. Chem. Phys.* **84** 4103
- [58] Shapiro M and Brumer P 1989 Coherent chemistry—Controlling chemical reactions with lasers *Acc. Chem. Res.* **22** 407
- [59] Judson R S and Rabitz H 1992 Teaching lasers to control molecules *Phys. Rev. Lett.* **68** 1500
- [60] Bardeen C J, Yakovlev V V, Wilson K R, Carpenter S D, Weber P M and Warren W S 1997 Feedback quantum control of molecular electronic population transfer *Chem. Phys. Lett.* **280** 151
- [61] Yelin D, Meshulach D and Silberberg Y 1997 Adaptive femtosecond pulse compression *Opt. Lett.* **22** 1793–5
- [62] Assion A, Baumert T, Bergt M, Brixner T, Kiefer B, Seyfried V, Strehle M and Gerber G 1998 Control of chemical reactions by feedback-optimized phase-shaped femtosecond laser pulses *Science* **282** 919
- [63] Weinacht T C, White J L and Bucksbaum P H 1999 Toward strong field mode-selective chemistry *J. Phys. Chem. A* **103** 10 166–8
- [64] Gordon R J and Rice S A 1997 Active control of the dynamics of atoms and molecules *Annu. Rev. Phys. Chem.* **48** 601
-

-71-

- [65] Rice S A and Zhao M 2000 *Optical Control of Molecular Dynamics* (New York: Wiley)
- [66] Cohen-Tannoudji C N and Phillips W D 1990 New mechanisms for laser cooling *Phys. Today* **43** 33–40
- [67] Cohen-Tannoudji C 1991 Atomic motion in laser light *Fundamental Systems in Quantum Optics* ed J Dalibard *et al* (Oxford: Elsevier)
- [68] Tannor D J and Bartana A 1999 On the interplay of control fields and spontaneous emission in laser cooling *J. Phys. Chem. A* **103** 10 359–63
-

FURTHER READING

Loudon R 1983 *The Quantum Theory of Light* (Oxford: Oxford University Press)

An excellent and readable discussion of all aspects of the interaction of light with matter, from blackbody radiation to lasers and nonlinear optics.

Herzberg G 1939, 1945, 1966 *Molecular Spectra and Molecular Structure* (New York: van Nostrand) 3 vols

This is the classic work on molecular rotational, vibrational and electronic spectroscopy. It provides a comprehensive coverage of all aspects of infrared and optical spectroscopy of molecules from the traditional viewpoint and, both for perspective and scope, is an invaluable supplement to this section.

Steinfeld J I 1986 *Molecules and Radiation* (Cambridge, MA: MIT)

A good introduction to the use of coherent optical techniques and their use to probe molecular spectra.

Shen Y R 1984 *The Principles of Non-linear Optics* (New York: Wiley)

A clear, comprehensive discussion of the many facets of nonlinear optics. The emphasis is on optical effects, such as harmonic generation. The treatment of nonlinear spectroscopy, although occupying only a fraction of the book, is clear and physically well-motivated.

Mukamel S 1995 *Principles of Non-linear Optical Spectroscopy* (New York: Oxford University Press)

A valuable handbook describing the many uses of nonlinear optics for spectroscopy. The focus of the book is a unified treatment of $P^{(3)}$, and methods for modelling the $P^{(3)}$ signal.

Heller E J 1981 The semiclassical way to molecular spectroscopy *Acc. Chem. Res.* **14** 368

A beautiful, easy-to-read introduction to wavepackets and their use in interpreting molecular absorption and resonance Raman spectra.

-72-

Rice S A and Zhao M 2000 *Optical Control of Molecular Dynamics* (New York: Wiley)

A valuable resource, reviewing both theoretical and experimental progress on coherent control to date.

Tannor D J 2001 *Introduction to Quantum Mechanics: A Time Dependent Perspective* (Mill Valley, CA: University Science Books)

A comprehensive discussion of wavepackets, classical-quantum correspondence, optical spectroscopy, coherent control and reactive scattering from a unified, time dependent perspective.

-1-

A1.7 Surfaces and interfaces

J A Yarmoff

A1.7.1 INTRODUCTION

Some of the most interesting and important chemical and physical interactions occur when dissimilar materials meet, i.e. at an interface. The understanding of the physics and chemistry at interfaces is one of the most challenging and important endeavors in modern science.

Perhaps the most intensely studied interface is that between a solid and vacuum, i.e. a surface. There are a number of reasons for this. For one, it is more experimentally accessible than other interfaces. In addition, it is

conceptually simple, as compared to interfaces between two solids or between a solid and a liquid, so that the vacuum–solid interface is more accessible to fundamental theoretical investigation. Finally, it is the interface most easily accessible for modification, for example by photons or charged particle beams that must be propagated in vacuum.

Studies of surfaces and surface properties can be traced to the early 1800s [1]. Processes that involved surfaces and surface chemistry, such as heterogeneous catalysis and Daguerre photography, were first discovered at that time. Since then, there has been a continual interest in catalysis, corrosion and other chemical reactions that involve surfaces. The modern era of surface science began in the late 1950s, when instrumentation that could be used to investigate surface processes on the molecular level started to become available.

Since the modern era began, the study of solid surfaces has been one of the fastest growing areas in solid-state research. The geometric, electronic and chemical structure at the surface of a solid is generally quite different from that of the bulk material. It is now possible to measure the properties of a surface on the atomic scale and, in fact, to image individual atoms on a surface. The theoretical understanding of the chemistry and physics at surfaces is also improving dramatically. Much of the theoretical work has been motivated by the experimental results, as well as by the vast improvements in computer technology that are required to carry out complex numerical calculations.

Surface studies address important issues in basic physics and chemistry, but are also relevant to a variety of applications. One of the most important uses of a surface, for example, is in heterogeneous catalysis. Catalysis occurs via adsorption, diffusion and reaction on a solid surface, so that delineation of surface chemical mechanisms is critical to the understanding of catalysis. Microelectronic devices are manufactured by processing of single-crystal semiconductor surfaces. Most dry processes that occur during device manufacture involve surface etching or deposition. Thus, understanding how molecules adsorb and react on surfaces and how electron and ion beams modify surfaces is crucial to the development of manufacturing techniques for semiconductor and, more recently, micro-electromechanical (MEMS), devices. Surfaces are also the active component in tribology, i.e. solid lubrication. In order to design lubricants that will stick to one surface, yet have minimal contact with another, one must understand the fundamental surface interactions involved. In addition, the movement of pollutants through the environment is controlled by the interactions of chemicals with the surfaces encountered in the soil. Thus, a fundamental understanding of the surface chemistry of metal oxide materials is needed in order to properly evaluate and solve environmental problems.

-2-

Surfaces are found to exhibit properties that are different from those of the bulk material. In the bulk, each atom is bonded to other atoms in all three dimensions. In fact, it is this infinite periodicity in three dimensions that gives rise to the power of condensed matter physics. At a surface, however, the three-dimensional periodicity is broken. This causes the surface atoms to respond to this change in their local environment by adjusting their geometric and electronic structures. The physics and chemistry of clean surfaces is discussed in section A1.7.2.

The importance of surface science is most often exhibited in studies of adsorption on surfaces, especially in regards to technological applications. Adsorption is the first step in any surface chemical reaction or film-growth process. The mechanisms of adsorption and the properties of adsorbate-covered surfaces are discussed in section A1.7.3.

Most fundamental surface science investigations employ single-crystal samples cut along a low-index plane. The single-crystal surface is prepared to be nearly atomically flat. The surface may also be modified in vacuum. For example, it may be exposed to a gas that adsorbs (sticks) to the surface, or a film can be grown onto a sample by evaporation of material. In addition to single-crystal surfaces, many researchers have investigated vicinal, i.e. stepped, surfaces as well as the surfaces of polycrystalline and disordered materials.

In [section A1.7.4](#), methods for the preparation of surfaces are discussed.

Surfaces are investigated with surface-sensitive techniques in order to elucidate fundamental information. The approach most often used is to employ a variety of techniques to investigate a particular materials system. As each technique provides only a limited amount of information, results from many techniques must be correlated in order to obtain a comprehensive understanding of surface properties. In [section A1.7.5](#), methods for the experimental analysis of surfaces in vacuum are outlined. Note that the interactions of various kinds of particles with surfaces are a critical component of these techniques. In addition, one of the more interesting aspects of surface science is to use the tools available, such as electron, ion or laser beams, or even the tip of a scanning probe instrument, to modify a surface at the atomic scale. The physics of the interactions of particles with surfaces and the kinds of modifications that can be made to surfaces are an integral part of this section.

The liquid–solid interface, which is the interface that is involved in many chemical and environmental applications, is described in [section A1.7.6](#). This interface is more complex than the solid–vacuum interface, and can only be probed by a limited number of experimental techniques. Thus, obtaining a fundamental understanding of its properties represents a challenging frontier for surface science.

A1.7.2 CLEAN SURFACES

The study of clean surfaces encompassed a lot of interest in the early days of surface science. From this, we now have a reasonable idea of the geometric and electronic structure of many clean surfaces, and the tools are readily available for obtaining this information from other systems, as needed.

-3-

When discussing geometric structure, the macroscopic morphology must be distinguished from the microscopic atomic structure. The morphology is the macroscopic shape of the material, which is a collective property of groups of atoms determined largely by surface and interfacial tension. The following discussion, however, will concentrate on the structure at the atomic level. Note that the atomic structure often plays a role in determining the ultimate morphology of the surface. What is most important about the atomic structure, however, is that it affects the manner in which chemistry occurs on a surface at the molecular level.

A1.7.2.1 SURFACE CRYSTALLOGRAPHY

To first approximation, a single-crystal surface is atomically flat and uniform, and is composed of a regular array of atoms positioned at well defined lattice sites. Materials generally have of the order of 10^{15} atoms positioned at the outermost atomic layer of each square centimetre of exposed surface. A bulk crystalline material has virtually infinite periodicity in three dimensions, but infinite periodicity remains in only two dimensions when a solid is cut to expose a surface. In the third dimension, i.e. normal to the surface, the periodicity abruptly ends. Thus, the surface crystal structure is described in terms of a two-dimensional unit cell parallel to the surface.

In describing a particular surface, the first important parameter is the Miller index that corresponds to the orientation of the sample. Miller indices are used to describe directions with respect to the three-dimensional bulk unit cell [2]. The Miller index indicating a particular surface orientation is the one that points in the direction of the surface normal. For example, a Ni crystal cut perpendicular to the [100] direction would be labelled Ni(100).

The second important parameter to consider is the size of the surface unit cell. A surface unit cell cannot be smaller than the projection of the bulk cell onto the surface. However, the surface unit cell is often bigger than

it would be if the bulk unit cell were simply truncated at the surface. The symmetry of the surface unit cell is easily determined by visual inspection of a low-energy electron diffraction (LEED) pattern (LEED is discussed in [section A1.7.5.1](#) and [section B1.21](#)).

There is a well defined nomenclature employed to describe the symmetry of any particular surface [1]. The standard notation for describing surface symmetry is in the form

$$M(hkl)-(p \times q)-A$$

where M is the chemical symbol of the substrate material, h , k , and l are the Miller indices that indicate the surface orientation, p and q relate the size of the surface unit cell to that of the substrate unit cell and A is the chemical symbol for an adsorbate (if applicable). For example, atomically clean Ni cut perpendicular to the $[100]$ direction would be notated as Ni(100)-(1 × 1), since this surface has a bulk-terminated structure. If the unit cell were bigger than that of the substrate in one direction or the other, then p and/or q would be larger than one. For example, if a Si single crystal is cleaved perpendicular to the $[111]$ direction, a Si(111)-(2 × 1) surface is produced. Note that p and q are often, but

-4-

are not necessarily, integers. If an adsorbate is involved in forming the reconstruction, then it is explicitly part of the nomenclature. For example, when silver is adsorbed on Si(111) under the proper conditions, the Si(111)-($\sqrt{3} \times \sqrt{3}$)-Ag structure is formed.

In addition, the surface unit cell may be rotated with respect to the bulk cell. Such a rotated unit cell is notated as

$$M(hkl)-(p \times q)Rr^\circ-A$$

where r is the angle in degrees between the surface and bulk unit cells. For example, when iodine is adsorbed onto the (111) face of silver, the Ag(111)-($\sqrt{3} \times \sqrt{3}$)R30°-I structure can be produced.

Finally, there is an abbreviation ‘c’, which stands for ‘centred’, that is used to indicate certain common symmetries. In a centred structure, although the primitive unit cell is rotated from the substrate unit cell, the structure can also be considered as a non-rotated unit cell with an additional atom placed in the centre. For example, a common adsorbate structure involves placing an atom at every other surface site of a square lattice. This has the effect of rotating the primitive unit cell by 45°, so that such a structure would ordinarily be notated as ($\sqrt{2} \times \sqrt{2}$)R45°. However, the unit cell can also be thought of as a (2 × 2) in registry with the substrate with an additional atom placed in the centre of the cell. Thus, in order to simplify the nomenclature, this structure is equivalently called a c(2 × 2). Note that the abbreviation ‘p’, which stands for ‘primitive’, is sometimes used for a unit cell that is in registry with the substrate in order to distinguish it from a centred symmetry. Thus, p(2 × 2) is just an unambiguous way of representing a (2 × 2) unit cell.

A1.7.2.2 TERRACES AND STEPS

For many studies of single-crystal surfaces, it is sufficient to consider the surface as consisting of a single domain of a uniform, well ordered atomic structure based on a particular low-Miller-index orientation. However, real materials are not so flawless. It is therefore useful to consider how real surfaces differ from the ideal case, so that the behaviour that is intrinsic to a single domain of the well ordered orientation can be distinguished from that caused by defects.

Real, clean, single-crystal surfaces are composed of terraces, steps and defects, as illustrated in [figure A1.7.1](#). This arrangement is called the TLK, or terrace-ledge-kink, model. A terrace is a large region in which the surface has a well-defined orientation and is atomically flat. Note that a singular surface is defined as one that is composed solely of one such terrace. It is impossible to orient an actual single-crystal surface to precise atomic flatness, however, and steps provide the means to reconcile the macroscopic surface plane with the microscopic orientation. A step separates singular terraces, or domains, from each other. Most steps are single atomic height steps, although for certain surfaces a double-height step is required in order that each terrace is equivalent. [Figure A1.7.1\(a\)](#) illustrates two perfect terraces separated by a perfect monoatomic step. The overall number and arrangement of the steps on any actual surface is determined by the misorientation, which is the angle between the nominal crystal axis direction and the actual surface normal. If the misorientation is not along a low-index direction, then there will be kinks in the steps to adjust for this, as illustrated in [figure A1.7.1\(b\)](#).

-5-

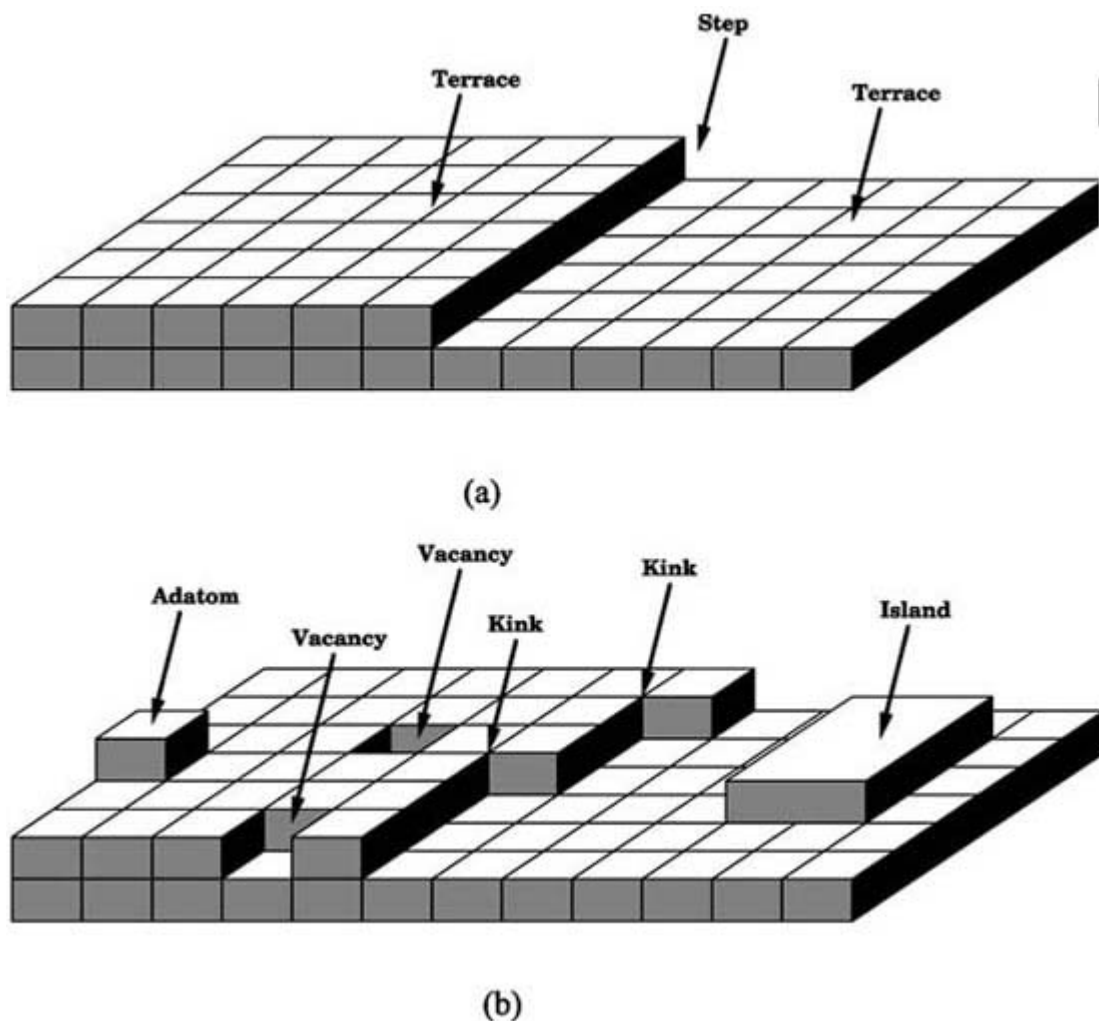


Figure A1.7.1. Schematic diagram illustrating terraces, steps, and defects. (a) Perfect flat terraces separated by a straight, monoatomic step. (b) A surface containing various defects.

A surface that differs from a singular orientation by a finite amount is called vicinal. Vicinal surfaces are composed of well oriented singular domains separated by steps. [Figure A1.7.2](#) shows a large-scale scanning tunnel microscope (STM) image of a stepped Si(111) surface (STM instruments are described in [section A1.7.5.3](#) and [section B1.20](#)). In this image, flat terraces separated by well defined steps are easily visible. It can be seen that the steps are all pointing along the same general direction.

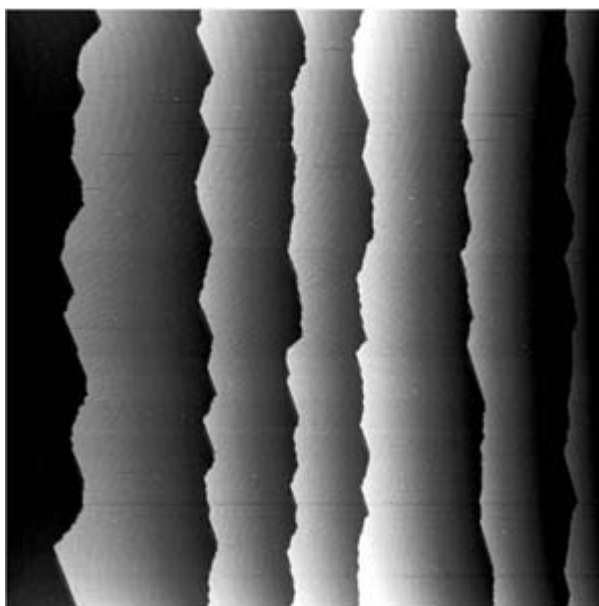


Figure A1.7.2. Large-scale ($5000 \text{ \AA} \times 5000 \text{ \AA}$) scanning tunnelling microscope image of a stepped Si (111)-(7×7) surface showing flat terraces separated by step edges (courtesy of Alison Baski).

Although all real surfaces have steps, they are not usually labelled as vicinal unless they are purposely misoriented in order to create a regular array of steps. Vicinal surfaces have unique properties, which make them useful for many types of experiments. For example, steps are often more chemically reactive than terraces, so that vicinal surfaces provide a means for investigating reactions at step edges. Also, it is possible to grow ‘nanowires’ by deposition of a metal onto a surface of another metal in such a way that the deposited metal diffuses to and attaches at the step edges [3].

Many surfaces have additional defects other than steps, however, some of which are illustrated in [figure A1.7.1\(b\)](#). For example, steps are usually not flat, i.e. they do not lie along a single low-index direction, but instead have kinks. Terraces are also not always perfectly flat, and often contain defects such as adatoms or vacancies. An *adatom* is an isolated atom adsorbed on top of a terrace, while a *vacancy* is an atom or group of atoms missing from an otherwise perfect terrace. In addition, a group of atoms called an *island* may form on a terrace, as illustrated.

Much surface work is concerned with the local atomic structure associated with a single domain. Some surfaces are essentially bulk-terminated, i.e. the atomic positions are basically unchanged from those of the bulk as if the atomic bonds in the crystal were simply cut. More common, however, are deviations from the bulk atomic structure. These structural adjustments can be classified as either relaxations or reconstructions. To illustrate the various classifications of surface structures, [figure A1.7.3\(a\)](#) shows a side-view of a bulk-terminated surface, [figure A1.7.3\(b\)](#) shows an oscillatory relaxation and [figure A1.7.3\(c\)](#) shows a reconstructed surface.

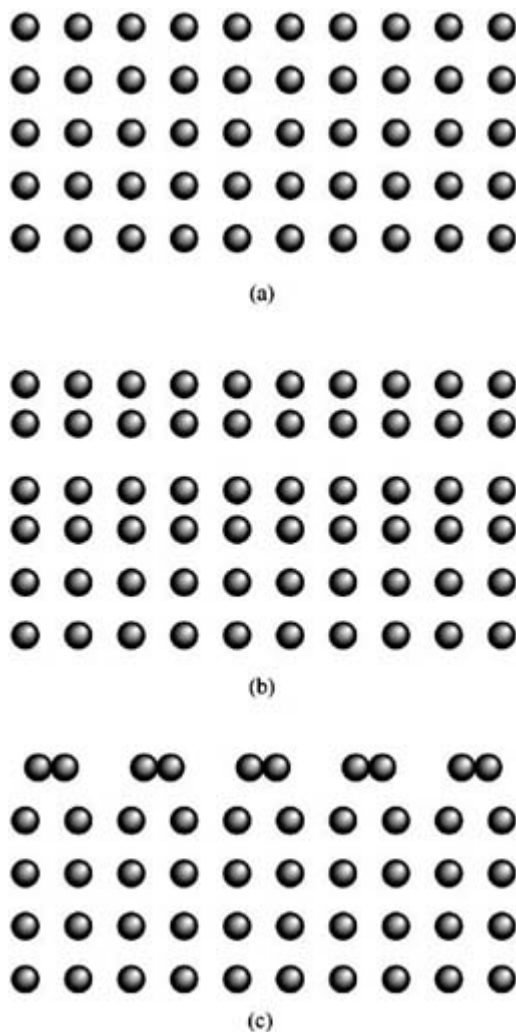


Figure A1.7.3. Schematic illustration showing side views of (a) a bulk-terminated surface, (b) a relaxed surface with oscillatory behaviour, and (c) a reconstructed surface.

A1.7.2.3 RELAXATION

Most metal surfaces have the same atomic structure as in the bulk, except that the interlayer spacings of the outermost few atomic layers differ from the bulk values. In other words, entire atomic layers are shifted as a whole in a direction perpendicular to the surface. This is called *relaxation*, and it can be either inward or outward. Relaxation is usually reported as a percentage of the value of the bulk interlayer spacing. Relaxation does not affect the two-dimensional surface unit cell symmetry, so surfaces that are purely relaxed have (1×1) symmetry.

The reason that relaxation occurs can be understood in terms of the free electron character of a metal. Because the electrons are free, they are relatively unperturbed by the periodic ion cores. Thus, the electron density is homogeneous

parallel to the surface. At the surface of a metal the solid abruptly stops, so that there is a net dipole perpendicular to the surface. This dipole field acts to attract electrons to the surface and is, in fact, responsible for the surface work function. The dipole field also interacts with the ion cores of the outermost atomic layer, however, causing them to move perpendicular to the surface. Note that some metals are also reconstructed since the assumption of perfectly free electrons unperturbed by the ion cores is not completely valid.

In many materials, the relaxations between the layers oscillate. For example, if the first-to-second layer spacing is reduced by a few percent, the second-to-third layer spacing would be increased, but by a smaller amount, as illustrated in [figure A1.7.3\(b\)](#). These oscillatory relaxations have been measured with LEED [4, 5] and ion scattering [6, 7] to extend to at least the fifth atomic layer into the material. The oscillatory nature of the relaxations results from oscillations in the electron density perpendicular to the surface, which are called Friedel oscillations [8]. The Friedel oscillations arise from Fermi–Dirac statistics and impart oscillatory forces to the ion cores.

A1.7.2.4 RECONSTRUCTION

The three-dimensional symmetry that is present in the bulk of a crystalline solid is abruptly lost at the surface. In order to minimize the surface energy, the thermodynamically stable surface atomic structures of many materials differ considerably from the structure of the bulk. These materials are still crystalline at the surface, in that one can define a two-dimensional surface unit cell parallel to the surface, but the atomic positions in the unit cell differ from those of the bulk structure. Such a change in the local structure at the surface is called a *reconstruction*.

For covalently bonded semiconductors, the largest driving force behind reconstructions is the need to pair up electrons. For example, as shown in [figure A1.7.4\(a\)](#) if a Si(100) surface were to be bulk-terminated, each surface atom would have two lone electrons pointing away from the surface (assuming that each atom remains in a tetrahedral configuration). Lone electrons protruding into the vacuum are referred to as *dangling bonds*. Instead of maintaining two dangling bonds at each surface atom, however, dimers can form in which electrons are shared by two neighbouring atoms. [Figure A1.7.4\(b\)](#) shows two symmetrically dimerized Si atoms, in which two dangling bonds have been eliminated, although the atoms still have one dangling bond each. [Figure A1.7.4\(c\)](#) shows the asymmetric arrangement that further lowers the energy by pairing up two lone electrons onto one atom. In this arrangement, the electrons at any instant are associated with one Si atom, while the other has an empty orbital. This distorts the crystal structure, as the upper atom is essentially sp^3 hybridized, i.e. tetrahedral, while the other is sp^2 , i.e. flat.

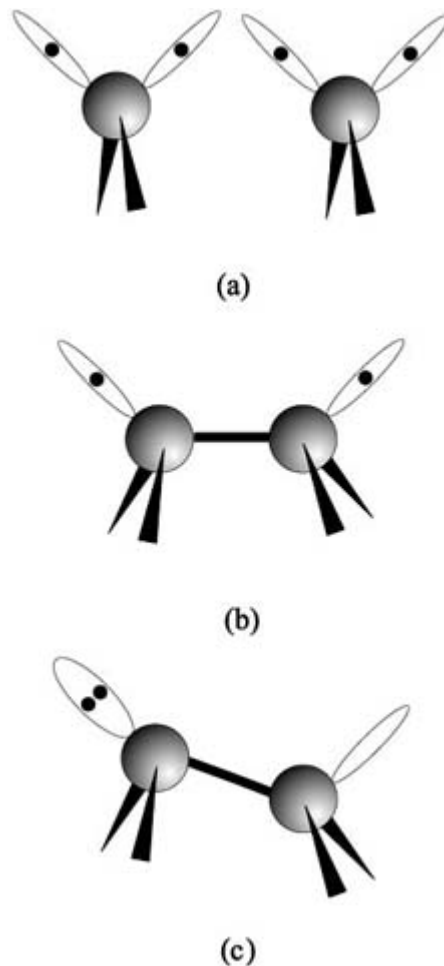


Figure A1.7.4. Schematic illustration of two Si atoms as they would be oriented on the (100) surface. (a) Bulk-terminated structure showing two dangling bonds (lone electrons) per atom. (b) Symmetric dimer, in which two electrons are shared and each atom has one remaining dangling bond. (c) Asymmetric dimer in which two electrons pair up on one atom and the other has an empty orbital.

[Figure A1.7.5\(a\)](#) shows a larger scale schematic of the Si(100) surface if it were to be bulk-terminated, while [figure A1.7.5\(b\)](#) shows the arrangement after the dimers have been formed. The dashed boxes outline the two-dimensional surface unit cells. The reconstructed Si(100) surface has a unit cell that is two times larger than the bulk unit cell in one direction and the same in the other. Thus, it has a (2×1) symmetry and the surface is labelled as Si(100)- (2×1) . Note that in actuality, however, any real Si(100) surface is composed of a mixture of (2×1) and (1×2) domains. This is because the dimer direction rotates by 90° at each step edge.

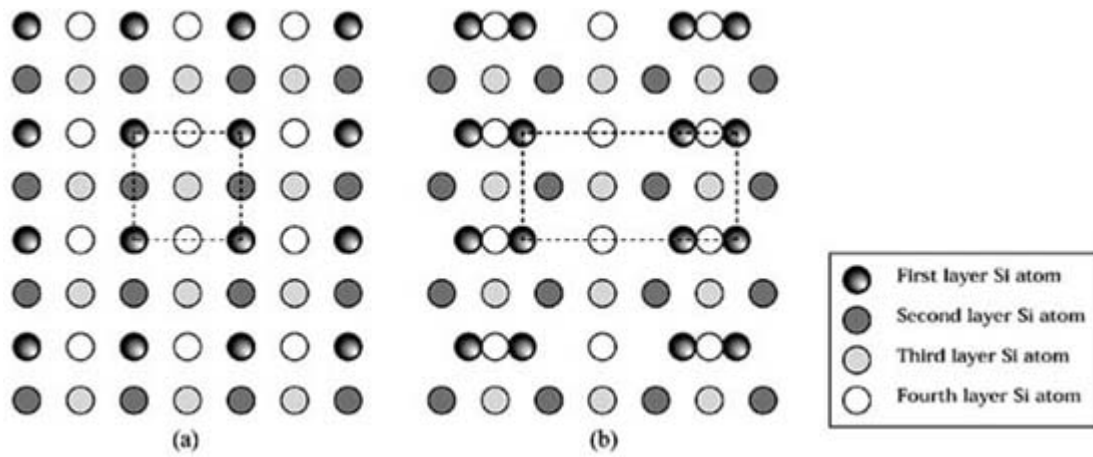


Figure A1.7.5. Schematic illustration showing the top view of the Si(100) surface. (a) Bulk-terminated structure. (b) Dimerized Si(100)-(2 × 1) structure. The dashed boxes show the two-dimensional surface unit cells.

The surface unit cell of a reconstructed surface is usually, but not necessarily, larger than the corresponding bulk-terminated two-dimensional unit cell would be. The LEED pattern is therefore usually the first indication that a reconstruction exists. However, certain surfaces, such as GaAs(110), have a reconstruction with a surface unit cell that is still (1 × 1). At the GaAs(110) surface, Ga atoms are moved inward perpendicular to the surface, while As atoms are moved outward.

The most celebrated surface reconstruction is probably that of Si(111)-(7 × 7). The fact that this surface has such a large unit cell had been known for some time from LEED, but the detailed atomic structure took many person-years of work to elucidate. Photoelectron spectroscopy [9], STM [10] and many other techniques were applied to the determination of this structure. It was transmission electron diffraction (TED), however, that provided the final information enabling the structure to be determined [11]. The structure now accepted is the so-called DAS, or dimer adatom stacking-fault, model, as shown in [figure A1.7.6](#). In this structure, there are a total of 19 dangling bonds per unit cell, which can be compared to the 49 dangling bonds that the bulk-terminated surface would have. [Figure A1.7.7](#) shows an atomic resolution STM image of the Si(111)-(7 × 7) surface. The bright spots in the image represent individual Si adatoms.

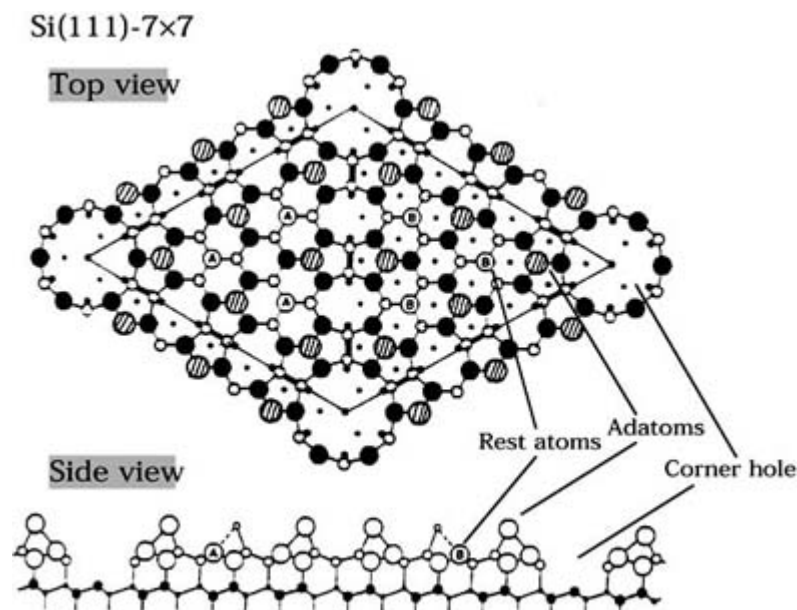


Figure A1.7.6. Schematic diagrams of the DAS model of the Si(111)-(7 × 7) surface structure. There are 12 ‘adatoms’ per unit cell in the outermost layer, which each have one dangling bond perpendicular to the surface. The second layer, called the rest layer, also has six ‘rest’ atoms per unit cell, each with a perpendicular dangling bond. The ‘corner holes’ at the edges of the unit cells also contain one atom with a dangling bond.

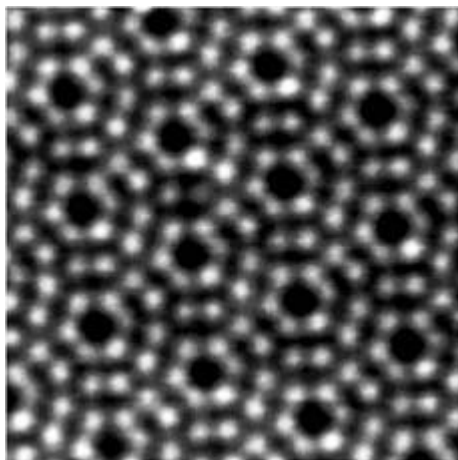


Figure A1.7.7. Atomic-resolution, empty-state STM image (100 Å × 100 Å) of the reconstructed Si(111)-7 × 7 surface. The bright spots correspond to a top layer of adatoms, with 12 adatoms per unit cell (courtesy of Alison Baski).

Although most metal surfaces exhibit only relaxation, some do have reconstructions. For example, the fcc metals, Pt(110), Au(110) and Ir(110), each have a (1 × 2) surface unit cell. The accepted structure of these surfaces is a

-12-

missing row model, in which every other surface row is missing. Also, as discussed below, when an adsorbate attaches to a metal surface, a reconstruction of the underlying substrate may be induced.

Reliable tables that list many known surface structures can be found in [1]. Also, the National Institute of Standards and Technology (NIST) maintains databases of surface structures and other surface-related information, which can be found at <http://www.nist.gov/srd/surface.htm>.

A1.7.2.5 SELF-DIFFUSION

The atoms on the outermost surface of a solid are not necessarily static, particularly as the surface temperature is raised. There has been much theoretical [12, 13] and experimental work (described below) undertaken to investigate surface self-diffusion. These studies have shown that surfaces actually have dynamic, changing structures. For example, atoms can diffuse along a terrace to or from step edges. When atoms diffuse across a surface, they may move by hopping from one surface site to the next, or by exchanging places with second layer atoms.

The field ion microscope (FIM) has been used to monitor surface self-diffusion in real time. In the FIM, a sharp, crystalline tip is placed in a large electric field in a chamber filled with He gas [14]. At the tip, He ions are formed, and then accelerated away from the tip. The angular distribution of the He ions provides a picture of the atoms at the tip with atomic resolution. In these images, it has been possible to monitor the diffusion of a single adatom on a surface in real time [15]. The limitations of FIM, however, include its applicability only to metals, and the fact that the surfaces are limited to those that exist on a sharp tip, i.e. diffusion along a large

terrace cannot be observed.

More recently, studies employing STM have been able to address surface self-diffusion across a terrace [16, 17, 18 and 19]. It is possible to image the same area on a surface as a function of time, and ‘watch’ the movement of individual atoms. These studies are limited only by the speed of the instrument. Note that the performance of STM instruments is constantly improving, and has now surpassed the 1 ps time resolution mark [20]. Not only has self-diffusion of surface atoms been studied, but the diffusion of vacancy defects on surfaces has also been observed with STM [18].

It has also been shown that sufficient surface self-diffusion can occur so that entire step edges move in a concerted manner. Although it does not achieve atomic resolution, the low-energy electron microscopy (LEEM) technique allows for the observation of the movement of step edges in real time [21]. LEEM has also been useful for studies of epitaxial growth and surface modifications due to chemical reactions.

A1.7.2.6 SURFACE ELECTRONIC STRUCTURE

At a surface, not only can the atomic structure differ from the bulk, but electronic energy levels are present that do not exist in the bulk band structure. These are referred to as ‘surface states’. If the states are occupied, they can easily be measured with photoelectron spectroscopy (described in [section A1.7.5.1](#) and [section B1.25.2](#)). If the states are unoccupied, a technique such as inverse photoemission or x-ray absorption is required [22, 23]. Also, note that STM has been used to measure surface states by monitoring the tunnelling current as a function of the bias voltage [24] (see [section B1.20](#)). This is sometimes called scanning tunnelling spectroscopy (STS).

-13-

Surface states can be divided into those that are intrinsic to a well ordered crystal surface with two-dimensional periodicity, and those that are extrinsic [25]. Intrinsic states include those that are associated with relaxation and reconstruction. Note, however, that even in a bulk-terminated surface, the outermost atoms are in a different electronic environment than the substrate atoms, which can also lead to intrinsic surface states. Extrinsic surface states are associated with imperfections in the perfect order of the surface region. Extrinsic states can also be formed by an adsorbate, as discussed below.

Note that in core-level photoelectron spectroscopy, it is often found that the surface atoms have a different binding energy than the bulk atoms. These are called surface core-level shifts (SCLS), and should not be confused with intrinsic surface states. An SCLS is observed because the atom is in a chemically different environment than the bulk atoms, but the core-level state that is being monitored is one that is present in all of the atoms in the material. A surface state, on the other hand, exists only at the particular surface.

A1.7.3 ADSORPTION

When a surface is exposed to a gas, the molecules can adsorb, or stick, to the surface. Adsorption is an extremely important process, as it is the first step in any surface chemical reaction. Some of the aspects of adsorption that surface science is concerned with include the mechanisms and kinetics of adsorption, the atomic bonding sites of adsorbates and the chemical reactions that occur with adsorbed molecules.

The coverage of adsorbates on a given substrate is usually reported in monolayers (ML). Most often, 1 ML is defined as the number of atoms in the outermost atomic layer of the unreconstructed, i.e. bulk-terminated, substrate. Sometimes, however, 1 ML is defined as the maximum number of adsorbate atoms that can stick to a particular surface, which is termed the saturation coverage. The saturation coverage can be much smaller

than the number of surface atoms, particularly with large adsorbates. Thus, in reading the literature, care must be taken to understand how a particular author defines 1 ML.

Molecular adsorbates usually cover a substrate with a single layer, after which the surface becomes passive with respect to further adsorption. The actual saturation coverage varies from system to system, and is often determined by the strength of the repulsive interactions between neighbouring adsorbates. Some molecules will remain intact upon adsorption, while others will adsorb dissociatively. This is often a function of the surface temperature and composition. There are also often multiple adsorption states, in which the stronger, more tightly bound states fill first, and the more weakly bound states fill last. The factors that control adsorbate behaviour depend on the complex interactions between adsorbates and the substrate, and between the adsorbates themselves.

The probability for sticking is known as the sticking coefficient, S . Usually, S decreases with coverage. Thus, the sticking coefficient at zero coverage, the so-called initial sticking coefficient, S_0 , reflects the interaction of a molecule with the bare surface.

In order to calibrate the sticking coefficient, one needs to determine the exposure, i.e. how many molecules have initially impacted a surface. The Langmuir (L) is a unit of exposure that is defined as 10^{-6} Torr s. An exposure of 1 L is approximately the number of incident molecules such that each outermost surface atom is impacted once. Thus, a

-14-

1 L exposure would produce 1 ML of adsorbates if the sticking coefficient were unity. Note that a quantitative calculation of the exposure per surface atom depends on the molecular weight of the gas molecules and on the actual density of surface atoms, but the approximations inherent in the definition of the Langmuir are often inconsequential.

A1.7.3.1 PHYSISORPTION

Adsorbates can physisorb onto a surface into a shallow potential well, typically 0.25 eV or less [25]. In physisorption, or physical adsorption, the electronic structure of the system is barely perturbed by the interaction, and the physisorbed species are held onto a surface by weak van der Waals forces. This attractive force is due to charge fluctuations in the surface and adsorbed molecules, such as mutually induced dipole moments. Because of the weak nature of this interaction, the equilibrium distance at which physisorbed molecules reside above a surface is relatively large, of the order of 3 Å or so. Physisorbed species can be induced to remain adsorbed for a long period of time if the sample temperature is held sufficiently low. Thus, most studies of physisorption are carried out with the sample cooled by liquid nitrogen or helium.

Note that the van der Waals forces that hold a physisorbed molecule to a surface exist for all atoms and molecules interacting with a surface. The physisorption energy is usually insignificant if the particle is attached to the surface by a much stronger chemisorption bond, as discussed below. Often, however, just before a molecule forms a strong chemical bond to a surface, it exists in a physisorbed precursor state for a short period of time, as discussed below in [section A1.7.3.3](#).

A1.7.3.2 CHEMISORPTION

Chemisorption occurs when the attractive potential well is large so that upon adsorption a strong chemical bond to a surface is formed. Chemisorption involves changes to both the molecule and surface electronic states. For example, when oxygen adsorbs onto a metal surface, a partially ionic bond is created as charge transfers from the substrate to the oxygen atom. Other chemisorbed species interact in a more covalent manner by sharing electrons, but this still involves perturbations to the electronic system.

Chemisorption is always an exothermic process. By convention, the heat of adsorption, ΔH_{ads} , has a positive sign, which is opposite to the normal thermodynamic convention [1]. Although the heat of adsorption has been directly measured with the use of a very sensitive microcalorimeter [26], it is more commonly measured via adsorption isotherms [1]. An isotherm is generated by measuring the coverage of adsorbates obtained by reaction at a fixed temperature as a function of the flux of incoming gas molecules. The flux is adjusted by regulating the pressure used during exposure. An analysis of the data then allows H_{ads} and other parameters to be determined. Heats of adsorption can also be determined from temperature programmed desorption (TPD) if the adsorption is reversible (TPD is discussed in [section A1.7.5.4](#) and [section B1.25](#)).

When a molecule adsorbs to a surface, it can remain intact or it may dissociate. Dissociative chemisorption is common for many types of molecules, particularly if all of the electrons in the molecule are tied up so that there are no electrons available for bonding to the surface without dissociation. Often, a molecule will dissociate upon adsorption, and then recombine and desorb intact when the sample is heated. In this case, dissociative chemisorption can be detected with TPD by employing isotopically labelled molecules. If mixing occurs during the adsorption/desorption sequence, it indicates that the initial adsorption was dissociative.

-15-

Atom abstraction occurs when a dissociation reaction occurs on a surface in which one of the dissociation products sticks to the surface, while another is emitted. If the chemisorption reaction is particularly exothermic, the excess energy generated by chemical bond formation can be channelled into the kinetic energy of the desorbed dissociation fragment. An example of atom abstraction involves the reaction of molecular halogens with Si surfaces [27, 28]. In this case, one halogen atom chemisorbs while the other atom is ejected from the surface.

A1.7.3.3 ADSORPTION KINETICS

When an atom or molecule approaches a surface, it feels an attractive force. The interaction potential between the atom or molecule and the surface, which depends on the distance between the molecule and the surface and on the lateral position above the surface, determines the strength of this force. The incoming molecule feels this potential, and upon adsorption becomes trapped near the minimum in the well. Often the molecule has to overcome an activation barrier, E_{act} , before adsorption can occur.

It is the relationship between the bound potential energy surface of an adsorbate and the vibrational states of the molecule that determine whether an adsorbate remains on the surface, or whether it desorbs after a period of time. The lifetime of the adsorbed state, τ , depends on the size of the well relative to the vibrational energy inherent in the system, and can be written as

$$\tau = \tau_0 \exp(\Delta H_{\text{ads}}/kT). \quad (\text{A1.7.1})$$

Such lifetimes vary from less than a picosecond to times greater than the age of the universe [29]. Thus, adsorbed states with short lifetimes can occur during a surface chemical reaction, or long-lived adsorbed states exist in which atoms or molecules remain attached to a surface indefinitely.

In this manner, it can also be seen that molecules will desorb as the surface temperature is raised. This is the phenomenon employed for TPD spectroscopy (see [section A1.7.5.4](#) and [section B1.25](#)). Note that some adsorbates may adsorb and desorb reversibly, i.e. the heats of adsorption and desorption are equal. Other adsorbates, however, will adsorb and desorb via different pathways.

Note that chemisorption often begins with physisorption into a weakly bound precursor state. While in this

state, the molecule can diffuse along the surface to find a likely site for chemisorption. This is particularly important in the case of dissociative chemisorption, as the precursor state can involve physisorption of the intact molecule. If a precursor state is involved in adsorption, a negative temperature dependence to the adsorption probability will be found. A higher surface temperature reduces the lifetime of the physisorbed precursor state, since a weakly bound species will not remain on the surface in the presence of thermal excitation. Thus, the sticking probability will be reduced at higher surface temperatures.

The kinetics of the adsorption process are important in determining the value and behaviour of S for any given system. There are several factors that come into play in determining S [25].

-16-

- (a) The activation barrier must be overcome in order for a molecule to adsorb. Thus, only the fraction of the incident particles whose energy exceeds E_{act} will actually stick.
- (b) The electronic orbitals of the incoming molecule must have the correct orientation with respect to the orbitals of the surface. Thus, only a fraction of the incoming molecules will immediately stick to the surface. Some of the incoming molecules may, however, diffuse across the surface while in a precursor state until they achieve the proper orientation. Thus, the details of how the potential energy varies across the surface are critical in determining the adsorption kinetics.
- (c) Upon adsorption, a molecule must effectively lose the remaining part of its kinetic energy, and possibly the excess energy liberated by an exothermic reaction, in a time period smaller than one vibrational period. Thus, excitations of the surface that can carry away this excess energy, such as plasmons or phonons, play a role in the adsorption kinetics.
- (d) Adsorption sites must be available for reaction. Thus, the kinetics may depend critically on the coverage of adsorbates already present on the surface, as these adsorbates may block or modify the remaining adsorption sites.

A1.7.3.4 ADSORPTION MODELS

The most basic model for chemisorption is that developed by Langmuir. In the Langmuir model, it is assumed that there is a finite number of adsorption sites available on a surface, and each has an equal probability for reaction. Once a particular site is occupied, however, the adsorption probability at that site goes to zero. Furthermore, it is assumed that the adsorbates do not diffuse, so that once a site is occupied it remains unreactive until the adsorbate desorbs from the surface. Thus, the sticking probability S goes to zero when the coverage, θ , reaches the saturation coverage, θ_0 . These assumptions lead to the following relationship between the sticking coefficient and the surface coverage,

$$S = S_0(1 - \theta/\theta_0). \quad (\text{A1.7.2})$$

The straight line in [figure A1.7.8](#) shows the relationships between S and θ expected for various models, with the straight line indicating Langmuir adsorption.

-17-

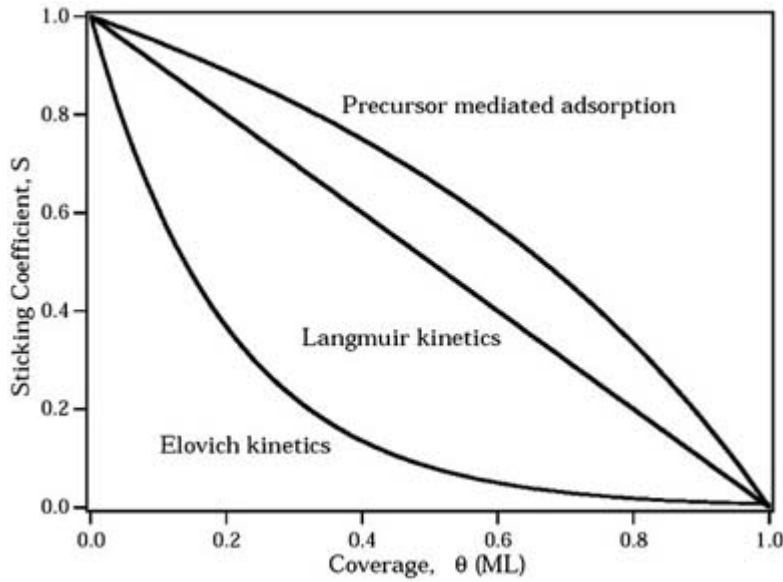


Figure A1.7.8. Sticking probability as a function of surface coverage for three different adsorption models.

Adsorbate atoms have a finite lifetime, τ , for remaining on a surface. Thus, there will always be a flux of molecules leaving the surface even as additional molecules are being adsorbed. If the desorption rate is equal to the rate of adsorption, then an isotherm can be collected by measuring the equilibrium coverage at a fixed temperature as a function of pressure, p . From the assumptions of the Langmuir model, one can derive the following expression relating the equilibrium coverage to pressure [29].

$$\theta = \frac{\chi p}{1 + \chi p} \quad (\text{A1.7.3})$$

where χ is a constant that depends on the adsorbate lifetime and surface temperature, T , as

$$\chi \propto \tau T^{1/2}. \quad (\text{A1.7.4})$$

If Langmuir adsorption occurs, then a plot of θ versus p for a particular isotherm will display the form of equation (A1.7.3). Measurements of isotherms are routinely employed in this manner in order to determine adsorption kinetics.

Langmuir adsorption adequately describes the behaviour of many systems in which strong chemisorption takes place, but it has limitations. For one, the sticking at surface sites actually does depend on the occupancy of neighbouring sites. Thus, sticking probability usually changes with coverage. A common observation, for example, is that the sticking probability is reduced exponentially with coverage, i.e.

$$S \propto \exp(-\alpha\theta/kT) \quad (\text{A1.7.5})$$

which is called the Elovich equation [25]. This is compared to the Langmuir model in [figure A1.7.8](#).

If adsorption occurs via a physisorbed precursor, then the sticking probability at low coverages will be enhanced due to the ability of the precursor to diffuse and find a lattice site [30]. The details depend on parameters such as strength of the lateral interactions between the adsorbates and the relative rates of desorption and reaction of the precursor. In [figure A1.7.8](#) an example of a plot of S versus θ for precursor mediated adsorption is presented.

Another limitation of the Langmuir model is that it does not account for multilayer adsorption. The Braunauer, Emmett and Teller (BET) model is a refinement of Langmuir adsorption in which multiple layers of adsorbates are allowed [29, 31]. In the BET model, the particles in each layer act as the adsorption sites for the subsequent layers. There are many refinements to this approach, in which parameters such as sticking coefficient, activation energy, etc, are considered to be different for each layer.

A1.7.3.5 ADSORPTION SITES

When atoms, molecules, or molecular fragments adsorb onto a single-crystal surface, they often arrange themselves into an ordered pattern. Generally, the size of the adsorbate-induced two-dimensional surface unit cell is larger than that of the clean surface. The same nomenclature is used to describe the surface unit cell of an adsorbate system as is used to describe a reconstructed surface, i.e. the symmetry is given with respect to the bulk terminated (unreconstructed) two-dimensional surface unit cell.

When chemisorption takes place, there is a strong interaction between the adsorbate and the substrate. The details of this interaction determine the local bonding site, particularly at the lowest coverages. At higher coverages, adsorbate–adsorbate interactions begin to also play a role. Most non-metallic atoms will adsorb above the surface at specific lattice sites. Some systems have multiple bonding sites. In this case, one site will usually dominate at low coverage, but a second, less stable site will be filled at higher coverages. Some adsorbates will interact with only one surface atom, i.e. be singly coordinated, while others prefer multiple coordinated adsorption sites. Other systems may form alloys or intermix during adsorption.

Local adsorption sites can be roughly classified either as on-top, bridge or hollow, as illustrated for a four-fold symmetric surface in [figure A1.7.9](#). In the on-top configuration, a singly coordinated adsorbate is attached directly on top of a substrate atom. A bridge site is the two-fold site between two neighbouring surface atoms. A hollow site is positioned between three or four surface atoms, for surfaces with three- or four-fold symmetry, respectively.

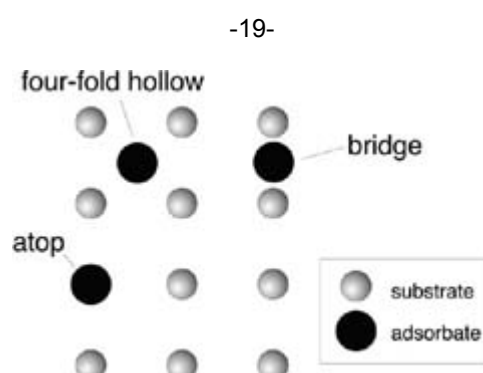


Figure A1.7.9. Schematic diagram illustrating three types of adsorption sites.

There are interactions between the adsorbates themselves, which greatly affect the structure of the adsorbates [32]. If surface diffusion is sufficiently facile during or following the adsorption step, attractive interactions can induce the adsorbates to form islands in which the local adsorbate concentration is quite high. Other adsorbates may repel each other at low coverages forming structures in which the distance between adsorbates

is maximized. Certain co-adsorption systems form complex ordered overlayer structures. The driving force in forming ordered overlayers are these adsorbate–adsorbate interactions. These interactions dominate the long-range structure of the surface in the same way that long-range interactions cause the formation of three-dimensional solid crystals.

Adsorbed atoms and molecules can also diffuse across terraces from one adsorption site to another [33]. On a perfect terrace, adatom diffusion could be considered as a ‘random walk’ between adsorption sites, with a diffusivity that depends on the barrier height between neighbouring sites and the surface temperature [29]. The diffusion of adsorbates has been studied with FIM [14], STM [34, 35] and laser-induced thermal desorption [36].

A1.7.3.6 ADSORPTION-INDUCED RECONSTRUCTION

When an adsorbate attaches to a surface, the substrate itself may respond to the perturbation by either losing its relaxation or reconstruction, or by forming a new reconstruction. This is not surprising, considering the strength of a chemisorption bond. Chemisorption bonds can provide electrons to satisfy the requirements for charge neutrality or electron pairing that may otherwise be missing at a surface.

For a reconstructed surface, the effect of an adsorbate can be to provide a more bulk-like environment for the outermost layer of substrate atoms, thereby lifting the reconstruction. An example of this is As adsorbed onto Si(111)-(7 × 7) [37]. Arsenic atoms have one less valence electron than Si. Thus, if an As atom were to replace each outermost Si atom in the bulk-terminated structure, a smooth surface with no unpaired electrons would be produced, with a second layer consisting of Si atoms in their bulk positions. Arsenic adsorption has, in fact, been found to remove the reconstruction and form a Si(111)-(1 × 1)–As structure. This surface has a particularly high stability due to the absence of dangling bonds.

An example of the formation of a new reconstruction is given by certain fcc (110) metal surfaces. The clean surfaces have (1 × 1) symmetry, but become (2 × 1) upon adsorption of oxygen [16, 38]. The (2 × 1) symmetry is not just due to oxygen being adsorbed into a (2 × 1) surface unit cell, but also because the substrate atoms rearrange themselves

-20-

into a new configuration. The reconstruction that occurs is sometimes called the ‘missing-row’ structure because every other row of surface atoms along the 2× direction is missing. A more correct terminology, however, is the ‘added-row’ structure, as STM studies have shown that it is formed by metal atoms diffusing away from a step edge and onto a terrace to create a new first layer, rather than by atoms being removed [16]. In this case, the (2 × 1) symmetry results not just from the long-range structure of the adsorbed layer, but also from a rearrangement of the substrate atoms.

A more dramatic type of restructuring occurs with the adsorption of alkali metals onto certain fcc metal surfaces [39]. In this case, multilayer composite surfaces are formed in which the alkali and metal atoms are intermixed in an ordered structure. These structures involve the substitution of alkali atoms into substrate sites, and the details of the structures are found to be coverage-dependent. The structures are influenced by the repulsion between the dipoles formed by neighbouring alkali adsorbates and by the interactions of the alkalis with the substrate itself [40].

There is also an interesting phenomenon that has been observed following the deposition of the order of 1 ML of a metal onto another metallic substrate. For certain systems, this small coverage is sufficient to alter the surface energy so that a large-scale faceting of the surface occurs [41]. The morphology of such a faceted

surface can be seen in the STM image of figure A1.7.10 which was collected from an annealed W(111) surface onto which a small amount of Pd had been deposited.

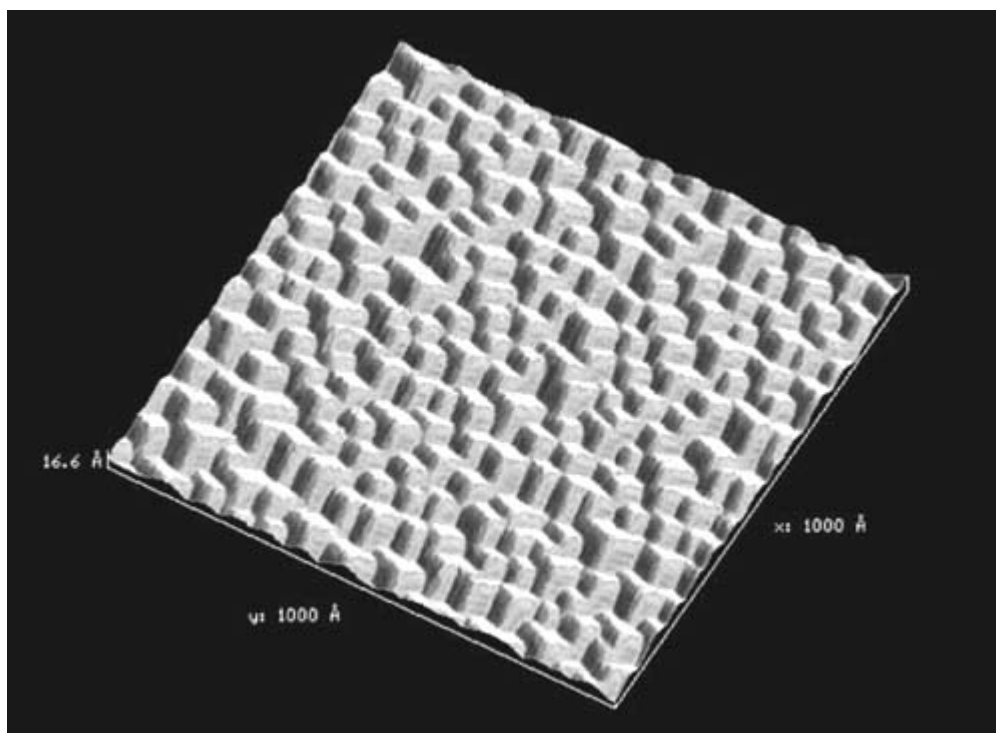


Figure A1.7.10. STM image ($1000 \text{ \AA} \times 1000 \text{ \AA}$) of the (111) surface of a tungsten single crystal, after it had been coated with a very thin film of palladium and heated to about 800 K (courtesy of Ted Madey).

A1.7.3.7 WORK FUNCTION CHANGES INDUCED BY ADSORBATES

The surface work function is formally defined as the minimum energy needed in order to remove an electron from a solid. It is often described as being the difference in energy between the Fermi level and the vacuum level of a solid. The work function is a sensitive measure of the surface electronic structure, and can be measured in a number of ways, as described in [section B1.26.4](#). Many processes, such as catalytic surface reactions or resonant charge transfer between ions and surfaces, are critically dependent on the work function.

When an electropositive or electronegative adsorbate attaches itself to a surface, there is usually a change in the surface dipole, which, in turn, affects the surface work function. Thus, very small coverages of adsorbates can be used to modify the surface work function in order to ascertain the role that the work function plays in a given process. Conversely, work function measurements can be used to accurately determine the coverage of these adsorbates.

For example, alkali ions adsorbed onto surfaces donate some or all of their valence electron to the solid, thereby producing dipoles pointing away from the surface [40, 42]. This has the effect of substantially lowering the work function for coverages as small as 0.01 ML. When the alkali coverage is increased to the point at which the alkali adsorbates can interact with each other, they tend to depolarize. Thus, the work function initially decreases as alkali atoms are adsorbed until a minimum in the work function is attained. At higher alkali coverages, the work function may increase slightly due to the adsorbate–adsorbate interactions. Note that it is very common to use alkali adsorption as a means of modifying the surface work function.

A1.7.3.8 SURFACE CHEMICAL REACTIONS

Surface chemical reactions can be classified into three major categories [29]:

- (a) corrosion reactions,
- (b) crystal growth reactions,
- (c) catalytic reactions.

All three types of reactions begin with adsorption of species onto a surface from the gas phase.

In corrosion, adsorbates react directly with the substrate atoms to form new chemical species. The products may desorb from the surface (volatilization reaction) or may remain adsorbed in forming a corrosion layer. Corrosion reactions have many industrial applications, such as dry etching of semiconductor surfaces. An example of a volatilization reaction is the etching of Si by fluorine [43]. In this case, fluorine reacts with the Si surface to form SiF_4 gas. Note that the crystallinity of the remaining surface is also severely disrupted by this reaction. An example of corrosion layer formation is the oxidation of Fe metal to form rust. In this case, none of the products are volatile, but the crystallinity of the surface is disrupted as the bulk oxide forms. Corrosion and etching reactions are discussed in more detail in [section A3.10](#) and [section C2.9](#).

The growth of solid films onto solid substrates allows for the production of artificial structures that can be used for many purposes. For example, film growth is used to create pn junctions and metal–semiconductor contacts during semiconductor manufacture, and to produce catalytic surfaces with properties that are not found in any single material. Lubrication can be applied to solid surfaces by the appropriate growth of a solid lubricating film. Film growth is also

used to fabricate quantum-wells and other types of layered structures that have unique electronic properties. These reactions may involve dissociative or non-dissociative adsorption as the first step. The three basic types of film growth reactions are physical vapour deposition (PVD), chemical vapour deposition (CVD) and molecular beam epitaxy (MBE). In PVD, an atomic gas is condensed onto a surface forming a solid. In CVD, a molecular gas dissociates upon adsorption. Some of the dissociation fragments solidify to form the material, while other dissociation fragments are evolved back into the gas phase. In MBE, carefully controlled atomic and/or molecular beams are condensed onto a surface in the proper stoichiometry in order to grow a desired material [44]. MBE is particularly important in the growth of III–V semiconductor materials.

In crystal growth reactions, material is deposited onto a surface in order to extend the surface crystal structure, or to grow a new material, without disruption of the underlying substrate. Growth mechanisms can be roughly divided into three categories. If the film grows one atomic layer at a time such that a smooth, uniform film is created, it is called *Frank von der Merwe* growth. Such layer-by-layer growth will occur if the surface energy of the overlayer is lower than that of the substrate. If the film grows in a von der Merwe growth mode such that it forms a single crystal in registry with the substrate, it is referred to as epitaxial. The smaller the lattice mismatch between the overlayer and the substrate, the more likely it is that epitaxial growth can be achieved. If the first ML is deposited uniformly, but subsequent layers agglomerate into islands, it is called *Stranski–Krastanov* growth. In this case, the surface energy of the first layer is lower than that of the substrate, but the surface energy of the bulk overlayer material is higher. If the adsorbate agglomerates into islands immediately, without even wetting the surface, it is referred to as *Vollmer–Weber* growth. In this case, the surface energy of the substrate is lower than that of the overlayer. Growth reactions are discussed in more detail in [section A3.10](#).

The desire to understand catalytic chemistry was one of the motivating forces underlying the development of surface science. In a catalytic reaction, the reactants first adsorb onto the surface and then react with each other to form volatile product(s). The substrate itself is not affected by the reaction, but the reaction would not occur without its presence. Types of catalytic reactions include exchange, recombination, unimolecular decomposition, and bimolecular reactions. A reaction would be considered to be of the *Langmuir–Hinshelwood* type if both reactants first adsorbed onto the surface, and then reacted to form the products. If one reactant first adsorbs, and the other then reacts with it directly from the gas phase, the reaction is of the *Eley–Rideal* type. Catalytic reactions are discussed in more detail in [section A3.10](#) and [section C2.8](#).

A tremendous amount of work has been done to delineate the detailed reaction mechanisms for many catalytic reactions on well characterized surfaces [1, 45]. Many of these studies involved impinging molecules onto surfaces at relatively low pressures, and then interrogating the surfaces in vacuum with surface science techniques. For example, a useful technique for catalytic studies is TPD, as the reactants can be adsorbed onto the sample in one step, and the products formed in a second step when the sample is heated. Note that catalytic surface studies have also been performed by reacting samples in a high-pressure cell, and then returning them to vacuum for measurement.

Recently, *in situ* studies of catalytic surface chemical reactions at high pressures have been undertaken [46, 47]. These studies employed sum frequency generation (SFG) and STM in order to probe the surfaces as the reactions are occurring under conditions similar to those employed for industrial catalysis (SFG is a laser-based technique that is described in [section A1.7.5.5](#) and [section B1.22](#)). These studies have shown that the highly stable adsorbate sites that are probed under vacuum conditions are not necessarily the same sites that are active in high-pressure catalysis. Instead, less stable sites that are only occupied at high pressures are often responsible for catalysis. Because the active

-23-

adsorption sites are not populated at low pressures, they are not seen in vacuum surface science experiments. Despite this, however, the low-pressure experiments are necessary in order to calibrate the spectroscopy so that the high-pressure results can be properly interpreted.

A1.7.4 PREPARATION OF CLEAN SURFACES

The exact methods employed to prepare any particular surface for study vary from material to material, and are usually determined empirically. In some respects, sample preparation is more of an art than a science. Thus, it is always best to consult the literature to look for preparation methods before starting with a new material.

Most samples require some initial *ex situ* preparation before insertion into a vacuum chamber [45]. A bulk single crystal must first be oriented [48], which is usually done with back-reflection Laue x-ray diffraction, and then cut to expose the desired crystal plane. Samples are routinely prepared to be within $\pm 1^\circ$ of the desired orientation, but an accuracy of $\pm 1/4^\circ$ or better can be routinely obtained. Cutting is often done using an electric discharge machine (spark cutter) for metals or a diamond saw or slurry drill for semiconductors. The surface must then be polished. Most polishing is done mechanically, with alumina or diamond paste, by polishing with finer and finer grits until the finest available grit is employed, which is usually of the order of $0.5\ \mu\text{m}$. Often, as a final step, the surface is electrochemically or chemi-mechanically polished. In addition, some samples are chemically reacted in solution in order to remove a large portion of the oxide layer that is present due to reaction with the atmosphere. Note that this layer is referred to as the *native oxide*.

In order to maintain the cleanliness of a surface at the atomic level, investigations must be carried out in ultra-high vacuum (UHV). UHV is usually considered to be a pressure of the order of 1×10^{-10} Torr or below. Surface science techniques are often sensitive to adsorbate levels as small as 1% of ML or less, so that great care must be taken to keep the surface contamination to a minimum. Even at moderate pressures, many contaminants will easily adsorb onto a surface. For example, at 1×10^{-6} Torr, which is a typical pressure realized by many diffusion-pumped systems, a 1 L exposure to the background gases will occur in 1 s. Thus, any molecule that is present in the background and has a high sticking probability, such as water or oxygen, will cover the surface within seconds. It is for this reason that extremely low pressures are necessary in order to keep surfaces contaminant-free at the atomic level.

Once a sample is properly oriented and polished, it is placed into a UHV chamber for the final preparation steps. Samples are processed *in situ* by a variety of methods in order to produce an atomically clean and flat surface. Ion bombardment and annealing (IBA) is the most common method used. Other methods include cleaving and film growth.

In IBA, the samples are first irradiated for a period of time with noble gas ions, such as Ar^+ or Ne^+ , that have kinetic energies in the range of 0.5–2.0 keV. This removes the outermost layers of adsorbed contaminants and oxides by the process of sputtering. In sputtering, ions directly collide with the atoms at the surface of the sample, physically knocking out material. Usually the sample is at room temperature during sputtering and the ion beam is incident normal to the surface. Certain materials, however, are better prepared by sputtering at elevated temperature or with different incidence directions.

-24-

Because keV ions penetrate several layers deep into a solid, a side effect of sputtering is that it destroys the crystallinity of the surface region. In the preparation of a single-crystal surface, the damage is removed by annealing (heating) the surface in UHV in order to re-crystallize it. Care must be taken to not overheat the sample for (at least) two reasons. First, surfaces will melt and/or sublime well below the melting point of the bulk material. Second, contaminants sometimes diffuse to the surface from the bulk at high temperatures. If the annealing temperature is not high enough, however, the material will not be sufficiently well ordered. Thus, care must be taken to determine the optimal annealing temperature for any given material.

After a sample has been sputtered to remove the contaminants and then annealed at the proper temperature to re-crystallize the surface region, a clean, atomically smooth and homogeneous surface can be produced. Note, however, that it usually takes many cycles of IBA to produce a good surface. This is because a side effect of annealing is that the chamber pressure is raised as adsorbed gases are emitted from the sample holder, which causes additional contaminants to be deposited on the surface. Also, contaminants may have diffused to the surface from the bulk during annealing. Another round of sputtering is then needed to remove these additional contaminants. After a sufficient number of cycles, the contaminants in either the sample holder or the bulk solid are depleted to the point that annealing does not significantly contaminate the surface.

For some materials, the most notable being silicon, heating alone suffices to clean the surface. Commercial Si wafers are produced with a thin layer of silicon dioxide covering the surface. This native oxide is inert to reaction with the atmosphere, and therefore keeps the underlying Si material clean. The native oxide layer is desorbed, i.e. removed into the gas phase, by heating the wafer in UHV to a temperature above approximately 1100 °C. This procedure directly forms a clean, well ordered Si surface.

At times, *in situ* chemical treatments are used to remove particular contaminants. This is done by introducing a low pressure ($\sim 10^6$ Torr) of gas to the vacuum chamber, which causes it to adsorb (stick) to the sample surface, followed by heating the sample to remove the adsorbates. The purpose is to induce a chemical

reaction between the contaminants and the adsorbed gas to form a volatile product. For example, carbon can be removed by exposing a surface to hydrogen gas and then heating it. This procedure produces methane gas, which desorbs from the surface into the vacuum. Similarly, hydrogen adsorption can be used to remove oxygen by forming gaseous water molecules.

Certain materials, most notably semiconductors, can be mechanically cleaved along a low-index crystal plane *in situ* in a UHV chamber to produce an ordered surface without contamination. This is done using a sharp blade to slice the sample along its preferred cleavage direction. For example, Si cleaves along the (111) plane, while III–V semiconductors cleave along the (110) plane. Note that the atomic structure of a cleaved surface is not necessarily the same as that of the same crystal face following treatment by IBA.

In addition, ultra-pure films are often grown *in situ* by evaporation of material from a filament or crucible, by molecular beam epitaxy (MBE), or with the use of chemical methods. Since the films are grown in UHV, the surfaces as grown will be atomically clean. Film growth has the advantage of producing a much cleaner and/or more highly ordered surface than could be obtained with IBA. In addition, certain structures can be formed with MBE that cannot be produced by any other preparation method. Film growth is discussed more explicitly above in [section A1.7.3.8](#) and in [section A3.10](#).

-25-

A1.7.5 TECHNIQUES FOR THE INVESTIGATION OF SURFACES

Because surface science employs a multitude of techniques, it is necessary that any worker in the field be acquainted with at least the basic principles underlying the most popular ones. These will be briefly described here. For a more detailed discussion of the physics underlying the major surface analysis techniques, see the appropriate chapter in this encyclopedia, or [49].

With the exception of the scanning probe microscopies, most surface analysis techniques involve scattering of one type or another, as illustrated in figure A1.7.11. A particle is incident onto a surface, and its interaction with the surface either causes a change to the particles' energy and/or trajectory, or the interaction induces the emission of a secondary particle(s). The particles that interact with the surface can be electrons, ions, photons or even heat. An analysis of the mass, energy and/or trajectory of the emitted particles, or the dependence of the emitted particle yield on a property of the incident particles, is used to infer information about the surface. Although these probes are indirect, they do provide reliable information about the surface composition and structure.

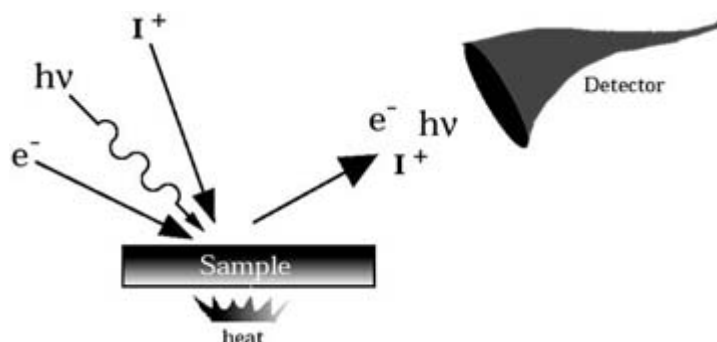


Figure A1.7.11. Schematic diagram of a generic surface science experiment. Particles, such as photons, electrons, or ions, are incident onto a solid surface, while the particles emitted from the surface are collected and measured by the detector.

Energetic particles interacting can also modify the structure and/or stimulate chemical processes on a surface. Absorbed particles excite electronic and/or vibrational (phonon) states in the near-surface region. Some surface scientists investigate the fundamental details of particle–surface interactions, while others are concerned about monitoring the changes to the surface induced by such interactions. Because of the importance of these interactions, the physics involved in both surface analysis and surface modification are discussed in this section.

The instrumentation employed for these studies is almost always housed inside a stainless-steel UHV chamber. One UHV chamber usually contains equipment for performing many individual techniques, each mounted on a different port, so that they can all be applied to the same sample. The sample is mounted onto a manipulator that allows for movement of the sample from one port to another, as well as for *in situ* heating and often cooling with liquid nitrogen (or helium). The chamber contains facilities for sample preparation, such as sputtering and annealing, as well as the possibility for gaseous exposures and/or film growth. Many instruments also contain facilities for the transfer of the

-26-

sample from one chamber to another while maintaining UHV. This allows for the incorporation of even more techniques, as well as the easy introduction of new samples into the chamber via a load-lock mechanism. Sample transfer into a reaction chamber also allows for the exposure of samples at high pressures or with corrosive gases or liquids that could not otherwise be introduced into a UHV chamber.

Below are brief descriptions of some of the particle–surface interactions important in surface science. The descriptions are intended to provide a basic understanding of how surfaces are probed, as most of the information that we have about surfaces was obtained through the use of techniques that are based on such interactions. The section is divided into some general categories, and the important physics of the interactions used for analysis are emphasized. All of these techniques are described in greater detail in subsequent sections of the encyclopaedia. Also, note that there are many more techniques than just those discussed here. These particular techniques were chosen not to be comprehensive, but instead to illustrate the kind of information that can be obtained from surfaces and interfaces.

A1.7.5.1 ELECTRON SPECTROSCOPY

Electrons are extremely useful as surface probes because the distances that they travel within a solid before scattering are rather short. This implies that any electrons that are created deep within a sample do not escape into vacuum. Any technique that relies on measurements of low-energy electrons emitted from a solid therefore provides information from just the outermost few atomic layers. Because of this inherent surface sensitivity, the various electron spectroscopies are probably the most useful and popular techniques in surface science.

Electrons interact with solid surfaces by elastic and inelastic scattering, and these interactions are employed in electron spectroscopy. For example, electrons that elastically scatter will diffract from a single-crystal lattice. The diffraction pattern can be used as a means of structural determination, as in LEED. Electrons scatter inelastically by inducing electronic and vibrational excitations in the surface region. These losses form the basis of electron energy loss spectroscopy (EELS). An incident electron can also knock out an inner-shell, or core, electron from an atom in the solid that will, in turn, initiate an Auger process. Electrons can also be used to induce stimulated desorption, as described in [section A1.7.5.6](#).

[Figure A1.7.12](#) shows the scattered electron kinetic energy distribution produced when a monoenergetic electron beam is incident on an Al surface. Some of the electrons are elastically backscattered with essentially

no energy loss, as evidenced by the elastic peak. Others lose energy inelastically, however, by inducing particular excitations in the solid, but are then emitted from the surface by elastic backscattering. The plasmon loss features seen in [figure A1.7.12](#) represent scattered electrons that have lost energy inelastically by excitation of surface plasmons. A plasmon is a collective excitation of substrate electrons, and a single plasmon excitation typically has an energy in the range of 5–25 eV. A small feature due to the emission of Auger electrons is also seen in the figure. Finally, the largest feature in the spectrum is the inelastic tail. The result of all of the electronic excitations is the production of a cascade of secondary electrons that are ejected from the surface. The intensity of the secondary electron ‘tail’ increases as the kinetic energy is reduced, until the cutoff energy is reached. The exact position of the cutoff is determined by the surface work function, and, in fact, is often used to measure the work function changes as the surface composition is modified.

-27-

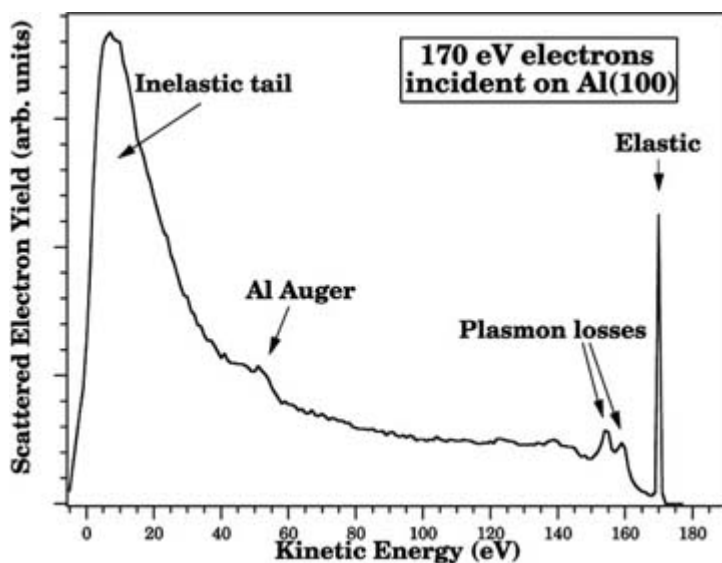


Figure A1.7.12. Secondary electron kinetic energy distribution, obtained by measuring the scattered electrons produced by bombardment of Al(100) with a 170 eV electron beam. The spectrum shows the elastic peak, loss features due to the excitation of plasmons, a signal due to the emission of Al LMM Auger electrons and the inelastic tail. The exact position of the cutoff at 0 eV depends on the surface work function.

The inelastic mean free path (IMFP) is often used to quantify the surface sensitivity of electron spectroscopy. The IMFP is the average distance that an electron travels through a solid before it is annihilated by inelastic scattering. The minimum in the IMFP for electrons travelling in a solid occurs just above the plasmon energy, as these electrons have the highest probability for excitation. Thus, for most materials, the electrons with the smallest mean free path are those with approximately 25–50 eV of kinetic energy [50]. When performing electron spectroscopy for quantitative analysis, it is necessary to define the mean escape depth (MED), rather than just use the IMFP [51]. The MED is the average depth below the surface from which electrons have originated, and includes losses by all possible elastic and inelastic mechanisms. Typical values of the MED for 10–1000 eV electrons are in the range of 4–10 Å, which is of the order of the interlayer spacings of a solid [52, 53]. Electron attenuation is modelled by assuming that the yield of electrons originating from a particular depth within the sample decreases exponentially with increasing depth, i.e.,

$$\text{Number of electrons} = \exp(-d/\lambda). \tag{A1.7.6}$$

Where λ is the MED for the particular material and d is the distance below the surface from which the

electron originated. This consideration allows measurements of depth distributions by changing either the electron kinetic energy or the emission angle in order to vary λ .

A popular electron-based technique is Auger electron spectroscopy (AES), which is described in [section B1.25.2.2](#). In AES, a 3–5 keV electron beam is used to knock out inner-shell, or core, electrons from atoms in the near-surface region of the material. Core holes are unstable, and are soon filled by either fluorescence or Auger decay. In the Auger

-28-

process, one valence, or lower-lying core, electron fills the hole while another is emitted from the sample, in order to satisfy conservation of energy. The emitted Auger electrons have kinetic energies that are characteristic of a particular element. The Perkin–Elmer Auger handbook contains sample spectra of each element, along with information on the relative sensitivity of each Auger line [54]. AES is most useful as a quantitative measure of the surface atomic composition, and is a standard technique employed to determine sample cleanliness. The ratio of the AES signal from an adsorbate to that of the substrate is also commonly used to quantify the coverage of an adsorbate.

LEED is used primarily to ascertain the crystallinity and symmetry of a single-crystal surface, but can also be used to obtain detailed structural information [55, 56]. LEED is described in detail in [section B1.21](#). In LEED, a 20–200 eV electron beam is incident upon a single-crystal surface along the sample normal. The angular distribution of the elastically scattered electrons is then measured, usually by viewing a phosphorescent screen. At certain angles, there are spots that result from the diffraction of electrons. The symmetry of the pattern of spots is representative of the two-dimensional unit cell of the surface. Note, however, that the spacings between LEED spots provide distances in inverse space, i.e. more densely packed LEED spots correspond to larger surface unit cells. The sharpness of the spots is an indication of the average size of the ordered domains on the surface. In order to extract detailed atomic positions from LEED, the intensity of the spots as a function of the electron energy, or intensity–voltage (I – V) curves, are collected and then compared to theoretical predictions for various surface structures [55, 56]. LEED I – V analysis is capable of providing structural details to an accuracy of 0.01 Å. LEED is probably the most accurate structural technique available, but it will only work for structures that are not overly complex.

The excitation of surface quanta can be monitored directly with EELS, as discussed in [section B1.7](#) and [section B1.25.5](#). In EELS, a monoenergetic electron beam is incident onto a surface and the kinetic energy distribution of the scattered electrons is collected. The kinetic energy distribution will display peaks corresponding to electrons that have lost energy by exciting transitions in the near-surface region, such as the plasmon loss peaks shown in [figure A1.7.12](#). EELS can be used to probe electronic transitions, in which case incident electron energies in the range of 10–100 eV are used. More commonly, however, EELS is used to probe low-energy excitations, such as molecular vibrations or phonon modes [57]. In this case, very low incident electron energies (<10 eV) are employed and a very high-energy resolution is required. When EELS is performed in this manner, the technique is known as high-resolution electron energy loss spectroscopy (HREELS).

Photoelectron spectroscopy provides a direct measure of the filled density of states of a solid. The kinetic energy distribution of the electrons that are emitted via the photoelectric effect when a sample is exposed to a monochromatic ultraviolet (UV) or x-ray beam yields a photoelectron spectrum. Photoelectron spectroscopy not only provides the atomic composition, but also information concerning the chemical environment of the atoms in the near-surface region. Thus, it is probably the most popular and useful surface analysis technique. There are a number of forms of photoelectron spectroscopy in common use.

X-ray photoelectron spectroscopy (XPS), also called electron spectroscopy for chemical analysis (ESCA), is described in [section B1.25.2.1](#). The most commonly employed x-rays are the Mg K α (1253.6 eV) and the Al K α (1486.6 eV) lines, which are produced from a standard x-ray tube. Peaks are seen in XPS spectra that correspond to the bound core-level electrons in the material. The intensity of each peak is proportional to the abundance of the emitting atoms in the near-surface region, while the precise binding energy of each peak depends on the chemical oxidation state and local environment of the emitting atoms. The Perkin–Elmer XPS handbook contains sample spectra of each element and binding energies for certain compounds [58].

-29-

XPS is also often performed employing synchrotron radiation as the excitation source [59]. This technique is sometimes called soft x-ray photoelectron spectroscopy (SXPS) to distinguish it from laboratory XPS. The use of synchrotron radiation has two major advantages: (1) a much higher spectral resolution can be achieved and (2) the photon energy of the excitation can be adjusted which, in turn, allows for a particular electron kinetic energy to be selected.

One of the more recent advances in XPS is the development of photoelectron microscopy [60]. By either focusing the incident x-ray beam, or by using electrostatic lenses to image a small spot on the sample, spatially-resolved XPS has become feasible. The limits to the spatial resolution are currently of the order of 1 μm , but are expected to improve. This technique has many technological applications. For example, the chemical makeup of micromechanical and microelectronic devices can be monitored on the scale of the device dimensions.

Ultraviolet photoelectron spectroscopy (UPS) is a variety of photoelectron spectroscopy that is aimed at measuring the valence band, as described in [section B1.25.2.3](#). Valence band spectroscopy is best performed with photon energies in the range of 20–50 eV. A He discharge lamp, which can produce 21.2 or 40.8 eV photons, is commonly used as the excitation source in the laboratory, or UPS can be performed with synchrotron radiation. Note that UPS is sometimes just referred to as photoelectron spectroscopy (PES), or simply valence band photoemission.

A particularly useful variety of UPS is angle-resolved photoelectron spectroscopy (ARPES), also called angle-resolved ultraviolet photoelectron spectroscopy (ARUPS) [61, 62]. In this technique, measurements are made of the valence band photoelectrons emitted into a small angle as the electron emission angle or photon energy is varied. This allows for the simultaneous determination of the kinetic energy and momentum of the photoelectrons with respect to the two-dimensional surface Brillouin zone. From this information, the electronic band structure of a single-crystal material can be experimentally determined.

The diffraction of photoelectrons (or Auger electrons) is also used as a structural tool [63, 64]. When electrons of a well defined energy are created at a particular atomic site, such as in XPS or AES, then the emitted electrons interact with other atoms in the crystal structure prior to leaving the surface. The largest effect is ‘forward scattering’, in which the intensity of an electron wave emitted from one atom is enhanced when it passes through another atom. Thus, the angular distribution of the emitted electron intensity provides a ‘map’ of the surface crystal structure. More generally, however, there is a complex multiple scattering behaviour, which produces variations of the emitted electron intensity with respect to both angle and energy such that the intensity modulations do not necessarily relate to the atomic bond directions. In order to determine a surface structure from such diffraction data, the measured angular and/or energy distributions of the Auger or photoelectrons is compared to a theoretical prediction for a given structure. Similar to LEED analysis, the structure employed for the calculation is varied until the best fit to the data is found.

A1.7.5.2 ION SPECTROSCOPY

Ions scattered from solid surfaces are useful probes for elemental identification of surface species and for measurements of the three-dimensional atomic structure of a single-crystal surface. Ions used for surface studies can be roughly divided into low (0.5–10 keV), medium (10–100 keV) and high (100 keV–1 MeV) energy regimes. In each regime, ions have distinct interactions with solid material and each regime is used for different types of measurements. The use of particle scattering for surface structure determination is described in detail in [section B1.23](#).

-30-

The fundamental interactions between ions and surfaces can be separated into elastic and inelastic processes. When an ion undergoes a direct collision with a single atom in a solid, it loses energy elastically by transferring momentum to the target atom. As an ion travels through a material, it also loses energy inelastically by initiating various electronic and vibrational excitations. The elastic and inelastic energy losses can usually be treated independently from each other.

Elastic losses result from binary collisions between the ions and unbound target atoms positioned at the lattice sites. For keV and higher energy ions, the cross sections for collisions are small enough that the ions essentially ‘see’ each atom in the solid individually, i.e. the trajectory can be considered as a sequence of events in which the ion interacts with one target atom at a time. This is the so-called binary collision approximation (BCA). The energy of a scattered particle is determined by conservation of energy and momentum during the single collision (the binding energy of the target atom to the surface can be neglected since it is considerably smaller than the energy of the ions). The smaller the mass of the target atom relative to the projectile, the more the energy that is lost during an elastic collision and the lower the scattered energy. Peaks are seen in scattered ion energy spectra, called single scattering peaks (SSP), or quasi-single (QS) scattering peaks, that result from these binary collisions. In this manner, ion scattering produces a mass spectrum of the surface region, as the position of each SSP indicates the mass of the target atom.

Ions in the low-energy range have reasonably short penetration depths, and therefore provide a surface-sensitive means for probing a material. Low-energy ion scattering (LEIS), often called ion scattering spectroscopy (ISS), is generally used as a measure of the surface composition. The surface sensitivity when using noble gas ions for standard ISS results from the high probability for neutralization for any ions that have penetrated past the first atomic layer. The intensity of an SSP is related to the surface concentration of the particular element, but care must be taken in performing quantitative analysis to properly account for ion neutralization. Energy losses due to inelastic excitations further modify the ion energies and charge states of scattered particles. In the low-energy regime, these effects are often neglected, as they only slightly alter the shapes of the SSP and shift it to a lower energy. In the high-energy regime, however, inelastic excitations are dominant in determining the shape of the scattered ion energy spectrum, as in Rutherford backscattering spectroscopy (RBS) [[65](#), [66](#)], which is discussed in [section B1.24](#).

Measurements of the angular distributions of scattered ions are often used as a structural tool, as they depend strongly on the relative positions of the atoms in the near-surface region. Ion scattering is used for structure determination by consideration of the shadow cones and blocking cones. These ‘cones’ are the regions behind each atom from which incoming ions are excluded because of scattering. A shadow cone is formed when an ion is incident onto the surface, while a blocking cone is formed when an ion that has scattered from a deeply-lying atom interacts with a surface atom along the outgoing trajectory. The ion flux is increased at the edges of the cones. Thus, rotating the ion beam or detector relative to the sample alters the flux of ions that scatter from any particular atom. The angular distributions are usually analysed by comparing the measured distributions to those obtained by computer simulation for a given geometry. Shadow/blocking cone analysis is used in both low- and medium-energy ion scattering to provide the atomic structure, and is accurate to about 0.1 Å [[67](#), [68](#)].

In the high-energy ion regime, ion channelling is used for surface structure determination [65, 66]. In this technique, the incident ion beam is aligned along a low-index direction in the crystal. Thus, most of the ions will penetrate into ‘channels’ created by the crystal structure. Those few ions that do backscatter from a surface atom are collected. The number of these scattering events is dependent on the detailed atomic structure. For performing a structure determination, the data is usually collected as ‘rocking curves’ in which the backscattered ion yield is collected as the crystal is precisely rotated about the channelling direction. The measured rocking curves are then compared to the

-31-

results of computer simulations performed for particular model surface structures. As in LEED I - V analysis, the structure employed for the simulation that most closely matches the experimental data is deemed to be correct.

Ions are also used to initiate secondary ion mass spectrometry (SIMS) [69], as described in [section B1.25.3](#). In SIMS, the ions sputtered from the surface are measured with a mass spectrometer. SIMS provides an accurate measure of the surface composition with extremely good sensitivity. SIMS can be collected in the ‘static’ mode in which the surface is only minimally disrupted, or in the ‘dynamic’ mode in which material is removed so that the composition can be determined as a function of depth below the surface. SIMS has also been used along with a shadow and blocking cone analysis as a probe of surface structure [70].

A1.7.5.3 SCANNING PROBE METHODS

Scanning probe microscopies have become the most conspicuous surface analysis techniques since their invention in the mid-1980s and the awarding of the 1986 Nobel Prize in Physics [71, 72]. The basic idea behind these techniques is to move an extremely fine tip close to a surface and to monitor a signal as a function of the tip’s position above the surface. The tip is moved with the use of piezoelectric materials, which can control the position of a tip to a sub-Ångström accuracy, while a signal is measured that is indicative of the surface topography. These techniques are described in detail in [section B1.20](#).

The most popular of the scanning probe techniques are STM and atomic force microscopy (AFM). STM and AFM provide images of the outermost layer of a surface with atomic resolution. STM measures the spatial distribution of the surface electronic density by monitoring the tunnelling of electrons either from the sample to the tip or from the tip to the sample. This provides a map of the density of filled or empty electronic states, respectively. The variations in surface electron density are generally correlated with the atomic positions. AFM measures the spatial distribution of the forces between an ultrafine tip and the sample. This distribution of these forces is also highly correlated with the atomic structure. STM is able to image many semiconductor and metal surfaces with atomic resolution. AFM is necessary for insulating materials, however, as electron conduction is required for STM in order to achieve tunnelling. Note that there are many modes of operation for these instruments, and many variations in use. In addition, there are other types of scanning probe microscopies under development.

Scanning probe microscopies have afforded incredible insight into surface processes. They have provided visual images of surfaces on the atomic scale, from which the atomic structure can be observed in real time. All of the other surface techniques discussed above involve averaging over a macroscopic region of the surface. From STM images, it is seen that many surfaces are actually not composed of an ideal single domain, but rather contain a mixture of domains. STM has been able to provide direct information on the structure of atoms in each domain, and at steps and defects on surfaces. Furthermore, STM has been used to monitor the movement of single atoms on a surface. Refinements to the instruments now allow images to be collected over temperatures ranging from 4 to 1200 K, so that dynamical processes can be directly investigated. An

STM has also been adapted for performing single-atom vibrational spectroscopy [73].

One of the more interesting new areas of surface science involves manipulation of adsorbates with the tip of an STM. This allows for the formation of artificial structures on a surface at the atomic level. In fact, STM tips are being investigated for possible use in lithography as part of the production of very small features on microcomputer chips [74].

-32-

Some of the most interesting work in this area has involved physisorbed molecules at temperatures as low as 4 K [75]. Note that it takes a specialized instrument to be able to operate at these low temperatures. An STM tip is brought into contact with the physisorbed species by lightly pushing down on it. Then, the STM tip is translated parallel to the surface while pressure is maintained on the adsorbate. In this manner, the adsorbates can be moved to any location on the surface. Manipulation of this type has led to the writing of 'IBM' with single atoms [76], as well as to the formation of structures such as the 'quantum corral' [77]. The quantum corral is so named, as it is an oval-shaped enclosure made from adsorbate atoms that provides a barrier for the free electrons of the metal substrate. Inside the corral, standing wave patterns are set up that can be imaged with the STM.

There are many other experiments in which surface atoms have been purposely moved, removed or chemically modified with a scanning probe tip. For example, atoms on a surface have been induced to move via interaction with the large electric field associated with an STM tip [78]. A scanning force microscope has been used to create three-dimensional nanostructures by 'pushing' adsorbed particles with the tip [79]. In addition, the electrons that are tunnelling from an STM tip to the sample can be used as sources of electrons for stimulated desorption [80]. The tunnelling electrons have also been used to promote dissociation of adsorbed O₂ molecules on metal or semiconductor surfaces [81, 82].

A1.7.5.4 THERMAL DESORPTION

Temperature programmed desorption (TPD), also called thermal desorption spectroscopy (TDS), provides information about the surface chemistry such as surface coverage and the activation energy for desorption [49]. TPD is discussed in detail in section B1.25. In TPD, a clean surface is first exposed to a gaseous molecule that adsorbs. The surface is then quickly heated (on the order of 10 K s⁻¹), while the desorbed molecules are measured with a mass spectrometer. An analysis of TPD spectra basically provides three types of information: (1) The identities of the desorbed product(s) are obtained directly from the mass spectrometer. (2) The area of a TPD peak provides a good measure of the surface coverage. In cases where there are multiple species desorbed, the ratios of the TPD peaks provide the stoichiometry. (3) The shapes of the peaks, and how they change with surface coverage, provide detailed information on the kinetics of desorption. For example, the shapes of TPD curves differ for zeroth-, first- or second-order processes.

A1.7.5.5 LASER-SURFACE INTERACTIONS

Lasers have been used to both modify and probe surfaces. When operated at low fluxes, lasers can excite electronic and vibrational states, which can lead to photochemical modification of surfaces. At higher fluxes, the laser can heat the surface to extremely high temperatures in a region localized at the very surface. A high-power laser beam produces a very non-equilibrium situation in the near-surface region, during which the effective electron temperature can be extremely high. Thus, lasers can also be used to initiate thermal desorption. Laser-induced thermal desorption (LITD) has some advantages over TPD as an analytical technique [36]. When a laser is used to heat the surface, the heat is localized in the surface region and the temperature rise is extremely fast. It is also possible to produce excitations that involve multiple photons

because of the high flux available with lasers. Furthermore, there are nonlinear effects that occur with laser irradiation of surfaces that allow for surface sensitive probes that do not require UHV, such as second harmonic generation (SHG) and sum frequency generation (SFG) [83, 84]. Optical techniques in surface science are discussed in [section B1.22](#).

-33-

Surface photochemistry can drive a surface chemical reaction in the presence of laser irradiation that would not otherwise occur. The types of excitations that initiate surface photochemistry can be roughly divided into those that occur due to direct excitations of the adsorbates and those that are mediated by the substrate. In a direct excitation, the adsorbed molecules are excited by the laser light, and will directly convert into products, much as they would in the gas phase. In substrate-mediated processes, however, the laser light acts to excite electrons from the substrate, which are often referred to as ‘hot electrons’. These hot electrons then interact with the adsorbates to initiate a chemical reaction.

Femtosecond lasers represent the state-of-the-art in laser technology. These lasers can have pulse widths of the order of 100 fs. This is the same time scale as many processes that occur on surfaces, such as desorption or diffusion. Thus, femtosecond lasers can be used to directly measure surface dynamics through techniques such as two-photon photoemission [85]. Femtochemistry occurs when the laser imparts energy over an extremely short time period so as to directly induce a surface chemical reaction [86].

A1.7.5.6 STIMULATED DESORPTION

An electron or photon incident on a surface can induce an electronic excitation. When the electronic excitation decays, an ion or neutral particle can be emitted from the surface as a result of the excitation. Such processes are known as desorption induced by electronic transitions (DIET) [87]. The specific techniques are known as electron-stimulated desorption (ESD) and photon-stimulated desorption (PSD), depending on the method of excitation.

A DIET process involves three steps: (1) an initial electronic excitation, (2) an electronic rearrangement to form a repulsive state and (3) emission of a particle from the surface. The first step can be a direct excitation to an antibonding state, but more frequently it is simply the removal of a bound electron. In the second step, the surface electronic structure rearranges itself to form a repulsive state. This rearrangement could be, for example, the decay of a valence band electron to fill a hole created in step (1). The repulsive state must have a sufficiently long lifetime that the products can desorb from the surface before the state decays. Finally, during the emission step, the particle can interact with the surface in ways that perturb its trajectory.

There are two main theoretical descriptions applied to stimulated desorption. The Menzel–Gomer–Redhead (MGR) model is used to describe low-energy valence excitations, while the Knotek–Feibelman mechanism is used to describe a type of desorption that occurs with ionically-bound species. In the MGR model, it is assumed that the initial excitation occurs by absorption of a photon or electron to directly create an excited, repulsive state. This excited state can be neutral or ionic. It simply needs to have a sufficient lifetime so that desorption can occur before the system relaxes to the ground state. Thus, the MGR mechanism can be applied to positive or negative ion emission, or to the emission of a neutral atom. The Knotek–Feibelman mechanism applies when there is an ionic bond at the surface. In this case, the incident electron kicks out an inner-shell electron, and an Auger process then fills the resulting core hole. In the Auger process, one electron drops down to fill the hole, while another electron is emitted from the surface in order to satisfy conservation of energy. Thus, the system has lost at least two electrons, which is sufficient to turn the negatively charged anion into a positive ion. Finally, Coulomb repulsion between this positive ion and the cation leads to the emission of a positive ion from the surface. Although this mechanism was originally proposed for maximally

valent bonding, it has since been observed to occur in a variety of systems providing that there is at least a moderate amount of charge transfer involved in the bonding. Note that this mechanism is often referred to as Auger-stimulated desorption (ASD).

-34-

Electron stimulated desorption angular distributions (ESDIAD) [88] provide a quick measure of the bond angles for a lightly bound adsorbate. ESDIAD patterns are recorded by impinging an electron beam onto a surface and then measuring the angular distributions of the desorbed ions with an imaging analyser. The measured ion emission angles are related to the original surface bond angles. The initial excitation responsible for ESD is normally directly along the bond axis. As an ion is exiting from a surface, however, there are two effects that act to alter the ion's trajectory. First, the ion is attracted to its image charge, which tends to spread out the ESDIAD pattern. Second, however, is that there is inhomogeneous neutralization of the emitted ions, in that the ions emitted at more grazing angles are preferentially neutralized. This acts to compress the observed pattern. Thus, a balance between these competing effects produces the measured angular distribution, and it is therefore difficult, although not impossible, to quantitatively determine the bond angle.

The ESDIAD pattern does, however, provide very useful information on the nature and symmetry of an adsorbate. As an example, figure A1.7.13(a) shows the ESDIAD pattern of desorbed F^+ collected from a 0.25 ML coverage of PF_3 on Ru(0001) [89]. The F^+ pattern displays a ring of emission, which indicates that the molecule adsorbs intact and is bonded through the P end. It freely rotates about the P–Ru bond so that the F^+ emission occurs at all azimuthal angles, regardless of the substrate structure. In figure A1.7.13(b), the ESDIAD pattern is shown following sufficient e^- -beam damage to remove much of the fluorine and produce adsorbed PF_2 and PF. Now, the F^+ emission shows six lobes along particular azimuths and one lobe along the surface normal. The off-normal lobes arise from PF_2 , and indicate that PF_2 adsorbs in registry with the substrate, with the F atoms pointing away from the surface at an off-normal angle. The centre lobe arises from PF and indicates that the PF moiety is bonded through the P end, with F pointing normal to the surface.

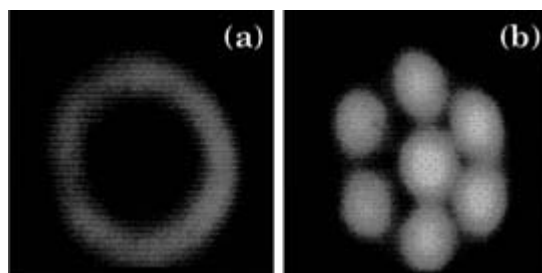


Figure A1.7.13. ESDIAD patterns showing the angular distributions of F^+ emitted from PF_3 adsorbed on Ru(0001) under electron bombardment. (a) 0.25 ML coverage, (b) the same surface following electron beam damage.

Some recent advances in stimulated desorption were made with the use of femtosecond lasers. For example, it was shown by using a femtosecond laser to initiate the desorption of CO from Cu while probing the surface with SHG, that the entire process is completed in less than 325 fs [90]. The mechanism for this kind of laser-induced desorption has been termed desorption induced by multiple electronic transitions (DIMET) [91]. Note that the mechanism must involve a multiphoton process, as a single photon at the laser frequency has insufficient energy to directly induce desorption. DIMET is a modification of the MGR mechanism in which each photon excites the adsorbate to a higher vibrational level, until a sufficient amount of vibrational energy has been amassed so that the particle can escape the surface.

A1.7.6 LIQUID–SOLID INTERFACE

One of the less explored frontiers in atomic-scale surface science is the study of the liquid–solid interface. This interface is critically important in many applications, as well as in biological systems. For example, the movement of pollutants through the environment involves a series of chemical reactions of aqueous groundwater solutions with mineral surfaces. Although the liquid–solid interface has been studied for many years, it is only recently that the tools have been developed for interrogating this interface at the atomic level. This interface is particularly complex, as the interactions of ions dissolved in solution with a surface are affected not only by the surface structure, but also by the solution chemistry and by the effects of the electrical double layer [31]. It has been found, for example, that some surface reconstructions present in UHV persist under solution, while others do not.

The electrical double layer basically acts as a capacitor by storing charge at the surface that is balanced by ions in solution [92]. The capacitance of the double layer is a function of the electrochemical potential of the solution, and has a maximum at the potential of zero-charge (pzc). The pzc in solution is essentially equivalent to the work function of that surface in vacuum. In solution, however, the electrode potential can be used to vary the surface charge in much the same way that alkali adsorbates are used to vary the work function of a surface in vacuum. The difference is that in solution the surface charge can be varied, while the surface composition is unchanged. The surface energy, which effects the atomic structure and reactivity, is directly related to the surface charge. It has been shown, for example, that by adjusting the electrode potential the reconstructions of certain surfaces in solution can be altered in a reversible manner. Electrochemistry can also be used to deposit and remove adsorbates from solution in a manner that is controlled by the electrode potential.

Studies of the liquid–solid interface can be divided into those that are performed *ex situ* and those performed *in situ*. In an *ex situ* experiment, a surface is first reacted in solution, and then removed from the solution and transferred into a UHV spectrometer for measurement. There has recently been, however, much work aimed at interrogating the liquid–solid interface *in situ*, i.e. while chemistry is occurring rather than after the fact.

In performing *ex situ* surface analysis, the transfer from solution to the spectrometer sometimes occurs either through the air or within a glove bag filled with an inert atmosphere. Many *ex situ* studies of chemical reactions at the liquid–solid interface, however, have been carried out using special wet cells that are directly attached to a UHV chamber [93, 94]. With this apparatus, the samples can be reacted and then immediately transferred to UHV without encountering air. Note that some designs enable complete immersion of the sample into solution, while others only allow the sample surface to interact with a meniscus. Although these investigations do not probe the liquid–solid interface directly, they can provide much information on the surface chemistry that has taken place.

One of the main uses of these wet cells is to investigate surface electrochemistry [94, 95]. In these experiments, a single-crystal surface is prepared by UHV techniques and then transferred into an electrochemical cell. An electrochemical reaction is then run and characterized using cyclic voltammetry, with the sample itself being one of the electrodes. In order to be sure that the electrochemical measurements all involved the same crystal face, for some experiments a single-crystal cube was actually oriented and polished on all six sides! Following surface modification by electrochemistry, the sample is returned to UHV for

measurement with standard techniques, such as AES and LEED. It has been found that the chemisorbed layers that are deposited by electrochemical reactions are stable and remain adsorbed after removal from solution. These studies have enabled the determination of the role that surface structure plays in electrochemistry.

The force between two adjacent surfaces can be measured directly with the surface force apparatus (SFA), as described in [section B1.20 \[96\]](#). The SFA can be employed in solution to provide an *in situ* determination of the forces. Although this instrument does not directly involve an atomically resolved measurement, it has provided considerable insight into the microscopic origins of surface friction and the effects of electrolytes and lubricants [\[97\]](#).

Scanning probe microscopies are atomically resolved techniques that have been successfully applied to measurements of the liquid–solid interface *in situ* [\[98, 99, 100, 101 and 102\]](#). The STM has provided atomically resolved images of surface reconstructions and adsorption geometry under controlled conditions in solution, and the dependence of these structures on solution composition and electrode potential. Note that in order to perform STM under solution, a special tip coated with a dielectric must be used in order to reduce the Faradaic current that would otherwise transmit through the solution. As an example, [figure A1.7.14](#) shows an STM image collected in solution from docosanol physisorbed on a graphite surface. The graphite lattice and the individual atoms in the adsorbed molecules can be imaged with atomic resolution. In addition, scanning probe microscopies have been used to image the surfaces of biological molecules and even living cells in solution [\[103\]](#).

-37-

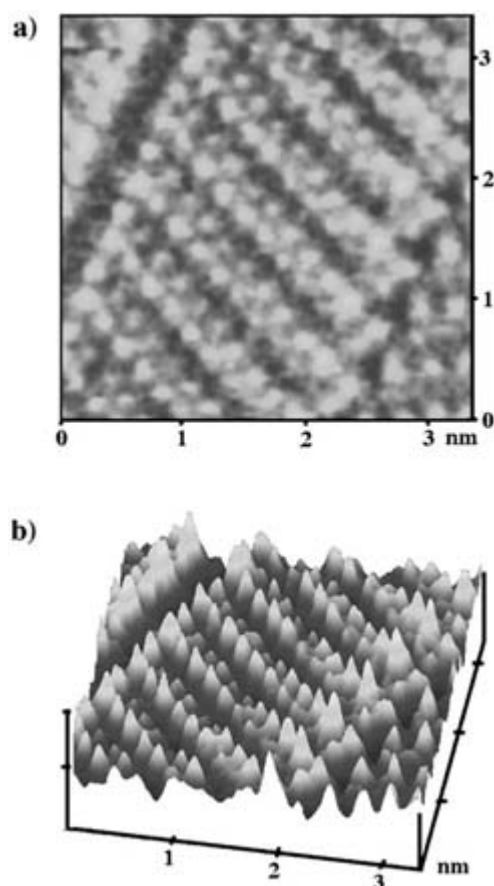


Figure A1.7.14. 3.4 nm × 3.4 nm STM images of 1-docosanol physisorbed onto a graphite surface in solution. This image reveals the hydrogen-bonding alcohol molecules assembled in lamellar fashion at the liquid–solid interface. Each ‘bright’ circular region is attributed to the location of an individual hydrogen

atom protruding upward out of the plane of the all-trans hydrocarbon backbone, which is lying flat on the surface. (a) Top view, and (b) a perspective image (courtesy of Leanna Giancarlo and George Flynn).

Since water is transparent to visible light, optical techniques can be used to interrogate the liquid–solid interface *in situ* [104]. For example, SFG has been used to perform IR spectroscopy directly at the liquid–solid interface [105, 106]. The surface sensitivity of SFG arises from the breaking of centrosymmetry at the interface, rather than from electron attenuation as in more traditional surface techniques, so that the information obtained is relevant to atomic-scale processes at the solid–liquid interface. This allows for the identification of the adsorbed species while a reaction is occurring. Note that these techniques can be extended to the liquid–liquid interface, as well [107]. In addition, x-ray scattering employing synchrotron radiation is being developed for use at the liquid–solid interface. For example, an *in situ* electrochemical cell for x-ray scattering has been designed [108].

REFERENCES

- [1] Somorjai G A 1994 *Introduction to Surface Chemistry and Catalysis* (New York: Wiley)
- [2] Kittel C 1996 *Introduction to Solid State Physics* 7th edn (New York: Wiley)
- [3] Himpsel F J, Jung T and Ortega J E 1997 Nanowires on stepped metal surfaces *Surf. Rev. Lett.* **4** 371
- [4] Noonan J R and Davis H L 1984 Truncation-induced multilayer relaxation of the Al(110) surface *Phys. Rev. B* **29** 4349
- [5] Adams D L, Jensen V, Sun X F and Vollesen J H 1988 Multilayer relaxation of the Al(210) surface *Phys. Rev. B* **38** 7913
- [6] Holub-Krappe E, Horn K, Frenken J W M, Krans R L and van der Veen J F 1987 Multilayer relaxation at the Ag(110) surface *Surf. Sci.* **188** 335
- [7] Busch B W and Gustafsson T 1998 Oscillatory relaxation of Al(110) reinvestigated by using medium-energy ion scattering *Surf. Sci.* **415** L1074
- [8] Cho J-H, Ismail, Zhang Z and Plummer E W 1999 Oscillatory lattice relaxation at metal surfaces *Phys. Rev. B* **59** 1677
- [9] Himpsel F J, McFeely F R, Morar J F, Taleb-Ibrahimi A and Yarmoff J A 1990 Core level spectroscopy at silicon surfaces and interfaces *Proc. Enrico Fermi School on 'Photoemission and Adsorption Spectroscopy and Interfaces with Synchrotron Radiation'* vol course CVIII, eds M Campagna and R Rosei (Amsterdam: Elsevier) p 203
- [10] Hamers R J, Tromp R M and Demuth J M 1986 Surface electronic structure of Si(111)-7 × 7 resolved in real space *Phys. Rev. Lett.* **56** 1972
- [11] Takayanagi K, Tanishiro Y, Takahashi M and Takahashi S 1985 Structural analysis of Si(111)-7 × 7 by UHV-transmission electron diffraction and microscopy *J. Vac. Sci. Technol. A* **3** 1502
- [12] Liu C L, Cohen J M, Adams J B and Voter A F 1991 EAM study of surface self-diffusion of single adatoms of fcc metals Ni, Cu, Al, Ag, Au, Pd, and Pt *Surf. Sci.* **253** 334
- [13] Bonig L, Liu S and Metiu H 1996 An effective medium theory study of Au islands on the Au(100) surface: reconstruction, adatom diffusion, and island formation *Surf. Sci.* **365** 87
- [14] Tsong T T 1988 Experimental studies of the behaviour of single adsorbed atoms on solid surfaces *Rep. Prog. Phys.* **51** 759
- [15] Chen C-L and Tsong T T 1991 Self-diffusion on the reconstructed and nonreconstructed Ir(110) surfaces *Phys. Rev. Lett.* **66** 1610

- [16] Jensen F, Besenbacher F, Laesgaard E and Stensgaard I 1990 Surface reconstruction of Cu (110) induced by oxygen chemisorption *Phys. Rev. B* **41** 10 233
- [17] Besenbacher F, Jensen F, Laegsgaard E, Mortensen K and Stensgaard I 1991 Visualization of the dynamics in surface reconstructions *J. Vac. Sci. Technol. B* **9** 874
-

-39-

- [18] Kitamura N, Lagally M G and Webb M B 1993 Real-time observations of vacancy diffusion on Si(100)-(2 × 1) by scanning tunneling microscopy *Phys. Rev. Lett.* **71** 2082
- [19] Linderoth T R, Horsch S, Laesgaard E, Stensgaard I and Besenbacher F 1997 Surface diffusion of Pt on Pt(110): Arrhenius behavior of long jumps *Phys. Rev. Lett.* **78** 4978
- [20] Botkin D, Glass J, Chemla D S, Ogletree D F, Salmeron M and Weiss S 1996 Advances in ultrafast scanning tunneling microscopy *Appl. Phys. Lett.* **69** 1321
- [21] Bauer E 1994 Low energy electron microscopy *Rep. Prog. Phys.* **57** 895
- [22] Himpsel F J 1990 Inverse photoemission from semiconductors *Surf. Sci. Rep.* **12** 1
- [23] Himpsel F J 1991 Unoccupied electronic states at surfaces *Surface Physics and Related Topics. Festschrift for Xide Xie* ed F-J Yang, G-J Ni, X Wang, K-M Zhang and D Lu (Singapore: World Scientific) p 179
- [24] Avouris P and Wolkow R 1989 Atom-resolved chemistry studied by scanning tunneling microscopy and spectroscopy *Phys. Rev. B* **39** 5091
- [25] Lüth H 1995 *Surfaces and Interfaces of Solid Materials* 3rd edn (Berlin: Springer)
- [26] Borroni-Bird C E, Al-Sarraf N, Andersson S and King D A 1991 Single crystal adsorption microcalorimetry *Chem. Phys. Lett.* **183** 516
- [27] Li Y L *et al* 1995 Experimental verification of a new mechanism for dissociative chemisorption: atom abstraction *Phys. Rev. Lett.* **74** 2603
- [28] Jensen J A, Yan C and Kummel A C 1996 Direct chemisorption site selectivity for molecular halogens on the Si(111)-(7 × 7) surface *Phys. Rev. Lett.* **76** 1388
- [29] Hudson J B 1992 *Surface Science: An Introduction* (Boston: Butterworth-Heinemann)
- [30] Kang H C and Weinberg W H 1994 Kinetic modeling of surface rate processes *Surf. Sci.* **299–300** 755
- [31] Adamson A W and Gast A P 1997 *Physical Chemistry of Surfaces* 6th edn (New York: Wiley-Interscience)
- [32] Over H 1998 Crystallographic study of interaction between adspecies on metal surfaces *Prog. Surf. Sci.* **58** 249
- [33] Gomer R 1990 Diffusion of adsorbates on metal surfaces *Rep. Prog. Phys.* **53** 917
- [34] Lagally M G 1993 Atom motion on surfaces *Physics Today* **46** 24
- [35] Dunphy J C, Sautet P, Ogletree D F, Dabbousi O and Salmeron M B 1993 Scanning-tunneling-microscopy study of the surface diffusion of sulfur on Re(0001) *Phys. Rev. B* **47** 2320
- [36] George S M, DeSantolo A M and Hall R B 1985 Surface diffusion of hydrogen on Ni(100) studied using laser-induced thermal desorption *Surf. Sci.* **159** L425
- [37] Olmstead M A, Bringans R D, Uhrberg R I G and Bachrach R Z 1986 Arsenic overlayer on Si(111): removal of surface reconstruction *Phys. Rev. B* **34** 6041
-

-40-

- [38] Yarmoff J A, Cyr D M, Huang J H, Kim S and Williams R S 1986 Impact-collision ion-scattering spectroscopy of Cu(110) and Cu(110)-(2 × 1)-O using 5-keV ⁶Li⁺ *Phys. Rev. B* **33** 3856
- [39] Tochihara H and Mizuno S 1998 Composite surface structures formed by restructuring-type adsorption of alkali-metals on FCC metals *Prog. Surf. Sci.* **58** 1
- [40] Diehl R D and McGrath R 1996 Structural studies of alkali metal adsorption and coadsorption on metal surfaces *Surf. Sci. Rep.* **23** 43
- [41] Madey T E, Guan J, Nien C-H, Dong C-Z, Tao H-S and Campbell R A 1996 Faceting induced by ultrathin metal films on W(111) and Mo(111): structure, reactivity, and electronic properties *Surf. Rev. Lett.* **3** 1315
- [42] Bonzel H P, Bradshaw A M and Ertl G 1989 *Physics and Chemistry of Alkali Metal Adsorption* (Amsterdam: Elsevier)
- [43] Winters H F and Coburn J W 1992 Surface science aspects of etching reactions *Surf. Sci. Rep.* **14** 161
- [44] Herman M A and Sitter H 1996 *Molecular Beam Epitaxy: Fundamentals and Current Status* (Berlin: Springer)
- [45] Somorjai G A 1981 *Chemistry in Two Dimensions: Surfaces* (Ithaca: Cornell University Press)
- [46] Somorjai G A 1996 Surface science at high pressures *Z. Phys. Chem.* **197** 1
- [47] Somorjai G A 1998 Molecular concepts of heterogeneous catalysis *J. Mol. Struct. (Theochem)* **424** 101
- [48] Wood E A 1963 *Crystal Orientation Manual* (New York: Columbia University Press)
- [49] Woodruff D P and Delchar T A 1994 *Modern Techniques of Surface Science* 2nd edn (Cambridge: Cambridge University Press)
- [50] Seah M P and Dench W A 1979 Quantitative electron spectroscopy of surfaces: a standard data base for electron inelastic mean free paths in solids *Surf. Interface Anal.* **1** 2
- [51] Powell C J, Jablonski A, Tilinin I S, Tanuma S and Penn D R 1999 Surface sensitivity of Auger-electron spectroscopy and x-ray photoelectron spectroscopy *J. Electron Spec. Relat. Phenom.* **98-9** 1
- [52] Duke C B 1994 Interaction of electrons and positrons with solids: from bulk to surface in thirty years *Surf. Sci.* **299-300** 24
- [53] Powell C J 1994 Inelastic interactions of electrons with surfaces: applications to Auger-electron spectroscopy and x-ray photoelectron spectroscopy *Surf. Sci.* **299-300** 34
- [54] Davis L E, MacDonald N C, Palmberg P W, Riach G E and Weber R E 1976 *Handbook of Auger Electron Spectroscopy* 2nd edn (Eden Prairie, MN: Perkin-Elmer Corporation)
- [55] Pendry J B 1974 *Low Energy Electron Diffraction: The Theory and its Application to Determination of Surface Structure* (London: Academic)
- [56] van Hove M A, Weinberg W H and Chan C-M 1986 *Low-Energy Electron Diffraction: Experiment, Theory, and Surface Structure Determination* (Berlin: Springer)
- [57] Ibach H and Mills D L 1982 *Electron Energy Loss Spectroscopy and Surface Vibrations* (New York: Academic)
- [58] Wagner C D, Riggs W M, Davis L E, Moulder J F and Muilenberg G E (eds) 1979 *Handbook of X-ray Photoelectron Spectroscopy* (Eden Prairie, MN: Perkin-Elmer Corporation)

- [59] Margaritondo G 1988 *Introduction to Synchrotron Radiation* (New York: Oxford University Press)
- [60] Tonner B P, Dunham D, Droubay T, Kikuma J, Denlinger J, Rotenberg E and Warwick A 1995 The development of electron spectromicroscopy *J. Electron Spectrosc.* **75** 309

- [61] Smith N V and Himpfel F J 1983 Photoelectron spectroscopy *Handbook on Synchrotron Radiation* ed E E Koch (Amsterdam: North-Holland)
- [62] Plummer E W and Eberhardt W 1982 Angle-resolved photoemission as a tool for the study of surfaces *Adv. Chem. Phys.* **49** 533
- [63] Egelhoff W F Jr 1990 X-ray photoelectron and Auger electron forward scattering: a new tool for surface crystallography *CRC Crit. Rev. Solid State Mater. Sci.* **16** 213
- [64] Fadley C S 1993 Diffraction and holography with photoelectrons and Auger electrons: some new directions *Surf. Sci. Rep.* **19** 231
- [65] Chu W-K, Mayer J W and Nicolet M-A 1978 *Backscattering Spectrometry* (New York: Academic)
- [66] Feldman L C, Mayer J W and Picraux S T 1982 *Materials Analysis by Ion Channeling: Submicron Crystallography* (New York: Academic)
- [67] Niehus H, Heiland W and Taglauer E 1993 Low-energy ion scattering at surfaces *Surf. Sci. Rep.* **17** 213
- [68] Fauster T 1988 Surface geometry determination by large-angle ion scattering *Vacuum* **38** 129
- [69] Benninghoven A, Rüdener F G and Werner H W 1987 *Secondary Ion Mass Spectrometry: Basic Concepts, Instrumental Aspects, Applications, and Trends* (New York: Wiley)
- [70] Chang C-C and Winograd N 1989 Shadow-cone-enhanced secondary-ion mass-spectrometry studies of Ag(110) *Phys. Rev. B* **39** 3467
- [71] Binnig G and Rohrer H 1987 Scanning tunneling microscopy—from birth to adolescence *Rev. Mod. Phys.* **59** 615
- [72] Wiesendanger R 1994 *Scanning Probe Microscopy and Spectroscopy: Methods and Applications* (New York: Cambridge University Press)
- [73] Stipe B C, Rezaei M A and Ho W 1998 Single-molecule vibrational spectroscopy and microscopy *Science* **280** 1732
- [74] Marrian C R K, Perkins F K, Brandow S L, Koloski T S, Dobisz E A and Calvert J M 1994 Low voltage electron beam lithography in self-assembled ultrathin films with the scanning tunneling microscope *Appl. Phys. Lett.* **64** 390
- [75] Stroscio J A and Eigler D M 1991 Atomic and molecular manipulation with the scanning tunneling microscope *Science* **254** 319
- [76] Eigler D M and Schweizer E K 1990 Positioning single atoms with a scanning tunneling microscope *Nature* **344** 524
- [77] Crommie M F, Lutz C P and Eigler D M 1993 Confinement of electrons to quantum corrals on a metal surface *Science* **262** 218
- [78] Boland J J 1993 Manipulating chlorine atom bonding on the Si(100)-(2 × 1) surface with the STM *Science* **262** 1703
- [79] Resch R, Baur C, Bugacov A, Koel B E, Madhukar A, Requicha A A G and Will P 1998 Building and manipulating three-dimensional and linked two-dimensional structures of nanoparticles using scanning force microscopy *Langmuir* **14** 6613

- [80] Shen T-C, Wang C, Abeln G C, Tucker J R, Lyding J W, Avouris P and Walkup R E 1995 Atomic-scale desorption through electronic and vibrational excitation mechanisms *Science* **268** 1590
- [81] Martel R, Avouris Ph and Lyo I-W 1996 Molecularly adsorbed oxygen species on Si(111)-(7 × 7): STM-induced dissociative attachment studies *Science* **272** 385
- [82] Stipe B C, Rezaei M A, Ho W, Gao S, Persson M and Lundqvist B I 1997 Single-molecule dissociation by tunneling electrons *Phys. Rev. Lett.* **78** 4410
- [83] Shen Y R 1994 Nonlinear optical studies of surfaces *Appl. Phys. A* **59** 541

- [84] Shen Y R 1994 Surfaces probed by nonlinear optics *Surf. Sci.* **299–300** 551
- [85] Petek H and Ogawa S 1997 Femtosecond time-resolved two-photon photoemission studies of electron dynamics in metals *Prog. Surf. Sci.* **56** 239
- [86] Her T-H, Finlay R J, Wu C and Mazur E 1998 Surface femtochemistry of CO/O₂/Pt(111): the importance of nonthermalized substrate electrons *J. Chem. Phys.* **108** 8595
- [87] Ramsier R D and Yates J T Jr 1991 Electron-stimulated desorption: principles and applications *Surf. Sci. Rep.* **12** 243
- [88] Madey T E 1986 Electron- and photon-stimulated desorption: probes of structure and bonding at surfaces *Science* **234** 316
- [89] Madey T E *et al* 1993 Structure and kinetics of electron beam damage in a chemisorbed monolayer: PF₃ on Ru(0001) *Desorption Induced by Electronic Transitions DIET V* vol 31, ed A R Burns, E B Stechel and D R Jennison (Berlin: Springer)
- [90] Prybyla J A, Tom H W K and Aumiller G D 1992 Femtosecond time-resolved surface reaction: desorption of Co from Cu(111) in <325 .fsec *Phys. Rev. Lett.* **68** 503
- [91] Misewich J A, Heinz T F and Newns D M 1992 Desorption induced by multiple electronic transitions *Phys. Rev. Lett.* **68** 3737
- [92] Kolb D M 1996 Reconstruction phenomena at metal–electrolyte interfaces *Prog. Surf. Sci.* **51** 109
- [93] Chusuei C C, Murrell T S, Corneille J S, Nooney M G, Vesecky S M, Hossner L R and Goodman D W 1999 Liquid reaction apparatus for surface analysis *Rev. Sci. Instrum.* **70** 2462
- [94] Soriaga M P 1992 Ultra-high vacuum techniques in the study of single-crystal electrode surfaces *Prog. Surf. Sci.* **39** 325
- [95] Hubbard A T 1990 Surface electrochemistry *Langmuir* **6** 97
- [96] Craig V S J 1997 An historical review of surface force measurement techniques *Colloids Surf. A: Physicochem. Eng. Aspects* **129–30** 75
- [97] Kumacheva E 1998 Interfacial friction measurements in surface force apparatus *Prog. Surf. Sci.* **58** 75
- [98] Itaya K 1998 *In situ* scanning tunneling microscopy in electrolyte solutions *Prog. Surf. Sci.* **58** 121
- [99] Cyr D M, Venkataraman B and Flynn G W 1996 STM investigations of organic molecules physisorbed at the liquid–solid interface *Chem. Mater.* **8** 1600

- [100] Drake B, Sonnenfeld R, Schneir J and Hansma P K 1987 Scanning tunneling microscopy of process at liquid–solid interfaces *Surf. Sci.* **181** 92
- [101] Giancarlo L C and Flynn G W 1988 Scanning tunneling and atomic force microscopy probes of self-assembled, physisorbed monolayers *Ann. Rev. Phys. Chem.* **49** 297
- [102] Schneir J, Harary H H, Dagata J A, Hansma P K and Sonnenfeld R 1989 Scanning tunneling microscopy and fabrication of nanometer scale structure at the liquid–gold interface *Scanning Microsc.* **3** 719
- [103] Vansteenkiste S O, Davies M C, Roberts C J, Tendler S J B and Williams P M 1998 Scanning probe microscopy of biomedical interfaces *Prog. Surf. Sci.* **57** 95
- [104] Iwasita T and Nart F C 1997 *In situ* infrared spectroscopy at electrochemical interfaces *Prog. Surf. Sci.* **55** 271
- [105] Raduge C, Pflumio V and Shen Y R 1997 Surface vibrational spectroscopy of sulfuric acid–water mixtures at the liquid–vapor interface *Chem. Phys. Lett.* **274** 140
- [106] Shen Y R 1998 Sum frequency generation for vibrational spectroscopy: applications to water interfaces and films of water and ice *Solid State Commun.* **108** 399

- [107] Gragson D E and Richmond G I 1998 Investigations of the structure and hydrogen bonding of water molecules at liquid surfaces by vibrational sum frequency spectroscopy *J. Phys. Chem.* **102** 3847
- [108] Koop T, Schindler W, Kazimirov A, Scherb G, Zegenhagen J, Schulz T, Feidenhans'l R and Kirschner J 1998 Electrochemical cell for *in situ* x-ray diffraction under ultrapure conditions *Rev. Sci. Instrum.* **69** 1840
-

A2.1 Classical thermodynamics

Robert L Scott

A2.1.1 INTRODUCTION

Thermodynamics is a powerful tool in physics, chemistry and engineering and, by extension, to substantially all other sciences. However, its power is narrow, since it says nothing whatsoever about time-dependent phenomena. It can demonstrate that certain processes are impossible, but it cannot predict whether thermodynamically allowed processes will actually take place.

It is important to recognize that thermodynamic laws are generalizations of experimental observations on systems of macroscopic size; for such bulk systems the equations are exact (at least within the limits of the best experimental precision). The validity and applicability of the relations are independent of the correctness of any model of molecular behaviour adduced to explain them. Moreover, the usefulness of thermodynamic relations depends crucially on *measurability*; unless an experimenter can keep the constraints on a system and its surroundings under control, the measurements may be worthless.

The approach that will be outlined here is due to Carathéodory [1] and Born [2] and should present fresh insights to those familiar only with the usual development in many chemistry, physics or engineering textbooks. However, while the formulations differ somewhat, the equations that finally result are, of course, identical.

A2.1.2 THE ZEROth LAW

A2.1.2.1 THE STATE OF A SYSTEM

First, a few definitions: a system is any region of space, any amount of material for which the boundaries are clearly specified. At least for thermodynamic purposes it must be of macroscopic size and have a topological integrity. It may not be only part of the matter in a given region, e.g. all the sucrose in an aqueous solution. A system could consist of two non-contiguous parts, but such a specification would rarely be useful.

To define the thermodynamic state of a system one must specify the values of a minimum number of variables, enough to reproduce the system with all its macroscopic properties. If special forces (surface effects, external fields—electric, magnetic, gravitational, etc) are absent, or if the bulk properties are insensitive to these forces, e.g. the weak terrestrial magnetic field, it ordinarily suffices—for a one-component system—to specify three variables, e.g. the temperature T , the pressure p and the number of moles n , or an equivalent set. For example, if the volume of a surface layer is negligible in comparison with the total volume, surface effects usually contribute negligibly to bulk thermodynamic properties.

In order to specify the size of the system, at least one of these variables ought to be *extensive* (one that is proportional to the size of the system, like n or the total volume V). In the special case of several phases in equilibrium several extensive properties, e.g. n and V for two phases, may be required to determine the relative amounts of the two phases. The rest of the variables can be *intensive* (independent of the size of the system) like T , p , the molar volume $\bar{V} = V/n$, or the density ρ . For multicomponent systems, additional variables, e.g. several n s, are needed to specify composition.

For example, the definition of a system as 10.0 g H₂O at 10.0°C at an applied pressure $p = 1.00$ atm is sufficient to specify that the water is liquid and that its other properties (energy, density, refractive index, even non-thermodynamic properties like the coefficients of viscosity and thermal conductivity) are uniquely fixed.

Although classical thermodynamics says nothing about time effects, one must recognize that nearly all thermodynamic systems are *metastable* in the sense that over long periods of time—much longer than the time to perform experiments—they may change their properties, e.g. perhaps by a very slow chemical reaction. Moreover, the time scale is merely relative; if a thermodynamic measurement can be carried out fast enough that it is finished before some other reaction can perturb the system, but slow enough for the system to come to internal equilibrium, it will be valid.

A2.1.2.2 WALLS AND EQUATIONS OF STATE

Of special importance is the nature of the boundary of a system, i.e. the wall or walls enclosing it and separating it from its *surroundings*. The concept of ‘surroundings’ can be somewhat ambiguous, and its thermodynamic usefulness needs to be clarified. It is not the rest of the universe, but only the external neighbourhood with which the system may interact. Moreover, unless this neighbourhood is substantially at internal equilibrium, its thermodynamic properties cannot be exactly specified. Examples of ‘surroundings’ are a thermostatic bath or the external atmosphere.

If neither matter nor energy can cross the boundary, the system is described as *isolated*; if only energy (but not matter) can cross the boundary, the system is *closed*; if both matter and energy can cross the boundary, the system is *open*.

(Sometimes, when defining a system, one must be careful to clarify whether the walls are part of the system or part of the surroundings. Usually the contribution of the wall to the thermodynamic properties is trivial by comparison with the bulk of the system and hence can be ignored.)

Consider two distinct closed thermodynamic systems each consisting of n moles of a specific substance in a volume V and at a pressure p . These two distinct systems are separated by an idealized wall that may be either *adiabatic* (heat-impermeable) or *diathermic* (heat-conducting). However, because the concept of heat has not yet been introduced, the definitions of adiabatic and diathermic need to be considered carefully. Both kinds of walls are impermeable to matter; a *permeable* wall will be introduced later.

If a system at equilibrium is enclosed by an adiabatic wall, the only way the system can be disturbed is by moving part of the wall; i.e. the only coupling between the system and its surroundings is by work, normally mechanical. (The adiabatic wall is an idealized concept; no real wall can prevent any conduction of heat over a long time. However, heat transfer must be negligible over the time period of an experiment.)

-3-

The diathermic wall is defined by the fact that two systems separated by such a wall cannot be at equilibrium at arbitrary values of their variables of state, p^α , V^α , p^β and V^β . (The superscripts are not exponents; they symbolize different systems, subsystems or phases; numerical subscripts are reserved for components in a mixture.) Instead there must be a relation between the four variables, which can be called an *equation of state*:

$$F(p^\alpha, V^\alpha, p^\beta, V^\beta) = 0. \quad (\text{A2.1.1})$$

Equation (A2.1.1) is essentially an expression of the concept of *thermal equilibrium*. Note, however, that, in this formulation, this concept precedes the notion of temperature.

To make the differences between the two kinds of walls clearer, consider the situation where both are ideal gases, each satisfying the ideal-gas law $pV = nRT$. If the two were separated by a diathermic wall, one would observe experimentally that $p^\alpha V^\alpha / p^\beta V^\beta = C$ where the constant C would be n^α / n^β . If the wall were adiabatic, the two pV products could be varied independently.

A2.1.2.3 TEMPERATURE AND THE ZEROth LAW

The concept of temperature derives from a fact of common experience, sometimes called the ‘zeroth law of thermodynamics’, namely, *if two systems are each in thermal equilibrium with a third, they are in thermal equilibrium with each other*. To clarify this point, consider the three systems shown schematically in figure A2.1.1, in which there are diathermic walls between systems α and γ and between systems β and γ , but an adiabatic wall between systems α and β .

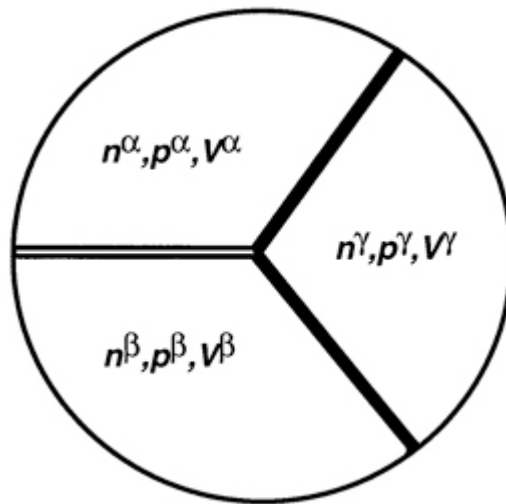


Figure A2.1.1. Illustration of the zeroth law. Three systems with two diathermic walls (solid) and one adiabatic wall (open).

-4-

Equation (A2.1.1) governs the diathermic walls, so one may write

$$F_A(p^\alpha, V^\alpha, p^\gamma, V^\gamma) = 0 \quad (\text{A2.1.2a})$$

$$F_B(p^\beta, V^\beta, p^\gamma, V^\gamma) = 0. \quad (\text{A2.1.2b})$$

It is a universal experimental observation, i.e. a ‘law of nature’, that the equations of state of systems 1 and 2 are then coupled as if the wall separating them were diathermic rather than adiabatic. In other words, there is a relation

$$F_C(p^\alpha, V^\alpha, p^\beta, V^\beta) = 0. \quad (\text{A2.1.2c})$$

It may seem that equation (A2.1.2c) is just a mathematical consequence of equation (2.1.2a) and equation (2.1.2b), but it is not; it conveys new physical information. If one rewrites equation (2.1.2a) and equation (2.1.2b) in the form

$$p^\gamma = \phi_\alpha(p^\alpha, V^\alpha, V^\gamma) = \phi_\beta(p^\beta, V^\beta, V^\gamma)$$

it is evident that this does not reduce to equation (A2.1.2c) unless one can separate V^γ out of the equation. This is not possible unless $\phi_\alpha = f_\alpha(p^\alpha, V^\alpha)g(V^\gamma) + h(V^\alpha)$ and $\phi_\beta = f_\beta(p^\beta, V^\beta)g(V^\gamma) + h(V^\beta)$. If equation (A2.1.2c) is a statement of a general experimental result, then $f_\alpha(p^\alpha, V^\alpha) = f_\beta(p^\beta, V^\beta)$ and the symmetry of equation (2.1.2a), equation (2.1.2b) and equation (A2.1.2c) extends the equality to $f_\gamma(p^\gamma, V^\gamma)$:

$$f_\alpha(p^\alpha, V^\alpha) = f_\beta(p^\beta, V^\beta) = f_\gamma(p^\gamma, V^\gamma) = \theta. \quad (\text{A2.1.3})$$

The three systems share a common property θ , the numerical value of the three functions f_α , f_β and f_γ , which can be called the *empirical temperature*. The equations (A2.1.3) are *equations of state* for the various systems, but the choice of θ is entirely arbitrary, since any function of f (e.g. f^2 , $\log f$, $\cos^2 f - 3f^3$, etc) will satisfy equation (A2.1.3) and could serve as ‘temperature’.

Redlich [3] has criticized the ‘so-called zeroth law’ on the grounds that the argument applies equally well for the introduction of any generalized force, mechanical (pressure), electrical (voltage), or otherwise. The difference seems to be that the physical nature of these other forces has already been clearly defined or postulated (at least in the conventional development of physics) while in classical thermodynamics, especially in the Born–Carathéodory approach, the existence of temperature has to be inferred from experiment.

For convenience, one of the systems will be taken as an ideal gas whose equation of state follows *Boyle’s law*,

$$pV = nf(\theta) = nC\theta_{\text{ig}} \quad (\text{A2.1.4})$$

and which defines an ideal-gas temperature θ_{ig} proportional to pV/n . Later this will be identified with the thermodynamic temperature T . It is now possible to use the pair of variables V and θ instead of p and V to define the state of the system (for fixed n). [The pair p and θ would also do unless there is more than one phase present, in which case some variable or variables (in addition to n) must be extensive.]

A2.1.3 THE FIRST LAW

A2.1.3.1 WORK

There are several different forms of work, all ultimately reducible to the basic definition of the infinitesimal work $Dw = fdl$ where f is the force acting to produce movement along the distance dl . Strictly speaking, both f and dl are vectors, so Dw is positive when the extension dl of the system is in the same direction as the applied force; if they are in opposite directions Dw is negative. Moreover, this definition assumes (as do all the equations that follow in this section) that there is a substantially equal and opposite force resisting the movement. Otherwise the actual work done on the system or by the system on the surroundings will be less or even zero. As will be shown later, the maximum work is obtained when the process is essentially ‘reversible’.

The work depends on the detailed path, so Dw is an ‘inexact differential’ as symbolized by the capitalization. (There is no established convention about this symbolism; some books—and all mathematicians—use the same symbol for all differentials; some use δ for an inexact differential; others use a bar through the d ; still others—as in this article—use D .) The difference between an *exact* and an *inexact* differential is crucial in thermodynamics. In general, the integral of a differential depends on the path taken from the initial to the final state. However, for some special but important cases, the integral is independent of the path; then and only then can one write

$$\int_i^f dF = F_f - F_i = \Delta F.$$

One then speaks of F as a ‘state function’ because it is a function only of those variables that define the state of the system, and not of the path by which the state was reached. An especially important feature of such functions is that if one writes DF as a function of several variables, say x, y, z ,

$$DF = X(x, y, z) dx + Y(x, y, z) dy + Z(x, y, z) dz$$

then, for exact differentials *only*, $X = (\partial F / \partial x)_{y,z}$, $Y = (\partial F / \partial y)_{x,z}$ and $Z = (\partial F / \partial z)_{x,y}$. Since these exact differentials are path-independent, the order of differentiation is immaterial and one can then write

$$(\partial^2 F / \partial x \partial y)_z = (\partial X / \partial y)_{x,z} = (\partial Y / \partial x)_{y,z} \quad \text{etc.}$$

One way of verifying the exactness of a differential is to check the validity of expressions like that above.

(A) GRAVITATIONAL WORK

What is probably the simplest form of work to understand occurs when a force is used to raise the system in a gravitational field:

$$Dw_{\text{grav}} = mg dh$$

-6-

where m is the mass of the system, g is the acceleration of gravity, and dh is the infinitesimal increase in height. Gravitational work is rarely significant in most thermodynamic applications except when a falling weight outside the system drives a paddle wheel inside the system, as in one of the famous experiments in which Joule (1849) compared the work done with the increase in temperature of the system, and determined the ‘mechanical equivalent of heat’. Note that, in this example, positive work is done on the system as the potential energy of the falling weight decreases. Note also that, in free fall, the potential energy of the weight decreases, but no work is done.

(B) ONE-DIMENSIONAL WORK

When a spring is stretched or compressed, work is done. If the spring is the system, then the work done on it is simply

$$Dw_1 = f dl.$$

Note that a displacement from the initial equilibrium, either by compression or by stretching, produces positive work on the system. A situation analogous to the stretching of a spring is the stretching of a chain polymer.

(C) TWO-DIMENSIONAL (SURFACE) WORK

When a surface is compressed by a force $f = \pi L$, the ‘surface pressure’ $\pi = f/L$ is the force per unit width L producing a decrease in length dl . (Note that L and l are not the same; indeed they are orthogonal.) The work is then

$$Dw_2 = -\pi dA$$

where $dA = L dl$ is the change in the surface area. This kind of work and the related thermodynamic functions for surfaces are important in dealing with monolayers in a Langmuir trough, and with membranes and other materials that are quasi-two-dimensional.

(D) THREE-DIMENSIONAL (PRESSURE-VOLUME) WORK

When a piston of area A , driven by a force $f = pA$, moves a distance $dl = -dV/A$, it produces a compression of the system by a volume dV . The work is then

$$Dw_3 = -p dV. \tag{A2.1.5}$$

It is this type of work that is ubiquitous in chemical thermodynamics, principally because of changes of the volume of the system under the external pressure of the atmosphere. The negative sign of the work done *on* the system is, of course, because the application of excess pressure produces a decrease in volume. (The negative sign in the two-dimensional case is analogous.)

-7-

(E) OTHER MECHANICAL WORK

One can also do work by stirring, e.g. by driving a paddle wheel as in the Joule experiment above. If the paddle is taken as part of the system, the energy input (as work) is determined by appropriate measurements on the electric motor, falling weights or whatever drives the paddle.

(F) ELECTRICAL WORK

When a battery (or a generator or other power supply) outside the system drives current, i.e. a flow of electric charge, through a wire that passes through the system, work is done on the system:

$$Dw_{\text{elec}} = \mathcal{E} dQ$$

where dQ is the infinitesimal charge that crosses the boundary of the system and \mathcal{E} is the electric potential (voltage) across the system, i.e. between the point where the wire enters and the point where it leaves.

Converting to current $\mathcal{I} = dQ/dt$ where dt is an infinitesimal time interval and to resistance $\mathcal{R} = \mathcal{E}/\mathcal{I}$ one can rewrite this equation in the form

$$Dw_{\text{elec}} = \mathcal{E}\mathcal{I} dt = (\mathcal{E}^2/\mathcal{R}) dt.$$

Such a resistance device is usually called an ‘electrical heater’ but, since there is no means of measurement at the boundary between the resistance and the material in contact with it, it is easier to regard the resistance as being inside the system, i.e. a part of it. Energy enters the system in the form of work where the wire breaches the wall, i.e. enters the container.

(G) ELECTROCHEMICAL WORK

A special example of electrical work occurs when work is done on an electrochemical cell or by such a cell on the surroundings ($-w$ in the convention of this article). Thermodynamics applies to such a cell when it is at equilibrium with its surroundings, i.e. when the electrical potential (*electromotive force* emf) of the cell is

balanced by an external potential.

(H) ELECTROMAGNETIC WORK

This poses a special problem because the source of the electromagnetic field may lie outside the defined boundaries of the system. A detailed discussion of this is outside the scope of this section, but the basic features can be briefly summarized.

When a specimen is moved in or out of an electric field or when the field is increased or decreased, the total work done on the whole system (charged condenser + field + specimen) in an infinitesimal change is

$$Dw_{el} = \int dV(\mathbf{E} \cdot d\mathbf{D}),$$

-8-

where E is the electric field vector, $D = \epsilon E$ is the electric displacement vector, and ϵ is the electric susceptibility tensor. The integration is over the whole volume encompassed by the total system, which must in principle extend as far as measurable fields exist.

Similarly, when a specimen is moved in or out of a magnetic field or when the magnetic field is increased or decreased, the total work done on the whole system (coil + field + specimen) in an infinitesimal change is

$$Dw_{mag} = \int dV(\mathbf{H} \cdot d\mathbf{B})$$

where H is the magnetic field vector, $B = \mu H$ is the magnetic induction vector and μ is the magnetic permeability tensor. (Some modern discussions of magnetism regard B as the fundamental magnetic field vector, but usually fail to give a new name to H .) As before the integration is over the whole volume.

For the special but familiar case of an isotropic specimen in a uniform external field E_0 or B_0 , it can be shown [4] that

$$Dw_{el} = \int dV(\epsilon_0 \mathbf{E}_0 \cdot d\mathbf{E}_0 - \mathbf{P} \cdot d\mathbf{E}_0) \tag{A2.1.6}$$

$$Dw_{mag} = \int dV(\mathbf{B}_0 \cdot d\mathbf{B}_0/\mu_0 + \mathbf{B}_0 \cdot d\mathbf{M}) \tag{A2.1.7}$$

where P is the polarization vector and M the magnetization vector; ϵ_0 and μ_0 are the susceptibility and permeability of the vacuum in the absence of the specimen. The vector notation could now be dropped since the external field and the induced field are parallel and the scalar product of two vectors oriented identically is simply the product of their scalar magnitudes; this will not be done in this article to avoid confusion with other thermodynamic quantities. (Note that equation (A2.1.7) is not the analogue of equation (A2.1.6).)

The work done increases the energy of the total system and one must now decide how to divide this energy between the field and the specimen. This separation is not measurably significant, so the division can be made arbitrarily; several self-consistent systems exist. The first term on the right-hand side of equation (A2.1.6) is obviously the work of creating the electric field, e.g. charging the plates of a condenser in the absence of the specimen, so it appears logical to consider the second term as the work done on the specimen.

By analogy, one is tempted to make the same division in equation (A2.1.7), regarding the first term as the work of creating the magnetic field in the absence of the specimen and the second, $\int dV(\mathbf{B}_0 \cdot d\mathbf{M})$, as the work done on the specimen. This is the way most books on thermodynamics present the problem and it is an acceptable convention, except that it is inconsistent with the measured spectroscopic energy levels and with one's intuitive idea of work. For example, equation (A2.1.7) says that the work done in moving a permanent magnet (constant magnetization M) into or out of an electromagnet of constant B_0 is exactly zero! This is actually correct if one considers the extra electrochemical work done on the battery driving the current through the electromagnet while the permanent magnet is moving; this exactly balances the mechanical work. A careful analysis [5, 6] shows that, if one writes equation (A2.1.7) in the following form:

-9-

$$Dw_{\text{mag}} = \int dV \left[\underbrace{\mathbf{B}_0 \cdot d\mathbf{B}_0 / \mu_0}_{\text{A}} - \underbrace{M d\mathbf{B}_0}_{\text{B}} + \underbrace{d(\mathbf{B}_0 \cdot \mathbf{M})}_{\text{C}} \right]$$

then term A is the work of creating the field in the absence of the specimen; term B is the work done on the specimen by 'ponderable forces', e.g. by a spring or by a physical push or pull; this is directly reflected in a change of the kinetic energy of the electrons; and term C is the work done by the electromotive force in the coil in creating the interaction field between B_0 and M . We elect to consider term B as the only work done on the specimen and write for the electromagnetic work

$$Dw_{\text{electromag}} = \int dV (-\mathbf{P} \cdot d\mathbf{E}_0 - \mathbf{M} \cdot d\mathbf{B}_0).$$

If in addition the specimen is assumed to be spherical as well as isotropic, so that P and M are uniform throughout the volume V , one can then write for the electromagnetic work

$$Dw_{\text{electromag}} = V(-\mathbf{P} \cdot d\mathbf{E}_0 - \mathbf{M} \cdot d\mathbf{B}_0). \quad (\text{A2.1.8})$$

Equation (A2.1.8) turns out to be consistent with the changes of the energy levels measured spectroscopically, so the energy produced by work defined this way is frequently called the 'spectroscopic energy'. Note that the electric and magnetic parts of the equations are now symmetrical.

A2.1.3.2 ADIABATIC WORK

One may now consider how changes can be made in a system across an adiabatic wall. The first law of thermodynamics can now be stated as another generalization of experimental observation, but in an unfamiliar form: *the work required to transform an adiabatic (thermally insulated) system from a completely specified initial state to a completely specified final state is independent of the source of the work (mechanical, electrical, etc.) and independent of the nature of the adiabatic path.* This is exactly what Joule observed; the same amount of work, mechanical or electrical, was always required to bring an adiabatically enclosed volume of water from one temperature θ_1 to another θ_2 .

This can be illustrated by showing the net work involved in various adiabatic paths by which one mole of helium gas (4.00 g) is brought from an initial state in which $p = 1.000$ atm, $V = 24.62$ l [$T = 300.0$ K], to a final state in which $p = 1.200$ atm, $V = 30.779$ l [$T = 450.0$ K]. Ideal-gas behaviour is assumed (actual experimental measurements on a slightly non-ideal real gas would be slightly different). Information shown in brackets could be measured or calculated, but is not essential to the experimental verification of the first law.

Path I	(a) Do electrical work on the system at constant $V = 24.62$ l until the pressure has risen to 1.500 atm. [$\Delta T = 150.0$ K, $w = (3/2)R\Delta T$]	$w_{\text{elec}} = 1871$ J
	(b) Expand the gas into a vacuum (i.e. against zero external pressure) until the total volume V is 30.77 l and $p = 1.200$ atm. [$\Delta T = 0$]	$w_{\text{exp}} = 0$ J $w_{\text{tot}} = 1871$ J
Path II	(a) Compress the gas reversibly and adiabatically from 1.000 atm to 1.200 atm. [At the end of the compression $T = 322.7$ K, $V = 22.07$ l, $w = (3/2)R\Delta T$]	$w_{\text{comp}} = 283$ J
	(b) Do electrical work on the system, holding the pressure constant at 1.200 atm, until the volume V has increased to 30.77 l; under these circumstances the system also does expansion work against the external pressure. [Electrical work = $(5/2)R\Delta T$ [Expansion work = $-p\Delta V = -10.45$ l atm]	$w_{\text{elec}} = 2646$ J $w_{\text{exp}} = -1058$ J $w_{\text{tot}} = 1871$ J
Path III	(a) Do electrical work on the system, holding the pressure constant at 1.000 atm, until the volume V has increased to 34.33 l; under these circumstances, the system also does expansion work against the external pressure. [Final $T = 418.4$ K] [Electrical work = $(5/2)RT$ [Expansion work = $-p\Delta V = -9.71$ l atm]	$w_{\text{elec}} = 2460$ J $w_{\text{exp}} = -984$ J
	(b) Compress the gas reversibly and adiabatically from 1.000 atm to 1.200 atm. [At the end of the compression $T = 450.0$ K, $V = 30.77$ l, $\Delta T = 31.65$ K, $w = (3/2)RT$]	$w_{\text{comp}} = 395$ J $w_{\text{tot}} = 1871$ J

For all of these adiabatic processes, the total (net) work is exactly the same.

(As we shall see, because of the limitations that the second law of thermodynamics imposes, it may be impossible to find any adiabatic paths from a particular state A to another state B because $S_A - S_B < 0$. In this situation, however, there will be several adiabatic paths from state B to state A.)

If the adiabatic work is independent of the path, it is the integral of an exact differential and suffices to define a change in a function of the state of the system, the *energy* U . (Some thermodynamicists call this the ‘internal energy’, so as to exclude any kinetic energy of the motion of the system as a whole.)

$$dU = dw_{\text{adiabatic}}$$

or

$$\Delta U = U_f(V_f, \theta_f) - U_i(V_i, \theta_i) = \int dw_{\text{adiabatic}} = w_{\text{adiabatic}} \quad (\text{A2.1.9})$$

Here the subscripts i and f refer to the initial and final states of the system and the work w is defined as the work performed on the system (the opposite sign convention—with w as work done by the system on the surroundings—is also in common use). Note that a cyclic process (one in which the system is returned to its initial state) is *not* introduced; as will be seen later, a cyclic adiabatic process is possible only if every step is reversible. Equation (A2.1.9), i.e. the introduction of U as a state function, is an expression of the law of conservation of energy.

A2.1.3.3 NON-ADIABATIC PROCESSES. HEAT

Not all processes are adiabatic, so when a system is coupled to its environment by diathermic walls, the heat q absorbed by the system is defined as the difference between the actual work performed and that which would have been required had the change occurred adiabatically.

$$Dq = dw_{\text{adiabatic}} - Dw = dU - Dw$$

or

$$q = w_{\text{adiabatic}} - w = \Delta U - w. \quad (\text{A2.1.10})$$

Note that, since Dw is inexact, so also must be Dq .

This definition may appear eccentric because many people have an intuitive feeling for ‘heat’ as a certain kind of energy flow. However, thoughtful reconsideration supports a suspicion that the intuitive feeling is for the heat absorbed in a particular kind of process, e.g. constant pressure, for which, as we shall see, the heat q_p is equal to the change in a state function, the enthalpy change ΔH . For another example, the ‘heats’ measured in modern calorimeters are usually determined either by a *measurement* of electrical or mechanical work or by comparing one process with another so calibrated (as in an ice calorimeter). Indeed one can argue that one never measures q directly, that all ‘measurements’ require equation (A2.1.10); one always infers q from other measurements.

A2.1.4 THE SECOND LAW

In this and nearly all subsequent sections, the work Dw will be restricted to pressure–volume work, $-p dV$, and the fact that the ‘heat’ Dq may in some cases be electrical work will be ignored.

A2.1.4.1 REVERSIBLE PROCESSES

A particular path from a given initial state to a given final state is the reversible process, one in which after each infinitesimal step the system is in equilibrium with its surroundings, and one in which an infinitesimal change in the conditions (constraints) would reverse the direction of the change.

A simple example (figure A2.1.2) consists of a gas confined by a movable piston supporting a pile of sand whose weight produces a downward force per unit area equal to the pressure of the gas. Removal of a grain of sand decreases the downward pressure by an amount δp and the piston rises with an increase of volume δV sufficient to decrease the gas pressure by the *same* δp ; the system is now again at equilibrium. Restoration of the grain of sand will drive the piston and the gas back to their initial states. Conversely, the successive removal of additional grains of sand will produce additional small decreases in pressure and small increases in volume; the sum of a very large number of such small steps can produce substantial changes in the thermodynamic properties of the system. Strictly speaking, such experimental processes are never quite reversible because one can never make the small changes in pressure and volume infinitesimally small (in such a case there would be no tendency for change and the process would take place only at an infinitely slow rate). The true reversible process is an idealized concept; however, one can usually devise processes sufficiently close to reversibility that no measurable differences will be observed.

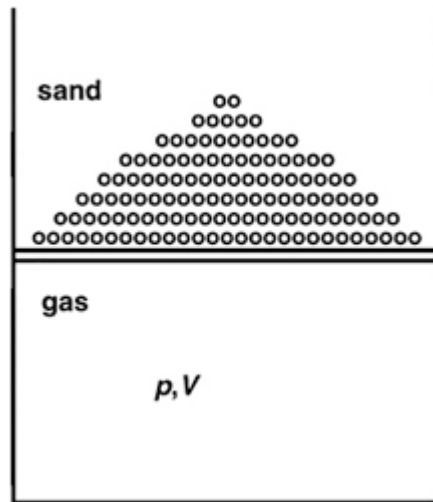


Figure A2.1.2. Reversible expansion of a gas with the removal one-by-one of grains of sand atop a piston.

The mere fact that a substantial change can be broken down into a very large number of small steps, with equilibrium (with respect to any applied constraints) at the end of each step, does not guarantee that the process is reversible. One can modify the gas expansion discussed above by restraining the piston, not by a pile of sand, but by the series of stops (pins that one can withdraw one-by-one) shown in figure A2.1.3. Each successive state is indeed an equilibrium one, but the pressures on opposite sides of the piston are not equal, and pushing the pins back in one-by-one will not drive the piston back down to its initial position. The two processes are, in fact, quite different even in the infinitesimal limit of their small steps; in the first case work is done by the gas to raise the sand pile, while in the second case there is no such work. Both the processes may be called ‘quasi-static’ but only the first is anywhere near reversible. (Some thermodynamics texts restrict the term ‘quasi-static’ to a more restrictive meaning equivalent to ‘reversible’, but this then leaves no term for the slow irreversible process.)

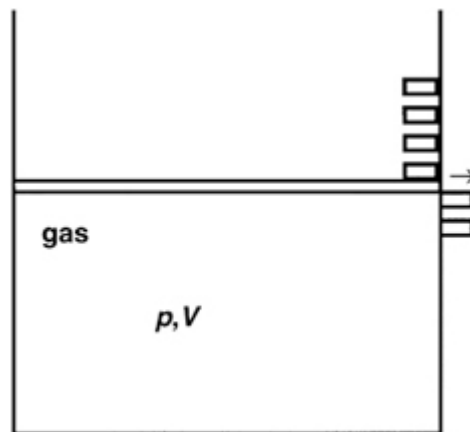


Figure A2.1.3. Irreversible expansion of a gas as stops are removed.

If a system is coupled with its environment through an adiabatic wall free to move without constraints (such as the stops of the second example above), mechanical equilibrium, as discussed above, requires equality of the pressure p on opposite sides of the wall. With a diathermic wall, thermal equilibrium requires that the temperature θ of the system equal that of its surroundings. Moreover, it will be shown later that, if the wall is permeable and permits exchange of matter, material equilibrium (no tendency for mass flow) requires equality of a chemical potential μ .

Obviously the first law is not all there is to the structure of thermodynamics, since some adiabatic changes occur spontaneously while the reverse process never occurs. An aspect of the second law is that a state function, the *entropy* S , is found that increases in a spontaneous adiabatic process and remains unchanged in a reversible adiabatic process; it cannot decrease in any adiabatic process.

The next few sections deal with the way these experimental results can be developed into a mathematical system. A reader prepared to accept the second law on faith, and who is interested primarily in applications, may skip section A2.1.4.2 and [section A2.1.4.6](#) and perhaps even [A2.1.4.7](#), and go to the final statement in [section A2.1.4.8](#).

A2.1.4.2 ADIABATIC REVERSIBLE PROCESSES AND INTEGRABILITY

In the example of the previous section, the release of the stop always leads to the motion of the piston in one direction, to a final state in which the pressures are equal, never in the other direction. This obvious experimental observation turns out to be related to a mathematical problem, the integrability of differentials in thermodynamics. The differential Dq , even Dq_{rev} , is inexact, but in mathematics many such expressions can be converted into exact differentials with the aid of an integrating factor.

In the example of pressure–volume work in the previous section, the adiabatic reversible process consisted simply of the sufficiently slow motion of an adiabatic wall as a result of an infinitesimal pressure difference. The work done on the system during an infinitesimal reversible change in volume is then $-p dV$ and one can write equation (A2.1.11) in the form

$$Dq_{\text{rev}} = dU + p dV = 0. \quad (\text{A2.1.11})$$

If U is expressed as a function of two variables of state, e.g. V and θ , one can write $dU = (\partial U/\partial V)_{\theta} dV + (\partial U/\partial \theta)_{V} d\theta$ and transform equation (A2.1.11) into the following:

$$Dq_{\text{rev}} = [(\partial U/\partial V)_{\theta} + p] dV + (\partial U/\partial \theta)_{V} d\theta = Y dV + Z d\theta = 0. \quad (\text{A2.1.12})$$

The coefficients Y and Z are, of course, functions of V and θ and therefore state functions. However, since in general $(\partial p/\partial \theta)_{V}$ is not zero, $\partial Y/\partial \theta$ is not equal to $\partial Z/\partial V$, so Dq_{rev} is not the differential of a state function but rather an inexact differential.

For a system composed of two subsystems α and β separated from each other by a diathermic wall and from the surroundings by adiabatic walls, the equation corresponding to equation (A2.1.12) is

$$\begin{aligned} Dq_{\text{rev}} &= Dq^{\alpha} + Dq^{\beta} \\ &= [(\partial U^{\alpha}/\partial V^{\alpha})_{\theta} + p^{\alpha}] dV^{\alpha} + [(\partial U^{\beta}/\partial V^{\beta})_{\theta} + p^{\beta}] dV^{\beta} + [(\partial U^{\alpha}/\partial \theta)_{V^{\alpha}} \\ &\quad + (\partial U^{\beta}/\partial \theta)_{V^{\beta}}] d\theta \\ &= X dV^{\alpha} + Y dV^{\beta} + Z d\theta = 0. \end{aligned} \quad (\text{A2.1.13})$$

One must now examine the integrability of the differentials in [equation \(A2.1.12\)](#) and [equation \(A2.1.13\)](#), which are examples of what mathematicians call *Pfaff differential equations*. If the equation is integrable, one can find an integrating denominator λ , a function of the variables of state, such that $Dq_{\text{rev}}/\lambda = d\phi$ where $d\phi$ is the exact differential of a function ϕ that defines a surface (line in the case of [equation \(A2.1.12\)](#)) in which the reversible adiabatic path must lie.

All equations of two variables, such as [equation \(A2.1.12\)](#), are necessarily integrable because they can be written in the form $dy/dx = f(x, y)$, which determines a unique value of the slope of the line through any point (x, y) . Figure A2.1.4 shows a set of non-intersecting lines in V - θ space representing solutions of [equation \(A2.1.12\)](#).

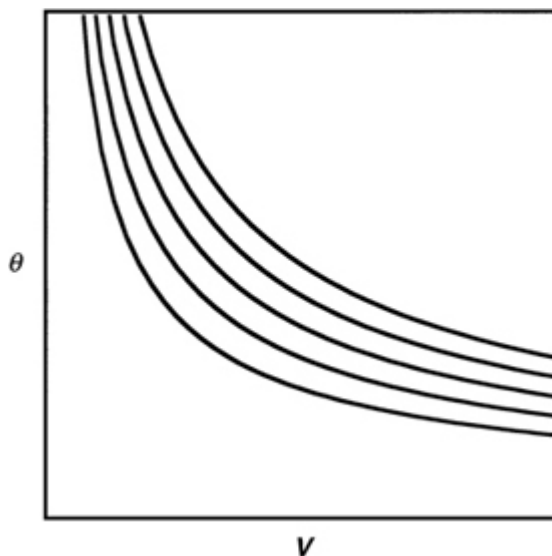


Figure A2.1.4. Adiabatic reversible (isentropic) paths that do not intersect. (The curves have been calculated for the isentropic expansion of a monatomic ideal gas.)

For equations such as (A2.1.13) involving more than two variables the problem is no longer trivial. Most such equations are not integrable.

(Born [2] cites as an example of a simple expression for which no integrating factor exists

$$DF = dy + x dz \stackrel{?}{=} \lambda(x, y, z) d\phi.$$

-15-

If an integrating factor exists $\partial\phi/\partial x = 0$, $\partial\phi/\partial y = 1/\lambda$ and $\partial\phi/\partial z = x/\lambda$. From the first of these relations one concludes that ϕ depends only on y and z . Using this result in the second relation one concludes that λ depends only on y and z . Given that ϕ and λ are both functions only of y and z , the third relation is a contradiction, so no factor λ can exist.)

There are now various adiabatic reversible paths because one can choose to vary dV^α or dV^β in any combination of steps. The paths can cross and interconnect. The question of integrability is tied to the question of whether all regions of $V^\alpha, V^\beta, \theta$ space are accessible by a series of connected adiabatic reversible paths or whether all such paths lie in a series of non-crossing surfaces. To distinguish, one must use a theorem of Carathéodory (the proof can be found in [1] and [2] and in books on differential equations):

If a Pfaff differential expression $DF = X dx + Y dy + Z dz$ has the property that every arbitrary neighbourhood of a point $P(x, y, z)$ contains points that are inaccessible along a path corresponding to a solution of the equation $DF = 0$, then an integrating denominator exists. Physically this means that there are two mutually exclusive possibilities: either (a) a hierarchy of non-intersecting surfaces $\phi(x, y, z) = C$, each with a different value of the constant C , represents the solutions $DF = 0$, in which case a point on one surface is inaccessible

by a path that is confined to another, or (b) any two points can be connected by a path, each infinitesimal segment of which satisfies the condition $DF = 0$. One must perform some experiments to determine which situation prevails in the physical world.

It suffices to carry out *one* such experiment, such as the expansion or compression of a gas, to establish that there are states inaccessible by adiabatic reversible paths, indeed even by any adiabatic irreversible path. For example, if one takes one mole of N_2 gas in a volume of 24 litres at a pressure of 1.00 atm (i.e. at 25 °C), there is no combination of adiabatic reversible paths that can bring the system to a final state with the same volume and a different temperature. A higher temperature (on the ideal-gas scale θ_{ig}) can be reached by an adiabatic irreversible path, e.g. by doing electrical work on the system, but a state with the same volume and a lower temperature θ_{ig} is inaccessible by *any* adiabatic path.

A2.1.4.3 ENTROPY AND TEMPERATURE

One concludes, therefore, that equation (A2.1.13) is integrable and there exists an integrating factor λ . For the general case $Dq_{rev} = \lambda d\phi$ it can be shown [1, 2] that

$$\ln \lambda = \int g(\theta) d\theta + \ln I(\phi)$$

where $I(\phi)$ is a constant of integration. It then follows that one may define two new quantities by the relations:

$$\ln(T/C) = \int g(\theta) d\theta \quad S = (1/C) \int I(\phi) d\phi.$$

-16-

and one can now write

$$Dd q_{rev} = \lambda d\phi = T dS. \tag{A2.1.14}$$

There are an infinite number of other integrating factors λ with corresponding functions ϕ ; the new quantities T and S are chosen for convenience. S is, of course, the *entropy* and T , a function of θ only, is the ‘absolute temperature’, which will turn out to be the ideal-gas temperature, θ_{ig} . The constant C is just a scale factor determining the size of the degree.

The surfaces in which the paths satisfying the condition $Dq_{rev} = 0$ must lie are, thus, surfaces of constant entropy; they do not intersect and can be arranged in an order of increasing or decreasing numerical value of the constant S . One half of the second law of thermodynamics, namely that for reversible changes, is now established.

Since $Dw_{rev} = -pdV$, one can utilize the relation $dU = Dq_{rev} + Dw_{rev}$ and write

$$dU = T dS - p dV. \tag{A2.1.15}$$

Equation (A2.1.15) involves only state functions, so it applies to any infinitesimal change in state whether the actual process is reversible or not (although, as equation (A2.1.14) suggests, dS is not experimentally accessible unless some reversible path exists).

A2.1.4.4 THERMODYNAMIC TEMPERATURE AND THE IDEAL-GAS THERMOMETER

So far, the thermodynamic temperature T has appeared only as an integrating denominator, a function of the empirical temperature θ . One now can show that T is, except for an arbitrary proportionality factor, the *same* as the empirical ideal-gas temperature θ_{ig} introduced earlier. Equation (A2.1.15) can be rewritten in the form

$$T dS = dU + p dV = (\partial U / \partial \theta)_V d\theta + [(\partial U / \partial V)_\theta + p] dV. \quad (\text{A2.1.16})$$

One assumes the existence of a fluid that obeys Boyle's law (equation (A2.1.4)) and that, on adiabatic expansion into a vacuum, shows no change in temperature, i.e. for which $pV = f(\theta)$ and $(\partial U / \partial V)_\theta = 0$. (All real gases satisfy this condition in the limit of zero pressure.) Equation (A2.1.16) then simplifies to

$$T dS = (dU/d\theta)d\theta + [f(\theta)/V]dV = f(\theta)\{[(dU/d\theta)/f(\theta)]d\theta + dV/V\}.$$

The factor in wavy brackets is obviously an exact differential because the coefficient of $d\theta$ is a function only of θ and the coefficient of dV is a function only of V . (The cross-derivatives vanish.) Manifestly then

$$\left. \begin{aligned} T &= Cf(\theta) = C(pV) \\ dS &= \frac{dU/d\theta}{Cf(\theta)} d\theta + \frac{1}{CV} dV = \frac{dU}{T} + \frac{dV}{CV} \end{aligned} \right\} \text{ ideal gas only.}$$

-17-

If the arbitrary constant C is set equal to $(nR)^{-1}$ where n is the number of moles in the system and R is the gas constant per mole, then the thermodynamic temperature $T = \theta_{\text{ig}}$ where θ_{ig} is the temperature measured by the ideal-gas thermometer depending on the equation of state

$$pV = nR\theta_{\text{ig}} = nRT. \quad (\text{A2.1.17})$$

Now that the identity has been proved θ_{ig} need not be used again.

A2.1.4.5 IRREVERSIBLE CHANGES AND THE SECOND LAW

It is still necessary to consider the role of entropy in irreversible changes. To do this we return to the system considered earlier in section A2.1.4.2, the one composed of two subsystems in thermal contact, each coupled with the outside through movable adiabatic walls. Earlier this system was described as a function of three independent variables, V^α , V^β and θ (or T). Now, instead of the temperature, the entropy $S = S^\alpha + S^\beta$ will be used as the third variable. A final state $V^{\alpha'}$, $V^{\beta'}$, S' can always be reached from an initial state $V^{\alpha 0}$, $V^{\beta 0}$, S^0 by a two-step process.

- (1) The volumes are changed adiabatically and reversibly from $V^{\alpha 0}$ and $V^{\beta 0}$ to $V^{\alpha'}$ and $V^{\beta'}$, during which change the entropy remains constant at S^0 .
- (2) At constant volumes $V^{\alpha'}$ and $V^{\beta'}$, the state is changed by the adiabatic performance of work (stirring, rubbing, electrical 'heating') until the entropy is changed from S^0 to S' .

If the entropy change in step (2) could be at times greater than zero and at other times less than zero, every neighbouring state $V^{\alpha'}$, $V^{\beta'}$, S' would be accessible, for there is no restriction on the adjustment of volumes in

step (1). This contradicts the experimental fact that allowed the integration of [equation \(A2.1.13\)](#) and established the entropy S as a state function. It must, therefore, be true that either $S' > S^0$ always or that $S' < S^0$ always. One experiment demonstrates that the former is the correct alternative; if one takes the absolute temperature as a positive number, one finds that the entropy cannot decrease in an adiabatic process. This completes the specification of temperature, entropy and part of the second law of thermodynamics. One statement of the second law of thermodynamics is therefore:

$$\text{for any adiabatic process } (Dq = 0) \, dS \geq 0. \quad (\text{A2.1.18})$$

(This is frequently stated for an isolated system, but the same statement about an adiabatic system is broader.)

A2.1.4.6 IRREVERSIBLE CHANGES AND THE MEASUREMENT OF ENTROPY

Thermodynamic measurements are possible only when both the initial state and the final state are essentially at equilibrium, i.e. internally and with respect to the surroundings. Consequently, for a spontaneous thermodynamic change to take place, some constraint—internal or external—must be changed or released. For example, the expansion of a gas requires the release of a pin holding a piston in place or the opening of a stopcock, while a chemical reaction can be initiated by mixing the reactants or by adding a catalyst. One often finds statements that ‘at equilibrium in an isolated system (constant U, V, n), the entropy is maximized’. What does this mean?

-18-

Consider two ideal-gas subsystems α and β coupled by a movable diathermic wall (piston) as shown in [figure A2.1.5](#). The wall is held in place at a fixed position l by a stop (pin) that can be removed; then the wall is free to move to a new position l' . The total system ($\alpha + \beta$) is adiabatically enclosed, indeed isolated ($q = w = 0$), so the total energy, volume and number of moles are fixed.

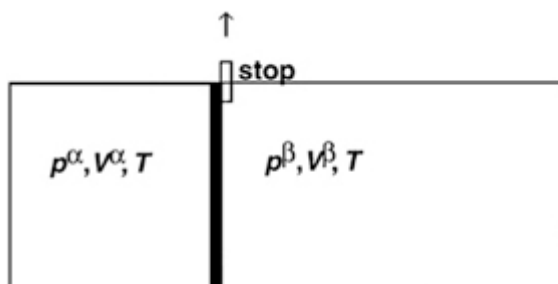


Figure A2.1.5. Irreversible changes. Two gases at different pressures separated by a diathermic wall, a piston that can be released by removing a stop (pin).

When the pin is released, the wall will either (a) move to the right, or (b) move to the left, or (c) remain at the original position l . It is evident that these three cases correspond to initial situations in which $p^\alpha > p^\beta, p^\alpha < p^\beta$ and $p^\alpha = p^\beta$, respectively; if there are no other stops, the piston will come to rest in a final state where $p^{\alpha'} = p^{\beta'}$. For the two spontaneous adiabatic changes (a) and (b), the second law requires that $\Delta S > 0$, but one does not yet know the magnitude. (Nothing happens in case (c), so $\Delta S = 0$.)

The change of case (a) can be carried out in a series of small steps by having a large number of stops separated by successive distances Δl . For any intermediate step, $p^{\alpha'} > p^{\alpha''} > p^{\beta''} > p^{\beta'}$, but since the steps, no matter how small, are never reversible, one still has no information about ΔS .

The only way to determine the entropy change is to drive the system back from the final state to the initial state along a reversible path. One reimposes a constraint, not with simple stops, but with a gear system that permits one to do mechanical work driving the piston back to its original position l_0 along a reversible path; this work can be measured in various conventional ways. During this reverse change the system is no longer isolated; the total V and the total n remain unchanged, but the work done on the system adds energy. To keep the total energy constant, an equal amount of energy must leave the system in the form of heat:

$$dU = Dq_{\text{rev}} + Dw_{\text{rev}} = 0$$

or

$$-\Delta S_{\text{forward}} = \Delta S_{\text{reverse}} = \int \frac{Dq_{\text{rev}}}{T} = - \int \frac{Dw_{\text{rev}}}{T}.$$

-19-

For an ideal gas and a diathermic piston, the condition of constant energy means constant temperature. The reverse change can then be carried out simply by relaxing the adiabatic constraint on the external walls and immersing the system in a thermostatic bath. More generally the initial state and the final state may be at different temperatures so that one may have to have a series of temperature baths to ensure that the entire series of steps is reversible.

Note that although the change in state has been reversed, the system has not returned along the same detailed path. The forward spontaneous process was adiabatic, unlike the driven process and, since it was not reversible, surely involved some transient temperature and pressure gradients in the system. Even for a series of small steps ('quasi-static' changes), the infinitesimal forward and reverse paths must be different in detail. Moreover, because q and w are different, there are changes in the surroundings; although the system has returned to its initial state, the surroundings have not.

One can, in fact, drive the piston in both directions from the equilibrium value $l = l_e$ ($p^\alpha = p^\beta$) and construct a curve of entropy S (with an arbitrary zero) as a function of the piston position l (figure A2.1.6). If there is a series of stops, releasing the piston will cause l to change in the direction of increasing entropy until the piston is constrained by another stop or until l reaches l_e . It follows that at $l = l_e$, $dS/dl = 0$ and $d^2S/dl^2 < 0$; i.e. S is maximized when l is free to seek its own value. Were this not so, one could find spontaneous processes to jump from the final state to one of still higher entropy.

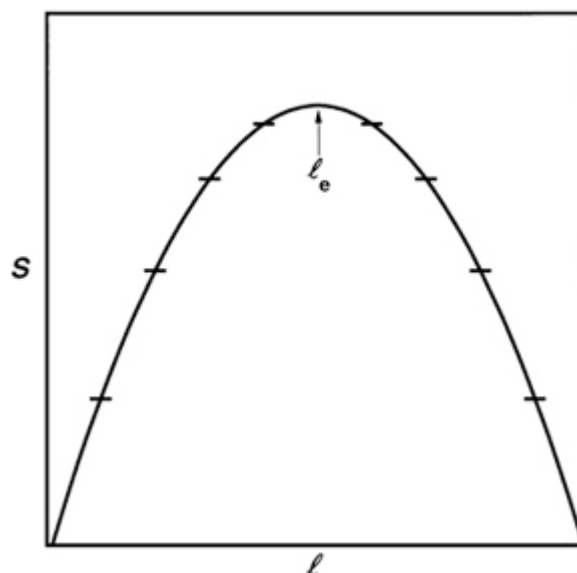


Figure A2.1.6. Entropy as a function of piston position l (the piston held by stops). The horizontal lines mark possible positions of stops, whose release produces an increase in entropy, the amount of which can be measured by driving the piston back reversibly.

Thus, the spontaneous process involves the release of a constraint while the driven reverse process involves the imposition of a constraint. The details of the reverse process are irrelevant; any series of reversible steps by which one can go from the final state back to the initial state will do to measure ΔS .

A2.1.4.7 IRREVERSIBLE PROCESSES: WORK, HEAT AND ENTROPY CREATION

One has seen that thermodynamic measurements can yield information about the change ΔS in an irreversible process (and thereby the changes in other state functions as well). What does thermodynamics tell one about work and heat in irreversible processes? Not much, in spite of the assertion in many thermodynamics books that

$$Dw = -p_{\text{ext}} dV = p_{\text{ext}} dV_{\text{ext}} \quad (\text{A2.1.19})$$

and

$$Dq = -T_{\text{ext}} dS_{\text{ext}} = -dU_{\text{ext}} - Dw \quad (\text{A2.1.20})$$

where p_{ext} and T_{ext} are the external pressure and temperature, i.e. those of the surroundings in which the changes $dV_{\text{ext}} = -dV$ and dS_{ext} occur.

Consider the situation illustrated in [figure A2.1.5](#), with the modification that the piston is now an adiabatic wall, so the two temperatures need not be equal. Energy is transmitted from subsystem α to subsystem β only in the form of work; obviously $dV^\alpha = -dV^\beta$ so, in applying equation (A2.1.20), is $dU^{\alpha \rightarrow \beta}$ equal to $-p^\alpha dV^\beta = p^\alpha dV^\alpha$ or equal to $p^\beta dV^\alpha$, or is it something else entirely? One can measure the changes in temperature, $T^{\alpha'} - T^\alpha$ and $T^{\beta'} - T^\beta$ and thus determine $\Delta U^{\alpha \rightarrow \beta}$ after the fact, but could it have been predicted in advance, at least for ideal gases? If the piston were a diathermic wall so the final temperatures are equal, the energy transfer $\Delta U^{\alpha \rightarrow \beta}$ would be calculable, but even in this case it is unclear how this transfer should be divided between heat and work.

In general, the answers to these questions are ambiguous. When the pin in [figure A2.1.5](#) is released, the potential energy inherent in the pressure difference imparts a kinetic energy to the piston. Unless there is a mechanism for the dissipation of this kinetic energy, the piston will oscillate like a spring; frictional forces, of course, dissipate this energy, but the extent to which the dissipation takes place in subsystem α or subsystem β depends on details of the experimental design not uniquely fixed by specifying the initial thermodynamic state. (For example, one can introduce braking mechanisms that dissipate the energy largely in subsystem α or, conversely, largely in subsystem β .) Only in one special case is there a clear prediction: if one subsystem (β) is empty no work can be done by α on β ; for expansion into a vacuum necessarily $w = 0$. A more detailed discussion of the work involved in irreversible expansion has been given by Kivelson and Oppenheim [7].

The paradox involved here can be made more understandable by introducing the concept of entropy creation. Unlike the energy, the volume or the number of moles, the entropy is not conserved. The entropy of a system (in the example, subsystems α or β) may change in two ways: first, by the transport of entropy across the boundary (in this case, from α to β or *vice versa*) when energy is transferred in the form of heat, and second,

by the creation of entropy within the subsystem during an irreversible process. Thus one can write for the change in the entropy of subsystem α in which some process is occurring

$$dS^\alpha = d_t S^\alpha + d_i S^\alpha$$

-21-

where $d_t S^\alpha = -d_t S^\beta$ is the change in entropy due to heat transfer to subsystem α and $d_i S^\alpha$ is the irreversible entropy creation inside subsystem α . (In the adiabatic example the dissipation of the kinetic energy of the piston by friction creates entropy, but no entropy is transferred because the piston is an adiabatic wall.)

The total change dS^α can be determined, as has been seen, by driving the subsystem α back to its initial state, but the separation into $d_t S^\alpha$ and $d_i S^\alpha$ is sometimes ambiguous. Any statistical mechanical interpretation of the second law requires that, at least for any volume element of macroscopic size, $d_i S \geq 0$. However, the total entropy change dS^α can be either positive or negative since the second law places no limitation on either the sign or the magnitude of $d_t S^\alpha$. (In the example above, the piston's adiabatic wall requires that $d_t S^\alpha = d_t S^\beta = 0$.)

In an irreversible process the temperature and pressure of the system (and other properties such as the chemical potentials μ_γ to be defined later) are not necessarily definable at some intermediate time between the equilibrium initial state and the equilibrium final state; they may vary greatly from one point to another. One can usually define T and p for each small volume element. (These volume elements must not be too small; e.g. for gases, it is impossible to define T , p , S , etc for volume elements smaller than the cube of the mean free path.) Then, for each such sub-subsystem, $d_i S$ (but not the total dS) must not be negative. It follows that $d_t S^\alpha$, the sum of all the $d_i S$ s for the small volume elements, is zero or positive. A detailed analysis of such irreversible processes is beyond the scope of classical thermodynamics, but is the basis for the important field of 'irreversible thermodynamics'.

The assumption (frequently unstated) underlying [equations \(A2.1.19\)](#) and [equation \(A2.1.20\)](#) for the measurement of irreversible work and heat is this: in the surroundings, which will be called subsystem β , internal equilibrium (uniform T^β , p^β and μ_i^β throughout the subsystem; i.e. no temperature, pressure or concentration gradients) is maintained throughout the period of time in which the irreversible changes are taking place in subsystem α . If this condition is satisfied $d_i S^\beta = 0$ and all the entropy creation takes place entirely in α . In any thermodynamic measurement that purports to yield values of q or w for an irreversible change, one must ensure that this condition is very nearly met. (Obviously, in the expansion depicted in [figure A2.1.5](#) neither subsystem α nor subsystem β satisfied this requirement.)

Essentially this requirement means that, during the irreversible process, immediately inside the boundary, i.e. on the system side, the pressure and/or the temperature are only infinitesimally different from that outside, although substantial pressure or temperature gradients may be found outside the vicinity of the boundary. Thus an infinitesimal change in p_{ext} or T_{ext} would instantly reverse the direction of the energy flow, i.e. the sign of w or q . That part of the total process occurring *at the boundary* is then 'reversible'.

Subsystem β may now be called the 'surroundings' or as Callen (see further reading at the end of this article) does, in an excellent discussion of this problem, a 'source'. To formulate this mathematically one notes that, if $d_i S^\beta = 0$, one can then write

$$d_t S^\beta = Dq^\beta / T^\beta$$

and thus

$$d_t S^\alpha = -d_t S^\beta = -Dq^\beta / T^\beta = Dq^\alpha / T^\beta$$

-22-

because Dq^α , the energy received by α in the form of heat, must be the negative of that lost by β . Note, however, that the temperature specified is still T^β , since only in the β subsystem has no entropy creation been assumed ($d_t S^\beta = 0$). Then

If one adds Dq^α / T^β to both sides of the inequality one has

A2.1.4.8 FINAL STATEMENT

If one now considers α as the ‘system’ and β as the ‘surroundings’ the second law can be reformulated in the form:

There exists a state function S , called the entropy of a system, related to the heat Dq absorbed from the surroundings during an infinitesimal change by the relations

where T_{surr} is a positive quantity depending only on the (empirical) temperature of the surroundings. It is understood that for the surroundings $d_t S_{\text{surr}} = 0$. For the integral to have any meaning T_{surr} must be constant, or one must change the surroundings in each step. The above equations can be written in the more compact form

(A2.1.21)

where, in this and subsequent similar expressions, the symbol \geq (‘greater than or equal to’) implies the equality for a reversible process and the inequality for a spontaneous (irreversible) process.

Equation (A2.1.21) includes, as a special case, the statement $dS \geq 0$ for adiabatic processes (for which $Dq = 0$) and, *a fortiori*, the same statement about processes that may occur in an isolated system ($Dq = Dw = 0$). If the universe is an isolated system (an assumption that, however plausible, is not yet subject to experimental verification), the first and second laws lead to the famous statement of Clausius: ‘The energy of the universe is constant; the entropy of the universe tends always toward a maximum.’

It must be emphasized that equation (A2.1.21) permits the entropy of a particular system to decrease; this can occur if more entropy is transferred to the surroundings than is created within the system. The entropy of the system cannot decrease, however, without an equal or greater increase in entropy somewhere else.

-23-

There are many equivalent statements of the second law, some of which involve statements about heat engines and ‘perpetual motion machines of the second kind’ that appear superficially quite different from [equation \(A2.1.21\)](#). They will not be dealt with here, but two variant forms of [equation \(A2.1.21\)](#) may be noted: in

view of the definition $dS = Dq_{\text{rev}}/T_{\text{surr}}$ one can also write for an infinitesimal change

$$Dq_{\text{rev}} \geq Dq$$

and, because $dU = Dq_{\text{rev}} + Dw_{\text{rev}} = Dq + Dw$,

$$Dw_{\text{rev}} \leq Dw.$$

Since w is defined as work done on the system, the minimum amount of work necessary to produce a given change in the system is that in a reversible process. Conversely, the amount of work done by the system on the surroundings is maximal when the process is reversible.

One may note, in concluding this discussion of the second law, that in a sense the zeroth law (thermal equilibrium) presupposes the second. Were there no irreversible processes, no tendency to move toward equilibrium rather than away from it, the concepts of thermal equilibrium and of temperature would be meaningless.

A2.1.5 OPEN SYSTEMS

A2.1.5.1 PERMEABLE WALLS AND THE CHEMICAL POTENTIAL

We now turn to a new kind of boundary for a system, a wall permeable to matter. Molecules that pass through a wall carry energy with them, so [equation \(A2.1.15\)](#) must be generalized to include the change of the energy with a change in the number of moles dn :

$$dU = T dS - p dV + \mu dn. \quad (\text{A2.1.22})$$

Here μ is the ‘chemical potential’ just as the pressure p is a mechanical potential and the temperature T is a thermal potential. A difference in chemical potential $\Delta\mu$ is a driving ‘force’ that results in the transfer of molecules through a permeable wall, just as a pressure difference Δp results in a change in position of a movable wall and a temperature difference ΔT produces a transfer of energy in the form of heat across a diathermic wall. Similarly equilibrium between two systems separated by a permeable wall must require equality of the chemical potential on the two sides. For a multicomponent system, the obvious extension of [equation \(A2.1.22\)](#) can be written

$$dU = T dS - p dV + \sum_i \mu_i dn_i \quad (\text{A2.1.23})$$

-24-

where μ_i and n_i are the chemical potential and number of moles of the i th species. [Equation \(A2.1.23\)](#) can also be generalized to include various forms of work (such as gravitational, electrochemical, electromagnetic, surface formation, etc., as well as the familiar pressure–volume work), in which a generalized force X_j produces a displacement dx_j along the coordinate x_j , by writing

$$dU = T dS + \sum_j X_j dx_j + \sum_i \mu_i dn_i.$$

As a particular example, one may take the electromagnetic work terms of [equation \(A2.1.8\)](#) and write

$$dU = T dS - V(P \cdot dE_0 + M \cdot dB_0) + \sum_i \mu_i dn_i. \quad (\text{A2.1.24})$$

The chemical potential now includes any such effects, and one refers to the *gravochemical potential*, the *electrochemical potential*, etc. For example, if the system consists of a gas extending over a substantial difference in height, it is the gravochemical potential (which includes a term mgh) that is the same at all levels, not the pressure. The electrochemical potential will be considered later.

A2.1.5.2 INTERNAL EQUILIBRIUM

Two subsystems α and β , in each of which the potentials T , p , and all the μ_i s are uniform, are permitted to interact and come to equilibrium. At equilibrium all infinitesimal processes are reversible, so for the overall system ($\alpha + \beta$), which may be regarded as isolated, the quantities conserved include not only energy, volume and numbers of moles, but also entropy, i.e. there is no entropy creation in a system at equilibrium. One now considers an infinitesimal reversible process in which small amounts of entropy $dS^{\alpha \rightarrow \beta}$, volume $dV^{\alpha \rightarrow \beta}$ and numbers of moles $dn_i^{\alpha \rightarrow \beta}$ are transferred from subsystem α to subsystem β . For this reversible change, one may use [equation A2.1.23](#) and write for dU^α and dU^β

$$\begin{aligned} dU^\alpha &= -T^\alpha dS^{\alpha \rightarrow \beta} + p^\alpha dV^{\alpha \rightarrow \beta} - \sum_i \mu_i^\alpha dn_i^{\alpha \rightarrow \beta} \\ dU^\beta &= T^\beta dS^{\alpha \rightarrow \beta} - p^\beta dV^{\alpha \rightarrow \beta} + \sum_i \mu_i^\beta dn_i^{\alpha \rightarrow \beta}. \end{aligned}$$

Combining, one obtains for dU

$$\begin{aligned} dU = dU^\alpha + dU^\beta = 0 &= (T^\beta - T^\alpha) dS^{\alpha \rightarrow \beta} - (p^\beta - p^\alpha) dV^{\alpha \rightarrow \beta} \\ &\quad + \sum_i (\mu_i^\beta - \mu_i^\alpha) dn_i^{\alpha \rightarrow \beta}. \end{aligned}$$

Thermal equilibrium means free transfer (exchange) of energy in the form of heat, mechanical (hydrostatic) equilibrium means free transfer of energy in the form of pressure–volume work, and material equilibrium means free transfer

of energy by the motion of molecules across the boundary. Thus it follows that at equilibrium our choices of $dS^{\alpha \rightarrow \beta}$, $dV^{\alpha \rightarrow \beta}$, and $dn_i^{\alpha \rightarrow \beta}$ are independent and arbitrary. Yet the total energy must be kept unchanged, so the conclusion that the coefficients of $dS^{\alpha \rightarrow \beta}$, $dV^{\alpha \rightarrow \beta}$ and $dn_i^{\alpha \rightarrow \beta}$ must vanish is inescapable.

$$T^\alpha = T^\beta \quad p^\alpha = p^\beta \quad \mu_i^\alpha = \mu_i^\beta.$$

If there are more than two subsystems in equilibrium in the large isolated system, the transfers of S , V and n_i between any pair can be chosen arbitrarily; so it follows that at equilibrium all the subsystems must have the same temperature, pressure and chemical potentials. The subsystems can be chosen as *very small* volume elements, so it is evident that the criterion of internal equilibrium within a system (asserted earlier, but without proof) is uniformity of temperature, pressure and chemical potentials throughout. It has now been

demonstrated conclusively that T , p and μ_i are *potentials*; they are intensive properties that measure ‘levels’; they behave like the (equal) heights of the water surface in two interconnected reservoirs at equilibrium.

A2.1.5.3 INTEGRATION OF DU

Equation (A2.1.23) can be integrated by the following trick: One keeps T , p , and all the chemical potentials μ_i constant and increases the number of moles n_i of each species by an amount $n_i d\xi$ where $d\xi$ is the same fractional increment for each. Obviously one is increasing the size of the system by a factor $(1 + d\xi)$, increasing all the extensive properties (U , S , V , n_i) by this factor and leaving the relative compositions (as measured by the mole fractions) and all other intensive properties unchanged. Therefore, $dS = S d\xi$, $dV = V d\xi$, $dn_i = n_i d\xi$, etc, and

$$dU = U d\xi = TS d\xi - pV d\xi + \sum_i \mu_i n_i d\xi.$$

Dividing by $d\xi$ one obtains

$$U = TS - pV + \sum_i \mu_i n_i. \quad (\text{A2.1.25})$$

Mathematically equation (A2.1.25) is the direct result of the statement that U is homogeneous and of first degree in the extensive properties S , V and n_i . It follows, from a theorem of Euler, that

$$U = (\partial U / \partial S)_{V, n_i} S + (\partial U / \partial V)_{S, n_i} V + \sum_i (\partial U / \partial n_i)_{V, S, n_j} n_i. \quad (\text{A2.1.26})$$

(The expression $(\partial U / \partial n_i)_{V, S, n_j}$ signifies, by common convention, the partial derivative of U with respect to the number of moles n_i of a particular species, holding S , V and the number of moles n_j of all other species ($j \neq i$) constant.)

-26-

Equation (A2.1.26) is equivalent to equation (A2.1.25) and serves to identify T , p , and μ_i as appropriate partial derivatives of the energy U , a result that also follows directly from equation (A2.1.23) and the fact that dU is an exact differential.

$$T = (\partial U / \partial S)_{V, n_i} \quad p = -(\partial U / \partial V)_{S, n_i} \quad \mu = (\partial U / \partial n_i)_{V, S, n_j}.$$

If equation (A2.1.25) is differentiated, one obtains

$$dU = T dS + S dT - p dV - V dp + \sum_i \mu_i dn_i + \sum_i n_i d\mu_i$$

which, on combination with equation (A2.1.23), yields a very important relation between the differentials of the potentials:

$$S dT - V dp + \sum_i n_i d\mu_i = 0. \quad (\text{A2.1.27})$$

The special case of equation (A2.1.27) when T and p are constant ($dT = 0$, $dp = 0$) is called the Gibbs–Duhem equation, so equation (A2.1.27) is sometimes called the ‘generalized Gibbs–Duhem equation’.

A2.1.5.4 ADDITIONAL FUNCTIONS AND DIFFERING CONSTRAINTS

The preceding sections provide a substantially complete *summary* of the fundamental concepts of classical thermodynamics. The basic equations, however, can be expressed in terms of other variables that are frequently more convenient in dealing with experimental situations under which different constraints are applied. It is often not convenient to use S and V as independent variables, so it is useful to define other quantities that are also functions of the thermodynamic state of the system. These include the enthalpy (or sometimes unfortunately called ‘heat content’) $H = U + pV$, the Helmholtz free energy (or ‘work content’) $A = U - TS$ and the Gibbs free energy (or ‘Lewis free energy’, frequently just called the ‘free energy’) $G = A + pV$. The usefulness of these will become apparent as some special situations are considered. In what follows it shall be assumed that there is no entropy creation in the surroundings, whose temperature and pressure can be controlled, so that [equation \(A2.1.19\)](#) and [equation \(A2.1.20\)](#) can be used to determine dw and dq . Moreover, for simplicity, the equations will be restricted to include only pressure–volume work; i.e. to [equation \(A2.1.5\)](#); the extension to other forms of work should be obvious.

(A) CONSTANT-VOLUME (ISOCHORIC) PROCESSES

If there is no volume change ($dV = 0$), then obviously there is no pressure–volume work done ($dw = 0$) irrespective of the pressure, and it follows from [equation \(A2.1.10\)](#) that the change in energy is due entirely to the heat absorbed, which can be designated as q_V :

$$(\Delta V = 0) \quad dU = dq \quad \Delta U = q_V. \quad (\text{A2.1.28})$$

-27-

Note that in this special case, the heat absorbed directly measures a state function. One still has to consider how this constant-volume ‘heat’ is measured, perhaps by an ‘electric heater’, but then is this not really work? Conventionally, however, if work is restricted to pressure–volume work, any remaining contribution to the energy transfers can be called ‘heat’.

(B) CONSTANT-PRESSURE (ISOBARIC) PROCESSES

For such a process the pressure p_{ext} of the surroundings remains constant and is equal to that of the system in its initial and final states. (If there are transient pressure changes within the system, they do not cause changes in the surroundings.) One may then write

$$\begin{aligned} dw &= -p_{\text{ext}} dV \\ dq &= dU + p_{\text{ext}} dV. \end{aligned}$$

However, since $dp_{\text{ext}} = 0$ and the initial and final pressures inside equal p_{ext} , i.e. $\Delta p = 0$ for the change in state,

$$\begin{aligned} (\Delta p = 0) \quad dq &= d(U + p_{\text{ext}} dV) = d(U + pV) = dH \\ \Delta(U + pV) &= \Delta H = q_p. \end{aligned} \quad (\text{A2.1.29})$$

Thus for isobaric processes a new function, the *enthalpy* H , has been introduced and its change ΔH is more directly related to the heat that must have been absorbed than is the energy change ΔU . The same reservations about the meaning of heat absorbed apply in this process as in the constant-volume process.

(C) CONSTANT-TEMPERATURE CONSTANT-VOLUME (ISOTHERMAL–ISOCHORIC) PROCESSES

In analogy to the constant-pressure process, constant temperature is defined as meaning that the temperature T of the surroundings remains constant and equal to that of the system in its initial and final (equilibrium) states. First to be considered are constant-temperature constant-volume processes (again $dw = 0$). For a reversible process

$$(\Delta T = 0, \Delta V = 0) \quad dU = dq_{\text{rev}} = T dS.$$

For an irreversible process, invoking the notion of entropy transfer and entropy creation, one can write

$$dU = dq = T d_t S < T(d_t S + d_i S) = T dS = d(TS) = dq_{\text{rev}} \tag{A2.1.30}$$

which includes the inequality of [equation \(A2.1.21\)](#). Expressed this way the inequality $dU < T dS$ looks like a contradiction of [equation \(A2.1.15\)](#) until one realizes that the right-hand side of [equation \(A2.1.30\)](#) refers to the measurement of the entropy by a totally different process, a reverse (driven) process in which some work must be done on the system. If [equation \(A2.1.30\)](#) is integrated to obtain the isothermal change in state one obtains

$$(\Delta T = 0, \Delta V = 0) \quad \Delta U = q < q_{\text{rev}} = T \Delta S = \Delta(TS)$$

or, rearranging the inequality,

$$\tag{A2.1.31}$$

Thus, for spontaneous processes at constant temperature and volume a new quantity, the *Helmholtz free energy* A , decreases. At equilibrium under such restrictions $dA = 0$.

(D) CONSTANT-TEMPERATURE CONSTANT-PRESSURE (ISOTHERMAL–ISOBARIC) PROCESSES

The constant-temperature constant-pressure situation yields an analogous result. One can write for the reversible process

$$(\Delta T = 0, \Delta p = 0) \quad dU = dq_{\text{rev}} + dw_{\text{rev}} = T dS - p dV$$

and for the irreversible process

$$dU = dq + dw = T d_t S - p dV < T dS - p dV$$

which integrated becomes

$$\begin{aligned} (\Delta T = 0, \Delta p = 0) \quad \Delta U < T \Delta S - p \Delta V = \Delta(TS - pV) \\ \Delta(U + pV - TS) = \Delta G < 0. \end{aligned} \tag{A2.1.32}$$

For spontaneous processes at constant temperature and pressure it is the *Gibbs free energy* G that decreases, while at equilibrium under such conditions $dG = 0$.

More generally, without considering the various possible kinds of work, one can write for an isothermal change in a closed system ($dn_i = 0$)

$$\Delta U = q + w = T\Delta S + \Delta A.$$

Now, as has been shown, $q = T\Delta S$ for an isothermal *reversible* process only; for an isothermal *irreversible* process $\Delta S = \Delta_r S + \Delta_i S$, and $q = T\Delta_r S$. Since $\Delta_i S$ is positive for irreversible changes and zero only for reversible processes, one concludes

$$\begin{array}{lll} q = T\Delta S & w = \Delta A & \text{(isothermal reversible changes)} \\ q < T\Delta S & w > \Delta A & \text{(isothermal irreversible changes).} \end{array}$$

-29-

Another statement of the second law would be: ‘The maximum work from (i.e. $-w$) a given isothermal change in thermodynamic state is obtained when the change in state is carried out reversibly; for irreversible isothermal changes, the work obtained is less’. Thus, in the expression $U = TS + A$, one may regard the TS term as that part of the energy of a system that is unavailable for conversion into work in an isothermal process, while A measures the ‘free’ energy that is available for isothermal conversion into work to be done on the surroundings. In isothermal changes some of A may be transferred quantitatively from one subsystem to another, or it may spontaneously decrease (be destroyed), but it cannot be created. Thus one may transfer the available part of the energy of an isothermal system (its ‘free’ energy) to a more convenient container, but one cannot increase its amount. In an irreversible process some of this ‘free’ energy is lost in the creation of entropy; some capacity for doing work is now irretrievably lost.

The usefulness of the Gibbs free energy G is, of course, that most changes of chemical interest are carried out under constant atmospheric pressure where work done on (or by) the atmosphere is not under the experimenter’s control. In an isothermal–isobaric process (constant T and p), the *maximum* available ‘useful’ work, i.e. work other than pressure–volume work, is $-\Delta G$; indeed Guggenheim (1950) suggested the term ‘useful energy’ for G to distinguish it from the Helmholtz ‘free energy’ A . (Another suggested term for G is ‘free enthalpy’ from $G = H - TS$.) An international recommendation is that A and G simply be called the ‘Helmholtz function’ and the ‘Gibbs function’, respectively.

A2.1.5.5 USEFUL INTERRELATIONS

By differentiating the defining equations for H , A and G and combining the results with [equation \(A2.1.25\)](#) and [equation \(A2.1.27\)](#) for dU and U (which are repeated here) one obtains general expressions for the differentials dH , dA , dG and others. One differentiates the defined quantities on the left-hand side of [equation \(A2.1.34\)](#), [equation \(A2.1.35\)](#), [equation \(A2.1.36\)](#), [equation \(A2.1.37\)](#), [equation \(A2.1.38\)](#) and [equation \(A2.1.39\)](#) and then substitutes the right-hand side of [equation \(A2.1.33\)](#) to obtain the appropriate differential. These are examples of *Legendre transformations*:

$$U = TS - pV + \sum_i \mu_i n_i \quad dU = T dS - p dV + \sum_i \mu_i dn_i. \quad (\text{A2.1.33})$$

$$H = U + pV = TS + \sum_i \mu_i n_i \quad dH = T dS + V dp + \sum_i \mu_i dn_i. \quad (\text{A2.1.34})$$

$$(\text{A2.1.35})$$

$$\begin{aligned}
A &= U - TS = -pV + \sum_i \mu_i n_i & dA &= -S dT - p dV + \sum_i \mu_i dn_i. \\
G &= U + pV - TS = \sum_i \mu_i n_i & dG &= -S dT + V dp + \sum_i \mu_i dn_i.
\end{aligned} \tag{A2.1.36}$$

$$-pV = U - TS - \sum_i \mu_i n_i \quad -d(pV) = -S dT - p dV - \sum_i n_i d\mu_i. \tag{A2.1.37}$$

$$TS = U + pV - \sum_i \mu_i n_i \quad -d(TS) = T dS + V dp - \sum_i n_i d\mu_i. \tag{A2.1.38}$$

$$0 = U - TS + pV - \sum_i \mu_i n_i \quad -d(0) = -S dT + V dp - \sum_i n_i d\mu_i. \tag{A2.1.39}$$

-30-

Equation (A2.1.39) is the ‘generalized Gibbs–Duhem equation’ previously presented (equation (A2.1.27)). Note that the Gibbs free energy is just the sum over the chemical potentials.

If there are other kinds of work, similar expressions apply. For example, with electromagnetic work (equation (A2.1.8)) instead of pressure–volume work, one can write for the Helmholtz free energy

$$dA = -S dT - V(\mathbf{P} \cdot d\mathbf{E}_0 + \mathbf{M} \cdot d\mathbf{B}_0) + \sum_i \mu_i dn_i. \tag{A2.1.40}$$

It should be noted that the differential expressions on the right-hand side of equation (A2.1.33), equation (A2.1.34), equation (A2.1.35), equation (A2.1.36), equation (A2.1.37), equation (A2.1.38), equation (A2.1.39) and equation (A2.1.40) express for each function the appropriate independent variables for that function, i.e. the variables—read constraints—that are kept constant during a spontaneous process.

All of these quantities are state functions, i.e. the differentials are exact, so each of the coefficients is a partial derivative. For example, from equation (A2.1.35) $p = -(\partial A/\partial V)_{T,n_i}$, while from equation (A2.1.36) $S = -(\partial G/\partial T)_{p,n_i}$. Moreover, because the order of partial differentiation is immaterial, one obtains as cross-differentiation identities from equation (A2.1.33), equation (A2.1.34), equation (A2.1.35), equation (A2.1.36), equation (A2.1.37), equation (A2.1.38), equation (A2.1.39) and equation (A2.1.40) a whole series of useful equations usually known as ‘Maxwell relations’. A few of these are: from equation (A2.1.33):

$$\begin{aligned}
(\partial^2 U/\partial S \partial V)_{n_i} &= \partial(\partial U/\partial V)/\partial S = \partial(\partial U/\partial S)/\partial V \\
&= -(\partial p/\partial S)_{V,n_i} = (\partial T/\partial V)_{S,n_i}
\end{aligned}$$

from equation (A2.1.35):

$$(\partial^2 A/\partial T \partial V)_{n_i} = -(\partial S/\partial V)_{T,n_i} = -(\partial p/\partial T)_{V,n_i} \tag{A2.1.41}$$

from equation (A2.1.36):

$$(\partial^2 G/\partial T \partial p)_{n_i} = -(\partial S/\partial p)_{T,n_i} = (\partial V/\partial T)_{p,n_i} \tag{A2.1.42}$$

and from equation (A2.1.40)

$$\begin{aligned}(\partial^2 A/\partial T \partial E_0)_n &= -(\partial S/\partial E_0)_{T,n} = -V(\partial P_0/\partial T)_{E_0,n} \\(\partial^2 A/\partial T \partial B_0)_n &= -(\partial S/\partial B_0)_{T,n} = -V(\partial M_0/\partial T)_{B_0,n}.\end{aligned}\tag{A2.1.43}$$

-31-

(Strictly speaking, differentiation with respect to a vector quantity is not allowed. However for the isotropic spherical samples for which [equation \(A2.1.8\)](#) is appropriate, the two vectors have the same direction and could have been written as scalars; the vector notation was kept to avoid confusion with other thermodynamic quantities such as energy, pressure, etc. It should also be noted that the Maxwell equations above are correct for either of the choices for electromagnetic work discussed earlier; under the other convention A is replaced by a generalized G .)

A2.1.5.6 FEATURES OF EQUILIBRIUM

Earlier in this section it was shown that, when a constraint, e.g. fixed l , was released in a system for which U , V and n were held constant, the entropy would seek a maximum value consistent with the remaining restrictions (e.g. $dS/dl = 0$ and $d^2S/dl^2 < 0$). One refers to this, a result of [equation \(A2.1.33\)](#), as a ‘feature of equilibrium’. We can obtain similar features of equilibrium under other conditions from [equation \(A2.1.34\)](#), [equation \(A2.1.35\)](#), [equation \(A2.1.36\)](#), [equation \(A2.1.37\)](#), [equation \(A2.1.38\)](#) and [equation \(A2.1.39\)](#). Since at equilibrium all processes are reversible, all these equations are valid at equilibrium. Each equation is a linear relation between differentials; so, if all but one are fixed equal to zero, at equilibrium the remaining differential quantity must also be zero. That is to say, the function of which it is the differential must have an equilibrium value that is either maximized or minimized and it is fairly easy, in any particular instance, to decide between these two possibilities. To summarize the more important of these equilibrium features:

for fixed U, V, n_i	S is a maximum
for fixed H, p, n_i	S is a maximum
for fixed S, V, n_i	U is a minimum
for fixed S, p, n_i	H is a minimum
for fixed T, V, n_i	A is a minimum
for fixed T, p, n_i	G is a minimum.

Of these the last condition, minimum Gibbs free energy at constant temperature, pressure and composition, is probably the one of greatest practical importance in chemical systems. (This list does not exhaust the mathematical possibilities; thus one can also derive other apparently unimportant conditions such as that at constant U, S and n_i , V is a minimum.) However, an experimentalist will wonder how one can hold the entropy constant and release a constraint so that some other state function seeks a minimum.

A2.1.5.7 THE CHEMICAL POTENTIAL AND PARTIAL MOLAR QUANTITIES

From [equation \(A2.1.33\)](#), [equation \(A2.1.34\)](#), [equation \(A2.1.35\)](#) and [equation \(A2.1.36\)](#) it follows that the chemical potential may be defined by any of the following relations:

$$\begin{aligned}\mu_i &= (\partial U/\partial n_i)_{S,V,n_j} = (\partial H/\partial n_i)_{S,p,n_j} \\ &= (\partial A/\partial n_i)_{T,V,n_j} = (\partial G/\partial n_i)_{S,V,n_j} = \tilde{G}_i.\end{aligned}\tag{A2.1.44}$$

-32-

In experimental work it is usually most convenient to regard temperature and pressure as the independent variables, and for this reason the term *partial molar quantity* (denoted by a bar above the quantity) is always restricted to the derivative with respect to n_i holding T , p , and all the other n_j constant. (Thus $\bar{V}_i = (\partial V/\partial n_i)_{T,p,n_j}$.) From the right-hand side of [equation \(A2.1.44\)](#) it is apparent that the chemical potential is the same as the partial molar Gibbs free energy \bar{G}_i and, therefore, some books on thermodynamics, e.g. Lewis and Randall (1923), do not give it a special symbol. Note that the partial molar Helmholtz free energy is *not* the chemical potential; it is

$$\bar{A}_i = (\partial A/\partial n_i)_{T,p,n_j} = [\partial(G - pV)/\partial n_i]_{T,p,n_j} = \mu_i - p\bar{V}_i.$$

On the other hand, in the theoretical calculations of statistical mechanics, it is frequently more convenient to use volume as an independent variable, so it is important to preserve the general importance of the chemical potential as something more than a quantity \bar{G}_i whose usefulness is restricted to conditions of constant temperature and pressure.

From cross-differentiation identities one can derive some additional Maxwell relations for partial molar quantities:

$$\begin{aligned} (\partial^2 G/\partial T \partial n_i)_{p,n_j} &= (\partial \mu_i/\partial T)_{p,n_i} = -(\partial S/\partial n_i)_{p,n_j} = -\bar{S}_i \\ (\partial^2 G/\partial p \partial n_i)_{T,n_j} &= (\partial \mu_i/\partial p)_{T,n_i} = (\partial V/\partial n_i)_{T,p,n_j} = \bar{V}_i. \end{aligned}$$

In passing one should note that the method of expressing the chemical potential is arbitrary. The amount of matter of species i in this article, as in most thermodynamics books, is expressed by the number of moles n_i ; it can, however, be expressed equally well by the number of molecules N_i (convenient in statistical mechanics) or by the mass m_i (Gibbs' original treatment).

A2.1.5.8 SOME ADDITIONAL IMPORTANT QUANTITIES

As one raises the temperature of the system along a particular path, one may define a *heat capacity* $C_{\text{path}} = Dq_{\text{path}}/dT$. (The term 'heat capacity' is almost as unfortunate a name as the obsolescent 'heat content' for H ; alas, no alternative exists.) However several such paths define state functions, e.g. [equation \(A2.1.28\)](#) and [equation \(A2.1.29\)](#). Thus we can define the heat capacity at constant volume C_V and the heat capacity at constant pressure C_p as

$$C_V = (\partial U/\partial T)_{V,n_i} = T(\partial S/\partial T)_{V,n_i} \tag{A2.1.45}$$

$$C_p = (\partial H/\partial T)_{p,n_i} = T(\partial S/\partial T)_{p,n_i}. \tag{A2.1.46}$$

The right-hand equalities in these two equations arise directly from [equation \(A2.1.33\)](#) and [equation \(A2.1.34\)](#).

Two other important quantities are the *isobaric expansivity* ('coefficient of thermal expansion') α_p and the *isothermal compressibility* κ_T , defined as

$$\alpha_p = (1/V)(\partial V/\partial T)_{p,n_i}$$

$$\kappa_T = -(1/V)(\partial V/\partial p)_{T,n_i}.$$

The adjectives ‘isobaric’ and ‘isothermal’ and the corresponding subscripts are frequently omitted, but it is important to distinguish between the isothermal compressibility and the *adiabatic compressibility*.

A relation between C_p and C_V can be obtained by writing

$$\begin{aligned}(\partial S/\partial T)_{p,n_i} &= (\partial S/\partial T)_{V,n_i} + (\partial S/\partial V)_{T,n_i}(\partial V/\partial T)_{p,n_i} \\ &= (\partial S/\partial T)_{V,n_i} + (\partial p/\partial T)_{V,n_i}(\partial V/\partial T)_{p,n_i} \\ (\partial p/\partial T)_{V,n_i} &= -(\partial p/\partial V)_{T,n_i}(\partial V/\partial T)_{p,n_i} = \alpha_p/\kappa_T \quad \text{(from the cyclic rule).}\end{aligned}$$

Combining these, we have

$$(\partial H/\partial T)_{p,n_i} - (\partial U/\partial T)_{V,n_i} = T(\partial p/\partial T)_{V,n_i}(\partial V/\partial T)_{p,n_i}$$

or

$$C_p - C_V = TV\alpha_p^2/\kappa_T. \quad (\text{A2.1.47})$$

For the special case of the ideal gas (equation (A2.1.17)), $\alpha_p = 1/T$ and $\kappa_T = 1/p$,

$$C_p - C_V = TVp/T^2 = nR \quad \text{(ideal gas only).}$$

A similar derivation leads to the difference between the isothermal and adiabatic compressibilities:

$$\kappa_T - \kappa_S = TV\alpha_p^2/C_p. \quad (\text{A2.1.48})$$

A2.1.5.9 THERMODYNAMIC EQUATIONS OF STATE

Two exact equations of state can be derived from equation (A2.1.33) and equation (A2.1.34)

$$\begin{aligned}(\partial U/\partial V)_{T,n_i} &= -p + T(\partial S/\partial V)_{T,n_i} = -p + T(\partial p/\partial T)_{V,n_i} \\ \text{or } p &= T(\partial p/\partial T)_{V,n_i} - (\partial U/\partial V)_{T,n_i}.\end{aligned} \quad (\text{A2.1.49})$$

-34-

$$\begin{aligned}(\partial H/\partial p)_{T,n_i} &= V + T(\partial S/\partial p)_{T,n_i} = V - T(\partial V/\partial T)_{p,n_i} \\ \text{or } V &= T(\partial V/\partial T)_{p,n_i} + (\partial H/\partial p)_{T,n_i}.\end{aligned} \quad (\text{A2.1.50})$$

It is interesting to note that, when the van der Waals equation for a fluid,

$$p = nRT/(V - nb) - n^2a/V^2,$$

is compared with [equation \(A2.1.49\)](#), the right-hand sides separate in the same way:

$$T(\partial p/\partial T)_{V,n} = nRT/(V - nb) \quad \text{and} \quad (\partial U/\partial V)_{T,n} = n^2a/V^2.$$

A2.1.6 APPLICATIONS

A2.1.6.1 PHASE EQUILIBRIA

When two or more phases, e.g. gas, liquid or solid, are in equilibrium, the principles of internal equilibrium developed in [section A2.1.5.2](#) apply. If transfers between two phases α and β can take place, the appropriate potentials must be equal, even though densities and other properties can be quite different.

$$T^\alpha = T^\beta \quad p^\alpha = p^\beta \quad \mu_i^\alpha = \mu_i^\beta.$$

As shown in preceding sections, one can have equilibrium of some kinds while inhibiting others. Thus, it is possible to have thermal equilibrium ($T^\alpha = T^\beta$) through a fixed impermeable diathermic wall; in such a case p^α need not equal p^β , nor need μ_i^α equal μ_i^β . It is possible to achieve mechanical equilibrium ($p^\alpha = p^\beta$) through a movable impermeable adiabatic wall; in such a case the transfer of heat or matter is prevented, so T and μ_i can be different on opposite sides. It is possible to have both thermal and mechanical equilibrium ($p^\alpha = p^\beta$, $T^\alpha = T^\beta$) through a movable diathermic wall. For a one-component system $\mu = f(T, p)$, so $\mu^\alpha = \mu^\beta$ even if the wall is impermeable. However, for a system of two or more components one can have $p^\alpha = p^\beta$ and $T^\alpha = T^\beta$, but the chemical potential is now also a function of composition, so μ_i^α need not equal μ_i^β . It does not seem experimentally possible to permit material equilibrium ($\mu_i^\alpha = \mu_i^\beta$) without simultaneously achieving thermal equilibrium ($T^\alpha = T^\beta$).

Finally, in membrane equilibria, where the wall is permeable to some species, e.g. the solvent, but not others, thermal equilibrium ($T^\alpha = T^\beta$) and solvent equilibrium ($\mu_i^\alpha = \mu_i^\beta$) are found, but $\mu_j^\alpha \neq \mu_j^\beta$ and $p^\alpha \neq p^\beta$; the difference $p^\beta - p^\alpha$ is the osmotic pressure.

-35-

For a one-component system, $\Delta G^{\alpha \rightarrow \beta} = \mu^\beta - \mu^\alpha = 0$, so one may write

$$\Delta G^{\alpha \rightarrow \beta} = \Delta H^{\alpha \rightarrow \beta} - T \Delta S^{\alpha \rightarrow \beta} = 0 \quad \text{or} \quad \Delta S^{\alpha \rightarrow \beta} = \Delta H^{\alpha \rightarrow \beta} / T. \quad (\text{A2.1.51})$$

THE CLAPEYRON EQUATION

Moreover, using the generalized Gibbs–Duhem [equations \(A2.1.27\)](#) for each of the two one-component phases,

$$S^\alpha dT - V^\alpha dp + n^\alpha d\mu = 0$$

or

$$d\mu = \bar{V}^\alpha dp - \bar{S}^\alpha dT = \bar{V}^\beta dp - \bar{S}^\beta dT$$

one obtains the Clapeyron equation for the change of pressure with temperature as the two phases continue to coexist:

$$dp/dT = \Delta \bar{S}^{\alpha \rightarrow \beta} / \Delta \bar{V}^{\alpha \rightarrow \beta} = \Delta \bar{H}^{\alpha \rightarrow \beta} / T \Delta \bar{V}^{\alpha \rightarrow \beta}. \quad (\text{A2.1.52})$$

The analogue of the Clapeyron equation for multicomponent systems can be derived by a complex procedure of systematically eliminating the various chemical potentials, but an alternative derivation uses the Maxwell relation (A2.1.41)

$$(\partial^2 A / \partial T \partial V)_{n_i} = -(\partial S / \partial V)_{T, n_i} = -(\partial p / \partial T)_{V, n_i}. \quad (\text{A2.1.41})$$

Applied to a two-phase system, this says that the change in pressure with temperature is equal to the change in entropy at constant temperature as the total volume of the system ($\alpha + \beta$) is increased, which can only take place if some α is converted to β :

$$dp/dT = \Delta S^{\alpha \rightarrow \beta} / \Delta V^{\alpha \rightarrow \beta} = \Delta H^{\alpha \rightarrow \beta} / T \Delta V^{\alpha \rightarrow \beta}.$$

In this case, whatever n_i moles of each species are required to accomplish the ΔV are the same n_i s that determine ΔS or ΔH . Note that this general equation includes the special one-component case of equation (A2.1.52).

When, for a one-component system, one of the two phases in equilibrium is a sufficiently dilute gas, i.e. is at a pressure well below 1 atm, one can obtain a very useful approximate equation from equation (A2.1.52). The molar volume of the gas is at least two orders of magnitude larger than that of the liquid or solid, and is very nearly an ideal gas. Then one can write

$$\Delta \bar{V}^{l \rightarrow g} \approx \bar{V}^g \approx RT/p$$

-36-

which can be substituted into [equation \(A2.1.52\)](#) to obtain

$$dp/dT \approx \Delta \bar{H}^{l \rightarrow g} (p/RT^2)$$

or

$$d \ln p / dT \approx \Delta \bar{H}^{l \rightarrow g} / RT^2 = \Delta \bar{H}_{\text{vap}} / RT^2 \quad (\text{A2.1.53})$$

or

$$d \ln p / d(1/T) \approx -\Delta \bar{H}_{\text{vap}} / R,$$

where $\Delta \bar{H}_{\text{vap}}$ is the molar enthalpy of vaporization at the temperature T . The corresponding equation for the vapour pressure of the solid is identical except for the replacement of the enthalpy of vaporization by the

enthalpy of sublimation.

(Equation (A2.1.53) is frequently called the *Clausius–Clapeyron equation*, although this name is sometimes applied to [equation \(A2.1.52\)](#). Apparently Clapeyron first proposed [equation \(A2.1.52\)](#) in 1834, but it was derived properly from thermodynamics decades later by Clausius, who also obtained the approximate equation (A2.1.53).)

It is interesting and surprising to note that, although the molar enthalpy $\Delta \bar{H}_{\text{vap}}$ and the molar volume of vaporization $\Delta \bar{V}_{\text{vap}}$ both decrease to zero at the critical temperature of the fluid (where the fluid is very non-ideal), a plot of $\ln p$ against $1/T$ for most fluids is very nearly a straight line all the way from the melting point to the critical point. For example, for krypton, the slope $d \ln p/d(1/T)$ varies by less than 1% over the entire range of temperatures; even for water the maximum variation of the slope is only about 15%.

THE PHASE RULE

Finally one can utilize the generalized Gibbs–Duhem [equations \(A2.1.27\)](#) for each phase

$$S^\alpha dT - V^\alpha dp + \sum_i n_i^\alpha d\mu_i = 0$$

$$S^\beta dT - V^\beta dp + \sum_i n_i^\beta d\mu_i = 0$$

etc to obtain the ‘Gibbs phase rule’. The number of variables (potentials) equals the number of components \mathcal{C} plus two (temperature and pressure), and these are connected by an equation for each of the \mathcal{P} phases. It follows that the number of potentials that can be varied independently (the ‘degrees of freedom’ \mathcal{F}) is the number of variables minus the number of equations:

-37-

$$\mathcal{F} = \mathcal{C} + 2 - \mathcal{P}.$$

From this equation one concludes that the maximum number of phases that can coexist in a one-component system ($\mathcal{C} = 1$) is three, at a unique temperature and pressure ($\mathcal{F} = 0$). When two phases coexist ($\mathcal{F} = 1$), selecting a temperature fixes the pressure. Conclusions for other situations should be obvious.

A2.1.6.2 REAL AND IDEAL GASES

Real gases follow the ideal-gas [equation \(A2.1.17\)](#) only in the limit of zero pressure, so it is important to be able to handle the thermodynamics of real gases at non-zero pressures. There are many semi-empirical equations with parameters that purport to represent the physical interactions between gas molecules, the simplest of which is the van der Waals [equation \(A2.1.50\)](#). However, a completely general form for expressing gas non-ideality is the series expansion first suggested by Kamerlingh Onnes (1901) and known as the *virial equation of state*:

$$pV/nRT = 1 + B(n/V) + C(n/V)^2 + D(n/V)^3 + \dots$$

The equation is more conventionally written expressing the variable n/V as the inverse of the molar volume, $1/\bar{V}$, although n/V is just the molar concentration c , and one could equally well write the equation as

$$p/RT = c + Bc^2 + Cc^3 + Dc^4 + \dots \quad (\text{A2.1.54})$$

The coefficients B , C , D , etc for each particular gas are termed its second, third, fourth, etc. *virial coefficients*, and are functions of the temperature only. It can be shown, by statistical mechanics, that B is a function of the interaction of an isolated pair of molecules, C is a function of the simultaneous interaction of three molecules, D , of four molecules, etc., a feature suggested by the form of equation (A2.1.54).

While volume is a convenient variable for the calculations of theoreticians, the pressure is normally the variable of choice for experimentalists, so there is a corresponding equation in which the equation of state is expanded in powers of p :

$$pV/n = RT + B'p + C'p^2 + D'p^3 + \dots \quad (\text{A2.1.55})$$

The pressure coefficients can be related to the volume coefficients by reverting the series and one finds that

$$B' = B \quad C' = (C - B^2)/RT \quad D' = (D - 3BC + 2B^3)/(RT)^2 \quad \text{etc.}$$

According to [equation \(A2.1.39\)](#) $(\partial\mu/\partial p)_T = V/n$, so equation (A2.1.55) can be integrated to obtain the chemical potential:

-38-

$$\mu(T, p) - \mu^0(T, p^0) = RT \ln(p/p^0) + B'p + C'p^2/2 + D'p^3/3 + \dots \quad (\text{A2.1.56})$$

Note that a constant of integration μ^0 has come into the equation; this is the chemical potential of the hypothetical *ideal gas* at a reference pressure p^0 , usually taken to be one atmosphere. In principle this involves a process of taking the real gas down to zero pressure and bringing it back to the reference pressure as an ideal gas. Thus, since $d\mu = (V/n) dp$, one may write

$$\mu(T, p) - \mu^0(T, p^0) = \int_0^{p^0} [(V/n) - (RT/p)] dp = B'p^0 + C'(p^0)^2 + \dots$$

If $p^0 = 1$ atm, it is sufficient to retain only the first term on the right. However, one does not need to know the virial coefficients; one may simply use volumetric data to evaluate the integral.

The molar entropy and the molar enthalpy, also with constants of integration, can be obtained, either by differentiating equation (A2.1.56) or by integrating [equation \(A2.1.42\)](#) or [equation \(A2.1.50\)](#):

$$\begin{aligned} \bar{S}(T, p) - \bar{S}^0(T, p^0) &= -R \ln(p/p^0) - (dB'/dT)p - (dC'/dT)p^2/2 - (dD'/dT)p^3/3 - \dots \\ \bar{H}(T, p) - \bar{H}^0(T, p^0) &= [B' - T(dB'/dT)]p + [C' - T(dC'/dT)]p^2/2 + [D' - T(dD'/dT)]p^3/3 - \dots \end{aligned} \quad (\text{A2.1.57})$$

where, as in the case of the chemical potential, the reference molar entropy \bar{S}^0 and reference molar enthalpy \bar{H}^0 are for the hypothetical ideal gas at a pressure p^0 .

It is sometimes convenient to retain the generality of the limiting ideal-gas equations by introducing the *activity* a , an 'effective' pressure (or, as we shall see later in the case of solutions, an effective mole fraction,

concentration, or molality). For gases, after Lewis (1901), this is usually called the *fugacity* and symbolized by f rather than by a . One can then write

$$\mu(T, p) - \mu^0(T, p^0) = RT \ln(f/p^0).$$

One can also define an *activity coefficient* or *fugacity coefficient* $\gamma = fp$; obviously

$$RT \ln \gamma = B'p + C'p^2/2 + D'p^3/3 + \dots$$

TEMPERATURE DEPENDENCE OF THE SECOND VIRIAL COEFFICIENT

Figure A2.1.7 shows schematically the variation of $B = B'$ with temperature. It starts strongly negative (theoretically at minus infinity for zero temperature, but of course unmeasurable) and decreases in magnitude until it changes sign at the *Boyle temperature* ($B = 0$, where the gas is more nearly ideal to higher pressures). The slope dB/dT remains

-39-

positive, but decreases in magnitude until very high temperatures. Theory requires the virial coefficient finally to reach a maximum and then slowly decrease, but this has been experimentally observed only for helium.

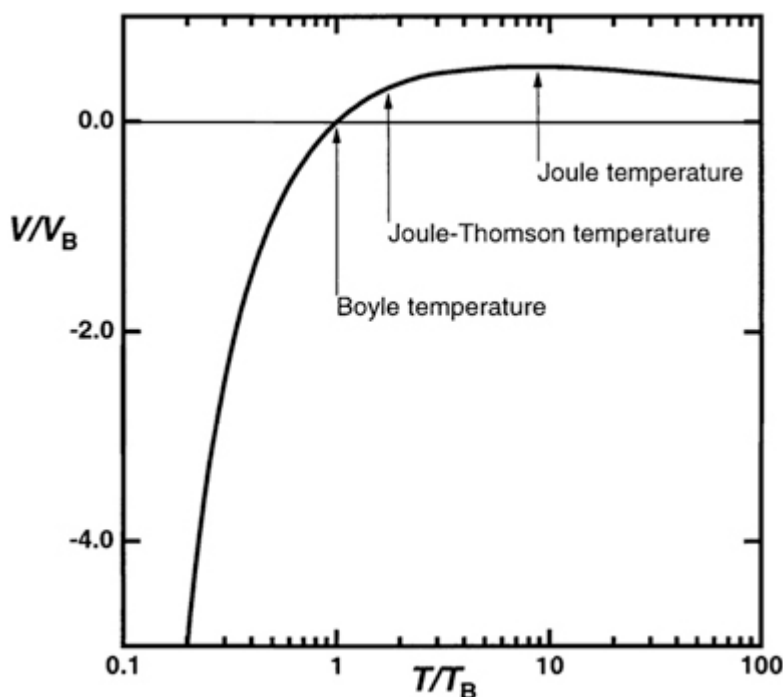


Figure A2.1.7. The second virial coefficient B as a function of temperature T/T_B . (Calculated for a gas satisfying the Lennard-Jones potential [8].)

It is widely believed that gases are virtually ideal at a pressure of one atmosphere. This is more nearly true at relatively high temperatures, but at the normal boiling point (roughly 20% of the Boyle temperature), typical gases have values of pV/nRT that are 5 to 15% lower than the ideal value of unity.

THE JOULE–THOMSON EFFECT

One of the classic experiments on gases was the measurement by Joule and Thomson (1853) of the change in temperature when a gas flows through a porous plug (throttling valve) from a pressure p_1 to a pressure p_2 (figure A2.1.8). The total system is jacketed in such a way that the process is adiabatic ($q = 0$), and the pressures are constant (other than an infinitesimal δp) in the two parts. The work done on the gas in the right-hand region to bring it through is $p_2 dV_2$, while that in the left-hand region is $-p_1 dV_1$ (because dV_1 is negative). The two volumes are of course unequal, but no assumption about the ideality of the gas is necessary (or even desirable). The total energy change can then be written as the loss of energy from the left-hand region plus the gain in energy by the right-hand region

$$dU = -dU_1 + dU_2 = p_1 dV_1 - p_2 dV_2$$

-40-

$g(\omega)$ is essentially the density of states and the above expression corresponds to the Debye model.

In general, the phonon density of states $g(\omega)$, $d\omega$ is a complicated function which can be directly measured from experiments, or can be computed from the results from computer simulations of a crystal. The explicit analytic expression of $g(\omega)$ for the Debye model is a consequence of the two assumptions that were made above for the frequency and velocity of the elastic waves. An even simpler assumption about $g(\omega)$ leads to the Einstein model, which first showed how quantum effects lead to deviations from the classical equipartition result as seen experimentally. In the Einstein model, one assumes that only one level at frequency ω_E is appreciably populated by phonons so that $g(\omega) = \delta(\omega - \omega_E)$ and, for each of the Einstein modes, ω_E/k_b is called the Einstein temperature θ_E .

High-temperature behaviour. Consider T much higher than a characteristic temperature like θ_D or θ_E . Since $\beta \hbar\omega$ is then small compared to 1, one can expand the exponential to obtain

$$\frac{\hbar\omega}{e^{\beta\hbar\omega} - 1} \approx \frac{1}{\beta}$$

and

$$U = \sum_n \frac{\hbar\omega_n}{e^{\beta\hbar\omega_n} - 1} \approx k_b T \sum_n 1 = 3Nk_b T \quad (\text{A2.2.95})$$

as expected by the equipartition law. This leads to a value of $3Nk_b$ for the heat capacity C_V . This is known as the Dulong and Petit's law.

Low-temperature behaviour. In the Debye model, when $T \ll \theta_D$, the upper limit, x_D , can be approximately replaced by ∞ , the integral over x then has a value $\pi^4/15$ and the total phonon energy reduces to

$$U(T) \approx \frac{3V}{2\pi^2 v^3 \hbar^3 \beta^4} \frac{\pi^4}{15} = \frac{3\pi^3 N k_b}{5\theta_D^3} T^4$$

proportional to T^4 . This leads to the heat capacity, for $T \ll \theta_D$,

$$C_V = \left(\frac{\partial U}{\partial T} \right)_{V,N} = \frac{12\pi^4 N k_b}{5\theta_D^3} T^3 \equiv A_{ph} T^3. \quad (\text{A2.2.96})$$

This result is called the Debye T^3 law. [Figure A2.2.4](#) compares the experimental and Debye model values for the heat capacity C_p . It also gives Debye temperatures for various solids. One can also evaluate C_V for the Einstein model: as expected it approaches the equipartition result at high temperatures but decays exponentially to zero as T goes to zero.

-41-

$$(\partial T/\partial V)_U = -(\partial U/\partial V)_T/(\partial U/\partial T)_V = -RT^2[(dB/dT)(n/V)^2 + \dots]/C_V.$$

Unlike $(\partial H/\partial p)_T$, $(\partial U/\partial V)_T$ does indeed vanish for real gases as the pressure goes to zero, but this is because the derivative is with respect to V , not because of the difference between U and H . At appreciable pressures $(\partial T/\partial V)_U$ is almost invariably negative, because the Joule temperature, at which dB/dT becomes negative, is extremely high (see [figure A2.1.7](#)).

A2.1.6.3 GASEOUS MIXTURES

According to Dalton's *law of partial pressures*, observed experimentally at sufficiently low pressures, the pressure of a gas mixture in a given volume V is the sum of the pressures that each gas would exert alone in the same volume at the same temperature. Expressed in terms of moles n_i

$$p(n_1, n_2, \dots, n_j, V, T) = \sum_{i=1}^j p_i(n_i, V, T),$$

or, given the validity of the ideal-gas law ([equation \(A2.1.18\)](#)) at these pressures,

$$p(n_1, n_2, \dots, n_j, V, T) = \sum_i (n_i RT/V) = \left(\sum_i n_i \right) (RT/V).$$

The *partial pressure* p_i of a component in an ideal-gas mixture is thus

$$p_i = \left(n_i / \sum_i n_i \right) p = x_i p \quad (\text{A2.1.59})$$

where $x_i = n_i/n$ is the mole fraction of species i in the mixture. (The partial pressure is always defined by [equation \(A2.1.59\)](#) even at higher pressures where Dalton's law is no longer valid.)

Given this experimental result, it is plausible to assume (and is easily shown by statistical mechanics) that the chemical potential of a substance with partial pressure p_i in an ideal-gas mixture is equal to that in the one-component ideal gas at pressure $p' = p_i$

$$\mu_i(p, T, x_i) = \mu'_i(p' = x_i p, T, x'_i = 1). \quad (\text{A2.1.60})$$

What thermodynamic experiments can be cited to support such an assumption? There are several:

- (1) There are a few semipermeable membranes that separate a gas mixture from a pure component gas. One is palladium, which is permeable to hydrogen, but not (in any reasonable length of time) to other gases. Another is rubber, through which carbon dioxide or ammonia diffuses rapidly, while gases like nitrogen or argon diffuse much more slowly. In such cases, at equilibrium (when the chemical potential of the diffusing gas must be the same on both sides of the membrane) the pressure of the one-component gas on one side of the membrane is found to be equal to its partial pressure in the gas mixture on the other side.
- (2) In the phase equilibrium between a pure solid (or a liquid) and its vapour, the addition of other gases, as long as they are insoluble in the solid or liquid, has negligible effect on the partial pressure of the vapour.
- (3) In electrochemical cells (to be discussed later), if a particular gas participates in a chemical reaction at an electrode, the observed electromotive force is a function of the partial pressure of the reactive gas and not of the partial pressures of any other gases present.

For precise measurements, there is a slight correction for the effect of the slightly different pressure on the chemical potentials of the solid or of the components of the solution. More important, corrections must be made for the non-ideality of the pure gas and of the gaseous mixture. With these corrections, [equation \(A2.1.60\)](#) can be verified within experimental error.

Given [equation \(A2.1.60\)](#) one can now write for an ideal-gas mixture

$$\begin{aligned}
 \mu_i(p, T, x_i) &= \mu_i' = \mu_i^0(p^0, T) + RT \ln(p_i/p^0) \\
 &= \mu_i^0(p^0, T) + RT \ln(x_i p/p^0) \\
 &= \mu_i^0(p^0, T) + RT \ln(p/p^0) + RT \ln x_i.
 \end{aligned}
 \tag{A2.1.61}$$

Note that this has resulted in the separation of pressure and composition contributions to chemical potentials in the ideal-gas mixture. Moreover, the thermodynamic functions for ideal-gas mixing at constant pressure can now be obtained:

$$\left. \begin{aligned}
 \Delta G_m(T, p) &= \sum_i n_i [\mu_i(T, p, x_i) - \mu_i(T, p, 1)] = RT \sum_i n_i \ln x_i \\
 \Delta S_m(T, p) &= -(\partial \Delta G_m / \partial T)_{p, n_i} = -R \sum_i n_i \ln x_i \\
 \Delta H_m(T, p) &= \Delta G_m(T, p) + T \Delta S_m(T, p) = 0.
 \end{aligned} \right\} \text{ideal gas only}$$

Gas mixtures are subject to the same degree of non-ideality as the one-component ('pure') gases that were discussed in the previous section. In particular, the second virial coefficient for a gas mixture can be written as a quadratic average

$$B(T, x_1, \dots, x_k) = \sum_{i=1}^k \sum_{j=1}^k x_i x_j B_{ij}.
 \tag{A2.1.62}$$

where B_{ij} , a function of temperature only, depends on the interaction of an i, j pair. Thus, for a binary mixture

of gases, one has B_{11} and B_{22} from measurements on the pure gases, but one needs to determine B_{12} as well. The corresponding third virial coefficient is a cubic average over the C_{ijk} s, but this is rarely needed. Appropriate differentiation of [equation \(A2.1.62\)](#) will lead to the non-ideal corrections to the equations for the chemical potentials and the mixing functions.

A2.1.6.4 DILUTE SOLUTIONS AND HENRY'S LAW

Experiments on sufficiently dilute solutions of non-electrolytes yield *Henry's law*, that the vapour pressure of a volatile solute, i.e. its partial pressure in a gas mixture in equilibrium with the solution, is directly proportional to its concentration, expressed in *any* units (molar concentrations, molality, mole fraction, weight fraction, etc.) because in sufficiently dilute solution these are all proportional to each other.

$$p_i = k_c c_i = k_m m_i = k_x x_i$$

where c_i is the molar concentration of species i (conventionally, but not necessarily, expressed in units of moles per litre of *solution*), m_i is its *molality* (conventionally expressed as moles per kilogram of *solvent*), and x_i is its mole fraction. The Henry's law constants k_c , k_m and k_x differ, of course, with the choice of units.

It follows that, because phase equilibrium requires that the chemical potential μ_i be the same in the solution as in the gas phase, one may write for the chemical potential in the solution:

$$\mu_i(T, c_i) - \mu_i(T, c^0) = RT \ln(c_i/c^0). \quad (\text{A2.1.63})$$

Here the composition is expressed as concentration c_i and the reference state is for unit concentration c^0 (conventionally 1 mol l^{-1}) but it could have been expressed using any other composition variable and the corresponding reference state.

It seems appropriate to assume the applicability of [equation \(A2.1.63\)](#) to sufficiently dilute solutions of non-volatile solutes and, indeed, to electrolyte species. This assumption can be validated by other experimental methods (e.g. by electrochemical measurements) and by statistical mechanical theory.

Just as increasing the pressure of a gas or a gas mixture introduces non-ideal corrections, so does increasing the concentration. As before, one can introduce an activity a_i and an activity coefficient γ_i and write $a_i = c_i \gamma_i$ and

$$\mu_i(T, c_i) - \mu_i(T, c^0) = RT \ln(a_i/c^0) = RT \ln(c_i/c^0) + RT \ln \gamma_i.$$

In analogy to the gas, the reference state is for the ideally dilute solution at c^0 , although at c^0 the real solution may be far from ideal. (Technically, since this has now been extended to non-volatile solutes, it is defined at the reference pressure p^0 rather than at the vapour pressure; however, because $(\partial \mu_i / \partial p)_T = \bar{V}_i$, and molar volumes are small in condensed systems, this is rarely of any consequence.)

Using the Gibbs–Duhem equation ([\(A2.1.27\)](#) with $dT = 0$, $dp = 0$), one can show that the solvent must obey *Raoult's law* over the same concentration range where Henry's law is valid for the solute (or solutes):

$$p_0 = p_0^0 x_0.$$

where x_0 is the mole fraction of solvent, p_0 is its vapour pressure, and p_0^0 is the vapour pressure of pure solvent, i.e. at $x_0 = 1$. A more careful expression of Raoult's law might be

$$\lim_{x_0 \rightarrow 1} (\partial p_0 / \partial x_0) = p_0^0.$$

It should be noted that, whatever the form of Henry's law (i.e. in whatever composition units), Raoult's law must necessarily be expressed in mole fraction. This says nothing about the appropriateness of mole fractions in condensed systems, e.g. in equilibrium expressions; it arises simply from the fact that it is a statement about the *gas phase*.

The reference state for the solvent is normally the pure solvent, so one may write

$$\mu_0(T, p^0, x_0) - \mu_0^0(T, p^0, 1) = RT \ln a_0 = RT \ln x_0 + RT \ln \gamma_0.$$

Finally, a brief summary of the known behaviour of activity coefficients:

Binary non-electrolyte mixtures:

$$\begin{array}{ll} \text{solvent:} & \ln \gamma_0 = kc_1^2 + O(c_1^3) \\ \text{solute:} & \ln \gamma_1 = k'c_1 + O(c_1^2). \end{array}$$

(Theory shows that these equations must be simple power series in the concentration (or an alternative composition variable) and experimental data can always be fitted this way.)

Single electrolyte solution:

$$\begin{array}{ll} \text{solvent:} & \ln \gamma_0 = k''m_1^{3/2} + O(m_1^2) \\ \text{solute:} & \ln \gamma_1 = k'''m_1^{1/2} + O(m_1). \end{array}$$

(The situation for electrolyte solutions is more complex; theory confirms the limiting expressions (originally from Debye–Hückel theory), but, because of the long-range interactions, the resulting equations are non-analytic rather than simple power series.) It is evident that electrolyte solutions are 'ideally dilute' only at extremely low concentrations. Further details about these activity coefficients will be found in other articles.

A2.1.6.5 CHEMICAL EQUILIBRIUM

If a thermodynamic system includes species that may undergo chemical reactions, one must allow for the fact that, even in a closed system, the number of moles of a particular species can change. If a chemical reaction (e.g. $\text{N}_2 + 3\text{H}_2 \rightarrow 2\text{NH}_3$) is represented by the symbolic equation



it is obvious that any changes in the numbers of moles of the species must be proportional to the coefficients ν . Thus if n_A^0, n_B^0 , etc, are the numbers of moles of the species at some initial point, we may write for the number of moles at some subsequent point

$$\begin{array}{ll} n_A = n_A^0 - \nu_A \xi & dn_A = -\nu_A d\xi \\ n_B = n_B^0 - \nu_B \xi & dn_B = -\nu_B d\xi \\ \dots & \dots \\ n_Y = n_Y^0 + \nu_Y \xi & dn_Y = \nu_Y d\xi \\ n_Z = n_Z^0 + \nu_Z \xi & dn_Z = \nu_Z d\xi \\ \dots & \dots \end{array}$$

where the parameter ξ is called the ‘degree of advancement’ of the reaction. (If the variable ξ goes from 0 to 1, one unit of the reaction represented by equation (A2.1.64) takes place, but $\xi = 0$ does not necessarily mean that only reactants are present, nor does $\xi = 1$ mean that only products remain.) More generally one can write

$$n_i = n_i^0 + \nu_i \xi \quad dn_i = \nu_i d\xi \quad (\text{A2.1.65})$$

where positive values of ν_i designate products and negative values of ν_i designate reactants. Equation (A2.1.33), Equation (A2.1.34), Equation (A2.1.35) and Equation (A2.1.36) can be rewritten in new forms appropriate for these closed, but chemically reacting systems. Substitution from equation (A2.1.65) yields

$$\begin{aligned} dU &= T dS - p dV + \sum_i \mu_i dn_i = T dS - p dV + \left(\sum_i \nu_i \mu_i \right) d\xi \\ dH &= T dS + V dp + \sum_i \mu_i dn_i = T dS + V dp + \left(\sum_i \nu_i \mu_i \right) d\xi \\ dA &= -S dT - p dV + \sum_i \mu_i dn_i = -S dT - p dV + \left(\sum_i \nu_i \mu_i \right) d\xi \\ dG &= -S dT + V dp + \sum_i \mu_i dn_i = -S dT + V dp + \left(\sum_i \nu_i \mu_i \right) d\xi. \end{aligned}$$

We have seen that equilibrium in an isolated system ($dU = 0, dV = 0$) requires that the entropy S be a maximum, i.e. that $(\partial S / \partial \xi)_{U,V} = 0$. Examination of the first equation above shows that this can only be true if $\sum_i \nu_i \mu_i$ vanishes. Exactly the same conclusion applies for equilibrium under the other constraints. Thus, for constant temperature and pressure, minimization of the Gibbs free energy requires that $(\partial G / \partial \xi)_{T,p} = \sum_i \nu_i \mu_i = 0$.

THE AFFINITY

This new quantity $\sum_i \nu_i \mu_i$, the negative of which De Donder (1920) has called the ‘affinity’ and given the symbol of a script A , is obviously the important thermodynamic function for chemical equilibrium:

$$-A = \sum_i \nu_i \mu_i = -T(\partial S / \partial \xi)_{U,V} = (\partial U / \partial \xi)_{S,V} = (\partial H / \partial \xi)_{S,p} = (\partial A / \partial \xi)_{T,V} = (\partial G / \partial \xi)_{T,p}.$$

Figure A2.1.9 illustrates how the entropy S and the affinity A vary with ξ in a constant U, V system. It is

apparent that when the slope $(\partial S/\partial \xi)_{U,V}$ is positive (positive affinity), ξ will spontaneously increase; when it is negative, ξ will spontaneously decrease; when it is zero, ξ has no tendency to change and the system is at equilibrium. Moreover, one should note the feature that $\mathcal{A}=0$ is the criterion for equilibrium for all these sets of constraints, whether U and V are fixed, or T and p are fixed.

Instead of using the chemical potential μ_i one can use the *absolute activity* $\lambda_i = \exp(\mu_i/RT)$. Since at equilibrium $\mathcal{A}=0$,

$$-\mathcal{A} = \sum_i v_i \mu_i = RT \sum_i v_i \ln \lambda_i = 0 \quad \text{or} \quad \prod_i \lambda_i^{v_i} = 1.$$

It is convenient to define a *relative activity* a_i in terms of the standard states of the reactants and products at the same temperature and pressure, where $\lambda_i = \lambda_i^0 a_i$, $\mu_i = \mu_i^0$

$$a_i = \lambda_i / \lambda_i^0 = \lambda_i / \exp(\mu_i^0/RT).$$

Thus, at equilibrium

$$-\mathcal{A} = \sum_i v_i \mu_i = \sum_i v_i \ln(\lambda_i^0 a_i) = \sum_i v_i \mu_i^0 + RT \sum_i v_i \ln a_i = 0.$$

If we define an equilibrium constant K as

$$K = \prod_i a_i^{v_i} = \prod_i (\lambda_i^0)^{-v_i} \tag{A2.1.66}$$

it can now be related directly to \mathcal{A}^0 or ΔG^0 :

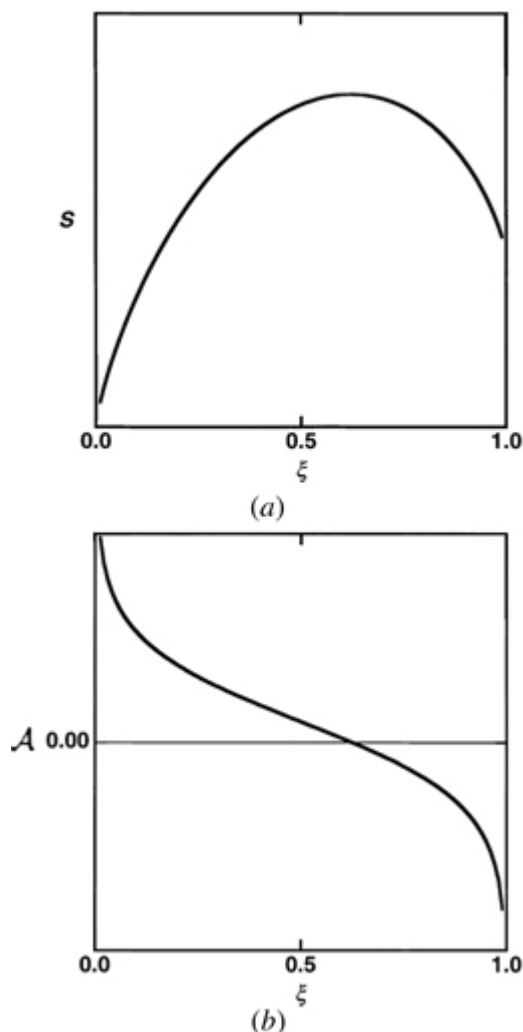


Figure A2.1.9. Chemically reacting systems. (a) The entropy S as a function of the degree of advancement ξ of the reaction at constant U and V . (b) The affinity \mathcal{A} as a function of ξ for the same reacting system. Equilibrium is reached at $\xi=0.623$ where U is a maximum and $\mathcal{A}=0$.

$$\begin{aligned}
 RT \ln K &= RT \sum_i v_i \ln a_i = -RT \sum_i v_i \ln \lambda_i^0 = -\sum_i v_i \mu_i^0 \\
 &= -\Delta G^0 = \mathcal{A}^0.
 \end{aligned}
 \tag{A2.1.67}$$

To proceed further, to evaluate the standard free energy ΔG^0 , we need information (experimental or theoretical) about the particular reaction. One source of information is the equilibrium constant for a chemical reaction involving gases. Previous sections have shown how the chemical potential for a species in a gaseous mixture or in a dilute solution (and the corresponding activities) can be defined and measured. Thus, if one can determine (by some kind of analysis)

the partial pressures of the reacting gases in an equilibrium mixture or the concentrations of reacting species in a dilute solution equilibrium (and, where possible, adjust them to activities so one allows for non-ideality), one can obtain the thermodynamic equilibrium constant K and the standard free energy of reaction ΔG^0 from [equation \(A2.1.66\)](#) and [equation \(A2.1.67\)](#).

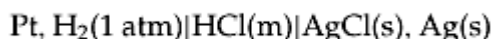
A cautionary word about units: equilibrium constants are usually expressed in units, because pressures and concentrations have units. Yet the argument of a logarithm must be dimensionless, so the activities in [equation \(A2.1.66\)](#), defined in terms of the absolute activities (which are dimensionless) are dimensionless. The value of the standard free energy ΔG^0 depends on the choice of reference state, as does the equilibrium constant. Thus it would be safer to write the equilibrium constant K for a gaseous reaction as

$$K = \prod_i (p_i/p^0)^{v_i} = (p^0)^{\sum v_i} \prod_i p^{v_i} = (p^0)^{\sum v_i} K_p.$$

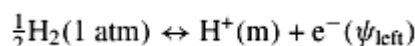
Here K is dimensionless, but K_p is not. Conversely, the factor $(p^0)^{\sum v_i}$ has units (unless $\sum v_i = 0$) but the value unity if $p^0 = 1$ atm. Similar considerations apply to equilibrium constants expressed in concentrations or molalities.

A2.1.6.6 REVERSIBLE GALVANIC CELLS

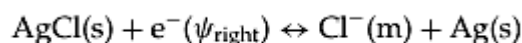
A second source of standard free energies comes from the measurement of the electromotive force of a galvanic cell. Electrochemistry is the subject of other articles ([A2.4](#) and [B1.28](#)), so only the basics of a reversible chemical cell will be presented here. For example, consider the cell conventionally written as



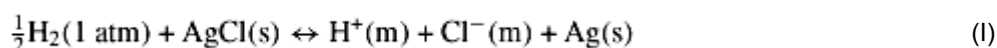
for which the electrode reactions are oxidation at the left electrode, the anode,



and reduction at the right electrode, the cathode,



which can be rewritten as two concurrent reactions:



The chemical reaction (I) cannot come to equilibrium directly; it can come to equilibrium only if the two electrodes are connected so that electrons can flow. One can use this feature to determine the affinity (or the ΔG) of reaction (I) by determining the affinity of reaction (II) which balances it.

In these equations the electrostatic potential ψ might be thought to be the potential at the actual electrodes, the platinum on the left and the silver on the right. However, electrons are not the hypothetical test particles of physics, and the electrostatic potential difference at a junction between two metals is unmeasurable. What *is* measurable is the difference in the *electrochemical* potential μ of the electron, which at equilibrium must be the same in any two wires that are in electrical contact. One assumes that the electrochemical potential can be written as the combination of two terms, a ‘chemical’ potential minus the electrical potential ($-\psi$ because of the negative charge on the electron). When two copper wires are connected to the two electrodes, the

'chemical' part of the electrochemical potential is assumed to be the same in both wires; then the potentiometer measures, under conditions of zero current flow, the electrostatic potential difference $\Delta\psi = \psi_{\text{right}} - \psi_{\text{left}}$ between the two copper wires, which is called the *electromotive force* (emf) \mathcal{E} of the cell.

For reaction (I) the two solids and the hydrogen gas are in their customary standard states ($a = 1$), so

$$\Delta G_{\text{I}} = \Delta G_{\text{I}}^0 + RT \ln \left(\frac{a_{\text{Ag}} a_{\text{H}^+} a_{\text{Cl}^-}}{a_{\text{H}_2}^{1/2} a_{\text{AgCl}}} \right) = \Delta G_{\text{I}}^0 + RT \ln(a_{\text{H}^+} a_{\text{Cl}^-})$$

while for the electrons

$$\Delta G_{\text{II}} = \mathcal{F}(\psi_{\text{right}} - \psi_{\text{left}}) = \mathcal{F}\mathcal{E}$$

where \mathcal{F} is the Faraday constant (the amount of charge in one mole of electrons).

When no current flows, there is a constrained equilibrium in which the chemical reaction cannot proceed in either direction, and \mathcal{E} can be measured. With this constraint, for the *overall reaction* $\Delta G = \Delta G_{\text{I}} + \Delta G_{\text{II}} = 0$, so

$$\Delta G_{\text{I}}^0 + RT \ln(a_{\text{H}^+} a_{\text{Cl}^-}) = -\mathcal{F}\mathcal{E}.$$

Were the HCl in its standard state, ΔG_{I}^0 would equal $-\mathcal{F}\mathcal{E}^0$, where \mathcal{E}^0 is the standard emf for the reaction. In general, for any reversible chemical cell without transference, i.e. one with a single electrolyte solution, not one with any kind of junction between two solutions,

$$\Delta G_{\text{I}}^0 + RT \ln \left(\sum_i a_i^{v_i} \right) = -n\mathcal{F}\mathcal{E} \quad (\text{A2.1.68})$$

$$\Delta G_{\text{I}}^0 = -n\mathcal{F}\mathcal{E}^0 \quad (\text{A2.1.69})$$

-50-

where n is the number of electrons associated with the cell reaction as written. By combining [equation \(A2.1.68\)](#) and [equation \(A2.1.69\)](#) one obtains the *Nernst equation*

$$\mathcal{E} = \mathcal{E}^0 - (RT/n\mathcal{F}) \ln \left(\sum_i a_i^{v_i} \right).$$

Thus, if the activities of the various species can be determined or if one can extrapolate to infinite dilution, the measurement of the emf yields the standard free energy of the reaction.

A2.1.6.7 STANDARD STATES AND STANDARD FREE ENERGIES OF FORMATION

With several experimental methods for determining the ΔG^0 s of chemical reactions, one can start to organize the information in a systematic way. (To complete this satisfactorily, or at least efficiently and precisely, one needs the third law to add third-law entropies to calorimetrically determined ΔH^0 s. Discussion of this is deferred to the next section, but it will be assumed for the purpose of this section that all necessary information is available.)

STANDARD STATES

Conventions about standard states (the reference states introduced earlier) are necessary because otherwise the meaning of the standard free energy of a reaction would be ambiguous. We summarize the principal ones:

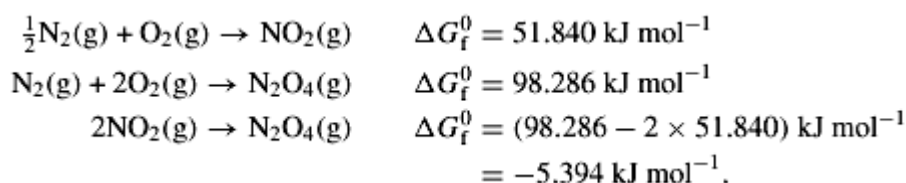
- (1) All standard states, both for pure substances and for components in mixtures and solutions, are defined for a pressure of exactly 1 atmosphere. However the temperature must be specified. (There is some movement towards metricating this to a pressure of 1 bar = 100 kPa = 0.986 924 atm. This would make a significant difference only for gases; at $T = 298$ K, this would decrease a μ^0 by 32.6 J mol⁻¹.)
- (2) As noted earlier, the standard state of a gas is the hypothetical ideal gas at 1 atmosphere and the specified temperature T .
- (3) The standard state of a substance in a condensed phase is the real liquid or solid at 1 atm and T .
- (4) The standard state of an electrolyte is the hypothetical ideally dilute solution (Henry's law) at a molarity of 1 mol kg⁻¹. (Actually, as will be seen, electrolyte data are conventionally reported as for the formation of individual ions.) Standard states for non-electrolytes in dilute solution are rarely invoked.
- (5) For a free energy of formation, the preferred standard state of the element should be the thermodynamically stable (lowest chemical potential) form of it; e.g. at room temperature, graphite for carbon, the orthorhombic crystal for sulfur.

Compounds that are products in reactions are sometimes reported in standard states for phases that are not the most stable at the temperature in question. The stable standard state of H₂O at 298 K (and 1 atm) is, of course, the liquid, but ΔG^0 s are sometimes reported for reactions leading to gaseous H₂O at 298 K. Moreover the standard functions for the formation of some metastable states, e.g. C(diamond) or S(monoclinic) at 298 K, are sometimes reported in tables.

-51-

The useful thermodynamic functions (e.g. G , H , S , C_p , etc) are all state functions, so their values in any particular state are independent of the path by which the state is reached. Consequently, one can combine (by addition or subtraction) the ΔG^0 s for several chemical reactions at the same temperature, to obtain the ΔG^0 of another reaction that is not directly measurable. (Indeed one experimentalist has commented that the principal usefulness of thermodynamics arises 'because some quantities are easier to measure than others'.)

In particular, one can combine reactions to yield the ΔG_f^0 , ΔH_f^0 and ΔS_f^0 for formation of compounds from their elements, quantities rarely measurable directly. (Many ΔH_f^0 s for formation of substances are easily calculated from the calorimetric measurement of the enthalpies of combustion of the compound and of its constituent elements.) For example, consider the dimerization of NO₂ at 298 K. In appropriate tables one finds



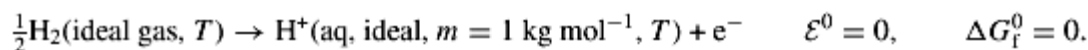
With this information one can now use [equation \(A2.1.67\)](#) to calculate the equilibrium constant at 298 K. One

finds $K = 75.7$ or, using the dimensional constant, $K_p = 75.7 \text{ atm}^{-1}$. (In fact, the free energies of formation were surely calculated using the experimental data on the partial pressures of the gases in the equilibrium. One might also note that this is one of the very few equilibria involving only gaseous species at room temperature that have constants K anywhere near unity.)

Thermodynamic tables usually report at least three quantities: almost invariably the standard enthalpy of formation at 298 K, $\Delta H_f^0(298 \text{ K})$; usually the standard entropy at 298 K, $S^0(298 \text{ K})$ (not $\Delta S_f^0(298 \text{ K})$, but the entropy based on the third-law convention (see subsequent section) that $S^0(0 \text{ K}) = 0$); and some form of the standard free energy of formation, usually either $\Delta G_f^0(298 \text{ K})$ or $\log_{10} K_f$. Many tables will include these quantities at a series of temperatures, as well as the standard heat capacity $C_{p,m}^0$, and enthalpies and entropies of various transitions (phase changes).

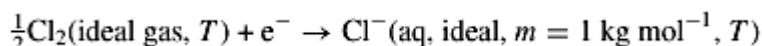
THE STANDARD FREE ENERGY OF FORMATION OF IONS

A special convention exists concerning the free energies of ions in aqueous solution. Most thermodynamic information about strong (fully dissociated) electrolytes in aqueous solutions comes, as has been seen, from measurements of the emf of reversible cells. Since the ions in very dilute solution (or in the hypothetical ideally dilute solution at $m = 1 \text{ mol kg}^{-1}$) are essentially independent, one would like to assign free energy values to individual ions and add together the values for the anion and the cation to get that for the electrolyte. Unfortunately the emf of a half cell is unmeasurable, although there have been some attempts to estimate it theoretically. Consequently, the convention that the standard half-cell emf of the hydrogen electrode is exactly zero has been adopted.

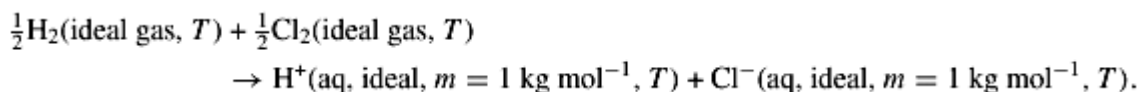


-52-

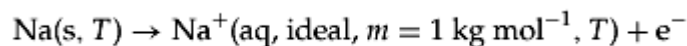
Thus, when tables report the standard emf or standard free energy of the chloride ion,



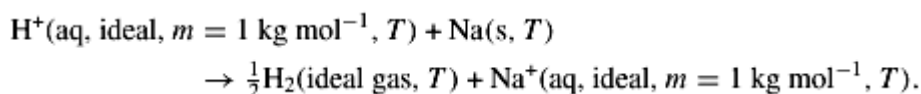
it is really that of the reaction



Similarly, the standard free energy or standard emf of the sodium ion, reported for



is really that for



TEMPERATURE DEPENDENCE OF THE EQUILIBRIUM CONSTANT

Since equation (A2.1.67) relates the equilibrium constant K to the standard free energy ΔG^0 of the reaction,

one can rewrite the equation as

$$\ln K = -\Delta G^0/RT$$

and differentiate with respect to temperature

$$d \ln K/dT = -(d\Delta G^0/dT)/RT + \Delta G^0/RT^2 = (T\Delta S^0 + \Delta G^0)/RT^2 = \Delta H^0/RT^2. \quad (\text{A2.1.70})$$

This important relation between the temperature derivative of the equilibrium constant K and the standard enthalpy of the reaction ΔH^0 is sometimes known as the *van't Hoff equation*. (Note that the derivatives are not expressed as partial derivatives at constant pressure, because the quantities involved are all defined for the standard pressure p^0 . Note also that in this derivation one has not assumed—as is sometimes alleged—that ΔH^0 and ΔS^0 are independent of temperature.)

The validity of equation (A2.1.70) has sometimes been questioned when enthalpies of reaction determined from calorimetric experiments fail to agree with those determined from the temperature dependence of the equilibrium constant. The thermodynamic equation is rigorously correct, so doubters should instead examine the experimental uncertainties and whether the two methods actually relate to exactly the same reaction.

-53-

A2.1.7 THE THIRD LAW

A2.1.7.1 HISTORY; THE NERNST HEAT THEOREM

The enthalpy, entropy and free energy changes for an isothermal reaction near 0 K cannot be measured directly because of the impossibility of carrying out the reaction reversibly in a reasonable time. One can, however, by a suitable combination of measured values, calculate them indirectly. In particular, if the value of ΔH^0 , ΔS^0 or ΔG^0 is known at a specified temperature T , say 298 K, its value at another temperature T' can be computed using this value and the changes involved in bringing the products and the reactants separately from T' to T . If these measurements can be extrapolated to 0 K, the isothermal changes for the reaction at 0 K can be calculated.

If, in going from 0 K to T , a substance undergoes phase changes (fusion, vaporization, etc) at T_A and T_B with molar enthalpies of transition $\Delta \bar{H}_A$ and $\Delta \bar{H}_B$, one can write

$$\begin{aligned} \bar{H}^0(T) - \bar{H}^0(0) &= \int_0^{T_A} \bar{C}_p^0 dT + \Delta \bar{H}_A + \int_{T_A}^{T_B} \bar{C}_p^0 dT + \Delta \bar{H}_B + \int_{T_B}^T \bar{C}_p^0 dT \\ \bar{S}^0(T) - \bar{S}^0(0) &= \int_0^{T_A} (\bar{C}_p^0/T) dT + \Delta \bar{H}_A/T_A + \int_{T_A}^{T_B} (\bar{C}_p^0/T) dT + \Delta \bar{H}_B/T_B + \int_{T_B}^T (\bar{C}_p^0/T) dT. \end{aligned} \quad (\text{A2.1.71})$$

It is manifestly impossible to measure heat capacities down to exactly 0 K, so some kind of extrapolation is necessary. Unless C_p were to approach zero as T approaches zero, the limiting value of C_p/T would not be finite and the first integral in equation (A2.1.71) would be infinite. Experiments suggested that C_p might approach zero and Nernst (1906) noted that computed values of the entropy change ΔS^0 for various reactions appeared to approach zero as the temperature approached 0 K. This empirical discovery, known as the *Nernst heat theorem*, can be expressed mathematically in various forms as

$$\lim_{T \rightarrow 0} (d\Delta G^0/dT) = \lim_{T \rightarrow 0} \Delta S^0 = \lim_{T \rightarrow 0} (d\Delta H^0/dT) = 0. \quad (\text{A2.1.72})$$

However, the possibility that C_p might not go to zero could not be excluded before the development of the quantum theory of the heat capacity of solids. When Debye (1912) showed that, at sufficiently low temperatures, C_p is proportional to T^3 , this uncertainty was removed, and a reliable method of extrapolation for most crystalline substances could be developed. (For metals there is an additional term, proportional to T , a contribution from the electrons.) If the temperature T' is low enough that $C_p = \alpha T^3$, one may write

$$\bar{S}^0(T') - \bar{S}^0(0) = \int_0^{T'} (\bar{C}_p^0/T) dT = \int_0^{T'} \alpha T^2 dT = \alpha T'^3/3 = \bar{C}_p^0(T')/3. \quad (\text{A2.1.73})$$

With this addition, better entropy determinations, e.g. measurements plus extrapolations to 0 K, became available.

-54-

The evidence in support of [equation \(A2.1.72\)](#) is of several kinds:

- (1) Many substances exist in two or more solid allotropic forms. At 0 K, the thermodynamically stable form is of course the one of lowest energy, but in many cases it is possible to make thermodynamic measurements on another (metastable) form down to very low temperatures. Using the measured entropy of transition at equilibrium, the measured heat capacities of both forms and [equation \(A2.1.73\)](#) to extrapolate to 0 K, one can obtain the entropy of transition at 0 K. Within experimental error ΔS^0 is zero for the transitions between β - and γ -phosphine, between orthorhombic and monoclinic sulfur and between different forms of cyclohexanol.
- (2) As seen in previous sections, the standard entropy ΔS^0 of a chemical reaction can be determined from the equilibrium constant K and its temperature derivative, or equivalently from the temperature derivative of the standard emf of a reversible electrochemical cell. As in the previous case, calorimetric measurements on the separate reactants and products, plus the usual extrapolation, will yield $\Delta S^0(0)$.

The limiting ΔS^0 so calculated is usually zero within experimental error, but there are some disturbing exceptions. Not only must solutions and some inorganic and organic glasses be excluded, but also crystalline CO, NO, N₂O and H₂O. It may be easy to see, given the most rudimentary statistical ideas of entropy, that solutions and glasses have some positional disorder frozen in, and one is driven to conclude that the same situation must occur with these few simple crystals as well. For these substances in the gaseous state at temperature T there is a disagreement between the ‘calorimetric’ entropy calculated using [equation \(A2.1.71\)](#) and the ‘spectroscopic’ entropy calculated by statistical mechanics using the rotational and vibrational constants of the gas molecule; this difference is sometimes called ‘residual entropy’. However, it can be argued that, because such a substance or mixture is frozen into a particular disordered state, its entropy is in fact zero. In any case, it is not in internal equilibrium (unless some special hypothetical constraints are applied), and it cannot be reached along a reversible path.

It is beyond the scope of this article to discuss the detailed explanation of these exceptions; suffice it to say that there are reasonable explanations in terms of the structure of each crystal.

A2.1.7.2 FIRST STATEMENT ($\Delta S^0 \rightarrow 0$)

Because it is necessary to exclude some substances, including some crystals, from the Nernst heat theorem, Lewis and Gibson (1920) introduced the concept of a ‘perfect crystal’ and proposed the following modification as a definitive statement of the ‘third law of thermodynamics’ (exact wording due to Lewis and Randall (1923)):

If the entropy of each element in some crystalline state be taken as zero at the absolute zero of temperature, every substance has a finite positive entropy, but at the absolute zero of temperature the entropy may become zero, and does so become in the case of perfect crystalline substances.

Because of the Nernst heat theorem and the third law, standard thermodynamic tables usually do not report entropies of formation of compounds; instead they report the molar entropy $\hat{S}^0(T)$ for each element and compound. The entropies reported for those substances that show ‘residual entropy’ (the ‘imperfect’ crystalline substances) are ‘spectroscopic’ entropies, not ‘calorimetric’ entropies.

-55-

For those who are familiar with the statistical mechanical interpretation of entropy, which asserts that at 0 K substances are normally restricted to a single quantum state, and hence have zero entropy, it should be pointed out that the conventional thermodynamic zero of entropy is not quite that, since most elements and compounds are mixtures of isotopic species that in principle should separate at 0 K, but of course do not. The thermodynamic entropies reported in tables ignore the entropy of isotopic mixing, and in some cases ignore other complications as well, e.g. ortho- and para-hydrogen.

A2.1.7.3 SECOND STATEMENT (UNATTAINABILITY OF 0 K)

In the Lewis and Gibson statement of the third law, the notion of ‘a perfect crystalline substance’, while understandable, strays far from the macroscopic logic of classical thermodynamics and some scientists have been reluctant to place this statement in the same category as the first and second laws of thermodynamics. Fowler and Guggenheim (1939), noting that the first and second laws both state universal limitations on processes that are experimentally possible, have pointed out that the principle of the unattainability of absolute zero, first enunciated by Nernst (1912) expresses a similar universal limitation:

It is impossible by any procedure, no matter how idealized, to reduce the temperature of any system to the absolute zero of temperature in a finite number of operations.

No one doubts the correctness of either of these statements of the third law and they are universally accepted as equivalent. However, there seems to have been no completely satisfactory proof of their equivalence; some additional, but very plausible, assumption appears necessary in making the connection.

Consider how the change of a system from a thermodynamic state α to a thermodynamic state β could decrease the temperature. (The change in state $\alpha \rightarrow \beta$ could be a chemical reaction, a phase transition, or just a change of volume, pressure, magnetic field, etc). Initially assume that α and β are always in complete internal equilibrium, i.e. neither has been cooled so rapidly that any disorder is frozen in. Then the Nernst heat theorem requires that $S^\alpha(0) = S^\beta(0)$ and the plot of entropy versus temperature must look something like the sketch in [figure A2.1.10a](#).

-56-

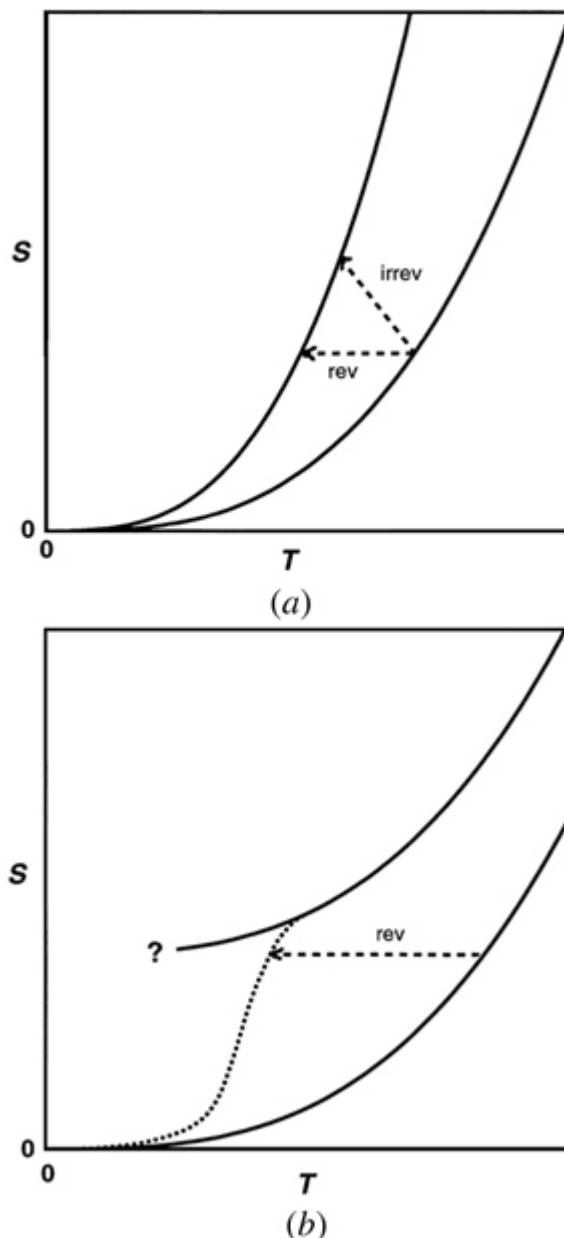


Figure A2.1.10. The impossibility of reaching absolute zero. (a) Both states α and β in complete internal equilibrium. Reversible and irreversible paths (dashed) are shown. (b) State β not in internal equilibrium and with ‘residual entropy’. The true equilibrium situation for β is shown dotted.

The most effective cooling process will be an adiabatic reversible one ($\Delta S = 0$). In any non-adiabatic process, heat will be absorbed from the surroundings (which are at a higher temperature), thus defeating the cooling process. Moreover, according to the second law, for an irreversible adiabatic process $\Delta S > 0$; it is obvious from [figure A2.1.10a](#) that the reversible process gets to the lower temperature. It is equally obvious that the process must end with β at a non-zero temperature.

But what if the thermodynamic state β is not in ‘complete internal equilibrium’, but has some ‘residual entropy’ frozen in? One might then imagine a diagram like [figure A2.1.10b](#) with a path for β leading to a positive entropy at 0 K. But β is not the true internal equilibrium situation at low temperature; it was obtained by freezing in what was equilibrium at a much higher temperature. In a process that generates β at a much lower temperature, one will not get this same frozen disorder; one will end on something more nearly like the

true internal equilibrium curve (shown dotted). This inconceivability of the low-temperature process yielding the higher temperature's frozen disorder is the added assumption needed to prove the equivalence of the two statements. (Most ordinary processes become increasingly unlikely at low temperatures; only processes with essentially zero activation energy can occur and these are hardly the kinds of processes that could generate 'frozen' situations.)

The principle of the unattainability of absolute zero in no way limits one's ingenuity in trying to obtain lower and lower thermodynamic temperatures. The third law, in its statistical interpretation, essentially asserts that the ground quantum level of a system is ultimately non-degenerate, that some energy difference $\Delta\varepsilon$ must exist between states, so that at equilibrium at 0 K the system is certainly in that non-degenerate ground state with zero entropy. However, the $\Delta\varepsilon$ may be very small and temperatures of the order of $\Delta\varepsilon/k$ (where k is the Boltzmann constant, the gas constant per molecule) may be obtainable.

MAGNETIC COOLING

A standard method of attaining very low temperatures is *adiabatic demagnetization*, a procedure suggested independently by Debye and by Giauque in 1926. A paramagnetic solid is cooled to a low temperature (one can reach about 1 K by the vaporization of liquid helium) and the solid is then magnetized isothermally in a high magnetic field B_0 . (Any heat developed is carried away by contact with dilute helium gas.) As shown in [figure A2.1.11](#), the entropy obviously decreases (compare [equation \(A2.1.43\)](#)).

-58-

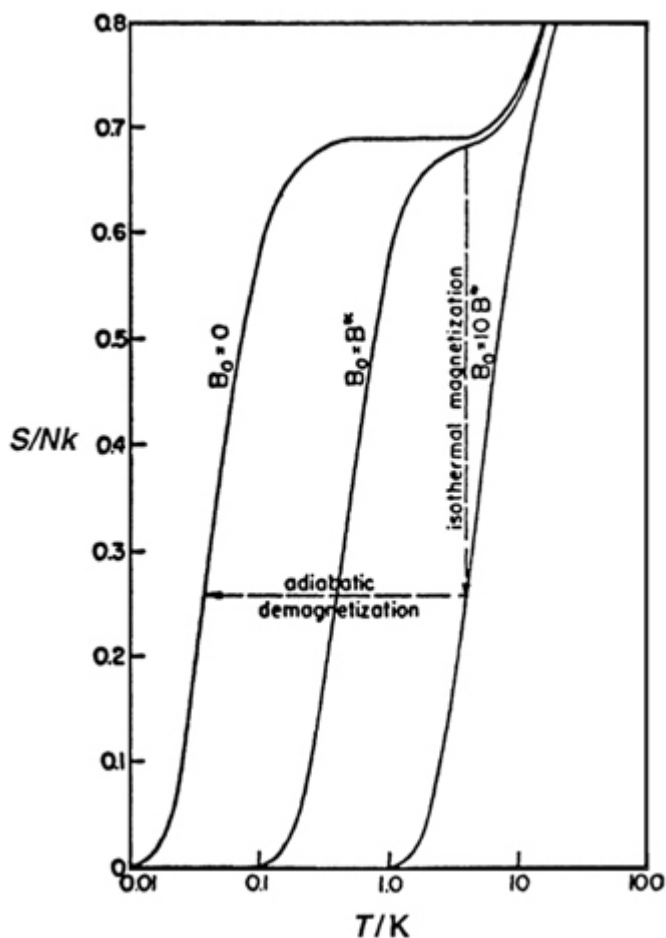


Figure A2.1.11. Magnetic cooling: isothermal magnetization at 4 K followed by adiabatic demagnetization to 0.04 K. (Constructed for a hypothetical magnetic substance with two magnetic states with an energy

separation $\Delta\varepsilon = k(0.1 \text{ K})$ at $\mathbf{B}_0 = 0$ and \mathbf{B}^* (the field at which the separation is $10\Delta\varepsilon$ or 7400 gauss); the crystalline Stark effect has been ignored. The entropy above $S/Nk = 0.69 = \ln 2$ is due to the vibration of a Debye crystal.)

Now the system is thermally insulated and the magnetic field is decreased to zero; in this adiabatic, essentially reversible (isentropic) process, the temperature necessarily decreases since

$$(\partial T/\partial B_0)_S = -(\partial S/\partial B_0)_T/(\partial S/\partial T)_{B_0}.$$

(($\partial S/\partial B_0)_T$ is negative and $(\partial S/\partial T)_{B_0}$, a heat capacity divided by temperature, is surely positive, so $(\partial T/\partial B_0)_S$ is positive.)

-59-

A2.1.7.4 THIRD STATEMENT (SIMPLE LIMITS AND STATISTICAL THERMODYNAMICS)

As we have seen, the third law of thermodynamics is closely tied to a statistical view of entropy. It is hard to discuss its implications from the exclusively macroscopic view of classical thermodynamics, but the problems become almost trivial when the molecular view of statistical thermodynamics is introduced. Guggenheim (1949) has noted that the usefulness of a molecular view is not unique to the situation of substances at low temperatures, that there are other limiting situations where molecular ideas are helpful in interpreting general experimental results:

- (1) Substances at high dilution, e.g. a gas at low pressure or a solute in dilute solution, show simple behaviour. The ideal-gas law and Henry's law for dilute solutions antedate the development of the formalism of classical thermodynamics. Earlier sections in this article have shown how these experimental laws lead to simple thermodynamic equations, but these results are added to thermodynamics; they are not part of the formalism. Simple molecular theories, even if they are not always recognized as statistical mechanics, e.g. the 'kinetic theory of gases', make the experimental results seem trivially obvious.
- (2) The entropy of mixing of very similar substances, i.e. the ideal solution law, can be derived from the simplest of statistical considerations. It too is a limiting law, of which the most nearly perfect example is the entropy of mixing of two isotopic species.

With this in mind Guggenheim suggested still another statement of the 'third law of thermodynamics':

By the standard methods of statistical thermodynamics it is possible to derive for certain entropy changes general formulas that cannot be derived from the zeroth, first, and second laws of classical thermodynamics. In particular one can obtain formulae for entropy changes in highly disperse systems, for those in very cold systems, and for those associated with the mixing of very similar substances.

A2.1.8 THERMODYNAMICS AND STATISTICAL MECHANICS

Any detailed discussion of statistical mechanics would be inappropriate for this section, especially since other sections (A2.2 and A2.3) treat this in detail. However, a few aspects that relate to classical thermodynamics deserve brief mention.

A2.1.8.1 ENSEMBLES AND THE CONSTRAINTS OF CLASSICAL THERMODYNAMICS

It is customary in statistical mechanics to obtain the average properties of members of an *ensemble*, an essentially infinite set of systems subject to the same constraints. Of course each of the systems contains the

same substance or group of substances, but in addition the constraints placed on a particular ensemble are parallel to those encountered in classical thermodynamics.

The *microcanonical ensemble* is a set of systems each having the same number of molecules N , the same volume V and the same energy U . In such an ensemble of isolated systems, any allowed quantum state is equally probable. In classical thermodynamics at equilibrium at constant n (or equivalently, N), V , and U , it is the entropy S that is a maximum. For the microcanonical ensemble, the entropy is directly related to the number of allowed quantum states $\Omega(N, V, U)$:

-60-

$$S(N, V, U) = k \ln \Omega(N, V, U).$$

The *canonical ensemble* is a set of systems each having the same number of molecules N , the same volume V and the same temperature T . This corresponds to putting the systems in a thermostatic bath or, since the number of systems is essentially infinite, simply separating them by diathermic walls and letting them equilibrate. In such an ensemble, the probability of finding the system in a particular quantum state l is proportional to $e^{-U_l/kT}$ where $U_l(N, V)$ is the energy of the l th quantum state and k , as before, is the Boltzmann constant. In classical thermodynamics, the appropriate function for fixed N , V and T is the Helmholtz free energy A , which is at a minimum at equilibrium and in statistical mechanics it is A that is directly related to the *canonical partition function* Q for the canonical ensemble.

$$Q(N, V, T) = \sum_l e^{-U_l(N, V)/kT}$$

$$A(N, V, T) = -kT \ln Q(N, V, T).$$

The *grand canonical ensemble* is a set of systems each with the same volume V , the same temperature T and the same chemical potential μ (or if there is more than one substance present, the same set of μ_i s). This corresponds to a set of systems separated by diathermic and permeable walls and allowed to equilibrate. In classical thermodynamics, the appropriate function for fixed μ , V , and T is the product pV (see [equation \(A2.1.37\)](#)) and statistical mechanics relates pV directly to the *grand canonical partition function* Ξ .

$$\Xi(\mu, V, T) = \sum_N Q_N(N, V, T) e^{-N\mu/kT} = \sum_N \lambda^N Q_N(N, V, T)$$

$$pV = kT \ln \Xi(\mu, V, T).$$

where λ is the absolute activity of [section 2.1.6.5](#).

Since other sets of constraints can be used, there are other ensembles and other partition functions, but these three are the most important.

A2.1.8.2 FLUCTUATIONS; THE 'EXACTNESS' OF THERMODYNAMICS

In defining the thermodynamic state of a system in terms of fixed macroscopic constraints, classical thermodynamics appears to assume the identical character of two states subject to the identical set of constraints. However, any consideration of the fact that such systems are composed of many molecules in constant motion suggests that this must be untrue. Surely, fixing the number of gas molecules N in volume V at temperature T does not guarantee that the molecules striking the wall create exactly the same pressure at all times and in all such systems. If the pressure p is just an average, what can one say about the magnitude of fluctuations about this average?

According to statistical mechanics, for the canonical ensemble one may calculate $\langle U \rangle$, the average energy of all the members of the ensemble, while for the grand canonical ensemble one can calculate two averages, $\langle N \rangle$ and $\langle U \rangle$. Of crucial importance, however, is the probability of observing significant variations (fluctuations) from these averages in any particular member of the ensemble. Fortunately, statistical mechanics yields an answer to these questions.

-61-

Probability theory shows that the standard deviation σ_x of a quantity x can be written as

$$\sigma_x^2 = \langle x^2 \rangle - \langle x \rangle^2$$

and statistical mechanics can relate these averages to thermodynamic quantities. In particular, for the canonical ensemble

$$\sigma_U^2 = kT^2 C_V$$

while for the grand canonical ensemble

$$\sigma_N^2 = kT\kappa_T(N^2/V).$$

All the quantities in these equations are intensive (independent of the size of the system) except C_V , N , and V , which are extensive (proportional to the size of the system, i.e. to N). It follows that σ_U^2 is of the order of N , so σ_U/N is of the order of $N^{-1/2}$, as is σ_N/U . Since a macroscopic system described by thermodynamics probably has at least about 10^{20} molecules, the uncertainty, i.e. the typical fluctuation, of a measured thermodynamic quantity must be of the order of 10^{-10} times that quantity, orders of magnitude below the precision of any current experimental measurement. Consequently we may describe thermodynamic laws and equations as 'exact'.

(An exception to this conclusion is found in the immediate vicinity of critical points, where fluctuations become much more significant, although—with present experimental precision—still not of the order of N .)

REFERENCES

- [1] Carathéodory C 1909 Untersuchungen über die Grundlagen der Thermodynamik *Math. Ann.* **67** 355–86
- [2] Born M 1921 Kritische Betrachtungen zur traditionellen Darstellung der Thermodynamik *Physik. Z.* **22** 218–24, 249–54, 282–6
- [3] Redlich O 1970 The so-called zeroth law of thermodynamics *J. Chem. Educ.* **47** 740
- [4] Guggenheim E A 1957 *Thermodynamics* 3rd edn (Amsterdam: North-Holland) p 441
- [5] Fokker A D 1939 Remark on the fundamental relations of thermomagnetism *Physica* **6** 791–6
- [6] Broer L J F 1946 On the statistical mechanics in a magnetic field *Physica* **12** 49–60
- [7] Kivelson D and Oppenheim I 1966 Work in irreversible expansions *J. Chem. Educ.* **43** 233–5

FURTHER READING

Callen H B 1960 *Thermodynamics, an Introduction to the Physical Theories of Equilibrium Thermostatistics and Irreversible Thermodynamics* (New York: Wiley)

General principles, representative applications, fluctuations and irreversible thermodynamics. Chapter 4 discusses quasi-static processes, reversible work and heat sources, and thermodynamic engines.

Guggenheim E A 1967 *Thermodynamics, An Advanced Treatment for Chemists and Physicists* 5th edn (Amsterdam: North-Holland, New York: Interscience)

Basic thermodynamics, statistical thermodynamics, third-law entropies, phase transitions, mixtures and solutions, electrochemical systems, surfaces, gravitation, electrostatic and magnetic fields. (In some ways the 3rd and 4th editions (1957 and 1960) are preferable, being less idiosyncratic.)

Lewis G N and Randall M 1961 *Thermodynamics* 2nd edn, ed K S Pitzer and L Brewer (New York: McGraw-Hill)

Classical thermodynamics with many applications and much experimental information. Tables and graphs. Results of statistical mechanical calculations. The first edition (1923) was the first real presentation of thermodynamics for chemists.

McGlashan M L 1979 *Chemical Thermodynamics* (London: Academic)

An idiosyncratic book with comparatively little emphasis on irreversible processes, but with much emphasis upon experimental procedures and experimental results.

Pippard A E 1957 (reprinted with corrections 1964) *The Elements of Classical Thermodynamics* (Cambridge: Cambridge University Press)

Fundamentals of thermodynamics. Applications to phase transitions. Primarily directed at physicists rather than chemists.

Reid C E 1990 *Chemical Thermodynamics* (New York: McGraw-Hill)

Basic laws, using the Carathéodory approach. Applications to gases, mixtures and solutions, chemical and phase equilibria, electrochemical systems, surfaces.

Reiss H 1965 *Methods of Thermodynamics* (New York: Blaisdell) (reissued unabridged with a few corrections by Dover) (Mineola, NY: 1996)

A careful analysis of the fundamentals of classical thermodynamics, using the Born–Carathéodory approach. Emphasis on constraints, chemical potentials. Discussion of difficulties with the third law. Few applications.

A2.2.1 INTRODUCTION

Thermodynamics is a phenomenological theory based upon a small number of fundamental laws, which are deduced from the generalization and idealization of experimental observations on macroscopic systems. The goal of statistical mechanics is to deduce the macroscopic laws of thermodynamics and other macroscopic theories (e.g. hydrodynamics and electromagnetism) starting from mechanics at a microscopic level and combining it with the rules of probability and statistics. As a branch of theoretical physics, statistical mechanics has extensive applications in physics, chemistry, biology, astronomy, materials science and engineering. Applications have been made to systems which are in thermodynamic equilibrium, to systems in steady state and also to non-equilibrium systems. Even though the scope of statistical mechanics is quite broad, this section is mostly limited to basics relevant to equilibrium systems.

At its foundation level, statistical mechanics involves some profound and difficult questions which are not fully understood, even for systems in equilibrium. At the level of its applications, however, the rules of calculation that have been developed over more than a century have been very successful.

The approach outlined here will describe a viewpoint which leads to the standard calculational rules used in various applications to systems in thermodynamic (thermal, mechanical and chemical) equilibrium. Some applications to ideal and weakly interacting systems will be made, to illustrate how one needs to think in applying statistical considerations to physical problems.

Equilibrium is a macroscopic phenomenon which implies a description on a length and time scale much larger than those appropriate to the molecular motion. The concept of 'absolute equilibrium' is an idealization and refers to a state of an absolutely isolated system and an infinitely long observation time. In non-equilibrium systems with slowly varying macroscopic properties, it is often useful to consider 'local equilibrium' where the 'macroscopic' time and length scales are determined in the context of an observation, or an experiment, and the system. A typical value of an experimentally measured property corresponds to the average value over the observation time of the corresponding physical observable; the physical properties of a system in equilibrium are time invariant and should be independent of the observation time. The observation time of an equilibrium state is typically quite long compared to the time characteristic of molecular motions.

Conservation laws at a microscopic level of molecular interactions play an important role. In particular, energy as a conserved variable plays a central role in statistical mechanics. Another important concept for equilibrium systems is the law of detailed balance. Molecular motion can be viewed as a sequence of collisions, each of which is akin to a reaction. Most often it is the momentum, energy and angular momentum of each of the constituents that is changed during a collision; if the molecular structure is altered, one has a chemical reaction. The law of detailed balance implies that, in equilibrium, the number of each reaction in the forward direction is the same as that in the reverse direction; i.e. each microscopic reaction is in equilibrium. This is a consequence of the time reversal symmetry of mechanics.

A2.2.2 MECHANICS, MICROSTATES AND THE DEGENERACY FUNCTION

Macroscopic systems contain a large number, N , of microscopic constituents. Typically N is of the order of

10^{20} – 10^{25} . Thus many aspects of statistical mechanics involve techniques appropriate to systems with large N . In this respect, even the non-interacting systems are instructive and lead to non-trivial calculations. The degeneracy function that is considered in this subsection is an essential ingredient of the formal and general methods of statistical mechanics. The degeneracy function is often referred to as the density of states.

We first consider three examples as a prelude to the general discussion of basic statistical mechanics. These are: (i) N non-interacting spin- $\frac{1}{2}$ particles in a magnetic field, (ii) N non-interacting point particles in a box, and (iii) N non-interacting harmonic oscillators. For each example the results of quantum mechanics are used to enumerate the microstates of the N -particle system and then obtain the degeneracy function (density of states) of the system's energy levels. Even though these three examples are for ideal non-interacting systems, there are many realistic systems which turn out to be well approximated by them.

A microstate (or a microscopic state) is one of the quantum states determined from $\mathcal{H}\Phi_l = E_l\Phi_l$, ($l = 1, 2, \dots$), where \mathcal{H} is the Hamiltonian of the system, E_l is the energy of the quantum state l and Φ_l is the wavefunction representing the quantum state l . The large- N behaviour of the degeneracy function is of great relevance. The calculation of the degeneracy function in these three examples is a useful precursor to the conceptual use of the density of states of an arbitrary interacting system in the general framework of statistical mechanics.

(i) N non-interacting spin- $\frac{1}{2}$ particles in a magnetic field. Each particle can be considered as an elementary magnet (its magnetic moment has a magnitude equal to μ) which can point along two possible directions in space ($+z$ or 'up' and $-z$ or 'down'). A microstate of such a model system is given by giving the orientation ($+$ or $-$) of each magnet. It is obvious that for N such independent magnets, there are 2^N different microstates for this system. Note that this number grows exponentially with N . The total magnetic moment \mathcal{M} of the model system is the vector sum of the magnetic moments of its N constituents. The component of \mathcal{M} along the z direction varies between $-N\mu$ and $+N\mu$, and can take any of the $(N+1)$ possible values $N\mu, (N-2)\mu, (N-4)\mu, \dots, -N\mu$. This number of possible values of \mathcal{M} is much less than the total number of microstates 2^N , for large N . The number of microstates for a given \mathcal{M} is the degeneracy function. In each of these microstates, there will be $\frac{1}{2}N + m$ spins up and $\frac{1}{2}N - m$ spins down, such that the difference between the two is $2m$, which is called the spin excess and equals \mathcal{M}/μ . If x is the probability for a particle to have its spin up and $y = (1-x)$ is the probability for its spin down, the degeneracy function $g(N, m)$ can be obtained by inspection from the binomial expansion

$$(x + y)^N = \sum_{m=-\frac{1}{2}N}^{\frac{1}{2}N} \frac{N!}{(\frac{1}{2}N + m)!(\frac{1}{2}N - m)!} x^{\frac{1}{2}N+m} y^{\frac{1}{2}N-m}.$$

-3-

That is

$$g(N, m) = \frac{N!}{(\frac{1}{2}N + m)!(\frac{1}{2}N - m)!}. \quad (\text{A2.2.1})$$

By setting $x = y = \frac{1}{2}$, one can see that $\sum_m g(N, m) = 2^N$. For typical macroscopic systems, the number N of constituent molecules is very large: $N \sim 10^{23}$. For large N , $g(N, m)$ is a very sharply peaked function of m . In order to see this one needs to use the Stirling approximation for $N!$ which is valid when $N \gg 1$:

$$N! \approx (2\pi N)^{\frac{1}{2}} N^N \exp[-N + 1/(12N) + \dots]. \quad (\text{A2.2.2})$$

For sufficiently large N , the terms $1/(12N) + \dots$ can be neglected in comparison with N and one obtains

$$\log N! \approx \frac{1}{2} \log(2\pi) + (N + \frac{1}{2}) \log N - N \quad (\text{A2.2.3})$$

$$\approx N \log N - N \quad (\text{A2.2.4})$$

since both $\frac{1}{2}\log(2\pi)$ and $\frac{1}{2}$ are negligible compared to N . Using (A2.2.3), for $\log g(N, m)$ one obtains

$$(\partial T / \partial \mathbf{B}_0)_S = -(\partial S / \partial \mathbf{B}_0)_T / (\partial S / \partial T)_{\mathbf{B}_0},$$

which reduces to a Gaussian distribution for $g(N, m)$:

$$g(N, m) \approx g(N, 0) \exp(-2m^2/N) \quad (\text{A2.2.5})$$

with

$$g(N, 0) = \frac{N!}{(\frac{1}{2}N)!(\frac{1}{2}N)!} \approx \left(\frac{2}{\pi N}\right)^{\frac{1}{2}} 2^N. \quad (\text{A2.2.6})$$

When $m^2 = N/2$, the value of g is decreased by a factor of e from its maximum at $m = 0$. Thus the fractional width of the distribution is $\Delta(m/N) \sim (1/N)^{\frac{1}{2}}$. For $N \sim 10^{22}$ the fractional width is of the order of 10^{-11} . It is the sharply peaked behaviour of the degeneracy functions that leads to the prediction that the thermodynamic properties of macroscopic systems are well defined.

For this model system the magnetic potential energy in the presence of a uniform magnetic field \vec{H} is given by $-\mathcal{M} \cdot \vec{H}$ and, for \vec{H} pointing in $+z$ direction, it is $-\mathcal{M}H$ or $-2m\mu H$. A fixed magnetic potential energy thus implies a fixed value of m . For a given H , the magnetic potential energy of the system is bounded from above and below. This is not the

case for the next two examples. Besides the example of an ideal paramagnet that this model explicitly represents, there are many other systems which can be modelled as effective two-state systems. These include the lattice gas model of an ideal gas, binary alloy and a simple two-state model of a linear polymer. The time dependence of the mean square displacement of a Brownian particle can also be analysed using such a model.

(ii) N non-interacting point particles in a box. The microstate (orbital) of a free particle of mass M confined in a cube of volume L^3 is specified by three integers (quantum numbers): $(n_x, n_y, n_z) \equiv n$; $n_x, n_y, n_z = 1, 2, 3, \dots$. Its wavefunction is

$$\phi_n = (2/L)^{3/2} \sin(xp_x) \sin(yp_y) \sin(zp_z)$$

with $p = (h\pi/L)n$, and the energy $\epsilon = p^2/(2M) = (h\pi/L)^2 n^2/(2M)$. The energy grows quadratically with n , and without bounds. One can enumerate the orbitals by considering the positive octant of a sphere in the space defined by n_x, n_y and n_z for free particle orbitals. With every unit volume $\Delta n_x \Delta n_y \Delta n_z = 1$, there is one orbital per spin orientation of the particle. For particles of spin I , there are $\gamma = (2I + 1)$ independent spin orientations. The energy of an orbital on the surface of a sphere of radius n_0 in the n space is $\epsilon_0 = (h\pi/L)^2 n_0^2/(2M)$. The

degeneracy function, or equivalently, the number of orbitals in the allowed (positive) octant of a spherical shell of thickness Δn is $\gamma \frac{1}{8} 4\pi n^2 \Delta n = \frac{1}{2} \gamma \pi n^2 \Delta n$. This is an approximate result valid asymptotically for large n_0 . Often one needs the number of orbitals with energy between ϵ and $\epsilon + d\epsilon$. If it is denoted by $\mathcal{D}(\epsilon) d\epsilon$, it is easy to show by using

$$\mathcal{D}(\epsilon) d\epsilon = \frac{1}{2} \gamma \pi n^2 \frac{dn}{d\epsilon} d\epsilon \quad (\text{A2.2.7})$$

that

$$\mathcal{D}(\epsilon) = \frac{\gamma V}{4\pi^2} \left(\frac{2M}{\hbar^2} \right)^{\frac{3}{2}} \epsilon^{\frac{1}{2}}. \quad (\text{A2.2.8})$$

This is the density of microstates for one free particle in volume $V = L^3$.

For N non-interacting particles in a box, the result depends on the particle statistics: Fermi, Bose or Boltzmann. The state of a quantum system can be specified by the wavefunction for that state, $\psi_v(q_1, q_2, \dots, q_N)$. ψ_v is the v th eigensolution to the Schrödinger equation for an N -particle system. If the particles are non-interacting, then the wavefunction can be expressed in terms of the single-particle wavefunctions (ϕ_n given above is an example). Let these be denoted by $\phi_1(q), \phi_2(q), \dots, \phi_j(q), \dots$. For a specific state v , $\psi_v(q_1, q_2, \dots, q_N)$ will be the appropriately symmetrized product containing n_1 particles with the single-particle wavefunction $\phi_1(q)$, n_2 particles with $\phi_2(q)$, etc. For Fermi particles (with half integral spin) the product is antisymmetric and for Bose particles (with integer spin) it is symmetric. The antisymmetry of the Fermi particle wavefunction implies that fermions obey the Pauli exclusion principle. The numbers $n_1, n_2, \dots, n_j, \dots$ are the occupation numbers of the respective single-particle states, and this set of occupation numbers $\{n_j\}$ completely specify the state v of the system. If there are N_v particles in this state then

-5-

$N_v = \sum_j n_j$, and if the j th single-particle state has energy ϵ_j , then the energy of the system in the state v is $E_v = \sum_j \epsilon_j n_j$.

In an ideal molecular gas, each molecule typically has translational, rotational and vibrational degrees of freedom. The example of 'one free particle in a box' is appropriate for the translational motion. The next example of oscillators can be used for the vibrational motion of molecules.

(iii) N non-interacting harmonic oscillators. Energy levels of a harmonic oscillator are non-degenerate, characterized by a quantum number l and are given by the expression $\epsilon_l = (l + \frac{1}{2})\hbar\omega$, $l = 0, 1, 2, 3, \dots, \infty$. For a system of N independent oscillators, if the i th oscillator is in the state n_i , the set $\{n_i\}$ gives the microstate of the system, and its total energy E is given by $E = \frac{1}{2}N\hbar\omega + \sum_{i=1}^N n_i \hbar\omega$. Consider the case when the energy of the system, above its zero point energy $\frac{1}{2}N\hbar\omega$, is a fixed amount $E_0 \equiv (E - \frac{1}{2}N\hbar\omega)$. Define n such that $E_0 = n\hbar\omega$. Then, one needs to find the number of ways in which a set of $\{n_i\}$ can be chosen such that $\sum_{i=1}^N n_i = n$. This number is the degeneracy function $g(N, n)$ for this system. For a single-oscillator case, $n_1 = n$ and $g(1, n) = 1$, for all n . Consider a sum $\sum_{n=0}^{\infty} g(1, n)t^n$; it is called a generating function. Since $g(1, n) = 1$, it sums to $(1-t)^{-1}$, if $|t| < 1$. For n independent oscillators, one would therefore use $(1-t)^{-N}$, rewriting it as

$$S(N, V, U) = k \ln \Omega(N, V, U).$$

Now since in general,

$$e^{-U_i/kT}$$

its use for the specific example of N oscillators gives

$$Q(N, V, T) = \sum_i e^{-U_i(N, V)/kT}$$

$$A(N, V, T) = -kT \ln Q(N, V, T).$$

with the final result that the degeneracy function for the N -oscillator system is

$$g(N, n) = \frac{(N + n - 1)!}{n!(N - 1)!}. \quad (\text{A2.2.9})$$

The model of non-interacting harmonic oscillators has a broad range of applicability. Besides vibrational motion of molecules, it is appropriate for phonons in harmonic crystals and photons in a cavity (black-body radiation).

A2.2.2.1 CLASSICAL MECHANICS

The set of microstates of a finite system in quantum statistical mechanics is a finite, discrete denumerable set of quantum states each characterized by an appropriate collection of quantum numbers. In classical statistical mechanics, the set of microstates form a continuous (and therefore infinite) set of points in Γ space (also called phase space).

-6-

Following Gibbs, the Γ space is defined as a $2f$ -dimensional space for a system with f degrees of freedom: $(p_1, p_2, \dots, p_f; q_1, q_2, \dots, q_f)$, abbreviated as (p, q) . Here (p_i, q_i) , $i = 1, \dots, f$ are the canonical momenta and canonical coordinates of the f degrees of freedom of the system. Given a precise initial state (p^0, q^0) , a system with the Hamiltonian $\mathcal{H}(p, q)$ evolves deterministically according to the canonical equations of motion:

$$\frac{\partial \mathcal{H}(p, q)}{\partial p_i} = \dot{q}_i \quad \frac{\partial \mathcal{H}(p, q)}{\partial q_i} = -\dot{p}_i. \quad (\text{A2.2.10})$$

Now, if D/Dt represents time differentiation along the deterministic trajectory of the system in the Γ space, it follows that

$$\dot{\mathcal{H}} = \frac{D\mathcal{H}}{Dt} = \frac{\partial \mathcal{H}}{\partial t} + \sum_k \left(\dot{q}_k \frac{\partial \mathcal{H}}{\partial q_k} + \dot{p}_k \frac{\partial \mathcal{H}}{\partial p_k} \right) = \frac{\partial \mathcal{H}}{\partial t} \quad (\text{A2.2.11})$$

where the last equality is obtained using the equations of motion. Thus, when \mathcal{H} does not depend on time explicitly, i.e. when $\partial \mathcal{H} / \partial t = 0$, the above equation implies that $\mathcal{H}(p, q) = E = \text{constant}$. The locus of points in Γ space satisfying this condition defines a $(2f - 1)$ -dimensional energy hypersurface S , and the trajectory of such a system in Γ space would lie on this hypersurface. Furthermore, since a given trajectory is uniquely determined by the equations of motion and the initial conditions, two trajectories in Γ space can never intersect.

A2.2.2.2 LIOUVILLE'S THEOREM

The volume of a Γ -space-volume-element does not change in the course of time if each of its points traces out a trajectory in Γ space determined by the equations of motion. Equivalently, the Jacobian

$$J(t, t_0) \equiv \frac{\partial(p, q)}{\partial(p^0, q^0)} = 1. \tag{A2.2.12}$$

Liouville's theorem is a restatement of mechanics. The proof of the theorem consists of two steps.

(1) Expand $J(t, t_0)$ around t_0 to obtain:

$$\sigma_x^2 = \langle x^2 \rangle - \langle x \rangle^2$$

Hence,

$$\sigma_U^2 = kT^2 C_V$$

-7-

(2) From the multiplication rule of Jacobians, one has for any t_1 between t_0 and t ,

$$\sigma_N^2 = kT\kappa_T(N^2/V).$$

Now let t_1 approach t . Then, by the result of the first step, the first factor on the right-hand side vanishes. Hence,

$$\frac{\partial J(t, t_0)}{\partial t} = 0 \quad \text{and} \quad J(t, t_0) = \text{constant}.$$

Finally, since $J(t_0, t_0) = 1$, one obtains the result $J(t, t_0) = 1$, which concludes the proof of Liouville's theorem.

Geometrically, Liouville's theorem means that if one follows the motion of a small phase volume in Γ space, it may change its shape but its volume is invariant. In other words the motion of this volume in Γ space is like that of an incompressible fluid. Liouville's theorem, being a restatement of mechanics, is an important ingredient in the formulation of the theory of statistical ensembles, which is considered next.

A2.2.3 STATISTICAL ENSEMBLES

In equilibrium statistical mechanics, one is concerned with the thermodynamic and other macroscopic properties of matter. The aim is to derive these properties from the laws of molecular dynamics and thus create a link between microscopic molecular motion and thermodynamic behaviour. A typical macroscopic system is composed of a large number N of molecules occupying a volume V which is large compared to that occupied by a molecule:

$$N \approx 10^{23} \text{ molecules} \quad V \approx 10^{23} \text{ molecular volumes.}$$

Due to such large numbers, it is useful to consider the limiting case of the *thermodynamic limit*, which is defined as

$$N \rightarrow \infty \quad V \rightarrow \infty \quad \frac{V}{N} = v \quad (\text{A2.2.13})$$

where the specific volume v is a given finite number. For a three-dimensional system with N *point* particles, the total number of degrees of freedom $f = 3N$.

A statistical ensemble can be viewed as a description of how an experiment is repeated. In order to describe a macroscopic system in equilibrium, its thermodynamic state needs to be specified first. From this, one can infer the macroscopic constraints on the system, i.e. which macroscopic (thermodynamic) quantities are held fixed. One can also deduce, from this, what are the corresponding microscopic variables which will be constants of motion. A macroscopic system held in a specific thermodynamic equilibrium state is typically consistent with a very large number (classically infinite) of microstates. Each of the repeated experimental measurements on such a system, under ideal

-8-

conditions and with identical macroscopic constraints, would correspond to the system in a different accessible microstate which satisfies the macroscopic constraints. It is natural to represent such a collection of microstates by an ensemble (a mental construct) of systems, which are identical in composition and macroscopic conditions (constraints), but each corresponding to a different microstate. For a properly constructed ensemble, each of its member systems satisfies the macroscopic constraints appropriate to the experimental conditions. Collectively the ensemble then consists of all the microstates that satisfy the macroscopic constraints (all accessible states). The simplest assumption that one can make in order to represent the repeated set of experimental measurements by the ensemble of accessible microstates is to give each an equal weight. The fundamental assumption in the ensemble theory is then the *Postulate of 'Equal a priori probabilities'*. It states that 'when a macroscopic system is in thermodynamic equilibrium, its microstate is equally likely to be any of the accessible states, each of which satisfy the macroscopic constraints on the system'.

Such an ensemble of systems can be geometrically represented by a distribution of representative points in the Γ space (classically a continuous distribution). It is described by an ensemble density function $\rho(p, q, t)$ such that $\rho(p, q, t)d^{2f}\Omega$ is the number of representative points which at time t are within the infinitesimal phase volume element $d^f p d^f q$ (denoted by $d^{2f}\Omega$) around the point (p, q) in the Γ space.

Let us consider the consequence of mechanics for the ensemble density. As in [subsection A2.2.2.1](#), let D/Dt represent differentiation along the trajectory in Γ space. By definition,

$$\frac{D}{Dt}(\rho d^{2f}\Omega) = 0.$$

According to Liouville's theorem,

$$\frac{D}{Dt}(d^{2f}\Omega) = 0.$$

Therefore,

$$\frac{D\rho}{Dt} = 0$$

or, equivalently,

$$\frac{\partial \rho}{\partial t} + \sum_k \left(\dot{q}_k \frac{\partial \rho}{\partial q_k} + \dot{p}_k \frac{\partial \rho}{\partial p_k} \right) = 0$$

which can be rewritten in terms of Poisson brackets using the equations of motion, (A2.2.10):

$$\frac{\partial \rho}{\partial t} + \sum_k \left(\frac{\partial \mathcal{H}}{\partial p_k} \frac{\partial \rho}{\partial q_k} - \frac{\partial \mathcal{H}}{\partial q_k} \frac{\partial \rho}{\partial p_k} \right) = 0.$$

-9-

This is same as

$$\frac{\partial \rho}{\partial t} + [\mathcal{H}, \rho]_{\text{P.B.}} = 0. \quad (\text{A2.2.14})$$

For the quantum mechanical case, ρ and \mathcal{H} are operators (or matrices in appropriate representation) and the Poisson bracket is replaced by the commutator $[\mathcal{H}, \rho]$. If the distribution is stationary, as for the systems in equilibrium, then $\partial \rho / \partial t = 0$, which implies

$$[\mathcal{H}, \rho]_{\text{P.B.}} = 0 \quad \text{classically} \quad \text{and} \quad [\mathcal{H}, \rho]_- = 0 \quad \text{quantum mechanically.} \quad (\text{A2.2.15})$$

A stationary ensemble density distribution is constrained to be a functional of the constants of motion (globally conserved quantities). In particular, a simple choice is $\rho(p, q) = \rho^*(\mathcal{H}(p, q))$, where $\rho^*(\mathcal{H})$ is some functional (function of a function) of \mathcal{H} . Any such functional has a vanishing Poisson bracket (or a commutator) with \mathcal{H} and is thus a stationary distribution. Its dependence on (p, q) through $\mathcal{H}(p, q) = E$ is expected to be reasonably smooth. Quantum mechanically, $\rho^*(\mathcal{H})$ is the density operator which has some functional dependence on the Hamiltonian \mathcal{H} depending on the ensemble. It is also normalized: $\text{Tr} \rho = 1$. The density matrix is the matrix representation of the density operator in some chosen representation of a complete orthonormal set of states. If the complete orthonormal set of eigenstates of the Hamiltonian is known:

$$\mathcal{H}|v\rangle = E_v|v\rangle \quad \langle v|v'\rangle = \delta_{vv'}$$

then the density operator is

$$\rho = \sum_v |v\rangle \langle v| \rho^*(E_v).$$

Often the eigenstates of the Hamiltonian are not known. Then one uses an appropriate set of states $|u\rangle$ which

are complete and orthonormal. In any such representation the density matrix, given as $\langle \nu | \rho^*(\mathcal{H}) | \nu \rangle$, is not diagonal.

A2.2.3.1 MICROCANONICAL ENSEMBLE

An explicit example of an equilibrium ensemble is the microcanonical ensemble, which describes closed systems with adiabatic walls. Such systems have constraints of fixed N , V and $E < \mathcal{H} < E + dE$. dE is very small compared to E , and corresponds to the assumed very weak interaction of the ‘isolated’ system with the surroundings. dE has to be chosen such that it is larger than $(\delta E)_{qu} \sim h/t_{ob}$ where h is the Planck’s constant and t_{ob} is the duration of the observation time. In such a case, even though dE may be small, there will be a great number of microstates for a macroscopic size system. For a microcanonical ensemble, the ‘equal *a priori* probability’ postulate gives its density distribution as:

-10-

classically,

$$\rho(p, q) = \begin{cases} \text{constant} & \text{if } E < \mathcal{H}(p, q) < E + dE \\ 0 & \text{otherwise} \end{cases} \quad (\text{A2.2.16})$$

quantum mechanically, if the system microstate is denoted by l , then

$$\rho_l = \begin{cases} \text{constant} & \text{if } E < E_l < E + dE \\ 0 & \text{otherwise.} \end{cases} \quad (\text{A2.2.17})$$

One considers systems for which the energy shell is a closed (or at least finite) hypersurface S . Then the energy shell has a finite volume:

$$\int_{E < \mathcal{H} < E + dE} d^{2f} \Omega.$$

For each degree of freedom, classical states within a small volume $\Delta p_i \Delta q_i \sim h$ merge into a single quantum state which cannot be further distinguished on account of the uncertainty principle. For a system with f degrees of freedom, this volume is h^f . Furthermore, due to the indistinguishability of identical particles in quantum mechanics, there are $N!$ distinguishable classical states for each quantum mechanical state, which are obtained by simple permutations of the N particles in the system. Then the number of microstates $\Gamma(E)$ in Γ space occupied by the microcanonical ensemble is given by

$$\Gamma(E) \equiv \mathcal{D}(E) d(E) = [h^f N!]^{-1} \int_{E < \mathcal{H} < E + dE} d^{2f} \Omega \quad (\text{A2.2.18})$$

where f is the total number of degrees of freedom for the N -particle system, and $\mathcal{D}(E)$ is the density of states of the system at energy E . If the system of N particles is made up of N_A particles of type A, N_B particles of type B, . . . , then $N!$ is replaced by $N_A! N_B! \dots$. Even though dE is conceptually essential, it does not affect the thermodynamic properties of macroscopic systems. In order that the ensemble density ρ is normalized, the ‘constant’ above in (A2.2.16) has to be $[\Gamma(E)]^{-1}$. Quantum mechanically $\Gamma(E)$ is simply the total number of microstates within the energy interval $E < E_l < E + dE$, and fixes the ‘constant’ in (A2.2.17). $\Gamma(E)$ is the microcanonical partition function; in addition to its indicated dependence on E , it also depends on N and V .

Consider a measurable property $\mathcal{B}(p, q)$ of the system, such as its energy or momentum. When a system is in

equilibrium, according to Boltzmann, what is observed macroscopically are the time averages of the form

$$\bar{\mathcal{B}}^t = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T \mathcal{B}(p_t, q_t) dt. \quad (\text{A2.2.19})$$

It was assumed that, apart from a vanishingly small number of exceptions, the initial conditions do not have an effect on these averages. However, since the limiting value of the time averages cannot be computed, an ergodic hypothesis

-11-

was introduced: *time averages are identical with statistical averages over a microcanonical ensemble, for reasonable functions \mathcal{B} , except for a number of initial conditions, whose importance is vanishingly small compared with that of all other initial conditions.* Here the ensemble average is defined as

$$\langle \mathcal{B} \rangle = \frac{\int d^{2f} \Omega \mathcal{B}(p, q) \rho(p, q)}{\int d^{2f} \Omega \rho(p, q)}. \quad (\text{A2.2.20})$$

The thinking behind this was that, over a long time period, a system trajectory in Γ space passes through every configuration in the region of motion (here the energy shell), i.e. the system is ergodic, and hence the infinite time average is equal to the average in the region of motion, or the average over the microcanonical ensemble density. The ergodic hypothesis is meant to provide justification of the ‘equal *a priori* probability’ postulate. It is a strong condition. For a system of N particles, the infinitely long time must be much longer than $O(e^N)$, whereas the usual observation time window is $O(1)$. (When one writes $y = O(x)$ and $z = o(x)$, it implies that $\lim_{x \rightarrow \infty} y/z = \text{finite} \neq 0$ and $\lim_{x \rightarrow \infty} z/x = 0$.) However, if by ‘reasonable functions \mathcal{B} ’ one means large variables $O(N)$, then their values are nearly the same everywhere in the region of motion and the trajectory need not be truly ergodic for the time average to be equal to the ensemble average. Ergodicity of a trajectory is a difficult mathematical problem in mechanics.

The microcanonical ensemble is a certain model for the repetition of experiments: in every repetition, the system has ‘exactly’ the same energy, N and V ; but otherwise there is no experimental control over its microstate. Because the microcanonical ensemble distribution depends only on the total energy, which is a constant of motion, it is time independent and mean values calculated with it are also time independent. This is as it should be for an equilibrium system. Besides the ensemble average value $\langle \mathcal{B} \rangle$, another commonly used ‘average’ is the most probable value, which is the value of $\mathcal{B}(p, q)$ that is possessed by the largest number of systems in the ensemble. The ensemble average and the most probable value are nearly equal if the *mean square fluctuation* is small, i.e. if

$$\frac{\langle \mathcal{B}^2 \rangle - \langle \mathcal{B} \rangle^2}{\langle \mathcal{B} \rangle^2} \ll 1. \quad (\text{A2.2.21})$$

If this condition is not satisfied, there is no unique way of calculating the observed value of \mathcal{B} , and the validity of the statistical mechanics should be questioned. In all physical examples, the mean square fluctuations are of the order of $1/N$ and vanish in the thermodynamic limit.

A2.2.3.2 MIXING

In the last subsection, the microcanonical ensemble was formulated as an ensemble from which the *equilibrium* properties of a dynamical system can be determined by its energy alone. We used the postulate of

equal *a priori* probability and gave a discussion of the ergodic hypothesis. The ergodicity condition, even though a strong condition, does not ensure that if one starts from a *non-equilibrium* ensemble the expectation values of dynamical functions will approach their equilibrium values as time proceeds. For this, one needs a stronger condition than ergodicity, a condition of *mixing*. Every mixing system is ergodic, but the reverse is not true.

-12-

Consider, at $t = 0$, some non-equilibrium ensemble density $\rho_{\text{ne}}(p^o, q^o)$ on the constant energy hypersurface S , such that it is normalized to one. By Liouville's theorem, at a later time t the ensemble density becomes $\rho_{\text{ne}}(\phi_{-t}(p, q))$, where $\phi_{-t}(p, q)$ is the function that takes the current phase coordinates (p, q) to their initial values time (t) ago; the function ϕ is uniquely determined by the equations of motion. The expectation value of any dynamical variable \mathcal{B} at time t is therefore

$$\int_S d^{2f} \Omega \mathcal{B}(p, q) \rho_{\text{ne}}(\phi_{-t}(p, q)). \quad (\text{A2.2.22})$$

As t becomes large, this should approach the equilibrium value $\langle \mathcal{B} \rangle$, which for an ergodic system is

$$\frac{\int d^{2f} \Omega \mathcal{B}(p, q) \rho(p, q)}{\int d^{2f} \Omega \rho(p, q)} = \frac{\int_S d^{2f} \Omega \mathcal{B}(p, q)}{\int_S d^{2f} \Omega} \quad (\text{A2.2.23})$$

where S is the hypersurface of the energy shell for the microcanonical ensemble. This equality is satisfied if the system is *mixing*.

A system is mixing if, for every pair of functions f and g whose squares are integrable on S ,

$$\lim_{t \rightarrow \pm\infty} \int_S d^{2f} \Omega f(p, q) g(\phi_{-t}(p, q)) = \frac{\int_S d^{2f} \Omega f(p, q) \int_S d^{2f} \Omega g(p, q)}{\int_S d^{2f} \Omega}. \quad (\text{A2.2.24})$$

The statement of the mixing condition is equivalent to the following: if Q and R are arbitrary regions in S , and an ensemble is initially distributed uniformly over Q , then the fraction of members of the ensemble with phase points in R at time t will approach a limit as $t \rightarrow \infty$, and this limit equals the fraction of area of S occupied by R .

The ensemble density $\rho_{\text{ne}}(p_r, q_r)$ of a mixing system does not approach its equilibrium limit in the pointwise sense. It is only in a 'coarse-grained' sense that the average of $\rho_{\text{ne}}(p_r, q_r)$ over a region R in S approaches a limit to the equilibrium ensemble density as $t \rightarrow \infty$ for each fixed R .

In the condition of mixing, equation (A2.2.24), if the function g is replaced by ρ_{ne} , then the integral on the left-hand side is the expectation value of f , and at long times approaches the equilibrium value, which is the microcanonical ensemble average $\langle f \rangle$, given by the right-hand side. The condition of mixing is a sufficient condition for this result. The condition of mixing for equilibrium systems also has the implication that every equilibrium time-dependent correlation function, such as $\langle f(p, q) g(\phi_t(p, q)) \rangle$, approaches a limit of the uncorrelated product $\langle f \rangle \langle g \rangle$ as $t \rightarrow \infty$.

A2.2.3.3 ENTROPY

For equilibrium systems, thermodynamic entropy is related to ensemble density distribution ρ as

$$S = -k_b \langle \log \rho \rangle \quad (\text{A2.2.25})$$

-13-

where k_b is a universal constant, the Boltzmann's constant. For equilibrium systems, ρ is a functional of constants of motion like energy and is time independent. Thus, the entropy as defined here is also invariant over time. In what follows it will be seen that this definition of entropy obeys all properties of thermodynamic entropy.

A low-density gas which is not in equilibrium, is well described by the one-particle distribution $f(\vec{v}, \vec{r}, t)$, which describes the behaviour of a particle in μ space of the particle's velocity \vec{v} and position \vec{r} . One can obtain $f(\vec{v}, \vec{r}, t)$ from the classical density distribution $\rho(p, q, t)$ (defined earlier following the postulate of equal *a priori* probabilities) by integrating over the degrees of freedom of the remaining $(N - 1)$ particles. Such a coarse-grained distribution f satisfies the Boltzmann transport equation (see section A3.1). Boltzmann used the crucial assumption of molecular chaos in deriving this equation. From f one can define an H function as $H(t) = \langle \log f \rangle_{\text{ne}}$ where the non-equilibrium average is taken over the time-dependent f . Boltzmann proved an H theorem, which states that 'if at a given instant t , the state of gas satisfies the assumption of molecular chaos, then at the instant $t + \epsilon$ ($\epsilon \rightarrow 0$), $dH(t) / dt = 0$; the equality $dH(t) / dt = 0$ is satisfied if and only if f is the equilibrium Maxwell-Boltzmann distribution'; i.e. $H(t)$ obtained from the Boltzmann transport equation is a monotonically decreasing function of t . Thus a generalization of the equilibrium definition of entropy to systems slightly away from equilibrium can be made:

$$S(t) = -k_b \langle \log f \rangle_{\text{ne}} \equiv -k_b H(t). \quad (\text{A2.2.26})$$

Such a generalization is consistent with the Second Law of Thermodynamics, since the H theorem and the generalized definition of entropy together lead to the conclusion that the entropy of an isolated non-equilibrium system increases monotonically, as it approaches equilibrium.

A2.2.3.4 ENTROPY AND TEMPERATURE IN A MICROCANONICAL ENSEMBLE

For a microcanonical ensemble, $\rho = [\Gamma(E)]^{-1}$ for each of the allowed $\Gamma(E)$ microstates. Thus for an isolated system in equilibrium, represented by a microcanonical ensemble,

$$S = k_b \log \Gamma(E). \quad (\text{A2.2.27})$$

Consider the microstates with energy E_l such that $E_l \leq E$. The total number of such microstates is given by

$$\Sigma(E) = [h^f N!]^{-1} \int_{0 < \mathcal{H} < E} d^{2f} \Omega. \quad (\text{A2.2.28})$$

Then $\Gamma(E) = \Sigma(E + dE) - \Sigma(E)$, and the density of states $\mathcal{D}(E) = d\Sigma/dE$. A system containing a large number of particles N , or an indefinite number of particles but with a macroscopic size volume V , normally has the number of states Σ , which approaches asymptotically to

$$\Sigma \sim \exp \left(N \phi \left(\frac{E}{N}, \frac{V}{N} \right) \right) \quad \text{or} \quad \Sigma \sim \exp \left(V \psi \left(\frac{E}{V}, \frac{N}{V} \right) \right) \quad (\text{A2.2.29})$$

where E/N , E/V and ϕ or ψ are each $\sim O(1)$, and

$$\phi > 0 \quad \phi' > 0 \quad \phi'' < 0. \quad (\text{A2.2.30})$$

Consider the three examples considered in (section A2.2.2). For examples (i) and (iii), the degeneracy function $g(N, n)$ is a discrete analogue of $\Gamma(E)$. Even though $\Sigma(E)$ can be obtained from $g(N, n)$ by summing it over n from the lowest-energy state up to energy E , the largest value of n dominates the sum if N is large, so that $g(N, n)$ is also like $\Sigma(E)$. For example (ii), $\Sigma(E)$ can be obtained from the density of states $\mathcal{D}(\epsilon)$ by an integration over ϵ from zero to E . $\Sigma(E)$ so obtained conforms to the above asymptotic properties for large N , for the last two of the three examples. For the first example of ‘ N non-interacting spin- $\frac{1}{2}$ particles in a magnetic field’, this is the case only for the energy states corresponding to $0 < m < N/2$. (The other half state space with $0 > m > (-N/2)$, corresponding to the positive magnetic potential energy in the range between zero and $\mu H N$, corresponds to the system in non-equilibrium states, which have sometimes been described using ‘negative temperatures’ in an equilibrium description. Such peculiarities often occur in a model system which has a finite upper bound to its energy.)

Using the asymptotic properties of $\Sigma(E)$ for large N , one can show that, within an additive constant $\sim O(\log N)$ or smaller, the three quantities $k_b \log \Gamma(E)$, $k_b \log \mathcal{D}(E)$ and $k_b \log \Sigma(E)$ are equivalent and thus any of the three can be used to obtain S for a large system. This leads to the result that $S = k_b \log \Gamma(E) = k_b N \phi$, so that *the entropy as defined is an extensive quantity*, consistent with the thermodynamic behaviour.

For an isolated system, among the independent macroscopic variables N , V and E , only V can change. Now V cannot decrease without compressing the system, and that would remove its isolation. Thus V can only increase, as for example is the case for the free expansion of a gas when one of the containing walls is suddenly removed. For such an adiabatic expansion, the number of microstates in the final state is larger; thus the entropy of the final state is larger than that of the initial state. More explicitly, note that $\Sigma(E)$ is a non-decreasing function of V , since if $V_1 > V_2$, then the integral in the defining equation, (A2.2.28), for $V = V_1$ extends over a domain of integration that includes that for $V = V_2$. Thus $S(E, V) = k_b \log \Sigma(E)$ is a non-decreasing function of V . This is *also consistent* with the Second Law of Thermodynamics.

Next, let x_i be either p_i or q_i , ($i = 1, \dots, f$). Consider the ensemble average $\langle x_i \partial \mathcal{H} / \partial x_j \rangle$:

$$\begin{aligned} \left\langle x_i \frac{\partial \mathcal{H}}{\partial x_j} \right\rangle &= \frac{\int_{E < \mathcal{H} < E + dE} d^{2f} \Omega x_i (\partial \mathcal{H} / \partial x_j)}{\int_{E < \mathcal{H} < E + dE} d^{2f} \Omega} \\ &= \frac{dE (\partial / \partial E) \int_{\mathcal{H} < E} d^{2f} \Omega x_i (\partial \mathcal{H} / \partial x_j)}{\mathcal{D} E dE}. \end{aligned}$$

Since $\partial E / \partial x_j = 0$,

$$\begin{aligned} \int_{\mathcal{H} < E} d^{2f} \Omega x_i \frac{\partial \mathcal{H}}{\partial x_j} &= \int_{\mathcal{H} < E} d^{2f} \Omega x_i \frac{\partial (\mathcal{H} - E)}{\partial x_j} \\ &= \int_{\mathcal{H} < E} d^{2f} \Omega \frac{\partial}{\partial x_j} [x_i (\mathcal{H} - E)] - \delta_{ij} \int_{\mathcal{H} < E} d^{2f} \Omega (\mathcal{H} - E). \end{aligned}$$

The first integral on the right-hand side is zero: it becomes a surface integral over the boundary where $(\mathcal{H} - E) = 0$. Using the result in the previous equation, one obtains

$$\begin{aligned}
\left\langle x_i \frac{\partial \mathcal{H}}{\partial x_j} \right\rangle &= \frac{\delta_{ij}}{\mathcal{D}(E)} \frac{\partial}{\partial E} \int_{\mathcal{H} < E} d^{2f} \Omega (E - \mathcal{H}) \\
&= \frac{\delta_{ij}}{\mathcal{D}(E)} \int_{\mathcal{H} < E} d^{2f} \Omega = \frac{\delta_{ij}}{\mathcal{D}(E)} \Sigma(E) \\
&= \delta_{ij} \frac{\Sigma(E)}{\partial \Sigma(E) / \partial E} = \delta_{ij} \left[\frac{\partial}{\partial E} \log \Sigma(E) \right]^{-1} \\
&= \delta_{ij} \frac{k_b}{\partial S / \partial E}.
\end{aligned} \tag{A2.2.31}$$

In a microcanonical ensemble, the internal energy of a system is

$$U \equiv \langle \mathcal{H} \rangle \sim E. \tag{A2.2.32}$$

Since the temperature T relates U and S as $T = (\partial U / \partial S)_V$, it is appropriate to make the identification

$$\frac{\partial S}{\partial E} \equiv \frac{1}{T}. \tag{A2.2.33}$$

Since T is positive for systems in thermodynamic equilibrium, S and hence $\log \Sigma$ should both be monotonically increasing functions of E . This is the case as discussed above.

With this identification of T , the above result reduces to the generalized equipartition theorem:

$$\left\langle x_i \frac{\partial \mathcal{H}}{\partial x_j} \right\rangle = \delta_{ij} k_b T. \tag{A2.2.34}$$

-16-

For $i = j$ and $x_i = p_i$, one has

$$\left\langle p_i \frac{\partial \mathcal{H}}{\partial p_i} \right\rangle = k_b T \tag{A2.2.35}$$

and for $i = j$ and $x_i = q_i$,

$$\left\langle q_i \frac{\partial \mathcal{H}}{\partial q_i} \right\rangle = k_b T. \tag{A2.2.36}$$

Now since $\partial \mathcal{H} / \partial q_i = -\dot{p}_i$, one gets the virial theorem:

$$\left\langle \sum_{i=1}^f q_i \dot{p}_i \right\rangle = -f k_b T. \tag{A2.2.37}$$

There are many physical systems which are modelled by Hamiltonians, which can be transformed through a canonical transformation to a quadratic form:

$$\mathcal{H} = \sum_i (a_i p_i^2 + b_i q_i^2) \quad (\text{A2.2.38})$$

where p_i and q_i are canonically conjugate variables and a_i and b_i are constants. For such a form of a Hamiltonian:

$$\sum_i \left(p_i \frac{\partial \mathcal{H}}{\partial p_i} + q_i \frac{\partial \mathcal{H}}{\partial q_i} \right) = 2\mathcal{H}. \quad (\text{A2.2.39})$$

If f of the constants a_i and b_i are non-zero, then it follows from above that

$$\langle \mathcal{H} \rangle = \frac{1}{2} f k_b T. \quad (\text{A2.2.40})$$

Each harmonic term in the Hamiltonian contributes $\frac{1}{2} k_b T$ to the average energy of the system, which is the theorem of the equipartition of energy. Since this is also the internal energy U of the system, one can compute the heat capacity

$$\frac{C_V}{k_b} = \frac{f}{2}. \quad (\text{A2.2.41})$$

This is a classical result valid only at high temperatures. At low temperatures, quantum mechanical attributes of a degree of freedom can partially or fully freeze it, thereby modifying or removing its contribution to U and C_V .

A2.2.3.5 THERMODYNAMICS IN A MICROCANONICAL ENSEMBLE: CLASSICAL IDEAL GAS

The definition of entropy and the identification of temperature made in the last subsection provides us with a connection between the microcanonical ensemble and thermodynamics.

A quasistatic thermodynamic process corresponds to a slow variation of E , V and N . This is performed by coupling the system to external agents. During such a process the ensemble is represented by uniformly distributed points in a region in Γ space, and this region slowly changes as the process proceeds. The change is slow enough that at every instant we have a microcanonical ensemble. Then the change in the entropy during an infinitesimal change in E , V and N during the quasistatic thermodynamic process is

$$dS(E, V, N) = \left(\frac{\partial S}{\partial E} \right)_{V, N} dE + \left(\frac{\partial S}{\partial V} \right)_{E, N} dV + \left(\frac{\partial S}{\partial N} \right)_{E, V} dN. \quad (\text{A2.2.42})$$

The coefficient of dE is the inverse absolute temperature as identified above. We now define the pressure and chemical potential of the system as

$$(\text{A2.2.43})$$

$$P \equiv T \left(\frac{\partial S}{\partial V} \right)_{E,N} \quad \mu \equiv -T \left(\frac{\partial S}{\partial N} \right)_{E,V}.$$

Then one has

$$dS = \frac{1}{T} (dE + P dV - \mu dN) \quad \text{or } dE = T dS - P dV + \mu dN. \quad (\text{A2.2.44})$$

This is the First Law of Thermodynamics.

The complete thermodynamics of a system can now be obtained as follows. Let the isolated system with N particles, which occupies a volume V and has an energy E within a small uncertainty dE , be modelled by a microscopic Hamiltonian \mathcal{H} . First, find the density of states $\mathcal{D}(E)$ from the Hamiltonian. Next, obtain the entropy as $S(E, V, N) = k_b \log \mathcal{D}(E)$ or, alternatively, by either of the other two equivalent expressions involving $\Gamma(E)$ or $\Sigma(E)$. Then, solve for E in terms of S, V and N . This is the internal energy of the system: $U(S, V, N) = E(S, V, N)$. Finally, find other thermodynamic functions as follows: the absolute temperature from $T = (\partial U / \partial S)_{V, N}$, the pressure from $P = T (\partial S / \partial V)_{E, N} = -(\partial U / \partial V)_{S, N}$, the Helmholtz free energy from $A = U - T S$, the enthalpy from $H = U + P V$, the Gibbs free energy from $G = U + P V - T S$, $\mu = G / N$ and the heat capacity at constant volume from $C_V = (\partial U / \partial T)_{V, N}$.

To illustrate, consider an ideal classical gas of N molecules occupying a volume V and each with mass M and three degrees of translational motion. The Hamiltonian is

-18-

$$\mathcal{H} = \frac{1}{2M} \sum_{i=1}^N p_i^2. \quad (\text{A2.2.45})$$

Calculate the $\Sigma(E)$ first. It is

$$\Sigma(E) = [h^{3N} N!]^{-1} \int_{0 < \mathcal{H} < E} d^{6N} \Omega = \frac{1}{N!} \left(\frac{V}{h^3} \right)^N \int_{\mathcal{H} < E} d^3 p_1 \dots d^3 p_N.$$

If $P_0 = (2ME)^{\frac{1}{2}}$, then the integral is the volume of a $3N$ -sphere of radius P_0 which is also equal to $C_{3N} P_0^{3N}$ where C_{3N} is the volume of a unit sphere in $3N$ dimensions. It can be shown that

$$C_{3N} = \frac{\pi^{3N/2}}{(3N/2)!}.$$

For large N , $N! \sim N^N e^{-N}$ and C_{3N} reduces to

$$C_{3N} = \left(\frac{2\pi}{3N} \right)^{3N/2} \exp(3N/2).$$

This gives

$$\Sigma(E) = \frac{C_{3N}}{N!} \left(\frac{V}{h^3} (2ME)^{\frac{3}{2}} \right)^N. \quad (\text{A2.2.46})$$

Now one can use $S = k_b \log \Sigma$. Then, for large N , for entropy one obtains

$$S(E, V, N) = Nk_b \left[\frac{5}{2} + \log \left(\frac{V}{N} \left(\frac{4\pi ME}{3h^2 N} \right)^{\frac{3}{2}} \right) \right]. \quad (\text{A2.2.47})$$

It is now easy to invert this result to obtain $E(S, V, N) \equiv U(S, V, N)$:

$$U(S, V, N) = \frac{3h^2}{4\pi} N \left(\frac{N}{V} \right)^{\frac{2}{3}} \exp \left(\frac{2}{3} \frac{S}{Nk_b} - \frac{5}{3} \right). \quad (\text{A2.2.48})$$

As expected, S and U are extensive, i.e. are proportional to N . From U one can obtain the temperature

-19-

$$T = \left(\frac{\partial U}{\partial S} \right)_{V,N} = \frac{2}{3} \frac{U}{Nk_b}. \quad (\text{A2.2.49})$$

From this result it follows that

$$C_V = \left(\frac{\partial U}{\partial T} \right)_{V,N} = \frac{3}{2} Nk_b. \quad (\text{A2.2.50})$$

Finally, the equation of state is

$$P = - \left(\frac{\partial U}{\partial V} \right)_{S,N} = \frac{2}{3} \frac{U}{V} = \frac{Nk_b T}{V} \quad (\text{A2.2.51})$$

and the chemical potential is

$$\mu = k_b T \log \left(\frac{N}{V} \left(\frac{h^2}{2\pi M k_b T} \right)^{\frac{3}{2}} \right). \quad (\text{A2.2.52})$$

For practical calculations, the microcanonical ensemble is not as useful as other ensembles corresponding to more commonly occurring experimental situations. Such equilibrium ensembles are considered next.

A2.2.3.6 INTERACTION BETWEEN SYSTEMS

Between two systems there can be a variety of interactions. Thermodynamic equilibrium of a system implies thermal, chemical and mechanical equilibria. It is therefore logical to consider, in sequence, the following interactions between two systems: *thermal contact*, which enables the two systems to share energy; *material contact*, which enables exchange of particles between them; and *pressure transmitting contact*, which allows an exchange of volume between the two systems. In each of the cases, the combined composite system is

supposed to be isolated (surrounded by adiabatic walls as described in section A2.1).

In addition, there could be a mechanical or electromagnetic interaction of a system with an external entity which may do work on an otherwise isolated system. Such a contact with a work source can be represented by the Hamiltonian $\mathcal{H}(p, q, x)$ where x is the coordinate (for example, the position of a piston in a box containing a gas, or the magnetic moment if an external magnetic field is present, or the electric dipole moment in the presence of an external electric field) describing the interaction between the system and the external work source. Then the force, canonically conjugate to x , which the system exerts on the outside world is

$$X = \frac{\partial \mathcal{H}(p, q, x)}{\partial x}. \quad (\text{A2.2.53})$$

-20-

A thermal contact between two systems can be described in the following way. Let two systems with Hamiltonians \mathcal{H}_I and \mathcal{H}_{II} be in contact and interact with Hamiltonian \mathcal{H}' . Then the composite system (I + II) has Hamiltonian $\mathcal{H} = \mathcal{H}_I + \mathcal{H}_{II} + \mathcal{H}'$. The interaction should be weak, such that the microstate of the composite system, say l , is specified by giving the microstate l' of system I and the microstate l'' of system II, with the energy E_l of the composite system given, to a good approximation, by $E_l = E_{l'} + E_{l''}$ where $l = (l', l'')$. The existence of the weak interaction is supposed to allow a sufficiently frequent exchange of energy between the two systems in contact. Then, after sufficient time, one expects the composite system to reach a final state regardless of the initial states of the subsystems. In the final state, every microstate (l', l'') of the composite system will be realized with equal probability, consistent with the postulate of equal *a priori* probability. Any such final state is a state of statistical equilibrium, the corresponding ensemble of states is called a *canonical ensemble*, and corresponds to thermal equilibrium in thermodynamics. The thermal contact as described here corresponds to a diathermic wall in thermodynamics (see section A2.1).

Contacts between two systems which enable them to exchange energy (in a manner similar to thermal contact) and to exchange particles are other examples of interaction. In these cases, the microstates of the composite system can be given for the case of a weak interaction by, $(N, l) = (N', l'; N'', l'')$. The sharing of the energy and the number of particles lead to the constraints: $E_l(N) = E_{l'}(N') + E_{l''}(N'')$ and $N = N' + N''$. The corresponding equilibrium ensemble is called a *grand canonical ensemble*, or a $T - \mu$ ensemble.

Finally, if two systems are separated by a movable diathermic (perfectly conducting) wall, then the two systems are able to exchange energy and volume: $E_l(V) = E_{l'}(V') + E_{l''}(V'')$ and $V = V' + V''$. If the interaction is weak, the microstate of the composite system is $(V, l) = (V', l'; V'', l'')$, and the corresponding equilibrium ensemble is called the $T - P$ ensemble.

A2.2.4 CANONICAL ENSEMBLE

Consider two systems in thermal contact as discussed above. Let the system II (with volume V_R and particles N_R) correspond to a reservoir R which is much larger than the system I (with volume V and particles N) of interest. In order to find the canonical ensemble distribution one needs to obtain the probability that the system I is in a specific microstate ν which has an energy E_ν . When the system is in this microstate, the reservoir will have the energy $E_R = E_T - E_\nu$ due to the constraint that the total energy of the isolated composite system I+II is fixed and denoted by E_T ; but the reservoir can be in any one of the $\Gamma_R(E_T - E_\nu)$ possible states that the mechanics within the reservoir dictates. Given that the microstate of the system of

interest is specified to be ν , the total number of accessible states for the composite system is clearly $\Gamma_{\mathbf{R}}(E_T - E_\nu)$. Then, by the postulate of equal *a priori* probability, the probability that the system will be in state ν (denoted by P_ν) is proportional to $\Gamma_{\mathbf{R}}(E_T - E_\nu)$:

$$P_\nu(E_\nu) = \frac{1}{C} \Gamma_{\mathbf{R}}(E_T - E_\nu)$$

where the proportionality constant is obtained by the normalization of P_ν ,

-21-

$$C = \sum_{\nu} \Gamma_{\mathbf{R}}(E_T - E_\nu)$$

where the sum is over all microstates accessible to the system I . Thus

$$P_\nu(E_\nu) = \frac{\Gamma_{\mathbf{R}}(E_T - E_\nu)}{\sum_{\nu} \Gamma_{\mathbf{R}}(E_T - E_\nu)}$$

which can be rewritten as

$$P_\nu(E_\nu) = \frac{\exp[\log \Gamma_{\mathbf{R}}(E_T - E_\nu)]}{\sum_{\nu} \exp[\log \Gamma_{\mathbf{R}}(E_T - E_\nu)]} \equiv \frac{\exp[S_{\mathbf{R}}(E_T - E_\nu)/k_{\mathbf{b}}]}{\sum_{\nu} \exp[S_{\mathbf{R}}(E_T - E_\nu)/k_{\mathbf{b}}]}$$

where the following definition of statistical entropy is introduced

$$S(E, V, N) \equiv k_{\mathbf{b}} \log \Gamma(E, V, N). \quad (\text{A2.2.54})$$

Now, since the reservoir is much bigger than the system I , one expects $E_T \gg E_\nu$. Thermal equilibrium between the reservoir and the system implies that their temperatures are equal. Therefore, using the identification of T in section A2.1.4, one has

$$\frac{\partial S_{\mathbf{R}}(E_{\mathbf{R}})}{\partial E_{\mathbf{R}}} = \frac{\partial S_I(E_\nu)}{\partial E_\nu} = \frac{\partial S_T}{\partial E_T} = \frac{1}{T}. \quad (\text{A2.2.55})$$

Then it is natural to use the expansion of $S_{\mathbf{R}}(E_{\mathbf{R}})$ around the maximum value of the reservoir energy, E_T :

$$S_{\mathbf{R}}(E_T - E_\nu) = S_{\mathbf{R}}(E_T) - \frac{\partial S_{\mathbf{R}}(E_T)}{\partial E_T} E_\nu + \dots$$

Using the leading terms in the expansion and the identification of the common temperature T , one obtains

$$S_{\mathbf{R}}(E_T - E_\nu) = S_{\mathbf{R}}(E_T) - E_\nu/(k_{\mathbf{b}}T)$$

from which it follows that

$$P_v(E_v) = \frac{\exp(-E_v/(k_b T))}{\sum_v \exp(-E_v/(k_b T))}. \quad (\text{A2.2.56})$$

-22-

Note that in this normalized probability, *the properties of the reservoir enter the result only through the common equilibrium temperature T* . The accuracy of the expansion used above can be checked by considering the next term, which is

$$\frac{1}{2} \frac{\partial^2 S}{\partial E^2} E_v^2.$$

Its ratio to the first term can be seen to be $(\partial T / \partial E_T) E_v / 2T$. Since E_v is proportional to the number of particles in the system N and E_T is proportional to the number of particles in the composite system $(N + N_R)$, the ratio of the second-order term to the first-order term is proportional to $N/(N + N_R)$. Since the reservoir is assumed to be much bigger than the system. (i.e. $N_R \gg N$) this ratio is negligible, and the truncation of the expansion is justified. The combination $1/(k_b T)$ occurs frequently and is denoted by β below.

The above derivation leads to the identification of the canonical ensemble density distribution. More generally, consider a system with volume V and N_A particles of type A, N_B particles of type B, etc., such that $N = N_A + N_B + \dots$, and let the system be in thermal equilibrium with a much larger heat reservoir at temperature T . Then if \mathcal{H} is the system Hamiltonian, the canonical distribution is (quantum mechanically)

$$\rho = \frac{\exp(-\beta \mathcal{H})}{\text{Tr}[\exp(-\beta \mathcal{H})]}. \quad (\text{A2.2.57})$$

The corresponding classical distribution is

$$\rho(p, q) d^{2f} \Omega = \frac{e^{-\beta \mathcal{H}(p, q)} d^{2f} \Omega}{h^f N_A! N_B! \dots Q_N} \quad (\text{A2.2.58})$$

where f is the total number of degrees of freedom for the N -particle system and

$$Q_N(\beta, V) = \frac{1}{h^f N_A! N_B! \dots} \int e^{-\beta \mathcal{H}(p, q)} d^{2f} \Omega \quad (\text{A2.2.59})$$

which, for a one-component system, reduces to

$$Q_N(\beta, V) = \frac{1}{h^f N!} \int e^{-\beta \mathcal{H}(p, q)} d^{2f} \Omega \quad (\text{A2.2.60})$$

This result is the classical analogue of

$$Q_N(\beta, V) = \sum_v \exp(-\beta E_v) \equiv \text{Tr}[\exp(-\beta \mathcal{H})]. \quad (\text{A2.2.61})$$

$Q_N(\beta, V)$ is called the canonical partition function, and plays a central role in determining the thermodynamic behaviour of the system. The constants in front of the integral in (A2.2.59) and (A2.2.60) can be understood in terms of the uncertainty principle and indistinguishability of particles, as was discussed earlier in [section A2.2.3.1](#) while obtaining (A2.2.18). Later, in [section A2.2.5.5](#), the classical limit of an ideal quantum gas is considered, which also leads to a similar understanding of these multiplicative constants, which arise on account of overcounting of microstates in classical mechanics.

The canonical distribution corresponds to the probability density for the system to be in a specific microstate with energy $E \sim \mathcal{H}$; from it one can also obtain the probability $\mathcal{P}(E)$ that the system has an energy between E and $E + dE$ if the density of states $\mathcal{D}(E)$ is known. This is because, classically,

$$[h^f N_A! N_B! \dots]^{-1} \int_{E < \mathcal{H} < E + dE} d^{2f} \Omega = \mathcal{D}(E) dE \quad (\text{A2.2.62})$$

and, quantum mechanically, the sum over the degenerate states with $E < \mathcal{H} < E + dE$ also yields the extra factor $\mathcal{D}(E) dE$. The result is

$$\mathcal{P}(E) d(E) = [Q_N]^{-1} e^{-\beta E} \mathcal{D}(E) dE. \quad (\text{A2.2.63})$$

Then, the partition function can also be rewritten, as

$$Q_N = \int e^{-\beta E} \mathcal{D}(E) dE. \quad (\text{A2.2.64})$$

A2.2.4.1 THERMODYNAMICS IN A CANONICAL ENSEMBLE

In the microcanonical ensemble, one has specified $E \sim U(S, V, N)$ and T, P and μ are among the derived quantities. In the canonical ensemble, the system is held at fixed T , and the change of a thermodynamic variable from S in a microcanonical ensemble to T in a canonical ensemble is achieved by replacing the internal energy $U(S, V, N)$ by the Helmholtz free energy $A(T, V, N) \equiv (U - TS)$. The First Law statement for dU , [equation \(A2.2.44\)](#) now leads to

$$dA = \mu dN - P dV - S dT. \quad (\text{A2.2.65})$$

If one denotes the averages over a canonical distribution by $\langle \dots \rangle$, then the relation $A = U - TS$ and $U = \langle \mathcal{H} \rangle$ leads to the statistical mechanical connection to the thermodynamic free energy A :

$$A = -k_b T \log Q_N. \quad (\text{A2.2.66})$$

To see this, note that $S = -k_b \langle \log \rho \rangle$. Thus

$$S = -k_b \langle \log(Q_N^{-1} e^{-\beta \mathcal{H}}) \rangle = k_b \log Q_N + T^{-1} \langle \mathcal{H} \rangle = k_b \log Q_N + T^{-1} U$$

which gives the result $A = U - TS = -k_b T \log Q_N$. For any canonical ensemble system, its thermodynamic properties can be found once its partition function is obtained from the system Hamiltonian. The sequence can be

$$\mathcal{H} \rightarrow Q_N \rightarrow A \rightarrow (\mu, P, S) \quad (\text{A2.2.67})$$

where the last connection is obtained from the differential relations

$$\mu = \left(\frac{\partial A}{\partial N} \right)_{V,T} \quad P = \left(\frac{\partial A}{\partial V} \right)_{N,T} \quad S = \left(\frac{\partial A}{\partial T} \right)_{N,V}. \quad (\text{A2.2.68})$$

One can trivially obtain the other thermodynamic potentials U , H and G from the above. It is also interesting to note that the internal energy U and the heat capacity $C_{V,N}$ can be obtained directly from the partition function. Since $Q_N(\beta, V) = \sum_v \exp(-\beta E_v)$, one has

$$\begin{aligned} U \equiv \langle E_v \rangle &= \frac{\sum_v E_v \exp(-\beta E_v)}{\sum_v \exp(-\beta E_v)} \\ &= - \frac{\partial}{\partial \beta} \log Q_N(\beta, V) = \frac{\partial}{\partial \beta} (\beta A). \end{aligned} \quad (\text{A2.2.69})$$

Fluctuations in energy are related to the heat capacity $C_{V,N}$, and can be obtained by twice differentiating $\log Q_N$ with respect to β , and using equation (A2.2.69):

$$\begin{aligned} \langle (E_v - \langle E \rangle)^2 \rangle &= \langle E_v^2 \rangle - \langle E \rangle^2 \\ &= - \frac{\partial U}{\partial \beta} = k_b T^2 \frac{\partial U}{\partial T} = k_b T^2 C_{V,N}. \end{aligned} \quad (\text{A2.2.70})$$

Both $\langle E \rangle$ and $C_{V,N}$ are extensive quantities and proportional to N or the system size. The root mean square fluctuation in energy is therefore proportional to $N^{\frac{1}{2}}$, and the relative fluctuation in energy is

$$\frac{\langle (E_v - \langle E \rangle)^2 \rangle^{\frac{1}{2}}}{\langle E \rangle} \sim \frac{1}{N^{\frac{1}{2}}}. \quad (\text{A2.2.71})$$

This behaviour is characteristic of thermodynamic fluctuations. This behaviour also implies the equivalence of various ensembles in the thermodynamic limit. Specifically, as $N \rightarrow \infty$ the energy fluctuations vanish, the partition of energy between the system and the reservoir becomes uniquely defined and the thermodynamic properties in microcanonical and canonical ensembles become identical.

A2.2.4.2 EXPANSION IN POWERS OF \hbar

In the relation (A2.2.66), one can use the partition function evaluated using either (A2.2.59) or (A2.2.61). The use of (A2.2.59) gives the first term in an expansion of the quantum mechanical A in powers of \hbar in the quasi-

classical limit. In this section the next non-zero term in this expansion is evaluated. For this consider the partition function (A2.2.61). The trace of $\exp(-\beta\mathcal{H})$ can be obtained using the wavefunctions of free motion of the ideal gas of N particles in volume V :

$$\psi_p = V^{-N/2} e^{(i/\hbar) \sum_j p_j q_j} \quad (\text{A2.2.72})$$

where q_j are the coordinates and $p_j = \hbar k_j$ are the corresponding momenta of the N particles, whose $3N$ degrees of freedom are labelled by the suffix j . The particles may be identical (with same mass M) or different. For identical particles, the wavefunctions above have to be made symmetrical or antisymmetrical in the corresponding $\{q_j\}$ depending on the statistics obeyed by the particles. This effect, however, leads to exponentially small correction in A and can be neglected. The other consequence of the indistinguishability of particles is in the manner of how the momentum sums are done. This produces a correction which is third order in \hbar , obtained in section A2.2.5.5, and does not affect the $O(\hbar^2)$ term that is calculated here. In each of the wavefunctions ψ , the momenta p_j are definite constants and form a dense discrete set with spacing between the neighbouring p_j proportional to V^{-1} . Thus, the summation of the matrix elements $\langle \psi_p | \exp(-\beta\mathcal{H}) | \psi_p \rangle$ with respect to all p_j can be replaced by an integration:

$$Q_N(\beta, V) = \text{Tr} \langle \psi_p | \exp(-\beta\mathcal{H}) | \psi_p \rangle \quad (\text{A2.2.73})$$

$$= \frac{1}{h^{3N} N_A! N_B! \dots} \int d^{3N} p d^{3N} q I \quad (\text{A2.2.74})$$

where

$$\chi_1 = -\frac{1}{2} i \beta^2 \sum_j \frac{p_j}{M_j} \frac{\partial U}{\partial q_j}$$

When $\beta = 0$, $I = 1$. For systems in which the Hamiltonian \mathcal{H} can be written as

$$\mathcal{H} = \sum_j \frac{p_j^2}{2M_j} + U = -\frac{1}{2} \hbar^2 \sum_j \frac{1}{M_j} \frac{\partial^2}{\partial q_j^2} + U \quad (\text{A2.2.75})$$

with $U = U(\{q_j\})$ as the potential energy of interaction between N particles, the integral I can be evaluated by considering its derivative with respect to β , (note that the operator \mathcal{H} will act on all factors to its right):

-26-

$$\frac{\partial I}{\partial \beta} = -e^{-(i/\hbar) \sum_j p_j q_j} \mathcal{H} \{ e^{(i/\hbar) \sum_j p_j q_j} I \} \quad (\text{A2.2.76})$$

$$= -E(p, q)I + \sum_j \frac{\hbar^2}{2M_j} \left(\frac{2i}{\hbar} p_j \frac{\partial I}{\partial q_j} + \frac{\partial^2 I}{\partial q_j^2} \right) \quad (\text{A2.2.77})$$

where $E(p, q) = (\sum_j p_j^2 / 2M_j) + U$ is the classical form of the energy. By using the substitution, $I = \exp(-\beta E(p, q))\chi$ and expanding $\chi = 1 + \hbar\chi_1 + \hbar^2\chi_2 + \dots$, one can obtain the quantum corrections to the classical

partition function. Since for $\beta = 0$, $I = 1$, one also has for $\beta = 0$, $\chi = 1$, and $\chi_1 = \chi_2 = 0$. With this boundary condition, one obtains the result that

$$\chi_1 = -\frac{1}{2}i\beta^2 \sum_j \frac{p_j}{M_j} \frac{\partial U}{\partial q_j}$$

and

$$\begin{aligned} \chi_2 = & -\frac{1}{8}\beta^4 \left(\sum_j \frac{p_j}{M_j} \frac{\partial U}{\partial q_j} \right)^2 + \frac{1}{6}\beta^3 \sum_j \sum_k \frac{p_j}{M_j} \frac{p_k}{M_k} \frac{\partial^2 U}{\partial q_j \partial q_k} \\ & + \frac{1}{6}\beta^3 \sum_j \frac{1}{M_j} \left(\frac{\partial U}{\partial q_j} \right)^2 - \frac{1}{4}\beta^2 \sum_j \frac{1}{M_j} \frac{\partial^2 U}{\partial q_j^2}. \end{aligned}$$

For the partition function, the contribution from χ_1 , which is the first-order correction in \hbar , vanishes identically. One obtains

$$Q_N(\beta, V) = Q_N^{\text{cl}}(1 + \hbar^2 \langle \chi_2 \rangle^{\text{cl}} + \dots) \quad (\text{A2.2.78})$$

where the superscript (cl) corresponds to the classical value, and $\langle \chi_2 \rangle^{\text{cl}}$ is the classical canonical ensemble average of χ_2 . The free energy A can then be inferred as

$$A = A^{\text{cl}} - \beta^{-1} \log(1 + \hbar^2 \langle \chi_2 \rangle^{\text{cl}} + \dots) \quad (\text{A2.2.79})$$

$$\approx A^{\text{cl}} - \beta^{-1} \hbar^2 \langle \chi_2 \rangle^{\text{cl}}. \quad (\text{A2.2.80})$$

One can formally evaluate $\langle \chi_2 \rangle^{\text{cl}}$. Since $\langle p_j p_k \rangle = M_j \beta^{-1} \delta_{jk}$, one obtains

$$\langle \chi_2 \rangle^{\text{cl}} = \frac{\beta^3}{24} \sum_j \frac{1}{M_j} \left\langle \left(\frac{\partial U}{\partial q_j} \right)^2 \right\rangle - \frac{\beta^2}{12} \sum_j \frac{1}{M_j} \left\langle \frac{\partial^2 U}{\partial q_j^2} \right\rangle. \quad (\text{A2.2.81})$$

-27-

This can be further simplified by noting that

$$\int \frac{\partial^2 U}{\partial q_j^2} e^{-\beta U} dq_j = \frac{\partial U}{\partial q_j} e^{-\beta U} + \beta \int \left(\frac{\partial U}{\partial q_j} \right)^2 e^{-\beta U} dq_j \quad (\text{A2.2.82})$$

which implies that

$$\left\langle \frac{\partial^2 U}{\partial q_j^2} \right\rangle = \beta \left\langle \left(\frac{\partial U}{\partial q_j} \right)^2 \right\rangle. \quad (\text{A2.2.83})$$

It follows that

$$\langle \chi_2 \rangle^{\text{cl}} = -\frac{\beta^3}{24} \sum_j \frac{1}{M_j} \left\langle \left(\frac{\partial U}{\partial q_j} \right)^2 \right\rangle \quad (\text{A2.2.84})$$

with the end result that

$$Q_N(\beta, V) = Q_N^{\text{cl}} \left(1 - \hbar^2 \frac{\beta^3}{24} \sum_j \frac{1}{M_j} \left\langle \left(\frac{\partial U}{\partial q_j} \right)^2 \right\rangle \right) \quad (\text{A2.2.85})$$

and

$$A = A^{\text{cl}} + \hbar^2 \frac{\beta^2}{24} \sum_j \frac{1}{M_j} \left\langle \left(\frac{\partial U}{\partial q_j} \right)^2 \right\rangle. \quad (\text{A2.2.86})$$

The leading order quantum correction to the classical free energy is always positive, is proportional to the sum of mean square forces acting on the particles and decreases with either increasing particle mass or increasing temperature. The next term in this expansion is of order \hbar^4 . This feature enables one to independently calculate the leading correction due to quantum statistics, which is $O(\hbar^3)$. The result calculated in [section A2.2.5.5](#) is

$$A_3 = \pm \frac{\pi^{\frac{3}{2}} N^2 \beta^{\frac{1}{2}} \hbar^3}{2\gamma V M^{\frac{3}{2}}} \quad (\text{A2.2.87})$$

for an ideal quantum gas of N identical particles. The upper sign is for Fermi statistics, the lower is for Bose statistics and γ is the degeneracy factor due to nuclear and electron spins.

In the following three subsections, the three examples described in [A2.2.2](#) are considered. In each case the model system is thermal equilibrium with a large reservoir at temperature $T = (k_b \beta)^{-1}$. Then the partition function for each system is evaluated and its consequences for the thermodynamic behaviour of the model system are explored.

A2.2.4.3 APPLICATION TO IDEAL SYSTEMS: TWO-STATE MODEL

Let us consider first the two-state model of non-interacting spin- $\frac{1}{2}$ particles in a magnetic field. For a system with only one such particle there are two non-degenerate energy levels with energies $\pm \mu H$, and the partition function is $Q_1 = \exp(-\beta \mu H) + \exp(\beta \mu H) = 2 \cosh(\beta \mu H)$. For N such indistinguishable spin- $\frac{1}{2}$ particles, the canonical partition function is

$$Q_N = \frac{(Q_1)^N}{N!} = \frac{2^N}{N!} (\cosh(\beta \mu H))^N$$

The internal energy is

$$U = -\frac{\partial \log Q_N}{\partial \beta} = -N\mu H \tanh(\beta\mu H).$$

If H is ∞ (very large) or T is zero, the system is in the lowest possible and a non-degenerate energy state and $U = -N\mu H$. If either H or β is zero, then $U = 0$, corresponding to an equal number of spins up and down. There is a symmetry between the positive and negative values of $\beta\mu H$, but negative β values do not correspond to thermodynamic equilibrium states. The heat capacity is

$$C_{H,N} = -\frac{1}{k_b T^2} \left(\frac{\partial U}{\partial \beta} \right)_{H,N} = Nk_b ((\beta\mu H) \operatorname{sech}(\beta\mu H))^2.$$

Figure A2.2.1 shows $C_{H,N}$, in units of Nk_b , as a function of $(\beta\mu H)$. $C_{H,N}$ is zero in the two limits of zero and infinite values of $(\beta\mu H)$, which also implies the limits of $T = \infty$ and $T = 0$. For small $(\beta\mu H)$, it approaches zero as $\sim (\beta\mu H)^2$ and for large $(\beta\mu H)$ as $(\beta\mu H)^2 \exp(-2\beta\mu H)$. It has a maximum value of $0.439Nk_b$ around $\beta\mu H = 1.2$. This behaviour is characteristic of any two-state system, and the maximum in the heat capacity is called a Schottky anomaly.

-29-

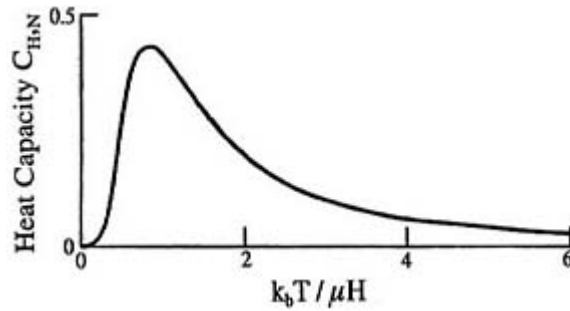


Figure A2.2.1. Heat capacity of a two-state system as a function of the dimensionless temperature, $k_b T / (\mu H)$.

From the partition function, one also finds the Helmholtz free energy as

$$\beta F = -\log Q_N = -N[\log(2) + \log(N!) + \log(\cosh(\beta\mu H))].$$

One can next obtain the entropy either from $S = (U - F)/T$ or from $S = -(\partial F / \partial T)_{V, N, H}$ and one can verify that the result is the same.

It is also instructive to start from the expression for entropy $S = k_b \log(g(N, m))$ for a specific energy partition between the two-state system and the reservoir. Using the result for $g(N, m)$ in section A2.2.2, and noting that $E = -(2\mu H)m$, one gets (using the Stirling approximation $N! \approx (2\pi N)^{1/2} N^N e^{-N}$),

$$k_b^{-1} S = -N \left[\left(\frac{1}{2} + \frac{m}{N} \right) \log \left(\frac{1}{2} + \frac{m}{N} \right) + \left(\frac{1}{2} - \frac{m}{N} \right) \log \left(\frac{1}{2} - \frac{m}{N} \right) \right].$$

Since $E = -(2\mu H)m$, a given spin excess value $2m$ implies a given energy partition. The free energy for such a specific energy partition is

$$\beta F = N \left[- (2\beta\mu H) \frac{m}{N} + \left(\frac{1}{2} + \frac{m}{N} \right) \log \left(\frac{1}{2} + \frac{m}{N} \right) + \left(\frac{1}{2} - \frac{m}{N} \right) \log \left(\frac{1}{2} - \frac{m}{N} \right) \right].$$

This has to be minimized with respect to E or equivalently m/N to obtain the thermal equilibrium result. The value of m/N that corresponds to equilibrium is found to be

$$\log \left(\frac{N + 2m}{N - 2m} \right) = 2\beta\mu H$$

which corresponds to $\langle 2m \rangle = N \tanh(\beta\mu H)$ and leads to the same U as above. It also gives the equilibrium magnetization as $\mathcal{M} = N\mu \tanh(\beta\mu H)$.

-30-

A2.2.4.4 APPLICATION TO IDEAL SYSTEMS: CLASSICAL IDEAL GAS

Consider a system of N non-interacting point particles in a three-dimensional cubical box of volume $V = L^3$. First consider one classical particle with energy $E = p^2/(2M)$. The partition function is

$$\begin{aligned} Q_1 &= h^{-3} \int_V dV \int_{-\infty}^{+\infty} dp_x \int_{-\infty}^{+\infty} dp_y \int_{-\infty}^{+\infty} dp_z e^{-(p_x^2 + p_y^2 + p_z^2)/(2Mk_b T)} \\ &= \frac{V}{h^3} \left(\int_{-\infty}^{+\infty} dp_x e^{-p_x^2/(2Mk_b T)} \right)^3 \\ &= \frac{V}{h^3} (2\pi M k_b T)^{\frac{3}{2}} = V \left(\frac{2\pi M k_b T}{h^2} \right)^{\frac{3}{2}} \equiv \frac{V}{V_q} \end{aligned} \quad (\text{A2.2.88})$$

where the definition of the quantum volume V_q associated with the thermal deBroglie wavelength, $\lambda_T \sim h/(2\pi M k_b T)^{\frac{1}{2}}$, is introduced. The same result is obtained using the density of states $\mathcal{D}(\epsilon)$ obtained for this case in section A2.2.2. Even though this $\mathcal{D}(\epsilon)$ was obtained using quantum considerations, the sum over \mathbf{n} was replaced by an integral which is an approximation that is valid when $k_b T$ is large compared to energy level spacing. This high-temperature approximation leads to a classical behaviour.

For an ideal gas of N indistinguishable point particles one has $Q_N = Q_1^N/N! = (V/V_q)^N/N!$. For large N one can again use the Stirling approximation for $N!$ and obtain the Helmholtz free energy

$$F = -k_b T \log Q_N = N k_b T \log \left(e \frac{N}{V} V_q \right) = N k_b T \log \left(e \frac{N}{V} \left(\frac{2\pi M k_b T}{h^2} \right)^{-\frac{3}{2}} \right).$$

(The term $\frac{1}{2} \log(2\pi N) k_b T$ is negligible compared to terms proportional to $N k_b T$.) The entropy obtained from the relation $S = -(\partial F/\partial T)_{N, V}$ agrees with the expression, equation (A2.2.47), obtained for the microcanonical ensemble, and one also obtains $U = F + T S = \frac{3}{2} N k_b T$ consistent with the equipartition law. The ideal equation of state $P = N k_b T/V$ is also obtained from evaluating $P = -(\partial F/\partial V)_{N, T}$. Thus one obtains the same thermodynamic behaviour from the canonical and microcanonical ensembles. This is generally the case when

N is very large since the fluctuations around the average behave as $N^{-\frac{1}{2}}$. A quantum ideal gas with either Fermi or Bose statistics is treated in [subsection A2.2.5.4](#), [subsection A2.2.5.5](#), [subsection A2.2.5.6](#) and [subsection A2.2.5.7](#).

A2.2.4.5 IDEAL GAS OF DIATOMIC MOLECULES

Consider a gas of N non-interacting diatomic molecules moving in a three-dimensional system of volume V . Classically, the motion of a diatomic molecule has six degrees of freedom—three translational degrees corresponding to the centre of mass motion, two more for the rotational motion about the centre of mass and one additional degree for the vibrational motion about the centre of mass. The equipartition law gives $\langle E_{\text{trans}} \rangle = \frac{3}{2}Nk_bT$. In a similar manner,

-31-

since the rotational Hamiltonian has rotational kinetic energy from two orthogonal angular momentum components, in directions each perpendicular to the molecular axis, equipartition gives $\langle E_{\text{rot}} \rangle = Nk_bT$. For a rigid dumb-bell model, one would then get $\langle E_{\text{total}} \rangle = \frac{5}{2}Nk_bT$, since no vibration occurs in a rigid dumb-bell. The corresponding heat capacity per mole (where $N = N_a$ is the Avogadro's number and $R = N_a k_b$ is the gas constant), is $C_v = \frac{5}{2}R$ and $C_p = \frac{7}{2}R$. If one has a vibrating dumb-bell, the additional vibrational motion has two quadratic terms in the associated Hamiltonian—one for the kinetic energy of vibration and another for the potential energy as in a harmonic oscillator. The vibrational motion thus gives an additional $\langle E_{\text{vib}} \rangle = Nk_bT$ from the equipartition law, which leads to $\langle E_{\text{total}} \rangle = \frac{7}{2}Nk_bT$ and heat capacities per mole as $C_v = \frac{7}{2}R$ and $C_p = \frac{9}{2}R$.

These results do not agree with experimental results. At room temperature, while the translational motion of diatomic molecules may be treated classically, the rotation and vibration have quantum attributes. In addition, quantum mechanically one should also consider the electronic degrees of freedom. However, typical electronic excitation energies are very large compared to k_bT (they are of the order of a few electronvolts, and 1 eV corresponds to $T \approx 10\,000$ K). Such internal degrees of freedom are considered frozen, and an electronic cloud in a diatomic molecule is assumed to be in its ground state ϵ_0 with degeneracy g_0 . The two nuclei A and B, which along with the electronic cloud make up the molecule, have spins I_A and I_B , and the associated degeneracies $(2I_A + 1)$ and $(2I_B + 1)$, respectively. If the molecule is homonuclear, A and B are indistinguishable and, by interchanging the two nuclei, but keeping all else the same, one obtains the same configuration. Thus for a homonuclear molecule, the configurations can be overcounted by a factor of two if the counting scheme used is the same as that for heteronuclear molecules. Thus, the degeneracy factor in counting the internal states of a diatomic molecule is $g = g_0(2I_A + 1)(2I_B + 1)/(1 + \delta_{AB})$ where δ_{AB} is zero for the heteronuclear case and one for the homonuclear case.

The energy of a diatomic molecule can be divided into translational and internal contributions: $\epsilon_j = (\hbar k)^2/(2M) + \epsilon_{\text{int}}$, and $\epsilon_{\text{int}} = \epsilon_0 + \epsilon_{\text{rot}} + \epsilon_{\text{vib}}$. In the canonical ensemble for an ideal gas of diatomic molecules in thermal equilibrium at temperature $T = (k_b\beta)^{-1}$ the partition function then factorizes:

$$Q_N = (N!)^{-1}[Q_{\text{trans}}]^N[Q_{\text{int}}]^N$$

where the single molecule translational partition function Q_{trans} is the same as Q_1 in [equation \(A2.2.88\)](#) and the single-molecule internal partition function is

$$Q_{\text{int}} = g e^{-\beta \epsilon_0} Q_{\text{rot}} Q_{\text{vib}}.$$

The rotational and vibrational motions of the nuclei are uncoupled, to a good approximation, on account of a mismatch in time scales, with vibrations being much faster than the rotations (electronic motions are even faster than the vibrational ones). One typically models these as a rigid rotation plus a harmonic oscillation, and obtains the energy eigenstates for such a model diatomic molecule. The resulting vibrational states are non-degenerate, are characterized by a vibrational quantum number $v = 0, 1, 2, \dots$ and with an energy $\epsilon_{\text{vib}} \equiv \epsilon_v = (\frac{1}{2} + v)\hbar\omega_0$ where ω_0 is the characteristic vibrational frequency. Thus

-32-

$$Q_{\text{vib}} = \sum_{v=0}^{\infty} e^{-\beta \hbar \omega_0 (\frac{1}{2} + v)} = [e^{\frac{1}{2}\beta \hbar \omega_0} - e^{\frac{1}{2}\beta \hbar \omega_0}]^{-1}.$$

The rotational states are characterized by a quantum number $J = 0, 1, 2, \dots$ are degenerate with degeneracy $(2J + 1)$ and have energy $\epsilon_{\text{rot}} \equiv \epsilon_J = J(J + 1)\hbar^2/(2I_0)$ where I_0 is the molecular moment of inertia. Thus

$$Q_{\text{rot}} = \sum_{J=0}^{\infty} (2J + 1) e^{-J(J+1)\theta_r/T}$$

where $\theta_r = \hbar^2/(2I_0 k_b)$. If the spacing between the rotational levels is small compared to $k_b T$, i.e. if $T \gg \theta_r$, the sum can be replaced by an integral (this is appropriate for heavy molecules and is a good approximation for molecules other than hydrogen):

$$Q_{\text{rot}} \approx \int_{J=0}^{\infty} dJ (2J + 1) e^{-J(J+1)\theta_r/T} = \frac{T}{\theta_r}$$

which is the high-temperature, or classical, limit. A better evaluation of the sum is obtained with the use of the Euler–Maclaurin formula:

$$\sum_{J=0}^{\infty} f(J) = \int_0^{\infty} dJ f(J) + \frac{1}{2}f(0) - \frac{1}{12}f'(0) + \frac{1}{720}f'''(0) - \frac{1}{30240}f^{(5)}(0) + \dots$$

Putting $f(J) = (2J + 1) \exp(-J(J + 1) \theta_r/T)$, one obtains

$$Q_{\text{rot}} \approx \frac{T}{\theta_r} + \frac{1}{3} + \frac{1}{15} \frac{\theta_r}{T} + \frac{4}{315} \left(\frac{\theta_r}{T}\right)^2 + \dots$$

If $T \ll \theta_r$, then only a first few terms in the sum need to be retained:

$$Q_{\text{rot}} \approx 1 + 3e^{-2\theta_r/T} + 5e^{-6\theta_r/T} + \dots$$

Once the partition function is evaluated, the contributions of the internal motion to thermodynamics can be evaluated. Q_{int} depends only on T , and has no effect on the pressure. Its effect on the heat capacity C_v can be obtained from the general expression $C_v = (k_b T^2)^{-1} (\partial^2 \log Q_N / \partial \beta^2)$. Since the partition function factorizes, its logarithm and, hence, heat capacity, reduces to additive contributions from translational, rotational and vibrational contributions: $C_v = C_v^{\text{trans}} + C_v^{\text{rot}} + C_v^{\text{vib}}$ where the translational motion (treated classically) yields $C_v^{\text{trans}} = \frac{3}{2} N k_b T$.

-33-

The rotational part at high temperatures gives

$$C_v^{\text{rot}} = N k_b \left(1 + \frac{1}{45} \left(\frac{\theta_r}{T} \right)^2 + \frac{16}{945} \left(\frac{\theta_r}{T} \right)^3 + \dots \right)$$

which shows that C_v^{rot} decreases at high T , reaching the classical equipartition value from above at $T = \infty$. At low temperatures,

$$C_v^{\text{rot}} \approx 12 N k_b \left(\frac{\theta_r}{T} \right)^2 e^{-2\theta_r/T}$$

so that as $T \rightarrow 0$, C_v^{rot} drops to zero exponentially. The vibrational contribution C_v^{vib} is given by

$$C_v^{\text{vib}} = N k_b \left(\frac{\theta_v}{T} \right)^2 \frac{e^{\theta_v/T}}{(e^{\theta_v/T} - 1)^2} \quad \text{where } \theta_v = \frac{\hbar \omega_0}{k_b}.$$

For $T \gg \theta_v$, C_v^{vib} is very nearly $N k_b$, the equipartition value, and for $T \ll \theta_v$, C_v^{vib} tends to zero as $(\theta_v/T)^2 \exp(-\theta_v/T)$. For most diatomic molecules θ_v is of the order of 1000 K and θ_r is less than 100 K. For HCl, $\theta_r = 15$ K; for N_2 , O_2 and NO it is between 2 and 3 K; for H_2 , D_2 and HD it is, respectively, 85, 43 and 64 K. Thus, at room temperature, the rotational contribution could be nearly $N k_b$ and the vibrational contribution could be only a few per cent of the equipartition value. [Figure A2.2.2](#) shows the temperature dependence of C_p for HD, HT and DT, various isotopes of the hydrogen molecule.

-34-

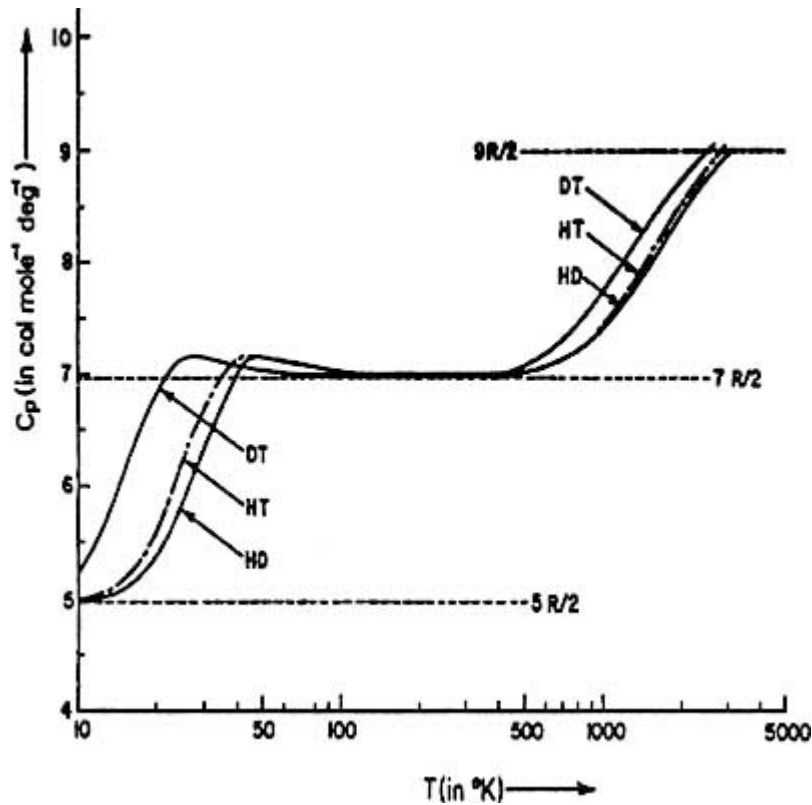


Figure A2.2.2. The rotational–vibrational specific heat, C_p , of the diatomic gases HD, HT and DT as a function of temperature. From *Statistical Mechanics* by Raj Pathria. Reprinted by permission of Butterworth Heinemann.

A2.2.4.6 APPLICATION TO IDEAL SYSTEMS: BLACK BODY RADIATION

This subsection, and the next, deals with a system of N non-interacting harmonic oscillators.

Electromagnetic radiation in thermal equilibrium within a cavity is often approximately referred to as the black-body radiation. A classical black hole is an ideal black body. Our own star, the Sun, is pretty black! A perfect black body absorbs all radiation that falls onto it. By Kirchhoff's law, which states that 'a body must emit at the same rate as it absorbs radiation if equilibrium is to be maintained', the emissivity of a black body is highest. As shown below, the use of classical statistical mechanics leads to an infinite emissivity from a black body. Planck quantized the standing wave modes of the electromagnetic radiation within a black-body cavity and solved this anomaly. He considered the distribution of energy U among N oscillators of frequency ω . If U is viewed as divisible without limit, then an infinite number of distributions are possible. Planck considered ' U as made up of an entirely determined number of finite equal parts' of value $\hbar\omega$. This quantization of the electromagnetic radiation leads to the concept of *photons* of energy quanta $\hbar\omega$, each of which having a Hamiltonian of the form of a harmonic oscillator. A state of the free electromagnetic field is specified by the number, n , for each of such oscillators and n then corresponds to the number of photons in a state with energy $\hbar\omega$. Photons obey Bose–Einstein statistics. Denote by n_j the number of photons with energy $\epsilon_j \equiv \hbar\omega_j$.

Then $n_j = 0, 1, 2, \dots$ and the canonical partition function is

$$Q = \sum_v e^{-\beta E_v} = \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \dots \sum_{n_j=0}^{\infty} \dots e^{-\beta(n_1 \epsilon_1 + n_2 \epsilon_2 + \dots + n_j \epsilon_j + \dots)}$$

Here the zero point energy is temporarily suppressed. Now the exponential is a product of independent factors. Thus one gets

$$Q = \prod_j \left(\sum_{n_j=0}^{\infty} e^{-\beta n_j \epsilon_j} \right) = \prod_j \left(\frac{1}{1 - e^{-\beta \epsilon_j}} \right)$$

on account of the geometric nature of the series being summed. One should note that photons are massless and their total number is indeterminate. Since $\log Q = -\beta A$, one can obtain various properties of the photon gas. Specifically consider the average occupation number of the j th state:

$$\begin{aligned} \langle n_j \rangle &= \frac{\sum_v n_j e^{-\beta E_v}}{\sum_v e^{-\beta E_v}} = \frac{\sum_{n_1, n_2, \dots} n_j e^{-\beta(n_1 \epsilon_1 + \dots + n_j \epsilon_j + \dots)}}{Q} \\ &= \frac{\partial \log Q}{\partial(-\beta \epsilon_j)} = \frac{\partial}{\partial(-\beta \epsilon_j)} \left\{ \sum_j [-\log(1 - e^{-\beta \epsilon_j})] \right\} \\ &= \frac{1}{e^{\beta \epsilon_j} - 1} \equiv \frac{1}{e^{\beta \hbar \omega_j} - 1}. \end{aligned} \tag{A2.2.89}$$

This is the Planck distribution function. The thermal average energy in the j th mode is (including the zero point energy)

$$\langle \epsilon_j \rangle = \frac{1}{2} \hbar \omega_j + \frac{\hbar \omega_j}{e^{\beta \hbar \omega_j} - 1}.$$

Since for small $\beta \hbar \omega_j = y$, $(\exp(y)-1)^{-1} \approx [y(1 + y/2 + \dots)]^{-1} \approx y^{-1}(1 - y/2) = (y^{-1} - 1/2)$, one obtains, when $\epsilon_j \ll k_b T$, the result for the high-temperature limit: $\langle \epsilon_j \rangle \rightarrow k_b T$. This is also the average energy for a classical harmonic oscillator with two quadratic degrees of freedom (one kinetic and one potential) in the Hamiltonian, an equipartition result. For low temperatures one has $\epsilon_j \gg k_b T$ and $\langle \epsilon_j \rangle \rightarrow (\frac{1}{2} \hbar \omega_j + \hbar \omega_j e^{-\beta \hbar \omega_j})$. The oscillator settles down in the ground state at zero temperature.

Any cavity contains an infinite number of electromagnetic modes. For radiation confined to a perfectly conducting cubical cavity of volume $V = L^3$, the modes are given by the electric field components of the form:

$$\begin{aligned}
E_x &= E_{x0} \sin \omega t \cos(n_x \pi x/L) \sin(n_y \pi x/L) \sin(n_z \pi x/L) \\
E_y &= E_{y0} \sin \omega t \sin(n_x \pi x/L) \cos(n_y \pi x/L) \sin(n_z \pi x/L) \\
E_z &= E_{z0} \sin \omega t \sin(n_x \pi x/L) \sin(n_y \pi x/L) \cos(n_z \pi x/L).
\end{aligned}$$

Within the cavity $\vec{\nabla} \cdot \vec{E} = 0$, which in Fourier space is $\vec{k} \cdot \vec{E} = 0$. Thus, only two of the three components of \vec{E} are independent. The electromagnetic field in a cavity is a transversely polarized field with two independent polarization directions, which are mutually perpendicular and are each normal to the propagation direction \vec{k} of the \vec{E} field, which satisfies the electromagnetic wave equation, $c^2 \nabla^2 \vec{E} = \partial^2 \vec{E} / \partial t^2$. Substituting the form of the \vec{E} field above, one gets

$$c^2 \pi^2 n^2 = \omega^2 L^2 \quad \text{where } n = (n_x^2 + n_y^2 + n_z^2)^{1/2}$$

so that the quantized photon modes have frequencies of the form $\omega_n = n\pi c/L$. The total energy of the photons in the cavity is then

$$Q_{\text{tot}} \approx 1 + 3e^{-2\theta_r/T} + 5e^{-6\theta_r/T} + \dots$$

Here the zero point energy is ignored, which is appropriate at reasonably large temperatures when the average occupation number is large. In such a case one can also replace the sum over \mathbf{n} by an integral. Each of the triplet (n_x, n_y, n_z) can take the values $0, 1, 2, \dots, \infty$. Thus the sum over (n_x, n_y, n_z) can be replaced by an integral over the volume element dn_x, dn_y, dn_z which is equivalent to an integral in the positive octant of the three-dimensional \mathbf{n} -space. Since there are two independent polarizations for each triplet (n_x, n_y, n_z) , one has

$$\sum_{\mathbf{n}} (\dots) = 2 \frac{1}{8} \int_0^\infty 4\pi n^2 dn (\dots).$$

Then

$$U = \pi \int_0^\infty dn n^2 \frac{\hbar \omega_n}{e^{\beta \hbar \omega_n} - 1} = V \frac{\hbar}{\pi^2 c^3} \int_0^\infty d\omega \frac{\omega^3}{e^{\beta \hbar \omega} - 1} \equiv V \int d\omega u_\omega.$$

Since $\int_0^\infty dx x^3 / (e^x - 1) = \pi^4/15$, one obtains the result for the energy per unit volume as

$$\frac{U}{V} = \frac{\pi^2 k_b^4}{15 \hbar^3 c^3} T^4. \quad (\text{A2.2.90})$$

This is known as the Stefan–Boltzmann law of radiation. If in this calculation of total energy U one uses the classical equipartition result $\langle \epsilon_n \rangle = k_b T$, one encounters the integral $\int_0^\infty d\omega \omega^2$ which is infinite. This divergence, which is the Rayleigh–Jeans result, was one of the historical results which collectively led to the inevitability of a quantum hypothesis. This divergence is also the cause of the infinite emissivity prediction for a black body according to classical mechanics.

The quantity u_ω introduced above is the spectral density defined as the energy per unit volume per unit frequency range and is

$$u_\omega = \frac{\hbar}{\pi^2 c^3} \frac{\omega^3}{e^{\beta \hbar \omega} - 1}. \quad (\text{A2.2.91})$$

This is known as the Planck radiation law. Figure A2.2.3 shows this spectral density function. The surface temperature of a hot body such as a star can be estimated by approximating it by a black body and measuring the frequency at which the maximum emission of radiant energy occurs. It can be shown that the maximum of the Planck spectral density occurs at $\hbar\omega_{\text{max}}/(k_B T) \approx 2.82$. So a measurement of ω_{max} yields an estimate of the temperature of the hot body. From the total energy U , one can also obtain the entropy of the photon gas (black-body radiation). At a constant volume, $dS = dU/T = (4\pi^2 k_b^4 V)/(15\hbar^3 c^3) T^2 dT$. This can be integrated with the result

$$S = \frac{4\pi^2 k_b^4 V}{45\hbar^3 c^3} T^3.$$

The constant of integration is zero: at zero temperature all the modes go to the unique non-degenerate ground state corresponding to the zero point energy. For this state $S \sim \log(g) = \log(1) = 0$, a confirmation of the Third Law of Thermodynamics for the photon gas.

-38-

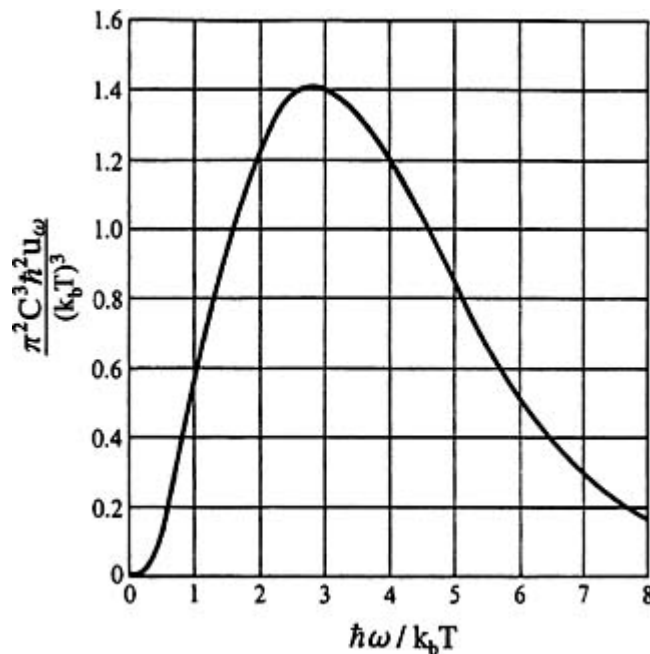


Figure A2.2.3. Planck spectral density function as a function of the dimensionless frequency $\hbar\omega/(k_b T)$.

A2.2.4.7 APPLICATION TO IDEAL SYSTEMS: ELASTIC WAVES IN A SOLID

The energy of an elastic wave in a solid is quantized just as the energy of an electromagnetic wave in a cavity.

The quanta of the elastic wave energy are called *phonons*. The thermal average number of phonons in an elastic wave of frequency ω is given, just as in the case of photons, by

$$\langle n(\omega) \rangle = (\exp(\beta\hbar\omega) - 1)^{-1}.$$

Phonons are normal modes of vibration of a low-temperature solid, where the atomic motions around the equilibrium lattice can be approximated by harmonic vibrations. The coupled atomic vibrations can be diagonalized into uncoupled normal modes (phonons) if a harmonic approximation is made. In the simplest analysis of the contribution of phonons to the average internal energy and heat capacity one makes two assumptions: (i) the frequency of an elastic wave is independent of the strain amplitude and (ii) the velocities of all elastic waves are equal and independent of the frequency, direction of propagation and the direction of polarization. These two assumptions are used below for all the modes and leads to the famous Debye model.

There are differences between photons and phonons: while the total number of photons in a cavity is infinite, the number of elastic modes in a finite solid is finite and equals $3N$ if there are N atoms in a three-dimensional solid. Furthermore, an elastic wave has three possible polarizations, two transverse and one longitudinal, in contrast to only

-39-

two transverse polarizations for photons. Thus the sum of a quantity over all phonon modes is approximated by

$$\sum_n (\dots) = \frac{3}{8} \int_0^{n_D} 4\pi n^2 dn (\dots)$$

where the maximum number n_D is obtained from the constraint that the total number of phonon modes is $3N$:

$$\frac{3}{8} \int_0^{n_D} 4\pi n^2 dn = 3N$$

which gives $n_D = (6N/\pi)^{1/3}$. Keeping in mind the differences noted above, the total thermal energy contributed by phonons can be calculated in a manner analogous to that used above for photons. In place of the velocity of light c , one has the velocity of sound v and $\omega_n = n\pi v/L$. The maximum value n_D corresponds to the highest allowed mode frequency $\omega_D = n_D\pi v/L$, and ω_D is referred to as the Debye frequency. The calculation for U then proceeds as

$$\begin{aligned} U &= \sum_n \langle \epsilon_n \rangle = \sum_n \frac{\hbar\omega_n}{e^{\beta\hbar\omega_n} - 1} \\ &= \frac{3\pi}{2} \int_0^{n_D} dn n^2 \frac{\hbar\omega_n}{e^{\beta\hbar\omega_n} - 1} \\ &= \frac{3V}{2\pi^2 v^3} \int_0^{\omega_D} d\omega \omega^2 \frac{\hbar\omega}{e^{\beta\hbar\omega} - 1} = \frac{3V}{2\pi^2 v^3 \hbar^3 \beta^4} \int_0^{x_D} dx \frac{x^3}{e^x - 1}. \end{aligned}$$

The upper limit of the dimensionless variable x_D is typically written in terms of the Debye temperature θ_D as $x_D = \theta_D/T$, where using $x_D = \beta\hbar\omega_D = \beta\hbar\pi v n_D/L = \beta\hbar v(6\pi^2 N/V)^{1/3}$, one identifies the Debye temperature as

$$\theta_D = (\hbar v/k_b)(6\pi^2 N/V)^{1/3}. \quad (\text{A2.2.92})$$

Since $\omega_D^3 = 6\pi^2 v^3 N/V$, one can also write

$$U = \int_0^\infty d\omega g(\omega) \frac{\hbar\omega}{e^{\beta\hbar\omega} - 1} \quad (\text{A2.2.93})$$

where $g(\omega)d\omega$ is the number of phonon states with a frequency between ω and $\omega + d\omega$, and is given by

$$g(\omega) = \begin{cases} \frac{9N}{\omega_D^3} \omega^2 & \text{if } \omega < \omega_D \\ 0 & \text{if } \omega > \omega_D. \end{cases} \quad (\text{A2.2.94})$$

-40-

$g(\omega)$ is essentially the density of states and the above expression corresponds to the Debye model.

In general, the phonon density of states $g(\omega)$, $d\omega$ is a complicated function which can be directly measured from experiments, or can be computed from the results from computer simulations of a crystal. The explicit analytic expression of $g(\omega)$ for the Debye model is a consequence of the two assumptions that were made above for the frequency and velocity of the elastic waves. An even simpler assumption about $g(\omega)$ leads to the Einstein model, which first showed how quantum effects lead to deviations from the classical equipartition result as seen experimentally. In the Einstein model, one assumes that only one level at frequency ω_E is appreciably populated by phonons so that $g(\omega) = \delta(\omega - \omega_E)$ and $U = (\hbar\omega_E)/(e^{\beta\hbar\omega_E} - 1)$, for each of the Einstein modes. $\hbar\omega_E/k_b$ is called the Einstein temperature θ_E .

High-temperature behaviour. Consider T much higher than a characteristic temperature like θ_D or θ_E . Since $\beta\hbar\omega$ is then small compared to 1, one can expand the exponential to obtain

$$\frac{\hbar\omega}{e^{\beta\hbar\omega} - 1} \approx \frac{1}{\beta}$$

and

$$U = \sum_n \frac{\hbar\omega_n}{e^{\beta\hbar\omega_n} - 1} \approx k_b T \sum_n 1 = 3Nk_b T \quad (\text{A2.2.95})$$

as expected by the equipartition law. This leads to a value of $3Nk_b$ for the heat capacity C_V . This is known as the Dulong and Petit's law.

Low-temperature behaviour. In the Debye model, when $T \ll \theta_D$, the upper limit, x_D , can be approximately replaced by ∞ , the integral over x then has a value $\pi^4/15$ and the total phonon energy reduces to

$$U(T) \approx \frac{3V}{2\pi^2 v^3 \hbar^3 \beta^4} \frac{\pi^4}{15} = \frac{3\pi^3 N k_b}{5\theta_D^3} T^4$$

proportional to T^4 . This leads to the heat capacity, for $T \ll \theta_D$,

$$C_V = \left(\frac{\partial U}{\partial T} \right)_{V,N} = \frac{12\pi^4 N k_b}{5\theta_D^3} T^3 \equiv A_{ph} T^3. \quad (\text{A2.2.96})$$

This result is called the Debye T^3 law. Figure A2.2.4 compares the experimental and Debye model values for the heat capacity C_p . It also gives Debye temperatures for various solids. One can also evaluate C_V for the Einstein model: as expected it approaches the equipartition result at high temperatures but decays exponentially to zero as T goes to zero.

The Debye model is more appropriate for the acoustic branches of the elastic modes of a harmonic solid. For molecular solids one has in addition optical branches in the elastic wave dispersion, and the Einstein model is more appropriate to describe the contribution to U and C_V from the optical branch. The above discussion for phonons is suitable for non-metallic solids. In metals, one has, in addition, the contribution from the electronic motion to U and C_V . This is discussed later, in section (A2.2.5.6).

Debye Temperatures (K)

<i>Substance</i>	Pb	Tl	Hg	I	Cd	Na	KBr	Ag	Ca
	88	96	97	106	168	172	177	215	226
<i>Substance</i>	KCl	Zn	NaCl	Cu	Al	Fe	CaF	FeS	C
	230	235	281	315	398	453	474	645	1860

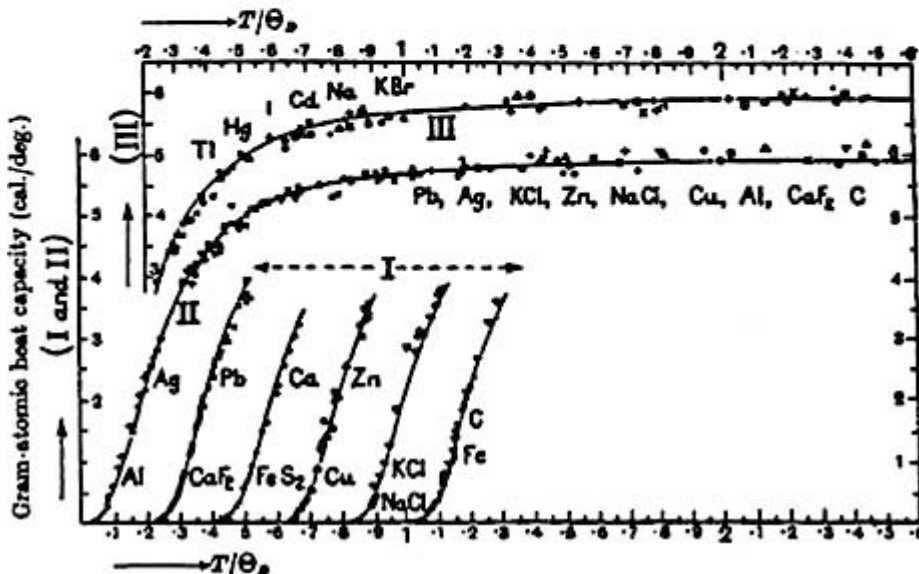


Figure A2.2.4. Experimental and Debye values for the heat capacity C_p . From Born and Huang [1].

A2.2.5 GRAND CANONICAL ENSEMBLE

Now consider two systems that are in thermal and diffusive contact, such that there can be sharing of both energy and particles between the two. Again let I be the system and II be a much larger reservoir. Since the composite system is isolated, one has the situation in which the volume of each of the two are fixed at V' and V'' , respectively, and the total energy and total number of particles are shared: $E_l = E_l^I + E_l^{II}$ where $l = (I, II)$ and $N = N' + N''$. We shall use the

-42-

notation $E = E' + E''$ for the former of these two constraints. For a given partition the allowed microstates of the system I is given by $\Gamma_I(E', N')$ and that for the system II by $\Gamma_{II}(E'', N'') \equiv \Gamma_{II}(E - E', N - N')$. Then the total number of allowed microstates for the composite system, subject to the two constraints, is

$$\Gamma_C(E, N) = \sum_{N'} \sum_{E'} \Gamma_I(E', N') \Gamma_{II}(E - E', N - N').$$

Among all possible partitions in the above expression, the equilibrium partition corresponds to the most probable partition, for which $d\Gamma_C = 0$. Evaluating this differential yields the following relation:

$$0 = \frac{d\Gamma_C}{\Gamma_I \Gamma_{II}} = \left(\frac{1}{\Gamma_I} \frac{\partial \Gamma_I}{\partial E'} - \frac{1}{\Gamma_{II}} \frac{\partial \Gamma_{II}}{\partial E''} \right) dE' + \left(\frac{1}{\Gamma_I} \frac{\partial \Gamma_I}{\partial N'} - \frac{1}{\Gamma_{II}} \frac{\partial \Gamma_{II}}{\partial N''} \right) dN'.$$

Since E' and N' are independent variables, their variations are arbitrary. Hence, for the above equality to be satisfied, each of the two bracketed expressions must vanish when the (E, N) partition is most probable. The vanishing of the coefficient of dE' implies the equality of temperatures of I and II, consistent with thermal equilibrium:

$$\beta_I \equiv \frac{\partial \log \Gamma_I}{\partial E'} = \frac{\partial \log \Gamma_{II}}{\partial E''} \equiv \beta_{II}. \quad (\text{A2.2.97})$$

The result that the coefficient of dN' is zero for the most probable partition is the consequence of the chemical equilibrium between the system and the reservoir. It leads us to identify the chemical potential μ as

$$\frac{\partial \log \Gamma}{\partial N} \equiv -\beta \mu \quad (\text{A2.2.98})$$

in analogy to the thermodynamic definition. Then, since $\beta_I = \beta_{II}$, the vanishing of the coefficient of dN' leads to the equality of chemical potentials: $\mu_I = \mu_{II}$. In a manner similar to that used to obtain the canonical distribution, one can expand

$$\begin{aligned}
\Gamma_{\text{II}}(E - E', N - N') &= \exp[S_{\text{II}}(E - E', N - N')/k_{\text{b}}] \\
&= \Gamma_{\text{II}}(E, N) \exp\left(-\frac{1}{k_{\text{b}}}\left(E' \frac{\partial S}{\partial E} + N' \frac{\partial S}{\partial N}\right)\right) \\
&\propto \exp[-\beta(E' - \mu N')].
\end{aligned}$$

With this result and arguments similar to those used in the last section, one finds the grand canonical ensemble distribution as (quantum mechanically)

-43-

$$\rho = \frac{\exp(-\beta[\mathcal{H} - \mu N])}{\sum_{N=0}^{\infty} \text{Tr}[\exp(-\beta[\mathcal{H} - \mu N])]} \quad (\text{A2.2.99})$$

The corresponding classical distribution is

$$\rho(p, q; N) d^{2f} \Omega = \frac{e^{-\beta[\mathcal{H}(p, q; N) - \mu N]} d^{2f} \Omega}{h^f N! \Xi} \quad (\text{A2.2.100})$$

where f is the total number of degrees of freedom if the system has N particles, and the grand partition function $\Xi(\beta, \mu, V)$ is given by

$$\Xi(\beta, \mu, V) = \sum_{N=0}^{\infty} \frac{1}{h^f N!} \int e^{-\beta[\mathcal{H}(p, q) - \mu N]} d^{2f} \Omega \quad (\text{A2.2.101})$$

which is the classical analogue of

$$\sum_{N=0}^{\infty} \sum_{\nu} \exp(-\beta[E_{\nu} - \mu N]) \equiv \sum_{N=0}^{\infty} \text{Tr}[\exp(-\beta[\mathcal{H} - \mu N])]. \quad (\text{A2.2.102})$$

In the above, the sum over N has the upper limit of infinity. This is clearly correct in the thermodynamic limit. However, for a system with finite volume, V , depending on the 'hard core size' of its constituents, there will be a maximum number of particles, $M(V)$, that can be packed in volume V . Then, for all N such that $N > M(V)$, the value of $(-\beta\mathcal{H})$ becomes infinity and all terms in the N sum with $N > M(V)$ vanish. Thus, provided the inter-particle interactions contain a strongly repulsive part, the N sum in the above discussion can be extended to infinity.

If, in this ensemble, one wants to find only the probability that the system has N particles, one sums the distribution over the energy microstates to obtain:

$$\mathcal{P}(N) = \frac{e^{\beta\mu N} Q_N(\beta, V)}{\Xi(\beta, \mu, V)}. \quad (\text{A2.2.103})$$

The combination $e^{\beta\mu}$ occurs frequently. It is called the fugacity and is denoted by z . The grand canonical ensemble is also known as $T - \mu$ ensemble.

A2.2.5.1 T-P ENSEMBLE

In many experiments the sample is in thermodynamic equilibrium, held at constant temperature and pressure, and various properties are measured. For such experiments, the $T-P$ ensemble is the appropriate description. In this case the system has fixed N and shares energy and volume with the reservoir: $E = E' + E''$ and $V = V' + V''$, i.e. the system

-44-

and the reservoir are connected by a pressure transmitting movable diathermic membrane which enables the sharing of the energy and the volume. The most probable partition leads to the conditions for thermal (equality of temperatures) and mechanical (equality of pressures) equilibria. The later condition is obtained after identifying pressure P as

$$\frac{\partial \log \Gamma}{\partial V} \equiv \beta P. \quad (\text{A2.2.104})$$

The $T-P$ ensemble distribution is obtained in a manner similar to the grand canonical distribution as (quantum mechanically)

$$\rho = \frac{\exp(-\beta[H + PV])}{\int_0^\infty dV \text{Tr}[\exp(-\beta[H + PV])]} \quad (\text{A2.2.105})$$

and classically as

$$\rho(p, q; V) d^{2f} \Omega = \frac{e^{-\beta[\mathcal{H}(p, q; V) + PV]} d^{2f} \Omega}{h^f N! Y} \quad (\text{A2.2.106})$$

where the $T-P$ partition function $Y(T, P, N)$ is given by

$$Y(T, P, N) = \frac{1}{h^f N!} \int_0^\infty dV \int e^{-\beta[\mathcal{H}(p, q; V) + PV]} d^{2f} \Omega. \quad (\text{A2.2.107})$$

Its quantum mechanical analogue is

$$Y(T, P, N) = \int_0^\infty dV \sum_v \exp(-\beta[E_v(V) + PV]) \quad (\text{A2.2.108})$$

$$\equiv \int_0^\infty dV \text{Tr}[\exp(-\beta[\mathcal{H} + PV])]. \quad (\text{A2.2.109})$$

The $T-P$ partition function can also be written in terms of the canonical partition function Q_N as:

$$Y(T, P, N) = \int_0^\infty Q_N(\beta, V) e^{-\beta PV} dV \quad (\text{A2.2.110})$$

and the probability that the system will have a volume between V and $V + dV$ is given by

-45-

$$\mathcal{P}(V) d(V) = Q_N(\beta, V) \frac{e^{-\beta PV} dV}{Y(T, P, N)}. \quad (\text{A2.2.111})$$

From the canonical ensemble where (V, T, N) are held fixed, one needs to change V to P as an independent variable in order to obtain the $T-P$ ensemble where (P, T, N) are fixed. This change is done through a Legendre transform, $G = A + PV$, which replaces the Helmholtz free energy by the Gibbs free energy as the relevant thermodynamic potential for the $T-P$ ensemble. Now, the internal energy U and its natural independent variables S, V , and N are all extensive quantities, so that for an arbitrary constant a ,

$$U(aS, aV, aN) = aU(S, V, N). \quad (\text{A2.2.112})$$

Differentiating both sides with respect to a and using the differential form of the First Law, $dU = T dS - P dV + \mu dN$, one obtains the *Gibbs–Duhem equation*:

$$U = TS - PV + \mu N \quad (\text{A2.2.113})$$

which implies that $G = A + PV = U - TS + PV = \mu N$. The connection to thermodynamics in the $T-P$ ensemble is made by the identification

$$G(T, P, N) = -k_b T \log Y(T, P, N). \quad (\text{A2.2.114})$$

The average value and root mean square fluctuations in volume V of the $T-P$ ensemble system can be computed from the partition function $Y(T, P, N)$:

$$\langle V \rangle = - \left(\frac{\partial \log Y}{\partial (\beta P)} \right)_{T, N} \quad (\text{A2.2.115})$$

$$\langle V^2 \rangle - \langle V \rangle^2 = - \frac{1}{\beta} \left(\frac{\partial \langle V \rangle}{\partial P} \right)_{T, N}. \quad (\text{A2.2.116})$$

The entropy S can be obtained from

$$S = - \left(\frac{\partial G}{\partial T} \right)_{P, N} = k_b \left(\log Y - \beta \frac{\partial}{\partial \beta} \log Y \right). \quad (\text{A2.2.117})$$

A2.2.5.2 THERMODYNAMICS IN A GRAND CANONICAL ENSEMBLE

In a canonical ensemble, the system is held at fixed (V, T, N) . In a grand canonical ensemble the (V, T, μ) of the system are fixed. The change from N to μ as an independent variable is made by a Legendre transformation in which the dependent variable A , the Helmholtz free energy, is replaced by the grand potential

$$\Omega_G = A - \mu N = U - TS - \mu N = -PV. \quad (\text{A2.2.118})$$

Therefore, from the differential relation, equation (A2.2.65), one obtains,

$$d\Omega_G = -S dT - P dV - N d\mu \quad (\text{A2.2.119})$$

which implies

$$N = -\left(\frac{\partial \Omega_G}{\partial \mu}\right)_{V,T} \quad P = -\left(\frac{\partial \Omega_G}{\partial V}\right)_{\mu,T} \quad S = -\left(\frac{\partial \Omega_G}{\partial T}\right)_{\mu,V}. \quad (\text{A2.2.120})$$

Using equation (A2.2.101) and equation (A2.2.60), one has

$$\Xi(\beta, \mu, V) = \sum_{N=0}^{\infty} e^{\beta \mu N} Q_N(\beta, V) = \sum_{N=0}^{\infty} e^{\beta(\mu N + k_b T \log Q_N)}. \quad (\text{A2.2.121})$$

Using this expression for Ξ and the relation $A = -k_b T \log Q_N$, one can show that the average of $(\mu N - A)$ in the grand canonical ensemble is

$$\langle \mu N - A \rangle = \frac{\partial}{\partial \beta} (\log \Xi(\beta, \mu, V)). \quad (\text{A2.2.122})$$

The connection between the grand canonical ensemble and thermodynamics of fixed (V, T, μ) systems is provided by the identification

$$\log \Xi(\beta, \mu, V) = -\beta \Omega_G = \beta P V. \quad (\text{A2.2.123})$$

Then one has

$$k_b T \log \Xi(\beta, \mu, V) = \mu \langle N \rangle - \langle A \rangle. \quad (\text{A2.2.124})$$

In the grand canonical ensemble, the number of particles fluctuates. By differentiating $\log \Xi$, equation (A2.2.121) with respect to $\beta \mu$ at fixed V and β , one obtains

$$\langle N \rangle = \frac{1}{\beta} \frac{\partial \log \Xi}{\partial \mu} \quad (\text{A2.2.125})$$

and

$$\langle (N - \langle N \rangle)^2 \rangle = \langle N^2 \rangle - \langle N \rangle^2 = \frac{1}{\beta^2} \frac{\partial^2 \log \Xi}{\partial \mu^2} = \frac{1}{\beta} \frac{\partial \langle N \rangle}{\partial \mu}. \quad (\text{A2.2.126})$$

Since $\partial\langle N \rangle / \partial\mu \sim \langle N \rangle$, the fractional root mean square fluctuation in N is

$$\frac{\langle (N - \langle N \rangle)^2 \rangle^{\frac{1}{2}}}{\langle N \rangle} \sim \frac{1}{N^{\frac{1}{2}}}. \quad (\text{A2.2.127})$$

There are two further useful results related to $\langle (N - \langle N \rangle)^2 \rangle$. First is its connection to the isothermal compressibility $\kappa_T = -V^{-1} \partial P / \partial V_{\langle N \rangle, T}$, and the second to the spatial correlations of density fluctuations in a grand canonical system.

Now since $\Omega_G = -PV$, the Gibbs–Duhem equation gives $d\Omega_G = -S, dT - P, dV - \langle N \rangle d\mu = -P dV - V dp$, which implies that $d\mu = (V dp - S dT) / \langle N \rangle$. Let $v = V / \langle N \rangle$ be the specific volume, and express μ as $\mu(v, T)$. Then the result for $d\mu$ gives

$$\left(\frac{\partial \mu}{\partial v} \right)_T = v \left(\frac{\partial P}{\partial v} \right)_T.$$

Now a change in v can occur either through V or $\langle N \rangle$:

$$\left(\frac{\partial}{\partial v} \right)_{v, T} = -\frac{\langle N \rangle}{v} \left(\frac{\partial}{\partial \langle N \rangle} \right)_{v, T}$$

$$\left(\frac{\partial}{\partial v} \right)_{\langle N \rangle, T} = \langle N \rangle \left(\frac{\partial}{\partial V} \right)_{\langle N \rangle, T}$$

$$\frac{\langle N \rangle}{v} \left(\frac{\partial \mu}{\partial \langle N \rangle} \right)_{v, T} = -V \left(\frac{\partial P}{\partial V} \right)_{\langle N \rangle, T}$$

These two should lead to an equivalent change in v . Thus one obtains

$$\frac{\langle N \rangle}{v} \left(\frac{\partial \mu}{\partial \langle N \rangle} \right)_{v, T} = -V \left(\frac{\partial P}{\partial V} \right)_{\langle N \rangle, T}$$

-48-

the substitution of which yields, for the mean square number fluctuations, the result

$$\frac{\langle (N - \langle N \rangle)^2 \rangle}{\langle N \rangle} = \frac{\kappa_T}{\beta v}. \quad (\text{A2.2.128})$$

For homogeneous systems, the average number density is $n_0 = \langle N \rangle / V \equiv v^{-1}$. Let us define a local number density through

$$n(\vec{r}) = \sum_{i=1}^N \delta(\vec{r} - \vec{r}_i) \quad (\text{A2.2.129})$$

where \vec{r} is a point within the volume V of the grand ensemble system in which, at a given instant, there are N particles whose positions are given by the vectors \vec{r}_i , $i = 1, 2, \dots, N$. One has $N = \int dV n(\vec{r})$ and, for homogeneous systems, $\langle N \rangle = \int dV \langle n(\vec{r}) \rangle = \int dV n_0 = V n_0$. One can then define the fluctuations in the local number density as $\delta n = n - n_0$, and construct the spatial density–density correlation function as

$$G(\vec{r} - \vec{r}') \equiv n_0^{-2} \langle \delta n(\vec{r}) \delta n(\vec{r}') \rangle. \quad (\text{A2.2.130})$$

$G(\vec{r})$ is also called the pair correlation function and is sometimes denoted by $h(\vec{r})$. Integration over \vec{r} and \vec{r}' through the domain of system volume gives, on the one hand,

$$\int_V d\vec{r}' \int_V d\vec{r} G(\vec{r} - \vec{r}') = V \int d\vec{r} G(\vec{r})$$

and, on the other,

$$\begin{aligned} \int_V d\vec{r}' \int_V d\vec{r} G(\vec{r} - \vec{r}') &= n_0^{-2} \int_V d\vec{r}' \int_V d\vec{r} [\langle n(\vec{r}) n(\vec{r}') \rangle - n_0^2] \\ &= n_0^{-2} (\langle N^2 \rangle - \langle N \rangle^2) = \langle N \rangle n_0^{-1} \frac{\kappa_T}{\beta}. \end{aligned}$$

Comparing the two results and substituting the relation of the mean square number fluctuations to isothermal compressibility, equation (A2.2.128) one has

$$\int_V d\vec{r} G(\vec{r}) = k_b T \kappa_T. \quad (\text{A2.2.131})$$

-49-

The correlation function $G(\vec{r})$ quantifies the density fluctuations in a fluid. Characteristically, density fluctuations scatter light (or any radiation, like neutrons, with which they can couple). Then, if a radiation of wavelength λ is incident on the fluid, the intensity of radiation scattered through an angle θ is proportional to the structure factor

$$S(\vec{q}) = n_0 \int_V d\vec{r} e^{-i\vec{q} \cdot \vec{r}} G(\vec{r}) \quad (\text{A2.2.132})$$

where $|\vec{q}| = 4\pi \sin(\theta/2)/\lambda$. The limiting value of $S(\vec{q})$ as $q \rightarrow 0$ is then proportional to κ_T . Near the critical point of a fluid, anomalous density fluctuations create a divergence of κ_T which is the cause of the phenomenon of critical opalescence: density fluctuations become correlated over a lengthscale which is long compared to a molecular lengthscale and comparable to the wavelength of the incident light. This causes the light to be strongly scattered, whereby multiple scattering becomes dominant, making the fluid medium appear turbid or opaque.

For systems in which the constituent particles interact via short-range pair potentials, $W = \sum_{i=1}^N \sum_{j=1}^{(i-1)} u(|\vec{r}_i - \vec{r}_j|)$, there are two relations, that one can prove by evaluating the average of the total energy $E = K + W$, where K is the total kinetic energy, and the average pressure P , that are valid in general. These are

$$\frac{\langle E \rangle}{\langle N \rangle} = \frac{3}{2} k_b T + \frac{1}{2} n_o \int_V d\vec{r} g(r) u(r) \quad (\text{A2.2.133})$$

and the virial equation of state,

$$P = n_o k_b T \left(1 - \frac{n_o}{6 k_b T} \int_V d\vec{r} g(r) r \frac{du(r)}{dr} \right). \quad (\text{A2.2.134})$$

Here $g(r) = G(r) + 1$ is called a radial distribution function, since $n_o g(r)$ is the conditional probability that a particle will be found at \vec{r} if there is another at the origin. For strongly interacting systems, one can also introduce the potential of the mean force $w(r)$ through the relation $g(r) = \exp(-\beta w(r))$. Both $g(r)$ and $w(r)$ are also functions of temperature T and density n_o .

A2.2.5.3 DENSITY EXPANSION

For an imperfect gas, i.e. a low-density gas in which the particles are, most of the time, freely moving as in an ideal gas and only occasionally having binary collisions, the potential of the mean force is the same as the pair potential $u(r)$. Then, $g(r) \approx \exp(-\beta u(r)) [1 + O(n_o)]$, and from equation (A2.2.133) the change from the ideal gas energy, $\Delta U = \langle E \rangle - \langle E \rangle_{\text{ideal}}$, to leading order in n_o , is

-50-

$$\begin{aligned} \frac{\Delta U}{N} &\approx \frac{1}{2} n_o \int_V d\vec{r} u(r) e^{-\beta u(r)} = -\frac{1}{2} n_o \frac{\partial}{\partial \beta} \int_V d\vec{r} [e^{-\beta u(r)} - 1] \\ &= -\frac{1}{2} n_o \frac{\partial}{\partial \beta} \int_V d\vec{r} f(r) \end{aligned} \quad (\text{A2.2.135})$$

where

$$f(r) = e^{-\beta u(r)} - 1. \quad (\text{A2.2.136})$$

Figure A2.2.5 shows a sketch of $f(r)$ for Lennard-Jones pair potential. Now if ΔA is the excess Helmholtz free energy relative to its ideal gas value, then $(-\beta \Delta A) = \log(Q/Q_{\text{ideal}})$ and $\Delta U/N = [\partial(\beta \Delta A/N)/(\partial \beta)]$. Then, integrating with respect to β , one obtains

$$-\beta \Delta A/N = \frac{1}{2} n_o \int_V d\vec{r} f(r) + O(n_o^2). \quad (\text{A2.2.137})$$

One can next obtain pressure P from the above by

$$\beta P = n_0 + n_0^2 \frac{\partial(\beta \Delta A/N)}{\partial n_0} = n_0 + n_0^2 B_2(T) + O(n_0^3) \quad (\text{A2.2.138})$$

where

$$B_2(T) = -\frac{1}{2} \int d\vec{r} f(r). \quad (\text{A2.2.139})$$

The same result can also be obtained directly from the virial equation of state given above and the low-density form of $g(r)$. $B_2(T)$ is called the second virial coefficient and the expansion of P in powers of n_0 is known as the virial expansion, of which the leading non-ideal term is deduced above. The higher-order terms in the virial expansion for P and in the density expansion of $g(r)$ can be obtained using the methods of cluster expansion and cumulant expansion.

-51-

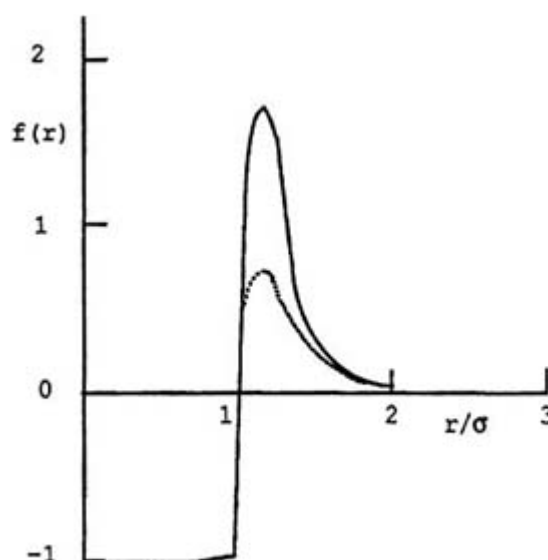


Figure A2.2.5. Sketch of $f(r)$ for the Lennard-Jones pair potential $u(r) = 4\epsilon[(\sigma/r)^{12} - (\sigma/r)^6]$; full curve $-\beta\epsilon = 1.0$ and broken curve $-\beta\epsilon = 0.5$. From Plischke and Bergersen 1985, further reading.

For purely repulsive potentials ($u(r) > 0$), $f(r)$ is negative and $B_2(T)$ is positive. For purely attractive potentials, on the other hand, $f(r)$ is always positive leading to a negative $B_2(T)$. Realistic interatomic potentials contain both a short-range repulsive potential (due to the strong short distance overlap of electronic wavefunctions of the two atoms) and a weaker longer range van der Waals attractive potential. The temperature dependence of $B_2(T)$ can be used to qualitatively probe the nature of interatomic potential. At a certain temperature T_B , known as the Boyle temperature, the effects of attractive and repulsive potentials balance exactly, giving $B_2(T_B) = 0$. A phenomenological extension of the ideal gas equation of state was made by van der Waals more than 100 years ago. For one mole of gas,

$$pV = RT \implies \left(P + \frac{a}{v^2} \right) (v - b) = RT. \quad (\text{A2.2.140})$$

Here b corresponds to the repulsive part of the potential, which is equivalent to the excluded volume due to the finite atomic size, and a/v^2 corresponds to the attractive part of the potential. The van der Waals equation

of state is a very good qualitative description of liquids as well as imperfect gases. Historically, it is the first example of a mean field theory. It fails only in the neighbourhood of a critical point due to its improper treatment of the density fluctuations.

A2.2.5.4 IDEAL QUANTUM GASES

Thermodynamics of ideal quantum gases is typically obtained using a grand canonical ensemble. In principle this can also be done using a canonical ensemble partition function, $Q = \sum_v \exp(-\beta E_v)$. For the photon and phonon gases, the canonical ensemble was used in [section A2.2.4.6](#) and [section A2.2.4.7](#). Photons and phonons are massless and their total number indeterminate, since they can be created or destroyed, provided the momentum and energy are conserved in the process. On the other hand, for an ideal gas consisting of particles with non-zero mass, in a canonical ensemble,

-52-

the total number of particles is fixed at N . Thus, in the occupation number representation of the single-particle states j , the sum of all n_j is constrained to be N :

$$Q_N = \sum_{n_1, n_2, \dots, n_j, \dots} \delta\left(N - \sum_j n_j\right) \exp\left(-\beta \sum_j n_j \epsilon_j\right)$$

where ϵ_j is the energy of the j th single-particle state. The restriction on the sum over n_j creates a complicated combinatorial problem, which even though solvable, is non-trivial. This constraint is removed by considering the grand canonical partition function:

$$\begin{aligned} \Xi(\beta, \mu, V) &= \sum_{N=0}^{\infty} e^{\beta\mu N} Q_N(\beta, V) \\ &= \sum_{N=0}^{\infty} e^{\beta\mu N} \sum_{n_1, n_2, \dots, n_j, \dots} \delta\left(N - \sum_j n_j\right) \exp\left(-\beta \sum_j n_j \epsilon_j\right) \\ &= \sum_{n_1, n_2, \dots, n_j, \dots} \exp\left(-\beta \sum_j (\epsilon_j - \mu) n_j\right). \end{aligned} \quad (\text{A2.2.141})$$

Now the exponential factors for various n_j within the sum are independent, which simplifies the result as

The sum over n_j can now be performed, but this depends on the statistics that the particles in the ideal gas obey. Fermi particles obey the Pauli exclusion principle, which allows only two possible values: $n_j = 0, 1$. For Bose particles, n_j can be any integer between zero and infinity. Thus the grand partition function is

$$\Xi = \prod_j [1 + e^{-\beta(\epsilon_j - \mu)}] \quad \text{for fermions} \quad (\text{A2.2.142})$$

and

$$\Xi = \prod_j [1 - e^{-\beta(\epsilon_j - \mu)}]^{-1} \quad \text{for bosons.} \quad (\text{A2.2.143})$$

This leads to, using equations (A2.2.123),

$$\beta PV = -\beta\Omega_G = \log \Xi = \pm \sum_j \log[1 \pm e^{-\beta(\epsilon_j - \mu)}] \quad (\text{A2.2.144})$$

-53-

where the upper sign corresponds to fermions and the lower sign to bosons. From equation (A2.2.141), the average occupation number $\langle n_j \rangle = \partial(\log \Xi)/\partial(\beta\mu)$. From this one obtains

$$\langle n_j \rangle = [e^{\beta(\epsilon_j - \mu)} \pm 1]^{-1} \quad (\text{A2.2.145})$$

where again the upper sign corresponds to fermions and the lower sign to bosons. From this, one has, for the total number of particles, $\langle N \rangle$,

$$\langle N \rangle = \sum_j \langle n_j \rangle = \sum_j [e^{\beta(\epsilon_j - \mu)} \pm 1]^{-1} \quad (\text{A2.2.146})$$

and for the total internal energy $U \equiv \langle E \rangle$

$$U = \sum_j \epsilon_j \langle n_j \rangle = \sum_j \epsilon_j [e^{\beta(\epsilon_j - \mu)} \pm 1]^{-1}. \quad (\text{A2.2.147})$$

When the single-particle states j are densely packed within any energy interval of $k_b T$, the sum over j can be replaced by an integral over energy such that

$$\sum_j \dots \rightarrow \int_0^\infty d\epsilon \mathcal{D}(\epsilon) \dots = \frac{\gamma V}{4\pi^2} \left(\frac{2M}{\hbar^2} \right)^{\frac{3}{2}} \int_0^\infty d\epsilon \epsilon^{\frac{1}{2}} \dots \quad (\text{A2.2.148})$$

Using equation (A2.2.88), this can be rewritten as

$$\sum_j \dots \rightarrow \frac{2\gamma V}{\pi^{\frac{1}{2}} V_q} \beta^{\frac{3}{2}} \int_0^\infty d\epsilon \epsilon^{\frac{1}{2}} \dots \quad (\text{A2.2.149})$$

Using this approximation, expressions for $\langle N \rangle$, U and P reduce to

$$\begin{aligned} n_o \equiv \frac{\langle N \rangle}{V} &= \frac{2\gamma}{\pi^{\frac{1}{2}} V_q} \beta^{\frac{3}{2}} \int_0^\infty d\epsilon \epsilon^{\frac{1}{2}} [e^{\beta(\epsilon_j - \mu)} \pm 1]^{-1} \\ &= \frac{2\gamma}{\pi^{\frac{1}{2}} V_q} \int_0^\infty dy y^{\frac{1}{2}} [e^{y - \beta\mu} \pm 1]^{-1} \end{aligned} \quad (\text{A2.2.150})$$

$$(\text{A2.2.151})$$

$$\equiv \frac{2\gamma}{\pi^{\frac{1}{2}} V_q} F_{\frac{3}{2}}(\beta\mu)$$

-54-

$$\begin{aligned} \frac{U}{V} &= \frac{2\gamma}{\pi^{\frac{1}{2}} V_q} \beta^{\frac{3}{2}} \int_0^\infty d\epsilon \epsilon^{\frac{3}{2}} [e^{\beta(\epsilon_j - \mu)} \pm 1]^{-1} \\ &= \frac{2\gamma}{\pi^{\frac{1}{2}} V_q} \beta^{-1} \int_0^\infty dy y^{\frac{3}{2}} [e^{y - \beta\mu} \pm 1]^{-1} \end{aligned} \quad (\text{A2.2.152})$$

$$\equiv \frac{2\gamma}{\pi^{\frac{1}{2}} V_q} \beta^{-1} F_{\frac{3}{2}}(\beta\mu) \quad (\text{A2.2.153})$$

and

$$P = \pm \frac{2\gamma}{\pi^{\frac{1}{2}} V_q} \beta^{\frac{1}{2}} \int_0^\infty d\epsilon \epsilon^{\frac{1}{2}} \log[1 \pm e^{\beta(\epsilon - \mu)}] \quad (\text{A2.2.154})$$

$$= \frac{2}{3} \frac{2\gamma}{\pi^{\frac{1}{2}} V_q} \beta^{\frac{3}{2}} \int_0^\infty d\epsilon \epsilon^{\frac{3}{2}} [e^{\beta(\epsilon_j - \mu)} \pm 1]^{-1} \quad (\text{A2.2.155})$$

$$\equiv \frac{2}{3} \frac{2\gamma}{\pi^{\frac{1}{2}} V_q} \beta^{-1} F_{\frac{3}{2}}(\beta\mu). \quad (\text{A2.2.156})$$

An integration by parts was used to deduce equation (A2.2.155) from equation (A2.2.154). Comparing the results for U and P , one finds that, just as for the classical gas, for ideal quantum gases, also, the relation $U = \frac{3}{2}PV$ is satisfied. In the above results it was found that $P = P(\beta\mu)$ and $\langle N \rangle / V \equiv n_0 = n_0(\beta\mu)$. In principle, one has to eliminate $(\beta\mu)$ between the two in order to deduce the equation of state, $P = P(\beta, n_0)$, for ideal quantum gases. Now $F_{\frac{3}{2}}(\beta\mu)$ is a function of a single variable. Therefore P is a homogeneous function of order $\frac{5}{2}$ in μ and $k_b T$. Similarly n_0 is a homogeneous function of order $\frac{3}{2}$ in μ and $k_b T$; and so is $S/V = (\partial P / \partial T)_{V, \mu}$. This means that $S/\langle N \rangle$ is a homogeneous function of order zero, i.e. $S/\langle N \rangle = \phi(\beta\mu)$, which in turn implies that for an adiabatic process $\beta\mu$ remains constant. Thus, from the expressions above for P and $\langle N \rangle / V$, one has for adiabatic processes, $P V^{\frac{5}{3}} = \text{constant}$, $V T^{\frac{2}{3}} = \text{constant}$ and $T^{\frac{5}{3}} / P = \text{constant}$.

A2.2.5.5 IDEAL QUANTUM GASES—CLASSICAL LIMIT

When the temperature is high and the density is low, one expects to recover the classical ideal gas limit. The number of particles is still given by $N = \sum n_j$. Thus the average number of particles is given by equation (A2.2.146). The average density $\langle N \rangle / V = n_0$ is the thermodynamic density. At low n_0 and high T one expects many more accessible single-particle states than the available particles, and $\langle N \rangle = \sum \langle n_j \rangle$ means that each $\langle n_j \rangle$ must be small compared to one. Thus, from equation (A2.2.145) for $\langle n_{j\bar{n}} \rangle$, the classical limit corresponds to the limit when $\exp(\beta(\epsilon_j - \mu)) \gg 1$. This has to be so for any ϵ_j , which means that the fugacity $z = \exp(-\beta\mu) \gg 1$ or $(-\beta\mu) \gg 1$ at low n_0 and high T . In this classical limit,

$\langle n_j \rangle = \exp(-\beta(\epsilon_j - \mu))$. The chemical potential μ is determined from $\langle N \rangle = \sum \langle n_j \rangle$, which leads to the result that

$$\mu = k_b T \log \left(\frac{\langle N \rangle}{\sum_j e^{-\beta \epsilon_j}} \right) \quad (\text{A2.2.157})$$

with the final result that

$$\langle n_j \rangle = \langle N \rangle \frac{e^{-\beta \epsilon_j}}{\sum_j e^{-\beta \epsilon_j}}.$$

This is the classical Boltzmann distribution in which $\langle n_j \rangle / \langle N \rangle$, the probability of finding a particle in the single-particle state j , is proportional to the classical Boltzmann factor $e^{-\beta \epsilon_j}$.

Now $\log Q_N = -\beta A$, $A = G - PV = \mu \langle N \rangle - PV$ and $\beta PV = \log \Xi$. Thus the canonical partition function is

$$\log Q(\langle N \rangle, V, T) = -\beta \mu \langle N \rangle + \log \Xi$$

which leads to the classical limit result:

$$\begin{aligned} \log Q &= -\beta \mu \langle N \rangle \pm \sum_j \log[1 \pm e^{-\beta(\epsilon_j - \mu)}] \\ &= -\beta \mu \langle N \rangle + \sum_j e^{-\beta(\epsilon_j - \mu)} \\ &= -\beta \mu \langle N \rangle + \sum_j \langle n_j \rangle = -\beta \mu \langle N \rangle + \langle N \rangle \end{aligned}$$

where the approximation $\log(1+x) \approx x$ for small x is used. Now, from the result for μ above, in equation (A2.2.157), one has

$$\beta \mu = \log \langle N \rangle - \log \sum_j e^{-\beta \epsilon_j}. \quad (\text{A2.2.158})$$

Thus

$$\log Q = -\langle N \rangle \log \langle N \rangle + \langle N \rangle + \langle N \rangle \log \sum_j e^{-\beta \epsilon_j}.$$

For large $\langle N \rangle$, $\langle N \rangle \log \langle N \rangle - \langle N \rangle \approx \log(\langle N \rangle!)$ whereby

$$Q(\langle N \rangle, V, T) = \frac{1}{\langle N \rangle!} \left(\sum_j e^{-\beta \epsilon_j} \right)^{\langle N \rangle}.$$

This result is identical to that obtained from a canonical ensemble approach in the thermodynamic limit, where the fluctuations in N vanish and $\langle N \rangle = N$. The single-particle expression for the canonical partition function $Q_1 = \sum_j e^{-\beta \epsilon_j}$ can be evaluated using $\epsilon_j = (\hbar\pi)^2 V^{-\frac{2}{3}} n_j^2 / (2M)$ for a particle in a cubical box of volume V . In the classical limit, the triplet of quantum numbers \mathbf{n}_j can be replaced by a continuous variable through the transformation

$\sum_j \rightarrow (\gamma V / \pi^3) \int_0^\infty dk_x \int_0^\infty dk_y \int_0^\infty dk_z$, and $\hbar n_j \rightarrow \hbar(V^{1/3} / \pi) \mathbf{k} \equiv \mathbf{p}$, which is the momentum of the classical particle. The transformation leads to the result that

$$Q_1 = (\gamma V / h^3) \int d^3 p \exp(-\beta p^2 / (2M)).$$

This is the same as that in the canonical ensemble. All the thermodynamic results for a classical ideal gas then follow, as in [section A2.2.4.4](#). In particular, since from [equation \(A2.2.158\)](#) the chemical potential is related to Q_1 , which was obtained in [equation \(A2.2.88\)](#), one obtains

$$\beta\mu = \log \langle N \rangle - \log Q_1 = \log[(\langle N \rangle V_q / (\gamma V))] = \log(n_o V_q / \gamma)$$

or, equivalently, $z = n_o V_q / \gamma$. The classical limit is valid at low densities when $n_o \ll \gamma / V_q$, i.e. when $z = \exp(\beta\mu) \ll 1$. For $n_o \geq (\gamma / V_q)$ one has a quantum gas. Equivalently, from the definition of V_q one has a quantum gas when $k_b T$ is below

$$k_b T_o \equiv (2\pi \hbar^2 / M)(n_o / \gamma)^{\frac{2}{3}}.$$

If $z = \exp(\beta\mu) \ll 1$, one can also consider the leading order quantum correction to the classical limit. For this consider the thermodynamic potential ω_G given in [equation \(A2.2.144\)](#). Using [equation \(A2.2.149\)](#), one can convert the sum to an integral, integrate by parts the resulting integral and obtain the result:

$$\Omega_G = -\frac{4\gamma V}{3\pi^{\frac{3}{2}} V_q \beta} \int_0^\infty dy \frac{y^{\frac{3}{2}}}{z^{-1} e^y \pm 1} \quad (\text{A2.2.159})$$

where $z = \exp(\beta\mu)$ is the fugacity, which has an ideal gas value of $n_o V_q / \gamma$. Note that the integral is the same as $F_{\frac{3}{2}}$. Since V_q is proportional to \hbar^3 , and z is small, the expansion of the integrand in powers of z is appropriate and leads to the leading quantum correction to the classical ideal gas limit. Using

$$[z^{-1} e^y \pm 1]^{-1} = z e^{-y} [1 \pm z e^{-y}]^{-1} \quad (\text{A2.2.160})$$

$$= z e^{-y} [1 \mp z e^{-y} + O(z^2)] \quad (\text{A2.2.161})$$

Ω_G can be evaluated with the result

$$\Omega_G = -PV = -\frac{\gamma V z}{V_q \beta} \left[1 \mp \frac{z}{2^{\frac{3}{2}}} + O(z^2) \right]. \quad (\text{A2.2.162})$$

The first term is the classical ideal gas term and the next term is the first-order quantum correction due to Fermi or Bose statistics, so that one can write

$$\Omega_G = \Omega_G^{\text{cl}} \pm \frac{\gamma V}{V_q \beta} \frac{z^2}{2^{\frac{3}{2}}} + O(z^3). \quad (\text{A2.2.163})$$

The small additions to all thermodynamic potentials are the same when expressed in terms of appropriate variables. Thus the first-order correction term when expressed in terms of V and β is the correction term for the Helmholtz free energy A :

$$A = A^{\text{cl}} \pm \frac{\pi^{\frac{3}{2}} N^2 \beta^{\frac{1}{2}}}{2\gamma V M^{\frac{3}{2}}} \hbar^3 + \dots \quad (\text{A2.2.164})$$

where the classical limiting value $z = n_0 V_q / \gamma$, and the definition in [equation \(A2.2.88\)](#) of V_q is used. Finally, one can obtain the correction to the ideal gas equation of state by computing $P = -(\partial A / \partial V)_{\beta, N}$. The result is

$$P = n_0 k_b T \left[1 \pm \frac{n_0}{2\gamma} \left(\frac{\pi \beta}{M} \right)^{\frac{3}{2}} \hbar^3 + \dots \right]. \quad (\text{A2.2.165})$$

The leading correction to the classical ideal gas pressure term due to quantum statistics is proportional to \hbar^3 and to n_0 . The correction at constant density is larger in magnitude at lower temperatures and lighter mass. The coefficient of n_0 can be viewed as an effective second virial coefficient $B_2^{\text{eff}}(T)$. The effect of quantum statistics at this order of correction is to add to a classical ideal gas some effective interaction. The upper sign is for a Fermi gas and yields a positive $B_2^{\text{eff}}(T)$ equivalent to an effective repulsive interaction which is a consequence of the Pauli exclusion rule. The lower sign is for a Bose gas which yields a negative $B_2^{\text{eff}}(T)$ corresponding to an effective attractive interaction. This is an indicator of the tendency of Bose particles to condense in the lowest-energy state. This phenomena is treated in [section A2.2.5.7](#).

A2.2.5.6 IDEAL FERMI GAS AND ELECTRONS IN METALS

The effects of quantum statistics are relevant in many forms of matter, particularly in solids at low temperatures. The presence of strong Coulombic interaction between electrons and ions leads one to expect that the behaviour of such systems will be highly complex and nonlinear. It is then remarkable that numerous metals and semiconductors can be well described in many respect in terms of models of effectively non-interacting ‘particles’. One such example is the thermal properties of conducting electrons in metals, which can be well approximated by an ideal gas of fermions. This approximation works at high densities on account of the Pauli exclusion principle. No two indistinguishable fermions occupy the same state, many single-particle states are filled and the lowest energy of unoccupied states is many times $k_b T$, so that the energetics of interactions between electrons become negligible. If the conduction electrons (mass m_e) in a metal are modelled by an ideal Fermi gas, the occupation number in the j th single-particle state is (from [equation \(A2.2.145\)](#))

$$\langle n_j \rangle = f(\epsilon_j) \quad \text{where } f(\epsilon_j) = [e^{\beta(\epsilon_j - \mu)} + 1]^{-1} \quad (\text{A2.2.166})$$

with $\epsilon_j = (\hbar k)^2 / (2m_e)$ and $\mathbf{k} = \mathbf{n}\pi V^{-\frac{1}{3}}$ as for a free particle in a cubical box. Consider first the situation at $T = 0$, $\beta = \infty$. Since μ depends on T , it is useful to introduce the Fermi energy and Fermi temperature as $\epsilon_F \equiv k_b T_F \equiv \mu(T = 0)$. At $T = 0$, $\langle n_j \rangle$ is one for $\epsilon_j \leq \epsilon_F$ and is zero for $\epsilon_j > \epsilon_F$. Due to the Pauli exclusion principle, each single-particle state \mathbf{n}_j is occupied by one spin-up electron and one spin-down electron. The total available, N , electrons fill up the single-particle states up to the Fermi energy ϵ_F which therefore depends on N . If n_F is defined via $\epsilon_F = (\hbar)^2 / (2m_e) V^{-\frac{2}{3}} (\pi n_F)^2$, then $N = 2 \left(\frac{1}{8}\right) (4\pi n_F^3 / 3) = \pi n_F^3 / 3$, which gives the relation between N and ϵ_F :

$$\epsilon_F \equiv k_b T_F = (\hbar)^2 / (2m_e) (3\pi^2 N / V)^{\frac{2}{3}}. \quad (\text{A2.2.167})$$

The total energy of the Fermi gas at $T = 0$ is

$$\begin{aligned} U_0 &= \sum_j \langle n_j \rangle \epsilon_j = \sum_{n < n_F} \epsilon_n = 2 \left(\frac{1}{8}\right) 4\pi \int_0^{n_F} dn n^2 \epsilon_n \\ &= \frac{\pi^3 \hbar^2}{2m_e V^{\frac{2}{3}}} \int_0^{n_F} dn n^4 = \frac{\pi^3 \hbar^2}{10m_e V^{\frac{2}{3}}} n_F^5 = \frac{3}{5} N \epsilon_F. \end{aligned} \quad (\text{A2.2.168})$$

The average kinetic energy per particle at $T = 0$, is $\frac{3}{5}$ of the Fermi energy ϵ_F . At constant N , the energy increases as the volume decreases since $\epsilon_F \sim V^{-\frac{2}{3}}$. Due to the Pauli exclusion principle, the Fermi energy gives a repulsive contribution to the binding of any material. This is balanced by the Coulombic attraction between ions and electrons in metals.

The thermal average of a physical quantity X can be computed at any temperature through

-59-

$$\langle X \rangle = \sum_{\mathbf{n}} f(\epsilon_{\mathbf{n}}, T, \mu) X_{\mathbf{n}}.$$

This can be expressed, in terms of the density of states $\mathcal{D}(\epsilon)$, as

$$\langle X \rangle = \int d\epsilon \mathcal{D}(\epsilon) f(\epsilon, T, \mu) X(\epsilon).$$

For the total number of particles N and total energy U one has

$$N = \int d\epsilon \mathcal{D}(\epsilon) f(\epsilon, T, \mu)$$

and

$$U = \int d\epsilon \mathcal{D}(\epsilon) \epsilon f(\epsilon, T, \mu).$$

For an ideal gas the density of states is computed in [section A2.2.2](#) (equation A2.2.8). Its use in evaluating N at $T = 0$ gives

$$N = \frac{\pi n_F^3}{3} = \frac{V}{3\pi^2} \left(\frac{2m_e \epsilon_F}{\hbar^2} \right)^{\frac{3}{2}} \quad \text{and} \quad \mathcal{D}(\epsilon_F) = \frac{V}{2\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{\frac{3}{2}} \epsilon_F^{\frac{1}{2}}. \quad (\text{A2.2.169})$$

This gives a simple relation

$$\mathcal{D}(\epsilon_F) = 3N/(2\epsilon_F) = 3N/(2k_b T_F). \quad (\text{A2.2.170})$$

At $T = 0$, N and U obtained above can also be found using

$$N = \int_0^{\epsilon_F} d\epsilon \mathcal{D}(\epsilon) \quad \text{and} \quad U = \int_0^{\epsilon_F} d\epsilon \mathcal{D}(\epsilon) \epsilon. \quad (\text{A2.2.171})$$

If the increase in the total energy of a system of N conduction electrons when heated from zero to T is denoted by ΔU , then

$$\Delta U = \int_0^{\infty} d\epsilon \mathcal{D}(\epsilon) \epsilon f(\epsilon) - \int_0^{\infty} d\epsilon \mathcal{D}(\epsilon) \epsilon.$$

Now multiplying the identity ($N = \int_0^{\infty} d\epsilon \mathcal{D}(\epsilon) f(\epsilon) = \int_0^{\epsilon_F} d\epsilon \mathcal{D}(\epsilon)$) by ϵ_F , one has

-60-

$$\left(\int_0^{\epsilon_F} + \int_{\epsilon_F}^{\infty} \right) d\epsilon \epsilon_F \mathcal{D}(\epsilon) f(\epsilon) = \int_0^{\epsilon_F} d\epsilon \epsilon_F \mathcal{D}(\epsilon).$$

Using this, one can rewrite the expression for ΔU in physically transparent form:

$$\Delta U = \int_{\epsilon_F}^{\infty} d\epsilon (\epsilon - \epsilon_F) \mathcal{D}(\epsilon) f(\epsilon) + \int_0^{\epsilon_F} d\epsilon (\epsilon_F - \epsilon) \mathcal{D}(\epsilon) [1 - f(\epsilon)]. \quad (\text{A2.2.172})$$

The first integral is the energy needed to move electrons from ϵ_F to orbitals with energy $\epsilon > \epsilon_F$, and the second integral is the energy needed to bring electrons to ϵ_F from orbitals below ϵ_F . The heat capacity of the electron gas can be found by differentiating ΔU with respect to T . The only T -dependent quantity is $f(\epsilon)$. So one obtains

$$C_{el} = \frac{\partial U}{\partial T} = \int_0^{\infty} d\epsilon (\epsilon - \epsilon_F) \mathcal{D}(\epsilon) \frac{\partial f}{\partial T}.$$

Now, typical Fermi temperatures in metals are of the order of 50 000 K. Thus, at room temperature, T/T_F is very small compared to one. So, one can ignore the T dependence of μ , to obtain

$$\frac{\partial f}{\partial T} = k_b \beta \frac{x e^x}{(e^x + 1)^2} \quad \text{where } x = \beta(\epsilon - \epsilon_F).$$

This is a very sharply peaked function around ϵ_F with a width of the order of $k_b T$. (At $T = 0$, $f(\epsilon)$ is a step function and its temperature derivative is a delta function at ϵ_F .) Thus in the integral for C_{el} , one can replace $\mathcal{D}(\epsilon)$ by its value at ϵ_F , transform the integration variable from ϵ to x and replace the lower limit of x , which is $(-\beta\epsilon_F)$, by $(-\infty)$. Then one obtains

$$\begin{aligned} C_{el} &= k_b^2 T \mathcal{D}(\epsilon_F) \int_{-\beta\epsilon_F}^{\infty} dx \frac{x^2 e^x}{(e^x + 1)^2} \approx k_b^2 T \mathcal{D}(\epsilon_F) \int_{-\infty}^{\infty} dx \frac{x^2 e^x}{(e^x + 1)^2} \\ &= \frac{\pi^2}{3} \mathcal{D}(\epsilon_F) k_b^2 T = \frac{\pi^2}{2} N k_b \frac{T}{T_F} \equiv A_{el} T \end{aligned} \quad (\text{A2.2.173})$$

where [equation \(A2.2.170\)](#) is used. This result can be physically understood as follows. For small T/T_F , the number of electrons excited at T from the $T = 0$ step-function Fermi distribution is of the order of NT/T_F and the energy of each of these electrons is increased by about $k_b T$. This gives $\Delta U \sim N k_b T^2/T_F$ and $C_{el} \sim N k_b T/T_F$.

In typical metals, both electrons and phonons contribute to the heat capacity at constant volume. The temperature-dependent expression

-61-

$$C_V = A_{el} T + A_{ph} T^3 \quad (\text{A2.2.174})$$

where A_{ph} is given in [equation \(A2.2.96\)](#) obtained from the Debye theory discussed in [section A2.2.4.7](#), fits the low-temperature experimental measurements of C_V for many metals quite well, as shown in [figure A2.2.6](#) for copper.

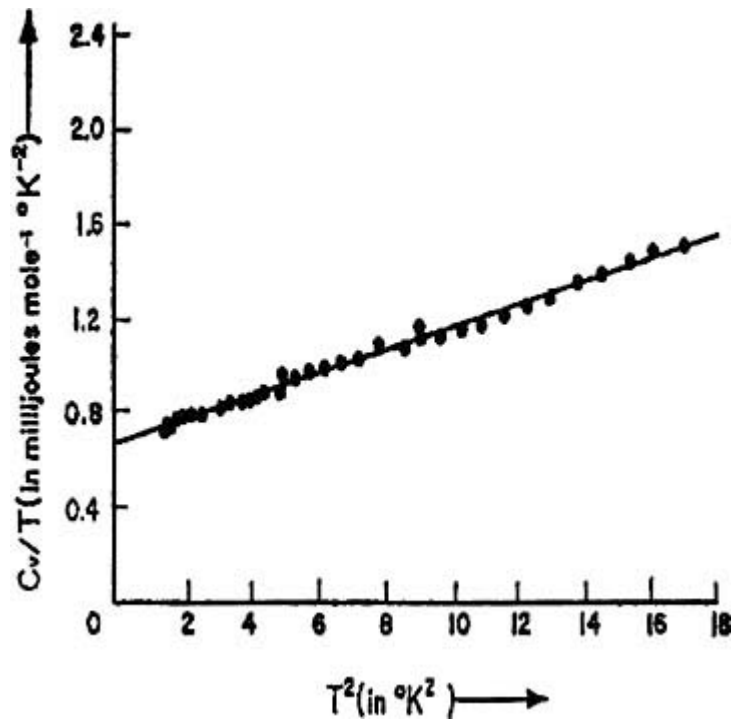


Figure A2.2.6. Electronic contribution to the heat capacity C_V of copper at low temperatures between 1 and 4 K. (From Corak *et al* [2]).

A2.2.5.7 IDEAL BOSE GAS AND BOSE-EINSTEIN CONDENSATION

In an ideal Bose gas, at a certain transition temperature a remarkable effect occurs: a macroscopic fraction of the total number of particles condenses into the lowest-energy single-particle state. This effect, which occurs when the Bose particles have non-zero mass, is called Bose-Einstein condensation, and the key to its understanding is the chemical potential. For an ideal gas of photons or phonons, which have zero mass, this effect does not occur. This is because their total number is arbitrary and the chemical potential is effectively zero for the photon or phonon gas.

From equation (A2.2.145), the average occupation number of an ideal Bose gas is

$$\langle n_j \rangle = [e^{\beta(\epsilon_j - \mu)} - 1]^{-1} \equiv [z^{-1} e^{\beta\epsilon_j} - 1]^{-1}.$$

There is clearly a possible singularity in $\langle n_j \rangle$ if $(\epsilon_j - \mu)$ vanishes. Let the energy scale be chosen such that the ground-state energy $\epsilon_0 = 0$. Then the ground-state occupancy is

$$\langle n_0 \rangle = [e^{\beta\mu} - 1]^{-1} \equiv [z^{-1} - 1]^{-1}.$$

At $T = 0$, it is expected that all the N particles will be in the ground state. Now if at low temperatures, $N \approx \langle n_0 \rangle$ is to be large, such as 10^{20} , then one must have z very close to one and $\beta\mu \ll 1$. Thus

$$N = \lim_{T \rightarrow 0} \frac{1}{e^{-\beta\mu} - 1} \approx \frac{1}{1 - \beta\mu - 1} = -\frac{k_b T}{\mu}$$

which gives the chemical potential of a Bose gas, as $T \rightarrow 0$, to be

$$\mu = -\frac{k_b T}{N} \quad \text{and the fugacity} \quad z = e^{\beta\mu} \approx 1 - \frac{1}{N}. \quad (\text{A2.2.175})$$

The chemical potential for an ideal Bose gas has to be lower than the ground-state energy. Otherwise the occupancy $\langle n_j \rangle$ of some state j would become negative.

Before proceeding, an order of magnitude calculation is in order. For $N = 10^{20}$ at $T = 1$ K, one obtains $\mu = -1.4 \times 10^{-36}$ ergs. For a He^4 atom (mass M) in a cube with $V = L^3$, the two lowest states correspond to $(n_x, n_y, n_z) = (1, 1, 1)$, and $(2, 1, 1)$. The difference in these two energies is $\Delta\epsilon = \epsilon(211) - \epsilon(111) = 3\hbar^2\pi^2/(2ML^2)$. For a box with $L = 1$ cm containing He^4 particles, $\Delta\epsilon = 2.5 \times 10^{-30}$ ergs. This is very small compared to $k_b T$, which even at 1 mK is 1.38×10^{-19} ergs. On the other hand, $\Delta\epsilon$ is large compared to μ , which at 1 mK is -1.4×10^{-39} ergs. Thus the occupancy of the (211) orbital is $\langle n_{211} \rangle \approx [\exp(\beta\Delta\epsilon) - 1]^{-1} \approx [\beta \Delta\epsilon]^{-1} \approx 0.5 \times 10^{11}$, and $\langle n_{111} \rangle \approx N \approx 10^{20}$, so that the ratio $\langle n_{211} \rangle / \langle n_{111} \rangle \approx \Sigma 10^{-9}$, a very small fraction.

For a spin-zero particle in a cubic box, the density of states is

$$\mathcal{D}(\epsilon) = \frac{V}{4\pi^2} \left(\frac{2M}{\hbar^2} \right)^{\frac{3}{2}} \epsilon^{\frac{1}{2}}.$$

The total number of particles in an ideal Bose gas at low temperatures needs to be written such that the ground-state occupancy is separated from the excited-state occupancies:

-63-

$$\begin{aligned} N &= \sum_j \langle n_j \rangle = \langle n_0 \rangle + \sum_{j \neq 0} \langle n_j \rangle \equiv N_0(T) + N_e(T) \\ &= N_0(T) + \int_0^\infty d\epsilon \mathcal{D}(\epsilon) [z^{-1} e^{\beta\epsilon} - 1]^{-1} \\ &= \frac{1}{z^{-1} - 1} + \frac{V}{4\pi^2} \left(\frac{2M}{\hbar^2} \right)^{\frac{3}{2}} \int_0^\infty d\epsilon \frac{\epsilon^{\frac{1}{2}}}{z^{-1} e^{\beta\epsilon} - 1}. \end{aligned} \quad (\text{A2.2.176})$$

Since $\mathcal{D}(\epsilon)$ is zero when $\epsilon = 0$, the ground state does not contribute to the integral for N_e . At sufficiently low temperatures, N_0 will be very large compared to one, which implies z is very close to one. Then one can approximate z by one in the integrand for N_e . Then the integral can be evaluated by using the transformation $x = \beta\epsilon$ and the known value of the integral

$$\int_0^\infty dx \frac{x^{\frac{1}{2}}}{e^x - 1} = 1.306\pi^{\frac{1}{2}}.$$

The result for the total number in the excited states is

$$N_e(T) = \frac{1.306V}{4} \left(\frac{2Mk_bT}{\pi\hbar^2} \right)^{\frac{3}{2}} = 2.612n_qV \quad (\text{A2.2.177})$$

where $n_q \equiv V_q^{-1} = [(Mk_bT)/(2\pi\hbar^2)]^{\frac{3}{2}}$ is the quantum concentration. The fraction $N_e/N \approx 2.612n_q/n$ where $n = N/V$. This ratio can also be written as

$$\frac{N_e}{N} = \left(\frac{T}{T_E} \right)^{\frac{3}{2}} \quad \text{where } T_E = \frac{2\pi\hbar^2}{Mk_b} \left(\frac{N}{2.612V} \right)^{\frac{2}{3}}. \quad (\text{A2.2.178})$$

T_E is called the Einstein temperature; $N_e(T_E) = N$. Above T_E the ground-state occupancy is not a macroscopic number. Below T_E , however, N_0 begins to become a macroscopic fraction of the total number of particles according to the relation

$$N_0 = N - N_e = N \left[1 - \left(\frac{T}{T_E} \right)^{\frac{3}{2}} \right]. \quad (\text{A2.2.179})$$

The function $N_0(T)$ is sketched in [figure A2.2.7](#). At zero temperature all the Bose particles occupy the ground state. This phenomenon is called the Bose–Einstein condensation and T_E is the temperature at which the transition to the condensation occurs.

-64-

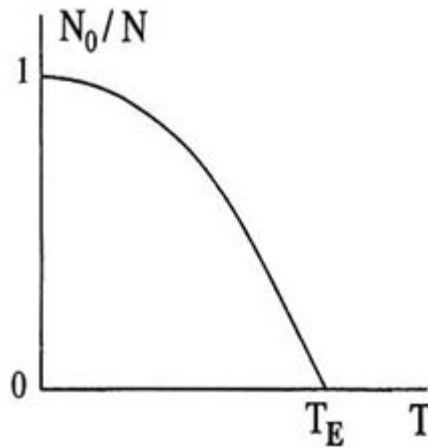


Figure A2.2.7. Fraction of Bose particles in the ground state as a function of the temperature.

A2.2.6 SUMMARY

In this chapter, the foundations of equilibrium statistical mechanics are introduced and applied to ideal and weakly interacting systems. The connection between statistical mechanics and thermodynamics is made by introducing ensemble methods. The role of mechanics, both quantum and classical, is described. In particular, the concept and use of the density of states is utilized. Applications are made to ideal quantum and classical gases, ideal gas of diatomic molecules, photons and the black body radiation, phonons in a harmonic solid, conduction electrons in metals and the Bose–Einstein condensation. Introductory aspects of the density

expansion of the equation of state and the expansion of thermodynamic quantities in powers of μ are also given. Other chapters deal with the applications to the strongly interacting systems, and the critical phenomena. Much of this section is restricted to equilibrium systems. Other sections discuss kinetic theory of fluids, chaotic and other dynamical systems, and other non-equilibrium phenomena.

REFERENCES

- [1] Born M and Huang K 1954 *Dynamical Theory of Crystal Lattices* (Oxford: Clarendon)
 - [2] Corak *et al* 1955 *Phys. Rev.* **98** 1699
-

FURTHER READING

- Chandler D 1987 *Introduction to Modern Statistical Mechanics* (Oxford: Oxford University Press)
- Hill T L 1960 *Introduction to Statistical Thermodynamics* (Reading, MA: Addison-Wesley)
-

-65-

- Huang K 1987 *Statistical Mechanics* 2nd edn (New York: Wiley)
- Kittel C and Kroemer H 1980 *Thermal Physics* 2nd edn (San Francisco, CA: Freeman)
- Landau L D and Lifshitz E M 1980 *Statistical Physics* part 1, 3rd edn (Oxford: Pergamon)
- Ma S-K 1985 *Statistical Mechanics* (Singapore: World Scientific)
- Pathria R K 1972 *Statistical Mechanics* (Oxford: Pergamon)
- Pauli W 1977 *Statistical mechanics Lectures on Physics* vol 4, ed C P Enz (Cambridge, MA: MIT)
- Plischke M and Bergersen B 1989 *Equilibrium Statistical Physics* (Englewood Cliffs, NJ: Prentice-Hall)
- Toda M, Kubo R and Saito N 1983 *Statistical Physics I* (Berlin: Springer)

-1-

A2.3 Statistical mechanics of strongly interacting systems: liquids and solids

Jayendran C Rasaiah
Had I been present at the creation, I would have given some useful hints for the better ordering of the universe.
Alphonso X, Learned King of Spain, 1252–1284

A2.3.1 INTRODUCTION

Statistical mechanics provides a link between the microscopic properties of a system at an atomic or

molecular level, and the equilibrium and dynamic properties measured in a laboratory. The statistical element follows from the enormously large number of particles involved, of the order of Avogadro's number (6.023×10^{23}), and the assumption that the measured properties (e.g. the pressure) are averages over instantaneous values. The equilibrium properties are determined from the partition function, while the transport coefficients of a system, not far from its equilibrium state, are related to equilibrium time correlation functions in the so-called linear response regime.

Fluctuations of observables from their average values, unless the observables are constants of motion, are especially important, since they are related to the response functions of the system. For example, the constant volume specific heat C_V of a fluid is a response function related to the fluctuations in the energy of a system at constant N , V and T , where N is the number of particles in a volume V at temperature T . Similarly, fluctuations in the number density ($\rho = N/V$) of an open system at constant μ , V and T , where μ is the chemical potential, are related to the isothermal compressibility κ_T , which is another response function. Temperature-dependent fluctuations characterize the dynamic equilibrium of thermodynamic systems, in contrast to the equilibrium of purely mechanical bodies in which fluctuations are absent.

In this chapter we discuss the main ideas and results of the equilibrium theory of strongly interacting systems. The partition function of a weakly interacting system, such as an ideal gas, is easily calculated to provide the absolute free energy and other properties (e.g. the entropy). The determination of the partition function of a strongly interacting system, however, is much more difficult, if not impossible, except in a few special cases. The special cases include several one-dimensional systems (e.g. hard rods, the one-dimensional (1D) Ising ferromagnet), the two-dimensional (2D) Ising model for a ferromagnet at zero magnetic field and the entropy of ice. Onsager's celebrated solution of the 2D Ising model at zero field profoundly influenced our understanding of strongly interacting systems near the critical point, where the response functions diverge. Away from this region, however, the theories of practical use to most chemists, engineers and physicists are approximations based on a mean-field or average description of the prevailing interactions. Theories of fluids in which, for example, the weaker interactions due to dispersion forces or the polarity of the molecules are treated as perturbations to the harsh repulsive forces responsible for the structure of the fluid, also fall into the mean-field category.

The structure of a fluid is characterized by the spatial and orientational correlations between atoms and molecules determined through x-ray and neutron diffraction experiments. Examples are the atomic pair correlation functions (g_{oo} , g_{oh} , g_{hh}) in liquid water. An important feature of these correlation functions is that the thermodynamic properties of a

-2-

system can be calculated from them. The information they contain is equivalent to that present in the partition function, and is more directly related to experimental observations. It is therefore natural to focus attention on the theory of these correlation functions, which is now well developed, especially in the region away from the critical point. Analytic and numerical approximations to the correlations functions are more readily formulated than for the corresponding partition functions from which they are derived. This has led to several useful theories, which include the scaled particle theory for hard bodies and integral equations approximations for the two body correlation functions of simple fluids. Examples are the Percus–Yevick, mean spherical and hypernetted chain approximations which are briefly described in this chapter and perturbation theories of fluids which are treated in greater detail.

We discuss classical non-ideal liquids before treating solids. The strongly interacting fluid systems of interest are hard spheres characterized by their harsh repulsions, atoms and molecules with dispersion interactions responsible for the liquid–vapour transitions of the rare gases, ionic systems including strong and weak electrolytes, simple and not quite so simple polar fluids like water. The solid phase systems discussed are ferromagnets and alloys.

A2.3.2 CLASSICAL NON-IDEAL FLUIDS

The main theoretical problem is to calculate the partition function given the classical Hamiltonian

$$H(\mathbf{r}^N, \mathbf{p}^N) = K(\mathbf{p}^N) + E_{\text{int}} + U_N(\mathbf{r}^N, \boldsymbol{\omega}^N) \quad (\text{A2.3.1})$$

where $K(\mathbf{p}^N)$ is the kinetic energy, E_{int} is the internal energy due to vibration, rotation and other internal degrees of freedom and

$$U_N(\mathbf{r}^N, \boldsymbol{\omega}^N) = \sum u_{ij}(r_{ij}, \omega_i, \omega_j) + \sum u_{ijk}(r_{ij}, r_{ik}, r_{kj}, \omega_i, \omega_j, \omega_k) + \dots \quad (\text{A2.3.2})$$

is the intermolecular potential composed of two-body, three-body and higher-order interactions. Here \mathbf{p}^N stands for the sets of momenta $\{p_1, p_2, \dots, p_N\}$ of the N particles, and likewise \mathbf{r}^N and $\boldsymbol{\omega}^N$ are the corresponding sets of the positions and angular coordinates of the N particles and r_{ij} is the distance between particles i and j . For an ideal gas $U_N(\mathbf{r}^N, \boldsymbol{\omega}^N) = 0$.

A2.3.2.1 INTERATOMIC POTENTIALS

Information about interatomic potentials comes from scattering experiments as well as from model potentials fitted to the thermodynamic and transport properties of the system. We will confine our discussion mainly to systems in which the total potential energy $U(\mathbf{r}^N, \boldsymbol{\omega}^N)$ for a given configuration $\{\mathbf{r}^N, \boldsymbol{\omega}^N\}$ is pairwise additive, which implies that the three- and higher-body potentials are ignored. This is an approximation because the fluctuating electron charge distribution in atoms and molecules determines their polarizability which is not pair-wise additive. However, the total potential can be approximated as the sum of effective pair potentials.

-3-

A few of the simpler pair potentials are listed below.

(a) The potential for hard spheres of diameter σ

$$u_{ij}(r) = \begin{cases} \infty & r < a_{ij} \\ 0 & r > a_{ij}. \end{cases} \quad (\text{A2.3.3})$$

(b) The square well or mound potential

$$u_{ij} = \begin{cases} \infty & r < a_{ij} \\ d_{ij} & a_{ij} < r < b_{ij} \\ 0 & b_{ij} < r. \end{cases} \quad (\text{A2.3.4})$$

(c) The Lennard-Jones potential

$$u_{ij}^{\text{LJ}}(r) = 4\varepsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right] \quad (\text{A2.3.5})$$

which has two parameters representing the atomic size σ and the well depth ε of the interatomic potential. The

r^{-6} dependence of the attractive part follows from the dispersive forces between the particles, while the r^{-12} dependence is a convenient representation of the repulsive forces. The potential is zero at $r = \sigma$ and $-\varepsilon$ at the minimum when $r = 2^{1/6}\sigma$. Typical values of ε and σ are displayed in table A2.3.1.

Table A2.3.1 Parameters for the Lennard-Jones potential.

Substance	σ (Å)	ε/k (K)
He	2.556	10.22
Ne	2.749	35.6
Ar	3.406	119.8
Kr	3.60	171
Xe	4.10	221
CH ₄	3.817	148.2

-4-

(d) The Coulomb potential between charges e_i and e_j separated by a distance r

$$u_{ij}^{\text{Coul}}(r) = \frac{e_i e_j}{r} \quad (\text{A2.3.6})$$

(e) The point dipole potential

$$u^{\text{DD}}(r, \omega_i, \omega_j) = -\frac{1}{r^3} [3(\mu_i \cdot \mathbf{r}_{ij})(\mu_j \cdot \mathbf{r}_{ij})/r^2 - (\mu_i \cdot \mu_j)] \quad (\text{A2.3.7})$$

where μ_i is the dipole moment of particle i and $r = |\mathbf{r}_{ij}|$ is the intermolecular separation between the point dipoles i and j .

Thermodynamic stability requires a repulsive core in the interatomic potential of atoms and molecules, which is a manifestation of the Pauli exclusion principle operating at short distances. This means that the Coulomb and dipole interaction potentials between charged and uncharged real atoms or molecules must be supplemented by a hard core or other repulsive interactions. Examples are as follows.

(f) The restricted primitive model (RPM) for ions in solution

$$u^{\text{RPM}}(r) = u_{ij}^{\text{HS}}(r) + u_{ij}^{\text{Coul}}(r) \quad (\text{A2.3.8})$$

in which the positive or negative charges are embedded in hard spheres of the same size in a continuum solvent of dielectric constant ϵ . An extension of this is the primitive model (PM) electrolyte in which the restriction of equal sizes for the oppositely charged ions is relaxed.

Other linear combinations of simple potentials are also widely used to mimic the interactions in real systems. An example is the following.

(g) The Stockmayer potential for dipolar molecules:

$$u_{ij}^{\text{S}}(r, \omega_i, \omega_j) = u^{\text{LJ}}(r) + u^{\text{DD}}(r, \omega_i, \omega_j) \quad (\text{A2.3.9})$$

which combines the Lennard-Jones and point dipole potentials.

Important applications of atomic potentials are models for water (TIP3, SPC/E) in which the intermolecular potential consists of atom–atom interactions between the oxygen and hydrogen atoms of distinct molecules, with the characteristic atomic geometry maintained (i.e. an HOH angle of 109° and a intramolecular OH distance of 1 \AA) by imposing constraints between atoms of the same molecule. For example, the effective simple point charge model (SPC/E) for water is defined as a linear combination of Lennard-Jones interactions between the oxygen atoms of distinct molecules and Coulombic interactions between the charges adjusted for a self-polarization correction.

-5-

The SPC/E model approximates many-body effects in liquid water and corresponds to a molecular dipole moment of 2.35 Debye (D) compared to the actual dipole moment of 1.85 D for an isolated water molecule. The model reproduces the diffusion coefficient and thermodynamics properties at ambient temperatures to within a few per cent, and the critical parameters (see below) are predicted to within 15%. The same model potential has been extended to include the interactions between ions and water by fitting the parameters to the hydration energies of small ion–water clusters. The parameters for the ion–water and water–water interactions in the SPC/E model are given in table A2.3.2.

Table A2.3.2 Halide–water, alkali metal cation–water and water–water potential parameters (SPC/E model). In the SPC/E model for water, the charges on H are at 1.000 \AA from the Lennard-Jones centre at O. The negative charge is at the O site and the HOH angle is 109.47° .

Ion/water	$\sigma_{io} \text{ (\AA)}$	$\epsilon_{io} \text{ (kJ mol}^{-1}\text{)}$	Charge (q)
F ⁻	3.143	0.6998	-1
Cl ⁻	3.785	0.5216	-1
Br ⁻	3.896	0.4948	-1
I ⁻	4.168	0.5216	-1

Li ⁺	2.337	0.6700	+1
Na ⁺	2.876	0.5216	+1
K ⁺	3.250	0.5216	+1
Rb ⁺	3.348	0.5216	+1
Cs ⁺	3.526	0.5216	+1

Water–water	σ_{oo} (Å)	ϵ_{oo} (kJ mol ⁻¹)	Charge (q)
O(H ₂ O)	3.169	0.6502	-0.8476
H(H ₂ O)			+0.4238

A2.3.2.2 EQUATIONS OF STATE, THE VIRIAL SERIES AND THE LIQUID–VAPOUR CRITICAL POINT

The equation of state of a fluid relates the pressure (P), density (ρ) and temperature (T),

$$P = P(\rho, T). \quad (\text{A2.3.10})$$

-6-

It is determined experimentally; an early study was the work of Andrews on carbon dioxide [1]. The exact form of the equation of state is unknown for most substances except in rather simple cases, e.g. a 1D gas of hard rods. However, the ideal gas law $P = \rho kT$, where k is Boltzmann's constant, is obeyed even by real fluids at high temperature and low densities, and systematic deviations from this are expressed in terms of the virial series:

$$Z = P/\rho kT = 1 + B_2(T)\rho + B_3(T)\rho^2 + \dots \quad (\text{A2.3.11})$$

which is an expansion of the compressibility factor $Z = P/\rho kT$ in powers of the number density ρ at constant temperature. Here $B_2(T)$, $B_3(T)$, \dots , $B_n(T)$ etc are the second, third, \dots and n th virial coefficients determined by the intermolecular potentials as discussed later in this chapter. They can be determined experimentally, but the radius of convergence of the virial series is not known. Figure A2.3.1 shows the second virial coefficient plotted as a function of temperature for several gases.

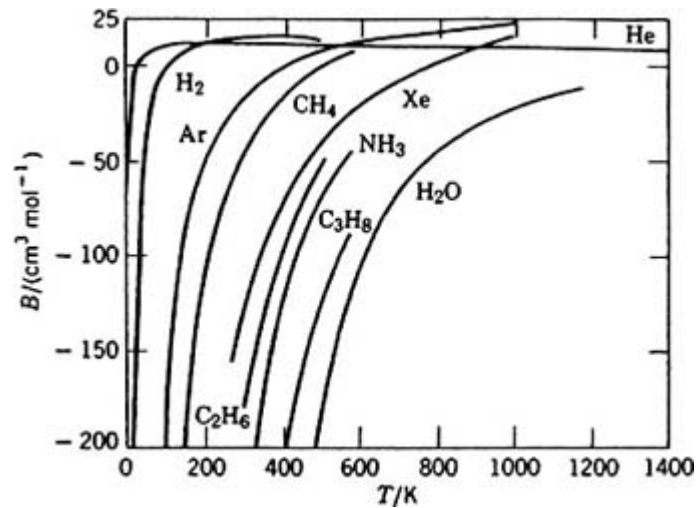


Figure A2.3.1 Second virial coefficient $B_2(T)$ of several gases as a function of temperature T . (From [10]).

The temperature at which $B_2(T)$ is zero is the Boyle temperature T_B . The excess Helmholtz free energy follows from the thermodynamic relation

$$\begin{aligned} \frac{\beta A^{\text{ex}}}{N} &= \int_0^\rho \left(\frac{P}{\rho kT} - 1 \right) d \ln \rho \\ &= \sum_{n=2}^{\infty} \frac{B_n(T)}{n-1} \rho^{n-1}. \end{aligned} \quad (\text{A2.3.12})$$

The first seven virial coefficients of hard spheres are positive and no Boyle temperature exists for hard spheres.

-7-

Statistical mechanical theory and computer simulations provide a link between the equation of state and the interatomic potential energy functions. A fluid–solid transition at high density has been inferred from computer simulations of hard spheres. A vapour–liquid phase transition also appears when an attractive component is present in the interatomic potential (e.g. atoms interacting through a Lennard-Jones potential) provided the temperature lies below T_c , the critical temperature for this transition. This is illustrated in [figure A2.3.2](#) where the critical point is a point of inflexion of the critical isotherm in the $P - V$ plane.

Below T_c , liquid and vapour coexist and their densities approach each other along the coexistence curve in the $T-V$ plane until they coincide at the critical temperature T_c . The coexisting densities in the critical region are related to $T - T_c$ by the power law

$$(\text{A2.3.13})$$

where β is called a critical exponent. The pressure P approaches the critical pressure P_c along the critical isotherm like

$$(\text{A2.3.14})$$

which defines another critical exponent δ . The isothermal compressibility κ_T and the constant volume specific heat C_V are response functions determined by fluctuations in the density and the energy. They diverge at the critical point, and determine two other critical exponents α and γ defined, along the critical isochore, by

(A2.3.15)

and

(A2.3.16)

As discussed elsewhere in this encyclopaedia, the critical exponents are related by the following expressions:

(A2.3.17)

The individual values of the exponents are determined by the symmetry of the Hamiltonian and the dimensionality of the system.

Although the exact equations of state are known only in special cases, there are several useful approximations collectively described as mean-field theories. The most widely known is van der Waals' equation [2]

(A2.3.18)

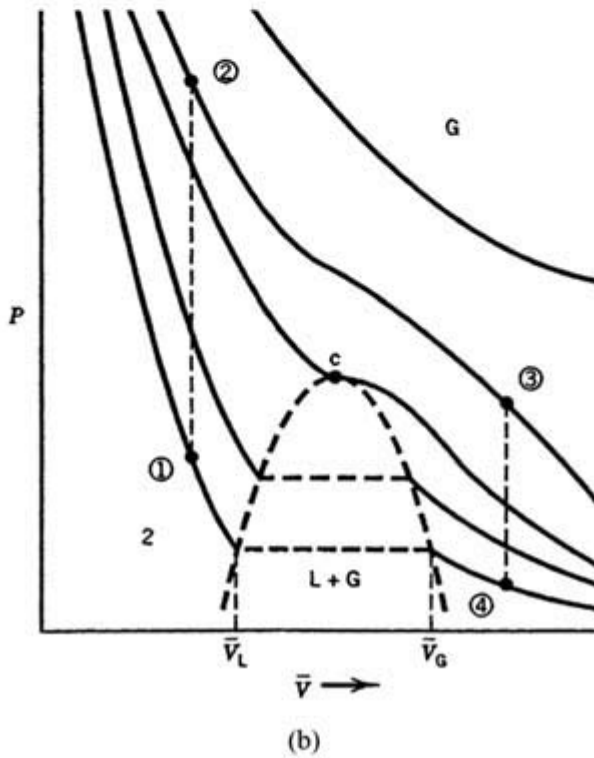
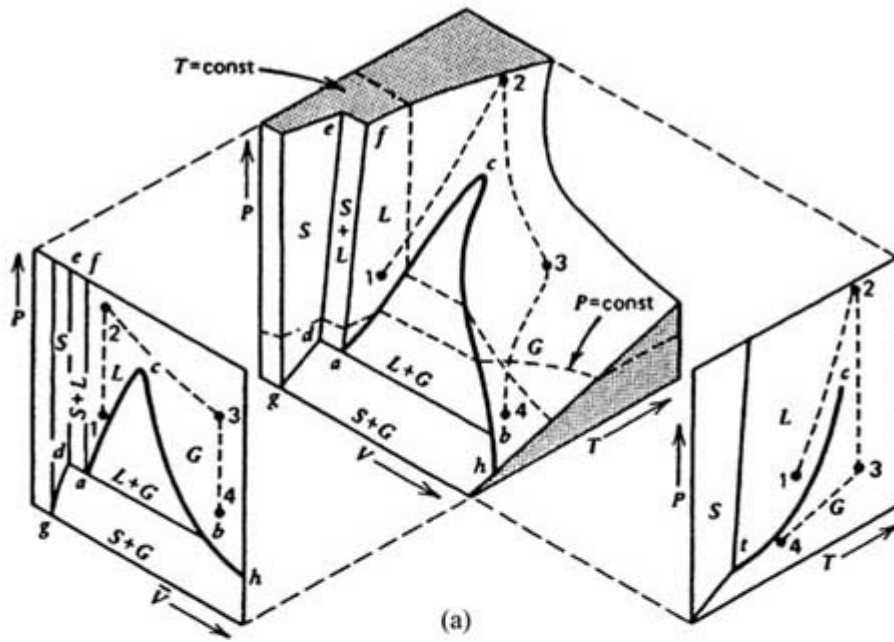


Figure A2.3.2 (a) P - V - T surface for a one-component system that contracts on freezing. (b) P - V isotherms in the region of the critical point.

The parameters a and b are characteristic of the substance, and represent corrections to the ideal gas law due to the attractive (dispersion) interactions between the atoms and the volume they occupy due to their repulsive cores. We will discuss van der Waals' equation in some detail as a typical example of a mean-field theory.

van der Waals' equation shows a liquid-gas phase transition with the critical constants $\rho_c = (1/3b)$, $P_c = a/(27b^2)$, $T_c = 8a/(27kb)$. This follows from the property that at the critical point on the P - V plane there is a point of inflexion and $(d^2P/dV^2)_c = (dP/dV)_c = 0$. These relations determine the parameters a and b from the

experimentally determined critical constants for a substance. The compressibility factor $Z_c = P_c/\rho_c kT_c$, however, is $3/8$ in contrast to the experimental value of about ≈ 0.30 for the rare gases. By expanding van der Waals' equation in powers of ρ one finds that the second virial coefficient

$$B_2(T) = b - \frac{a}{kT}. \quad (\text{A2.3.19})$$

This is qualitatively of the right form. As $T \rightarrow \infty$, $B_2(T) \rightarrow b$ and $B_2(T) = 0$ at the Boyle temperature, $T_B = a/(kb)$. This provides another route to determining the parameters a and b .

van der Waals' equation of state is a cubic equation with three distinct solutions for V at a given P and T below the critical values. Subcritical isotherms show a characteristic loop in which the middle portion corresponds to positive $(dP/dV)_T$ representing an unstable region.

The coexisting densities below T_c are determined by the equalities of the chemical potentials and pressures of the coexisting phases, which implies that the horizontal line joining the coexisting vapour and liquid phases obeys the condition

$$\mu_{\text{vapour}} - \mu_{\text{liquid}} = \int_{\text{liquid}}^{\text{vapour}} V dp = 0 \quad (T \text{ constant}). \quad (\text{A2.3.20})$$

This is the well known equal areas rule derived by Maxwell [3], who enthusiastically publicized van der Waal's equation (see [figure A2.3.3](#)). The critical exponents for van der Waals' equation are typical mean-field exponents $\alpha \approx 0$, $\beta = 1/2$, $\gamma = 1$ and $\delta = 3$. This follows from the assumption, common to van der Waals' equation and other mean-field theories, that the critical point is an analytic point about which the free energy and other thermodynamic properties can be expanded in a Taylor series.

-10-

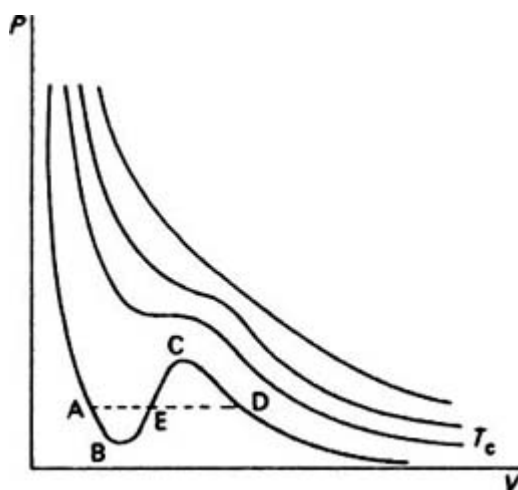


Figure A2.3.3 P - V isotherms for van der Waals' equation of state. Maxwell's equal areas rule (area ABE = area ECD) determines the volumes of the coexisting phases at subcritical temperatures.

van der Waals' equation can be written in the reduced form

$$P_R = \frac{8T_R}{(3V_R - 1)} - \frac{3}{V_R^2} \quad (\text{A2.3.21})$$

using the reduced variables $V_R = V/V_c$, $P_R = P/P_c$ and $T_R = T/T_c$. The reduced second virial coefficient

$$B_{2,R}(T_R) = \frac{B_2(T)}{\bar{V}_c} = \left[\frac{1}{3} - \frac{9}{8T_R} \right], \quad (\text{A2.3.22})$$

where $\bar{V}_c = V_c/N$ and the reduced Boyle temperature $T_{R,B} = 27/8$. The reduced forms for the equation of state and second virial coefficient are examples of the law of corresponding states. The statistical mechanical basis of this law is discussed in this chapter and has wider applicability than the van der Waals form.

A2.3.3 ENSEMBLES

An ensemble is a collection of systems with the same $r + 2$ variables, at least one of which must be extensive. Here r is the number of components. For a one-component system there are just three variables and the ensembles are characterized by given values of N, V, T (canonical), μ, V, T (grand canonical), N, V, E (microcanonical) and N, P, T (constant pressure). Our discussion of strongly interacting systems of classical fluids begins with the fundamental equations in canonical and grand canonical ensembles. The results are equivalent to each other in the thermodynamic limit. The particular choice of an ensemble is a matter of convenience in developing the theory, or in treating the fluctuations like the density or the energy.

-11-

A2.3.3.1 CANONICAL ENSEMBLE (N, V, T)

This is a collection of closed systems with the same number of particles N and volume V (constant density) for each system at temperature T . The partition function

$$(\text{A2.3.23})$$

where Q_{int} is the internal partition function (PF) determined by the vibration, rotation, electronic states and other degrees of freedom. It can be factored into contributions q_{int} from each molecule so that $Q_{\text{int}} = q_{\text{int}}^N$. The factor $1/\Lambda^{3N}$ is the translational PF in which $\Lambda = h/(2\pi mkT)^{1/2}$ is the thermal de Broglie wavelength. The configurational PF assuming classical statistics for this contribution is

$$Z(N, V, T) = \frac{1}{\Omega^N} \int \exp(-\beta U_N(r^N, \omega^N)) dr^N d\omega^N \quad (\text{A2.3.24})$$

where Ω is 4π for linear molecules and $8\pi^2$ for nonlinear molecules. The classical configurational PF is independent of the momenta and the masses of the particles. In the thermodynamic limit ($N \rightarrow \infty$, $V \rightarrow \infty$, $N/V = \rho$), the Helmholtz free energy

$$A = -kT \ln Q(N, V, T). \quad (\text{A2.3.25})$$

Other thermodynamic properties are related to the PF through the equation

$$dA = -SdT - p dV + \sum \mu_i dN_i \quad (\text{A2.3.26})$$

where μ_i is the chemical potential of species i , and the summation is over the different species. The pressure

$$P = -kT(\partial \ln Z / \partial V)_{N,T} \quad (\text{A2.3.27})$$

and since the classical configurational PF Z is independent of the mass, so is the equation of state derived from it. Differences between the equations of state of isotopes or isotopically substituted compounds (e.g. H_2O and D_2O) are due to quantum effects.

For an ideal gas, $U(\mathbf{r}^N, \omega^N) = 0$ and the configurational PF $Z(N, V, T) = V^N$. Making use of Sterling's approximation for $N! \approx (e/N)^N$ for large N , it follows that the Helmholtz free energy

$$A^{\text{ideal}} = -NkT \ln(q_{\text{int}} e / \Lambda^3) + NkT \ln \rho \quad (\text{A2.3.28})$$

-12-

and the chemical potential

$$\mu^{\text{ideal}} = (\partial A^{\text{ideal}} / \partial N)_{V,T} = kT \ln(\Lambda^3 / q_{\text{int}}) + kT \ln \rho. \quad (\text{A2.3.29})$$

The excess Helmholtz free energy

$$A^{\text{ex}} = A - A^{\text{ideal}} = -kT \ln[Z(N, V, T) / V^N] \quad (\text{A2.3.30})$$

and the excess chemical potential

$$\begin{aligned} \mu^{\text{ex}} &= (\partial A^{\text{ex}} / \partial N)_{V,T} \\ &\approx A^{\text{ex}}(N+1, V, T) - A^{\text{ex}}(N, V, T) \quad \text{for large } N \\ &= kT \ln[V Z(N, V, T) / Z(N+1, V, T)]. \end{aligned} \quad (\text{A2.3.31})$$

Confining our attention to angularly-independent potentials to make the argument and notation simpler,

$$Z(N+1, V, T) = \int \exp(-\beta U_{N+1}(\mathbf{r}^{N+1})) d\mathbf{r}^{N+1} = V \int \exp(-\beta U_{N+1}(\mathbf{r}^{N+1})) d\mathbf{r}^N$$

in which

$$U_{N+1}(\mathbf{r}^{N+1}) = U_N(\mathbf{r}^N) + \Delta U_N(\mathbf{r}_{N+1}, \mathbf{r}^N)$$

where $\Delta U_N(\mathbf{r}_{N+1}, \mathbf{r}^N)$ is the interaction of the $(N+1)$ th particle situated at \mathbf{r}_{N+1} with the remaining N particles at coordinates $\mathbf{r}^N = \{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N\}$. Substituting this into the expression for $Z(N+1, V, T)$ we see that the ratio

$$\frac{Z(N+1, V, T)}{Z(N, V, T)} = \frac{V \int \exp(-\beta U_N(\mathbf{r}^N)) \exp(-\beta \Delta U_N(\mathbf{r}_{N+1}, \mathbf{r}^N)) d\mathbf{r}_{N+1} d\mathbf{r}^N}{Z(N, V, T)}$$

But $\exp(-\beta U_N(\mathbf{r}^N)) / Z(N, V, T)$ is just the probability that the N particle system is in the configuration $\{\mathbf{r}^N\}$,

and it follows that

$$\begin{aligned}\mu^{\text{ex}} &= -kT \ln[Z(N+1, V, T)/VZ(N, V, T)] \\ &= -kT \ln\langle \exp(-\beta \Delta U_N(\mathbf{r}_{N+1}, \mathbf{r}^N)) \rangle_N\end{aligned}\quad (\text{A2.3.32})$$

-13-

where $\langle \dots \rangle_N$ represents a configurational average over the canonical N particle system. This expression for μ^{ex} , proposed by Widom, is the basis of the particle insertion method used in computer simulations to determine the excess chemical potential. The exponential term is either zero or one for hard core particles, and

$$\mu^{\text{ex}} = -kT \ln P(0, v) \quad (\text{A2.3.33})$$

where $P(0, v)$ is the probability of forming a cavity of the same size and shape as the effective volume v occupied by the particle.

A2.3.3.2 GRAND CANONICAL ENSEMBLE (μ, V, T)

This is a collection of systems at constant μ, V and T in which the number of particles can fluctuate. It is of particular use in the study of open systems. The PF

$$\Xi(\mu, V, T) = \sum_{n=0}^{\infty} Q(N, V, T) \lambda^N = \sum_{n=0}^{\infty} \frac{Z(N, V, T)}{N!} Z^N \quad (\text{A2.3.34})$$

where the absolute activity $\lambda = \exp(\mu/kT)$ and the fugacity $z = q_{\text{int}} \lambda / \Lambda^3$. The first equation shows that the grand PF is a generating function for the canonical ensemble PF. The chemical potential

$$\mu = kT \ln(\Lambda^3 / q_{\text{int}}) + kT \ln z. \quad (\text{A2.3.35})$$

For an ideal gas $Z(N, V, T) = V^N$, we have seen earlier that

$$\mu^{\text{ideal}} = kT \ln(\Lambda^3 / q_{\text{int}}) + kT \ln \rho. \quad (\text{A2.3.36})$$

The excess chemical potential

$$\mu^{\text{ex}} = \mu - \mu^{\text{ideal}} = kT \ln(z/\rho) \quad (\text{A2.3.37})$$

and $(z/\rho) \rightarrow 1$ as $\rho \rightarrow 0$. In the thermodynamic limit ($v \rightarrow \infty$), the pressure

$$p = \frac{kT}{V} \ln \Xi(\mu, V, T). \quad (\text{A2.3.38})$$

The characteristic thermodynamic equation for this ensemble is

$$(\text{A2.3.39})$$

$$d(pV) = S dT + p dv + \sum N_i d\mu_i$$

-14-

from which the thermodynamic properties follow. In particular, for a one-component system,

$$\langle N \rangle = kT \{ \partial(PV) / \partial \mu \}_{T,V} = kT \{ z \partial \ln \Xi / \partial z \}_{T,V}. \quad (\text{A2.3.40})$$

Note that the average density $\rho = \langle N \rangle / V$. Defining $\chi(z) = (1/V) \ln \Xi$, one finds that

$$\begin{aligned} P/kT &= \chi(z) \\ \rho &= z \partial \chi(z) / \partial z. \end{aligned} \quad (\text{A2.3.41})$$

On eliminating z between these equations, we get the equation of state $P = P(\rho, T)$ and the virial coefficients by expansion in powers of the density. For an ideal gas, $Z(N, V, T) = V^N$, $\Xi = \exp(zV)$ and $P/kT = z = \chi(z) = \rho$.

A2.3.3.3 THE VIRIAL COEFFICIENTS

The systematic calculation of the virial coefficients, using the grand canonical ensemble by eliminating z between the equations presented above, is a technically formidable problem. The solutions presented using graph theory to represent multidimensional integrals are elegant, but impractical beyond the first six or seven virial coefficients due to the mathematical complexity and labour involved in calculating the integrals. However, the formal theory led to the development of density expansions for the correlation functions, which have proved to be extremely useful in formulating approximations for them.

A direct and transparent derivation of the second virial coefficient follows from the canonical ensemble. To make the notation and argument simpler, we first assume pairwise additivity of the total potential with no angular contribution. The extension to angularly-independent non-pairwise additive potentials is straightforward. The total potential

$$U_N(\mathbf{r}^N) = \sum u_{ij}(r_{ij}) \quad (\text{A2.3.42})$$

and the configurational PF assuming classical statistics is

$$\begin{aligned} Z(N, V, T) &= \int \exp \left(-\beta \sum_{i<j} u(r_{ij}) \right) d\mathbf{r}_1 \dots d\mathbf{r}_N \\ &= \int \prod_{i<j} \exp(-\beta u(r_{ij})) d\mathbf{r}_1 \dots d\mathbf{r}_N. \end{aligned} \quad (\text{A2.3.43})$$

The Mayer function defined by

$$f_{ij}(r_{ij}) = \exp(-\beta u_{ij}(r_{ij})) - 1 \quad (\text{A2.3.44})$$

figures prominently in the theoretical development [7]. It is a step function for hard spheres:

$$f_{ij}^{\text{HS}}(r_{ij}) = \begin{cases} -1 & r < \sigma \\ 0 & r > \sigma \end{cases} \quad (\text{A2.3.45})$$

where σ is the sphere diameter. More generally, for potentials with a repulsive core and an attractive well, $f_{ij}(r_{ij})$ has the limiting values -1 and 0 as $r \rightarrow 0$ and ∞ , respectively, and a maximum at the interatomic distance corresponding to the minimum in $u_{ij}(r_{ij})$. Substituting $(1 + f_{ij}(r_{ij}))$ for $\exp(-\beta u_{ij}(r_{ij}))$ decomposes the configurational PF into a sum of terms:

$$\begin{aligned} Z(N, V, T) &= \int \prod_{i < j} (1 + f_{ij}(r_{ij})) \, d\mathbf{r}_1 \dots d\mathbf{r}_N \\ &= \int \left[1 + \sum_{i < j} f_{ij}(r_{ij}) + \dots \right] d\mathbf{r}_1 \dots d\mathbf{r}_N \\ &= V^N + \frac{N(N-1)}{2} V^{N-2} \int f_{12}(r_{12}) \, d\mathbf{r}_1 \, d\mathbf{r}_2 + \dots \\ &= V^N \left[1 + \frac{N(N-1)}{2V^2} \int f_{12}(r_{12}) \, d\mathbf{r}_1 \, d\mathbf{r}_2 + \dots \right] \\ &= V^N \left[1 - \frac{N(N-1)}{V} I_2(T) + \dots \right] \end{aligned} \quad (\text{A2.3.46})$$

where

$$I_2(T) = -(1/2V) \iint f_{12}(r_{12}) \, d\mathbf{r}_1 \, d\mathbf{r}_2.$$

The third step follows after interchanging summation and integration and recognizing that the $N(N-1)/2$ terms in the sum are identical. The pressure follows from the relation

$$\begin{aligned} P &= kT \left(\frac{\partial \ln Z(N, V, T)}{\partial V} \right)_{N, T} \\ &= \frac{NkT}{V} + \frac{kT}{1 - [N(N-1)/V] I_2(T)} \frac{N(N-1)}{V^2} I_2(T) + \dots \\ &= \frac{NkT}{V} \left[1 + \frac{N-1}{V} I_2(T) + \dots \right]. \end{aligned} \quad (\text{A2.3.47})$$

In the thermodynamic limit ($N \rightarrow \infty$, $V \rightarrow \infty$ with $N/V = \rho$), this is just the virial expansion for the pressure, with $I_2(T)$ identified as the second virial coefficient

$$\begin{aligned} B_2(T) &= -1/(2V) \iint f_{12}(r_{12}) \, d\mathbf{r}_1 \, d\mathbf{r}_2 \\ &= -1/(2V) \iint f_{12}(r_{12}) \, d\mathbf{r}_1 \, d\mathbf{r}_{12}. \end{aligned}$$

The second step involves a coordinate transformation to the origin centred at particle 1 with respect to which

the coordinates of particle 2 are defined. Since f_{12} depends only on the distance between particles 1 and 2, integration over the position of 1 gives a factor V that cancels the V in the denominator:

$$B_2(T) = -(1/2) \int f_{12}(r_{12}) \, d\mathbf{r}_{12}. \quad (\text{A2.3.48})$$

Finally, the assumed spherical symmetry of the interactions implies that the volume element $d\mathbf{r}_{12}$ is $4\pi r_{12}^2 \, dr_{12}$. For angularly-dependent potentials, the second virial coefficient

$$B_2(T) = -1/(2\Omega^2) \iint f_{12}(r_{12}, \omega_1, \omega_2) \, d\mathbf{r}_{12} \, d\omega_1 \, d\omega_2 \quad (\text{A2.3.49})$$

where $f_{12}(r_{12}, \omega_1, \omega_2)$ is the corresponding Mayer f -function for an angularly-dependent pair potential $u_{12}(r_{12}, \omega_1, \omega_2)$.

The n th virial coefficient can be written as sums of products of Mayer f -functions integrated over the coordinates and orientations of n particles. The third virial coefficient for spherically symmetric potentials is

$$B_3(T) = -1/(3\Omega^3) \iiint f_{12}(r_{12}) f_{13}(r_{13}) f_{23}(r_{23}) \, d\mathbf{r}_{12} \, d\mathbf{r}_{13}. \quad (\text{A2.3.50})$$

If we represent the f -bond by a line with two open circles to denote the coordinates of the particle 1 and 2, then the first two virial coefficients can be depicted graphically as

$$B_2(T) = -1/2 \bullet \text{---} \circ \quad (\text{A2.3.51})$$

and

$$B_3(T) = -1/3 \triangle \quad (\text{A2.3.52})$$

where blackening a circle implies integration over the coordinates of the particle represented by the circle. The higher virial coefficients can be economically expressed in this notation by extending it to include the symmetry number [5].

For hard spheres of diameter σ , $f_{12}(r_{12}) = -1$ for $r < \sigma$ and is zero otherwise. It follows that

$$B_2^{\text{HS}} = b_0 = 2\pi\sigma^3/3 \quad (\text{A2.3.53})$$

where the second virial coefficient, abbreviated as b_0 , is independent of temperature and is four times the volume of each sphere. This is called the excluded volume correction per molecule; the difference between the system volume V and the excluded volume of the molecules is the actual volume available for further occupancy. The factor four arises from the fact that σ is the distance of closest approach of the centers of the two spheres and the excluded volume for a pair is the volume of a sphere of radius σ . Each molecule

contributes half of this to the second virial coefficient which is equal to b_0 . The third and fourth virial coefficients for hard spheres have been calculated exactly and are

$$B_3^{\text{HS}}/b_0^2 = 5/8 \quad (\text{A2.3.54})$$

$$B_4^{\text{HS}}/b_0^3 = [2707\pi + 438\sqrt{2} - 4131 \arccos(1/3)]/4480 \quad (\text{A2.3.55})$$

while the fifth, sixth and seventh virial coefficients were determined numerically by Rhee and Hoover [8]:

$$\begin{aligned} B_5^{\text{HS}}/b_0^4 &\approx 0.1103 \\ B_6^{\text{HS}}/b_0^5 &\approx 0.0386 \\ B_7^{\text{HS}}/b_0^6 &\approx 0.0138. \end{aligned} \quad (\text{A2.3.56})$$

They are positive and independent of temperature.

The virial series in terms of the packing fraction $\eta = \pi\rho\sigma^3/3$ is then

$$P/\rho kT = 1 + 4\eta + 10\eta^2 + 18.36\eta^3 + 28.25\eta^4 + 39.5\eta^5 + \dots \quad (\text{A2.3.57})$$

which, as noticed by Carnahan and Starling [9], can be approximated as

$$P/\rho kT = 1 + 4\eta + 10\eta^2 + 18\eta^3 + 28\eta^4 + 40\eta^5 + \dots \quad (\text{A2.3.58})$$

This is equivalent to approximating the first few coefficients in this series by

$$C_n^{\text{HS}} \approx (n-1)^2 + 3(n-1) \quad \text{for } n \geq 2.$$

Assuming that this holds for all n enables the series to be summed exactly when

-18-

$$\begin{aligned} \frac{P_{\text{CS}}}{\rho kT} &= 1 + \sum_{n=2}^{\infty} [(n-1)^2 + 3(n-1)]\eta^{n-1} \\ &= \frac{1 + \eta + \eta^2 - \eta^3}{(1-\eta)^3}. \end{aligned} \quad (\text{A2.3.59})$$

This is Carnahan and Starling's (CS) equation of state for hard spheres; it agrees well with the computer simulations of hard spheres in the fluid region. The excess Helmholtz free energy

$$\frac{\beta A^{\text{ex}}}{N} = \int_0^\eta \left(\frac{P}{\rho kT} - 1 \right) d \ln \eta = \frac{\eta(4-3\eta)}{(1-\eta)^2}. \quad (\text{A2.3.60})$$

Figure A2.3.4 compares $P/\rho kT - 1$, calculated from the CS equation of state for hard spheres, as a function of

the reduced density $\rho\sigma^3$ with the virial expansion.

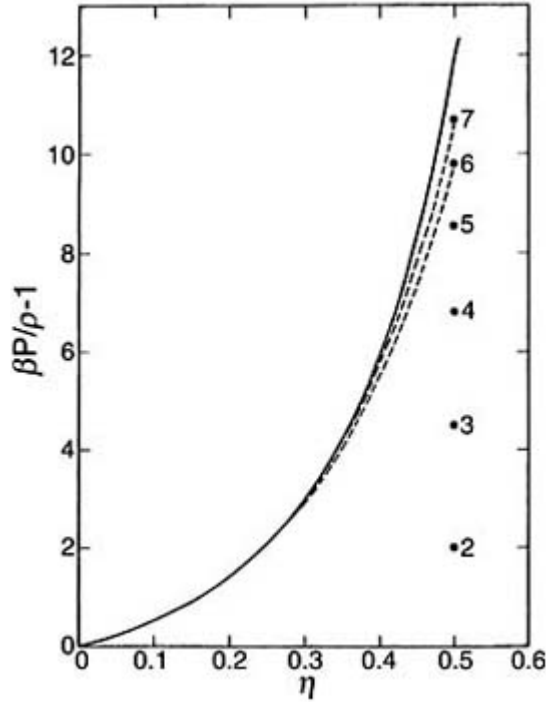


Figure A2.3.4 The equation of state $P/\rho kT - 1$, calculated from the virial series and the CS equation of state for hard spheres, as a function of $\eta = \pi\rho\sigma^3/6$ where $\rho\sigma^3$ is the reduced density.

These equations provide a convenient and accurate representation of the thermodynamic properties of hard spheres, especially as a reference system in perturbation theories for fluids.

A2.3.3.4 QUANTUM CORRECTIONS AND PATH INTEGRAL METHODS

We have so far ignored quantum corrections to the virial coefficients by assuming classical statistical mechanics in our discussion of the configurational PF. Quantum effects, when they are relatively small, can be treated as a perturbation (Friedman 1995) when the leading correction to the PF can be written as

$$Q(N, V, T) = Q_{\text{class}}(N, V, T) \left[1 - \frac{\beta^2 \hbar^2}{24} \sum_{i=1}^{3N} \frac{\langle U'' \rangle}{m_i} + O(\hbar^2) \right] \quad (\text{A2.3.61})$$

where $U'' = (\partial^2 U_N / \partial r_i^2)$ is the curvature in the potential energy function for a given configuration. The curvature of the pair potential is greatest near the minimum in the potential, and is analogous to a force constant, with the corresponding angular frequency given by $(\langle U'' \rangle / m_i)^{1/2}$. Expressing this as a wavenumber ν in cm^{-1} , the leading correction to the classical Helmholtz free energy of a system with a pairwise additive potential is given by

$$\beta(A - A_{\text{classical}}) = \left(\frac{298.16}{T} \right)^2 \sum_{i=1}^{3N} \left(\frac{\nu_i}{1015.1 \text{ cm}^{-1}} \right)^2 \quad (\text{A2.3.62})$$

which shows that the quantum correction is significant only if the mean curvature of the potential corresponds

to a frequency of 1000 cm^{-1} or more [4] and the temperature is low. Thus the quantum corrections to the second virial coefficient of light atoms or molecules He^4 , H_2 , D_2 and Ne are significant [6, 8]. For angularly-dependent potentials such as for water, the quantum effects of the rotational contributions to the second virial coefficient also contribute to significant deviations from the classical value (10 or 20%) at low temperatures [10].

When quantum effects are large, the PF can be evaluated by path integral methods [11]. Our exposition follows a review article by Gillan [12]. Starting with the canonical PF for a system of N particles

$$Q(N, V, T) = \text{Tr} e^{-\beta H} = \sum_n \langle n | e^{-\beta H} | n \rangle \quad (\text{A2.3.63})$$

where $\beta = 1/kT$ and the trace is taken over a complete set of orthonormal states $|n\rangle$. If the states $|n\rangle$ are the eigenstates of the Hamiltonian operator, the PF simplifies to

$$Q(N, V, T) = \sum_n \exp(-\beta E_n) \quad (\text{A2.3.64})$$

where the E_n are the eigenvalues. The average value of an observable A is given by

$$\langle A \rangle = Q(N, V, T)^{-1} \sum_n \langle n | \hat{A} | n \rangle \exp(-\beta E_n) \quad (\text{A2.3.65})$$

-20-

where \hat{A} is the operator corresponding to the observable A . The above expression for $\langle A \rangle$ is the sum of the expectation values of A in each state weighted by the probabilities of each state at a temperature T . The difficulty in evaluating this for all except simple systems lies in (a) the enormous number of variables required to represent the state of N particles which makes the sum difficult to determine and (b) the symmetry requirements of quantum mechanics for particle exchange which must be incorporated into the sum.

To make further progress, consider first the PF of a single particle in a potential field $V(x)$ moving in one dimension. The Hamiltonian operator

$$\hat{H} = -\frac{\hbar^2}{2m} \frac{d^2}{dx^2} + V(x). \quad (\text{A2.3.66})$$

The eigenvalues and eigenfunctions are E_n and $\phi_n(x)$ respectively. The density matrix

$$\rho(x, x'; \beta) = \sum_n \phi_n(x) \phi_n(x') \exp(-\beta E_n) \quad (\text{A2.3.67})$$

is a sum over states; it is a function of x and x' and the temperature. The PF

$$Q(V, T) = \sum_n \exp(-\beta E_n) = \int dx \rho(x, x : \beta) \quad (\text{A2.3.68})$$

which is equivalent to setting $x = x'$ in the density matrix and integrating over x . The average value of an observable A is then given by

$$\langle A \rangle = \int \hat{A} \rho(x, x')_{x=x'} dx \quad (\text{A2.3.69})$$

where $\hat{A} \rho(x, x')_{x=x'}$ means \hat{A} operates on $\rho(x, x'; \beta)$ and then x' is set equal to x . A differential equation for the density matrix follows from differentiating it with respect to β , when one finds

$$\frac{d}{d\beta} \rho(x, x'; \beta) = -\hat{H} \rho(x, x'; \beta) \quad (\text{A2.3.70})$$

which is the Bloch equation. This is similar in form to the time-dependent Schrödinger equation:

$$i\hbar \frac{d\psi(x, t)}{dt} = \hat{H} \psi(x, t). \quad (\text{A2.3.71})$$

-21-

The time evolution of the wavefunction $\psi(x, t)$ is given by

$$\psi(x, t) = \int dx' K(x, x'; t) \psi(x', 0) \quad (\text{A2.3.72})$$

where $K(x, x'; t)$ is the propagator which has two important properties. The first follows from differentiating the above equation with respect to time, when it is seen that the propagator obeys the equation

$$i\hbar \frac{dK(x, x'; t)}{dt} = \hat{H} K(x, x'; t) \quad (\text{A2.3.73})$$

with the boundary condition

$$K(x, x'; 0) = \delta(x - x'). \quad (\text{A2.3.74})$$

Comparing this with the Bloch equation establishes a correspondence between t and $i\beta\hbar$. Putting $t = i\beta\hbar$, one finds

$$\rho(x, x'; \beta) = K(x, x'; -i\beta\hbar). \quad (\text{A2.3.75})$$

The boundary condition is equivalent to the completeness relation

$$\sum_n \phi_n(x) \phi_n(x') = \delta(x - x'). \quad (\text{A2.3.76})$$

The second important property of the propagator follows from the fact that time is a serial quantity, and

$$(\text{A2.3.77})$$

$$\begin{aligned}\psi(x, t_1 + t_2) &= \int dx'' K(x, x''; t_2) \psi(x'', t_1) \\ &= \int dx' dx'' K(x, x''; t_2) K(x'', x'; t_1) \psi(x', 0)\end{aligned}$$

which implies that

$$K(x, x'; t_1 + t_2) = \int dx'' K(x, x''; t_2) K(x'', x'; t_1). \quad (\text{A2.3.78})$$

A similar expression applies to the density matrix, from its correspondence with the propagator. For example,

-22-

$$\rho(x, x'; \beta) = \int dx'' \rho(x, x''; \beta/2) \rho(x'', x'; \beta/2) \quad (\text{A2.3.79})$$

and generalizing this to the product of P factors at the inverse temperature β/P ,

$$\begin{aligned}\rho(x_0, x_P; \beta) &= \int dx_1 dx_2 \dots dx_{P-1} \rho(x_0, x_1; \beta/P) \rho(x_1, x_2; \beta/P) \\ &\dots \rho(x_{P-1}, x_P; \beta/P).\end{aligned} \quad (\text{A2.3.80})$$

Here P is an integer. By joining the ends x_0 and x_P and labelling this x , one sees that

$$\begin{aligned}Q(N, V, T) &= \int dx \rho(x, x; \beta) \\ &= \int dx_1 \dots dx_P \rho(x_1, x_2; \beta/P) \rho(x_2, x_3; \beta/P) \dots \rho(x_{P-1}, x_P; \beta/P)\end{aligned} \quad (\text{A2.3.81})$$

which has an obvious cyclic structure of P beads connecting P density matrices at an inverse temperature of β/P . This increase in temperature by a factor P for the density matrix corresponds to a short time approximation for the propagator reduced by the same factor P .

To evaluate the density matrix at high temperature, we return to the Bloch equation, which for a free particle ($V(x) = 0$) reads

$$\frac{d\rho(x, x'; \beta)}{d\beta} = \frac{\hbar^2}{2m} \frac{d^2 \rho(x, x'; \beta)}{dx^2} \quad (\text{A2.3.82})$$

which is similar to the diffusion equation

$$\frac{d\rho}{dt} = -D \frac{d^2 \rho}{dx^2}. \quad (\text{A2.3.83})$$

The solution to this is a Gaussian function, which spreads out in time. Hence the solution to the Bloch equation for a free particle is also a Gaussian:

$$\rho(x, x'; \beta) = \left(\frac{m}{2\pi\beta\hbar^2} \right)^{1/2} \exp \left[-\frac{m}{2\beta\hbar^2} (x - x')^2 - \beta V \right] \quad (\text{A2.3.84})$$

-23-

where V is zero. The above solution also applies to a single-particle, 1D system in a constant potential V . For a single particle acted on by a potential $V(x)$, we can treat it as constant in the region between x and x' , with a value equal to its mean, when β is small. One then has the approximation

$$\rho(x, x'; \beta/P) \approx \left(\frac{m}{2\pi\beta\hbar^2} \right)^{1/2} \exp \left[-\frac{mP}{2\beta\hbar^2} (x - x')^2 - \frac{\beta[V(x) + V(x')]}{2P} \right] \quad (\text{A2.3.85})$$

which leads to the following expression for the PF of a single particle in a potential $V(x)$:

$$Q(V, T) \approx Q_P(V, T) = \left(\frac{mP}{2\pi\beta\hbar} \right)^{P/2} \int dx_1 dx_2 \dots dx_P \exp \left[-\beta \sum_{s=1}^P \frac{mP}{2\beta\hbar^2} (x_{s+1} - x_s)^2 + P^{-1} V_s(x) \right]. \quad (\text{A2.3.86})$$

Feynman showed that this is exact in the limit $P \rightarrow \infty$. The expression for Q_P has the following important characteristics.

(a) It is the classical PF for a polymer chain of P beads coupled by the harmonic force constant

$$k = \frac{mP}{\beta^2\hbar^2} \quad (\text{A2.3.87})$$

with each bead acted on by a potential $V(x)/P$. The spring constant comes from the kinetic energy operator in the Hamiltonian.

(b) The cyclic polymer chain has a mean extension, characterized by the root mean square of its radius of gyration

$$\Delta^2 = P^{-1} \left\langle \sum_{s=1}^P \Delta x_s^2 \right\rangle \quad (\text{A2.3.88})$$

where $\Delta x_s = x_s - \bar{x}$ and \bar{x} is the instantaneous centre of mass of the chain, defined by

$$\bar{x} = P^{-1} \sum_{s=1}^P x_s. \quad (\text{A2.3.89})$$

For free particles, the mean square radius of gyration is essentially the thermal wavelength to within a numerical factor, and for a 1D harmonic oscillator in the $P \rightarrow \infty$ limit,

$$\Delta^2 = (\beta m \omega_0^2)^{-1} [(\beta \hbar \omega_0 / 2) \coth(\beta \hbar \omega_0 / 2) - 1] \quad (\text{A2.3.90})$$

where ω_0 is the frequency of the oscillator. As $T \rightarrow 0$, this tends to the mean square amplitude of vibration in the ground state.

(c) The probability distribution of finding the particle at x_1 is given by

$$P(x_1) \approx \frac{\int \dots \int dx_2 \dots dx_P \exp[-(k/2)(x_{s+1} - x_s)^2 + P^{-1}V(x_s)]}{\int \dots \int dx_1 \dots dx_P \exp[-(k/2)(x_{s+1} - x_s)^2 + P^{-1}V(x_s)]} \quad (\text{A2.3.91})$$

which shows that it is the same as the probability of finding one particular bead at x_1 in the classical isomorph, which is the same as $1/P$ times the probability of finding any particular one of the P beads at x_1 . The isomorphism between the quantum system and the classical polymer chain allows a variety of techniques, including simulations, to study these systems.

The eigenfunctions of a system of two particles are determined by their positions \mathbf{x} and \mathbf{y} , and the density matrix is generalized to

$$\rho(\mathbf{x}, \mathbf{y}; \mathbf{x}', \mathbf{y}'; \beta) = \sum_n \phi_n(\mathbf{x}, \mathbf{y}) \exp(-\beta E_n) \phi_n^*(\mathbf{x}', \mathbf{y}') \quad (\text{A2.3.92})$$

with the PF given by

$$Q(2, V, T) = \int d\mathbf{x} d\mathbf{y} \rho(\mathbf{x}, \mathbf{y}; \mathbf{x}, \mathbf{y}; \beta). \quad (\text{A2.3.93})$$

In the presence of a potential function $U(\mathbf{x}, \mathbf{y})$, the density matrix in the high-temperature approximation has the form

$$\rho(\mathbf{x}, \mathbf{y}; \mathbf{x}, \mathbf{y}; \beta/P) \approx \left(\frac{mP}{2\pi\beta\hbar^2} \right)^3 \exp \left\{ -\frac{mP}{2\beta\hbar^2} [(\mathbf{x} - \mathbf{x}')^2 + (\mathbf{y} - \mathbf{y}')^2] + \frac{\beta}{2P} [U(\mathbf{x}, \mathbf{y}) - U(\mathbf{x}', \mathbf{y}')] \right\}. \quad (\text{A2.3.94})$$

Using this in the expression for the PF, one finds

$$Q(2, V, T) \approx \left(\frac{mP}{2\pi\beta\hbar^2} \right)^{3P} \int d\mathbf{x}_1 d\mathbf{y}_1 \dots d\mathbf{x}_P d\mathbf{y}_P \\ \times \exp \left(-\beta \sum_{s=1}^P [(1/2)K(\mathbf{x}_{s+1} - \mathbf{x}_s)^2 + (1/2)K(\mathbf{y}_{s+1} - \mathbf{y}_s)^2 + P^{-1}U(\mathbf{x}_s, \mathbf{y}_s)] \right).$$

There is a separate cyclic polymer chain for each of the two particles, with the same force constant between adjacent beads on each chain. The potential acting on each bead in a chain is reduced, as before, by a factor $1/P$ but interacts only with the corresponding bead of the same label in the second chain. The generalization to many particles is straightforward, with one chain for each particle, each having the same number of beads coupled by harmonic springs between adjacent beads and with interactions between beads of the same label on different chains. This, however, is still an approximation as the exchange symmetry of the wavefunctions of the system is ignored.

The invariance of the Hamiltonian to particle exchange requires the eigenfunctions to be symmetric or antisymmetric, depending on whether the particles are bosons or fermions. The density matrix for bosons and fermions must then be sums over the corresponding symmetric and anti-symmetric states, respectively. Important applications of path integral simulations are to mixed classical and quantum systems, e.g. an electron in argon. For further discussion, the reader is referred to the articles by Gillan, Chandler and Wolynes, Berne and Thirumalai, Alavi and Makri in the further reading section.

We return to the study of classical systems in the remaining sections.

A2.3.3.5 1 D HARD RODS

This is an example of a classical non-ideal system for which the PF can be deduced exactly [13]. Consider N hard rods of length d in a 1D extension of length L which takes the place of the volume in a three-dimensional (3D) system. The canonical PF

$$Q(N, L, T) = \frac{Z(N, L, T)}{N! \Lambda^N} \quad (\text{A2.3.96})$$

where the configurational PF

$$Z(N, L, T) = \iint \dots \int \exp \left(-\beta \sum_{i<j} u_{ij}(r_{ij}) \right) dr_1 \dots dr_N \\ = \iint \dots \int \prod_{i<j} \exp(-\beta u_{ij}(r_{ij})) dr_1 \dots dr_N. \quad (\text{A2.3.97})$$

For hard rods of length d

$$u_{ij}(r_{ij}) = \begin{cases} \infty & r_{ij} < d \\ 0 & r_{ij} > d \end{cases} \quad (\text{A2.3.98})$$

so that an exponential factor in the integrand is zero unless $r_{ij} > d$. Another restriction in 1D systems is that since the particles are ordered in a line they cannot pass each other. Hence $d/2 < r_1 < L - Nd + d/2$, $3d/2 < r_2 < L - Nd + 3d/2$ etc. Changing variables to $x = r_1 - d/2$, $x = r_2 - 3d/2$, \dots , $x_N = r - (N-1)d/2$, we have

$$Z(N, L, T) = \int \dots \int_{0 < x_i < L - Nd} dx_1 \dots dx_N = (L - Nd)^N. \quad (\text{A2.3.99})$$

The pressure

$$P = kT(d \ln Z/dL)_{N,T} = \rho kT/(1 - \rho d) \quad (\text{A2.3.100})$$

where the density of rods $\rho = N/L$. This result is exact. Expanding the denominator (when $\rho d < 1$) leads to the virial series for the pressure:

$$P/\rho kT = 1 + d\rho + d^2\rho^2 + \dots \quad (\text{A2.3.101})$$

The n th virial coefficient $B_n^{\text{HR}} = d^{n-1}$ is independent of the temperature. It is tempting to assume that the pressure of hard spheres in three dimensions is given by a similar expression, with d replaced by the excluded volume b_0 , but this is clearly an approximation as shown by our previous discussion of the virial series for hard spheres. This is the excluded volume correction used in van der Waals' equation, which is discussed next. Other 1D models have been solved exactly in [14, 15 and 16].

A2.3.3.6 MEAN-FIELD THEORY—VAN DER WAALS' EQUATION

van der Waals' equation corrects the ideal gas equation for the attractive and repulsive interactions between molecules. The approximations employed are typical of mean-field theories. We consider a simple derivation, assuming that the total potential energy $U_N(\mathbf{r}^N)$ of the N molecules is pairwise additive and can be divided into a reference part and a remainder in any given spatial configuration $\{\mathbf{r}^N\}$. This corresponds roughly to repulsive and attractive contributions, and

$$U_N(\mathbf{r}^N) = U_N^0(\mathbf{r}^N) + U_N^{\text{attr}}(\mathbf{r}^N) \quad (\text{A2.3.102})$$

-27-

where $U_N^0(\mathbf{r}^N)$ is the energy of the reference system and

$$U_N^{\text{attr}}(\mathbf{r}^N) = \sum_{i < j} u_{ij}^{\text{attr}}(r_{ij}) \quad (\text{A2.3.103})$$

is the attractive component. This separation assumes the cross terms are zero. In the above sum, there are $N(N-1)/2$ terms that depend on the relative separation of the particles. The total attractive interaction of each molecule with the rest is replaced by an approximate average interaction, expressed as

$$U_N^{\text{attr}}(\mathbf{r}^N) = \frac{N(N-1)}{2V} \int u_{12}^{\text{attr}}(r_{12}) 4\pi r_{12}^2 dr_{12} \approx -a \frac{N^2}{V} \quad (\text{A2.3.104})$$

where

$$a = -\frac{1}{2} \int u_{12}^{\text{attr}}(r_{12}) 4\pi r_{12}^2 dr_{12}. \quad (\text{A2.3.105})$$

The PF

$$\begin{aligned} Z(N, V, T) &= \int \exp(-\beta U_N(\mathbf{r}^N)) d\mathbf{r}^N \\ &\approx \exp(\beta a N^2/V) Z^0(N, V, T) \end{aligned} \quad (\text{A2.3.106})$$

in which $Z^0(N, V, T)$ is the configurational PF of the reference system. The Helmholtz free energy and pressure follow from the fundamental equations for the canonical ensemble

$$A = -kT \ln Q(N, V, T) = A^0 + aN^2/V \quad (\text{A2.3.107})$$

and

$$P = kT (\ln Z(N, V, T)/T)_{N,T} = P^0 - aN^2/V^2. \quad (\text{A2.3.108})$$

Assuming a hard sphere reference system with the pressure given by

$$P^0 = NkT/(V - Nb) \quad (\text{A2.3.109})$$

-28-

we immediately recover van der Waals' equation of state

$$P = NkT/(V - Nb) - aN^2/V^2 \quad (\text{A2.3.110})$$

since $\rho = N/V$. An improvement to van der Waals' equation would be to use a more accurate expression for the hard sphere reference system, such as the CS equation of state discussed in the previous section. A more complete derivation that includes the Maxwell equal area rule was given in [18].

A2.3.3.7 THE LAW OF CORRESPONDING STATES

van der Waals' equation is one of several two-parameter, mean-field equations of state (e.g. the Redlich–Kwong equation) that obey the law of corresponding states. This is a scaling law in which the thermodynamic properties are expressed as functions of reduced variables defined in terms of the critical parameters of the system.

A theoretical basis for the law of corresponding states can be demonstrated for substances with the same intermolecular potential energy function but with different parameters for each substance. Conversely, the experimental verification of the law implies that the underlying intermolecular potentials are essentially similar in form and can be transformed from substance to substance by scaling the potential energy parameters. The potentials are then said to be conformal. There are two main assumptions in the derivation:

(a) quantum effects are neglected, i.e. classical statistics is assumed;

(b) the total potential is pairwise additive

$$U_N(\mathbf{r}^N) = \sum_{i < j} u_{ij}(r_{ij}) \quad (\text{A2.3.111})$$

and characterized by two parameters: a well depth ε and a size σ

$$u_{ij}(r_{ij}) = \varepsilon \phi(r_{ij}/\sigma). \quad (\text{A2.3.112})$$

The configurational PF

$$\begin{aligned} Z(N, V, T) &= \int \dots \int \exp(-\beta U_N(\mathbf{r}^N)) \, d\mathbf{r}_1 \, d\mathbf{r}_2 \dots d\mathbf{r}_N \\ &= \sigma^{3N} \int \dots \int \exp\left(-\beta \varepsilon \sum_{i < j} \phi(r_{ij}/\sigma)\right) \, d(\mathbf{r}_1/\sigma^3) \dots d(\mathbf{r}_N/\sigma^3) \\ &= \sigma^{3N} Z^*(N, V^*, T^*) \end{aligned} \quad (\text{A2.3.113})$$

where $V^* = V/\sigma^{3N}$, $T^* = kT/\varepsilon$ and $Z^*(N, V, T)$ is the integral in the second line.

-29-

The excess Helmholtz free energy

$$\begin{aligned} A^{\text{ex}} &= -kT \ln(Z(N, V, T)/V^N) = -kT \ln(Z^*(N, V^*, T^*)/V^{*N}) \\ &= -NkT[f(T^*, \rho^*)] \end{aligned} \quad (\text{A2.3.114})$$

where $\rho^* = \rho\sigma^3$, $\rho = N/V$ and

$$f(T^*, \rho^*) = [Z^*(N, V^*, T^*)]^{1/N}/N. \quad (\text{A2.3.115})$$

It follows that atoms or molecules interacting with the same pair potential $\varepsilon\phi(r_{ij}/\sigma)$, but with different ε and σ , have the same thermodynamic properties, derived from A^{ex}/NkT , at the same scaled temperature T^* and scaled density ρ^* . They obey the same scaled equation of state, with identical coexistence curves in scaled variables below the critical point, and have the same scaled vapour pressures and second virial coefficients as a function of the scaled temperature. The critical compressibility factor $P_c V_c / RT_c$ is the same for all substances obeying this law and provides a test of the hypothesis. Table A2.3.3 lists the critical parameters and the compressibility factors of rare gases and other simple substances.

Table A2.3.3 Critical constants.

Substance	T_c (K)	P_c (atm)	V_c (cm ³ mol ⁻¹)	$P_c V_c / RT_c$
He	5.21	2.26	57.76	0.305
Ne	44.44	26.86	41.74	0.307

Ar	150.7	48	75.2	0.292
Kr	209.4	54.3	92.2	0.291
Xe	289.8	58.0	118.8	0.290
CH ₄	190.6	45.6	98.7	0.288
H ₂	33.2	12.8	65.0	0.305
N ₂	126.3	33.5	126.3	0.292
O ₂	154.8	50.1	78.0	0.308
CO ₂	304.2	72.9	94.0	0.274
NH ₃	405.5	111.3	72.5	0.242
H ₂ O	647.4	218.3	55.3	0.227

-30-

The compressibility factor $Z_c = P_c V_c / RT_c$ of the rare gases Ar, Kr and Xe at the critical point is nearly 0.291, and they are expected to obey a law of corresponding states, but one very different from the prediction of van der Waals' equation discussed earlier, for which the compressibility factor at the critical point is 0.375. Deviations of Z_c from 0.291 for the other substances listed in the table are small, except for CO₂, NH₃ or H₂O, for which the molecular charge distributions contribute significantly to the intermolecular potential and lead to deviations from the law of corresponding states. The effect of hydrogen bonding in water contributes to its anomalous properties; this is mimicked by the charge distribution in the SPC/E or other models discussed in [section A2.3.2](#). The pair potentials of all the substances listed in the table, except CO₂, NH₃ or H₂O, are fairly well represented by the Lennard-Jones potential—see [table A2.3.1](#). The lighter substances, He, H₂ and to some extent Ne, show deviations due to quantum effects. The rotational PF of water in the vapour phase also has significant contribution from this source.

The equation of state determined by $Z^*(N, V^*, T^*)$ is not known in the sense that it cannot be written down as a simple expression. However, the critical parameters depend on ϵ and σ , and a test of the law of corresponding states is to use the reduced variables T_R , P_R and V_R as the scaled variables in the equation of state. [Figure A2.3.5 b](#)) illustrates this for the liquid–gas coexistence curves of several substances. As first shown by Guggenheim [19], the curvature near the critical point is consistent with a critical exponent β closer to 1/3 rather than the 1/2 predicted by van der Waals' equation. This provides additional evidence that the law of corresponding states obeyed is not the form associated with van der Waals' equation. [Figure A2.3.5 \(b\)](#) shows that $P/\rho kT$ is approximately the same function of the reduced variables T_R and P_R

$$P/\rho kT = f(T_R, P_R) \quad (\text{A2.3.116})$$

for several substances.

[Figure A2.3.6](#) illustrates the corresponding states principle for the reduced vapour pressure P_R and the second virial coefficient as functions of the reduced temperature T_R showing that the law of corresponding states is obeyed approximately by the substances indicated in the figures. The usefulness of the law also lies in its predictive value.

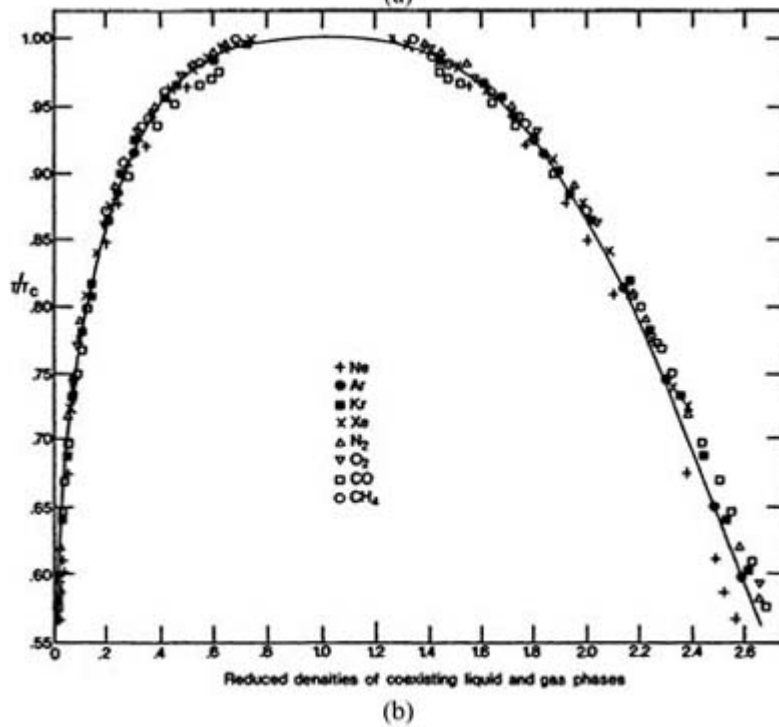
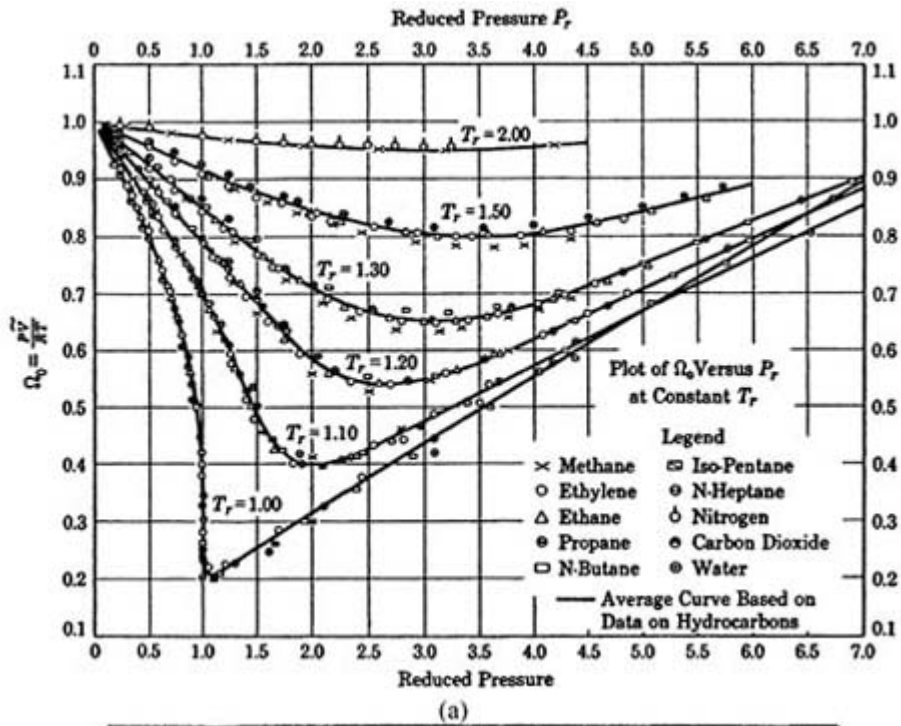


Figure A2.3.5 (a) $P/\rho kT$ as a function of the reduced variables T_R and P_R and (b) coexisting liquid and vapour densities in reduced units ρ_R as a function of T_R for several substances (after [19]).

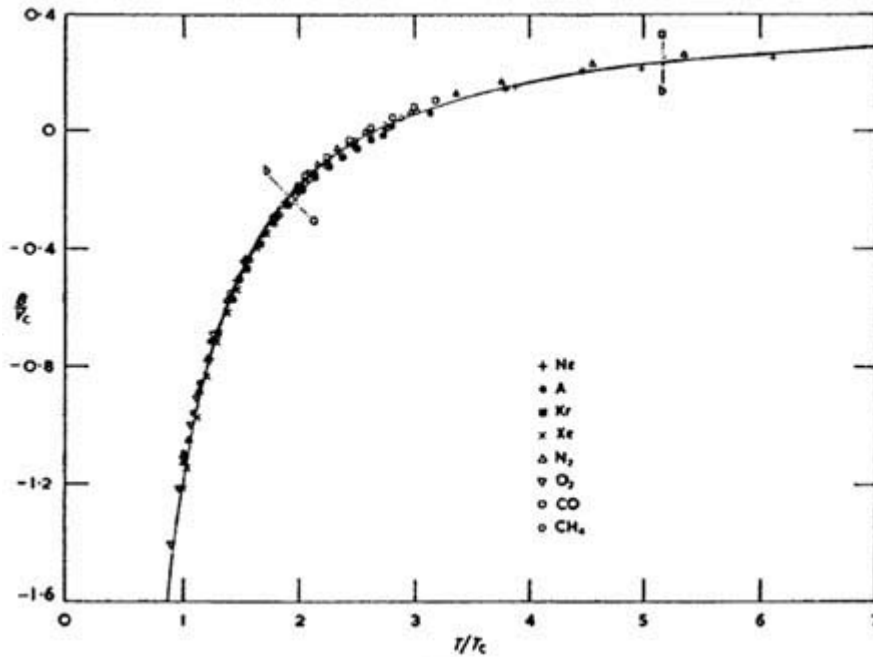


Figure A2.3.6 (a) Reduced second virial coefficient B_2/V_c as a function of T_R and (b) $\ln P_R$ versus $1/T_R$ for several substances (after [19]).

A2.3.4 CORRELATION FUNCTIONS OF SIMPLE FLUIDS

The correlation functions provide an alternate route to the equilibrium properties of classical fluids. In particular, the two-particle correlation function of a system with a pairwise additive potential determines all of its thermodynamic properties. It also determines the compressibility of systems with even more complex three-body and higher-order interactions. The pair correlation functions are easier to approximate than the PFs to which they are related; they can also be obtained, in principle, from x-ray or neutron diffraction experiments. This provides a useful perspective of fluid structure, and enables Hamiltonian models and approximations for the equilibrium structure of fluids and solutions to be tested by direct comparison with the experimentally determined correlation functions. We discuss the basic relations for the correlation functions in the canonical and grand canonical ensembles before considering applications to model systems.

A2.3.4.1 CANONICAL ENSEMBLE

The probability of observing the configuration $\{r^N\}$ in a system of given N , V and T is

$$P(\mathbf{r}^N) = \frac{\exp(-\beta U(\mathbf{r}^N))}{Z(N, V, T)} \quad (\text{A2.3.117})$$

where $Z(N, V, T)$ is the configurational PF and

$$\int P(\mathbf{r}^N) d\mathbf{r}^N = 1. \quad (\text{A2.3.118})$$

The probability function $P(\mathbf{r}^N)$ cannot be factored into contributions from individual particles, since they are coupled by their interactions. However, integration over the coordinates of all but a few particles leads to reduced probability functions containing the information necessary to calculate the equilibrium properties of the system.

Integration over the coordinates of all but one particle provides the one-particle correlation function:

$$\rho_N^{(1)}(\mathbf{r}_1) = N \int \dots \int P(\mathbf{r}^N) d\mathbf{r}_2 \dots d\mathbf{r}_N = \left\langle \sum_{i=1}^N \delta(\mathbf{r}_1 - \mathbf{r}'_i) \right\rangle_{\text{CE}} \quad (\text{A2.3.119})$$

where $\langle D \rangle_{\text{CE}}$ denotes the average value of a dynamical variable D in the canonical ensemble. Likewise, the two- and n -particle reduced correlation functions are defined by

-34-

$$\begin{aligned} \rho_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2) &= N(N-1) \int \dots \int P(\mathbf{r}^N) d\mathbf{r}_3 \dots d\mathbf{r}_N \\ &= \left\langle \sum_{\substack{i=1 \\ i \neq j}}^N \sum_{j=1}^N \delta(\mathbf{r}_1 - \mathbf{r}'_i) \delta(\mathbf{r}_2 - \mathbf{r}'_j) \right\rangle_{\text{CE}} \end{aligned} \quad (\text{A2.3.120})$$

$$\rho_N^{(n)}(\mathbf{r}_1, \dots, \mathbf{r}_n) = \frac{N!}{(N-n)!} \int \dots \int P(\mathbf{r}^N) d\mathbf{r}_{n+1} \dots d\mathbf{r}_N \quad (\text{A2.3.121})$$

where n is an integer. Integrating these functions over the coordinates $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n$ gives the normalization factor $N!/(N-n)!$. In particular,

$$\begin{aligned} \int \rho_N^{(1)}(\mathbf{r}_1) d\mathbf{r}_1 &= N \\ \int \rho_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 &= N-1. \end{aligned} \quad (\text{A2.3.122})$$

For an isotropic fluid, the one-particle correlation function is independent of the position and

$$\rho_N^{(1)}(\mathbf{r}_1) = N/V = \rho \quad (\text{A2.3.123})$$

which is the fluid density. The two-particle correlation function depends on the relative separation between particles. Assuming no angular dependence in the pair interaction,

$$\rho_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2) = \rho_N^{(2)}(r) = \rho^2 g(r) \quad (\text{A2.3.124})$$

where $r = |\mathbf{r}_{12}|$. The second relation defines the radial distribution function $g(r)$ which measures the correlation between the particles in the fluid at a separation r . Thus, we could regard $\rho g(r)$ as the average density of particles at \mathbf{r}_{12} , given that there is one at the origin \mathbf{r}_1 . Since the fluid is isotropic, the average local density in a shell of radius r and thickness Δr around each particle is independent of the direction, and

$$g(r) = \frac{\langle N(r, \Delta r) \rangle}{[(N - 1)/V]V_{\text{shell}}} \quad (\text{A2.3.125})$$

-35-

where $V_{\text{shell}} = 4\pi r^2 \Delta r$ is the volume of the shell and $\langle N(r, \Delta r) \rangle$ is the average number of particles in the shell. We see that the pair distribution function $g(r)$ is the ratio of the average number $\langle N(r, \Delta r) \rangle$ in the shell of radius r and thickness Δr to the number that would be present if there were no particle interactions. At large distances, this interaction is zero and $g(r) \rightarrow 1$ as $r \rightarrow \infty$. At very small separations, the strong repulsion between real atoms (the Pauli exclusion principle working again) reduces the number of particles in the shell, and $g(r) \rightarrow 0$ as $r \rightarrow 0$. For hard spheres of diameter σ , $g(r)$ is exactly zero for $r < \sigma$. Figure A2.3.7 illustrates the radial distribution function $g(r)$ of argon, a typical monatomic fluid, determined by a molecular dynamics (MD) simulation using the Lennard-Jones potential for argon at $T^* = 0.72$ and $\rho^* = 0.84$, and [figure A2.3.8](#) shows the corresponding atom–atom radial distribution functions $g_{\text{oo}}(r)$, $g_{\text{oh}}(r)$ and $g_{\text{hh}}(r)$ of the SPC/E model for water at 25 °C also determined by MD simulations. The correlation functions are in fair agreement with the experimental results obtained by x-ray and neutron diffraction.

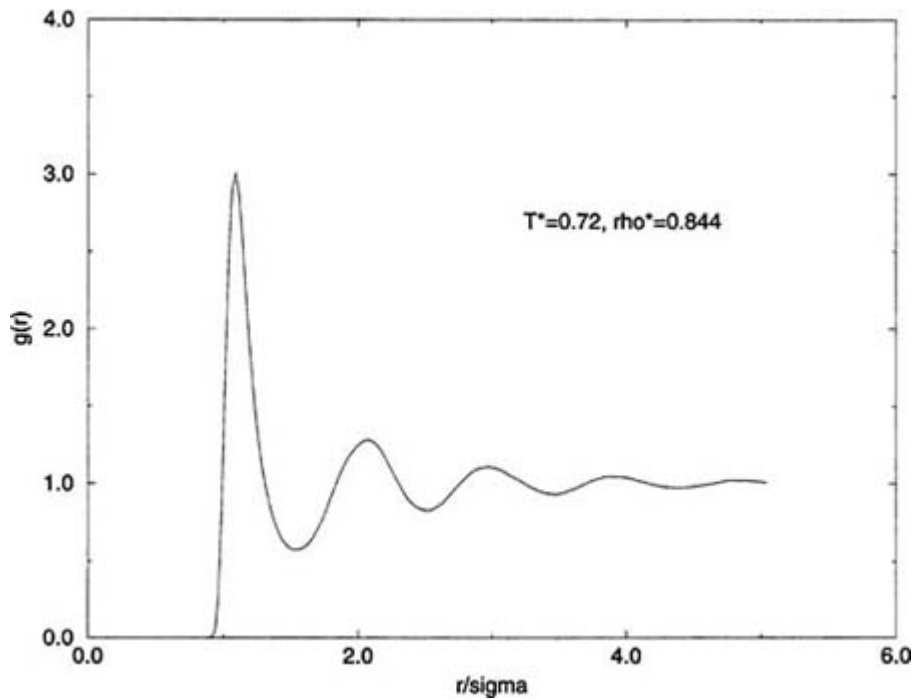


Figure A2.3.7 The radial distribution function $g(r)$ of a Lennard-Jones fluid representing argon at $T^* = 0.72$ and $\rho^* = 0.844$ determined by computer simulations using the Lennard-Jones potential.

-36-

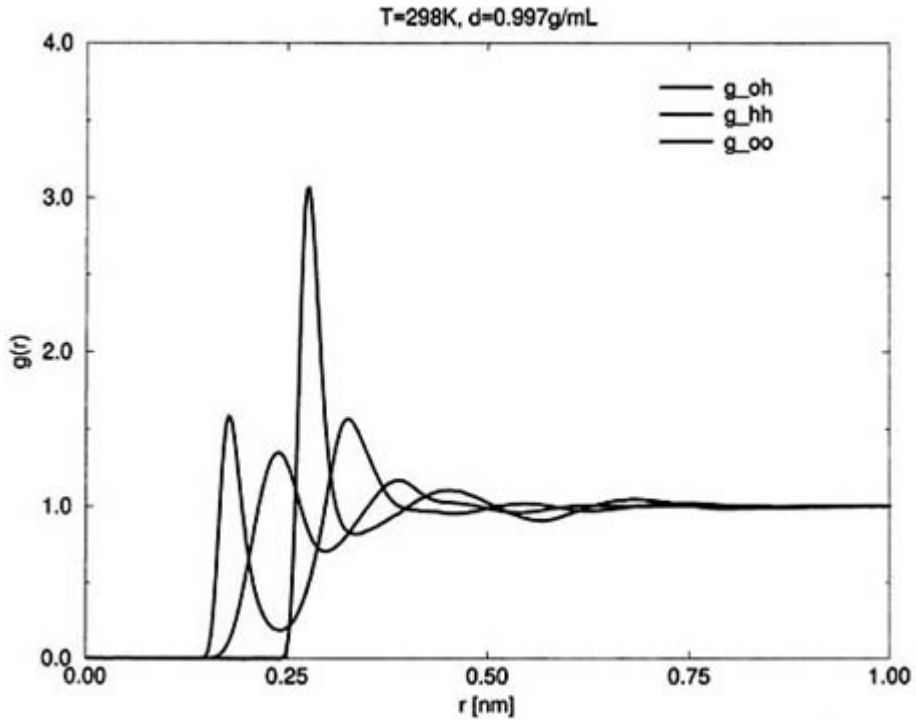


Figure A2.3.8 Atom–atom distribution functions $g_{oo}(r)$, $g_{oh}(r)$ and $g_{hh}(r)$ for liquid water at 25°C determined by MD simulations using the SPC/E model. Curves are from the leftmost peak: g_{oh} , g_{hh} , g_{oo} are red, green, blue, respectively.

Between the limits of small and large r , the pair distribution function $g(r)$ of a monatomic fluid is determined by the direct interaction between the two particles, and by the indirect interaction between the same two particles through other particles. At low densities, it is only the direct interaction that operates through the Boltzmann distribution and

$$g(r) \approx \exp(-\beta u(r)) \quad (\text{low-density approximation}). \quad (\text{A2.3.126})$$

At higher densities, the effect of indirect interactions is represented by the cavity function $y(r, \rho, T)$, which multiplies the Boltzmann distribution

$$g(r) \approx \exp(-\beta u(r)) y(r, \rho, T). \quad (\text{A2.3.127})$$

$y(r, \rho, T) \rightarrow 1$ as $\rho \rightarrow 0$, and it is a continuous function of r even for hard spheres at $r = \sigma$, the diameter of the spheres. It has a density expansion similar to the virial series:

$$y(r, \rho, T) = 1 + \sum_{n=1}^{\infty} y_n(r, T) \rho^n. \quad (\text{A2.3.128})$$

The coefficient of ρ is the convolution integral of Mayer f -functions:

$$(\text{A2.3.129})$$

$$y_1(r, T) = \int f_{13}(r_{13}) f_{32}(r_{32}) d\mathbf{r}_3 = \begin{array}{c} \diagup \quad \diagdown \\ \circ \quad \quad \circ \\ 1 \quad \quad 2 \end{array} .$$

In the graphical representation of the integral shown above, a line represents the Mayer function $f(r_{ij})$ between two particles i and j . The coordinates are represented by open circles that are labelled, unless it is integrated over the volume of the system, when the circle representing it is blackened and the label erased. The black circle in the above graph represents an integration over the coordinates of particle 3, and is not labelled. The coefficient of ρ^2 is the sum of three terms represented graphically as

$$y_2(r, T) = \begin{array}{c} \text{---} \\ \circ \quad \quad \circ \\ 1 \quad \quad 2 \end{array} + 2 \begin{array}{c} \diagup \quad \diagdown \\ \circ \quad \quad \circ \\ 1 \quad \quad 2 \end{array} + \begin{array}{c} \diagdown \quad \diagup \\ \circ \quad \quad \circ \\ 1 \quad \quad 2 \end{array} + \begin{array}{c} \diagup \quad \diagdown \\ \circ \quad \quad \circ \\ 1 \quad \quad 2 \end{array} \quad (\text{A2.3.130})$$

In general, each graph contributing to $y_n(r, T)$ has n black circles representing the number of particles through which the indirect interaction occurs; this is weighted by the n th power of the density in the expression for $g(r)$. This observation, and the symmetry number of a graph, can be used to further simplify the graphical notation, but this is beyond the scope of this article. The calculation or accurate approximation of the cavity function are important problems in the correlation function theory of non-ideal fluids.

For hard spheres, the coefficients $y_n(r)$ are independent of temperature because the Mayer f -functions, in terms of which they can be expressed, are temperature independent. The calculation of the leading term $y_1(r)$ is simple, but the determination of the remaining terms increases in complexity for larger n . Recalling that the Mayer f -function for hard spheres of diameter σ is -1 when $r < \sigma$, and zero otherwise, it follows that $y_1(r, T)$ is zero for $r > 2\sigma$. For $r < 2\sigma$, it is just the overlap volume of two spheres of radii 2σ and a simple calculation shows that

$$y_1(r) = \begin{cases} \pi\sigma^3 \left\{ \frac{4}{3} - \left(\frac{r}{\sigma}\right) + \frac{1}{12} \left(\frac{r}{\sigma}\right)^3 \right\} & r < 2\sigma \\ 0 & r > 2\sigma. \end{cases} \quad (\text{A2.3.131})$$

This leads to the third virial coefficient for hard spheres. In general, the n th virial coefficient of pairwise additive potentials is related to the coefficient $y_n(r, T)$ in the expansion of $g(r)$, except for Coulombic systems for which the virial coefficients diverge and special techniques are necessary to resum the series.

The pair correlation function has a simple physical interpretation as the potential of mean force between two particles separated by a distance r

$$w(r) = -kT \ln g(r) = u(r) - kT \ln y(r). \quad (\text{A2.3.132})$$

As $\rho \rightarrow 0$, $y(r) \rightarrow 1$ and $w(r) \rightarrow u(r)$. At higher densities, however, $w(r) \neq u(r)$. To understand its significance, consider the *mean force* between *two fixed* particles at \mathbf{r}_1 and \mathbf{r}_2 , separated by the distance $r = |\mathbf{r}_1 - \mathbf{r}_2|$. The mean force on particle 1 is the force averaged over the positions and orientations of all other particles, and is given by

$$\begin{aligned}
\langle F_1(r) \rangle &= -\langle \nabla_1 U_N(\mathbf{r}^N) \rangle \\
&= \frac{-\int \dots \int \nabla_1 U_N(\mathbf{r}^N) \exp(-\beta U_N(\mathbf{r}^N)) d\mathbf{r}_3 \dots d\mathbf{r}_N}{Z(N, V, T)} \\
&= kT \nabla_1 \ln \left[\int \dots \int \exp(-\beta U_N(\mathbf{r}^N)) d\mathbf{r}_3 \dots d\mathbf{r}_N \right] \\
&= kT \nabla_1 \ln \left[\frac{Z(N, V, T) \rho^{(2)}(\mathbf{r}_1, \mathbf{r}_2)}{N(N-1)} \right] \\
&= kT \nabla_1 \ln g(r) = -\nabla_1 w(r)
\end{aligned} \tag{A2.3.133}$$

where we have used the definition of the two-particle correlation function, $\rho^{(2)}(\mathbf{r}_1, \mathbf{r}_2)$, and its representation as $\rho^2 g(r)$ for an isotropic fluid in the last two steps. It is clear that the negative of the gradient of $w(r)$ is the force on the fixed particles, averaged over the motions of the others. This explains its characterization as the potential of mean force.

The concept of the potential of mean force can be extended to mixtures and solutions. Consider two ions in a sea of water molecules at fixed temperature T and solvent density ρ . The potential of mean force $w(r)$ is the direct interaction between the ions $u_{ij}(r) = u_{ij}^*(r) + q_i q_j / r$, plus the interaction between the ions through water molecules which is $-kT \ln y_{ij}(r)$. Here $u_{ij}^*(r)$ is the short-range potential and $q_i q_j / r$ is the Coulombic potential between ions. Thus,

$$w_{ij}(r) = u_{ij}^*(r) + q_i q_j / r - kT \ln y(r). \tag{A2.3.134}$$

At large distances, $u_{ij}^*(r) \rightarrow 0$ and $w_{ij}(r) \simeq q_i q_j / \epsilon r$ where ϵ is the macroscopic dielectric constant of the solvent. This shows that the dielectric constant ϵ of a polar solvent is related to the cavity function for two ions at large separations. One could extend this concept to define a local dielectric constant $\epsilon(r)$ for the interaction between two ions at small separations.

The direct correlation function $c(r)$ of a homogeneous fluid is related to the pair correlation function through the Ornstein–Zernike relation

$$h(r_{12}) = c(r_{12}) + \rho \int c(\mathbf{r}_{13}) h(\mathbf{r}_{32}) d\mathbf{r}_3 \tag{A2.3.135}$$

where $h(r) = g(r) - 1$ differs from the pair correlation function only by a constant term. $h(r) \rightarrow 0$ as $r \rightarrow \infty$ and is equal to -1 in the limit of $r = 0$. For hard spheres of diameter σ , $h(r) = -1$ inside the hard core, i.e. $r < \sigma$. The function $c(r)$ has the range of the intermolecular potential $u(r)$, and is generally easier to approximate. Figure A2.3.9 shows plots of $g(r)$ and $c(r)$ for a Lennard-Jones fluid at the triple point $T^* = 0.72$, $\rho^* = 0.84$, compared to $\beta u(r) = \phi(r)/\epsilon$.

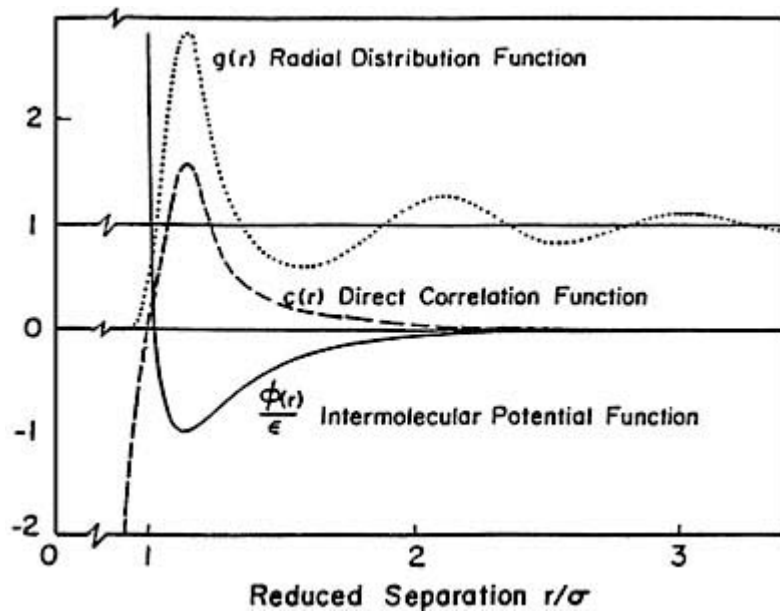


Figure A2.3.9 Plots of $g(r)$ and $c(r)$ versus r for a Lennard-Jones fluid at $T^* = 0.72$, $\rho^* = 0.84$, compared to $\beta u(r)$.

The second term in the Ornstein–Zernike equation is a convolution integral. Substituting for $h(r)$ in the integrand, followed by repeated iteration, shows that $h(r)$ is the sum of convolutions of c -functions or ‘bonds’ containing one or more c -functions in series. Representing this graphically with $c(r) = \circ - - - \circ$, we see that

$$h(r) = \circ - - - \circ + \rho \circ - - - \bullet - - - \circ + \rho^2 \circ - - - \bullet - - - \bullet - - - \circ + \dots \quad (\text{A2.3.136})$$

$h(r) - c(r)$ is the sum of series diagrams of c -bonds, with black circles signifying integration over the coordinates. It represents only part of the indirect interactions between two particles through other particles. The remaining indirect interactions cannot be represented as series diagrams and are called bridge diagrams. We now state, without proof, that the logarithm of the cavity function

$$\ln y(r) = h(r) - c(r) + B(r) \quad (\text{A2.3.137})$$

where the bridge diagram $B(r)$ has the f -bond density expansion

$$B(r_{12}) = \begin{array}{c} \diagup \quad \diagdown \\ \circ \quad \quad \circ \\ \diagdown \quad \diagup \\ 1 \quad \quad 2 \end{array} + \dots \quad (\text{A2.3.138})$$

Only the first term in this expansion is shown. It is identical to the last term shown in the equation for $y_2(r)$, which is the coefficient of ρ^2 in the expansion of the cavity function $y(r)$.

It follows that the exact expression for the pair correlation function is

$$g(r) = h(r) + 1 = \exp(-\beta u(r) + h(r) - c(r) + B(r)). \quad (\text{A2.3.139})$$

Combining this with the Ornstein–Zernike equation, we have two equations and three unknowns $h(r), c(r)$ and $B(r)$ for a given pair potential $u(r)$. The problem then is to calculate or approximate the bridge functions for which there is no simple general relation, although some progress for particular classes of systems has been made recently.

The thermodynamic properties of a fluid can be calculated from the two-, three- and higher-order correlation functions. Fortunately, only the two-body correlation functions are required for systems with pairwise additive potentials, which means that for such systems we need only a theory at the level of the two-particle correlations. The average value of the total energy

$$\langle E \rangle = \langle KE \rangle + \langle E_{\text{int}} \rangle + \langle U_N \rangle \quad (\text{A2.3.140})$$

where the translational kinetic energy $\langle KE \rangle = 3/2NkT$ is determined by the equipartition theorem. The rotational, vibrational and electronic contributions to $\langle E_{\text{int}} \rangle$ are separable and determined classically or quantum mechanically. The average potential energy

$$\langle U_N(\mathbf{r}_N) \rangle = \frac{\int \dots \int U_N(\mathbf{r}_N) \exp(-\beta U_N(\mathbf{r}_N)) \, d\mathbf{r}_N}{Z(N, V, T)}, \quad (\text{A2.3.141})$$

For a pairwise additive potential, each term in the sum of pair potentials gives the same result in the above expression and there are $N(N-1)/2$ such terms. It follows that

$$\begin{aligned} \langle U_N(\mathbf{r}_N) \rangle &= \frac{N(N-1)}{2} \frac{\int \dots \int u_{12}(r_{12}) \exp(-\beta U_N(\mathbf{r}_N)) \, d\mathbf{r}_1 \, d\mathbf{r}_2 \, d\mathbf{r}^{N-2}}{Z(N, V, T)} \\ &= \frac{1}{2} \int \dots \int u_{12}(r_{12}) \rho^{(2)}(\mathbf{r}_1, \mathbf{r}_2) \, d\mathbf{r}_1 \, d\mathbf{r}_2. \end{aligned} \quad (\text{A2.3.142})$$

For a fluid $\rho^{(2)}(r_1, r_2) = \rho^2 g(r_{12})$ where the number density $\rho = N/V$. Substituting this in the integral, changing to relative coordinates with respect to particle 1 as the origin, and integrating over r_1 to give V , leads to

$$\frac{\langle U_N \rangle}{N} = \frac{\rho}{2} \int \dots \int u_{12}(r_{12}) g(r_{12}) \, d\mathbf{r}_{12}. \quad (\text{A2.3.143})$$

-41-

The pressure follows from the virial theorem, or from the characteristic thermodynamic equation and the PF. It is given by

$$\frac{P_V}{\rho kT} = 1 - \frac{\rho}{6kT} \int r \frac{du_{12}(r)}{dr} g(r) \, d\mathbf{r}_{12} \quad (\text{A2.3.144})$$

which is called the virial equation for the pressure. The first term is the ideal gas contribution to the pressure and the second is the non-ideal contribution. Inserting $g(r) = \exp(-\beta u(r))y(r)$ and the density expansion of the cavity function $y(r)$ into this equation leads to the virial expansion for the pressure. The n th virial coefficient, $B_n(T)$, given in terms of the coefficients $y_n(r, T)$ in the density expansion of the cavity function is

$$B_n(T) = -\frac{1}{6kT} \int r \frac{du_{12}(r)}{dr} \exp(-\beta u(r)) y_{n-2}(r, T) \, d\mathbf{r}_{12}. \quad (\text{A2.3.145})$$

The virial pressure equation for hard spheres has a simple form determined by the density ρ , the hard sphere diameter σ and the distribution function at contact $g(\sigma+)$. The derivative of the hard sphere potential is discontinuous at $r = \sigma$, and

$$\phi(r) = \exp(-\beta u(r)) = \begin{cases} 1 & r < \sigma \\ 0 & r > \sigma \end{cases}$$

is a step function. The derivative of this with respect to r is a delta function

$$\frac{d\phi}{dr} = -\beta \frac{du(r)}{dr} \exp(-\beta(u(r))) = \delta(r - \sigma+)$$

and it follows that

$$\frac{du(r)}{dr} g(r) = -\frac{1}{\beta} \delta(r - \sigma+) y(r).$$

Inserting this expression in the virial pressure equation, we find that

$$\frac{P}{\rho kT} = 1 + \frac{2\pi}{3} \rho \sigma^3 g(\sigma+) \quad (\text{A2.3.146})$$

where we have used the fact that $y(\sigma) = g(\sigma+)$ for hard spheres. The virial coefficients of hard spheres are thus also related to the contact values of the coefficients $y_n(\sigma)$ in the density expansion of the cavity function. For example, the expression $y_2(r)$ for hard spheres given earlier leads to the third virial coefficient $B_3^{\text{HS}} = 5b_0^2/8$.

-42-

We conclude this section by discussing an expression for the excess chemical potential in terms of the pair correlation function and a parameter λ , which couples the interactions of one particle with the rest. The idea of a coupling parameter was introduced by Onsager [20] and Kirkwood [21]. The choice of λ depends on the system considered. In an electrolyte solution it could be the charge, but in general it is some variable that characterizes the pair potential. The potential energy of the system

$$U_N(\mathbf{r}^N; \lambda) = U_{N-1}(\mathbf{r}^{N-1}) + \lambda \sum_{j=1}^N u_{1j}(\mathbf{r}_{1j}) \quad (\text{A2.3.147})$$

where the particle at \mathbf{r}_1 couples with the remaining $N-1$ particles and $0 \leq \lambda \leq 1$. The configurational PF for this system

$$Z(N, V, T; \lambda) = \int \dots \int \exp(-\beta U_N(\mathbf{r}^N; \lambda)) d\mathbf{r}^N. \quad (\text{A2.3.148})$$

When $\lambda = 1$, we recover the PF $Z(N, V, T)$ for the fully coupled system. In the opposite limit of $\lambda = 0$, $Z(N, V, T; 0) = VZ(N-1, V, T)$, where $Z(N-1, V, T)$ refers to a fully coupled system of $N-1$ particles. Our previous discussion of the chemical potential showed that the excess chemical potential is related to the logarithm of the ratio $Z(N, V, T; 0)/VZ(N-1, V, T)$ for large N :

$$\begin{aligned}\mu^{\text{ex}} &= -kT \ln[Z(N, V, T)/VZ(N-1, V, T)] \\ &= -kT \int_0^1 \frac{\partial \ln Z(N, V, T; \lambda)}{\partial \lambda} d\lambda.\end{aligned}\tag{A2.3.149}$$

The integral is easily simplified for a pairwise additive system, and one finds

$$\begin{aligned}\frac{dZ(N, V, T; \lambda)}{d\lambda} &= -\beta \int \dots \int \sum_{j=2}^N u_{1j}(r_{1j}) \exp(-\beta U_N(\mathbf{r}^N, \lambda)) d\mathbf{r}^N \\ &= \beta(N-1) \int \dots \int u_{12}(r_{12}) \exp(-\beta U_N(\mathbf{r}^N, \lambda)) d\mathbf{r}^N.\end{aligned}$$

Dividing by $Z(N, V, T; \lambda)$ and recalling the definition of the correlation function

$$\frac{d \ln Z(N, V, T)}{d\lambda} = \frac{\beta}{N} \int \dots \int u_{12}(r_{12}) \rho^{(2)}(\mathbf{r}_1, \mathbf{r}_2; \lambda) d\mathbf{r}_1 d\mathbf{r}_2.$$

-43-

For a fluid, $\rho^{(2)}(\mathbf{r}_1, \mathbf{r}_2; \lambda) = \rho^2 g(r_{12}; \lambda)$. Changing to relative coordinates, integrating over the coordinates of particle 1 and inserting this in the expression for the excess chemical potential leads to the final result

$$\mu^{\text{ex}} = \rho \int_0^1 d\lambda \int u(r_{12}) g(r_{12}; \lambda) d\mathbf{r}_{12}.\tag{A2.3.150}$$

This is Kirkwood's expression for the chemical potential. To use it, one needs the pair correlation function as a function of the coupling parameter λ as well as its spatial dependence. For instance, if λ is the charge on a selected ion in an electrolyte, the excess chemical potential follows from a theory that provides the dependence of $g(r_{12}; \lambda)$ on the charge and the distance r_{12} . This method of calculating the chemical potential is known as the Guntelburg charging process, after Guntelburg who applied it to electrolytes.

By analogy with the correlation function for the fully coupled system, the pair correlation function $g(r; \lambda)$ for an intermediate values of λ is given by

$$g(r_{12}; \lambda) = \exp(-\beta \lambda u_{12}(r_{12})) y(r_{12}, \rho, T; \lambda)\tag{A2.3.151}$$

where $y(r, \rho, T; \lambda)$ is the corresponding cavity function for the partially coupled system. Kirkwood derived an integral equation for $g(r; \lambda)$ in terms of a three-body correlation function approximated as the product of two-body correlation functions called the superposition approximation. The integral equation, which can be solved numerically, gives results of moderate accuracy for hard spheres and Lennard-Jones systems. A similar approximation is due to Born and Green [23, 24] and Yvon [22]. Other approximations for $g(r)$ are discussed later in this chapter.

The presence of three-body interactions in the total potential energy leads to an additional term in the internal energy and virial pressure involving the three-body potential $u_{123}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)$, and the corresponding three-body correlation function $g^{(3)}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)$. The expression for the energy is then

$$\begin{aligned} \frac{\langle U_N \rangle}{N} = & \frac{\rho}{2} \int \dots \int u_{12}(r_{12})g(r_{12}) \, d\mathbf{r}_{12} + \frac{\rho}{6} \int \\ & \dots \int u_{123}(r_1, r_2, r_3)g^{(3)}(r_{12}, r_{13}, r_{23}) \, d\mathbf{r}_2 \, d\mathbf{r}_3. \end{aligned} \quad (\text{A2.3.152})$$

The virial equation for the pressure is also modified by the three-body and higher-order terms, and is given in general by

$$P = \rho kT - \frac{1}{DV} \left\langle \sum_{i < j} \mathbf{r}_i \cdot \nabla_i U_N(\mathbf{r}^N) \right\rangle \quad (\text{A2.3.153})$$

where D is the dimensionality of the system.

-44-

A2.3.4.2 GRAND CANONICAL ENSEMBLE (μ, V, T)

The grand canonical ensemble is a collection of open systems of given chemical potential μ , volume V and temperature T , in which the number of particles or the density in each system can fluctuate. It leads to an important expression for the compressibility κ_T of a one-component fluid:

$$\rho kT \kappa_T = \frac{\langle N^2 \rangle - \langle N \rangle^2}{\langle N \rangle} \quad (\text{A2.3.154})$$

where the compressibility can be determined experimentally from light scattering or neutron scattering experiments. Generalizations of the above expression to multi-component systems have important applications in the theory of solutions [25].

It was shown in [section A2.3.3.2](#) that the grand canonical ensemble (GCE) PF is a generating function for the canonical ensemble PF, from which it follows that correlation functions in the GCE are just averages of the fluctuating numbers N and $N - 1$

$$\begin{aligned} \int \langle \rho^{(1)}(\mathbf{r}_1) \rangle_{\text{GCE}} \, d\mathbf{r}_1 &= \langle N \rangle \\ \iint \langle \rho^{(2)}(\mathbf{r}_1, \mathbf{r}_2) \rangle_{\text{GCE}} \, d\mathbf{r}_1 \, d\mathbf{r}_2 &= \langle N - 1 \rangle. \end{aligned} \quad (\text{A2.3.155})$$

We see that

$$\iint [\langle \rho^{(2)}(\mathbf{r}_1, \mathbf{r}_2) \rangle - \langle \rho^{(1)}(\mathbf{r}_1) \rangle \langle \rho^{(1)}(\mathbf{r}_2) \rangle] \, d\mathbf{r}_1 \, d\mathbf{r}_2 = \langle N(N - 1) \rangle - \langle N \rangle^2$$

where the subscript GCE has been omitted for convenience. The right-hand side of this is just $\langle N^2 \rangle - \langle N \rangle^2 - \langle N \rangle$. The pair correlation function $g(\mathbf{r}_1, \mathbf{r}_2)$ is defined by

$$\langle \rho^{(2)}(\mathbf{r}_1, \mathbf{r}_2) \rangle = \langle \rho^{(1)}(\mathbf{r}_1) \rangle \langle \rho^{(1)}(\mathbf{r}_2) \rangle g(\mathbf{r}_1, \mathbf{r}_2) \quad (\text{A2.3.156})$$

and it follows that

$$\iint \langle \rho_1(\mathbf{r}_1) \rangle \langle \rho_1(\mathbf{r}_2) \rangle [g(\mathbf{r}_1, \mathbf{r}_2) - 1] d\mathbf{r}_1 d\mathbf{r}_2 = \langle N^2 \rangle - \langle N \rangle^2 - \langle N \rangle.$$

For an isotropic fluid, the singlet density is the density of the fluid, i.e. $\langle \rho^{(1)}(r) \rangle = \langle N \rangle / V = \rho$, and the pair correlation function $g(\mathbf{r}_1, \mathbf{r}_2)$ depends on the interparticle separation $r_{12} = |\mathbf{r}_1 - \mathbf{r}_2|$. Using this in the above integral, changing to relative coordinates with respect to particle 1 as the origin and integrating over its coordinates, one finds

$$\rho^2 V \int [g(r) - 1] d\mathbf{r}_{12} = \langle N^2 \rangle - \langle N \rangle^2 - \langle N \rangle.$$

Division by $\langle N \rangle$ and taking note of the fluctuation [formula \(A2.3.144\)](#) for the compressibility leads to the fundamental relation

$$\rho k T \kappa_T = 1 + \rho \int [g(r_{12}) - 1] d\mathbf{r}_{12} \quad (\text{A2.3.157})$$

called the compressibility equation which is not limited to systems with pairwise additive potentials. Integrating the compressibility with respect to the density provides an independent route to the pressure, aside from the pressure calculated from the virial equation. The exact pair correlation function for a given model system should give the same values for the pressure calculated by different routes. This serves as a test for the accuracy of an approximate $g(r)$ for a given Hamiltonian.

The first term in the compressibility equation is the ideal gas term and the second term, the integral of $g(r)-1 = h(r)$, represents the non-ideal contribution due to the correlation or interaction between the particles. The correlation function $h(r)$ is zero for an ideal gas, leaving only the first term. The correlations between the particles in a fluid displaying a liquid-gas critical point are characterized by a correlation length ζ that becomes infinitely large as the critical point is approached. This causes the integral in the compressibility equation and the compressibility κ_T to diverge.

The divergence in the correlation length ζ is characterized by the critical exponent ν defined by

$$\zeta = |T - T_c|^{-\nu} \quad (\text{A2.3.158})$$

while the divergence in the compressibility, near the critical point, is characterized by the exponent γ as discussed earlier. The correlation function near the critical region has the asymptotic form [\[26\]](#)

$$h(r) \approx \frac{f(r/\zeta)}{r^{D-2+\eta}} \quad (\text{A2.3.159})$$

where D is the dimensionality and η is a critical exponent. Substituting this in the compressibility equation, it follows with $D = 3$ that

$$k T \kappa_T \approx \xi^{2-\eta} \int f(x) x^{1-\eta} x dx \quad (\text{A2.3.160})$$

where $x = r/\zeta$. Inserting the expressions for the temperature dependence of the compressibility and the correlation length near the critical point, one finds that the exponents are related by

-46-

$$\gamma = \nu(2 - \eta). \quad (\text{A2.3.161})$$

Table A2.3.4 summarizes the values of these critical exponents in two and three dimensions and the predictions of mean field theory.

Table A2.3.4 The critical exponents γ , ν and η .

Exponent	MFT	Ising ($d = 2$)	Numerical ($d = 3$)
ν	1/2	1	0.630 ± 0.001
γ	1	7/4	1.239 ± 0.002
η	0	1/4	0.03

The compressibility equation can also be written in terms of the direct correlation function. Taking the Fourier transform of the Ornstein–Zernike equation

$$\tilde{h}(k) = \tilde{c}(k) + \rho \tilde{c}(k) \tilde{h}(k) \quad (\text{A2.3.162})$$

where we have used the property that the Fourier transform of a convolution integral is the product of Fourier transforms of the functions defining the convolution. Here the Fourier transform of a function $f(r)$ is defined by

$$\tilde{f}(k) = \int f(r) \exp(-ikr) \, dr. \quad (\text{A2.3.163})$$

From the Ornstein–Zernike equation in Fourier space one finds that

$$1 + \rho \tilde{h}(k) = [1 - \rho \tilde{c}(k)]^{-1}$$

when $k = 0$, $1 + \rho \tilde{h}(0)$ is just the right-hand side of the compressibility equation. Taking the inverse, it follows that

$$\beta \left(\frac{\partial P}{\partial \rho} \right)_T = [1 - \rho \tilde{c}(0)] = 1 - \rho \int c(r) \, dr. \quad (\text{A2.3.164})$$

At the critical point $\beta(\partial P/\partial \rho)_T = 0$, and the integral of the direct correlation function remains finite, unlike the integral of $h(r)$.

A2.3.4.3 INTEGRAL EQUATION APPROXIMATIONS FOR A FLUID

The equilibrium properties of a fluid are related to the correlation functions which can also be determined experimentally from x-ray and neutron scattering experiments. Exact solutions or approximations to these correlation functions would complete the theory. Exact solutions, however, are usually confined to simple systems in one dimension. We discuss a few of the approximations currently used for 3D fluids.

Successive n and $n + 1$ particle density functions of fluids with pairwise additive potentials are related by the Yvon–Born–Green (YBG) hierarchy [6]

$$\nabla_1 \rho^{(n)}(\mathbf{r}^n) = \beta \left(\mathbf{F}_1^{\text{ext}} + \sum_{j=2}^n \mathbf{F}_{1j} \right) \rho^{(n)}(\mathbf{r}^n) + \beta \int \mathbf{F}_{1,n+1} \rho^{(n+1)}(\mathbf{r}^{n+1}) d\mathbf{r}^{n+1} \quad (\text{A2.3.165})$$

where $\mathbf{F}_1^{\text{ext}} = -\nabla_1 \phi$ is the external force, $\mathbf{F}_{1j} = -\nabla_1 u(r_{1j})$ and $\mathbf{r}^n \equiv \{r_1, r_2, r_3 \dots r_n\}$ is the set of coordinates of n particles. The simplest of these occurs when $n = 1$, and it relates the one- and two-particle density functions of a fluid in an inhomogeneous field, e.g. a fluid near a wall:

$$kT \ln \rho(\mathbf{r}_1, [\phi]) = -\nabla_1 \phi(\mathbf{r}_1) - \int \rho(\mathbf{r}_2 | \mathbf{r}_1; [\phi]) \nabla_1 u(r_{12}) d\mathbf{r}_2 \quad (\text{A2.3.166})$$

where $\rho(\mathbf{r}_2 | \mathbf{r}_1; [\phi]) = \rho^{(2)}(\mathbf{r}_1, \mathbf{r}_2; [\phi]) \rho(\mathbf{r}_1; [\phi])$ and the superscript 1 is omitted from the one-particle local density. For an homogeneous fluid in the absence of an external field, $\mathbf{F}^{\text{ext}} = 0$ and $\rho^{(n)}(\mathbf{r}^n) = \rho^n g^{(n)}(\mathbf{r}^n)$ and the YBG equation leads to

$$\nabla_1 g^{(n)}(\mathbf{r}^n) = -\beta \sum_{j=2}^n \nabla_1 u(r_{1j}) g^{(n)}(\mathbf{r}^n) - \beta \int \nabla_1 u(r_{1,n+1}) g^{(n+1)}(\mathbf{r}^{n+1}) d\mathbf{r}^{n+1}. \quad (\text{A2.3.167})$$

Kirkwood derived an analogous equation that also relates two- and three-particle correlation functions but an approximation is necessary to uncouple them. The superposition approximation mentioned earlier is one such approximation, but unfortunately it is not very accurate. It is equivalent to the assumption that the potential of average force of three or more particles is pairwise additive, which is not the case even if the total potential is pair decomposable. The YBG equation for $n = 1$, however, is a convenient starting point for perturbation theories of inhomogeneous fluids in an external field.

We will describe integral equation approximations for the two-particle correlation functions. There is no single approximation that is equally good for all interatomic potentials in the 3D world, but the solutions for a few important models can be obtained analytically. These include the Percus–Yevick (PY) approximation [27, 28] for hard spheres and the mean spherical (MS) approximation for charged hard spheres, for hard spheres with point dipoles and for atoms interacting with a Yukawa potential. Numerical solutions for other approximations, such as the hypernetted chain (HNC) approximation for charged systems, are readily obtained by fast Fourier transform methods

The Ornstein–Zernike equation

$$h(r_{12}) = c(r_{12}) + \rho \int c(r_{13})h(r_{32}) \, d\mathbf{r}_3 \quad (\text{A2.3.168})$$

and the exact relation for the pair correlation function

$$g(r) = h(r) + 1 = \exp[-\beta u(r)] + h(r) - c(r) + B(r) \quad (\text{A2.3.169})$$

provide a convenient starting point for the discussion of these approximations. This equivalent to the exact relation

$$c(r) = \beta u(r) - \ln g(r) + h(r) + B(r) \quad (\text{A2.3.170})$$

for the direct correlation function. As $r \rightarrow \infty$, $c(r) \rightarrow -\beta u(r)$ except at $T = T_c$. Given the pair potential $u(r)$, we have two equations for the three unknowns $h(r)$, $c(r)$ and $B(r)$; one of these is the Ornstein–Zernike relation and the other is either one of the exact relations cited above. Each of the unknown functions has a density expansion which is the sum of integrals of products of Mayer f -functions, which motivates their approximation by considering different classes of terms. In this sense, the simplest approximation is the following.

(a) *Hypernetted chain approximation*

This sets the bridge function

$$B(r) = 0. \quad (\text{A2.3.171})$$

It is accurate for simple low valence electrolytes in aqueous solution at 25°C and for molten salts away from the critical point. The solutions are obtained numerically. A related approximation is the following.

(b) *Percus–Yevick (PY) approximation*

In this case [27, 28], the function $\exp[(h(r)-c(r))]$ in the exact relation for $g(r)$ is linearized after assuming $B(r) = 0$, when

$$g(r) \simeq \exp(-\beta u(r))[1 + h(r) - c(r)] = \exp(-\beta u(r))[g(r) - c(r)]. \quad (\text{A2.3.172})$$

Rearranging this, we have the PY approximation for the direct correlation function

$$c(r) = f(r)y(r). \quad (\text{A2.3.173})$$

This expression is combined with the Ornstein–Zernike equation to obtain the solution for $c(r)$.

For hard spheres of diameter σ , the PY approximation is equivalent to $c(r) = 0$ for $r > \sigma$ supplemented by the core condition $g(r) = 0$ for $r < \sigma$. The analytic solution to the PY approximation for hard spheres was obtained independently by Wertheim [32] and Thiele [33]. Solutions for other potentials (e.g. Lennard-Jones) are

obtained numerically.

(c) Mean spherical approximation

In the MS approximation, for hard core particles of diameter σ , one approximates the direct correlation function by

$$c(r) = -\beta u(r) \quad \text{for } r > \sigma \quad (\text{A2.3.174})$$

and supplements this with the exact relation

$$g(r) = 0 \quad \text{for } r < \sigma. \quad (\text{A2.3.175})$$

The solution determines $c(r)$ inside the hard core from which $g(r)$ outside this core is obtained via the Ornstein–Zernike relation. For hard spheres, the approximation is identical to the PY approximation. Analytic solutions have been obtained for hard spheres, charged hard spheres, dipolar hard spheres and for particles interacting with the Yukawa potential. The MS approximation for point charges (charged hard spheres in the limit of zero size) yields the Debye–Huckel limiting law distribution function.

It would appear that the approximations listed above are progressively more drastic. Their accuracy, however, is unrelated to this progression and depends on the nature of the intermolecular potential. Approximations that are good for systems with strong long-range interactions are not necessarily useful when the interactions are short ranged. For example, the HNC approximation is accurate for simple low valence electrolytes in aqueous solution in the normal preparative (0–2 M) range at 25°C, but fails near the critical region. The PY approximation, on the other hand, is poor for electrolytes, but is much better for hard spheres. The relative accuracy of these approximations is determined by the cancellation of terms in the density expansions of the correlation functions, which depends on the range of the intermolecular potential.

A2.3.5 EQUILIBRIUM PROPERTIES OF NON-IDEAL FLUIDS

A2.3.5.1 INTEGRAL EQUATION AND SCALED PARTICLE THEORIES

Theories based on the solution to integral equations for the pair correlation functions are now well developed and widely employed in numerical and analytic studies of simple fluids [6]. Further improvements for simple fluids would require better approximations for the bridge functions $B(r)$. It has been suggested that these functions can be scaled to the same functional form for different potentials. The extension of integral equation theories to molecular fluids was first accomplished by Chandler and Andersen [30] through the introduction of the site–site direct correlation function $c_{\alpha\beta}(r)$ between atoms in each molecule and a site–site Ornstein–Zernike relation called the reference interaction site

model (RISM) equation [31]. Approximations, corresponding to the closures for simple monatomic fluids, enable the site–site pair correlation functions $h_{\alpha\beta}(r)$ to be obtained. The theory has been successfully applied to simple molecules and to polymers.

Integral equation approximations for the distribution functions of simple atomic fluids are discussed in the following.

(a) *Hard spheres*

(i) *PY and MS approximations.* The two approximations are identical for hard spheres, as noted earlier. The solution yields the direct correlation function inside the hard core as a cubic polynomial:

$$c(r) = \begin{cases} -\lambda_1 - 6\lambda_2\eta(r/\sigma) - (1/2)\lambda_1\eta(r/\sigma)^3 & r < \sigma \\ = 0 & r > \sigma. \end{cases} \quad (\text{A2.3.176})$$

In this expression, the packing fraction $\eta = \pi\rho\sigma^3/6$, and the other two parameters are related to this by

$$\lambda_1 = (1 + 2\eta)^2/(1 - \eta)^4 \quad \lambda_2 = -(1 + \eta/2)^2/(1 - \eta)^4.$$

The solution was first obtained independently by Wertheim [32] and Thiele [33] using Laplace transforms. Subsequently, Baxter [34] obtained the same solutions by a Wiener–Hopf factorization technique. This method has been generalized to charged hard spheres.

The pressure from the virial equation is calculated by noting that $h(r)-c(r)$ is continuous at $r = \sigma$, and $c(r) = 0$ for $r > \sigma$. It follows that

$$h(\sigma+) - c(\sigma+) = h(\sigma-) - c(\sigma-)$$

and since $c(\sigma+) = 0$ and $h(\sigma-) = -1$, we have $g(\sigma+) = 1+h(\sigma+) = c(\sigma-)$. This gives an expression for the pressure of hard spheres in the PY approximation in terms of $c(\sigma-)$, equivalent to the virial pressure [equation A2.3.146](#)

$$\frac{P}{\rho kT} = 1 - \frac{2\pi}{3}\rho\sigma^3 c(\sigma-). \quad (\text{A2.3.177})$$

Setting $r = \sigma$ in the solution for $c(r)$, it follows that

$$\frac{P_V}{\rho kT} = \frac{1 + 2\eta + 3\eta^2}{(1 - \eta)^2} \quad (\text{A2.3.178})$$

-51-

where the subscript V denotes the pressure calculated from the virial pressure equation. The pressure from the compressibility equation follows from the expression for $(dP/d\rho)_T$ in terms of the integral of the direct correlation function $c(r)$; the upper limit of this integral is $r = \sigma$ in the PY approximation for hard spheres since $c(r) = 0$ for $r > \sigma$. One finds that the pressure P_c from the compressibility equation is given by

$$\frac{P_c}{\rho kT} = \frac{1 + \eta + \eta^2}{(1 - \eta)^3}. \quad (\text{A2.3.179})$$

The CS equation for the pressure is found to be the weighted mean of the pressure calculated from the virial and compressibility equations:

$$P_{CS} = (1/3)P_V + (2/3)P_c.$$

Figure A2.3.10 compares the virial and pressure equations for hard spheres with the pressure calculated from the CS equations and also with the pressures determined in computer simulations.

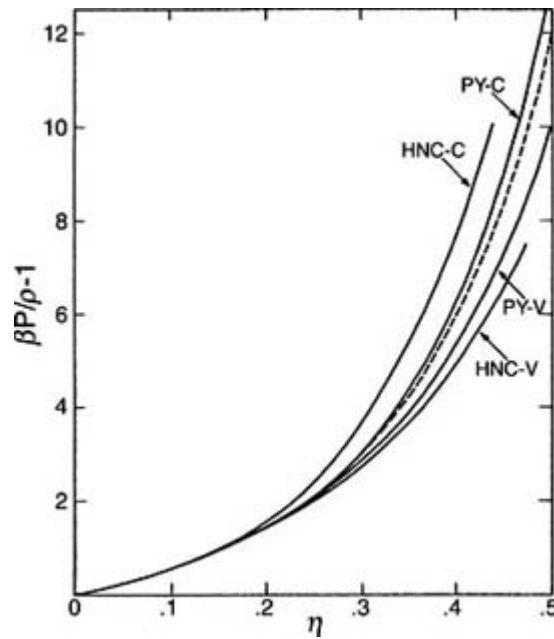


Figure A2.3.10 Equation of state for hard spheres from the PY and HNC approximations compared with the CS equation (---). C and V refer to the compressibility and virial routes to the pressure (after [6]).

The CS pressures are close to the machine calculations in the fluid phase, and are bracketed by the pressures from the virial and compressibility equations using the PY approximation. Computer simulations show a fluid–solid phase transition that is not reproduced by any of these equations of state. The theory has been extended to mixtures of hard spheres with additive diameters by Lebowitz [35], Lebowitz and Rowlinson [35], and Baxter [36].

-52-

(ii) *Scaled particle theory.* The virial equation for the pressure of hard spheres is determined by the contact value $g(\sigma^+)$ of the pair correlation functions, which is related to the average density of hard spheres in contact with a spherical cavity of radius σ , from which the spheres are excluded. The fixed cavity affects the fluid in the same way that a hard sphere at the centre of this cavity would influence the rest of the fluid. Reiss, Frisch and Lebowitz [37] developed an approximate method to calculate this, and found that the pressure for hard spheres is identical to the pressure from the compressibility equation in the PY approximation given in equation A2.3.178.

The method has been extended to mixtures of hard spheres, to hard convex molecules and to hard spherocylinders that model a nematic liquid crystal. For mixtures (m subscript) of hard convex molecules of the same shape but different sizes, Gibbons [38] has shown that the pressure is given by

$$\frac{P}{\rho kT} = \frac{1}{1 - \xi_m} + \frac{AB}{(1 - \xi_m)^2} + \frac{B^2 C}{3(1 - \xi_m)^3} \quad (\text{A2.3.180})$$

where

$$\begin{aligned}
\xi_m &= \rho \sum_i x_i V_i & A &= \sum_i x_i \bar{R}_i \\
B &= \sum_i x_i S_i & C &= \sum_i x_i \bar{R}_i^2
\end{aligned}
\tag{A2.3.181}$$

where \bar{R}_i is the radius of particle i averaged over all orientations, V_i and S_i are the volume and surface area of the particle i , respectively, and x_i is its mole fraction. The pressure corresponding to the PY compressibility equation is obtained for parameters corresponding to hard sphere mixtures. We refer the reader to the review article by Reiss in the further reading section for more detailed discussions.

(iii) *Gaussian statistics.* Chandler [39] has discussed a model for fluids in which the probability $P(N, \nu)$ of observing N particles within a molecular size volume ν is a Gaussian function of N . The moments of the probability distribution function are related to the n -particle correlation functions $g^{(n)}(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n)$, and

$$\alpha_n = \langle N(N-1) \dots N-n+1 \rangle = \rho^n \int_{\nu} \dots \int_{\nu} g^{(n)}(\mathbf{r}_1, \dots, \mathbf{r}_n) d\mathbf{r}_1 \dots d\mathbf{r}_n.$$

The inversion of this leads to an expression for $P(N, \nu)$:

$$P(N, \nu) = \sum_{n=N}^{\infty} (-1)^{n-N} \frac{\alpha_n}{N!(n-N)!}$$

involving all of the moments of the probability distribution function. The Gaussian approximation implies that only the first two moments $\langle N \rangle_{\nu}$ and $\langle N^2 \rangle_{\nu}$, which are determined by the density and the pair correlation function, are sufficient

to determine the probability distribution $P(N, \nu)$. Computer simulation studies of hard spheres by Crooks and Chandler [40] and even water by Hummer *et al* [41] have shown that the Gaussian model is accurate at moderate fluid densities; deviations for hard spheres begin to occur at very low and high densities near the ideal gas limit and close to the transition to a solid phase, respectively.

The assumption of Gaussian fluctuations gives the PY approximation for hard sphere fluids and the MS approximation on addition of an attractive potential. The RISM theory for molecular fluids can also be derived from the same model.

(b) *Strong electrolytes*

The long-range interactions between ions lie at the opposite extreme to the harsh repulsive interactions between hard spheres. The methods used to calculate the thermodynamic properties through the virial expansion cannot be directly applied to Coulombic systems since the integrals entering into the virial coefficients diverge. The correct asymptotic form of the thermodynamic properties at low concentrations was first obtained by Debye and Hückel in their classic study of charged hard spheres [42] by linearizing the Poisson–Boltzmann equation as discussed below. This immediately excludes serious consideration of ion pairing, but this defect, especially in low dielectric solvents, was taken into account by Bjerrum [43], who assumed that all oppositely charged ions within a distance $e_+ e_- / 2\epsilon kT$ were paired, while the rest were free. The free ions were treated in the Debye–Hückel approximation.

The Debye treatment is not easily extended to higher concentrations and special methods are required to

incorporate these improvements. One method, due to Mayer [44], resums the virial expansion to cancel out the divergences of the integrals. Mayer obtained the Debye–Hückel limiting law and the first correction to this as a convergent renormalized second virial coefficient that automatically incorporates the effect of ion pairing. Improvements due to Outhwaite, Bhuyian and others, involve modifications of Debye and Hückel’s original treatment of the Poisson–Boltzmann equation to yield a modified Poisson–Boltzmann (MPB) equation for the average electrostatic potential $\psi_i(r)$ of an ion. We refer to the review article by Outhwaite (1974) in the further reading section for a detailed discussion.

Two widely used theories of electrolytes at room temperature are the MS and HNC approximations for the pair correlation functions. The approximations fail or are less successful near the critical point. The solutions to the HNC approximation in the usual laboratory concentration range are obtained numerically, where fast Fourier transform methods are especially useful [45]. They are accurate for low valence electrolytes in aqueous solution at room temperature up to 1 or 2 M. However, the HNC approximation does not give a numerical solution near the critical point. The MS approximation of charged hard spheres can be solved analytically, as first shown by Waisman and Lebowitz [46]. This is very convenient and useful in mapping out the properties of electrolytes of varying charges over a wide range of concentrations. The solution has been extended recently to charged spheres of unequal size [47] and to sticky charged hard spheres [48, 49]. Ebeling [50] extended Bjerrum’s theory of association by using the law of mass action to determine the number of ion pairs while treating the free ions in the MS approximation supplemented with the second ionic virial coefficient. Ebeling and Grigoro [51] located a critical point from this theory. The critical region of electrolytes is known to be characterized by pairing and clustering of ions and it has been observed experimentally that dimers are abundant in the vapour phase of ionic fluids. The nature of the critical exponents in this region, whether they are classical or non-classical, and the possibilities of a crossover from one to the other are currently under study [52, 53, 54, 55 and 56]. Computer simulation studies of this region are also under active investigation [57, 58 and 59]. Koneshan and Rasaiah [60] have observed clusters of sodium and chloride ions in simulations of aqueous sodium chloride solutions under supercritical conditions.

-54-

Strong electrolytes are dissociated into ions that are also paired to some extent when the charges are high or the dielectric constant of the medium is low. We discuss their properties assuming that the ionized gas or solution is electrically neutral, i.e.

$$\sum_{i=1}^{\sigma} c_i e_i = 0 \quad (\text{A2.3.182})$$

where c_i is the concentration of the free ion i with charge e_i and σ is the number of ionic species. The local charge density at a distance r from the ion i is related to the ion concentrations c_j and pair distribution functions $g_{ij}(r)$ by

$$\rho_i(r) = \sum_{j=1}^{\sigma} c_j e_j g_{ij}(r). \quad (\text{A2.3.183})$$

The electroneutrality condition can be expressed in terms of the integral of the charge density by recognizing the obvious fact that the total charge around an ion is equal in magnitude and opposite in sign to the charge on the central ion. This leads to the zeroth moment condition

$$-e_i = \int \rho_i(r) \, d\mathbf{r}. \quad (\text{A2.3.184})$$

The distribution functions also satisfy a second moment condition, as first shown by Stillinger and Lovett [61]:

$$-\frac{3\varepsilon_0 kT}{2\pi} = \sum_{i=1}^{\sigma} c_i e_i \int \rho_i(r) r^2 dr \quad (\text{A2.3.185})$$

where ε_0 is the dielectric constant of the medium in which the ions are immersed. The Debye–Hückel limiting law and the HNC and MS approximations satisfy the zeroth and second moment conditions.

The thermodynamic properties are calculated from the ion–ion pair correlation functions by generalizing the expressions derived earlier for one-component systems to multicomponent ionic mixtures. For ionic solutions it is also necessary to note that the interionic potentials are solvent averaged ionic potentials of average force:

$$u_{ij}(r; T, P) = u_{ij}^*(r; T, P) + \frac{e_i e_j}{\varepsilon_0 r}. \quad (\text{A2.3.186})$$

Here $u_{ij}^*(r, T, P)$ is the short-range potential for ions, and ε_0 is the dielectric constant of the solvent. The solvent averaged potentials are thus actually free energies that are functions of temperature and pressure. The thermodynamic properties calculated from the pair correlation functions are summarized below.

-55-

(i) The virial equation provides the osmotic coefficient measured in isopiestic experiments:

$$\phi_v = 1 - \frac{1}{6ckT} \sum_{i=1}^{\sigma} \sum_{j=1}^{\sigma} c_i c_j \int r \frac{\partial u_{ij}}{\partial r} g_{ij}(r) dr. \quad (\text{A2.3.187})$$

(ii) The generalization of the compressibility equation, taking into account electroneutrality,

$$\frac{\partial \ln \gamma_{\pm}}{\partial \ln c} = \frac{1}{cG_{\pm}} - 1 \quad (\text{A2.3.188})$$

where

$$G_{\pm} = \int (g_{\pm} - 1) dr \quad (\text{A2.3.189})$$

provides the concentration dependence of the mean activity coefficient γ determined experimentally from cell EMFs.

(iii) The energy equation is related to the heat of dilution determined from calorimetric measurements

$$E^{\text{ex}} = \frac{1}{2} \sum_{i=1}^{\sigma} \sum_{j=1}^{\sigma} c_i c_j \int \frac{\partial [\beta u_{ij}(r)]}{\partial \beta} g_{ij}(r) dr. \quad (\text{A2.3.190})$$

For an ionic solution

$$\frac{\partial[\beta u_{ij}(r)]}{\partial r} = \frac{e_i e_j}{\epsilon_0 r} \left[1 + \frac{\partial \ln \epsilon_0}{\partial \ln T} \right] + \frac{\partial[\beta u_{ij}^*(r)]}{\partial r} \quad (\text{A2.3.191})$$

and $d \ln \epsilon_0 / d \ln T = -1.3679$ for water at 25°C.

(iv) The equation for the excess volume is related to the partial molar volumes of the solute determined from density measurements

$$V^{\text{ex}} = \frac{1}{2} \sum_{i=1}^{\sigma} \sum_{j=1}^{\sigma} c_i c_j \int \frac{\partial[\beta u_{ij}(r)]}{\partial P} g_{ij}(r) dr. \quad (\text{A2.3.192})$$

-56-

In an ionic solution

$$\frac{\partial[\beta u_{ij}(r)]}{\partial P} = \frac{e_i e_j}{\epsilon_0 r} \left[\frac{\partial \ln \epsilon_0}{\partial \ln P} \right] + \frac{\partial[\beta u_{ij}^*(r)]}{\partial P} \quad (\text{A2.3.193})$$

where $d \ln \epsilon_0 / d \ln P_0 = 47.1 \times 10^{-6}$ for water at 25°C.

The theory of strong electrolytes due to Debye and Hückel derives the exact limiting laws for low valence electrolytes and introduces the idea that the Coulomb interactions between ions are screened at finite ion concentrations.

(c) *The Debye–Hückel theory*

The model used is the RPM. The average electrostatic potential $\psi_i(r)$ at a distance r away from an ion i is related to the charge density $\rho_i(r)$ by Poisson's equation

$$\nabla^2 \psi_i(r) = -\frac{4\pi \rho_i(r)}{\epsilon_0} = -\frac{4\pi}{\epsilon_0} \sum_{j=1}^{\sigma} c_j e_j g_{ij}. \quad (\text{A2.3.194})$$

Debye and Hückel [42] assumed that the ion distribution functions are related to $\psi_i(r)$ by

$$g_{ij}(r) = \exp(-\beta e_j \psi_i(r))$$

which is an approximation. This leads to the PB equation

$$\nabla^2 \psi_i(r) = \begin{cases} -\frac{4\pi}{\epsilon_0} \sum_{j=1}^{\sigma} c_j e_j \exp(-\beta e_j \psi_i(r)) & r > \sigma \\ 0 & r < \sigma. \end{cases} \quad (\text{A2.3.195})$$

Linearizing the exponential,

$$g_{ij}(r) = 1 - \beta e_j \psi_i(r) \quad (\text{A2.3.196})$$

in the PB equation leads to the Debye–Hückel differential equation:

$$\nabla^2 \psi_i(r) = \begin{cases} \kappa^2(\psi_i(r)) & r > \sigma \\ 0 & r < \sigma \end{cases} \quad (\text{A2.3.197})$$

-57-

where κ is defined by

$$\kappa^2 = \frac{4\pi}{\epsilon_0 kT} \sum_{j=1}^{\sigma} c_j e_j^2. \quad (\text{A2.3.198})$$

The solution to this differential equation is

$$g_{ij}(r) = \begin{cases} 1 - \frac{e_i e_j}{\epsilon_0 kT r} \frac{\exp(-\kappa(r - \sigma))}{(1 + \kappa\sigma)} & r > \sigma \\ 0 & r < \sigma \end{cases} \quad (\text{A2.3.199})$$

which obeys the zeroth moment or electroneutrality condition, but not the second moment condition.

The mean activity coefficient γ_{\pm} of a single electrolyte in this approximation is given by

$$\ln \gamma_{\pm} = -\frac{A|z_+ z_-| \sqrt{I}}{1 + Ba\sqrt{I}} \quad (\text{A2.3.200})$$

where a is the effective distance of closest approach of the ions, and A and B are constants determined by the temperature T and the dielectric constant of the solvent ϵ_0 . This expression is widely used to calculate the activity coefficients of simple electrolytes in the usual preparative range. The contributions of the hard cores to non-ideal behaviour are ignored in this approximation.

When $\kappa\sigma \ll 1$ (i.e. at very low concentrations), we have the Debye–Hückel limiting law distribution function:

$$g_{ij}(r) = \begin{cases} 1 - \beta e_i e_j \exp(-\kappa r) / \epsilon_0 r & (r > \sigma) \\ = 0 & (r < \sigma) \end{cases} \quad (\text{A2.3.201})$$

which satisfies both the zeroth and second moment conditions. It also has an interesting physical interpretation. The total charge $P_i(r) dr$ in a shell of radius r and thickness dr around an ion is

$$P_i(r) dr = \rho_i(r) 4\pi r^2 dr = -\kappa^2 e_i r \exp(-\kappa r) dr \quad (\text{A2.3.202})$$

which has a maximum at a distance $r = 1/\kappa$, which is called the Debye length or the radius of the ‘ionic atmosphere’. Each ion is pictured as surrounded by a cloud or ‘ionic atmosphere’ whose net charge is opposite in sign to the central ion. The cloud charge $P_i(r)$ has a maximum at $r = 1/\kappa$. The limiting law distribution function implies that the electrostatic potential

$$\psi_i(r) = e_i \exp(-\kappa r) / \epsilon_0 r. \quad (\text{A2.3.203})$$

Expanding the exponential, one finds for small κr that

$$\psi_i(r) = \frac{e_i}{\epsilon_0 r} - \frac{e_i}{\epsilon_0(1/\kappa)}. \quad (\text{A2.3.204})$$

The first term is the Coulomb field of the ion, and the second is the potential due to the ion atmosphere at an effective distance equal to $1/\kappa$. For a univalent aqueous electrolyte at 298 K,

$$1/\kappa = 3.043/\sqrt{C} \text{ \AA}$$

where C is the total electrolyte concentration in moles per litre.

The thermodynamic properties derived from the limiting law distribution functions are

$$\frac{E^{\text{ex}}}{NkT} = \frac{\kappa^3}{8\pi c} \left[1 + \frac{\partial \ln \epsilon_0}{\partial \ln T} \right] \quad (\text{A2.3.205})$$

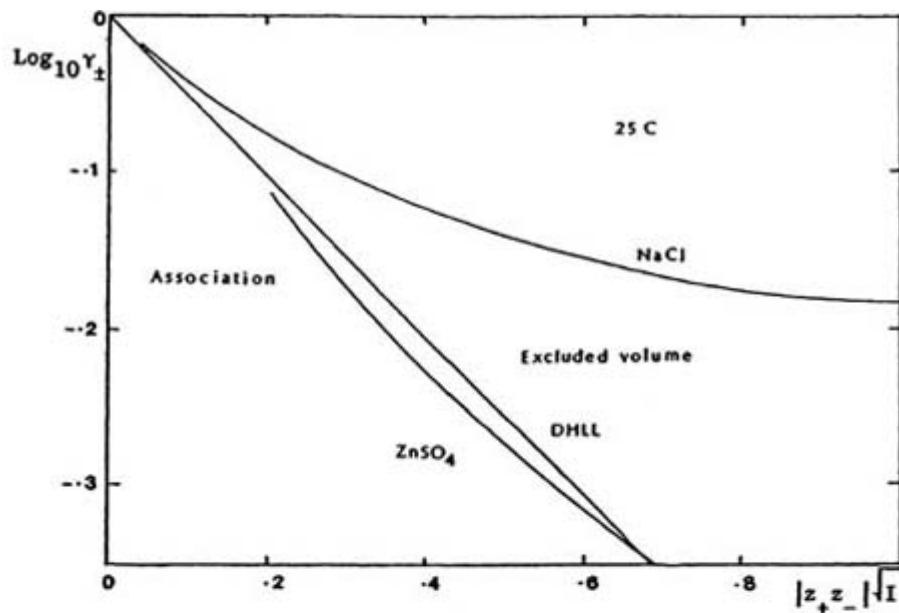
$$\ln \gamma_{\pm} = \ln \gamma^{\text{HS}} - \frac{\kappa^3}{8\pi c} \quad (\text{A2.3.206})$$

$$\phi = \phi^{\text{HS}} - \frac{\kappa^3}{24\pi c} \quad (\text{A2.3.207})$$

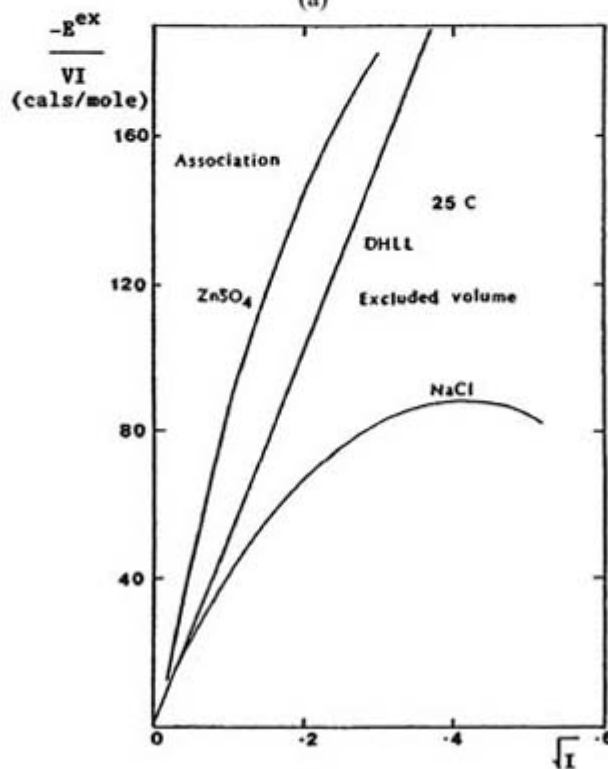
$$\frac{A^{\text{ex}}}{NkT} = \frac{A^{\text{ex,HS}}}{NkT} - \frac{\kappa^3}{24\pi c} \quad (\text{A2.3.208})$$

where $c = \sum c_i$ is the total ionic concentration and the superscript HS refers to the properties of the corresponding uncharged hard sphere system. Debye and Hückel assumed ideal behaviour for the uncharged system ($\phi^{\text{HS}} = \gamma^{\text{HS}} = 1$ and $A^{\text{ex,HS}} = 0$).

The Debye–Hückel limiting law predicts a square-root dependence on the ionic strength $I = 1/2 \sum c_i z_i^2$ of the logarithm of the mean activity coefficient ($\log \gamma_{\pm}$), the heat of dilution (E^{ex}/VT) and the excess volume (V^{ex}); it is considered to be an exact expression for the behaviour of an electrolyte at infinite dilution. Some experimental results for the activity coefficients and heats of dilution are shown in [figure A2.3.11](#) for aqueous solutions of NaCl and ZnSO₄ at 25°C; the results are typical of the observations for 1–1 (e.g. NaCl) and 2–2 (e.g. ZnSO₄) aqueous electrolyte solutions at this temperature.



(a)



(b)

Figure A2.3.11 The mean activity coefficients and heats of dilution of NaCl and ZnSO₄ in aqueous solution at 25°C as a function of $|z_+ z_-| \sqrt{I}$, where I is the ionic strength. DHLL = Debye–Hückel limiting law.

The thermodynamic properties approach the limiting law at infinite dilution, but deviate from it at low concentrations, in different ways for the two charge types. Evidence from the ionic conductivity of 2–2 electrolyte solutions suggests that the negative deviations from the limiting law observed for these solutions are due to ion pairing or association. The opposite behaviour found for aqueous 1–1 electrolytes, for which ion pairing is negligible at room temperature, is caused by the finite size of the ions and is the excluded volume effect. The Debye–Hückel theory ignores ion association and treats the effect of the sizes of the ions incompletely. The limiting law slopes and deviations from them depend strongly on the temperature and

dielectric constant of the solvent and on the charges on the ions. An aqueous solution of sodium chloride, for instance, behaves like a weak electrolyte near the critical temperature of water because the dielectric constant of the solvent decreases rapidly with increasing temperature.

As pointed out earlier, the contributions of the hard cores to the thermodynamic properties of the solution at high concentrations are not negligible. Using the CS equation of state, the osmotic coefficient of an uncharged hard sphere solute (in a continuum solvent) is given by

$$\phi^{\text{HS}} = 1 + \frac{4\eta - 2\eta^2}{(1 - \eta^3)} \quad (\text{A2.3.209})$$

where $\eta = c\sigma^3/6$. For a 1 M solution this contributes 0.03 to the deviation of the osmotic coefficient from ideal behaviour.

(d) Mayer's theory

The problem with the virial expansion when applied to ionic solutions is that the virial coefficients diverge. This difficulty was resolved by Mayer who showed how the series could be resummed to cancel the divergencies and yield a new expansion for a charged system. The terms in the new series are ordered differently from those in the original expansion, and the Debye-Hückel limiting law follows as the leading correction due to the non-ideal behaviour of the corresponding uncharged system. In principle, the theory enables systematic corrections to the limiting law to be obtained as at higher electrolyte concentrations. The results are quite general and are applicable to any electrolyte with a well defined short-range potential $u_{ij}^*(r)$, besides the RPM electrolyte.

The principle ideas and main results of the theory at the level of the second virial coefficient are presented below. The Mayer f -function for the solute pair potential can be written as the sum of terms:

$$f_{ij}(r) = f_{ij}^*(r) + (1 + f_{ij}^*(r)) \sum_{n=1}^{\infty} \frac{1}{n!} (-\beta e_i e_j / \epsilon_0 r)^n \quad (\text{A2.3.210})$$

where $f_{ij}^*(r)$ is the corresponding Mayer f -function for the short-range potential $u_{ij}^*(r)$ which we represent graphically as $i \circ \text{---} \circ_j$ and $\beta = 1/kT$. Then the above expansion can be represented graphically as

-61-

$$i \circ \text{---} \circ_j = i \circ \text{---} \text{---} \text{---} \circ_j + i \circ \text{~~~~~} \circ_j + i \circ \text{~~~~~} \circ_j + \dots$$

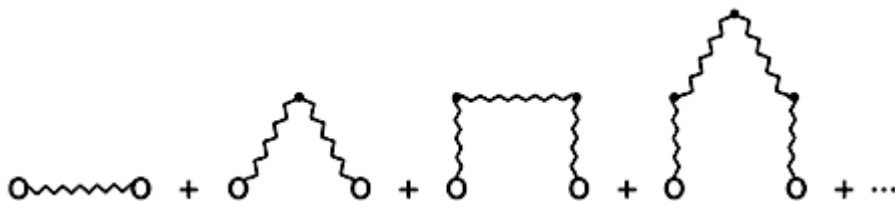
$$+ i \circ \text{---} \text{---} \text{---} \circ_j + i \circ \text{---} \text{---} \text{---} \circ_j + \dots \quad (\text{A2.3.211})$$

$i \circ \text{---} \circ_j$ represents $f_{ij}(r)$, the Mayer f -function for the pair potential $u_{ij}(r)$, and $i \circ \text{~~~~~} \circ_j$ represents the Coulomb potential multiplied by $-\beta$. The graphical representation of the virial coefficients in terms of Mayer f -bonds can now be replaced by an expansion in terms of f^* bonds ($i \circ \text{---} \circ_j$) and Coulomb bonds ($i \circ \text{~~~~~} \circ_j$). Each f -bond is replaced by an f^* -bond and the sum of one or more Coulomb bonds in parallel with or without an f^* -bond in parallel. The virial coefficients then have the following graphical representation:

$$\begin{aligned}
 i \bullet \text{---} \bullet j &= i \bullet \text{---} \bullet j + i \bullet \text{---} \bullet j + i \bullet \text{---} \bullet j + i \bullet \text{---} \bullet j + \dots \\
 \triangle &= \triangle + \triangle + \dots \\
 \square &= \square + \square + \dots \\
 \vdots & \\
 \vdots & \\
 \hline
 \frac{A^{\text{ex}}}{NkT} &= \frac{A^{\text{ex},0}}{NkT} - \frac{\kappa^3}{12\pi\epsilon} + \dots \text{HT}
 \end{aligned}
 \tag{A2.3.212}$$

where HT stands for higher-order terms. There is a symmetry number associated with each graph which we do not need to consider explicitly in this discussion. Each black circle denotes summation over the concentration c_i and integration over the coordinates of species i . The sum over all graphs in which the f -bond is replaced by an f^* -bond gives the free energy $A^{\text{ex}*}$ of the corresponding uncharged system. The effect of the Coulomb potential on the expansion is more complicated because of its long range. The second term in the expansion of the second virial coefficient is the bare Coulomb bond multiplied by $-\beta$. If we multiply this by a screening function and carry out the integration the result is finite, but it contributes nothing to the overall free energy because of electroneutrality. This is because the contribution of the charge e_i from the single Coulomb bond at a vertex when multiplied by c_i and summed over i is zero. The result for a graph with a cycle of Coulomb bonds, however, is finite. Each vertex in these graphs has two Coulomb bonds leading into it and instead of $c_i e_i$, we have $\sum c_i e_i^2$ (which appears as a factor in the definition of κ^2). This is not zero unless the ion concentration is also zero. Mayer summed all graphs with cycles of Coulomb bonds

and found that this leads to the Debye–Hückel limiting law expression for the excess free energy! The essential mechanism behind this astonishing result is that the long-range nature of the Coulomb interaction requires that the ions be considered collectively rather than in pairs, triplets etc, which is implied by the conventional virial expansion. The same mechanism is also responsible for the modification of the interaction between two charges by the presence of others, which is called ‘screening’. The sum of all chains of Coulomb bonds between two ions represents the direct interaction as well as the sum of indirect interactions (of the longest range) through other ions. The latter is a subset of the graphs which contribute to the correlation function $h_{ij}(r) = g_{ij}(r) - 1$ and has the graphical representation



Explicit calculation of this sum shows that it is the Debye screened potential

$$q_{ij}(r) = -\beta e_i e_j \exp(-\kappa r) / \epsilon_0 r. \quad (\text{A2.3.213})$$

Going beyond the limiting law it is found that the modified (or renormalized) virial coefficients in Mayer's theory of electrolytes are functions of the concentration through their dependence on κ . The ionic second virial coefficient $B_2(\kappa)$ is given by [62]

$$B_2(\kappa) = -\frac{1}{2} \sum_{i=1}^{\sigma} \sum_{j=1}^{\sigma} c_i c_j \int [(\exp(-\beta u_{ij}^*(r) + q_{ij}(r)) - 1 - q_{ij}(r) - q_{ij}(r)^2/2)] dr. \quad (\text{A2.3.214})$$

This expression contains the contribution of the short-range potential included earlier in A^{ex} , so that the excess free energy, to this level of approximation, is

$$\left(\frac{A^{\text{ex}}}{NkT} \right)_{\text{DHLL}+B_2} = -\frac{\kappa^3}{12\pi c} + \frac{B_2(\kappa)}{c}. \quad (\text{A2.3.215})$$

This is called the DHLL+ B_2 approximation. On carrying out the integrations over $q_{ij}(r)$ and $q_{ij}(r)^2/2$ and using the electroneutrality condition, this can be rewritten as [63]

$$\left(\frac{A^{\text{ex}}}{NkT} \right)_{\text{DHLL}+B_2} = -\frac{5\kappa^3}{96\pi c} + \frac{S_2(\kappa)}{c} \quad (\text{A2.3.216})$$

-63-

where

$$S_2(\kappa) = -\frac{1}{2} \sum_{i=1}^{\sigma} \sum_{j=1}^{\sigma} c_i c_j \int [(\exp(-\beta u_{ij}^*(r) + q_{ij}(r)) - 1)] dr. \quad (\text{A2.3.217})$$

This has the form of a second virial coefficient in which the Debye screened potential has replaced the Coulomb potential. Expressions for the other excess thermodynamic properties are easily derived.

Mayer's theory is formally exact within the radius of convergence of the virial series and it predicts the properties characteristic of all charge types without the need to introduce any additional assumptions. Unfortunately, the difficulty in calculating the higher virial coefficients limits the range of concentrations to which the theory can be applied with precision. The DHLL+ B_2 approximation is qualitatively correct in reproducing the association effects observed at low concentrations for higher valence electrolytes and the excluded volume effects observed for all electrolytes at higher concentrations.

(e) The MS approximation

The MS approximation for the RPM, i.e. charged hard spheres of the same size in a continuum dielectric, was solved by Waisman and Lebowitz [46] using Laplace transforms. The solutions can also be obtained [47] by an extension of Baxter's method to solve the PY approximation for hard spheres and sticky hard spheres. The method can be further extended to solve the MS approximation for unsymmetrical electrolytes (with hard cores of unequal size) and weak electrolytes, in which chemical bonding is mimicked by a delta function interaction. We discuss the solution to the MS approximation for the symmetrically charged RPM electrolyte.

For the RPM of an electrolyte the MS approximation is

$$c_{ij}(r) = -\beta u_{ij}(r) = -e_i e_j / \epsilon_0 r \quad \text{for } r > \sigma \quad (\text{A2.3.218})$$

with the exact relation

$$h_{ij}(r) = g_{ij}(r) - 1 = -1 \quad \text{for } r < \sigma. \quad (\text{A2.3.219})$$

The generalization of the Ornstein–Zernike equation to a mixture is

$$h_{ij}(r_{12}) = c_{ij}(r_{12}) + \sum_{k=1}^{\sigma} \rho_k \int c_{ik}(r_{13}) h_{kj}(r_{32}) dr_3 \quad (\text{A2.3.220})$$

where i and j refer to the ionic species (positive and negative ions), ρ_i is the concentration (or number density) of the i th species and σ is the number of ionic species. Taking Fourier transforms and using the convolution theorem puts this in matrix form

-64-

$$\tilde{\mathbf{H}} = \tilde{\mathbf{C}} + \tilde{\mathbf{H}}\tilde{\mathbf{P}}\tilde{\mathbf{C}} \quad (\text{A2.3.221})$$

where \mathbf{H} and \mathbf{C} are matrices whose elements are the Fourier transforms of h_{ij} and c_{ij} , and \mathbf{P} is a diagonal matrix whose elements are the concentrations ρ_i of the ions. The correlation function matrix is symmetric since $c_{+-} = c_{-+}$ and $h_{+-} = h_{-+}$. The RPM symmetrically charged electrolyte has the additional simplification

$$\begin{aligned} |e_+| &= |e_-| = e \\ \rho_+ &= \rho_- = \rho/2 \end{aligned} \quad (\text{A2.3.222})$$

and

$$c_{++} = c_{--} \quad h_{++} = h_{--} \quad (\text{A2.3.223})$$

where e is the magnitude of the charge and ρ is the total ion concentration. Defining the sum and difference functions

$$F_s = (F_+ + F_-)/2 \quad \text{and} \quad F_D = (F_+ - F_-)/2 \quad (\text{A2.3.224})$$

of the direct and indirect correlation functions c_{ij} and h_{ij} , the Ornstein–Zernike equation separates into two equations

$$h_s = c_s + \rho c_s * h_s \quad (\text{A2.3.225})$$

$$h_D = c_D - \rho c_D * h_D \quad (\text{A2.3.226})$$

where* stands for a convolution integral and the core condition is replaced by

$$h_s = -1 \quad h_D = 0 \quad \text{for } 0 < r < \sigma. \quad (\text{A2.3.227})$$

The MS solution for c_s turns out to be identical to the MS (or PY) approximation for hard spheres of diameter σ ; it is a cubic polynomial in r/σ . The solution for c_D is given by

$$c_D = \begin{cases} \frac{\beta e^2}{\epsilon_0 k T \sigma} \left[2B - B^2 \left(\frac{r}{\sigma} \right) \right] & 0 < r < \sigma \\ \frac{\beta e^2}{\epsilon_0 k T r} & r > \sigma \end{cases} \quad (\text{A2.3.228})$$

-65-

where

$$B = \frac{[(1+x) - (1+2x)^{1/2}]}{x} \quad (\text{A2.3.229})$$

and $x = \kappa\sigma$. The excess energy of a fully dissociated strong electrolyte in the MSA approximation is

$$\frac{E^{\text{ex}}}{NkT} = \frac{-x[(1+x) - (1+2x)^{1/2}]}{4\pi c\sigma^3}. \quad (\text{A2.3.230})$$

Integration with respect to β , from $\beta = 0$ to finite β , leads to the excess Helmholtz free energy:

$$\frac{A^{\text{ex}} - A^{\text{ex,HS}}}{NkT} = -\frac{[6x + 3x^2 + 2 - 2(1+2x)^{3/2}]}{12\pi\rho\sigma^3} \quad (\text{A2.3.231})$$

where $A^{\text{ex,HS}}$ is the excess free energy of hard spheres. The osmotic coefficient follows from this and is given by

$$\phi^E = \phi^{\text{HS}} + \frac{[3x + 3x(1+2x)^{1/2} - 2(1+2x)^{3/2} + 2]}{12\pi\rho\sigma^3} \quad (\text{A2.3.232})$$

where ϕ^{HS} is the osmotic coefficient of the uncharged hard spheres of diameter σ in the MS or PY approximation. The excess Helmholtz free energy is related to the mean activity coefficient γ_{\pm} by

$$A^{\text{ex}} = NkT[\ln \gamma_{\pm} + (1 - \phi)] \quad (\text{A2.3.233})$$

and the activity coefficient from the energy equation, calculated from ϕ^E and $A^{\text{ex,E}}$ is given by

$$\ln \gamma_{\pm}^E = \ln \gamma^{\text{HS}} + \frac{-x[(1+x) - (1+2x)^{1/2}]}{4\pi c\sigma^3}. \quad (\text{A2.3.234})$$

The second term on the right is $\beta E^{\text{ex}}/NkT$. This is true for any theory that predicts $\beta(A^{\text{ex}} - A^{\text{HS}})$ as a function of $x = \kappa\sigma$ only, which is the case for the MS approximation.

The thermodynamic properties calculated by different routes are different, since the MS solution is an approximation. The osmotic coefficient from the virial pressure, compressibility and energy equations are not the same. Of these, the energy equation is the most accurate by comparison with computer simulations of Card and Valleau [63]. The osmotic coefficients from the virial and compressibility equations are

$$\phi_V = \phi^{HS} + \frac{x^2 B}{12\pi\rho\sigma^3} \quad (\text{A2.3.235})$$

$$\phi_C = \phi^{HS}. \quad (\text{A2.3.236})$$

In the limit of zero ion size, i.e. as $\sigma \rightarrow 0$, the distribution functions and thermodynamic functions in the MS approximation become identical to the Debye–Hückel limiting law.

(f) The HNC approximation

The solutions to this approximation are obtained numerically. Fast Fourier transform methods and a reformulation of the HNC (and other integral equation approximations) in terms of the screened Coulomb potential by Allnatt [64] are especially useful in the numerical solution. Figure A2.3.12 compares the osmotic coefficient of a 1–1 RPM electrolyte at 25°C with each of the available Monte Carlo calculations of Card and Valleau [63].

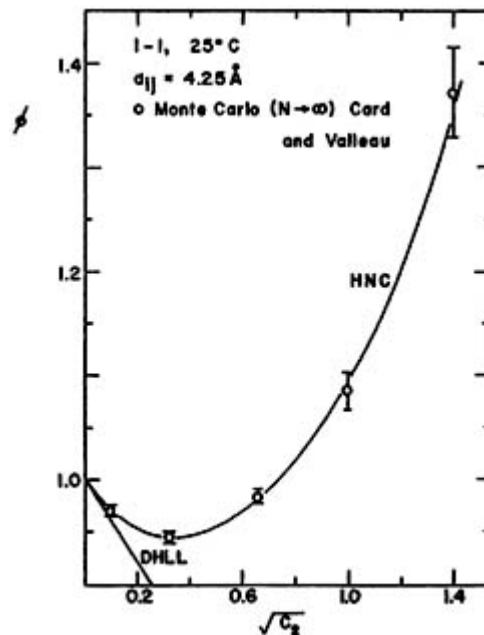


Figure A2.3.12 The osmotic coefficient of a 1–1 RPM electrolyte compared with the Monte Carlo results of [63].

The agreement is excellent up to a 1 molar concentration. The excess energies for 1–1, 2–1, 2–2 and 3–1 charge types calculated from the MS and HNC approximations are shown in figure A2.3.13. The Monte Carlo

results for 2-2 and 3-1 electrolytes are also shown in the same figure. The agreement is good, even for the energies of the higher valence electrolytes. However, as illustrated in [figure A2.3.14](#) the HNC and MS approximations deteriorate in accuracy as the charges on the ions are increased [67].

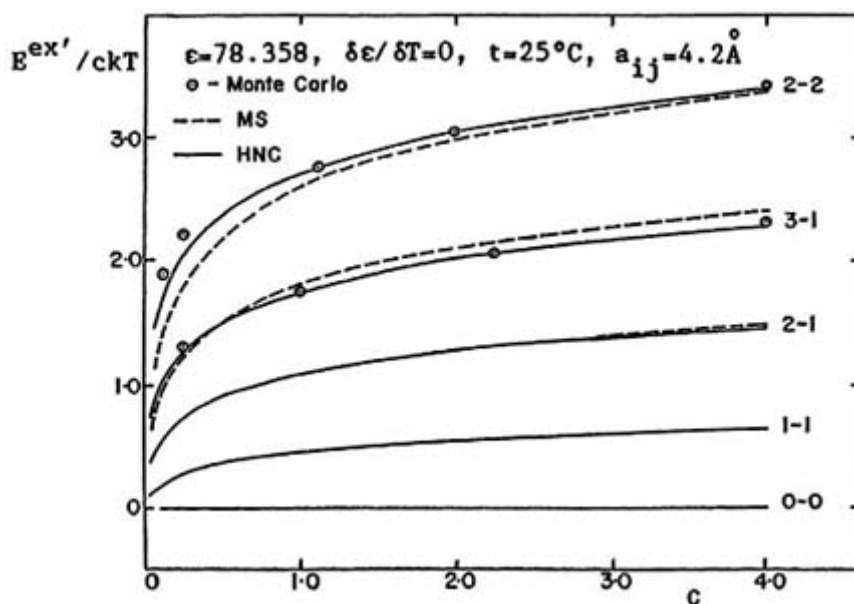


Figure A2.3.13 The excess energy of 1-1, 2-1, 3-1 and 2-2 RPM electrolytes in water at 25°C. The full and dashed curves are from the HNC and MS approximations, respectively. The Monte Carlo results of Card and Valleau [63] for the 1-3 and 2-2 charge types are also shown.

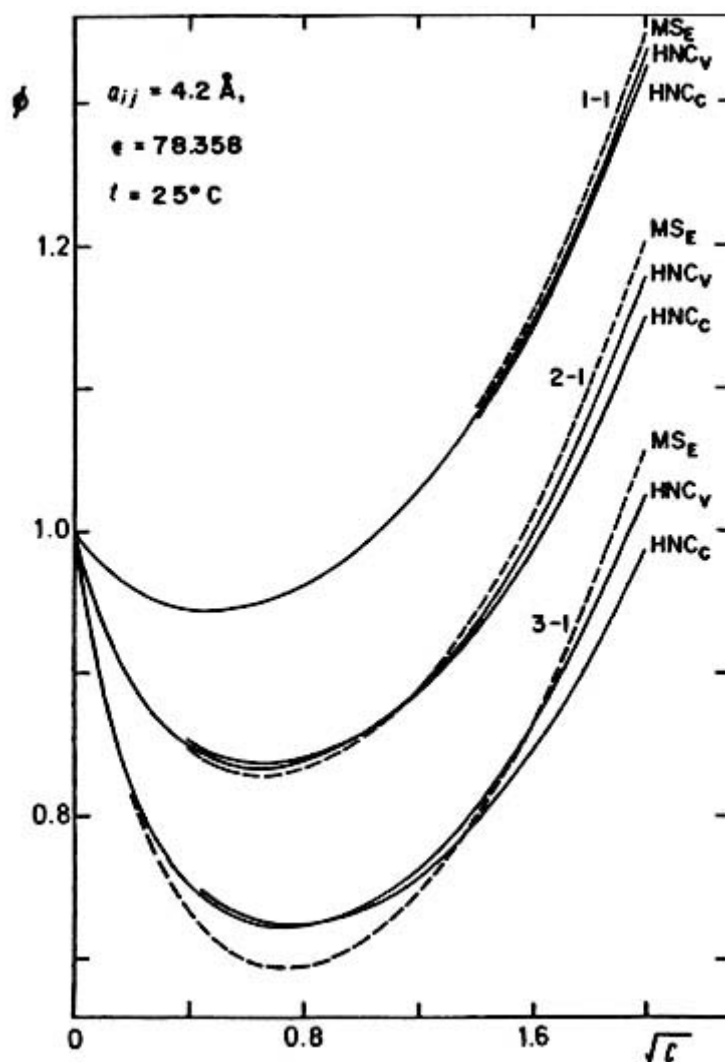


Figure A2.3.14 Osmotic coefficients for 1-1, 2-1 and 3-1 RPM electrolytes according to the MS and HNC approximations.

The osmotic coefficients from the HNC approximation were calculated from the virial and compressibility equations; the discrepancy between ϕ_V and ϕ_C is a measure of the accuracy of the approximation. The osmotic coefficients calculated via the energy equation in the MS approximation are comparable in accuracy to the HNC approximation for low valence electrolytes. [Figure A2.3.15](#) shows deviations from the Debye-Hückel limiting law for the energy and osmotic coefficient of a 2-2 RPM electrolyte according to several theories. The negative deviations from the limiting law are reproduced by the HNC and DHLL + B_2 equations but not by the MS approximation.

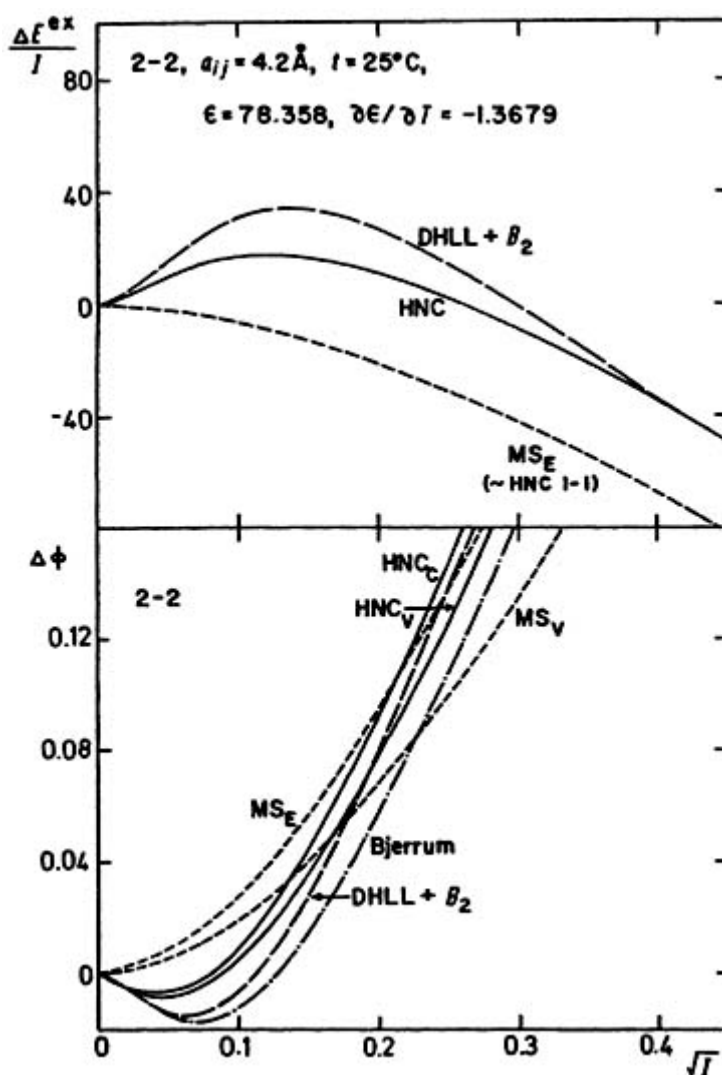


Figure A2.3.15 Deviations (Δ) of the heat of dilution E^{ex}/I and the osmotic coefficient ϕ from the Debye–Hückel limiting law for 1–1 and 2–2 RPM electrolytes according to the DHLL + B_2 , HNC and MS approximations.

In [figure A2.3.16](#) the theoretical HNC osmotic coefficients for a range of ion size parameters in the primitive model are compared with experimental data for the osmotic coefficients of several 1–1 electrolytes at 25°C. Choosing $a_{+-} = r_+ + r_-$ to fit the data at low concentrations, it is found that the calculated osmotic coefficients are too large at the higher concentrations. On choosing a_{+-} to be the sum of the Pauling radii of the ions, and a short-range potential given by a square well or mound d_{ij} equal to the width of a water molecule (2.76Å), it is found that the osmotic coefficients can be fitted to the accuracy shown in [figure A2.3.17](#) [65]. There are other models for the short-range potential which produce comparable fits for the osmotic coefficients showing that the square well approximation is by no means unique [66].

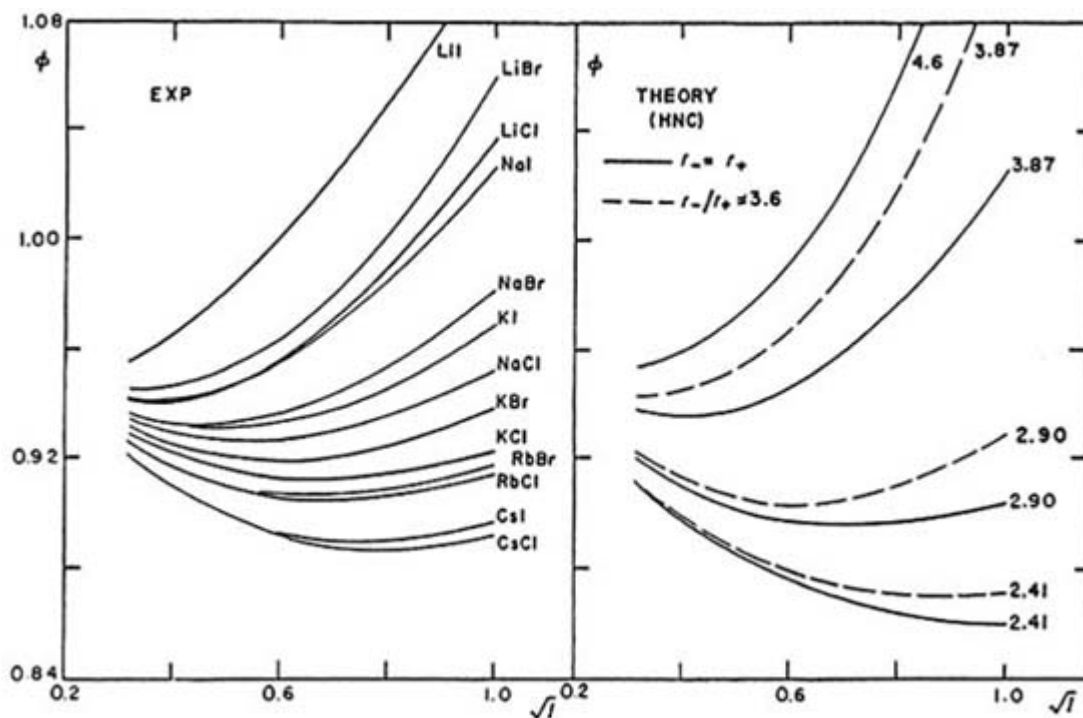


Figure A2.3.16. Theoretical HNC osmotic coefficients for a range of ion size parameters in the primitive model compared with experimental data for the osmotic coefficients of several 1-1 electrolytes at 25°C. The curves are labelled according to the assumed value of $a_{+-} = r_+ + r_-$

-71-

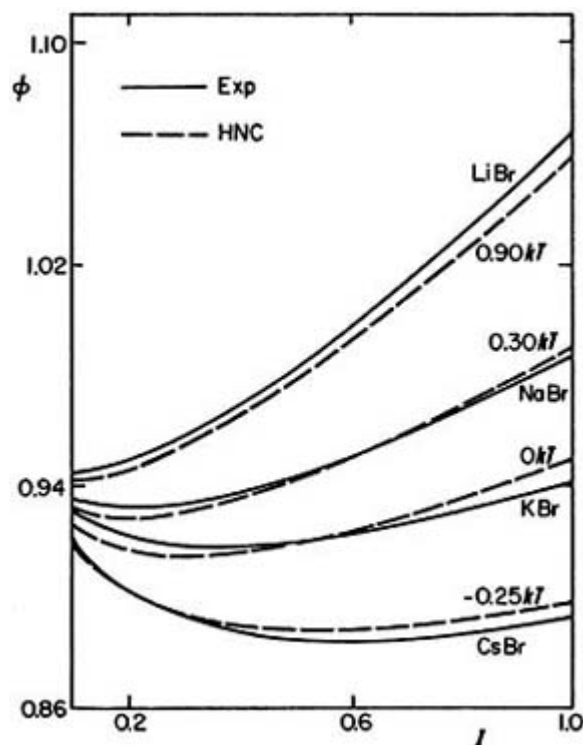


Figure A2.3.17 Theoretical (HNC) calculations of the osmotic coefficients for the square well model of an electrolyte compared with experimental data for aqueous solutions at 25°C. The parameters for this model are $a_{+-} = r_+ (\text{Pauling}) + r_- (\text{Pauling})$, $d_{++} = d_- = 0$ and d_{+-} as indicated in the figure.

A2.3.5.2 WEAK ELECTROLYTES

In a weak electrolyte (e.g. an aqueous solution of acetic acid) the solute molecules AB are incompletely dissociated into ions A^+ and B^- according to the familiar chemical equation



The forces binding the atoms in AB are chemical in nature and must be introduced, at least approximately, in the Hamiltonian in a theoretical treatment of this problem. The binding between A and B in the dimer AB is quite distinct from the formation of ion pairs in higher valence electrolytes (e.g. aqueous solutions of $ZnSO_4$ at room temperature) where the Coulomb interactions between the ions lead to ion pairs which account for the anomalous conductance and activity coefficients at low concentration. The greater shielding of the ion charges with increasing electrolyte concentration would induce the ion pairs to dissociate as the concentration rises, whereas the dimer population produced by the chemical bonding represented in the above chemical reaction would increase with the concentration of the solution.

-72-

Weak electrolytes in which dimerization (as opposed to ion pairing) is the result of chemical bonding between oppositely charged ions have been studied using a sticky electrolyte model (SEM). In this model, a delta function interaction is introduced in the Mayer f -function for the oppositely charged ions at a distance $L = \sigma$, where σ is the hard sphere diameter. The delta function mimics bonding and the Mayer f -function

$$f_{+-} = -1 + L\zeta\delta(r - L)/12 \quad r \leq \sigma \quad (\text{A2.3.238})$$

where ζ is the sticking coefficient. This induces a delta function in the correlation function $h_{+-}(r)$ for oppositely charged ions with a different coefficient λ :

$$h_{+-} = -1 + L\lambda\delta(r - L)/12 \quad r \leq \sigma. \quad (\text{A2.3.239})$$

The interaction between ions of the same sign is assumed to be a pure hard sphere repulsion for $r \leq \sigma$. It follows from simple steric considerations that an exact solution will predict dimerization only if $L < \sigma/2$, but polymerization may occur for $\sigma/2 < L = \sigma$. However, an approximate solution may not reveal the full extent of polymerization that occurs in a more accurate or exact theory. Cummings and Stell [69] used the model to study chemical association of uncharged atoms. It is closely related to the model for adhesive hard spheres studied by Baxter [70].

The association 'constant' K defined by $K = \rho_{AB}/\rho_+\rho_-$ is

$$K = \frac{\pi\lambda(L/\sigma)^3}{3(1 - \langle N \rangle)^2} \quad (\text{A2.3.240})$$

where the average number of dimers $\langle N \rangle = \eta\lambda(L/\sigma)^3$ and $\eta = \pi\rho\sigma^3/6$, in which ρ is the total ionic density. We can now distinguish three different cases:

$\lambda = 0$	no dimers	strong electrolyte (RPM)
$\lambda = (\sigma/L)^3/\eta$	all dimers if $L < \sigma/2$	dipolar dumb-bells
$0 < \lambda < (\sigma/L)^3/\eta$	ions + dimers	weak electrolyte (SEM).

Either the same or different approximations may be used to treat the binding at $r = L$ and the remaining electrical interactions between the ions. The excess energy of the sticky electrolyte is given by

$$\frac{E^{\text{ex}}}{NkT} = \frac{\langle N \rangle}{2} \frac{\partial \ln \zeta}{\partial \beta} - \left(1 + \frac{\partial \ln \epsilon_0}{\partial \ln T} \right) \frac{\kappa H}{2} \quad (\text{A2.3.241})$$

where

$$H = \kappa \int_{\sigma}^{\infty} h_{\text{D}}(r) r \, dr \quad (\text{A2.3.242})$$

-73-

and $h_{\text{D}}(r) = [h_{+-}(r) - h_{++}(r)]/2$. The first term is the binding energy and the second is the energy due to the interactions between the charges which can be determined analytically in the MS approximation and numerically in the HNC approximation. For any integer $n = \sigma/L$, $H' = H/\sigma$ in the MS approximation has the form [48]

$$H' = \frac{(a_1 + a_2 x) - (a_1^2 + 2x a_3)^{1/2}}{24 a_4 \eta} \quad (\text{A2.3.243})$$

where a_i ($i = 1$ to 4) are functions of the reduced ion concentration η , the association parameter λ and n . When $\lambda = 0$, $a_i = 1$, the average number of dimers $\langle N \rangle = 0$ and the energy of the RPM strong electrolyte in the MS approximation discussed earlier is recovered. The effect of a hard sphere solvent on the degree of dissociation of a weak electrolyte enhances the association parameter λ due to the packing effect of the solvent, while adding a dipole to the solvent has the opposite effect [71].

The PY approximation for the binding leads to negative results for λ ; the HNC approximation for this is satisfactory. Figure A2.3.18 shows the excess energy as a function of the weak electrolyte concentration for the RPM and SEM for a 2–2 electrolyte.

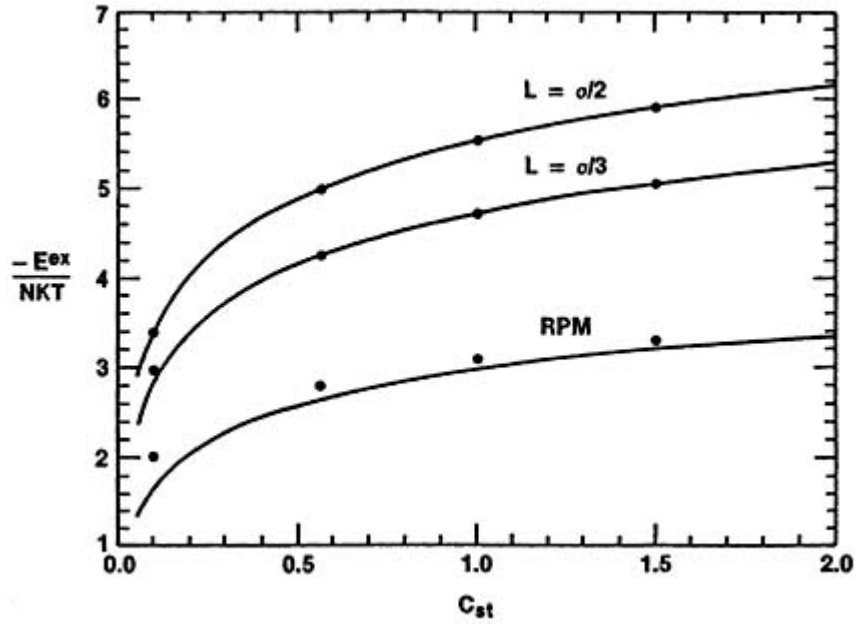


Figure A2.3.18 The excess energy E^{ex} in units of NkT as a function of the concentration c_{st} for the RPM and SEM 2–2 electrolyte. The curves and points are results of the HNC/MS and HNC approximations, respectively, for the binding and the electrical interactions. The ion parameters are $\sigma = 4.2 \text{ \AA}$, and $E = 73.4$. The sticking coefficients $\zeta = 1.6 \times 10^6$ and 2.44×10^6 for $L = \sigma/2$ and $\sigma/3$, respectively.

-74-

In the limit $\lambda = (\sigma/L)^3/\eta$ with $L < \sigma/2$, the system should consist of dipolar dumb-bells. The asymptotic form of the direct correlation function (defined through the Ornstein–Zernike equation) for this system (in the absence of a solvent) is given by

$$c_{ij}(r) = -\beta A e_i e_j / r \quad (\text{A2.3.244})$$

where $A = \varepsilon/(\varepsilon - 1)$ and ε is the dielectric constant of the system of dipolar dumb-bells. The energy of dipolar dumb-bells, excluding the binding energy, in the MS approximation is [48]

$$\frac{E^{\text{ex}}}{N_{\text{D}}kT} = \frac{-x[c_1 + c_2x'] - (c_1^2 + 2c_3x')^{1/2}}{24\eta} \quad (\text{A2.3.245})$$

where x' is the reduced dipole moment defined by

$$x' = \kappa\sigma = 2n(A\pi\rho/kT)^{1/2}\mu \quad (\text{A2.3.246})$$

dipole moment $\mu = eL = e\sigma/n$, $N_{\text{D}} = N/2$ is the number of dipoles, the coefficients c_i ($i = 1$ to 3) depend on the dipole elongation and n is an integer. This provides an analytic solution for the energy of dipolar dumb-bells in the MSA approximation; it suffers from the defect that it tends to a small but finite constant in the limit of zero density and should strictly be applicable only for $L < \sigma/2$.

A2.3.6 PERTURBATION THEORY

The attractive dispersive forces between the atoms of a simple homogeneous fluid increase their cohesive energy, but their vector sums nearly cancel, producing little alteration in the structure determined primarily by the repulsive part of the interatomic potential. Charges, dipoles and hydrogen bonding, as in water molecules, increase the cohesive energy of molecules and produce structural changes. Despite this, the harsh interatomic repulsions dominate the structure of simple fluids. This observation forms the physical basis of perturbation theory. van der Waals implicitly used this idea in his equation of state in which the attractive part of the interaction is treated as a perturbation to the repulsive part in a mean-field approximation.

In perturbation theories of fluids, the pair total potential is divided into a reference part and a perturbation

$$u(1, 2) = u^0(1, 2) + w(1, 2) \quad (\text{A2.3.247})$$

where $u^0(1, 2)$ is the pair potential of the reference system which usually has the features that determine the size and shape of the molecules, while the perturbation $w(1, 2)$ contains dispersive and attractive components which provide the cohesive energy of the fluid. The equilibrium properties of the system are calculated by expansion in a suitable parameter about the reference system, whose properties are assumed known to the extent that is necessary.

-75-

The reference system may be anisotropic, i.e. with $u^0(1, 2) = u^0(r_{12}, \Omega_1, \Omega_2)$, where (Ω_1, Ω_2) represent the angular coordinates of atoms 1 and 2, or it may be isotropic when $u^0(1, 2) = u^0(r_{12})$.

The most common choice for a reference system is one with hard cores (e.g. hard spheres or hard spheroidal particles) whose equilibrium properties are necessarily independent of temperature. Although exact results are lacking in three dimensions, excellent approximations for the free energy and pair correlation functions of hard spheres are now available to make the calculations feasible.

The two principal methods of expansion used in perturbation theories are the high-temperature λ expansion of Zwanzig [72], and the γ expansion introduced by Hemmer [73]. In the λ -expansion, the perturbation $w(1, 2)$ is modulated by the switching parameter λ which varies between 0 and 1, thereby turning on the perturbation. The free energy is expanded in powers of λ , and reduces to that of the reference system when $\lambda = 0$. In the γ expansion, the perturbation is long ranged of the form $w(r) = -\gamma^3 \phi(\gamma r)$, and the free energy is expanded in powers of γ about $\gamma = 0$. In the limit as $\gamma \rightarrow 0$, the free energy reduces to a mean-field van der Waals-like equation. The γ expansion is especially useful in understanding long-range perturbations, such as Coulomb and dipolar interactions, but difficulties in its practical implementation lie in the calculation of higher-order terms in the expansion. Another perturbation approach is the mode expansion of Andersen and Chandler [74], in which the configurational integral is expanded in terms of collective coordinates that are the Fourier transforms of the particle densities. The expansion is especially useful for electrolytes and has been optimized and improved by adding the correct second virial coefficient. Combinations of the λ and γ expansions, the union of the λ and virial expansions and other improvements have also been discussed in the literature. Our discussion will be mainly confined to the λ expansion and to applications of perturbation theory to determining free energy differences by computer simulation. We conclude the section with a brief discussion of perturbation theory of inhomogeneous fluids.

A2.3.6.1 THE λ EXPANSION

The first step is to divide the total potential into two parts: a reference part and the remainder treated as a perturbation. A coupling parameter λ is introduced to serve as a switch which turns the perturbation on or off. The total potential energy of N particles in a given configuration $(\mathbf{r}_1, \dots, \mathbf{r}_N)$ is

$$U_N(\mathbf{r}_1, \dots, \mathbf{r}_N; \lambda) = U_N^0(\mathbf{r}_1, \dots, \mathbf{r}_N) + \lambda W_N(\mathbf{r}, \dots, \mathbf{r}_N) \quad (\text{A2.3.248})$$

where $0 = \lambda = 1$, $U_N^0(\mathbf{r}_1, \dots, \mathbf{r}_N)$ is the reference potential and $W_N(\mathbf{r}_1, \dots, \mathbf{r}_N)$ is the perturbation. When λ is zero the perturbation is turned off, and it is on when $\lambda = 1$.

The configurational PF

$$\begin{aligned} Z(N, V, T; \lambda) &= \int \exp(-\beta U_N(\mathbf{r}^N; \lambda)) \, d\mathbf{r}^N \\ &= \int \exp(-\beta U_N^0(\mathbf{r}^N)) \exp(-\beta \lambda W_N(\mathbf{r}^N)) \, d\mathbf{r}^N. \end{aligned} \quad (\text{A2.3.249})$$

-76-

Multiplying and dividing by the configurational PF of the reference system

$$Z^0(N, V, T) = \int \exp(-\beta U_N^0(\mathbf{r}^N)) \, d\mathbf{r}^N \quad (\text{A2.3.250})$$

one finds that

$$Z(N, V, T; \lambda) = Z^0(N, V, T) \langle \exp(-\lambda W_N(\mathbf{r}^N)) \rangle_0 \quad (\text{A2.3.251})$$

where $\langle \dots \rangle_0$ is an average over the reference system. The Helmholtz free energy

$$A(N, V, T; \lambda) = -kT \ln Q(N, V, T; \lambda) = -kT \ln \frac{Z(N, V, T; \lambda)}{N! \Lambda^{3N}}. \quad (\text{A2.3.252})$$

It follows that the change in Helmholtz free energy due to the perturbation is

$$A(N, V, T; \lambda) - A^0(N, V, T; \lambda) = -kT \ln \langle \exp(-\beta \lambda W_N(\mathbf{r}^N)) \rangle_0. \quad (\text{A2.3.253})$$

This equation was first derived by Zwanzig [72]. Note that β and λ always occur together. Expanding about $\lambda = 0$ at constant β (or equivalently about $\beta = 0$ at constant λ) one finds

$$\begin{aligned} -\beta \Delta A(\lambda) &= \ln \langle \exp(-\beta \lambda W_N(\mathbf{r}^N)) \rangle_0 \\ &= \ln \left\langle \sum_{n=0}^{\infty} \frac{(-\beta \lambda)^n}{n!} W_N(\mathbf{r}^N) \right\rangle_0 = \sum_{n=1}^{\infty} (-\beta \lambda)^n a_n \end{aligned} \quad (\text{A2.3.254})$$

which defines the coefficients a_n . By comparing the coefficients of $(-\beta \lambda)^n$ for different n , one finds

$$(\text{A2.3.255})$$

$$\begin{aligned}
a_1 &= \langle W_N \rangle_0 \\
a_2 &= \frac{1}{2} [\langle W_N^2 \rangle_0 - \langle W_N \rangle_0^2] \\
a_3 &= \frac{1}{3!} [\langle W_N^3 \rangle_0 - 3\langle W_N \rangle_0 \langle W_N^2 \rangle_0 + 2\langle W_N \rangle_0^3] \text{ etc}
\end{aligned}$$

where the averages are over the reference system whose properties are assumed to be known.

The first term in the high-temperature expansion, a_1 , is essentially the mean value of the perturbation averaged over the reference system. It provides a strict upper bound for the free energy called the Gibbs–Bogoliubov inequality. It follows from the observation that $\exp(-x) \geq 1-x$ which implies that $\ln \langle \exp(-x) \rangle \geq \ln(1 - \langle x \rangle)$. Hence

-77-

$$\ln \langle \exp(-\beta \lambda W_N(\mathbf{r}^N)) \rangle_0 \geq -\beta \lambda \langle W_N(\mathbf{r}^N) \rangle_0.$$

Multiplying by -1 reverses this inequality, and we see that

$$\beta \Delta A(\lambda) \leq \beta \lambda \langle W_N(\mathbf{r}^N) \rangle_0 = \beta \lambda a_1 \quad (\text{A2.3.256})$$

which proves the result. The higher-order terms in the high-temperature expansion represent fluctuations about the mean.

Assuming the perturbing potential is pairwise additive,

$$W_N(\mathbf{r}^N) = \sum_{i < j} w_{ij}(r_{ij}) \quad (\text{A2.3.257})$$

we have

$$\begin{aligned}
a_1 &= \langle W_N \rangle_0 = \frac{\int \dots \int \sum_{i < j} w_{ij}(r_{ij}) \exp(-\beta U_N^0(\mathbf{r}^N)) \, d\mathbf{r}^N}{Z^0(N, V, T)} \\
&= \frac{N(N-1)}{2} \frac{\int \dots \int \exp(-\beta U_N^0(\mathbf{r}^N)) \, d\mathbf{r}_3 \dots d\mathbf{r}_N \, d\mathbf{r}_1 \, d\mathbf{r}_2}{Z^0(N, V, T)} \\
&= \int \dots \int \rho_N^{0,(2)}(\mathbf{r}_1, \mathbf{r}_2) \, d\mathbf{r}_1 \, d\mathbf{r}_2
\end{aligned}$$

where translational and rotational invariance of the reference fluid system implies that $\rho_N^{0,(2)}(\mathbf{r}_1, \mathbf{r}_2) = \rho^2 g_N^0(r_{12})$. Using this in the above expression, changing to relative coordinates and integrating over coordinates of 1, one has

$$a_1 = \frac{\rho N}{2} \int w_{12}(r_{12}) g_N^0(r_{12}) \, d\mathbf{r}_{12} \quad (\text{A2.3.258})$$

which was first obtained by Zwanzig. As discussed above, this provides an upper bound for the free energy,

so that

$$\frac{\Delta A(\lambda)}{N} \leq \frac{\rho}{2} \int \lambda w_{12}(r_{12}) g_N^0(r_{12}) \mathbf{d}r_{12}. \quad (\text{A2.3.259})$$

The high-temperature expansion, truncated at first order, reduces to van der Waals' equation, when the reference system is a fluid of hard spheres.

-78-

The second-order term, a_2 , was also obtained by Zwanzig, and involves two-, three- and four-body correlation functions for an N -particle system. Before passage to the thermodynamic limit,

$$\begin{aligned} a_2 = & \frac{1}{2} \int \dots \int \rho_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2) w_{12}(r_{12}) \mathbf{d}r_1 \mathbf{d}r_2 \\ & + \int \int \int \rho_N^{(3)}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) w_{12}(r_{12}) w_{23}(r_{23}) \mathbf{d}r_1 \mathbf{d}r_2 \mathbf{d}r_3 \\ & + \frac{1}{4} \int \int \int \int [\rho_N^{(4)}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \mathbf{r}_4) - \rho_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2) \rho_N^{(2)}(\mathbf{r}_3, \mathbf{r}_4)] w_{12}(r_{12}) w_{23}(r_{23}) \mathbf{d}r_1 \mathbf{d}r_2 \mathbf{d}r_3 \mathbf{d}r_4. \end{aligned} \quad (\text{A2.3.260})$$

Evaluating its contribution to the free energy of the system requires taking the thermodynamic limit ($N \rightarrow \infty$) for the four-particle distribution function. Lebowitz and Percus [75] and Hiroike [76] showed that the asymptotic behaviour of $\rho_N^{(4)}$ in the canonical ensemble, when the 1,2 and 3,4 pairs are widely separated, is given by

$$\rho_N^{(4)}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \mathbf{r}_4) = \rho_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2) \rho_N^{(2)}(\mathbf{r}_3, \mathbf{r}_4) + x(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \mathbf{r}_4)/N + O(N^{-2}) \quad (\text{A2.3.261})$$

where the $O(1/N)$ term makes a finite contribution to the last term in a_2 . This correction can also be evaluated in the grand canonical ensemble.

The high-temperature expansion could also be derived as a Taylor expansion of the free energy in powers of λ about $\lambda = 0$:

$$A(\lambda) = A_0 + \lambda(\delta A/\delta \lambda)_{\lambda=0} + (\lambda^2/2)(\delta^2 A/\delta \lambda^2)_{\lambda=0} + \dots \quad (\text{A2.3.262})$$

so that the coefficients of the various powers of λ are related to a_n , with

$$\begin{aligned} \frac{\partial A}{\partial \lambda} &= \frac{1}{Z(N, V, T; \lambda)} \int \dots \int W_N(\mathbf{r}^N) \exp(-\beta U(\mathbf{r}^N)) \mathbf{d}r^N \\ &= \langle W_N(\mathbf{r}^N) \rangle_\lambda. \end{aligned} \quad (\text{A2.3.263})$$

It follows that

$$A(\lambda) = A(0) + \int_0^\lambda \langle W_N(\mathbf{r}^N) \rangle_\lambda \mathbf{d}\lambda. \quad (\text{A2.3.264})$$

Assuming the perturbing potential is pairwise additive, an argument virtually identical to the calculation of $a_1 = \langle W_N(\mathbf{r}^N) \rangle_0$ shows that

-79-

$$\langle W_N(\mathbf{r}^N) \rangle_\lambda = \frac{1}{2} \int \dots \int \rho_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2; \lambda) w_{12}(r_{12}) d\mathbf{r}_1 d\mathbf{r}_2$$

where $\rho_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2; \lambda)$ is the two-particle density correlation function in an N -particle system with potential $U_N(\mathbf{r}_1, \dots, \mathbf{r}_N; \lambda)$. Substituting this in the expression for $A(\lambda)$, we have

$$\Delta A(\lambda) = \frac{1}{2} \int_0^\lambda d\lambda \int \dots \int \rho_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2; \lambda) w_{12}(r_{12}) d\mathbf{r}_1 d\mathbf{r}_2 \quad (\text{A2.3.265})$$

where $\Delta A(\lambda) = A(\lambda) - A(0)$. Expanding the two-particle density correlation function in powers of λ ,

$$\rho_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2; \lambda) = \rho_N^{0,(2)}(\mathbf{r}_1, \mathbf{r}_2) + \lambda \left(\frac{\partial \rho_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2; \lambda)}{\partial \lambda} \right)_{\lambda=0} + \dots \quad (\text{A2.3.266})$$

we see that the zeroth order term in λ yields the first-order term a_1 in the high-temperature expansion for the free energy, and the first-order term in λ gives the second-order term a_2 in this expansion. As is usual for a fluid, $\rho_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2; \lambda) = \rho^2 g_N^{(2)}(r_{12}; \lambda)$ and

$$g_N^{(2)}(r_{12}; \lambda) = g_N^{0,(2)}(r_{12}) + \lambda \left(\frac{\partial g_N^{(2)}(r_{12}; \lambda)}{\partial \lambda} \right)_{\lambda=0} + \dots \quad (\text{A2.3.267})$$

But by definition

$$g_N^{(2)}(r_{12}; \lambda) = \exp[-\beta(u_{12}^0(r_{12}) + \lambda w_{12}(r_{12}))] y_N(r_{12}; \lambda) \quad (\text{A2.3.268})$$

where $y(r_{12}; \lambda)$ is the cavity function, and $u_{12}^0(r_{12})$ is the pair potential of the reference system, from which it follows that

$$g_N^{(2)}(r_{12}; \lambda) \approx [1 - \beta \lambda w_{12}(r_{12})] g_N^{0,(2)}(r_{12}) \quad (\text{A2.3.269})$$

which suggests $g_N^{(2)}(r_{12}; \lambda) \simeq g_N^0(r_{12})$ when $\beta w_{12}(r_{12}) \beta \varepsilon \ll 1$, where ε is the depth of the potential well. It also suggests an improved approximation

$$c(r) = \beta u(r) - \ln g(r) + h(r) + B(r) \quad (\text{A2.3.270})$$

-80-

where $u_{12}(r_{12}; \lambda) = u_{12}^0(r_{12}) + \lambda w_{12}(r_{12})$ is the pair potential. The calculation of the second-order term in the high-temperature expansion involves the three- and four-body correlation functions which are generally not

known even for a hard sphere reference system. The situation becomes worse for the higher-order terms in the perturbation expansion. However, determination of the first-order term in this expansion requires only the pair correlation function of the reference system, for which a convenient choice is a fluid of hard spheres whose equilibrium properties are known. Barker and Henderson [77] suggested a hard sphere diameter defined by

$$d = \int_0^{\infty} [1 - \exp(-\beta u_{12}(r))] dr \quad (\text{A2.3.271})$$

where $u_{12}(r)$ is the pair potential. This diameter is temperature dependent and the free energy needs to be calculated to second order to obtain the best results.

Truncation at the first-order term is justified when the higher-order terms can be neglected. When $\beta\epsilon \ll 1$, a judicious choice of the reference and perturbed components of the potential could make the higher-order terms small. One choice exploits the fact that a_1 , which is the mean value of the perturbation over the reference system, provides a strict upper bound for the free energy. This is the basis of a variational approach [78, 79] in which the reference system is approximated as hard spheres, whose diameters are chosen to minimize the upper bound for the free energy. The diameter depends on the temperature as well as the density. The method was applied successfully to Lennard-Jones fluids, and a small correction for the softness of the repulsive part of the interaction, which differs from hard spheres, was added to improve the results.

A very successful first-order perturbation theory is due to Weeks, Chandler and Andersen [80], in which the pair potential $u(r)$ is divided into a reference part $u^0(r)$ and a perturbation $w(r)$

$$u(r) = u^0(r) + w(r) \quad (\text{A2.3.272})$$

in which

$$u^0(r) = \begin{cases} u(r) + \epsilon & (r < R_{\min}) \\ 0 & (r > R_{\min}) \end{cases} \quad (\text{A2.3.273})$$

and

$$w(r) = \begin{cases} -\epsilon & (r < R_{\min}) \\ u(r) & (r > R_{\min}) \end{cases} \quad (\text{A2.3.274})$$

where ϵ is the depth of the potential well which occurs at $r = R_{\min}$. This division into reference and perturbed parts is very fortuitous. The second step is to relate the reference system to an equivalent hard sphere fluid with a pair potential $u^{\text{HS}}(r)$. This is done by defining $v(r)$ by

$$\exp(-\beta v(r)) = \exp(-\beta u^{\text{HS}}(r)) + \alpha [\exp(-\beta u^0(r)) - \exp(-\beta u^{\text{HS}}(r))] \quad (\text{A2.3.275})$$

in which α is an expansion parameter with $0 \leq \alpha \leq 1$. The free energy of the system with the pair potential $v(r)$ is expanded in powers of α to $O(\alpha^2)$ to yield

$$A' = A^{\text{HS}} - \frac{N\rho\alpha}{2} \int [\exp(-\beta u^0(r)) - \exp(-\beta u^{\text{HS}}(r))] y^{\text{HS}}(r; d) dr + O(\alpha^2) \quad (\text{A2.3.276})$$

where $y^{\text{HS}}(r, d)$ is the cavity function of hard spheres of diameter d determined by annihilating the term of order α by requiring the integral to be zero. The diameter d is temperature and density dependent as in the variational theory. The free energy of the fluid with the pair potential $u(r)$ is now calculated to first order using the approximation

$$g^0(r) \exp(-\beta u(r)) y^{\text{HS}}(r, d). \quad (\text{A2.3.277})$$

This implies, with the indicated choice of hard sphere diameter d , that the compressibilities of the reference system and the equivalent of the hard sphere system are the same.

Another important application of perturbation theory is to molecules with anisotropic interactions. Examples are dipolar hard spheres, in which the anisotropy is due to the polarity of the molecule, and liquid crystals in which the anisotropy is due also to the shape of the molecules. The use of an anisotropic reference system is more natural in accounting for molecular shape, but presents difficulties. Hence, we will consider only isotropic reference systems, in which the reference potential $u^0(r_{12})$ is usually chosen in one of two ways. In the first choice, $u^0(r_{12})$ is defined by

$$\exp[-\beta u^0(r_{12})] = \Omega^{-2} \iint \exp[-\beta u(r_{12}, \Omega_1, \Omega_2)] d\Omega_1 d\Omega_2 \quad (\text{A2.3.278})$$

which can be applied even to hard non-spherical molecules. The ensuing reference potential, first introduced by Rushbrooke [81], is temperature dependent and was applied by Cook and Rowlinson [82] to spheroidal molecules. It is more complicated to use than the temperature-independent reference potential defined by the simple averaging

$$u^0(r_{12}) = \Omega^{-2} \iint u(r_{12}, \Omega_1, \Omega_2) d\Omega_1 d\Omega_2. \quad (\text{A2.3.279})$$

This choice was introduced independently by Pople [83] and Zwanzig [84].

We assume that the anisotropic pair interaction can be written as

$$u(r_{12}, \Omega_1, \Omega_2; \lambda) = u^0(r_{12}) + \lambda w(r_{12}, \Omega_1, \Omega_2) \quad (\text{A2.3.280})$$

-82-

where the switching function λ lies between zero and one. The perturbation is fully turned on when λ is one and is switched off when λ is zero. In the λ expansion for the free energy of the fluid,

$$\Delta A(\lambda) = \lambda A_1 + \lambda^2 A_2 + \lambda^3 A_3 + \dots \quad (\text{A2.3.281})$$

one finds that the leading term of order λ

$$A_1 = \frac{\rho N}{2\Omega^2} \iiint g^0(r_{12}) w(r_{12}, \Omega_1, \Omega_2) dr_{12} d\Omega_1 d\Omega_2 \quad (\text{A2.3.282})$$

vanishes on carrying out the angular integration due to the spherical symmetry of the reference potential. The expressions for the higher-order terms are

$$A_2 = -\frac{\beta\rho}{4\Omega^2} \iiint g^0(r_{12})w(r_{12}, \Omega_1, \Omega_2)^2 d\mathbf{r}_{12} d\Omega_1 d\Omega_2 \quad (\text{A2.3.283})$$

$$A_3 = \frac{\beta^2\rho}{12\Omega^2} \iiint g^0(r_{12})w(r_{12}, \Omega_1, \Omega_2)^3 d\mathbf{r}_{12} d\Omega_1 d\Omega_2 + \frac{\beta^2\rho^2}{6\Omega^3} \iiint g^{0,(3)}(r_{12}, r_{23}, r_{31}) \times w(r_{12}, \Omega_1, \Omega_2)w(r_{23}, \Omega_1, \Omega_2)w(r_{13}, \Omega_1, \Omega_2) d\mathbf{r}_{12} d\mathbf{r}_{13} d\Omega_1 d\Omega_2 d\Omega_3. \quad (\text{A2.3.284})$$

The expansion of the perturbation $w(r_{12}, \Omega_1, \Omega_2)$ in terms of multipole potentials (e.g. dipole–dipole, dipole–quadrupole, quadrupole–quadrupole) using spherical harmonics

$$w(r_{12}, \Omega_1, \Omega_2) = \sum \sum X^{l_1 l_2 m}(r) Y_{m_1}^{l_1}(\Omega_1) Y_{m_2}^{l_2}(\Omega_2) \quad (\text{A2.3.285})$$

leads to additional simplifications due to symmetry. For example, for molecules with only dipole and quadrupolar interactions, all terms in which the dipole moment appears an odd number of times at an integration vertex vanish. In particular, for a pure dipolar fluid, the two-body integral contributing to A_3 vanishes. Angular integration of the three-body integral leads to the Axelrod–Teller three-body potential:

$$u(r_1, r_2, r_3) = (3 \cos \theta_1 \cos \theta_2 \cos \theta_3 + 1)/(r_{12}r_{13}r_{23})^3 \quad (\text{A2.3.286})$$

-83-

where $\theta_1, \theta_2, \theta_3$ and r_{12}, r_{13}, r_{23} are the angles and sides of the triangle formed by the three dipoles so that only the spatial integration remains to evaluate A_3 . This is accomplished by invoking the superposition approximation

$$g^{0,(3)}(r_{12}, r_{13}, r_{23}) = g^0(r_{12})g^0(r_{13})g^0(r_{23}) \quad (\text{A2.3.287})$$

which makes only a small error when it contributes to the third-order perturbation term. Tables of the relevant integrals for dipoles and multipoles associated with a hard sphere reference system are available [85].

For many molecules the reduced dipole moment $\mu^* = (\mu^2/(\epsilon\sigma^3))^{1/2}$ is greater than 1 and the terms in the successive terms in the λ expansion oscillate widely. Stell, Rasaiah and Narang [85] suggested taming this by replacing the truncated expansion by the Padé approximant

$$\Delta A = A_2 \left(1 - \frac{A_2}{A_3}\right)^{-1} \quad (\text{A2.3.288})$$

which reproduces the expected behaviour that as μ becomes large the free energy A increases as μ^2 . The Padé approximant is quite successful in reproducing the thermodynamic behaviour of polar fluids. However, the critical exponents, as in all mean-field theories, are the classical exponents.

The generalization of the λ expansion to multicomponent systems is straightforward but requires knowledge

of the reference system pair correlation functions of all the different species. Application to electrically neutral Coulomb systems is complicated by the divergence of the leading term of order λ in the expansion, but this difficulty can be circumvented by exploiting the electroneutrality condition and using a screened Coulomb potential

$$u_{ij}(r) = u_{ij}^0(r) + \frac{e_i e_j}{\epsilon_0 r} \exp(-\alpha r) \quad (\text{A2.3.289})$$

where α is the screening parameter and $u_{ij}^0(r)$ is the reference potential. The term of order λ , generalized to the multicomponent electrolyte, is

$$A'_1 = \frac{V}{2\epsilon_0} \sum_{i=1}^{\sigma} \sum_{j=1}^{\sigma} e_i e_j c_i c_j \int_0^{\infty} g_{ij}^0(r) \frac{\exp(-\alpha r)}{r} dr. \quad (\text{A2.3.290})$$

-84-

The integrals are divergent in the limit $\alpha = 0$. However, substituting $g_{ij}^0(r) = h_{ij}^0(r) + 1$, and making use of the electroneutrality condition in the form

$$\sum_i^{\sigma} \sum_j^{\sigma} e_i e_j c_i c_j = \left(\sum_{i=1}^{\sigma} e_i c_i \right)^2 = 0 \quad (\text{A2.3.291})$$

one finds, on taking the limit $\alpha \rightarrow 0$, that

$$A_1 = \frac{V}{2\epsilon_0} \sum_{i=1}^{\sigma} \sum_{j=1}^{\sigma} e_i e_j c_i c_j \int_0^{\infty} \frac{h_{ij}^0(r)}{r} dr \quad (\text{A2.3.292})$$

in which the divergences have been subtracted out. It follows from the Gibbs–Bogoliubov inequality that the first two terms form an upper bound, and

$$A \leq A^0 + A_1. \quad (\text{A2.3.293})$$

For a symmetrical system in which the reference species are identical (e.g. hard spheres of the same size), the integral can be taken outside the summation, which then adds up to zero due to the electroneutrality condition, to yield

$$A_1 = 0. \quad (\text{A2.3.294})$$

The reference free energy in this case is an upper bound for the free energy of the electrolyte. A lower bound for the free energy difference ΔA between the charged and uncharged RPM system was derived by Onsager [86]; this states that $\Delta A/N > -e^2/\epsilon\sigma$. Improved upper and lower bounds for the free energy have been discussed by Gillan [87].

The expression for κ^2 shows that it is the product of the ionic concentration c and $e^2/\epsilon_0 kT$, which is called the Bjerrum parameter. The virial series is an expansion in the total ionic concentration c at a fixed value of $e^2/\epsilon_0 kT$. A theory due to Stell and Lebowitz (SL) [88], on the other hand, is an expansion in the Bjerrum parameter at constant c . The leading terms in this expansion are

$$\frac{A^{\text{ex}}}{NkT} = \frac{A^0}{NkT} + \frac{A_1}{NkT} - \frac{\kappa_1^3}{12\pi c} \quad (\text{A2.3.295})$$

-85-

where we have already seen that the first two terms form an upper bound. Here κ_1^{-1} is a modified Debye length defined by

$$\kappa_1^2 = \kappa^2 + \frac{4\pi}{\epsilon_0 kT} \sum_{i=1}^{\sigma} \sum_{j=1}^{\sigma} e_i e_j c_i c_j \int_0^{\infty} h_{ij}^0 dr. \quad (\text{A2.3.296})$$

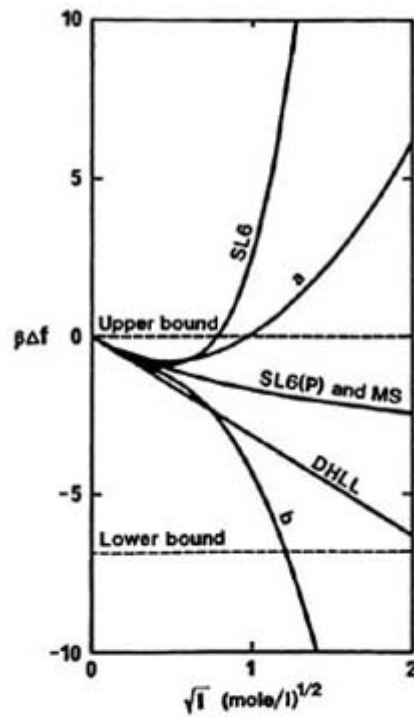
By the same argument as before, the integral may be taken outside the summations for a symmetrical reference system (e.g. the RPM electrolyte) and applying the electroneutrality condition one sees that in this case $\kappa_1 = \kappa$. Since the first two terms in the SL expansion form an upper bound for the free energy, the limiting law as $T \rightarrow \infty$ at constant c must always be approached from one side, unlike the Debye–Hückel limiting law which can be approached from above or below as $c \rightarrow 0$ at fixed temperature T (e.g. ZnSO_4 and HCl in aqueous solutions).

Examination of the terms to $O(\kappa^6)$ in the SL expansion for the free energy show that the convergence is extremely slow for a RPM 2–2 electrolyte in ‘aqueous solution’ at room temperature. Nevertheless, the series can be summed using a Padé approximant similar to that for dipolar fluids which gives results that are comparable in accuracy to the MS approximation as shown in [figure A2.3.19\(a\)](#). However, unlike the DHLL + B_2 approximation, neither of these approximations produces the negative deviations in the osmotic and activity coefficients from the DHLL observed for higher valence electrolytes at low concentrations. This can be traced to the absence of the complete renormalized second virial coefficient in these theories; it is present only in a linearized form. The union of the Pade approximant (SL6(P)), derived from the SL theory to $O(\kappa^6)$, and the Mayer expansion carried as far as DHLL + B_2

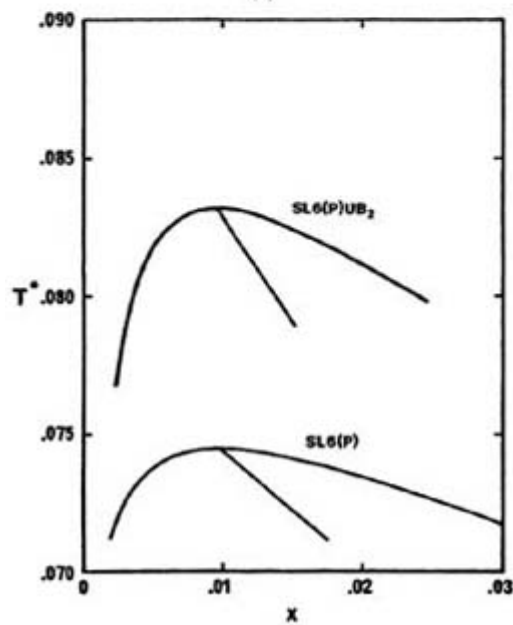
$$\text{SL6(P)} \cup B_2 = \text{SL6(P)} + B_2 - \text{SL6(P)} \cap B_2 \quad (\text{A2.3.297})$$

produces the right behaviour at low concentrations and has an accuracy comparable to the MS approximation at high concentrations. [Figure A2.3.19\(b\)](#) shows the coexistence curves for charged hard spheres predicted by SL6(P) and the $\text{SL6(P)} \cup B_2$.

-86-



(a)



(b)

Figure A2.3.19 Coexistence curve for the RPM predicted by SLR(P) and SL6(P)Y B₂. The reduced temperature $T^* = E kT\sigma/e^2$ and $x = \rho\sigma^3$ (after [85]).

By integrating over the hard cores in the SL expansion and collecting terms it is easily shown this expansion may be viewed as a correction to the MS approximation which still lacks the complete second virial coefficient. Since the MS approximation has a simple analytic form within an accuracy comparable to the Pade (SL6(P)) approximation it may be more convenient to consider the union of the MS approximation with Mayer theory. Systematic improvements to the MS approximation for the free energy were used to determine

the critical point and coexistence curves of charged, hard spheres by Stell, Wu and Larsen [89], and are discussed by Haksjold and Stell [90].

A2.3.6.2 COMPUTATIONAL ALCHEMY

Perturbation theory is also used to calculate free energy differences between distinct systems by computer simulation. This computational alchemy is accomplished by the use of a switching parameter λ , ranging from zero to one, that transforms the Hamiltonian of one system to the other. The linear relation

$$U(\lambda) = \lambda U_C + (1 - \lambda)U_B \quad (\text{A2.3.298})$$

interpolates between the energies $U_B(\mathbf{r}^N, \omega^N)$ and $U_C(\mathbf{r}^N, \omega^N)$ of the initial and final states of molecules C and B and allows for fictitious intermediate states also to be sampled. The switching parameter could be the dihedral angle in a peptide or polymer chain, the charge on an atom or one of the parameters ϵ or σ defining the size or well depth of its pair interaction with the environment.

It follows from our previous discussion of perturbation theory that

$$\Delta A(\mathbf{B} \rightarrow \mathbf{C}) = A_C - A_B = -kT \ln \langle \exp[-\beta(U_C - U_B)] \rangle_B \quad (\text{A2.3.299})$$

which is the free energy difference between the two states as a function of their energy difference sampled over the equilibrium configurations of one of the states. In the above expression, the averaging is over the equilibrium states of B, and C is treated as a perturbation of B. One could equally well sample over C and treat B as a perturbation of C. The averages are calculated using Monte Carlo or molecular dynamics discussed elsewhere; convergence is rapid when the initial and final states are similar. Since free energy differences are additive, the change in free energy between widely different states can also be determined through multiple simulations via closely-spaced intermediates determined by the switching function λ which gradually mutates B into C. The total free energy difference is the sum of these changes, so that

$$\Delta A(\mathbf{B} \rightarrow \mathbf{C}) = \int_0^1 \frac{\delta \Delta A(\lambda)}{\delta \lambda} d\lambda \quad (\text{A2.3.300})$$

and one calculates the derivative by using perturbation theory for small increments in λ . The accuracy can be improved by calculating incremental changes in both directions.

A closely-related method for determining free energy differences is characterized as thermodynamic integration. The configurational free energy of an intermediate state

$$A(\lambda) = -kT \ln Z(\lambda) \quad (\text{A2.3.301})$$

from which it follows that

$$\Delta A(\mathbf{B} \rightarrow \mathbf{C}) = \int_0^1 \left\langle \frac{\delta U(\lambda)}{\delta \lambda} \right\rangle_\lambda d\lambda \quad (\text{A2.3.302})$$

where the derivative pertains to equilibrated intermediate states. This forms the basis of the ‘slow growth’ method, in which the perturbation is applied linearly over a finite number of time steps and the free energy difference computed as the sum of energy differences. This method, however, samples over non-equilibrium intermediate states and the choice of the number of time steps over which the perturbation is applied and the corresponding accuracy of the calculation must be determined empirically.

Free energy perturbation (FEP) theory is now widely used as a tool in computational chemistry and biochemistry [91]. It has been applied to determine differences in the free energies of solvation of two solutes, free energy differences in conformational or tautomeric forms of the same solute by mutating one molecule or form into the other. Figure A2.3.20 illustrates this for the mutation of $\text{CH}_3\text{OH} \rightarrow \text{CH}_3\text{CH}_3$ [92].

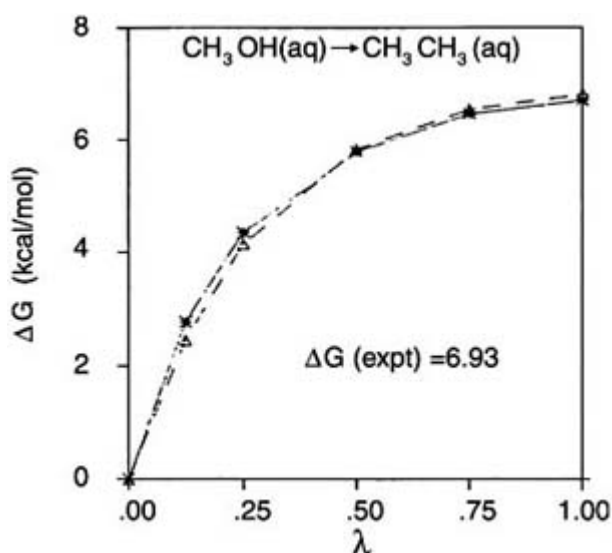
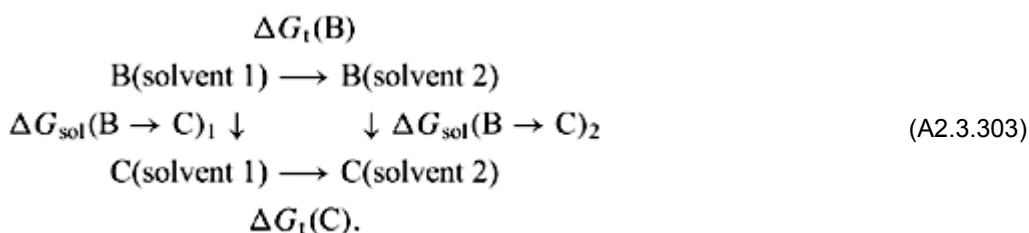


Figure A2.3.20 Free energy change in the transformation of CH_3OH to CH_3CH_3 (after [92]).

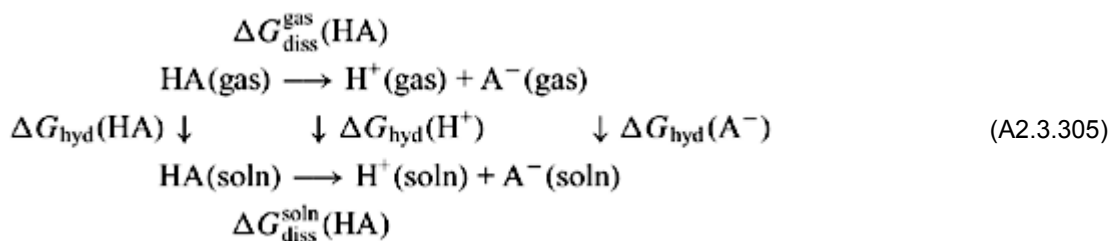
There are many other applications. They include determination of the ratios of the partition coefficients (P_B/P_C) of solutes B and C in two different solvents by using the thermodynamic cycle:



It follows that

$$\begin{aligned}
 2.3RT \ln(P_B/P_C) &= \Delta G_t(\text{C}) - \Delta G_t(\text{B}) \\
 &= \Delta G_{\text{sol}}(\text{B} \rightarrow \text{C})_2 - \Delta G_{\text{sol}}(\text{B} \rightarrow \text{C})_1
 \end{aligned} \tag{A2.3.304}$$

where $\Delta G_t(\text{B})$ and $\Delta G_t(\text{C})$ are the free energies of transfer of solute B and C, respectively, from solvent 1 to 2 and $\Delta G_{\text{sol}}(\text{B} \rightarrow \text{C})_1$ and $\Delta G_{\text{sol}}(\text{B} \rightarrow \text{C})_2$ are differences in the solvation energies of B and C in the respective solvents. Likewise, the relative pK_a s of two acids HA and HB in the same solvent can be calculated from the cycle depicted below for the acid HA



with a corresponding cycle for HB. From the cycle for HA, we see that

$$\begin{aligned}
2.3RT pK_a(\text{HA}) &= \Delta G_{\text{diss}}^{\text{soln}}(\text{HA}) \\
&= + \Delta G_{\text{hyd}}(\text{H}^+) + \Delta G_{\text{hyd}}(\text{A}^-) - \Delta G_{\text{hyd}}(\text{HA})
\end{aligned} \tag{A2.3.306}$$

with a similar expression for $pK_a(\text{HB})$. The difference in the two pK_a s is related to the differences in the gas phase acidities (free energies of dissociations of the acids), the free energies of hydration of B^- and A^- and the corresponding free energies of the undissociated acids:

$$\begin{aligned}
2.3RT[pK_a(\text{HA}) - pK_a(\text{HB})] &= \Delta G_{\text{diss}}^{\text{gas}}(\text{HA}) - \Delta G_{\text{diss}}^{\text{gas}}(\text{HB}) + \Delta G_{\text{hyd}}(\text{HA}) - \Delta G_{\text{hyd}}(\text{HB}) \\
&\quad + \Delta G_{\text{hyd}}(\text{A}^-) - \Delta G_{\text{hyd}}(\text{B}^-).
\end{aligned}$$

The relative acidities in the gas phase can be determined from *ab initio* or molecular orbital calculations while differences in the free energies of hydration of the acids and the cations are obtained from FEP simulations in which HA and A^- are mutated into HB and B^- , respectively.

-90-

Another important application of FEP is in molecular recognition and host-guest binding with its dependence on structural alterations. The calculations parallel our discussion of acid dissociation constants and have been used to determine the free energies of binding of A-T and C-G base pairs in solution from the corresponding binding energies in the gas phase. The relative free energies of binding two different ligands L_1 and L_2 to the same host are obtained from the following cycle:

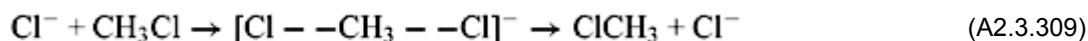


The difference in the free energy change when L_1 is replaced by L_2 is

$$\Delta G_2 - \Delta G_1 = \Delta G_4 - \Delta G_3 \tag{A2.3.308}$$

which is determined in FEP simulations by mutating L_1 to L_2 and EL_1 to EL_2 .

FEP theory has also been applied to modelling the free energy profiles of reactions in solution. An important example is the solvent effect on the SN_2 reaction



as illustrated from the work of Jorgenson [93] in figure A2.3.21.

The gas phase reaction shows a double minimum and a small barrier along the reaction coordinate which is the difference between the two C-Cl distances. The minima disappear in aqueous solution and this is accompanied by an increase in the height of the barrier. The behaviour in dimethyl formamide is intermediate between these two.

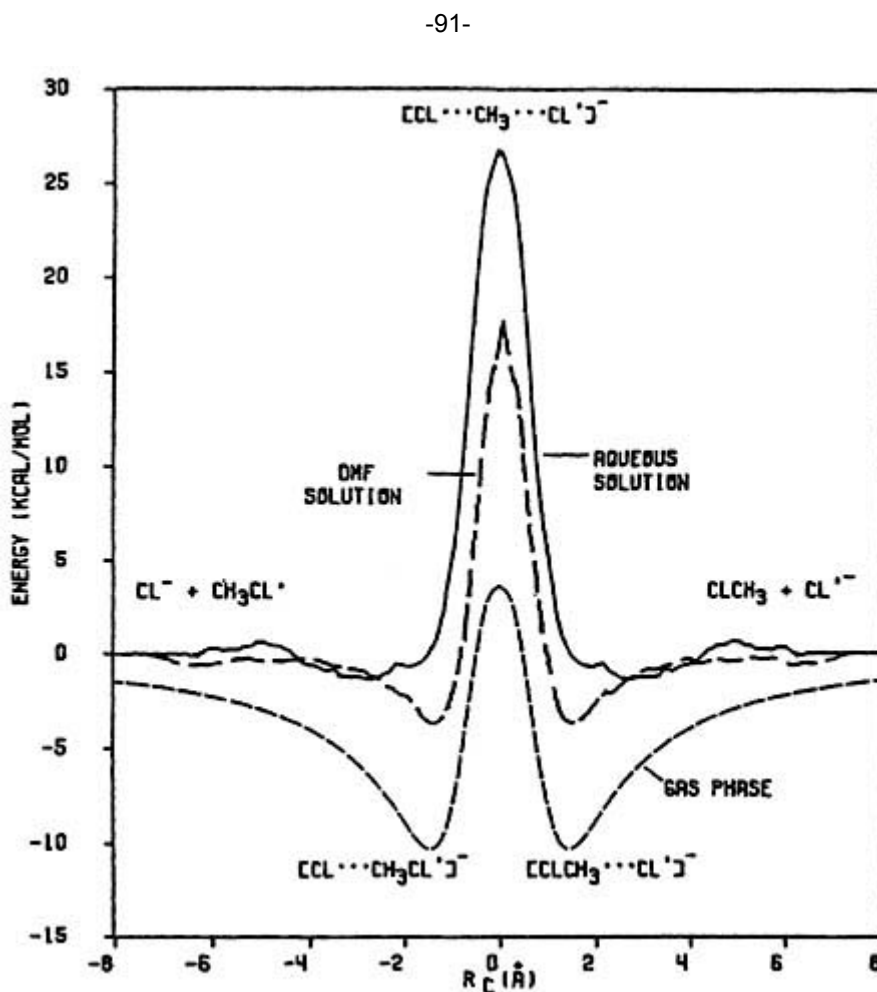


Figure A2.3.21 Free energy profile of the SN2 reaction $\text{Cl}^- + \text{CH}_3\text{Cl} \rightarrow [\text{Cl}-\text{CH}_3-\text{Cl}]^- \rightarrow \text{ClCH}_3 + \text{Cl}^-$ in the gas phase, dimethyl formamide and in water (from [93]).

A2.3.6.3 INHOMOGENEOUS FLUIDS

An inhomogeneous fluid is characterized by a non-uniform singlet density $\rho(r_1, [\phi])$ that changes with distance over a range determined by an external field. Examples of an external field are gravity, the walls enclosing a system or charges on an electrode. They are important in studies of interfacial phenomena such as wetting and the electrical double layer. The attractive interatomic forces in such systems do not effectively cancel due to the presence of the external field and perturbation theories applied to homogeneous systems are not very useful. Integral equation methods that ignore the bridge diagrams are also not very successful.

As discussed earlier, the singlet density $\rho(\mathbf{r}_1, [\phi])$ in an external field ϕ due to a wall is given by

$$kT \ln \rho(\mathbf{r}_1, [\phi]) = -\nabla_1 \phi(\mathbf{r}_1) - \int \rho(\mathbf{r}_2|\mathbf{r}_1; [\phi]) \nabla_1 u(r_{12}) d\mathbf{r}_2 \quad (\text{A2.3.310})$$

where the conditional density $\rho(\mathbf{r}_2|\mathbf{r}_1; [\phi]) = \rho^{(2)}(\mathbf{r}_1, \mathbf{r}_2; [\phi])/\rho(\mathbf{r}_1; [\phi])$. Weeks, Selinger and Broughton (WSB) [94] use this as a starting point of a perturbation theory in which the potential is separated into two parts,

$$u(r_{12}) = u^R(r_{12}) + w(r_{12}) \quad (\text{A2.3.311})$$

where $u^R(r_{12})$ is the reference potential and $w(r_{12})$ the perturbation. An effective field ϕ^R for the reference system is chosen so that the singlet density is unchanged from that of the complete system, implying the same average force on the atoms at \mathbf{r}_1 :

$$\rho^R(\mathbf{r}_1, [\phi^R]) = \rho(\mathbf{r}_1, [\phi]). \quad (\text{A2.3.312})$$

The effective field is determined by assuming that the conditional probabilities are the same, i.e.

$$\rho^R(\mathbf{r}_2|\mathbf{r}_1; [\phi^R]) = \rho(\mathbf{r}_2|\mathbf{r}_1; [\phi]) \quad (\text{A2.3.313})$$

when it follows that

$$\nabla_1 [\phi^R(\mathbf{r}_1) - \phi(\mathbf{r}_1)] = \int \rho^R(\mathbf{r}_2|\mathbf{r}_1; [\phi^R]) \nabla_1 w(r_{12}) d\mathbf{r}_2. \quad (\text{A2.3.314})$$

The conditional probability $\rho^R(\mathbf{r}_2|\mathbf{r}_1; [\phi^R])$ differs from the singlet density $\rho^R(\mathbf{r}_2; [\phi^R])$ mainly when 1 and 2 are close and the gradient of the perturbation $\nabla_1 w(r_{12})$ is small. Replacing $\rho^R(\mathbf{r}_2|\mathbf{r}_1; [\phi^R])$ by $\rho^R(\mathbf{r}_2; [\phi^R])$, taking the gradient outside the integral and integrating,

$$\begin{aligned} [\phi^R(\mathbf{r}_1) - \phi(\mathbf{r}_1)] &= \int \rho^R(\mathbf{r}_2; [\phi^R] - \rho^B) w(r_{12}) d\mathbf{r}_2 \\ &= \rho^B \int g^R(\mathbf{r}_2; [\phi^R]) - 1) w(r_{12}) d\mathbf{r}_2 \end{aligned} \quad (\text{A2.3.315})$$

where ρ^B is the bulk density far from the wall and $\rho^R(\mathbf{r}_2; [\phi^R]) = \rho^B g^R(\mathbf{r}_2; [\phi^R])$. This equation can be solved by standard methods provided the reference fluid distribution functions in the external field ϕ^R are known, for example through computer simulations. Other approximate methods to do this have also been devised by WSB [95].

A2.3.7.1 INTRODUCTION

Our discussion of solids and alloys is mainly confined to the Ising model and to systems that are isomorphic to it. This model considers a periodic lattice of N sites of any given symmetry in which a spin variable $s_i = \pm 1$ is associated with each site and interactions between sites are confined only to those between nearest neighbours. The total potential energy of interaction

$$U_N(\{s_k\}) = -J \sum_{\langle ij \rangle} s_i s_j - H \sum_i s_i \quad (\text{A2.3.316})$$

where $\{s_k\}$ denotes the spin variable $\{s_1, s_2, \dots, s_N\}$, J is the coupling constant between neighbouring sites and H is the external field which acts on the spins s_i at each site. The notation $\langle ij \rangle$ denotes summation over the nearest-neighbour sites; there are $Nq/2$ terms in this sum where q is the coordination number of each site. Ferromagnetic systems correspond to $J > 0$, for which the spins are aligned in domains either up $\uparrow\uparrow\uparrow\uparrow\uparrow$ or down $\downarrow\downarrow\downarrow\downarrow\downarrow$ at temperatures below the critical point, while in an antiferromagnet $J < 0$, and alternating spins $\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow$ on the lattice sites dominate at the lowest temperatures. The main theoretical problem in these systems is to predict the critical temperature and the phase diagram. Of added interest is the isomorphism to the lattice gas and to a two-component alloy so that the phase diagram for an Ising ferromagnetic can be mapped on to those for these systems as well. This analogy is further strengthened by the universality hypothesis, which states that the critical exponents and properties near the critical point are identical, to the extent that they depend only on the dimensionality of the system and the symmetry of the Hamiltonian. The details of the intermolecular interactions are thus of less importance.

A2.3.7.2 ISING MODEL

The partition function (PF) for the Ising [96] model for a system of given N , H and T is

$$\begin{aligned} Z(N, H, T) &= \exp(-\beta G) = \sum_{\{s_k\}} \exp[-\beta U_N(\{s_k\})] \\ &= \sum_{s_1=\pm 1} \sum_{s_2=\pm 1} \dots \sum_{s_N=\pm 1} \exp \beta \left[J \sum_{\langle ij \rangle} s_i s_j + H \sum_i s_i \right]. \end{aligned} \quad (\text{A2.3.317})$$

There are 2^N terms in the sum since each site has two configurations with spin either up or down. Since the number of sites N is finite, the PF is analytic and the critical exponents are classical, unless the thermodynamic limit ($N \rightarrow \infty$) is considered. This allows for the possibility of non-classical exponents and ensures that the results for different ensembles are equivalent. The characteristic thermodynamic equation for the variables N , H and T is

-94-

$$dG = -S dT - M dH + \mu dN \quad (\text{A2.3.318})$$

where M is the total magnetization and μ is the chemical potential. Since the sites are identical, the average magnetization per site is independent of its location, $m(H, T) = \langle s_i \rangle$ for all sites i . The total magnetization

$$M = Nm(H, T) = N \langle s_0 \rangle. \quad (\text{A2.3.319})$$

The magnetization per site

$$\langle s_0 \rangle = \frac{1}{Z} \sum_{\{s_i\}} s_0 \exp[-\beta(U_M(\{s_i\}))] = \frac{1}{Z} \left\{ \frac{\partial \ln Z(N, H, T)}{\partial H} \right\} \quad (\text{A2.3.320})$$

follows from the PF. As $H \rightarrow 0$

$$m(H, T) = \langle s_0 \rangle = 0 \quad (T > T_c) \quad (\text{A2.3.321})$$

unless the temperature is less than the critical temperature T_c when the magnetization lies between -1 and $+1$,

$$-1 \leq m(H, T) \leq 1 \quad (T < T_c) \quad (\text{A2.3.322})$$

and $m(H, T)$ versus H is a symmetrical odd function as shown in the [figure A2.3.22](#).

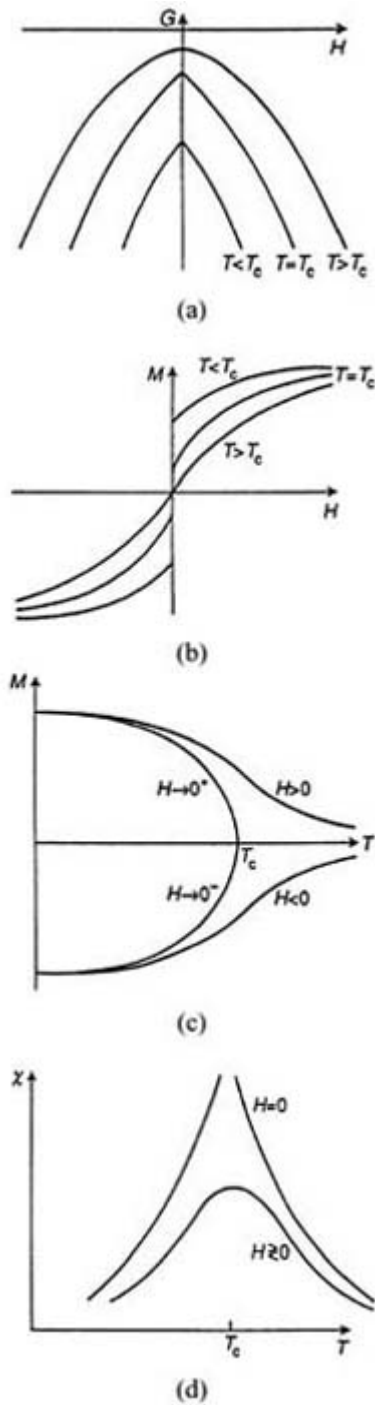


Figure A2.3.22 (a) The free energy G and (b) the magnetization $m(H, T)$ as a function of the magnetic field H at different temperatures. (c) The magnetization $m(H, T)$ and (d) the susceptibility χ as a function of temperature.

At $T = 0$, all the spins are either aligned up or down. The magnetization per site is an order parameter which vanishes at the critical point. Along the coexistence curve at zero field

$$m(H, T) \approx (T - T_c)^\beta$$

where β here is a critical exponent identical to that for a fluid in the same dimension due to the universality. Since $s_i = \pm 1$, at all temperatures $\langle s_i^2 \rangle = 1$.

To calculate the spin correlation functions $\langle s_i s_j \rangle$ between any two sites, multiply the expression for $\langle s_0 \rangle$ by Z when

$$\langle s_0 \rangle Z = \sum_{\{s_i\}} s_0 \exp \beta \left[J \sum_{\langle ij \rangle} s_i s_j + H \sum_i s_i \right].$$

Differentiating with respect to H and dividing by Z , we have

$$\left\{ \frac{\partial \langle s_0 \rangle}{\partial H} \right\}_T + N \beta \langle s_0 \rangle \langle s_i \rangle = \beta \sum_i \langle s_0 s_i \rangle$$

where the factor N comes from the fact that there are N identical sites. The magnetic susceptibility per site

$$\chi_T(H) = \left(\frac{\partial \langle s_0 \rangle}{\partial H} \right)_T = \beta \sum_i [\langle s_0 s_i \rangle - \langle s_0 \rangle \langle s_i \rangle]. \quad (\text{A2.3.324})$$

For $T > T_c$, $\langle s_0 \rangle = \langle s_i \rangle = 0$ when $H = 0$. Separating out the term $i = 0$ from the rest and noting that $\langle s_0^2 \rangle = 1$, we have

$$\chi_T(\mathbf{0}) = \beta \left[1 + \sum_{i \neq 0} \langle s_0 s_i \rangle \right] \quad \text{for } T > T_c \quad (\text{A2.3.325})$$

which relates the susceptibility at zero field to the sum of the pair correlation function over different sites.

This equation is analogous to the compressibility equation for fluids and diverges with the same exponent γ as the critical temperature is approached from above:

$$\chi_T(\mathbf{0}) \simeq |T - T_c|^{-\gamma}. \quad (\text{A2.3.326})$$

The correlation length $\zeta = |T - T_c|^{-\nu}$ diverges with the exponent ν . Assuming that when $T > T_c$ the site correlation function decays as [26]

$$\langle s_i s_0 \rangle = \frac{f(r/\xi)}{r^{D-2+\eta}} \quad (\text{A2.3.327})$$

where r is the distance between sites, D is the dimensionality and η is another critical exponent, one finds, as for fluids, that $(2 - \eta)\nu = \gamma$.

An alternative formulation of the nearest-neighbour Ising model is to consider the number of up $[\uparrow]$ and down $[\downarrow]$ spins, the numbers of nearest-neighbour pairs of spins $[\uparrow\uparrow]$, $[\downarrow\downarrow]$, $[\uparrow\downarrow]$ and their distribution over the lattice sites. Not all of the spin densities are independent since

$$N = [\uparrow] + [\downarrow] \quad (\text{A2.3.328})$$

and, if q is the coordination number of each site,

$$q[\uparrow] = 2[\uparrow\uparrow] + [\uparrow\downarrow] \quad (\text{A2.3.329})$$

$$q[\downarrow] = 2[\downarrow\downarrow] + [\uparrow\downarrow]. \quad (\text{A2.3.330})$$

Thus, only two of the five quantities $[\uparrow]$, $[\downarrow]$, $[\uparrow\uparrow]$, $[\downarrow\downarrow]$, $[\uparrow\downarrow]$ are independent. We choose the number of down spins $[\downarrow]$ and nearest-neighbour pairs of down spins $[\downarrow\downarrow]$ as the independent variables. Adding and subtracting the above two equations,

$$qN = 2([\uparrow\uparrow] + [\downarrow\downarrow] + [\uparrow\downarrow]) \quad (\text{A2.3.331})$$

$$q([\uparrow] - [\downarrow]) = 2([\uparrow\uparrow] - [\downarrow\downarrow]) \quad (\text{A2.3.332})$$

and

$$[\uparrow\uparrow] + [\downarrow\downarrow] = qN/2 - [\uparrow\downarrow]. \quad (\text{A2.3.333})$$

Defining the magnetization per site as the average number of up spins minus down spins,

$$\langle s_i \rangle = m(H, T) = \{[\uparrow] - [\downarrow]\}/N = 2([\uparrow\uparrow] - [\downarrow\downarrow])/N \quad (\text{A2.3.334})$$

where the last relation follows because we consider only nearest-neighbour interactions between sites. The lattice Hamiltonian

$$\begin{aligned} U_N(\{s_k\}) &= -J \sum_{\langle ij \rangle} s_i s_j - H \sum_i s_i \\ &= -J([\downarrow\downarrow] + [\uparrow\uparrow] - [\uparrow\downarrow]) - H([\uparrow] - [\downarrow]). \end{aligned} \quad (\text{A2.3.335})$$

-98-

Making use of the relations between the spin densities, the energy of a given spin configuration can be written in terms of the numbers of down spins $[\downarrow]$ and nearest-neighbour down spins $[\downarrow\downarrow]$:

$$\begin{aligned} U_N(\{s_k\}) &= -J \left(\frac{qN}{2} - 2[\uparrow\downarrow] \right) - H(N - 2[\downarrow]) \\ &= -N \left(\frac{qJ}{2} + H \right) + 2(qJ + H)[\downarrow] - 4J[\downarrow\downarrow]. \end{aligned} \quad (\text{A2.3.336})$$

For given J , H , q and N , the PF is determined by the numbers $[\downarrow]$ and $[\downarrow\downarrow]$ and their distribution over the sites, and is given by

$$Z(N, H, T) = e^{(\beta J q/2 + H)N} \sum_{[\downarrow]} e^{-2\beta(qJ+H)[\downarrow]} \sum_{[\downarrow\downarrow]} g_N([\downarrow], [\downarrow\downarrow]) e^{4\beta J[\downarrow\downarrow]} \quad (\text{A2.3.337})$$

where the sum over the number of nearest-neighbour down spins $[\downarrow\downarrow]$ is for a given number of down spins $[\downarrow]$, and $g_N([\downarrow], [\downarrow\downarrow])$ is the number of ways of distributing $[\downarrow]$ and $[\downarrow\downarrow]$ over N sites. Summing this over all $[\downarrow\downarrow]$ for fixed $[\downarrow]$ just gives the number of ways of distributing $[\downarrow]$ down spins over N sites, so that

$$\sum_{[\downarrow\downarrow]} g_N([\downarrow], [\downarrow\downarrow]) = \frac{N!}{([\downarrow])!(N - [\downarrow])!}. \quad (\text{A2.3.338})$$

In this formulation a central problem is the calculation of $g_N([\downarrow], [\downarrow\downarrow])$.

The Ising model is isomorphic with the lattice gas and with the nearest-neighbour model for a binary alloy, enabling the solution for one to be transcribed into solutions for the others. The three problems are thus essentially one and the same problem, which emphasizes the importance of the Ising model in developing our understanding not only of ferromagnets but other systems as well.

A2.3.7.3 LATTICE GAS

This model for a fluid was introduced by Lee and Yang [97]. The system is divided into cells with occupation numbers

$$n_i = \begin{cases} 1 & \text{cell } i \text{ is occupied} \\ 0 & \text{cell } i \text{ is not occupied.} \end{cases} \quad (\text{A2.3.339})$$

No more than one particle may occupy a cell, and only nearest-neighbour cells that are both occupied interact with energy $-\varepsilon$. Otherwise the energy of interactions between cells is zero. The total energy for a given set of occupation numbers $\{n\} = (n_1, n_2, \dots, n_N)$ of the cells is then

-99-

$$U_N(\{n\}) = -\varepsilon \sum_{\langle ij \rangle} n_i n_j \quad (\text{A2.3.340})$$

where the sum is over nearest-neighbour cells. The grand PF for this system is

$$\Xi(\mu, N, T) = \exp(\beta p N) = \sum_{n_1=0.1} \dots \sum_{n_N=0.1} \exp \beta \left[\varepsilon \sum_{\langle ij \rangle} n_i n_j + \mu \sum_i n_i \right]. \quad (\text{A2.3.341})$$

The relationship between the lattice gas and the Ising model follows from the observation that the cell occupation number

$$n_i = \frac{(1 + s_i)}{2} = \begin{cases} 1 & s_i = 1 \\ 0 & s_i = -1 \end{cases} \quad (\text{A2.3.342})$$

which associates the spin variable $s_i = \pm 1$ of the Ising model with the cell occupation number of the lattice gas. To calculate the energy, note that

$$\sum_{\langle ij \rangle} n_i n_j = \frac{1}{4} \sum_{\langle ij \rangle} (1 + s_i)(1 + s_j) = \frac{Nq}{8} + \frac{q}{4} \sum_i s_i + \frac{1}{4} \sum_{\langle ij \rangle} s_i s_j \quad (\text{A2.3.343})$$

where the second equality follows from

$$\sum_{\langle ij \rangle} 1 = \frac{Nq}{2} \quad \sum_{\langle ij \rangle} s_i = \frac{q}{2} \sum_i s_i \quad (\text{A2.3.344})$$

Also

$$\sum_i n_i = \sum_i \frac{(1 + s_i)}{2} = \frac{N}{2} + \frac{1}{2} \sum_i s_i. \quad (\text{A2.3.345})$$

It follows that the exponent appearing in the PF for the lattice gas,

$$\varepsilon \sum_{\langle ij \rangle} n_i n_j + \mu \sum_i n_i = \frac{\varepsilon Nq}{8} + \frac{\mu N}{2} + \frac{\varepsilon}{4} \sum_{\langle ij \rangle} s_i s_j + \left(\frac{\varepsilon q}{4} + \frac{\mu}{2} \right) \sum_i s_i. \quad (\text{A2.3.346})$$

-100-

Using this in the lattice gas grand PF,

$$\exp(\beta p N) = \exp \left[\beta \left(\frac{\varepsilon Nq}{8} + \frac{\mu N}{2} \right) \right] \sum_{\{s_i\}} \exp \left[\frac{\beta \varepsilon}{4} \sum_{\langle ij \rangle} s_i s_j + \beta \left(\frac{\varepsilon q}{4} + \frac{\mu}{2} \right) \sum_i s_i \right]. \quad (\text{A2.3.347})$$

Comparing with the PF of the Ising model

$$\exp(-\beta G) = \sum_{\{s_i\}} \exp \left[\beta J \sum_{\langle ij \rangle} s_i s_j + \beta H \sum_i s_i \right] \quad (\text{A2.3.348})$$

one sees that they are of the same form, with solutions related by the following transcription table:

Ising	Lattice gas
$4J$	ε
H	$\left(\frac{\varepsilon q}{4} + \frac{\mu}{2} \right) = Jq + \mu/2$
$-G$	$pN - \left(\frac{\varepsilon Nq}{8} + \frac{\mu N}{2} \right)$

It follows from this that

(A2.3.349)

$$\begin{aligned}\rho(\text{lattice gas}) &= \langle n_i \rangle = (1 + \langle s_i \rangle)/2 = (1 + m)/2 \\ \mu(\text{lattice gas}) &= 2H - 2Jq \\ P(\text{lattice gas}) &= -\frac{G}{N} + \frac{\varepsilon q}{8} + \frac{\mu}{2} = -\frac{G}{N} - \frac{Jq}{2} + H.\end{aligned}$$

At $H = 0$, $\mu(\text{lattice gas}) = -2Jq$ and the chemical potential is analytic even at $T = T_c$. From the thermodynamic relation,

$$d\mu = -S_M dT + V_M dp \quad (\text{A2.3.350})$$

where S_M and V_M are the molar entropy and volume, it follows that

$$\rho \left(\frac{\partial^2 \mu}{\partial T^2} \right)_\rho = -\frac{1}{T} \frac{C}{V_M} + \left(\frac{\partial^2 P}{\partial T^2} \right)_\rho. \quad (\text{A2.3.351})$$

The specific heat along the critical isochore hence has the same singularity as $(\partial^2 P / \partial T^2)_\rho$ for a lattice gas.

-101-

The relationship between the lattice gas and the Ising model is also transparent in the alternative formulation of the problem, in terms of the number of down spins [\downarrow] and pairs of nearest-neighbour down spins [$\downarrow\downarrow$]. For a given degree of site occupation [\downarrow],

$$U_{[\downarrow]} = -\varepsilon[\downarrow\downarrow] \quad (\text{A2.3.352})$$

and the lattice gas canonical ensemble PF

$$Q([\downarrow], N, T) = \sum_{[\downarrow\downarrow]} g_N([\downarrow], [\downarrow\downarrow]) \exp(\beta\varepsilon[\downarrow\downarrow]). \quad (\text{A2.3.353})$$

Removing the restriction on fixed [\downarrow], by considering the grand ensemble which sums over [\downarrow], one has

$$\exp(\beta p N) = \Xi(z, N, T) = \sum_{[\downarrow]} z^{[\downarrow]} \sum_{[\downarrow\downarrow]} g_N([\downarrow], [\downarrow\downarrow]) \exp(\beta\varepsilon[\downarrow\downarrow]) \quad (\text{A2.3.354})$$

where the fugacity $z = \exp(\beta\mu)$. Comparing this with the PF for the Ising model in this formulation, the entries in the transcription table given above are readily derived. Note that

$$m(T, H) \iff (1 - 2\rho) \quad (\text{A2.3.355})$$

and

$$2H \iff -kT \ln(z/\sigma) \quad (\text{A2.3.356})$$

where $\sigma = \exp(-2\beta q J) = \exp(-\beta q \varepsilon/2)$. Since m is an odd function of H , for the Ising ferromagnet $(1 - 2r)$

must be an odd function of $kT \ln(z/\sigma)$ for a lattice gas and $m = 0$ corresponds to $\rho = 1/2$ for a lattice gas. The liquid and vapour branches of the lattice gas are completely symmetrical about $\rho = 1/2$ when $T < T_c$. The phase diagram on the $\rho - \mu$ is illustrated in [figure A2.3.23](#).

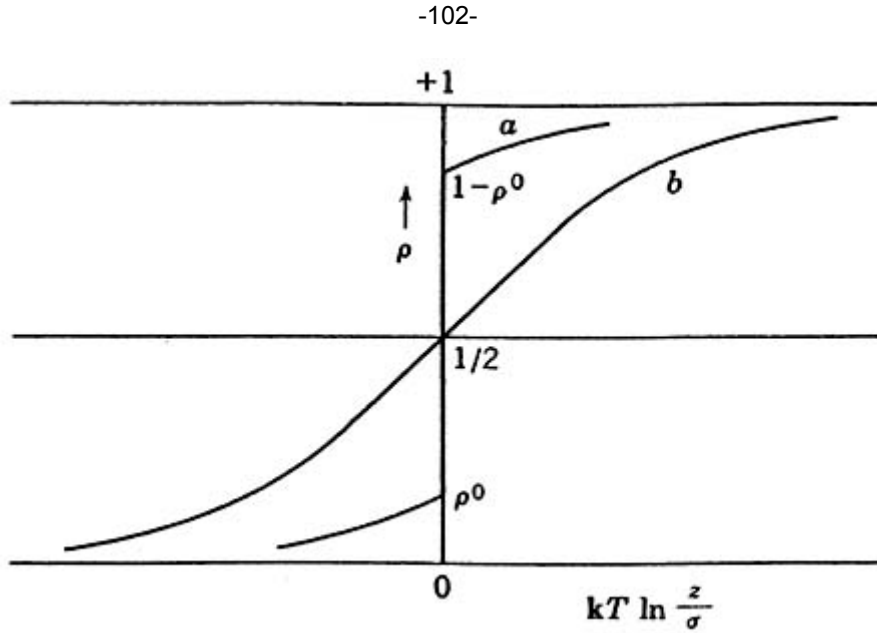


Figure A2.3.23 The phase diagram for the lattice gas.

For two symmetrically placed points A and B on the isotherm, i.e. conjugate phases,

$$\begin{aligned}
 \rho(z) &= 1 - \rho(z') \\
 \ln(z/\sigma) &= -\ln(z'/\sigma), \text{ i.e. } zz' = \sigma^2 \\
 \beta p(z) - (1/2) \ln z &= \beta p(z') - (1/2) \ln z' \\
 E/N - \rho(z)q\varepsilon/2 &= E'/N - \rho(z')q\varepsilon/2
 \end{aligned}
 \tag{A2.3.357}$$

from which it follows that for the conjugate phases

$$\begin{aligned}
 \rho(z) + \rho(z') &= 1 \\
 \mu(z, T) + \mu(z', T) &= \varepsilon \\
 p(z, T) - p(z', T) &= (1/2)[\mu(z, T) - \mu(z', T)] \\
 E(z, T) - E(z', T) &= (Nq\varepsilon/2)[\rho(z, T) - \rho(z', T)].
 \end{aligned}
 \tag{A2.3.358}$$

A2.3.7.4 BINARY ALLOY

A binary alloy of two components A and B with nearest-neighbour interactions ε_{AA} , ε_{BB} and ε_{AB} , respectively, is also isomorphic with the Ising model. This is easily seen on associating spin up with atom A and spin down with atom B. There are no vacant sites, and the occupation numbers of the site i are defined by

$$n_{i,A} = (1/2)(1 - s_i) \quad n_{i,B} = (1/2)(1 + s_i). \quad (\text{A2.3.359})$$

Summing over the sites

$$\sum_i (n_{i,A} + n_{i,B}) = N_A + N_B = N \quad (\text{A2.3.360})$$

where N_A and N_B are the number of atoms of A and B, respectively, distributed over N sites. For an open system,

$$\langle N_A \rangle = \left\langle (1/2) \left[1 - \sum_i s_i \right] \right\rangle = (N/2)[1 - m(H, T)] \quad (\text{A2.3.361})$$

$$\langle N_B \rangle = \left\langle (1/2) \left[1 + \sum_i s_i \right] \right\rangle = (N/2)[1 + m(H, T)].$$

The coordination number of each site is q , and

$$qN_A = 2N_{AA} + N_{AB} \quad (\text{A2.3.362})$$

$$qN_B = 2N_{BB} + N_{AB} \quad (\text{A2.3.363})$$

$$N = N_A + N_B = (2/q)[N_{AA} + N_{BB} + N_{AB}]. \quad (\text{A2.3.364})$$

On a given lattice of N sites, one number from the set $\{N_A, N_B\}$ and another from the set $\{N_{AA}, N_{BB}, N_{AB}\}$ determine the rest. We choose N_A and N_{AA} as the independent variables. Assuming only nearest-neighbour interactions, the energy of a given configuration

$$\begin{aligned} U_{[N_A]}(N_A, N_{AA}) &= \varepsilon_{AA}N_{AA} + \varepsilon_{BB}N_{BB} + \varepsilon_{AB}N_{AB} \\ &= \frac{qN\varepsilon_B}{2} + qN_A(\varepsilon_{AB} - \varepsilon_{AA}) + N_{AA}(\varepsilon_{AA} + \varepsilon_{BB} - 2\varepsilon_{AB}) \end{aligned} \quad (\text{A2.3.365})$$

which should be compared with the corresponding expressions for the lattice gas and the Ising model. The grand PF for the binary alloy is

$$\begin{aligned}
\Xi(N, z_A, z_B, T) &= \sum_{N_A=0}^N z_A^{N_A} z_B^{N_B} \sum_{N_{AA}, \text{fixed } N_A} q_N(N_A, N_{AA}) \exp[-\beta U_{N_A}(N_A, N_{AA})] \\
&= (z_B^{N_B} e^{\beta q \epsilon_{BB}}) \sum_{N_A=0}^N [(z_A/z_B) e^{-\beta q(\epsilon_{AB} - \epsilon_{AA})}] \\
&\quad \sum_{N_{AA}, \text{fixed } N_A} g_N(N_A, N_{AA}) e^{-\beta q(\epsilon_{AA} + \epsilon_{BB} - 2\epsilon_{AB})N_{AA}}
\end{aligned}$$

where $g_N(N_A, N_{AA})$ is the number of ways of distributing N_A and N_{AA} over N lattice sites and $z_A = \exp(\beta\mu_A)$ and $z_B = \exp(\beta\mu_B)$ are the fugacities of A and B, respectively. Comparing the grand PF for the binary alloy with the corresponding PFs for the lattice gas and Ising model leads to the following transcription table:

Ising model	Lattice gas	Binary alloy
$-4J$	ϵ	$\epsilon_{AA} + \epsilon_{BB} - 2\epsilon_{AB}$
$-2(qJ + H)$	μ	$\mu_A + \mu_B + q(\epsilon_{AA} - \epsilon_{AB})$

When $2\epsilon_{AB} > (\epsilon_{AA} + \epsilon_{BB})$, the binary alloy corresponds to an Ising ferromagnet ($J > 0$) and the system splits into two phases: one rich in A and the other rich in component B below the critical temperature T_c . On the other hand, when $2\epsilon_{AB} < (\epsilon_{AA} + \epsilon_{BB})$, the system corresponds to an antiferromagnet: the ordered phase below the critical temperature has A and B atoms occupying alternate sites.

A2.3.8 MEAN-FIELD THEORY AND EXTENSIONS

Our discussion shows that the Ising model, lattice gas and binary alloy are related and present one and the same statistical mechanical problem. The solution to one provides, by means of the transcription tables, the solution to the others. Historically, however, they were developed independently before the analogy between the models was recognized.

We now turn to a mean-field description of these models, which in the language of the binary alloy is the Bragg–Williams approximation and is equivalent to the Curie–Weiss approximation for the Ising model. Both these approximations are closely related to the van der Waals description of a one-component fluid, and lead to the same classical critical exponents $\alpha = 0$, $\beta = 1/2$, $\delta = 3$ and $\gamma = 1$.

As a prelude to discussing mean-field theory, we review the solution for non-interacting magnets by setting $J = 0$ in the Ising Hamiltonian. The PF

$$\begin{aligned}
Z(N, H, T) &= \sum_{s_1=\pm 1} \dots \sum_{s_N=\pm 1} e^{\beta H \sum_{i=1}^N s_i} = \sum_{s_1=\pm 1} \dots \sum_{s_N=\pm 1} \prod_{i=1}^N e^{\beta H s_i} \\
&= \left(\sum_{s_i=\pm 1} e^{-\beta H s_i} \right)^N = (e^{\beta H s_i} + e^{-\beta H s_i})^N = 2^N \cosh N(\beta H)
\end{aligned} \tag{A2.3.367}$$

where the third step follows from the identity of all the N lattice sites. The magnetization per site

$$m(H, T) = \frac{1}{N\beta} \left(\frac{\partial \ln Z}{\partial H} \right)_T = \tanh(\beta H) \quad (\text{A2.3.368})$$

and the graph of $m(H, T)$ versus H is a symmetrical sigmoid curve through the origin with no residual magnetization at any temperature when $H = 0$. This is because there are no interactions between the sites. We will see that a modification of the local field at a site that includes, even approximately, the effect of interactions between the sites leads to a critical temperature and residual magnetization below this temperature.

The local field at site i in a given configuration is

$$H_i = J \sum_{(j)} s_j + H = qJ \langle s_i \rangle + H - J \sum_{(j)} (s_j - \langle s_i \rangle) \quad (\text{A2.3.369})$$

where the last term represents a fluctuation from the average value of the spin $\langle s_i \rangle$ at site i which is the magnetization $m(H, T)$ per site. In the mean-field theory, this fluctuation is ignored and the effective mean field at all sites is

$$H_{\text{eff}} = qJm(H, T) + H. \quad (\text{A2.3.370})$$

Substituting this in the expressions for the PF for non-interacting magnets with the external field replaced by the effective field H_{eff} we have

$$Z_{\text{eff}}(N, H, T) = 2^N \cosh^N \beta[qJm(H, T) + H] \quad (\text{A2.3.371})$$

and by differentiation with respect to H ,

$$m(H, T) = \tanh[\beta(qJm(H, T) + H)] \quad (\text{A2.3.372})$$

from which it follows that

$$\left(\frac{1+m}{1-m} \right) = \exp[2\beta(qJm + H)]. \quad (\text{A2.3.373})$$

Since $dG = -S dT - M dH$, integration with respect to H yields the free energy per site:

$$\begin{aligned} \frac{G}{N} &= \frac{kT}{2} \ln \left[\frac{(1-m^2)}{4} \right] + \frac{qJm^2}{2} \\ &= - \left(\frac{Jqm^2}{2} + mH \right) + \frac{kT}{2} \left[(1+m) \ln \left(1 + \frac{m}{2} \right) + (1-m) \ln \left(1 - \frac{m}{2} \right) \right] \end{aligned} \quad (\text{A2.3.374})$$

where the first two terms represent the energy contribution, and the last term is the negative of the temperature

times the contribution of the entropy to the free energy. It is apparent that this entropy contribution corresponds to ideal mixing.

At zero field ($H = 0$),

$$m(0, T) = \tanh[\beta q J m(0, T)] \quad (\text{A2.3.375})$$

which can be solved graphically by plotting $\tanh[\beta q J m(0, T)]$ versus $m(0, T)$ and finding where this cuts the line through the origin with a slope of one, see figure A2.3.24.

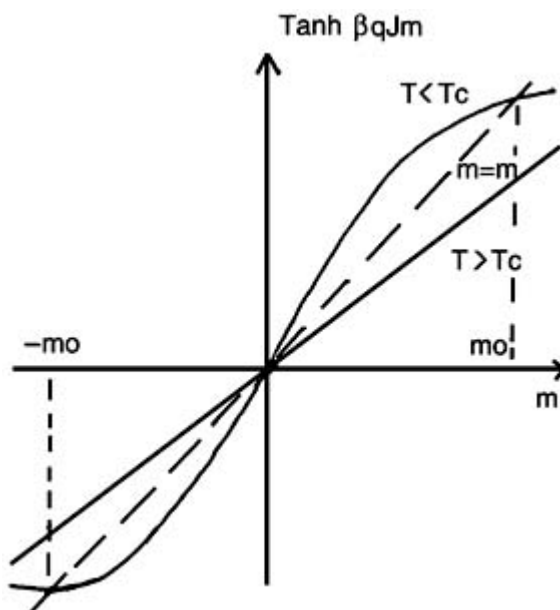


Figure A2.3.24 Plot of $\tanh[\beta q J m(0, T)]$ versus $m(0, T)$ at different temperatures.

Since

$$\tanh(\beta q J m) = \beta q J m - (1/3)(\beta q J m)^3 + \dots \quad (\text{A2.3.376})$$

the slope as $m \rightarrow 0$ is $\beta q J$. A solution exists for the residual magnetization when the slope is greater than 1. This implies that the critical temperature $T_c = qJ/k$, which depends on the coordination number q and is independent of the dimensionality. [Table A2.3.5](#) compares the critical temperatures predicted by mean-field theory with the ‘exact’ results. Mean-field theory is seriously in error for 1D systems but its accuracy improves with the dimensionality. For $D \geq 4$ it is believed to be exact.

It follows from our [equation \(A2.3.373\)](#) that

$$H = \frac{1}{2\beta} \ln \left(\frac{1+m}{1-m} \right) - q J m. \quad (\text{A2.3.377})$$

The magnetization is plotted as a function of the field in figure A2.3.25.

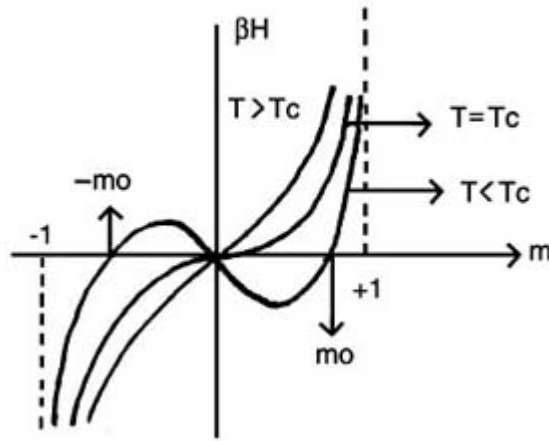


Figure A2.3.25 The magnetic field versus the magnetization $m(H, T)$ at different temperatures.

When $T > T_c$, $m = 0$ and the susceptibility at zero field $\chi_T(0) > 0$. At $T = T_c$, $(dH/dm)_{T, H=0} = 0$ which implies that the susceptibility diverges, i.e. $\chi_T(0) = \infty$.

When $T < T_c$, the graph of H versus m shows a van der Waals like loop, with an unstable region where the susceptibility $\chi_T(0) < 0$. In the limit $H \rightarrow 0$, there are three solutions for the residual magnetization $m = (-m_0, 0, m_0)$, of which the solution $m = 0$ is rejected as unphysical since it lies in the unstable region. The symmetrically disposed acceptable solutions for the residual magnetizations are solutions to

-108-

$$\begin{aligned} m_0 &= \tanh[\beta q J m_0] \\ &= \tanh[(T_c/T)m_0] = (T_c/T)m_0 + (1/3)[(T_c/T)m_0]^3 + \dots \end{aligned}$$

from which it follows that as $T \rightarrow T_c$

$$m_0 \simeq 3^{1/2}(T/T_c)[1 - T/T_c]^{1/2}. \quad (\text{A2.3.378})$$

This shows that the critical exponent $\beta = 1/2$.

The susceptibility at finite field H is given by

$$\chi_T(H) = (\partial m / \partial H)_T = \frac{\beta(1 - m^2)}{1 - \beta q J(1 - m^2)}. \quad (\text{A2.3.379})$$

Recalling that $\beta q J = T_c/T$, the susceptibility at zero field ($H \rightarrow 0$)

$$\chi_T(0) = \frac{(1 - m_0^2)}{k[(T - T_c) + T_c m_0^2]}. \quad (\text{A2.3.380})$$

For $T > T_c$, $m_0 = 0$ and

$$\chi_T(0) \simeq A_1(T - T_c)^{-\gamma} = (1/k)(T - T_c)^{-1} \quad (\text{A2.3.381})$$

which shows that the critical exponent $\gamma = 1$ and the amplitude A_1 of the divergence of $\chi_T(0)$ is N/k , when the critical point is approached from above T_c . When $T < T_c$, $m_0^2 \simeq 3(T_c - T)/T_c$ and

$$\chi_T(0) \simeq A_2(T - T_c)^{-\gamma} = (1/2k)(T - T_c)^{-1} \quad (\text{A2.3.382})$$

which shows that the critical exponent γ remains the same but the amplitude A_2 of the divergence is $1/2k$ when the critical point is approached from below T_c . This is half the amplitude when T_c is approached from above.

Along the critical isotherm, $T = T_c$,

$$\begin{aligned} H &= kT_c(1/2) \ln[(1 + m_0)/(1 - m_0) - m_0] \\ &\approx kT_c[m_0 + m_0^3/3 + m_0^5/5 + \dots - m_0] \approx kT_c m_0^3/3 + \dots \end{aligned} \quad (\text{A2.3.383})$$

It follows that the critical exponent δ defined by $H \approx m_0^\delta$ is 3.

-109-

Fluctuations in the magnetization are ignored by mean-field theory and there is no correlation between neighbouring sites, so that

$$\langle s_i s_j \rangle = \langle s_i \rangle \langle s_j \rangle \quad (\text{A2.3.384})$$

and the spins are randomly distributed over the sites. As seen earlier, the entropy contribution to the free energy is that of ideal mixing of up and down spins. The average energy

$$\begin{aligned} \langle U_N \rangle &= -J \left\langle \sum_{(ij)} s_i s_j \right\rangle - H \left\langle \sum_i s_i \right\rangle \\ &= J \sum_{(ij)} \langle s_i \rangle \langle s_j \rangle - H \sum_i \langle s_i \rangle \\ &= -J(Nq/2)m^2 - HNm \end{aligned} \quad (\text{A2.3.385})$$

which is in accord with our interpretation of the terms contributing to the free energy in the mean-field approximation. Since $m = 0$ at zero field for $T > T_c$ and $m = \pm m_0$ at zero field when $T < T_c$, the configurational energy at zero field ($H = 0$) is given by

$$\langle U_N(H = 0) \rangle = \begin{cases} 0 & T > T_c \\ J(Nq/2)m_0^2 & T < T_c. \end{cases} \quad (\text{A2.3.386})$$

This shows very clearly that the specific heat has a jump discontinuity at $T = T_c$:

$$(\text{A2.3.387})$$

$$C_{H=0}(T) = (1/N)\langle \partial U_N(H=0)/\partial T \rangle$$

$$= \begin{cases} 0 & T > T_c \\ -qJm_0(dm_0/dT) & T < T_c. \end{cases}$$

The neglect of fluctuations in mean-field theory implies that

$$[\downarrow\downarrow] \propto [\downarrow]^2 \quad [\uparrow\uparrow] \propto [\uparrow]^2 \quad [\uparrow\downarrow] \propto [\uparrow][\downarrow] \quad (\text{A2.3.388})$$

and it follows that

$$\frac{[\downarrow\downarrow][\uparrow\uparrow]}{[\uparrow\downarrow]^2} = \frac{1}{4}. \quad (\text{A2.3.389})$$

-110-

This is the equilibrium constant for the ‘reaction’



assuming the energy change is zero. An obvious improvement is to use the correct energy change ($4J$) for the ‘reaction’, when

$$\frac{[\downarrow\downarrow][\uparrow\uparrow]}{[\uparrow\downarrow]^2} = \frac{1}{4} \exp(4J/kT). \quad (\text{A2.3.391})$$

This is the quasi-chemical approximation introduced by Fowler and Guggenheim [98] which treats the nearest-neighbour pairs of sites, and not the sites themselves, as independent. It is exact in one dimension. The critical temperature in this approximation is

$$T_c = (2J/k)[1/\ln(q/(q-2))] \quad (\text{A2.3.392})$$

which predicts the correct result of $T_c = 0$ for the 1D Ising model, and better estimates than mean-field theory, as seen in table A2.3.5, for the same model in two and three dimensions ($d = 2$ and 3). Bethé [99] obtained equivalent results by a different method. Mean-field theory now emerges as an approximation to the quasi-chemical approximation, but the critical exponents in the quasi-chemical approximation are still the classical values. [Figure A2.3.26](#) shows mean-field and quasi-chemical approximations for the specific heat and residual magnetization of a square lattice ($d = 2$) compared to the exact results.

Table A2.3.5 Critical temperatures predicted by mean-field theory (MFT) and the quasi-chemical (QC) approximation compared with the exact results.

kT_c/J

D	q	MFT	QC	Exact
1	2	2	0	0
2	4(sq)	4	2.88	2.27
3	6(sc)	6	4.93	4.07

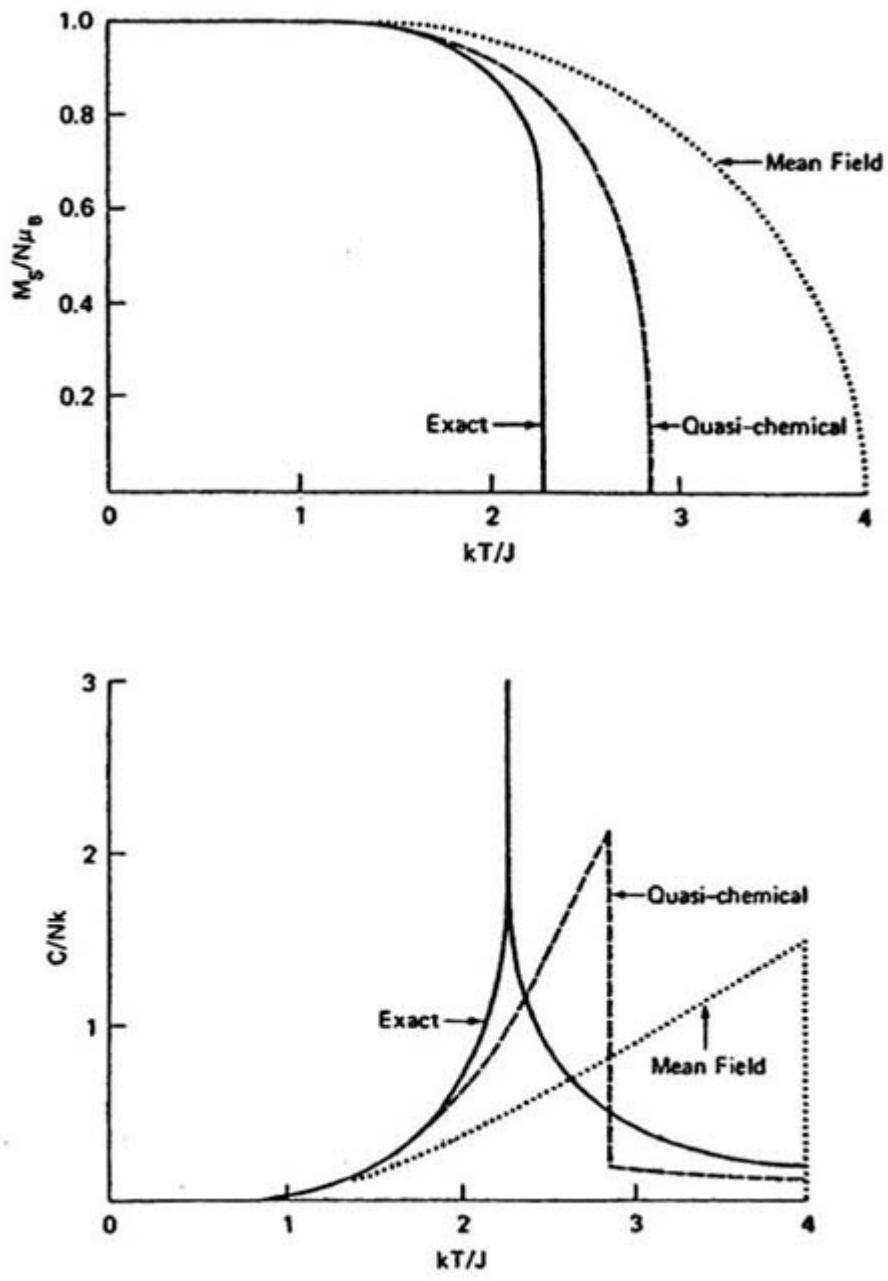


Figure A2.3.26 Mean-field and quasi-chemical approximations for the specific heat and residual magnetization of a square lattice ($d = 2$) compared to the exact results.

A2.3.8.1 LANDAU'S GENERALIZED MEAN-FIELD THEORY

An essential feature of mean-field theories is that the free energy is an analytical function at the critical point. Landau [100] used this assumption, and the up-down symmetry of magnetic systems at zero field, to analyse their phase behaviour and determine the mean-field critical exponents. It also suggests a way in which mean-field theory might be modified to conform with experiment near the critical point, leading to a scaling law, first proposed by Widom [101], which has been experimentally verified.

Assume that the free energy can be expanded in powers of the magnetization m which is the order parameter. At zero field, only even powers of m appear in the expansion, due to the up-down symmetry of the system, and

$$G = G_0 + a_2 m^2 + a_4 m^4 + \dots \tag{A2.3.393}$$

where the coefficients a_i are temperature dependent and $a_4 > 0$ but a_2 may be positive, negative or zero as illustrated in figure A2.3.27.

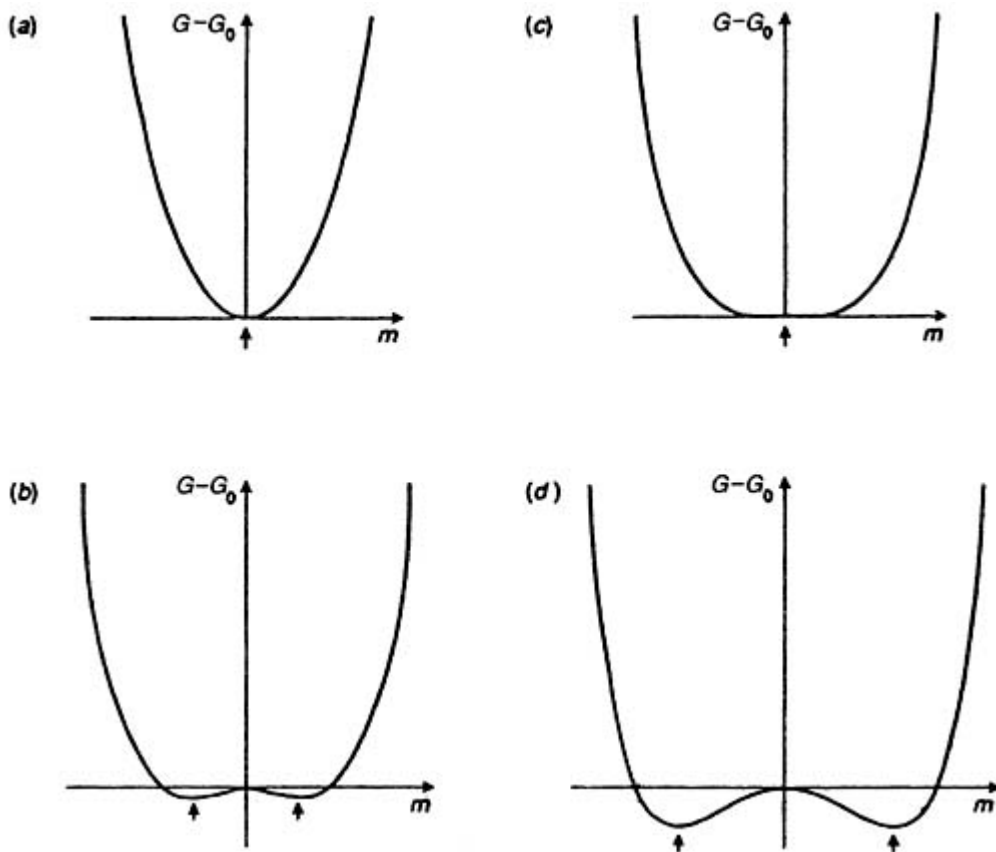


Figure A2.3.27 The free energy as a function of m in the Landau theory for (a) $a_2 > 0$, (b) $a_2 = 0$, (c) $a_2 < 0$ and (d) $a_2 < 0$ with $a_4 > 0$.

A finite residual magnetization of $\pm m_0$ is obtained only if $a_2 < 0$; and the critical temperature corresponds to $a_2 = 0$. Assume that a_2 is linear in $t = (T - T_c)/T_c$, near T_c , when

$$a_2 \approx a_2^* t. \quad (\text{A2.3.394})$$

The free energy expansion reads

$$G = G_0 + a_2^* t m^2 + a_4 m^4 + \dots. \quad (\text{A2.3.395})$$

The equilibrium magnetization corresponds to a minimum free energy which implies that

$$(dG/dm) = 0 = 2a_2^* t m + 4a_4 m^3 + \dots.$$

It follows that

$$m_0^2 = [a_2^*/(2a_4)](-t) \quad (\text{A2.3.396})$$

which implies that m_0 is real and finite if $t < 0$ (i.e. $T < T_c$) and the critical exponent $\beta = 1/2$. For t the only solution is $m_0 = 0$. Hence m_0 changes continuously with t and is zero for $t \geq 0$.

Along the coexistence curve, $t < 0$,

$$\begin{aligned} G &= G_0 + a_2^* t m_0^2 + a_4 m_0^4 \\ &= G_0 - a_2^* t^2 / (2a_4) + O(t^4) \end{aligned} \quad (\text{A2.3.397})$$

and the specific heat

$$C_{H=0} = \frac{1}{T} \left(\frac{\partial^2 G}{\partial T^2} \right) = \begin{cases} 0 & t > 0 \\ \frac{1}{T_c} \left(\frac{a_2^{*2}}{a_4} \right) & t < 0. \end{cases} \quad (\text{A2.3.398})$$

There is jump discontinuity in the specific heat as the temperature passes from below to above the critical temperature.

To determine the critical exponents γ and δ , a magnetic interaction term $-hm$ is added to the free energy and

$$G = G_0 - hm + a_2^* t m^2 + a_4 m^4 + \dots. \quad (\text{A2.3.399})$$

-114-

Minimizing the free energy with respect to m , one finds

$$h = 2a_2^* t m + 4a_4 m^3 + \dots. \quad (\text{A2.3.400})$$

Along the critical isochore, $t = 0$ and

$$h \approx 4a_4 m^3 \quad (\text{A2.3.401})$$

which implies that $\delta = 3$.

For $t \neq 0$, the inverse susceptibility

$$\chi_T(0)^{-1} = (dh/dm)_t = 2a_2^*t + 12a_4m^3 \quad (\text{A2.3.402})$$

which, as $h \rightarrow 0$, leads to

$$\chi_T(0)^{-1} = (dh/dm)_t = 2a_2^*t + 12a_4m^3. \quad (\text{A2.3.403})$$

When $t > 0$, $m_0 = 0$ and

$$\chi_T(0)_{H=0} = (1/2a_2^*)t^{-1} \quad (\text{A2.3.404})$$

while for $t < 0$, $m_0^2 = [a_2^*/(2a_4)](-t)$ and

$$\chi_T(0)_{H=0} = -(1/4a_2^*)t^{-1}. \quad (\text{A2.3.405})$$

This implies that the critical exponent $\gamma = 1$, whether the critical temperature is approached from above or below, but the amplitudes are different by a factor of 2, as seen in our earlier discussion of mean-field theory. The critical exponents are the classical values $\alpha = 0$, $\beta = 1/2$, $\delta = 3$ and $\gamma = 1$.

The assumption that the free energy is analytic at the critical point leads to classical exponents. Deviations from this require that this assumption be abandoned. In mean-field theory,

$$h = am(t + bm^3) + \dots \quad (\text{A2.3.406})$$

-115-

near the critical point, which implies $\beta = 1/2$ and $\gamma = 1$. Modifying this to

$$h = am(t + bm^{1+\gamma/\beta}) \quad (\text{A2.3.407})$$

implies that $\delta = 1 + \gamma/\beta$, which is correct. Widom postulated that

$$h = m\phi(t, m^{1/\beta}) \quad (\text{A2.3.408})$$

where ϕ is a generalized homogeneous function of degree γ

$$\phi(\lambda t, (\lambda m)^{1/\beta}) = \lambda^\gamma \phi(t, m^{1/\beta}). \quad (\text{A2.3.409})$$

This is Widom's scaling assumption. It predicts a scaled equation of state, like the law of corresponding states, that has been verified for fluids and magnets [102].

A2.3.9 HIGH- AND LOW-TEMPERATURE EXPANSIONS

Information about the behaviour of the 3D Ising ferromagnet near the critical point was first obtained from high- and low-temperature expansions. The expansion parameter in the high-temperature series is $\tanh K$, and the corresponding parameter in the low-temperature expansion is $\exp(-2K)$. A 2D square lattice is self-dual in the sense that the bisectors of the line joining the lattice points also form a square lattice and the coefficients of the two expansions, for the 2D square lattice system, are identical to within a factor of two. The singularity occurs when

$$\tanh K = \exp(-2K). \quad (\text{A2.3.410})$$

Kramers and Wannier [103] used this to locate the critical temperature $T_c = 2.27J/k$.

A2.3.9.1 THE HIGH-TEMPERATURE EXPANSION

The PF at zero field

$$Z(N, 0, T) = \sum_{\{s\}} \exp \sum_{\langle ij \rangle} K s_i s_j = \sum_{\{s\}} \prod_{\langle ij \rangle} \exp K s_i s_j \quad (\text{A2.3.411})$$

-116-

where $K = \beta J$ and $\{s\}$ implies summation over the spins on the lattice sites. Since $s_i s_j = \pm 1$,

$$\begin{aligned} \exp(K s_i s_j) &= \exp(\pm K) = \cosh K \pm \sinh K \\ &= \cosh K + s_i s_j \sinh K \\ &= \cosh K [1 + s_i s_j \tanh K] \end{aligned} \quad (\text{A2.3.412})$$

from which it follows that

$$\begin{aligned} Z(N, 0, T) &= (\cosh K)^{qN/2} \sum_{\{s\}} \prod_{\langle ij \rangle} (1 + s_i s_j \tanh K) \\ &= (\cosh K)^{qN/2} \sum_{\{s\}} \left[1 + \tanh K \sum_l s_i s_j \right. \\ &\quad \left. + \tanh^2 K \sum_l (s_i s_j)(s_k s_l) + \dots \right] \end{aligned} \quad (\text{A2.3.413})$$

where Σ_l is the sum over all possible sets of l pairs of nearest-neighbour spins. The expansion parameter is $\tanh K$ which $\rightarrow 0$ as $T \rightarrow \infty$ and becomes 1 as $T \rightarrow 0$. The expansion coefficients can be expressed graphically. A coefficient $(s_i s_j)(s_k s_l) \dots (s_p s_q)$ of $\tanh^r K$ is the product or sum of products of graphs with r bonds in which each bond is depicted as a line joining two sites. Note also that

$$\sum_{s_i = \pm 1} s_i^n = \begin{cases} 2 & n \text{ even} \\ 0 & n \text{ odd.} \end{cases} \quad (\text{A2.3.414})$$

Hence, on summing over the graphs, the only non-zero terms are closed polygons with an even number of bonds at each site, i.e. s_i must appear an even number of times at a lattice site in a graph that does not add up to zero on summing over the spins on the sites.

Each lattice point extraneous to the sites connected by graphs also contributes a factor of two on summing over spin states. Hence all lattice points contribute a factor of 2^N whether they are connected or not, and

$$Z(N, 0, T) = (\cosh K)^{qN/2} 2^N \sum_{r=0}^{qN/2} n(r, N) \tanh^r K \quad (\text{A2.3.415})$$

where (for $r \neq 0$), $n(r, N)$ is the number of distinct-side polygons (closed graphs) drawn on N sites such that there are an even number of bonds on each site. For $r = 0$, no lattice site is connected, but define $n(0, N) = 1$. Also, since closed polygons cannot be connected on one or two sites, $n(1, N) = n(2, N) = 0$. The problem then is to count $n(r, N)$ for all r . On an infinite lattice of identical sites, $n(r, N) = Np(r)$ where $p(r)$ is the number of r -side polygons that can be constructed on a given lattice site. This number is closely connected to the structure of the lattice.

-117-

A2.3.9.2 THE LOW-TEMPERATURE EXPANSION

At zero field (see [equation \(A2.3.335\)](#)),

$$\sum_{(ij)} K s_i s_j = [\uparrow\uparrow] + [\downarrow\downarrow] - [\uparrow\downarrow] = qN/2 - 2[\uparrow\downarrow] \quad (\text{A2.3.416})$$

and

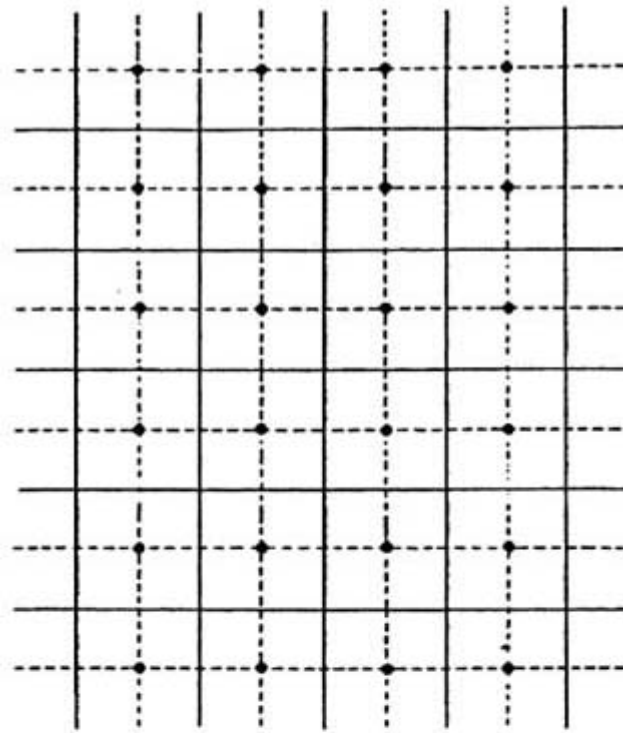
$$\begin{aligned} Z(N, 0, T) &= \exp(qN/2) \sum_{[\uparrow\downarrow]} \exp(-2K[\uparrow\downarrow]) \\ &= \exp(qN/2) \sum_r m(r, N) (-2Kr) \end{aligned} \quad (\text{A2.3.417})$$

where $m(r, N)$ is the number of configurations with $[\uparrow\downarrow] = r$. The high- and low-temperature expansions are complementary.

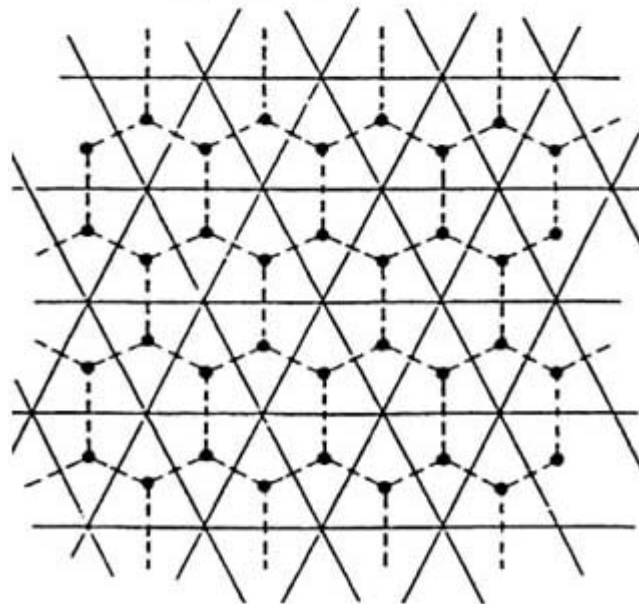
The dual lattice is obtained by drawing the bisectors of lines connecting neighbouring lattice points. Examples of lattices in two dimensions and their duals are shown in [figure A2.3.28](#). A square lattice is self-dual.

Consider a closed polygon which appears in the high-temperature expansion. Put up spins $[\uparrow]$ on the sites of the lattice inside the polygon and down spins $[\downarrow]$ on the lattice sites outside. The spins across the sides of the closed polygons are oppositely paired. The PF can be calculated equally well by counting closed polygons on the original lattice (high-temperature expansion) or oppositely paired spins on the dual lattice (low-temperature expansion). Both expansions are exact.

-118-



Square lattice and its dual



Triangular lattice and its hexagonal dual lattice

Figure A2.3.28 Square and triangular lattices and their duals. The square lattice is self-dual.

For a 2D square lattice $q = 4$, and the high- and low-temperature expansions are related in a simple way

$$2n(r, N) = m(r, N)^{\text{dual}} \tag{A2.3.418}$$

where the factor two comes from the fact that reversing all the spins does not change the number of oppositely paired spins. The dual of the lattice is a square lattice. Hence the PFs in the two expansions have the same

coefficients except for irrelevant constant factors:

$$\begin{aligned} Z(N, 0, T) &= 2^N (\cosh K)^{2N} \sum_{r=0}^{2N} n(r, N) \tanh K \\ &= 2 \exp(2KN) \sum_{r=0}^{2N} n(r, N) \exp(-2Kr). \end{aligned} \quad (\text{A2.3.419})$$

Both expansions are exact and assuming there is only one singularity, identified with the critical point, this must occur when

$$\tanh K = \exp(-2K). \quad (\text{A2.3.420})$$

With $x = \exp(-2K)$, this implies that

$$x = (1 - x)/(1 + x)$$

which leads to a quadratic equation

$$x^2 + 2x - 1 = 0. \quad (\text{A2.3.421})$$

The solutions are $x = -1 \pm \sqrt{2}$. Since $K = \beta J$ is necessarily not negative, the only acceptable solution is $x = -1 + \sqrt{2}$. Identifying the singularity with the critical point, the solution $x = \exp(2K_c) = -1 + \sqrt{2}$ is equivalent to the condition

$$\sinh(2K_c) = \sinh(2J/kT_c) = 1 \quad (\text{A2.3.422})$$

from which it follows that the critical temperature $T_c = 2.27J/k$. This result was known before Onsager's solution to the 2D Ising model at zero field.

More generally, for other lattices and dimensions, numerical analysis of the high-temperature expansion provides information on the critical exponents and temperature. The high-temperature expansion of the susceptibility may be written in powers of $K = \beta J$ as

-120-

$$\chi_T(0) = \sum_{n=0}^{\infty} a_n (\beta J)^n. \quad (\text{A2.3.423})$$

Suppose the first $n + 1$ coefficients are known, where $n \simeq 15$. The susceptibility diverges as $(1 - \beta/\beta_c)^{-\gamma}$ as $\beta \rightarrow \beta_c$ - and we have

$$\begin{aligned} \chi_T(0) &\approx A(1 - \beta/\beta_c)^{-\gamma} \\ &= A \left[1 + \gamma(\beta/\beta_c) + \dots + \frac{\gamma(\gamma + 1) \dots (\gamma + n - 1)}{n!} (\beta/\beta_c)^n + \dots \right]. \end{aligned} \quad (\text{A2.3.424})$$

For large n

$$a_n J^n \approx A \frac{\gamma(\gamma+1)\dots(\gamma+n-1)}{n! \beta_c^n} \quad (\text{A2.3.425})$$

Taking the ratio of successive terms and dividing by the coordination number q

$$r_n = \frac{a_n}{q a_{n-1}} = \frac{kT_c}{qJ} \left(1 + \frac{\gamma-1}{n} \right) \quad (\text{A2.3.426})$$

Plotting r versus $1/n$ gives kT_c/qJ as the intercept and $(kT_c/qJ)(1-\gamma)$ as the slope from which T_c and γ can be determined. Figure A2.3.29 illustrates the method for lattices in one, two and three dimensions and compares it with mean-field theory which is independent of the dimensionality.

-121-

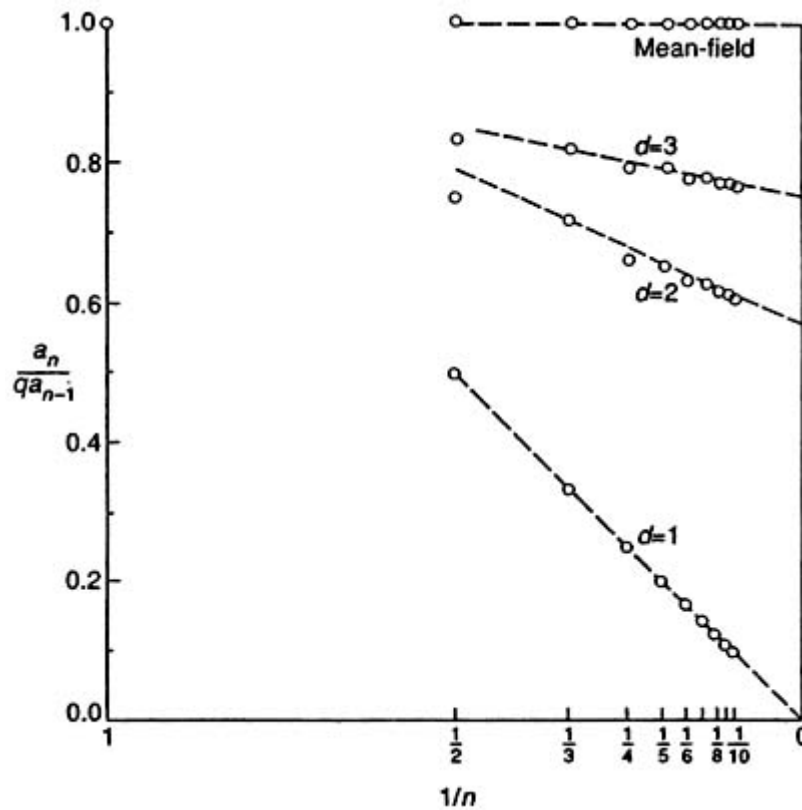


Figure A2.3.29 Calculation of the critical temperature T_c and the critical exponent γ for the magnetic susceptibility of Ising lattices in different dimensions from high-temperature expansions.

A2.3.10 EXACT SOLUTIONS TO THE ISING MODEL

The Ising model has been solved exactly in one and two dimensions; Onsager's solution of the model in two dimensions is only at zero field. Information about the Ising model in three dimensions comes from high- and low-temperature expansions pioneered by Domb and Sykes [104] and others. We will discuss the solution to the 1D Ising model in the presence of a magnetic field and the results of the solution to the 2D Ising model at zero field.

A2.3.10.1 ONE DIMENSION

We will describe two cases: open and closed chains of N sites. For an open chain of N sites, the energy of a spin configuration $\{s_k\}$ is

$$U_N(\{s_k\}) = -J \sum_{i=1}^{N-1} s_i s_{i+1} - H \sum_{i=1}^N s_i \quad (\text{A2.3.427})$$

-122-

and for a closed chain of N sites with periodic boundary conditions $s_{N+1} = s_1$

$$U_N(\{s_k\}) = -J \sum_{i=1}^N s_i s_{i+1} - \frac{H}{2} \sum_{i=1}^N (s_i + s_{i+1}). \quad (\text{A2.3.428})$$

Both systems give the same results in the thermodynamic limit. We discuss the solution for the open chain at zero field and the closed chain for the more general case of $H \neq 0$.

(a) *Open chain at zero field, i.e. $H = 0$*

The PF

$$\begin{aligned} Z(N, 0, T) &= \sum_{s_1=\pm 1} \dots \sum_{s_N=\pm 1} \exp\left(\beta J \sum_{i=1}^{N-1} s_i s_{i+1}\right) \\ &= \sum_{s_1=\pm 1} \dots \sum_{s_N=\pm 1} \exp\left(\beta J \sum_{i=1}^{N-2} s_i s_{i+1}\right) \sum_{s_N=\pm 1} \exp(\beta J s_{N-1} s_N). \end{aligned} \quad (\text{A2.3.429})$$

Doing the last sum

$$\begin{aligned} Z(N, 0, T) &= Z(N-1, 0, T) [\exp(\beta J s_{N-1}) + \exp(\beta J s_{N-1})] \\ &= Z(N-1, 0, T) 2 \cosh(\beta J) \end{aligned} \quad (\text{A2.3.430})$$

since $s_{N-1} = \pm 1$. Proceeding by iteration, starting from $N=1$, which has just two states with the spin up or down

$$\begin{aligned} Z(1, 0, T) &= 2 \\ Z(2, 0, T) &= Z(1, 0, T) 2 \cosh(\beta J) = 2^2 \cosh(\beta J) \\ Z(3, 0, T) &= 2^3 \cosh^2(\beta J) \\ &\dots \\ Z(N, 0, T) &= 2^N \cosh^{N-1}(\beta J). \end{aligned} \quad (\text{A2.3.431})$$

The free energy G in the thermodynamic limit ($N \rightarrow \infty$) follows from

$$\begin{aligned}
-\frac{\beta G}{N} &= \lim_{N \rightarrow \infty} \frac{1}{N} \ln Z(N, 0, T) \\
&= \ln 2 + \lim_{N \rightarrow \infty} \left(\frac{N-1}{N} \right) \ln \cosh(\beta J) = \ln[2 \cosh(\beta J)].
\end{aligned}
\tag{A2.3.432}$$

-123-

(b) Closed chain, $H \neq 0$

The PF in this case is

$$\begin{aligned}
Z(N, H, T) &= \sum_{s_1 = \pm 1} \dots \sum_{s_N = \pm 1} \exp \left[\left(\beta J \sum_{k=1}^N s_k s_{k+1} \right) + \frac{\beta H}{2} \sum_{k=1}^N (s_k + s_{k+1}) \right] \\
&= \sum_{s_1 = \pm 1} \dots \sum_{s_N = \pm 1} \prod_{k=1}^N \exp \beta \left[J s_k s_{k+1} + \frac{H}{2} (s_k + s_{k+1}) \right] \\
&= \sum_{s_1 = \pm 1} \dots \sum_{s_N = \pm 1} P_{s_1 s_2} P_{s_2 s_3} \dots P_{s_N s_1}
\end{aligned}
\tag{A2.3.433}$$

where $P_{s_1 s_2}$ are the elements of a 2×2 matrix called the transfer matrix

$$\mathbf{P} = \begin{pmatrix} P_{11} & P_{1-1} \\ P_{-11} & P_{-1-1} \end{pmatrix} = \begin{pmatrix} \exp \beta(J + H) & \exp(-\beta J) \\ \exp(-\beta J) & \exp \beta(J - H) \end{pmatrix}
\tag{A2.3.434}$$

with the property that $\sum_{s_2} P_{s_1 s_2} P_{s_2 s_3} = (\mathbf{P}^2)_{s_1 s_3}$. It follows for the closed chain that

$$Z(N, H, T) = \sum_{s_1 = \pm 1} (\mathbf{P}^N)_{s_1 s_1} = \text{Tr} \mathbf{P}^N
\tag{A2.3.435}$$

where \mathbf{P}^N is also a 2×2 matrix.

The trace is evaluated by diagonalizing the matrix \mathbf{P} using a similarity transformation \mathbf{S} :

$$\mathbf{P}' = \mathbf{S}^{-1} \mathbf{P} \mathbf{S} = \begin{pmatrix} \lambda_+ & 0 \\ 0 & \lambda_- \end{pmatrix}
\tag{A2.3.436}$$

where the diagonal elements of the matrix \mathbf{P}' are the eigenvalues of \mathbf{P} , and

$$\mathbf{P}'^N = \begin{pmatrix} \lambda_+^N & 0 \\ 0 & \lambda_-^N \end{pmatrix}.
\tag{A2.3.437}$$

-124-

Noting that

$$\mathbf{P}'^N = \mathbf{S}^{-1} \mathbf{P} \mathbf{S} \mathbf{S}^{-1} \mathbf{P} \mathbf{S} \dots \mathbf{S}^{-1} \mathbf{P} \mathbf{S} = \mathbf{S}^{-1} \mathbf{P}^N \mathbf{S}$$

by virtue of the property that $\mathbf{S} \mathbf{S}^{-1} = \mathbf{I}$, where \mathbf{I} is the identity matrix, we see that

$$\text{Tr}[\mathbf{P}'^N] = \text{Tr}[\mathbf{S}^{-1} \mathbf{P}^N \mathbf{S}] = \text{Tr}[\mathbf{S}^{-1} \mathbf{S} \mathbf{P}^N] = \text{Tr}[\mathbf{P}^N]$$

which leads to

$$Z(N, H, T) = \lambda_+^N + \lambda_-^N. \quad (\text{A2.3.438})$$

Assuming the eigenvalues are not degenerate and $\lambda_+ > \lambda_-$,

$$Z(N, H, T) = \lambda_+^N [1 + (\lambda_-/\lambda_+)^N].$$

In the thermodynamic limit of $N \rightarrow \infty$,

$$\frac{-\beta G}{N} = \lim_{N \rightarrow \infty} \frac{1}{N} \ln Z(N, H, T) = \ln \lambda_+. \quad (\text{A2.3.439})$$

This is an important general result which relates the free energy per particle to the largest eigenvalue of the transfer matrix, and the problem reduces to determining this eigenvalue.

The eigenvalues of the transfer matrix are the solutions to

$$\det |\mathbf{P} - \lambda \mathbf{I}| = 0.$$

This leads to a quadratic equation whose solutions are

$$\lambda_{\pm} = \exp(\beta J) \{ \cosh(\beta H) \pm [\sinh^2(\beta H) + \exp(-4\beta J)]^{1/2} \} \quad (\text{A2.3.440})$$

which confirms that the eigenvalues are not degenerate. The free energy per particle

$$\frac{-\beta G}{N} = \beta J + \ln \{ \cosh(\beta H) + [\sinh^2(\beta H) + \exp(-4\beta J)]^{1/2} \}. \quad (\text{A2.3.441})$$

-125-

This reduces to the results for the free energy at zero field ($H = 0$)

$$\frac{-\beta G}{N} = \ln [2 \cosh(\beta J)] \quad (\text{A2.3.442})$$

and the free energy of non-interacting magnets in an external field

(A2.3.443)

$$\frac{-\beta G}{N} = \ln[2 \cosh(\beta H)]$$

which were derived earlier. At finite T (i.e. $T > 0$), λ_+ is analytic and there is no phase transition. However, as $T \rightarrow 0$,

$$\begin{aligned} \lambda_+ &\rightarrow \exp(K)[\cosh(h) + (\sinh^2(h))^{1/2}(1 + O(\exp(-4K)))] \\ &= \exp(K)[\cosh(h) + |\sinh(h)|(1 + O(\exp(-4K)))] \end{aligned}$$

where $K = \beta J$ and $h = \beta H$. But $\cosh(h) + |\sinh(h)| = \exp|h|$, and it follows that,

$$\lambda_+ \rightarrow \exp(K + |h|)$$

as $T \rightarrow 0$. We see from this that as $T \rightarrow 0$

$$-\frac{G}{N} = kT \ln \lambda_+ = kT[K + |h|] = J + |H| \quad (\text{A2.3.444})$$

and

$$m = \frac{1}{N} \left(\frac{\partial G}{\partial H} \right)_T = \begin{cases} +1 & H > 0 \\ -1 & H < 0 \end{cases} \quad (\text{A2.3.445})$$

which implies a residual magnetization $m_0 = \pm 1$ at zero field and a first-order phase transition at $T = 0$. For $T \neq 0$, there is no discontinuity in m as H passes through zero from positive to negative values or *vice versa*, and differentiation of G with respect to H at constant T provides the magnetization per site

$$m(H, T) = \frac{\sinh(\beta H)}{[\sinh^2(\beta H) + \exp(-4\beta J)]^{1/2}} \quad (\text{A2.3.446})$$

which is an odd function of H with $m \rightarrow 0$ as $H \rightarrow 0$. Note that this reduces to the result

$$m(H, T) = \tanh(\beta H) \quad (\text{A2.3.447})$$

for non-interacting magnets.

-126-

As $H \rightarrow 0$, $\sinh(\beta J) \rightarrow \beta J$, $m(H, T) \rightarrow \beta H \exp(2\beta J)$ and

$$\begin{aligned} \text{As } H \rightarrow 0, \sinh(\beta J) &\rightarrow \beta J, m(H, T) \rightarrow \beta H \exp(2\beta J) \text{ and} \\ \chi_T(0) &= (dm/dH)_T = \beta \exp(2\beta J) \end{aligned} \quad (\text{A2.3.448})$$

which diverges exponentially as $T \rightarrow 0$, which is also characteristic of a phase transition at $T = 0$.

The average energy $\langle E \rangle$ follows from the relation

$$\langle E \rangle / N = -(1/N)(d \ln Z / d\beta)_{H,J} = -(d \ln \lambda_- / d\beta)_{H,J} \quad (\text{A2.3.449})$$

and at zero field

$$\langle E \rangle_{H=0} / N = -J \tanh(\beta J). \quad (\text{A2.3.450})$$

The specific heat at zero field follows easily,

$$C_{H=0} = -\frac{N}{kT^2} \left(\frac{\partial \langle E \rangle_{H=0}}{\partial \beta} \right) = Nk(\beta J)^2 \operatorname{sech}^2(\beta J) \quad (\text{A2.3.451})$$

and we note that it passes through a maximum as a function of T .

The spin correlation functions and their dependence on the distance between sites and the coupling between adjacent sites are of great interest in understanding the range of these correlations. In general, for a closed chain

$$\langle s_i s_{i+n} \rangle = Z(N, H, T)^{-1} \sum_{s_1=\pm 1} \dots \sum_{s_N=\pm 1} s_i s_{i+n} \exp \left(\sum_{j=1}^N K s_j s_{j+1} + h s_j \right). \quad (\text{A2.3.452})$$

For nearest-neighbour spins

$$\langle s_j s_{j+1} \rangle = [N Z(N, H, T)]^{-1} [dZ(N, H, T) / dK] \quad (\text{A2.3.453})$$

and making use of $Z(N, H, T) = \lambda_+^N [1 + (\lambda_- / \lambda_+)^N]$ in the thermodynamic limit ($N \rightarrow \infty$)

$$\begin{aligned} \langle s_i s_{i+1} \rangle &= (\partial \ln \lambda_+ / \partial K) \\ &= 1 - \frac{2 \exp(-4K) [\sinh^2 h + \exp(-4K)]^{-1/2}}{\cosh h + [\sinh^2 h + \exp(-4K)]^{1/2}}. \end{aligned} \quad (\text{A2.3.454})$$

-127-

At zero field ($H=0$), $h=0$ and

$$\langle s_i s_{i+1} \rangle = \tanh K \quad (\text{A2.3.455})$$

which shows that the correlation between neighbouring sites approaches 1 as $T \rightarrow 0$. The correlation between non-nearest neighbours is easily calculated by assuming that the couplings ($K_1, K_2, K_3, \dots, K_N$) between the sites are different, in which case a simple generalization of the results for equal couplings leads to the PF at zero field

$$Z(N, 0, T) = 2^N \prod_{j=1}^{N-1} \cosh K_j. \quad (\text{A2.3.456})$$

Repeating the earlier steps one finds, as expected, that the coupling K_i between the spins at the sites i and $i+1$ determines their correlation:

$$\langle s_i s_{i+1} \rangle = Z^{-1} (dZ(N, H, T)/dK_i) = \tanh K_i. \quad (\text{A2.3.457})$$

Now notice that since $s_{i+1}^2 = 1$,

$$\begin{aligned} \langle s_i s_{i+1} s_{i+1} s_{i+1} \rangle &= \langle s_i s_{i+2} \rangle = Z^{-1} \left(\frac{\partial^2 Z}{\partial K_i \partial K_{i+1}} \right) \\ &= \tanh K_i \tanh K_{i+1}. \end{aligned} \quad (\text{A2.3.458})$$

In the limit $K_i = K_{i+1} = K$,

$$\langle s_i s_{i+2} \rangle = \tanh^2 K \quad (\text{A2.3.459})$$

and repeating this argument serially for the spin correlations between i and $i + n$ sites

$$\langle s_i s_{i+n} \rangle = \tanh^n K \quad (\text{A2.3.460})$$

so the correlation between non-neighbouring sites approaches 1 as $T \rightarrow 0$ since the spins are all aligned in this limit.

The correlation length ζ follows from the above relation, since

$$\langle s_i s_{i+j} \rangle = \exp(j \ln \tanh K) = \exp(-j \ln \coth K) = \exp(-j/\zeta) \quad (\text{A2.3.461})$$

from which it follows that

$$\zeta = 1/\ln \coth(K). \quad (\text{A2.3.462})$$

-128-

As expected, as $T \rightarrow 0$, $K \rightarrow \infty$ and the correlation length $\zeta \approx \exp(\beta J)/2 \rightarrow \infty$, while in the opposite limit, as $T \rightarrow \infty$, $\zeta \rightarrow 0$.

A2.3.10.2 TWO DIMENSIONS

Onsager's solution to the 2D Ising model in zero field ($H = 0$) is one of the most celebrated results in theoretical chemistry [105]; it is the first example of critical exponents. Also, the solution for the Ising model can be mapped onto the lattice gas, binary alloy and a host of other systems that have Hamiltonians that are isomorphic to the Ising model Hamiltonian.

By a deft application of the transfer matrix technique, Onsager showed that the free energy is given by

$$-\frac{\beta G}{N} = \ln \cosh(2\beta J) + \frac{1}{2\pi} \int_0^\pi d\phi \ln \frac{[1 + (1 - \kappa^2 \sin^2 \phi)]^{1/2}}{2} \quad (\text{A2.3.463})$$

where

$$\kappa = \frac{2 \sinh(2\beta J)}{\cosh^2(2\beta J)} \quad (\text{A2.3.464})$$

which is zero at $T = 0$ and $T = \infty$ and passes through a maximum of 1 when $\beta J_c = 0.44069$. This corresponds to a critical temperature $T_c = 2.269 J/k$ when a singularity occurs in the Gibbs free energy, since $[1 + (1 - \kappa^2 \sin^2 \phi)^{1/2}] \rightarrow 0$ as $T \rightarrow T_c$ and $\phi \rightarrow \pi/2$. As $T \rightarrow T_c$,

$$C_{H=0} \approx \frac{8k}{\pi} \frac{J}{kT_c} \ln |T - T_c|^{-1} \quad (\text{A2.3.465})$$

so that the critical exponent $\alpha = 0_{\log}$. The spontaneous magnetization

$$m_0 = \begin{cases} 0 & T > T_c \\ [1 - \sinh^{-4}(2\beta J)]^{1/8} & T < T_c \end{cases} \quad (\text{A2.3.466})$$

and the critical exponent $\beta = 1/8$. This result was first written down by Onsager during a discussion at a scientific meeting, but the details of his derivation were never published. Yang [107] gave the first published proof of this remarkably simple result. The spin correlation functions at $T = T_c$ decay in a simple way as shown by Kaufman and Onsager [106],

$$\langle s_i s_{i+j} \rangle \sim 1/r^{1/4} \quad (\text{A2.3.467})$$

where r is the distance between the sites.

A2.3.11 SUMMARY

We have described the statistical mechanics of strongly interacting systems. In particular those of non-ideal fluids, solids and alloys. For fluids, the virial coefficients, the law of corresponding states, integral equation approximations for the correlation functions and perturbation theories are treated in some detail, along with applications to hard spheres, polar fluids, strong and weak electrolytes and inhomogeneous fluids. The use of perturbation theory in computational studies of the free energy of ligand binding and other reactions of biochemical interest is discussed. In treating solids and alloys, the Ising model and its equivalence to the lattice gas model and a simple model of binary alloys, is emphasized. Mean-field approximations to this model and the use of high- and low-temperature approximations are described. Solutions to the 1D Ising model with and without a magnetic field are derived and Onsager's solution to the 2D case is briefly discussed.

REFERENCES

- [1] Andrews T 1869 On the continuity of the gaseous and liquid states *Phil. Trans. R. Soc.* **159** 575
- [2] van der Waals J H 1873 Over de continuïteit van den gas-en vloeistof toestand *Thesis* University of Leiden (English transl. 1988 *Studies in Statistical Mechanics* ed J S Rowlinson (Amsterdam: North-Holland))

- [3] Maxwell J C 1874 Van der Waals on the continuity of the gaseous and liquid states *Nature* **10** 477
Maxwell J C 1875 On the dynamical evidence of the molecular constitution of bodies *Nature* **11** 357
- [4] Bett K E, Rowlinson J S and Saville G 1975 *Thermodynamics for Chemical Engineers* (Cambridge, MA: MIT Press)
- [5] Stell G 1964 Cluster expansions for classical systems in equilibrium *The Equilibrium Theory of Classical Fluids* ed H L Frisch and J L Lebowitz (New York: Benjamin)
- [6] Hansen J P and McDonald I 1976 *Theories of Simple Liquids* (New York: Academic)
- [7] Mayer J G and Mayer M G 1940 *Statistical Mechanics* (New York: Wiley)
- [8] Rhee F H and Hoover W G 1964 Fifth and sixth virial coefficients for hard spheres and hard disks *J. Chem. Phys.* **40** 939
Rhee F H and Hoover W G 1967 Seventh virial coefficients for hard spheres and hard disks *J. Chem. Phys.* **46** 4181
- [9] Carnahan N F and Starling K E 1969 Equation of state for nonattracting rigid spheres *J. Chem. Phys.* **51** 635
- [10] Harvey A N 1999 Applications of first-principles calculations to the correlation of water's second virial coefficient *Proc. 13th Int. Conf. of the Properties of Water and Steam (Toronto, 12–16 September 1999)*
- [11] Feynman R P 1972 *Statistical Mechanics, a Set of Lectures* (New York: Benjamin/Cummings)
-

-130-

- [12] Gillan M J 1990 Path integral simulations of quantum systems *Computer Modeling of Fluids and Polymers* ed C R A Catlow *et al* (Dordrecht: Kluwer)
- [13] Tonks L 1936 The complete equation of state of one, two and three dimensional gases of hard elastic spheres *Phys. Rev.* **50** 955
- [14] Takahashi H 1942 *Proc. Phys. Math. Soc. Japan* **24** 60
- [15] Lieb E H and Mattis D C 1966 *Mathematical Physics in One Dimension* (New York: Academic)
- [16] Cho C H, Singh S and Robinson G W 1996 An explanation of the density maximum in water *Phys. Rev. Lett.* **76** 1651
- [17] Kac M, Uhlenbeck G E and Hemmer P 1963 On van der Waals theory of vapor–liquid equilibrium. I. Discussion of a one-dimensional model *J. Math. Phys.* **4** 216
- [18] van Kampen N G 1964 Condensation of a classical gas with long-range attraction *Phys. Rev. A* **135** 362
- [19] Guggenheim E A 1945 The principle of corresponding states *J. Chem. Phys.* **13** 253
- [20] Onsager L 1933 Theories of concentrated electrolytes *Chem. Rev.* **13** 73
- [21] Kirkwood J G 1935 Statistical mechanics of fluid mixtures *J. Chem. Phys.* **3** 300
Kirkwood J G 1936 Statistical mechanics of liquid solutions *Chem. Rev.* **19** 275
- [22] Yvon J 1935 *Actualités Scientifiques et Industriel* (Paris: Herman et Cie)
- [23] Born M and Green H S 1946 A general kinetic theory of liquids: I. The molecular distribution functions *Proc. R. Soc. A* **188** 10

- [24] Born M and Green H S 1949 *A General Kinetic Theory of Liquids* (Cambridge: Cambridge University Press)
- [25] Kirkwood J G and Buff F P 1951 Statistical mechanical theory of solutions I *J. Chem. Phys.* **19** 774
- [26] Fisher M 1964 Correlation functions and the critical region of simple fluids *J. Math. Phys.* **5** 944
- [27] Percus J K 1982 Non uniform fluids *The Liquid State of Matter: Fluids, Simple and Complex* ed E W Montroll and J L Lebowitz (Amsterdam: North-Holland)
- [28] Percus J K and Yevick G J 1958 Analysis of classical statistical mechanics by means of collective coordinates *Phys. Rev.* **110** 1
- [29] Stell G 1977 Fluids with long-range forces: towards a simple analytic theory *Statistical Mechanics part A, Equilibrium Techniques* ed B Berne (New York: Plenum)
- [30] Chandler D and Andersen H C 1972 Optimized cluster expansions for classical fluids II. Theory of molecular liquids *J. Chem. Phys.* **57** 1930
- [31] Chandler D 1982 Equilibrium theory of polyatomic fluids *The Liquid State of Matter: Fluids, Simple and Complex* ed E W Montroll and J L Lebowitz (Amsterdam: North-Holland)
- [32] Wertheim M S 1963 Exact solution of the Percus–Yevick equation for hard spheres *Phys. Rev. Lett.* **10** 321
Wertheim M S 1964 Analytic solution of the Percus–Yevick equation *J. Math. Phys.* **5** 643

- [33] Thiele E 1963 Equation of state for hard spheres *J. Chem. Phys.* **39** 474
- [34] Baxter R J 1968 Ornstein Zernike relation for a disordered fluid *Aust. J. Phys.* **21** 563
- [35] Lebowitz J L 1964 Exact solution of the generalized Percus–Yevick equation for a mixture of hard spheres *Phys. Rev.* **133** A895
Lebowitz J L and Rowlinson J S 1964 Thermodynamic properties of hard sphere mixtures *J. Chem. Phys.* **41** 133
- [36] Baxter R J 1970 Ornstein Zernike relation and Percus–Yevick approximation for fluid mixtures *J. Chem. Phys.* **52** 4559
- [37] Reiss H, Frisch H I and Lebowitz J L 1959 Statistical mechanics of rigid spheres *J. Chem. Phys.* **31** 361
Reiss H 1977 Scaled particle theory of hard sphere fluids *Statistical Mechanics and Statistical Methods in Theory and Application* ed U Landman (New York: Plenum) pp 99–140
- [38] Gibbons R M 1969 Scaled particle theory for particles of arbitrary shape *Mol. Phys.* **17** 81
- [39] Chandler D 1993 Gaussian field model of fluids with an application to polymeric fluid *Phys. Rev. E* **48** 2989
- [40] Crooks G E and Chandler D 1997 Gaussian statistics of the hard sphere fluid *Phys. Rev. E* **56** 4217
- [41] Hummer G, Garde S, García A E, Pohorille A and Pratt L R 1996 An information theory model of hydrophobic interactions *Proc. Natl Acad. Sci.* **93** 8951
- [42] Debye P and Huckel E 1923 *Phys. Z.* **24** 305
- [43] Bjerrum N 1926 *Kgl. Dansk Videnskab, Selskab* **7** No 9

- [44] Mayer J 1950 Theory of ionic solutions *J. Chem. Phys.* **18** 1426
- [45] Rasaiah J C and Friedman H L 1968 Integral equation methods in computations of equilibrium properties of ionic solutions *J. Chem. Phys.* **48** 2742
Rasaiah J C and Friedman H L 1969 Integral equation computations for 1-1 electrolytes. Accuracy of the method *J. Chem. Phys.* **50** 3965
- [46] Waisman E and Lebowitz J K 1972 Mean spherical model integral equation for charged hard spheres I. Method of solution *J. Chem. Phys.* **56** 3086
Waisman E and Lebowitz J K 1972 Mean spherical model integral equation for charged hard spheres II. Results *J. Chem. Phys.* **56** 3093
- [47] Blum L 1980 Primitive electrolytes in the mean spherical model *Theoretical Chemistry: Advances and Perspectives* vol 5 (New York: Academic)
- [48] Rasaiah J C 1990 A model for weak electrolytes *Int. J. Thermophys.* **11** 1
- [49] Zhou Y and Stell G 1993 Analytic approach to molecular liquids V. Symmetric dissociative dipolar dumb-bells with the bonding length $\sigma/3 = L = \sigma/2$ and related systems *J. Chem. Phys.* **98** 5777
- [50] Ebeling W 1968 Zur Theorie der Bjerrumschen Ionenassoziation in Electrolyten *Z. Phys. Chem. (Leipzig)* **238** 400

-132-

- [51] Ebeling W and Grigoro M 1980 Analytical calculation of the equation of state and the critical point in a dense classical fluid of charged hard spheres *Phys. (Leipzig)* **37** 21
- [52] Pitzer K S 1995 Ionic fluids: near-critical and related properties *J. Phys. Chem.* **99** 13 070
- [53] Fisher M and Levin Y 1993 Criticality in ionic fluids: Debye Huckel Theory, Bjerrum and beyond *Phys. Rev. Lett.* **71** 3826
Fisher M 1996 The nature of criticality in ionic fluids *J. Phys.: Condens. Matter.* **8** 9103
- [54] Stell G 1995 Criticality and phase transitions in ionic fluids *J. Stat. Phys.* **78** 197
Stell G 1999 New results on some ionic fluid problems, new approaches to problems in liquid state theory *Proc. NATO Advanced Study Institute (Patte Marina, Messina, Italy 1998)* ed C Caccamo, J P Hansen and G Stell (Dordrecht: Kluwer)
- [55] Anisimov M A, Povodyrev A A, Sengers J V and Levelt-Sengers J M H 1997 Vapor-liquid equilibria, scaling and crossover in aqueous solutions of sodium chloride near the critical line *Physica A* **244** 298
- [56] Jacob J, Kumar A, Anisimov M A, Povodyrev A A . and Sengers J V 1998 Crossover from Ising to mean-field critical behavior in an aqueous electrolyte solution *Phys. Rev. E* **58** 2188
- [57] Orkoulas G and Panagiotopoulos A Z 1999 Phase behavior of the restricted primitive model and square-well fluids from Monte Carlo simulations in the grand canonical ensemble *J. Chem. Phys.* **110** 1581
- [58] Valleau J P and Torrie G M 1998 Heat capacity of the restricted primitive model *J. Chem. Phys.* **108** 5169
- [59] Camp P J and Patey G N 1999 Ion association in model ionic fluids *Phys. Rev. E* **60** 1063
Camp P J and Patey G N 1999 Ion association and condensation in primitive models of electrolytes *J. Chem. Phys.*
- [60] Koneshan S and Rasaiah J C 2000 Computer simulation studies of aqueous sodium chloride solutions at 298K and 683K *J. Chem. Phys.* **113** 8125

- [61] Stillinger F H and Lovett R 1968 General restriction on the distribution of ions in electrolytes *J. Chem. Phys.* **48** 1991
- [62] Friedman H L 1962 *Ionic Solution Theory* (New York: Interscience)
- [63] Card D N and Valleau J 1970 Monte Carlo study of the thermodynamics of electrolyte solutions *J. Chem. Phys.* **52** 6232
Rasaiah J C, Card D N and Valleau J 1972 Calculations on the 'restricted primitive model' for 1-1 electrolyte solutions *J. Chem. Phys.* **56** 248
- [64] Allnatt A 1964 Integral equations in ionic solution theory *Mol. Phys.* **8** 533
- [65] Rasaiah J C 1970 Equilibrium properties of ionic solutions; the primitive model and its modification for aqueous solutions of the alkali halides at 25°C *J. Chem. Phys.* **52** 704
- [66] Ramanathan P S and Friedman H L 1971 Study of a refined model for aqueous 1-1 electrolytes *J. Chem. Phys.* **54** 1086
- [67] Rasaiah J C 1972 Computations for higher valence electrolytes in the restricted primitive model *J. Chem. Phys.* **56** 3071

-133-

- [68] Valleau J P and Cohen L K 1980 Primitive model electrolytes. I. Grand canonical Monte Carlo computations *J. Chem. Phys.* **72** 5932
Valleau J P, Cohen L K and Card D N 1980 Primitive model electrolytes. II. The symmetrical electrolyte *J. Chem. Phys.* **72** 5942
- [69] Cummings P T and Stell G 1984 Statistical mechanical models of chemical reactions analytic solution of models of $A + B \rightleftharpoons AB$ in the Percus–Yevick approximation *Mol. Phys.* **51** 253
Cummings P T and Stell G 1984 Statistical mechanical models of chemical reactions II. Analytic solutions of the Percus–Yevick approximation for a model of homogeneous association *Mol. Phys.* **51** 253
- [70] Baxter R J 1968 Percus–Yevick equation for hard spheres with surface adhesion *J. Chem. Phys.* **49** 2770
- [71] Zhu J and Rasaiah J C 1989 Solvent effects in weak electrolytes II. Dipolar hard sphere solvent and the sticky electrolyte model with $L = \sigma$ *J. Chem. Phys.* **91** 505
- [72] Zwanzig R 1954 High temperature equation of state by a perturbation method I. Nonpolar Gases *J. Chem. Phys.* **22** 1420
- [73] Hemmer P C 1964 On van der Waals theory of vapor–liquid equilibrium IV. The pair correlation function and equation of state for long-range forces *J. Math. Phys.* **5** 75
- [74] Andersen H C and Chandler D 1970 Mode expansion in equilibrium statistical mechanics I. General theory and application to electron gas *J. Chem. Phys.* **53** 547
Chandler D and Andersen H C 1971 Mode expansion in equilibrium statistical mechanics II. A rapidly convergent theory of ionic solutions *J. Chem. Phys.* **54** 26
Andersen H C and Chandler D 1971 Mode expansion in equilibrium statistical mechanics III. Optimized convergence and application to ionic solution theory *J. Chem. Phys.* **55** 1497
- [75] Lebowitz J L and Percus J 1961 Long range correlations in a closed system with applications to nonuniform fluids *Phys. Rev.* **122** 1675
- [76] Hiroike K 1972 Long-range correlations of the distribution functions in the canonical ensemble *J. Phys. Soc. Japan* **32** 904

- [77] Barker J and Henderson D 1967 Perturbation theory and equation of state for a fluids II. A successful theory of liquids *J. Chem. Phys.* **47** 4714
- [78] Mansoori G A and Canfield F B 1969 Variational approach to the equilibrium properties of simple liquids I *J. Chem. Phys.* **51** 4958
Mansoori G A and Canfield F B 1970 *J. Chem. Phys.* **53** 1618
- [79] Rasaiah J C and Stell G 1970 Upper bounds on free energies in terms of hard sphere results *Mol. Phys.* **18** 249
- [80] Weeks J, Chandler D and Anderson H C 1971 Role of repulsive forces in determining the equilibrium structure of simple liquids *J. Chem. Phys.* **54** 5237
Chandler D, Weeks J D and Andersen H C 1983 The van der Waals picture of liquids, solids and phase transformations *Science* **220** 787
- [81] Rushbrooke G 1940 On the statistical mechanics of assemblies whose energy-levels depend on temperature *Trans. Faraday Soc.* **36** 1055
-

-134-

- [82] Cook and Rowlinson J S 1953 Deviations from the principles of corresponding states *Proc. R. Soc. A* **219** 405
- [83] Pople J 1954 Statistical mechanics of assemblies of axially symmetric molecules I. General theory *Proc. R. Soc. A* **221** 498
Pople J 1954 Statistical mechanics of assemblies of axially symmetric molecules II. Second virial coefficients *Proc. R. Soc. A* **221** 508
- [84] Zwanzig R 1955 High temperature equation of state by a perturbation method II. Polar gases *J. Chem. Phys.* **23** 1915
- [85] Larsen B, Rasaiah J C and Stell G 1977 Thermodynamic perturbation theory for multipolar and ionic fluids *Mol. Phys.* **33** 987
Stell G, Rasaiah J C and Narang H 1974 *Mol. Phys.* **27** 1393
- [86] Onsager L 1939 Electrostatic interaction of molecules *J. Phys. Chem.* **43** 189
- [87] Gillan M 1980 Upper bound on the free energy of the restricted primitive model for ionic liquids *Mol. Phys.* **41** 75
- [88] Stell G and Lebowitz J 1968 Equilibrium properties of a system of charged particles *J. Chem. Phys.* **49** 3706
- [89] Stell G, Wu K C and Larsen B 1976 Critical point in a fluid of charged hard spheres *Phys. Rev. Lett.* **211** 369
- [90] Haksjold B and Stell G 1982 The equilibrium studies of simple ionic liquids *The Liquid State of Matter: Fluids, Simple and Complex* ed E W Montroll and J L Lebowitz (Amsterdam: North-Holland)
- [91] Straatsma T P and McCammon J A 1992 Computational alchemy *Ann. Rev. Phys. Chem.* **43** 407
- [92] Jorgenson W L and Ravimohan C 1985 Monte Carlo simulation of the differences in free energy of hydration *J. Chem. Phys.* **83** 3050
- [93] Jorgenson W 1989 Free energy calculations: a breakthrough in modeling organic chemistry in solution *Accounts Chem. Res.* **22** 184
- [94] Weeks J D, Selinger R L B and Broughton J Q 1995 Self consistent treatment of attractive forces in

non uniform liquids *Phys. Rev. Lett.* **75** 2694

- [95] Weeks J D, Vollmayr K and Katsov K 1997 Intermolecular forces and the structure of uniform and non uniform fluids *Physica A* **244** 461
Weeks J D, Katsov K and Vollmayr K 1998 Roles of repulsive and attractive forces in determining the structure of non uniform liquids: generalized mean field theory *Phys. Rev. Lett.* **81** 4400
- [96] Ising E 1925 *Z. Phys.* 31 253
- [97] Lee T D and Yang C N 1952 Statistical theory of equations of state and phase transitions II. Lattice gas and Ising models *Phys. Rev.* **87** 410
- [98] Fowler R H and Guggenheim E A 1940 Statistical thermodynamics of super-lattices *Proc. R. Soc. A* **174** 189
- [99] Bethé H 1935 Statistical theory of superlattices *Proc. R. Soc. A* **150** 552
- [100] Landau L D 1935 quoted in Landau L D and Lifshitz E M 1958 *Statistical Physics* ch XIV, section 135 (Oxford: Pergamon)
-

-135-

- [101] Widom B 1965 Equation of state near the critical point *J. Chem. Phys.* **43** 3898
- [102] Neece G A and Widom B 1969 Theories of liquids *Ann. Rev. Phys. Chem.* **20** 167
- [103] Kramers H A and Wannier G H 1941 Statistics of the two-dimensional ferromagnet part I *Phys. Rev.* **60** 252
Kramers H A and Wannier G H 1941 Statistics of the two-dimensional ferromagnet part II *Phys. Rev.* **60** 263
- [104] Domb C and Sykes M F 1957 On the susceptibility of a ferromagnetic above the Curie point *Proc. R. Soc. A* **240** 214
Domb C and Sykes M F 1957 Specific heat of a ferromagnetic Substance above the Curie point *Phys. Rev.* **129** 567
- [105] Onsager L 1944 Crystal statistics I. A two-dimensional model with an order–disorder transition *Phys. Rev.* **65** 117
- [106] Kaufman B 1949 Crystal statistics II. Partition function evaluated by Spinor analysis *Phys. Rev.* **65** 1232
Onsager L and Kaufman B 1949 Crystal statistics III. Short range order in a binary Ising lattice *Phys. Rev.* **65** 1244
- [107] Yang C N 1952 The spontaneous magnetization of a two-dimensional Ising lattice *Phys. Rev.* **85** 809 (87 404)
-

FURTHER READING

Alavi A 1996 Path integrals and *ab initio* molecular dynamics *Monte Carlo and Molecular Dynamics of Condensed Matter Systems* ed K Binder and G Ciccotti (Bologna: SIF)

Anisimov M A and Sengers J V 1999 Crossover critical phenomena in aqueous solutions *Proc. 13th Int. Conf. on the Properties of Water and Steam (Toronto, September 12–16 1999)*

Barker J A and Henderson D 1976 What is a liquid? Understanding the states of matter *Rev. Mod. Phys.* **48** 587

Berne B J and Thirumalai D 1986 On the simulation of quantum systems: path integral methods *Ann. Rev. Phys. Chem.* **37** 401

Chandler D 1987 *Introduction to Modern Statistical Mechanics* (Oxford: Oxford University Press)

Chandler D and Wolynes P 1979 Exploiting the isomorphism between quantum theory and classical statistical mechanics of polyatomic fluids *J. Chem. Phys.* **70** 2914

Debenedetti P G 1996 *Metastable Liquids, Concepts and Principles* (Princeton, NJ: Princeton University Press)

Domb C 1996 *The Critical Point. A Historical Introduction to the Modern Theory of Critical Phenomena* (London: Taylor and Francis)

Eyring H, Henderson D, Stover B J and Eyring E 1982 *Statistical Mechanics and Dynamics* (New York: Wiley)

Fisher M 1983 Scaling, universality and renormalization group theory *Critical Phenomena (Lecture Notes in Physics vol 186)* (Berlin: Springer)

Friedman H 1985 *A Course in Statistical Mechanics* (Englewood Cliffs, NJ: Prentice-Hall)

-136-

Friedman H L and Dale W T 1977 Electrolyte solutions at equilibrium *Statistical Mechanics part A, Equilibrium Techniques* ed B J Berne (New York: Plenum)

Goldenfeld L 1992 *Lectures in Phase Transitions and Renormalization Group* (New York: Addison-Wesley)

Goodstein D L 1974 *States of Matter* (Englewood Cliffs, NJ: Prentice-Hall and Dover)

Guggenheim E A 1967 *Thermodynamics and Advanced Treatment* 5th edn (Amsterdam: North-Holland)

McQuarrie D 1976 *Statistical Mechanics* (New York: Harper and Row)

Rice S A and Gray P 1965 *The Statistical Mechanics of Simple Liquids* (New York: Interscience)

Wilde R E and Singh S 1998 *Statistical Mechanics* (New York: Wiley)

Hirschfelder J O, Curtiss C F and Bird R B 1954 *Molecular Theory of Gases and Liquids* (New York: Wiley)

Rowlinson J and Swinton J 1983 *Liquids and Liquid Mixtures* 3rd edn (London: Butterworth)

Voth G 1996 Path integral centroid methods *Advances in Chemical Physics, New methods in Computational Quantum Mechanics* vol XCIII, ed I Prigogine and S A Rice

Makri N 1999 Time dependent quantum methods for large systems *Ann. Rev. Phys. Chem.* **50** 167

Reiss H and Hammerich A D S 1986 Hard spheres: scaled particle theory and exact relations on the existence and structure of the fluid/solid phase transition *J. Phys. Chem.* **90** 6252

Stillinger F 1973 Structure in aqueous solutions from the standpoint of scaled particle theory *J. Solution Chem.* **2** 141

Widom B 1967 Intermolecular forces and the nature of the liquid state *Science* **375** 157

Longuet-Higgins H C and Widom B 1964 A rigid sphere model for the melting of argon *Mol. Phys.* **8** 549

Smith W R 1972 Perturbation theory in the classical statistical mechanics of fluids *Specialist Periodical Report* vol 1 (London: Chemical Society)

Watts R O 1972 Integral equation approximations in the theory of fluids *Specialist Periodical Report* vol 1 (London: Chemical

Society)

Mitchell D J, McQuarrie D A, Szabo A and Groeneveld J 1977 On the second-moment condition of Stillinger and Lovett *J. Stat. Phys.* **17** 1977

Vlachy V 1999 Ionic effects beyond Poisson–Boltzmann theory *Ann. Rev. Phys. Chem.* **50** 145

Outhwaite C W 1974 Equilibrium theories of electrolyte solutions *Specialist Periodical Report* (London: Chemical Society)

Rasaiah J C 1987 Theories of electrolyte solutions *The Liquid State and its Electrical Properties (NATO Advanced Science Institute Series Vol 193)* ed E E Kunhardt, L G Christophous and L H Luessen (New York: Plenum)

Rasaiah J C 1973 A view of electrolyte solutions *J. Solution Chem.* **2** 301

-137-

Stell G, Patey G N and Høye J S 1981 Dielectric constant of fluid models: statistical mechanical theory and its quantitative implementation *Adv. Chem. Phys.* **48** 183

Wertheim M 1979 Equilibrium statistical mechanics of polar fluids *Ann. Rev. Phys. Chem.* **30** 471

Reynolds C, King P M and Richards W G 1992 Free energy calculations in molecular biophysics *Mol. Phys.* **76** 251

Pratt L 1997 Molecular theory of hydrophobic effects *Encyclopedia of Computational Chemistry*

Lynden-Bell R M and Rasaiah J C 1997 From hydrophobic to hydrophilic behavior: a simulation study of solvation entropy and free energy of simple solutes *J. Chem. Phys.* **107** 1981

Hummer G, Garde S, Garcia A E, Paulitis M E and Pratt L R 1998 Hydrophobic effects on a molecular scale *J. Phys. Chem.* **102** 10 469

Lum K, Chandler D and Weeks J D 1999 Hydrophobicity at small and large length scales *J. Phys. Chem. B* **103** 4570

Pratt L R and Hummer G (eds) 1999 Simulation and theory of electrostatic interactions in solution; computational chemistry, biophysics and aqueous solutions *AIP Conf. Proc. (Sante Fe, NM, 1999)* vol 492 (New York: American Institute of Physics)

Stanley H E 1971 *Introduction to Phase Transitions and Critical Phenomena* (Oxford: Oxford University Press)

Ziman J M 1979 *Models of Disorder* (Cambridge: Cambridge University Press)

Yeomans Y M 1992 *Statistical Mechanics of Phase Transitions* (Oxford: Oxford University Press)

Stanley H E 1999 Scaling, universality and renormalization: three pillars of modern critical phenomena *Rev. Mod. Phys.* **71** S358

Kadanoff L P 1999 *Statistical Physics: Statics, Dynamics and Renormalization* (Singapore: World Scientific)

-138-

Stell G, Patey G N and Høye J S 1981 Dielectric constant of fluid models: statistical mechanical theory and its quantitative implementation *Adv. Chem. Phys.* **48** 183

Wertheim M 1979 Equilibrium statistical mechanics of polar fluids *Ann. Rev. Phys. Chem.* **30** 471

Reynolds C, King P M and Richards W G 1992 Free energy calculations in molecular biophysics *Mol. Phys.* **76** 251

Pratt L 1997 Molecular theory of hydrophobic effects *Encyclopedia of Computational Chemistry*

Lynden-Bell R M and Rasaiah J C 1997 From hydrophobic to hydrophilic behavior: a simulation study of solvation entropy and free energy of simple solutes *J. Chem. Phys.* **107** 1981

Hummer G, Garde S, Garcia A E, Paulitis M E and Pratt L R 1998 Hydrophobic effects on a molecular scale *J. Phys. Chem.* **102** 10 469

Lum K, Chandler D and Weeks J D 1999 Hydrophobicity at small and large length scales *J. Phys. Chem. B* **103** 4570

Pratt L R and Hummer G (eds) 1999 Simulation and theory of electrostatic interactions in solution; computational chemistry, biophysics and aqueous solutions *AIP Conf. Proc. (Sante Fe, NM, 1999)* vol 492 (New York: American Institute of Physics)

Stanley H E 1971 *Introduction to Phase Transitions and Critical Phenomena* (Oxford: Oxford University Press)

Ziman J M 1979 *Models of Disorder* (Cambridge: Cambridge University Press)

Yeomans Y M 1992 *Statistical Mechanics of Phase Transitions* (Oxford: Oxford University Press)

Stanley H E 1999 Scaling, universality and renormalization: three pillars of modern critical phenomena *Rev. Mod. Phys.* **71** S358

Kadanoff L P 1999 *Statistical Physics: Statics, Dynamics and Renormalization* (Singapore: World Scientific)

-1-

A 2.4 Fundamentals of electrochemistry

Andrew Hamnett

Electrochemistry is concerned with the study of the interface between an electronic and an ionic conductor and, traditionally, has concentrated on: (i) the nature of the ionic conductor, which is usually an aqueous or (more rarely) a non-aqueous solution, polymer or superionic solid containing mobile ions; (ii) the structure of the electrified interface that forms on immersion of an electronic conductor into an ionic conductor; and (iii) the electron-transfer processes that can take place at this interface and the limitations on the rates of such processes.

Ionic conductors arise whenever there are mobile ions present. In electrolyte solutions, such ions are normally formed by the dissolution of an ionic solid. Provided the dissolution leads to the complete separation of the ionic components to form essentially independent anions and cations, the electrolyte is termed *strong*. By contrast, *weak* electrolytes, such as organic carboxylic acids, are present mainly in the undissociated form in solution, with the total *ionic* concentration orders of magnitude lower than the formal concentration of the solute. Ionic conductivity will be treated in some detail below, but we initially concentrate on the equilibrium structure of liquids and ionic solutions.

A 2.4.1 THE ELEMENTARY THEORY OF LIQUIDS

Modern-day approaches to ionic solutions need to be able to contend with the following problems:

- (1) the nature of the solvent itself, and the interactions taking place in that solvent;
- (2) the changes taking place on the dissolution of an ionic electrolyte in the solvent;

(3) macroscopic and microscopic studies of the properties of electrolyte solutions.

Even the description of the solvent itself presents major theoretical problems: the partition function for a liquid can be written in the classical limit [1, 2] as

$$Q(T, V, N) = \frac{q^N}{(8\pi^2)^N N! \Lambda^{3N}} \int \cdots \int d\mathbf{X}^N \exp[-\beta U_N(\mathbf{X}^N)] \quad (\text{A2.4.1})$$

where $\beta = 1/kT$ and the integral in (1) is over both spatial and orientation coordinates (i.e. both Cartesian and Eulerian coordinates) of each of the N molecules, and is termed the *configurational integral*, Z_N . In this equation, q is the partition coefficient for the internal degrees of freedom in each molecule (rotational, vibrational and electronic), Λ is the translational partition function $h/(2\pi mkT)^{1/2}$ and $U_N(\mathbf{X}^N)$ is the energy associated with the instantaneous configuration of the N molecules defined by \mathbf{X}^N . Clearly, the direct evaluation of (1) for all but the simplest cases is quite impossible, and modern theories have made an indirect attack on (1) by defining *distribution functions*. If we consider, as an example, two particles, which we fix with total coordinates X_1 and X_2 , then the joint probability of finding particle 1 in volume dX_1 and particle 2 in volume dX_2 is

-2-

$$P^{(2)}(\mathbf{X}_1, \mathbf{X}_2) d\mathbf{X}_1 d\mathbf{X}_2 = d\mathbf{X}_1 d\mathbf{X}_2 \int \cdots \int d\mathbf{X}_3, \dots, d\mathbf{X}_N P(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \dots, \mathbf{X}_N) \quad (\text{A2.4.2})$$

where $P(\mathbf{X}_1, \dots, \mathbf{X}_N)$ is the Boltzmann factor:

$$P(\mathbf{X}^N) = \frac{\exp[-\beta U_N(\mathbf{X}^N)]}{\int \cdots \int d\mathbf{X}^N \exp[-\beta U_N(\mathbf{X}^N)]}$$

In fact, given that there are $N(N-1)$ ways of choosing a pair of particles, the pair distribution function, $\rho^{(2)}(X_1, X_2) dX_1 dX_2 = N(N-1) P^{(2)}(X_1, X_2) dX_1 dX_2$ is the probability of finding any particle at X_1 in volume dX_1 and a different particle at X_2 in volume dX_2 . A little reflection will show that for an isotropic liquid, the value of $\rho^{(1)}(X_1)$ is just the number density of the liquid, $\rho = N/V$, since we can integrate over all orientations and, if the liquid is isotropic, its density is everywhere constant.

A2.4.1.1 CORRELATION FUNCTIONS

The pair distribution function clearly has dimensions (density)², and it is normal to introduce the pair correlation function $g(X_1, X_2)$ defined by

$$g(\mathbf{X}_1, \mathbf{X}_2) = \frac{\rho^{(2)}(\mathbf{X}_1, \mathbf{X}_2)}{\rho^{(1)}(\mathbf{X}_1)\rho^{(1)}(\mathbf{X}_2)} \quad (\text{A2.4.3})$$

and we can average over the orientational parts of both molecules, to give

$$g(\mathbf{R}_1, \mathbf{R}_2) = \frac{1}{(8\pi^2)^2} \iint d\Omega_1 d\Omega_2 g(\mathbf{X}_1, \mathbf{X}_2). \quad (\text{A2.4.4})$$

Given that \mathbf{R}_1 can be arbitrarily chosen as anywhere in the sample volume of an isotropic liquid in the absence

of any external field, we can transform the variables $\mathbf{R}_1, \mathbf{R}_2$ into the variables $\mathbf{R}_1, \mathbf{R}_{12}$, where \mathbf{R}_{12} is the vector separation of molecules 1 and 2. This allows us to perform one of the two integrations in (equation A2.4.2) above, allowing us to write the probability of a second molecule being found at a distance r ($= |\mathbf{r}_{12}|$) from a central molecule as $\rho g(r) r dr \sin \theta d\theta d\phi$. Integration over the angular variables gives the number of molecules found in a spherical shell at a distance r from a central molecule as

$$N(r) dr = 4\pi r^2 \rho g(r) dr. \quad (\text{A2.4.5})$$

The function $g(r)$ is central to the modern theory of liquids, since it can be measured experimentally using neutron or x-ray diffraction and can be related to the interparticle potential energy. Experimental data [1] for two liquids, water and argon (iso-electronic with water) are shown in figure A2.4.1 plotted as a function of $R^* = R/\sigma$, where σ is the effective diameter of the species, and is roughly the position of the first maximum in $g(R)$. For water, $\sigma = 2.82 \text{ \AA}$,

-3-

very close to the intermolecular distance in the normal tetrahedrally bonded form of ice, and for argon, $\sigma = 3.4 \text{ \AA}$. The second peak for argon is at $R^* = 2$, as expected for a spherical molecular system consisting roughly of concentric spheres. However, for water, the second peak in $g(r)$ is found at $R^* = 1.6$, which corresponds closely to the second-nearest-neighbour distance in ice, strongly supporting the model for the structure of water that is ice-like over short distances. This strongly structured model for water in fact dictates many of its anomalous properties.

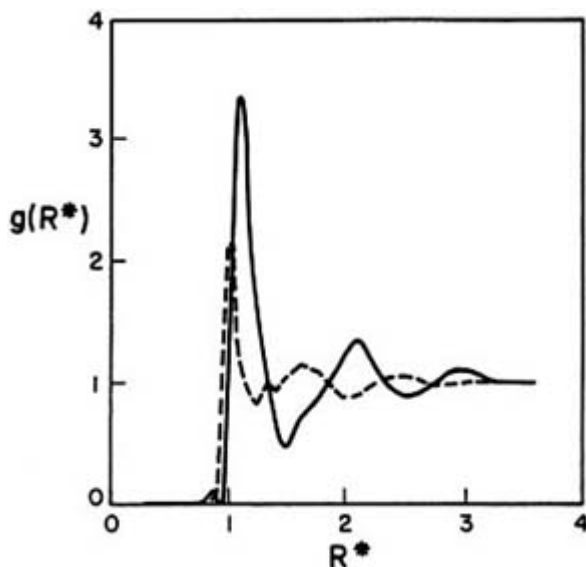


Figure A2.4.1. Radial distribution function $g(R^*)$ for water (dashed curve) at $4 \text{ }^\circ\text{C}$ and 1 atm and for liquid argon (full curve) at 84.25 K and 0.71 atm as functions of the reduced distance $R^* = R/\sigma$, where σ is the molecular diameter; from [1].

The relationship between $g(r)$ and the interparticle potential energy is most easily seen if we assume that the interparticle energy can be factorized into pairwise additive potentials as

$$U_N(\mathbf{X}^N) \equiv U_N(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N) = \sum_{i < j}^N u(r_{ij}) \quad (\text{A2.4.6})$$

where the summation is over all pairs i, j . From equation A2.4.1 we can calculate the total internal energy U

as

$$U = -\left(\frac{\partial \ln Q}{\partial \beta}\right)_{v,N} = -\frac{1}{Q} \left\{ \frac{1}{N! \Lambda^{3N}} \int d\mathbf{R}^N \left[-\sum_{i<j}^N u(r_{ij}) \right] \exp(-U_N) - \frac{3NQ}{\Lambda} \frac{\partial \Lambda}{\partial \beta} \right\} \quad (\text{A2.4.7})$$

and where, for simplicity, we have ignored internal rotational and orientational effects. For an isotropic liquid, the summation in A2.4.6 over pairs of molecules yields $N(N-1)/2$ equal terms, which can be written as the product of a two-particle integral over the $u(r_{12})$ and integrals of the type shown in [A2.4.2](#) above. After some algebra, we find

-4-

$$U = \frac{3}{2} NkT + \frac{1}{2} \int_V d\mathbf{r}_1 \int_V d\mathbf{r}_2 u(r_{12}) \rho^{(2)}(\mathbf{r}_1, \mathbf{r}_2) = \frac{3}{2} NkT + \frac{\rho N}{2} \int_0^\infty dr 4\pi r^2 g(r) u(r). \quad (\text{A2.4.8})$$

To the same order of approximation, the pressure P can be written as

$$\frac{\beta P}{\rho} = 1 - \frac{\beta \rho}{6} \int_0^\infty dr 4\pi r^3 \frac{du(r)}{dr} g(r) \quad (\text{A2.4.9})$$

where in both A2.4.8 and A2.4.9 the potential $u(r)$ is assumed to be sufficiently short range for the integrals to converge. Other thermodynamic functions can be calculated once $g(r)$ is known, and it is of considerable importance that $g(r)$ can be obtained from $u(r)$ and, ideally, that the inverse process can also be carried out. Unfortunately, this latter process is much more difficult to do in such a way as to distinguish different possible $u(r)$ s with any precision.

Clearly, the assumption of pairwise additivity is unlikely to be a good one for water; indeed, it will break down for any fluid at high density. Nonetheless, $g(r)$ remains a good starting point for any liquid, and we need to explore ways in which it can be calculated. There are two distinct methods: (a) solving equations relating $g(r)$ to $u(r)$ by choosing a specific $u(r)$; (b) by simulation methods using molecular dynamic or Monte Carlo methods.

There are two approaches commonly used to derive an analytical connection between $g(r)$ and $u(r)$: the Percus-Yevick (PY) equation and the hypernetted chain (HNC) equation. Both are derived from attempts to form functional Taylor expansions of different correlation functions. These auxiliary correlation functions include:

(i) the *total correlation function*,

$$h(\mathbf{r}_1, \mathbf{r}_2) = g(\mathbf{r}_1, \mathbf{r}_2) - 1; \quad (\text{A2.4.10})$$

(ii) the *background correlation function*,

$$y(\mathbf{r}_1, \mathbf{r}_2) = g(\mathbf{r}_1, \mathbf{r}_2) \exp[\beta u(\mathbf{r}_1, \mathbf{r}_2)]; \quad (\text{A2.4.11})$$

(iii) the *direct correlation function*, $C(\mathbf{r}_1, \mathbf{r}_2)$, defined through the Ornstein-Zernike relation:

$$h(\mathbf{r}_1, \mathbf{r}_2) - C(\mathbf{r}_1, \mathbf{r}_2) \equiv \rho \int d\mathbf{r}_3 h(\mathbf{r}_1, \mathbf{r}_3) C(\mathbf{r}_3, \mathbf{r}_2). \quad (\text{A2.4.12})$$

The singlet *direct correlation function* $C^{(1)}(\mathbf{r})$ is defined through the relationship

$$C^{(1)}(\mathbf{r}) \equiv \ln[\rho^{(1)}(\mathbf{r})\Lambda^3] + \beta[w(\mathbf{r}) - \mu] \quad (\text{A2.4.13})$$

where $\rho^{(1)}$ is as defined above, μ is the chemical potential and $w(\mathbf{r})$ the local one-body potential in an inhomogeneous system.

-5-

The PY equation is derived from a Taylor expansion of the direct correlation function, and has the form

$$y(\mathbf{r}_1, \mathbf{r}_2) \approx 1 + \rho \int d\mathbf{r}_3 C(\mathbf{r}_2, \mathbf{r}_3)h(\mathbf{r}_3, \mathbf{r}_1) \quad (\text{A2.4.14})$$

and comparison with the Ornstein-Zernike equation shows that $C(\mathbf{r}_1, \mathbf{r}_2) \approx g(\mathbf{r}_1, \mathbf{r}_2) - y(\mathbf{r}_1, \mathbf{r}_2) \approx y(\mathbf{r}_1, \mathbf{r}_2)f(\mathbf{r}_1, \mathbf{r}_2)$, where $f(\mathbf{r}_1, \mathbf{r}_2) \equiv \exp[-\beta u(\mathbf{r}_1, \mathbf{r}_2)] - 1$. Substitution of this expression into A2.4.14 finally gives us, in terms of the pair correlation coefficient alone

$$g(\mathbf{r}_1, \mathbf{r}_2) \exp[\beta u(\mathbf{r}_1, \mathbf{r}_2)] = 1 + \rho \int d\mathbf{r}_3 g(\mathbf{r}_2, \mathbf{r}_3) e^{\beta u(\mathbf{r}_2, \mathbf{r}_3)} [e^{-\beta u(\mathbf{r}_2, \mathbf{r}_3)} - 1][g(\mathbf{r}_3, \mathbf{r}_1) - 1]. \quad (\text{A2.4.15})$$

This integral equation can be solved by expansion of the integrand in bipolar coordinates [2, 3]. Further improvement to the PY equation can be obtained by analytical fit to simulation studies as described below.

The HNC equation uses, instead of the expression for $C(\mathbf{r}_1, \mathbf{r}_2)$ from A2.4.14 above, an expression $C(\mathbf{r}_1, \mathbf{r}_2) \approx h(\mathbf{r}_1, \mathbf{r}_2) - \ln(y(\mathbf{r}_1, \mathbf{r}_2))$, which leads to the first-order HNC equation:

$$\ln(y(\mathbf{r}_1, \mathbf{r}_2)) \approx \rho \int d\mathbf{r}_3 C(\mathbf{r}_1, \mathbf{r}_3)h(\mathbf{r}_3, \mathbf{r}_2). \quad (\text{A2.4.16})$$

Comparison with the PY equation shows that the HNC equation is nonlinear, and this does present problems in numerical work, as well as preventing any analytical solutions being developed even in the simplest of cases.

In the limit of low densities, A2.4.15 shows that the zeroth-order approximation for $g(r)$ has the form

$$g(\mathbf{r}_1, \mathbf{r}_2) \approx e^{-\beta u(\mathbf{r}_1, \mathbf{r}_2)} \quad (\text{A2.4.17})$$

a form that will be useful in our consideration of the electrolyte solutions described below.

A2.4.1.2 SIMULATION METHODS

Simulation methods for calculating $g(r)$ have come into their own in the past 20 years as the cost of computing has fallen. The Monte Carlo method is the simplest in concept: this depends essentially on identifying a statistical or Monte Carlo approach to the solution of (equation A2.4.2). As with all Monte Carlo integrations, a series of random values of the coordinates X_1, \dots, X_N is generated, and the integrand evaluated. The essential art in the technique is to pick predominantly configurations of high probability, or at least to

eliminate the wasteful evaluation of the integrand for configurations of high energy. This is achieved by moving one particle randomly from the previous configuration, i , and checking the energy difference $\Delta U = U_{i+1} - U_i$. If $\Delta U < 0$ the configuration is accepted, and if $\Delta U > 0$, the

-6-

value of $\exp(-\beta\Delta U)$ is compared to a second random number ξ , where $0 < \xi < 1$. If $\exp(-\beta\Delta U) > \xi$ the configuration is accepted, otherwise it is rejected and a new single-particle movement generated. A second difficulty is that the total number of particles that can be treated by Monte Carlo techniques is relatively small unless a huge computing resource is available: given this, boundary effects would be expected to be dominant, and so periodic boundary conditions are imposed, in which any particle leaving through one surface re-enters the system through the opposite surface. Detailed treatments of the Monte Carlo technique were first described by Metropolis *et al* [4]; the method has proved valuable not only in the simulation of realistic interparticle potentials, but also in the simulation of model potentials for comparison with the integral equation approaches above.

The alternative simulation approaches are based on molecular dynamics calculations. This is conceptually simpler than the Monte Carlo method: the equations of motion are solved for a system of N molecules, and periodic boundary conditions are again imposed. This method permits both the equilibrium and transport properties of the system to be evaluated, essentially by numerically solving the equations of motion

$$m \frac{d^2 \mathbf{R}_k}{dt^2} = \sum_{j=1, j \neq k}^N \mathbf{F}(\mathbf{R}_{kj}) = - \sum_{j=1, j \neq k}^N \nabla_k U(\mathbf{R}_{kj}) \quad (\text{A2.4.18})$$

by integrating over discrete time intervals δt . Details are given elsewhere [2].

A 2.4.2 IONIC SOLUTIONS

There is, in essence, no limitation, other than the computing time, to the accuracy and predictive capacity of molecular dynamic and Monte Carlo methods, and, although the derivation of realistic potentials for water is a formidable task in its own right, we can anticipate that accurate simulations of water will have been made relatively soon. However, there remain major theoretical problems in deriving any analytical theory for water, and indeed any other highly-polar solvent of the sort encountered in normal electrochemistry. It might be felt, therefore, that the extension of the theory to analytical descriptions of ionic solutions was a well-nigh hopeless task. However, a major simplification of our problem is allowed by the possibility, at least in more dilute solutions, of smoothing out the influence of the solvent molecules and reducing their influence to such average quantities as the dielectric permittivity, ϵ_m , of the medium. Such a viewpoint is developed within the McMillan-Mayer theory of solutions [1, 2], which essentially seeks to partition the interaction potential into three parts: that due to the interaction between the solvent molecules themselves, that due to the interaction between the solvent and the solute and that due to the interaction between the solute molecules dispersed within the solvent. The main difference from the dilute fluid results presented above is that the potential energy $u(r_{ij})$ is replaced by the potential of mean force $W(r_{ij})$ for two particles and, for N_a particles of solute in the solvent, by the expression

$$W(\mathbf{X}^{N_a}; z_a \rightarrow 0) = -kT \ln g^{(N_a)}(\mathbf{X}^{N_a}; z_a \rightarrow 0)$$

where Z_a is the so-called *activity* defined as $z_a = q_a e^{-\beta\mu_a} / \Lambda_a^3$ (cf equation A2.4.1); it has units of number density.

The McMillan-Mayer theory allows us to develop a formalism similar to that of a dilute interacting fluid for solute dispersed in the solvent provided that a sensible description of W can be given. At the limit of dilution, when intersolute interactions can be neglected, we know that the chemical potential of a can be written as $\mu_a = W(a|s) + kT \ln(\rho_a \Lambda_a^3 q_a^{-1})$, where $W(a|s)$ is the potential of mean force for the interaction of a solute molecule with the solvent. If we define $\gamma_a^0 = \lim_{\rho_a \rightarrow 0} (z_a 8\pi^2 / \rho_a) = e^{\beta W(a|s)}$ then the grand canonical partition function can be written in the form:

$$\Xi(T, V, \lambda_a, \lambda_s) = \sum_{N_a \geq 0} \frac{(z_a / \gamma_a^0)^{N_a}}{N_a!} \Xi(T, V, \lambda_s) \int e^{-\beta W(\mathbf{X}^{N_a}; z_a \rightarrow 0)} d\mathbf{X}^{N_a} \quad (\text{A2.4.19})$$

where we have successfully partitioned the solute-solute interactions into a modified configuration integral, the solute-solvent interactions into γ_a^0 and the solvent-solvent interactions into the partition coefficient $\Xi(T, V, \lambda_s)$.

A2.4.2.1 THE STRUCTURE OF WATER AND OTHER POLAR SOLVENTS

In terms of these three types of interactions, we should first consider the problems of water and other polar solvents in more detail. Of the various components of the interaction between water molecules, we may consider the following.

- (1) At very short distances, less than about 2.5 Å, a reasonable description of the interaction will be strongly repulsive, to prevent excessive interpenetration; a Lennard-Jones function will be adequate:

$$U_{\text{LJ}}(R) = 4\varepsilon \left[\left(\frac{\sigma}{R} \right)^{12} - \left(\frac{\sigma}{R} \right)^6 \right]. \quad (\text{A2.4.20})$$

- (2) At distances of a few molecular diameters, the interaction will be dominated by electric multipole interactions; for dipolar molecules, such as water, the dominant term will be the dipole-dipole interaction:

$$U_{\text{DD}}(\mathbf{X}_1, \mathbf{X}_2) = R_{12}^{-3} [\vec{\mu}_1 \cdot \vec{\mu}_2 - 3(\vec{\mu}_1 \cdot \mathbf{u}_{12})(\vec{\mu}_2 \cdot \mathbf{u}_{12})] \quad (\text{A2.4.21})$$

where \mathbf{u}_{12} is a unit vector in the direction of the vector $\mathbf{R}_2 - \mathbf{R}_1$.

- (3) At intermediate distances, $2.4 \text{ \AA} \leq R \leq 4 \text{ \AA}$, there is a severe analytical difficulty for water and other hydrogen-bonded solvents; in that the hydrogen-bond energy is quite large, but is extremely orientation dependent. If the water molecule is treated as tetrahedral, with two O–H vectors $\mathbf{h}_{i1}, \mathbf{h}_{i2}$ and two lone-pair vectors $\mathbf{l}_{i1}, \mathbf{l}_{i2}$ for the i th molecule, then the hydrogen-bond energy has the form

$$\begin{aligned} U_{\text{HB}}(\mathbf{X}_1, \mathbf{X}_2) &= \varepsilon_{\text{HB}} G(\mathbf{X}_1, \mathbf{X}_2) \\ &= \varepsilon_{\text{HB}} G_\sigma(R_{ij} - R_{\text{H}}) \left\{ \sum_{\alpha, \beta=1}^2 G_\sigma[(\mathbf{h}_{i\alpha} \cdot \mathbf{u}_{ij}) - 1] G_\sigma[(\mathbf{l}_{j\beta} \cdot \mathbf{u}_{ij}) + 1] \right. \\ &\quad \left. + G_\sigma[(\mathbf{l}_{i\alpha} \cdot \mathbf{u}_{ij}) - 1] G_\sigma[(\mathbf{h}_{j\beta} \cdot \mathbf{u}_{ij}) + 1] \right\} \end{aligned} \quad (\text{A2.4.22})$$

an expression that looks unwieldy but is quite straightforward to apply numerically. The function $G_\sigma(x)$ is defined either as unity for $|x| < \sigma$ and zero for $|x| \geq \sigma$ or in a Gaussian form: $G_\sigma = \exp(-x^2/2\sigma^2)$.

The form of the hydrogen-bonded potential leads to a strongly structured model for water, as discussed above. In principle, this structure can be defined in terms of the average number of hydrogen bonds formed by a single water molecule with its neighbours. In normal ice this is four, and we expect a value close to this in water close to the freezing point. We also intuitively expect that this number will decrease with increasing temperature, an expectation confirmed by the temperature dependence of $g(R)$ for water in figure A2.4.2 [1]. The picture should be seen as highly dynamic, with these hydrogen bonds forming and breaking continuously, with the result that the clusters of water molecules characterizing this picture are themselves in a continuous state of flux.

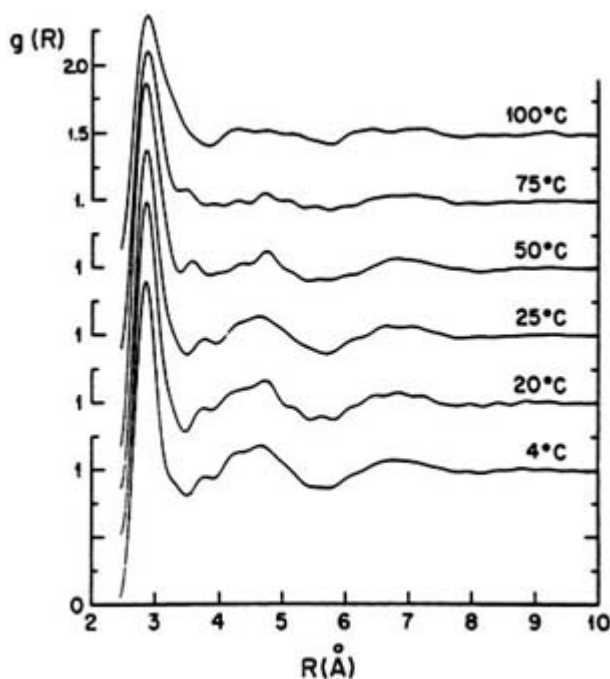


Figure A2.4.2. The temperature dependence of $g(R)$ of water. From [1]

A2.4.2.2 HYDRATION AND SOLVATION OF IONS

The solute-solvent interaction in [equation A2.4.19](#) is a measure of the solvation energy of the solute species at infinite dilution. The basic model for ionic hydration is shown in [figure A2.4.3](#) [5]: there is an inner hydration sheath of water molecules whose orientation is essentially determined entirely by the field due to the central ion. The number of water molecules in this inner sheath depends on the size and chemistry of the central ion; being, for example, four for Be^{2+} , but six for Mg^{2+} , Al^{3+} and most of the first-row transition ions. Outside this primary shell, there is a secondary sheath of more loosely bound water molecules oriented essentially by hydrogen bonding, the evidence for which was initially indirect and derived from ion mobility measurements. More recent evidence for this secondary shell has now come from x-ray diffraction and scattering studies and infrared (IR) measurements. A further highly diffuse region, the tertiary region, is probably present, marking a transition to the hydrogen-bonded structure of water described above. The ion, as it moves, will drag at least part of this solvation sheath with it, but the picture should be seen as essentially dynamic, with the well defined inner sheath structure of [figure A2.4.3](#) being mainly found in highly-charged ions of

high electronic stability, such as Cr^{3+} . The enthalpy of solvation of *cations* primarily depends on the charge on the central ion and the effective ionic radius, the latter being the sum of the normal Pauling ionic radius

and the radius of the oxygen atom in water (0.85 Å). A reasonable approximate formula has

$$\Delta H_{\text{hyd}}^0 = -695Z^2/(r_+ + 0.85) [\text{kJ mol}^{-1}]. \quad (\text{A2.4.23})$$

In general, anions are less strongly hydrated than cations, but recent neutron diffraction data have indicated that even around the halide ions there is a well defined primary hydration shell of water molecules, which, in the case of Cl^- varies from four to six in constitution; the exact number being a sensitive function of concentration and the nature of the accompanying cation.

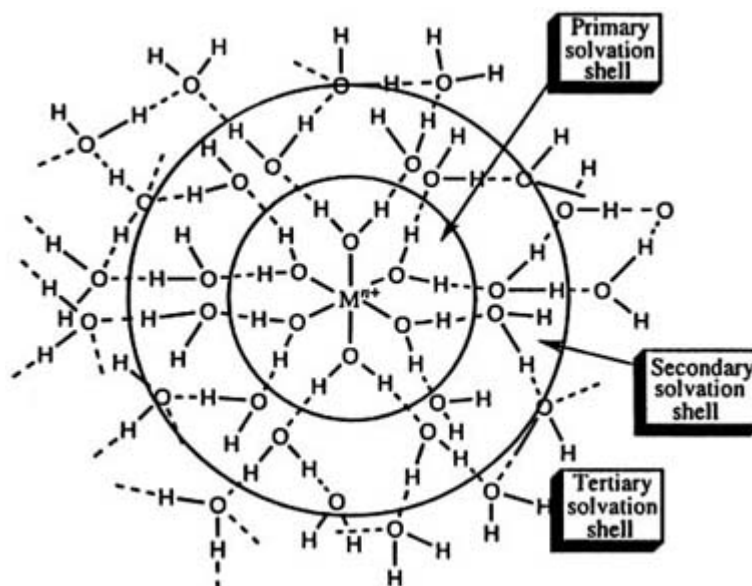


Figure A2.4.3. The localized structure of a hydrated metal cation in aqueous solution (the metal ion being assumed to have a primary hydration number of six). From [5].

(A) METHODS FOR DETERMINING THE STRUCTURE OF THE SOLVATION SHEATH

Structural investigations of metal-ion hydration have been carried out by spectroscopic, scattering and diffraction techniques, but these techniques do not always give identical results since they measure in different timescales. There are three distinct types of measurement:

- (1) those giving an average structure, such as neutron and x-ray scattering and diffraction studies;
- (2) those revealing dynamic properties of coordinated water molecules, such as nuclear magnetic resonance (NMR) and quasi-elastic scattering methods;
- (3) those based on energetic discrimination between water in the different hydration sheaths and bulk water, such as IR, Raman and thermodynamic studies.

First-order difference neutron scattering methods for the analysis of concentrated solutions of anions and cations were pioneered by Enderby [6] and co-workers and some results for Ni^{2+} plotted as $\Delta g(r)$ are shown in figure A2.4.4 [5]. The sharp M–O and M–D pair correlations are typical of long-lived inner hydration sheaths, with the broader structure showing the second hydration sheath being clearly present, but more diffuse. Note that the water molecule is tilted so that the D–O–D . . . M atoms are not coplanar. This tilt appears to be concentration dependent, and decreases to zero below 0.1 M NiCl_2 . It is almost certainly caused by interaction between the hydrogen-bonded secondary sheaths around the cations, a fact that will complicate the nature of the potential of the mean force, discussed in more detail below. The secondary hydration sheaths

have been studied by large-angle x-ray scattering (LAXS). For Cr^{3+} , a well defined secondary sheath containing 13 ± 1 molecules of water could be identified some $4.02 \pm 0.2 \text{ \AA}$ distant from the central ion. The extended x-ray absorption-edge fine structure (EXAFS) technique has also been used to study the local environment around anions and cations: in principle the technique is ideally suited for this, since it has high selectivity for the central ion and can be used in solutions more dilute than those accessible to neutron or x-ray scattering. However, the technique also depends on the capacity of the data to resolve different structural models that may actually give rise to rather similar EXAFS spectra. The sensitivity of the technique also falls away for internuclear distances in excess of 2.5 \AA .

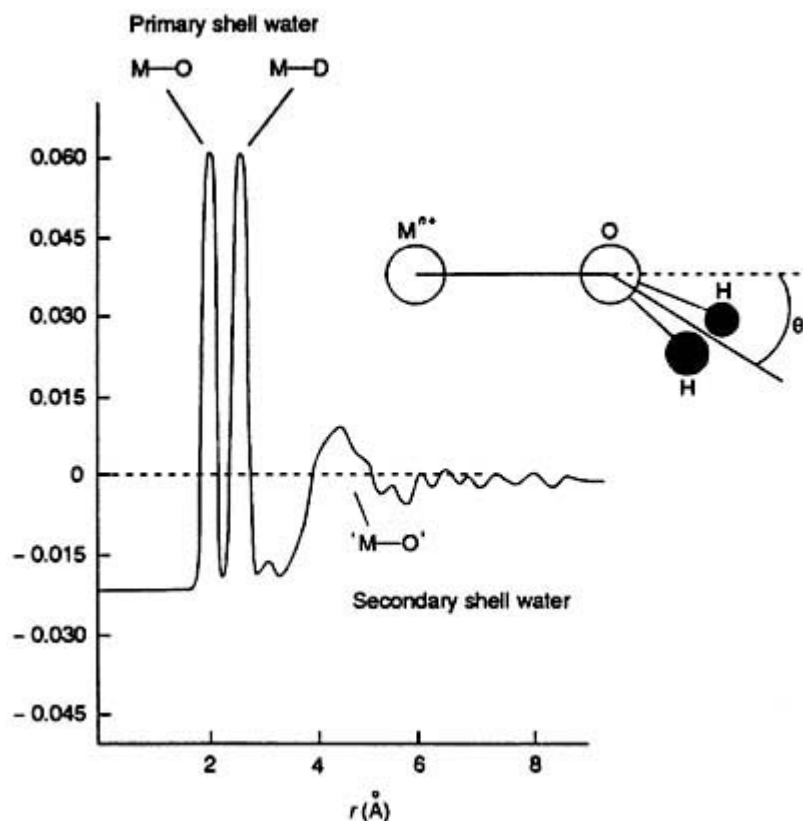


Figure A2.4. Plot of the radial distribution difference function $\Delta g(r)$ against distance r (pm) for a 1.46 M solution of NiCl_2 in D_2O . From [5].

-11-

The secondary hydration sheath has also been studied using vibrational spectroscopy. In the presence of highly-charged cations, such as Al^{3+} , Cr^{3+} and Rh^{3+} , frequency shifts can be seen due to the entire primary and secondary hydration structure, although the *number* of water molecules hydrating the cation is somewhat lower than that expected on the basis of neutron data or LAXS data. By contrast, comparison of the Raman and neutron diffraction data for Sc^{3+} indicates the presence of $[\text{Sc}(\text{H}_2\text{O})_7]^{3+}$ in solution, a result supported by the observation of pentagonal bipyramidal coordination in the x-ray structure of the aqua di- μ -hydroxo dimer $[\text{Sc}_2(\text{OH})_2]^{4+}$.

The hydration of more inert ions has been studied by ^{18}O labelling mass spectrometry. ^{18}O -enriched water is used, and an equilibrium between the solvent and the hydration around the central ion is first attained, after which the cation is extracted rapidly and analysed. The method essentially reveals the number of oxygen atoms that exchange slowly on the timescale of the extraction, and has been used to establish the existence of the stable $[\text{Mo}_3\text{O}_4]^{4+}$ cluster in aqueous solution.

One of the most powerful methods for the investigation of hydration is NMR, and both ^1H and ^{17}O nuclei

have been used. By using paramagnetic chemical shift reagents such as Co^{2+} and Dy^{3+} , which essentially shift the peak position of bulk water, hydration measurements have been carried out using ^1H NMR on a number of tripositive ions. ^{17}O NMR measurements have also been carried out and, by varying the temperature, the dynamics of water exchange can also be studied. The hydration numbers measured by this technique are those for the inner hydration sheath and, again, values of four are found for Be^{2+} and six for many other di- and tri-positive cations. The hydration numbers for the alkali metals' singly-positive cations have also been determined by this method, with values of around three being found.

Hydration and solvation have also been studied by conductivity measurements; these measurements give rise to an effective radius for the ion, from which a hydration number can be calculated. These effective radii are reviewed in the next section.

A2.4.3 IONIC CONDUCTIVITY

A2.4.3.1 THE BOLTZMANN TRANSPORT EQUATION

The motion of particles in a fluid is best approached through the Boltzmann transport equation, provided that the combination of internal and external perturbations does not substantially disturb the equilibrium. In other words, our starting point will be the statistical thermodynamic treatment above, and we will consider the effect of both the internal and external fields. Let the chemical species in our fluid be distinguished by the Greek subscripts α, β, \dots and let $f_\alpha(\mathbf{r}, \mathbf{c}, t)$ $dV d c_x d c_y d c_z$ be the number of molecules of type α located in volume dV at \mathbf{r} and having velocities between c_x and $c_x + d c_x$ etc. Note that we expect \mathbf{c} and \mathbf{r} are independent. Let the external force on molecules of type α be \mathbf{F}_α . At any space point, \mathbf{r} , the rate of increase of f_α , $(\partial f_\alpha / \partial t)$, will be determined by:

-12-

- (1) the nett flow of molecules of type α with velocity \mathbf{c} into dV , $-\nabla \cdot (\mathbf{c} f_\alpha) = -\mathbf{c} \cdot \text{grad}(f_\alpha)$;
- (2) acceleration of molecules in dV into and out of the range $d\mathbf{c}$ by \mathbf{F}_α , $-(1/m_\alpha) \nabla_{\mathbf{c}} \cdot (\mathbf{F}_\alpha f_\alpha)$ accelerations, de-excitations, etc of local molecules by intermolecular collisions. This is the most troublesome part analytically: it will be composed of terms corresponding to *gain* of molecules in dV at \mathbf{c} and a *loss* by collisions. We will not write down an explicit expression for the nett collision effect, but rather write $(\partial f_\alpha / \partial t)_{\text{coll}}$.

The nett result is

$$\left(\frac{\partial f_\alpha}{\partial t} \right) + \mathbf{c} \cdot \bar{\nabla} f_\alpha + \frac{1}{m_\alpha} \bar{\nabla}_{\mathbf{c}} \cdot \mathbf{F}_\alpha f_\alpha = \left(\frac{\partial f_\alpha}{\partial t} \right)_{\text{coll}} \quad (\text{A2.4.24})$$

which is Boltzmann's transport equation. To make progress, we make the assumption now that in first order, f_α on the left-hand side of the equation is the equilibrium value, f_α^0 . We further make the so-called relaxation-time approximation that $(\partial f_\alpha / \partial t)_{\text{coll}} = (f_\alpha^0 - f_\alpha) / \tau$, where τ is, in principle, a function of \mathbf{c} , or at least of $|\mathbf{c}|$. We then have, from A2.4.24 $(z_\alpha e_0 / m_\alpha) \mathbf{E} \cdot \bar{\nabla}_{\mathbf{c}} (f_\alpha) = ((f_\alpha^0 - f_\alpha) / \tau)$, where the charge on ions of type α is $z_\alpha e_0$ and the applied electric field is \mathbf{E} . Given that the current density, \mathbf{J} , in dV is

$$\iiint z_\alpha e_0 \mathbf{c} f_\alpha d\mathbf{c} \equiv \iiint z_\alpha e_0 \mathbf{c} (f_\alpha - f_\alpha^0) d\mathbf{c} \quad (\text{A2.4.25})$$

substituting (A2.4.25) from the Boltzmann equation and evaluating the conductivity, κ_α , from ions of type α , we have, after carrying out the spatial integrations

$$\kappa_\alpha = -\frac{z_\alpha^2 e_0^2}{m_\alpha} \iiint \tau \mathbf{c} \cdot \bar{\nabla}_c f_\alpha d\mathbf{c} = \frac{z_\alpha^2 e_0^2}{m_\alpha} \iiint f_\alpha \nabla_c (\tau \mathbf{c}) d\mathbf{c} \approx \frac{z_\alpha^2 e_0^2 N \tau}{m_\alpha} \quad (\text{A2.4.26})$$

where N is the number of ions per unit volume. From elementary analysis, if we define a mean ionic drift velocity v in the direction of the applied electric field, \mathbf{E} , the conductivity contribution from ions of type α will be $Nz_\alpha e_0 v / |\mathbf{E}| \equiv Nz_\alpha e_0 u$, where u is termed the *mobility*; from which we can see that $u = z_\alpha e_0 \tau / m_\alpha$.

A2.4.3.2 THE ELEMENTARY THEORY OF IONIC CONDUCTIVITY [7]

An alternative approach is to consider ions of charge $z_\alpha e_0$ accelerated by the electric field strength, \mathbf{E} , being subject to a frictional force, \mathbf{K}_R , that increases with velocity, \mathbf{v} , and is given, for simple spherical ions of radius r_α , by the Stokes formula, $\mathbf{K}_R = 6\pi\eta r_\alpha \mathbf{v}$, where η is the viscosity of the medium. After a short induction period, the velocity attains a limiting value, \mathbf{v}_{\max} , corresponding to the exact balance between the electrical and frictional forces:

$$z_\alpha e_0 \mathbf{E} = 6\pi\eta r_\alpha \mathbf{v}_{\max} \quad (\text{A2.4.27})$$

-13-

and the terminal velocity is given by

$$\mathbf{v}_{\max} = z_\alpha e_0 \mathbf{E} / (6\pi\eta r_\alpha) \quad (\text{A2.4.28})$$

and it is evident that $\tau = m_\alpha / (6\pi\eta r_\alpha)$. It follows that, for given values of η and \mathbf{E} , each type of ion will have a transport velocity dependent on the charge and the radius of the solvated ion and a direction of migration dependent on the sign of the charge.

For an electrolyte solution containing both anions and cations, with the terminal velocity of the cations being \mathbf{v}_{\max}^+ , and the number of ions of charge $z^+ e_0$ per unit volume being N^+ , the product $AN^+ \mathbf{v}_{\max}^+$ corresponds just to that quantity of positive ions that passes per unit time through a surface of area A normal to the direction of flow. The product $AN^- \mathbf{v}_{\max}^-$ can be defined analogously, and the amount of charge carried through this surface per unit time, or the current per area A , is given by

$$\begin{aligned} I &= I^+ + I^- = Ae_0(N^+ z^+ \mathbf{v}_{\max}^+ + N^- z^- \mathbf{v}_{\max}^-) \\ &= Ae_0(N^+ z^+ u^+ + N^- z^- u^-) \times |\mathbf{E}| \end{aligned} \quad (\text{A2.4.29})$$

where the u are the mobilities defined above. If the potential difference between the electrodes is ΔV , and the distance apart of the electrodes is l , then the magnitude of the electric field $|\mathbf{E}| = \Delta V / l$. Since $I = G\Delta V$, where G is the conductance, G is given by

$$G = (A/l)e_0(N^+ z^+ u^+ + N^- z^- u^-). \quad (\text{A2.4.30})$$

The conductivity is obtained from this by division by the geometric factor (A/l) , giving

$$\kappa = e_0(N^+ z^+ u^+ + N^- z^- u^-). \quad (\text{A2.4.31})$$

It is important to recognize the approximations made here: the electric field is supposed to be sufficiently small so that the equilibrium distribution of velocities of the ions is essentially undisturbed. We are also assuming that we can use the relaxation approximation, and that the relaxation time τ is independent of the ionic concentration and velocity. We shall see below that these approximations break down at higher ionic concentrations: a primary reason for this is that ion-ion interactions begin to affect both τ and F_{α} , as we shall see in more detail below. However, in very dilute solutions, the ion scattering will be dominated by solvent molecules, and in this limiting region A2.4.31 will be an adequate description.

Measurement of the conductivity can be carried out to high precision with specially designed cells. In practice, these cells are calibrated by first measuring the conductance of an accurately known standard, and then introducing the sample under study. Conductances are usually measured at about 1 kHz AC rather than with DC voltages in order to avoid complications arising from electrolysis at anode and cathode [8].

-14-

The conductivity of solutions depends, from A2.4.31, on both the concentration of ions and their mobility. Typically, for 1 M NaCl in water at 18°C, a value of $7.44 \Omega^{-1} \text{m}^{-1}$ is found: by contrast, 1 M H₂SO₄ has a conductivity of $36.6 \Omega^{-1} \text{m}^{-1}$ at the same temperature and acetic acid, a *weak* electrolyte, has a conductivity of only $0.13 \Omega^{-1} \text{m}^{-1}$.

In principle, the effects of the concentration of ions can be removed by dividing A2.4.31 by the concentration. Taking Avagadro's constant as L and assuming a concentration of solute $c \text{ mol m}^{-3}$, then from the electroneutrality principle we have $N^+ z^+ = N^- z^- = v_{\pm} z_{\pm} c L$ and clearly

$$\Lambda \equiv \frac{\kappa}{c} = v_{\pm} z_{\pm} L e_0 (u^+ + u^-) \equiv v_{\pm} z_{\pm} F (u^+ + u^-) \quad (\text{A2.4.32})$$

where Λ is termed the *molar conductivity* and F is the Faraday, which has the numerical value $96\,485 \text{ C mol}^{-1}$.

In principle, Λ should be independent of the concentration according to A2.4.31, but this is not found experimentally. At very low concentrations Λ is roughly constant, but at higher concentrations substantial changes in the mobilities of the ions are found, reflecting increasing ion-ion interactions. Even at low concentrations the mobilities are not constant and, empirically, for strong electrolytes, Kohlrausch observed that Λ decreased with concentration according to the expression

$$\Lambda = \Lambda_0 - k \sqrt{c/c^0} \quad (\text{A2.4.33})$$

where Λ_0 is the molar conductivity extrapolated to zero concentration and c^0 is the standard concentration (usually taken as 1 M). Λ_0 plays an important part in the theory of ionic conductivity since at high dilution the ions should be able to move completely independently, and as a result equation A2.4.32 expressed in the form

$$\Lambda_0 = v_+ z_+ F (u_0^+ + u_0^-) \equiv v_+ \lambda_0^+ + v_- \lambda_0^- \quad (\text{A2.4.34})$$

is exactly true.

The fraction of current carried by the cations is clearly $I^+/(I^+ + I^-)$; this fraction is termed the *transport number* of the cations, t^+ , and evidently

$$t^+ = \frac{u^+}{u^+ + u^-}. \quad (\text{A2.4.35})$$

In general, since the mobilities are functions of the concentration, so are the transport numbers, but limiting transport numbers can be defined by analogy to A2.4.34. The measurement of transport numbers can be carried out straightforwardly, allowing an unambiguous partition of the conductivity and assessment of the individual ionic mobilities at any concentration. Some typical transport numbers are given in [table A2.4.1 \[7\]](#), for aqueous solutions at 25° and some limiting single-ion molar conductivities are given in [table A2.4.2 \[7\]](#).

-15-

Table A2.4.1. Typical transport numbers for aqueous solutions.

Electrolyte	t_0^+	$t_0^- (= 1 - t_0^+)$
KCl	0.4906	0.5094
NH ₄ Cl	0.4909	0.5091
HCl	0.821	0.179
KOH	0.274	0.726
NaCl	0.3962	0.6038
NaOOCCH ₃	0.5507	0.4493
KOOCCH ₃	0.6427	0.3573
CuSO ₄	0.375	0.625

Table A2.4.2. Limiting single-ion conductivities.

Ion	λ_0^+, λ_0^- ($\Omega^{-1} \text{ mol}^{-1} \text{ cm}^2$)	Ion	λ_0^+, λ_0^- ($\Omega^{-1} \text{ mol}^{-1} \text{ cm}^2$)
H ⁺	349.8	Ag ⁺	62.2
OH ⁻	197	Na ⁺	50.11
K ⁺	73.5	Li ⁺	38.68
NH ₄ ⁺	73.7	[Fe(CN) ₆] ⁴⁻	440
Rb ⁺	77.5	[Fe(CN) ₆] ³⁻	303
Cs ⁺	77	[CrO ₄] ²⁻	166
Ba ²⁺	126.4	[SO ₄] ²⁻	161.6
Ca ²⁺	119.6	I ⁻	76.5
Mg ²⁺	106	Cl ⁻	76.4
		NO ₃ ⁻	71.5
		CH ₃ COO ⁻	40.9
		C ₆ H ₅ COO ⁻	32.4

A2.4.3.3 THE SOLVATION OF IONS FROM CONDUCTIVITY MEASUREMENTS

We know from [equation A2.4.32](#) and [equation A2.4.34](#) that the limiting ionic conductivities are directly proportional to the limiting ionic mobilities: in fact

$$\lambda_0^+ = z^+ F u_0^+ \quad (\text{A2.4.36})$$

$$\lambda_0^- = z^- F u_0^-. \quad (\text{A2.4.37})$$

At infinite dilution, the assumption of a constant relaxation time is reasonable and, using Stokes law as well, we have

$$u_0 = ze_0/6\pi\eta r. \quad (\text{A2.4.38})$$

At first sight, we would expect that the mobilities of more highly-charged ions would be larger, but it is apparent from [table A2.4.2](#) that this is not the case; the *mobilities* of Na^+ and Ca^{2+} are comparable, even though [equation A2.4.38](#) would imply that the latter should be about a factor of two larger. The explanation lies in the fact that r also increases with charge, which, in turn, can be traced to the increased size of the hydration sheath in the doubly-charged species, since there is an increased attraction of the water dipoles to the more highly-charged cations.

It is also possible to explain, from hydration models, the differences between equally-charged cations, such as the alkali metals ($\lambda_0^{\text{K}^+} = 73.5$, $\lambda_0^{\text{Na}^+} = 50.1$ and $\lambda_0^{\text{Li}^+} = 38.68$, all in units of $\Omega^{-1} \text{ mol}^{-1} \text{ cm}^2$). From atomic physics it is known that the radii of the bare ions is in the order $\text{Li}^+ < \text{Na}^+ < \text{K}^+$. The attraction of the water dipoles to the cation increases strongly as the distance between the charge centres of the cation and water molecule decreases, with the result that the total radius of the ion and bound water molecules actually increases in the order $\text{K}^+ < \text{Na}^+ < \text{Li}^+$, and this accounts for the otherwise rather strange order of mobilities.

The differing extent of hydration shown by the different types of ion can be determined experimentally from the amount of water carried over with each type of ion. A simple measurement can be carried out by adding an electrolyte such as LiCl to an aqueous solution of sucrose in a Hittorf cell. Such a cell consists of two compartments separated by a narrow neck [7]; on passage of charge the strongly hydrated Li^+ ions will migrate from the anode to the cathode compartment, whilst the more weakly hydrated Cl^- ions migrate towards the anode compartment; the result is a slight increase in the concentration of sucrose in the anode compartment, since the sucrose itself is essentially electrically neutral and does not migrate in the electric field. The change in concentration of the sucrose can either be determined analytically or by measuring the change in rotation of plane polarized light transmitted through the compartment. Measurements carried out in this way lead to hydration numbers for ions, these being the number of water molecules that migrate with each cation or anion. Values of 10–12 for Mg^{2+} , 5.4 for K^+ , 8.4 for Na^+ and 14 for Li^+ are clearly in reasonable agreement with the values inferred from the Stokes law arguments above. They are also in agreement with the measurements carried out using large organic cations to calibrate the experiment, since these are assumed not to be hydrated at all.

Anions are usually less strongly hydrated, as indicated above, and from [equation A2.4.38](#) this would suggest that increasing the charge on the anion should lead unequivocally to an increase in mobility and hence to an increase in limiting ionic conductivity. An inspection of [table A2.4.2](#) shows this to be borne out to some extent by the limited data

available. The rather low conductivities exhibited by organic anions is a result of their considerably larger size; even taking hydration into account, their total diameter normally exceeds that of the simple anions.

One anomaly immediately obvious from [table A2.4.2](#) is the much higher mobilities of the proton and hydroxide ions than expected from even the most approximate estimates of their ionic radii. The origin of this behaviour lies in the way in which these ions can be accommodated into the water structure described above. Free protons cannot exist as such in aqueous solution: the very small radius of the proton would lead to an enormous electric field that would polarize any molecule, and in an aqueous solution the proton immediately attaches itself to the oxygen atom of a water molecule, giving rise to an H_3O^+ ion. In this ion, however, the positive charge does not simply reside on a single hydrogen atom; NMR spectra show that all three hydrogen atoms are equivalent, giving a structure similar to that of the NH_3 molecule. The formation of a water cluster around the H_3O^+ ion and its subsequent fragmentation may then lead to the positive charge being transmitted across the cluster without physical migration of the proton, and the limiting factor in proton motion becomes hydrogen-bonded cluster formation and not conventional migration. It is clear that this model can be applied to the anomalous conductivity of the hydroxide ion without any further modification. Hydrogen-atom tunnelling from a water molecule to an OH^- ion will leave behind an OH^- ion, and the migration of OH^- ions is, in fact, traceable to the migration of H^+ in the opposite direction. This type of mechanism is supported by the observation of the effect of temperature. It is found that the mobility of the proton goes through a maximum at a temperature of 150°C (where, of course, the measurements are carried out under pressure). This arises because as the temperature is increased from ambient, the main initial effect is to loosen the hydrogen-bonded local structure that inhibits reorientation. However, at higher temperatures, the thermal motion of the water molecules becomes so marked that cluster formation becomes inhibited.

The complete hydration shell of the proton consists of both the central H_3O^+ unit and further associated water molecules; mass spectrometric evidence would suggest that a total of four water molecules form the actual H_9O_4^+ unit, giving a hydration number of four for the proton. Of course, the measurement of this number by the Hittorf method is not possible since the transport of protons takes place by a mechanism that does not involve the actual movement of this unit. By examining concentration changes and using large organic cations as calibrants, a hydration number of one is obtained, as would be expected.

From [equation A2.4.36](#) and [equation A2.4.37](#), we can calculate the magnitudes of the mobilities for cations and anions. As an example, from [table A2.4.2](#), the limiting ionic conductivity for the Na^+ ion is $50.11 \times 10^{-4} \text{ m}^2 \Omega^{-1} \text{ mol}^{-1}$. From this we obtain a value of $\mu_0^+ = \lambda_0^+ / F \approx 5.19 \times 10^{-8} \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$, which implies that in a field of 100 V m^{-1} , the sodium ion would move a distance of about 2 cm in 1 h. The mobilities of other ions have about the same magnitude ($(4-8) \times 10^{-8} \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$), with the marked exception of the proton. This has an apparent mobility of $3.63 \times 10^{-7} \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$, almost an order of magnitude higher, reflecting the different conduction mechanism described above. These mobilities give rise to velocities that are small compared to thermal velocities, at least for the small electric fields normally used, confirming the validity of the analysis carried out above.

With the knowledge now of the magnitude of the mobility, we can use [equation A2.4.38](#) to calculate the radii of the ions; thus for lithium, using the value of $0.00089 \text{ kg m}^{-1} \text{ s}^{-1}$ for the viscosity of pure water (since we are using the conductivity at infinite dilution), the radius is calculated to be $2.38 \times 10^{-10} \text{ m}$ ($\approx 2.38 \text{ \AA}$). This can be contrasted with the crystalline ionic radius of Li^+ , which has the value 0.78 \AA . The difference between these values reflects the presence of the hydration sheath of water molecules; as we showed above, the transport measurements suggest that Li^+ has a hydration number of 14.

From [equation A2.4.38](#) we can, finally, deduce Walden's rule, which states that the product of the ionic mobility at infinite dilution and the viscosity of the pure solvent is a constant. In fact

$$u_0\eta = ze_0/6\pi r = \text{constant} \quad (\text{A2.4.39})$$

whereby $\lambda_0\eta = \text{constant}$ and $\Lambda_0\eta = \text{constant}$. This rule permits us to make an estimate of the change of u_0 and λ_0 with a change in temperature and the alteration of the solvent, simply by incorporating the changes in viscosity.

A2.4.4 IONIC INTERACTIONS

The McMillan-Mayer theory offers the most useful starting point for an elementary theory of ionic interactions, since at high dilution we can incorporate all ion-solvent interactions into a limiting chemical potential, and deviations from solution ideality can then be explicitly connected with ion-ion interactions only. Furthermore, we may assume that, at high dilution, the interaction energy between two ions (assuming only two are present in the solution) will be of the form

$$u(r_{12}) = \begin{cases} +\infty & r \leq d \\ \frac{z_1 z_2 e_0^2}{4\pi\epsilon\epsilon_0 r_{12}} & r > d \end{cases} \quad (\text{A2.4.40})$$

where in the limiting dilution law, first calculated by Debye and Hückel (DH), d is taken as zero. It should be emphasized that $u(r)$ is not the potential of mean force, $W(r)$, defined in the McMillan-Mayer theory above; this latter needs to be worked out by calculating the *average electrostatic potential* (AEP), $\psi_i(r)$ surrounding a given ion, i , with charge $z_i e_0$. This is because although the interaction between any ion j and this central ion is given by A2.4.40, the work required to bring the ion j from infinity to a distance r from i is influenced by other ions surrounding i . Oppositely charged ions will tend to congregate around the central ion, giving rise to an ionic 'atmosphere' or *cosphere*, which intervenes between ions i and j , *screening* the interaction represented in A2.4.40. The resulting AEP is the sum of the central interaction and the interaction with the ionic cosphere, and it can be calculated by utilizing the Poisson equation:

$$\nabla^2 \psi_i(r) = -\frac{q_{(i)}(r)}{\epsilon\epsilon_0} \quad (\text{A2.4.41})$$

where $q_{(i)}(r)$ is the charge density (i.e. the number of charges per unit volume) at a distance r from the centre i . In terms of the pair correlation coefficient defined above:

$$q_{(i)}(r) = e_0 \sum_j z_j \rho_j g_{ji}(r) \quad (\text{A2.4.42})$$

where ρ_j is the number density of ions of type j . From [A2.4.20](#) above, we have $g_{ji} = e^{-\beta W_{ji}(r)}$, and it is the potential

of the mean force W that is related to the AEP. The first approximation in the DH theory is then to write

$$W_{ji}(r) \approx e_0 z_j \psi_i(r) \quad (\text{A2.4.43})$$

whence $g_{ji}(r) \approx e^{-\beta z_j e_0 \psi_i(r)}$, which was originally given by DH as a consequence of the Boltzmann law, but clearly has a deep connection with statistical thermodynamics of fluids. From (A2.4.41) and (A2.4.42), we have

$$\nabla^2 \psi_i(r) = -\frac{e_0}{\epsilon \epsilon_0} \sum_j z_j \rho_j e^{-\beta z_j e_0 \psi}. \quad (\text{A2.4.44})$$

The major deficiency of the equation as written is that there is no excluded volume, a deficiency DH could rectify for the central ion, but not for all ions around the central ion. This deficiency has been addressed within the DH framework by Outhwaite [9].

To solve A2.4.44, the assumption is made that $\beta z_j e_0 \psi_i(r) \ll 1$, so the exponential term can be expanded. Furthermore, we must have $\sum_j z_j \rho_j = 0$ since the overall solution is electroneutral. Finally we end up with

$$\nabla^2 \psi_i(r) = -\frac{e_0}{\epsilon \epsilon_0} \sum_j z_j \rho_j e^{-\beta z_j e_0 \psi_i(r)} \approx \frac{e_0^2 \psi_i(r)}{\epsilon \epsilon_0 k T} \sum_j \rho_j z_j^2 \equiv \kappa^2 \psi_i(r) \quad (\text{A2.4.45})$$

where κ has the units of inverse length. The *ionic strength*, I , is defined as

$$I = \frac{1}{2} \sum_j \rho_j (z_j e_0)^2 \quad (\text{A2.4.46})$$

so $\kappa^2 = 2I/\epsilon \epsilon_0 k T$. For aqueous solutions at 25°C, $\kappa^{-1} \text{ m}^{-1} = 3.046 \times 10^{-10}/(I)^{1/2}$. Equation A2.4.45 can be solved straightforwardly providing the assumption is made that the mean cosphere around each ion is spherical. On this basis A2.4.45 reduces to

$$\frac{1}{r^2} \frac{d}{dr} \left(r^2 \frac{d\psi_i}{dr} \right) = \kappa^2 \psi \quad (\text{A2.4.47})$$

which solves to give

$$\psi_i(r) = \frac{A}{r} e^{-\kappa r} + \frac{B}{r} e^{+\kappa r} \quad (\text{A2.4.48})$$

where A and B are constants of integration. B is clearly zero and, in the original DH model with no core repulsion term, A was fixed by the requirement that as $r \rightarrow 0$, $\psi_i(r)$ must behave as $z_i e_0 / 4\pi \epsilon \epsilon_0 r$. In the extended model,

equation A2.4.47 is also solved for the central ion, and the integration constants determined by matching ψ and its derivative at the ionic radius boundary. We finally obtain, for the limiting DH model:

$$\psi_i(r) = \frac{z_i e_0}{4\pi \epsilon \epsilon_0 r} e^{-\kappa r}. \quad (\text{A2.4.49})$$

Given that $W_{ji}(r) \approx e_0 z_j \psi_i(r)$, we finally obtain for the pair correlation coefficient

$$g_{ji}(r) = \exp[-\beta z_j e_0 \psi_i(r)] \approx 1 - \frac{z_i z_j e_0^2}{4\pi \epsilon \epsilon_0 k T r} e^{-\kappa r}. \quad (\text{A2.4.50})$$

An alternative derivation, which uses the Ornstein-Zernicke equation (equation (A2.4.12)), was given by Lee [2]. The Ornstein-Zernicke equation can be written as

$$h_{ij}(rr') - C_{ij}(rr') = \sum_l \rho_l \int ds C_{il}(rs) h_{lj}(sr'). \quad (\text{A2.4.51})$$

Given $y(\mathbf{r}, \mathbf{r}') \approx 1$ for very dilute solutions, the PY condition leads to

$$\begin{aligned} C_{ij} &\approx f_{ij} = e^{-\beta u_{ij}} - 1 \approx -u_{ij}/kT & \text{and} \\ h_{ij} &= e^{-\beta W_{ij}} - 1 \approx -W_{ij}/kT \end{aligned} \quad (\text{A2.4.52})$$

whence W_{ij} satisfies the integral equation

$$-W_{ij}(r) = -\frac{z_i z_j e_0^2}{4\pi \epsilon \epsilon_0 r} + \sum_l \beta \rho_l \int ds \left(\frac{z_l z_j e_0^2}{4\pi \epsilon \epsilon_0 s} \right) W_{lj}(|\mathbf{r} - \mathbf{s}|). \quad (\text{A2.4.53})$$

This can be solved by standard techniques to yield

$$W_{ij}(r) = \frac{z_i z_j e_0^2}{4\pi \epsilon \epsilon_0} \frac{e^{-\kappa r}}{r}. \quad (\text{A2.4.54})$$

A result identical to that above.

From these results, the thermodynamic properties of the solutions may be obtained within the McMillan-Mayer approximation; i.e. treating the dilute solution as a quasi-ideal gas, and looking at deviations from this model solely in terms of ion-ion interactions, we have

-21-

$$\begin{aligned} \frac{U - U^{\text{ideal}}}{V} &= \frac{1}{2} \sum_i \sum_j \rho_i \rho_j \int dr 4\pi r^2 u_{ij}(r) g_{ij}(r) = \frac{1}{2} \frac{e_0^2}{4\pi \epsilon \epsilon_0} \sum_i \sum_j \rho_i \rho_j z_i z_j \\ &\quad \cdot \int dr \frac{1}{r} [4\pi r^2 g_{ij}(r)] \end{aligned} \quad (\text{A2.4.55})$$

using (A2.4.50) and $\sum_j z_j \rho_j = 0$, this can be evaluated to give

$$\frac{U - U^{\text{ideal}}}{V} = -\frac{\kappa^3 k T}{8\pi}. \quad (\text{A2.4.56})$$

The chemical potential may be calculated from the expression

$$\frac{\mu_j}{kT} = \ln(\rho_j \Lambda_j^3) + \sum_i \frac{\rho_i}{kT} \int_0^1 d\xi \int_0^\infty dr 4\pi r^2 u_{ij}(r) g_{ij}(r; \xi) \quad (\text{A2.4.57})$$

where ξ is a coupling parameter which determines the extent to which ion j is coupled to the remaining ions in the solution. This is closely related to the work of charging the j th ion in the potential of all the other ions and, for the simple expression in (A2.4.57), the charging can be represented by writing the charge on the j th ion in the equation for g_{ij} as $z_j \xi$, with ξ increasing from zero to one as in (A2.4.57). Again, using (A2.4.50) and $\sum_j z_j \rho_j = 0$, we find

$$\frac{\mu_j}{kT} = \ln(\rho_j \Lambda_j^3) - \frac{z_j^2 e_0^2 \kappa}{8\pi \epsilon \epsilon_0 kT}. \quad (\text{A2.4.58})$$

This is, in fact, the main result of the DH analysis. The activity coefficient is clearly given by

$$\ln \gamma_j = -\frac{z_j^2 e_0^2 \kappa}{8\pi \epsilon \epsilon_0 kT} \quad (\text{A2.4.59})$$

where again the activity coefficient is referred to as a dilute non-interacting ‘gas’ of solvated ions in the solvent. From (equation A2.4.46), we can express (A2.4.59) in terms of the ionic strength I , and we find:

$$\ln \gamma_j = -Az_j^2 \sqrt{I} \equiv -1.172z_j^2 \sqrt{I} \quad (\text{A2.4.60})$$

where I is the ionic strength at a standard concentration of 1 mol kg^{-1} and A is a constant that depends solely on the properties of the solvent, and the equivalence refers to water at 25°C . It should be realized that separate ionic activity coefficients are not, in fact, accessible experimentally, and only the *mean* activity coefficient, defined for a binary electrolyte $A_{\nu^+} B_{\nu^-}$ by $\gamma_{\pm} = (\gamma_+^{\nu^+} \gamma_-^{\nu^-})^{1/(\nu^+ + \nu^-)}$ is accessible. Straightforward algebra gives

-22-

$$\ln \gamma_{\pm} = -A|z_+ z_-| \sqrt{I} \equiv -1.172|z_+ z_-| \sqrt{I}. \quad (\text{A2.4.61})$$

We have seen that the DH theory in the limiting case neglects excluded volume effects; in fact the excluded volume of the *central* ion can be introduced into the theory as explained after A2.4.48. If the radius of the ions is taken as a_0 for all ions, we have, in first order,

$$\ln \gamma_{\pm} = -\frac{|z_+ z_-| e_0^2 \kappa}{8\pi \epsilon \epsilon_0 kT (1 + a_0 \kappa)}. \quad (\text{A2.4.62})$$

For different electrolytes of the same charge, expression A2.4.61 predicts the same values for γ_{\pm} ; in other words, the limiting law does not make allowance for any differences in size or other ionic properties. For 1-1 electrolytes, this is experimentally found to be the case for concentrations below 10^{-2} M , although for multi-charged electrolytes, the agreement is less good, even for $10^{-3} \text{ mol kg}^{-1}$. In table A2.4.3 [7] some values of γ_{\pm} calculated from A2.4.61 are collected and compared to some measured activity coefficients for a few simple electrolytes. Figure A2.4.5 [7] shows these properties graphed for 1-1 electrolytes to emphasize the nature of the deviations from the limiting law. It is apparent from the data in both the table and the figure that deviations from the limiting law are far more serious for 2-1 electrolytes, such as H_2SO_4 and Na_2SO_4 . In the latter case, for example, the limiting law is in serious error even at 0.005 M .

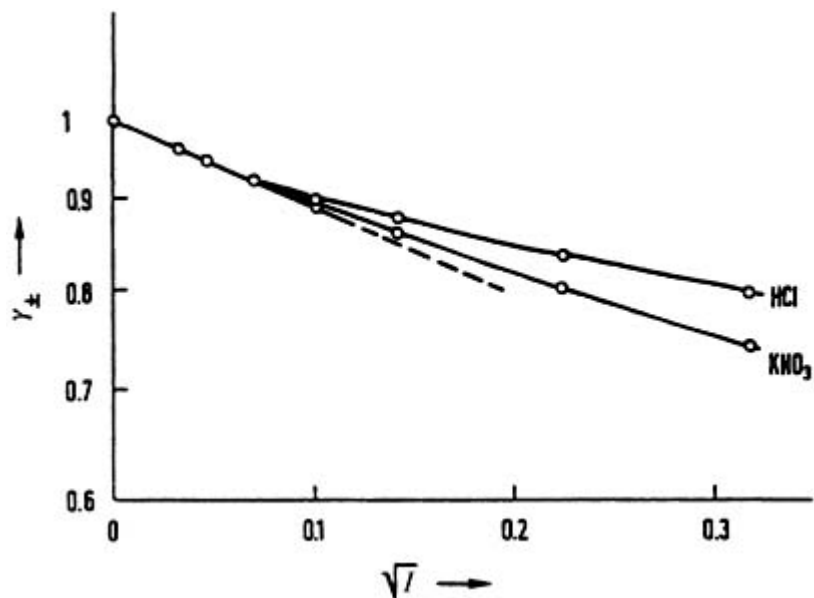


Figure A2.4.5. Theoretical variation of the activity coefficient γ_{\pm} with \sqrt{I} from equation (A2.4.61) and experimental results for 1-1 electrolytes at 25°C. From [7].

-23-

Table A2.4.3. γ_{\pm} values of various electrolytes at different concentration.

m (mol kg ⁻¹)	I	Equation (A2.4.60)	Electrolyte		
			HCl	KNO ₃	LiF
1-1 electrolytes					
0.001	0.001	0.9636	0.9656	0.9649	0.965
0.002	0.002	0.9489	0.9521	0.9514	0.951
0.005	0.005	0.9205	0.9285	0.9256	0.922
0.010	0.010	0.8894	0.9043	0.8982	0.889
0.020	0.020	0.8472	0.8755	0.8623	0.850
0.050	0.050		0.8304	0.7991	
0.100	0.100		0.7964	0.7380	
H ₂ SO ₄ Na ₂ SO ₄					
1-2 or 2-1 electrolytes					
0.001	0.003	0.8795	0.837	0.887	
0.002	0.006	0.8339	0.767	0.847	
0.005	0.015	0.7504	0.646	0.778	
0.010	0.030	0.6662	0.543	0.714	
0.020	0.060		0.444	0.641	
0.050	0.150			0.536	
0.100	0.300		0.379	0.453	
CdSO ₄ CuSO ₄					

2-2 electrolytes				
0.001	0.004	0.7433	0.754	0.74
0.002	0.008	0.6674	0.671	
0.005	0.020	0.5152	0.540	0.53
0.010	0.040		0.432	0.41
0.020	0.080		0.336	0.315
0.050	0.200		0.277	0.209
0.100	0.400		0.166	0.149

-24-

A2.4.4.1 BEYOND THE LIMITING LAW

At concentrations greater than $0.001 \text{ mol kg}^{-1}$, equation A2.4.61 becomes progressively less and less accurate, particularly for unsymmetrical electrolytes. It is also clear, from table A2.4.3, that even the properties of electrolytes of the same charge type are no longer independent of the chemical identity of the electrolyte itself, and our neglect of the factor κa_0 in the derivation of A2.4.61 is also not valid. As indicated above, a partial improvement in the DH theory may be made by including the effect of finite size of the central ion alone. This leads to the expression

$$\ln \gamma_{\pm} = -\frac{|z_+z_-|e_0^2\kappa}{8\pi\epsilon\epsilon_0kT(1+a_0\kappa)} = -\frac{A|z_+z_-|\sqrt{I}}{(1+Ba_0\sqrt{I})} \quad (\text{A2.4.63})$$

where the parameter B also depends only on the properties of the solvent, and has the value $3.28 \times 10^9 \text{ m}^{-1}$ for water at 25°C . The parameter a_0 is adjustable in the theory, and usually the product $B a_0$ is close to unity.

Even A2.4.63 fails at concentrations above about 0.1 M, and the mean activity coefficient for NaCl shown in figure A2.4.6 [2] demonstrates that in more concentrated solutions the activity coefficients begin to rise, often exceeding the value of unity. This rise can be traced to more than one effect. As we shall see below, the inclusion of ion-exclusion effects for all the ions gives rise to this phenomenon. In addition, the ion-ion interactions at higher concentrations cannot really be treated by a hard-sphere model anyway, and models taking into account the true ion-ion potential for solvated ions at close distances are required. Furthermore, the number of solvent molecules essentially immobilized in the solvent sheath about each ion becomes a significant fraction of the total amount of solvent present. This can be exemplified by the case of sulphuric acid: given that each proton requires four water molecules for solvation and the sulphate ion can be estimated to require one, each mole of H_2SO_4 will require 9 mol of water. One kilogram of water contains approximately 55 mol, so that a 1 mol kg^{-1} solution of H_2SO_4 will only leave 46 mol of 'free' water. The effective concentration of an electrolyte will, therefore, be appreciably higher than its analytical value, and this effect becomes more marked the higher the concentration. A further effect also becomes important at higher concentrations: implicit in our whole approach is the assumption that the free energy of the solvation of the ions is independent of concentration. However, if we look again at our example of sulphuric acid, it is clear that for $m > 6 \text{ mol kg}^{-1}$, apparently all the water is present in the solvation sheaths of the ions! Of course what actually occurs is that the extent of solvation of the ions changes, in effect decreasing the stability of the ions. However, this process essentially invalidates the McMillan-Mayer approach, or at the least requires the potential of mean force to be chosen in such a way as to reproduce the change in solvation energy.

-25-

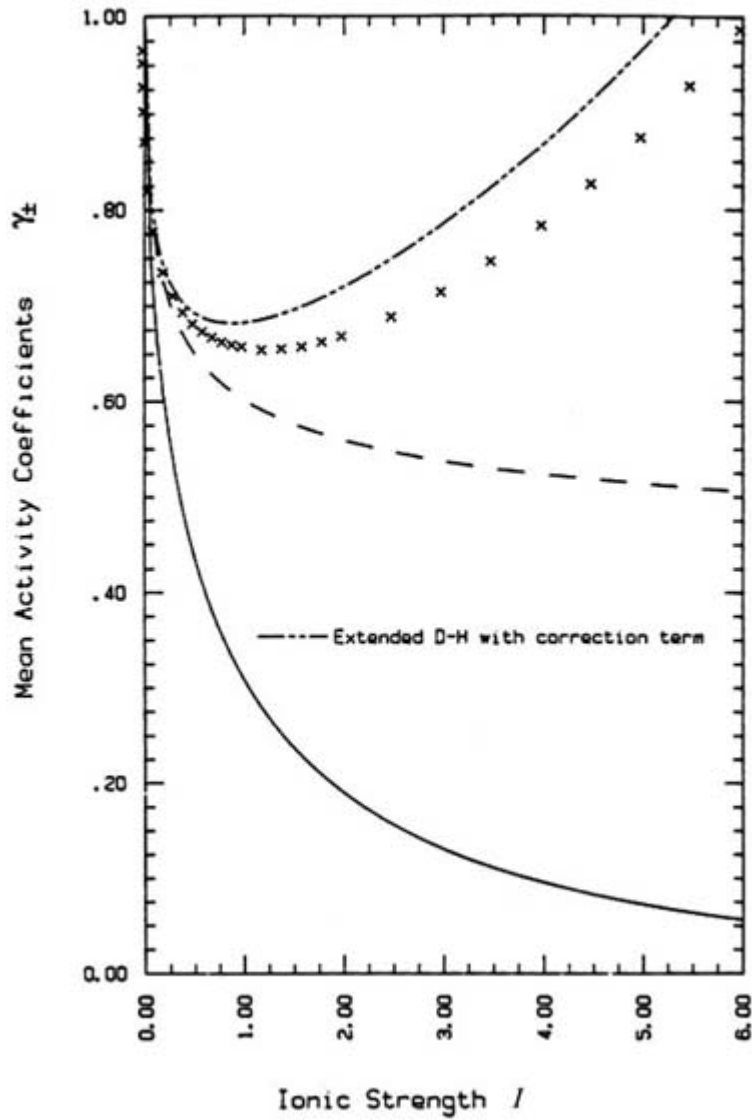


Figure A2.4.6. Mean activity coefficient for NaCl solution at 25 °C as a function of the concentration: full curve from ((A2.4.61)); dashed curve from ((A2.4.63)); dot-dashed curve from (A2.4.64). The crosses denote experimental data. From [2].

Within the general DH approach, [equation A2.4.63](#) may be further modified by adding a linear term, as suggested by Hitchcock [8]:

$$\ln \gamma_{\pm} = -\frac{A|z_+z_-|\sqrt{I}}{(1 + Ba_0\sqrt{I})} + bI \quad (\text{A2.4.64})$$

-26-

where b is a parameter to be fitted to the data. As can be seen from [figure A2.4.6](#) this accounts for the behaviour quite well, but the empirical nature of the parameter b and the lack of agreement on its interpretation mean that [A2.4.64](#) can only be used empirically.

The simplest extension to the DH equation that does at least allow the qualitative trends at higher concentrations to be examined is to treat the excluded volume rationally. This model, in which the ion of charge $z_i e_0$ is given an ionic radius d_i is termed the *primitive model*. If we assume an essentially spherical equation for the u_{ij} :

$$u_{ij} = \begin{cases} +\infty & r \leq d_{ij} \\ \frac{z_i z_j e_0^2}{4\pi \epsilon \epsilon_0 r_{ii}} & r > d_{ij}. \end{cases} \quad (\text{A2.4.65})$$

This can be treated analytically within the *mean spherical approximation* for which

$$g_{ij} = 0 \quad r < d_{ij}$$

$$C_{ij}(r) \approx -\frac{z_i z_j e_0^2}{4\pi \epsilon \epsilon_0 k T r} \quad r > d_{ij}. \quad (\text{A2.4.66})$$

These equations were solved by Blum [10], and a characteristic inverse length, 2Γ , appears in the theory. This length is implicitly given by the equation

$$2\Gamma = \alpha \left\{ \sum_{i=1}^n \rho_i \left[\frac{z_i - (\pi/2\Delta) d_i^2 P_n}{1 + \Gamma d_i} \right]^2 \right\}^{1/2} \quad (\text{A2.4.67})$$

where

$$P_n \equiv \frac{1}{\Omega} \sum_k \frac{\rho_k d_k z_k}{1 + \Gamma d_k}$$

$$\Omega \equiv 1 + \frac{\pi}{2\Delta} \sum_k \frac{\rho_k d_k^3}{1 + \Gamma d_k}$$

$$\zeta_n \equiv \sum_k \rho_k (d_k)^n$$

$$\Delta \equiv 1 - \frac{\pi \zeta_3}{6}$$

$$\alpha^2 \equiv \frac{e_0^2}{\epsilon \epsilon_0 k T}.$$

-27-

In this formalism, which is already far from transparent, the internal energy is given by

$$-\frac{U - U^{\text{HS}}}{V k T} = \frac{e_0^2}{4\pi \epsilon \epsilon_0 k T} \left\{ \Gamma \sum_{i=1}^n \frac{\rho_i z_i^2}{1 + \Gamma d_i} + \frac{\pi}{2\Delta} \Omega P_n^2 \right\} \quad (\text{A2.4.68})$$

and the mean activity coefficient by

$$\ln \gamma_{\pm} - \ln \gamma_{\pm}^{\text{HS}} = \frac{U - U^{\text{HS}}}{N k T} - \frac{\alpha^2}{8\rho} \left(\frac{P_n}{\Delta} \right)^2 \quad (\text{A2.4.69})$$

where the superscript HS refers to solutions of the pure hard-sphere model as given by Lee [2]

The integral equation approach has also been explored in detail for electrolyte solutions, with the PY equation proving less useful than the HNC equation. This is partly because the latter model reduces cleanly to the MSA model for small $h(12)$ since

$$C(12) = h(12) - \ln y(12) = h(12) - \ln[1 + h(12)] - \beta u(12) \approx -\beta u(12) + \frac{1}{2}(h(12))^2 + \dots$$

Using the Ornstein-Zernicke equation, numerical solutions for the restricted primitive model can be.

In principle, simulation techniques can be used, and Monte Carlo simulations of the primitive model of electrolyte solutions have appeared since the 1960s. Results for the osmotic coefficients are given for comparison in [table A2.4.4](#) together with results from the MSA, PY and HNC approaches. The primitive model is clearly deficient for values of r_{ij} close to the closest distance of approach of the ions. Many years ago, Gurney [11] noted that when two ions are close enough together for their solvation sheaths to overlap, some solvent molecules become freed from ionic attraction and are effectively returned to the bulk [12].

-28-

Table A2.4.4. Osmotic coefficients obtained by various methods.

Concentration (mol dm ⁻³)	Monte Carlo	MSA	PY	HNC
0.00911	0.97	0.969	0.97	0.97
0.10376	0.945	0.931	0.946	0.946
0.425	0.977	0.945	0.984	0.980
1.00	1.094	1.039	1.108	1.091
1.968	1.346	1.276	1.386	1.340

The potential model for this approach has the form

$$u_{ij} = \begin{cases} +\infty & r_{ij} \leq d_{ij} \\ \frac{z_i z_j e_0^2}{4\pi \epsilon \epsilon_0 r_{ij}} & r_{ij} > d_{ij} + 2r_w \\ (A_g)_{ij} + \frac{z_i z_j e_0^2}{4\pi \epsilon \epsilon_0 r_{ij}} & d_{ij} < r_{ij} < d_{ij} + 2r_w. \end{cases} \quad (\text{A2.4.70})$$

The A_g are essentially adjustable parameters and, clearly, unless some of the parameters in A2.4.70 are fixed by physical argument, then calculations using this model will show an improved fit for purely algebraic reasons. In principle, the radii can be fixed by using tables of ionic radii; calculations of this type, in which just the A_g are adjustable, have been carried out by Friedman and co-workers using the HNC approach [12]. Further refinements were also discussed by Friedman [13], who pointed out that an additional term u_{cavity} is required to account for the fact that each ion is actually in a cavity of low dielectric constant, ϵ_c , compared to that of the bulk solvent, ϵ . A real difficulty discussed by Friedman is that of making the potential continuous, since the discontinuous potentials above may lead to artefacts. Friedman [13] addressed this issue and derived formulae that use repulsion terms of the form $B_{ij} [(r_i + r_j) / r_{ij}]^n$, rather than the hard-sphere model presented above.

A quite different approach was adopted by Robinson and Stokes [8], who emphasized, as above, that if the solute dissociated into ν ions, and a total of h molecules of water are required to solvate these ions, then the real concentration of the ions should be corrected to reflect only the bulk solvent. Robinson and Stokes derive, with these ideas, the following expression for the activity coefficient:

$$\ln \gamma_{\pm} = -\frac{A|z_+z_-|\sqrt{I}}{(1 + Ba_0\sqrt{I})} - \frac{h}{\nu} \ln a_A - \ln[1 + 0.001 W_A(\nu - h)m] \quad (\text{A2.4.71})$$

-29-

where a_A is the activity of the solvent, W_A is its molar mass and m is the molality of the solution. Equation (A2.4.71) has been extensively tested for electrolytes and, provided h is treated as a parameter, fits remarkably well for a large range of electrolytes up to molalities in excess of two. Unfortunately, the values of h so derived, whilst showing some sensible trends, also show some rather counter-intuitive effects, such as an increase from Cl^- to I^- . Furthermore, the values of h are not additive between cations and anions in solution, leading to significant doubts about the interpretation of the equation. Although considerable effort has gone into finding alternative, more accurate expressions, there remains considerable doubt about the overall physical framework.

A2.4.4.2 INTERIONIC INTERACTIONS AND THE CONDUCTIVITY

Equation (A2.4.24) determines the mobility of a single ion in solution, and contains no correction terms for the interaction of the ions themselves. However, in solution, the application of an electric field causes positive and negative ions to move in opposite directions and the symmetrical spherical charge distribution of equation (A2.4.49) becomes distorted. Each migrating ion will attempt to rebuild its atmosphere during its motion, but this rebuilding process will require a certain time, termed the relaxation time, so that the central ion, on its progress through the solution, will always be a little displaced from the centre of charge of its ionic cloud. The result of this is that each central ion will experience a retarding force arising from its associated ionic cloud, which is migrating in the opposite direction, an effect termed the relaxation or asymmetry effect. Obviously this effect will be larger the nearer, on average, the ions are in solution; in other words, the effect will increase at higher ionic concentrations.

In addition to the relaxation effect, the theory of Debye, Hückel and Onsager also takes into account a second effect, which arises from the Stokes law discussed above. We saw that each ion travelling through the solution will experience a frictional effect owing to the viscosity of the liquid. However, this frictional effect itself depends on concentration, since, with increasing concentration, encounters between the solvent sheaths of oppositely charged ions will become more frequent. The solvent molecules in the solvation sheaths are moving with the ions, and therefore an individual ion will experience an additional drag associated with the solvent molecules in the solvation sheaths of oppositely charged ions; this is termed the electrophoretic effect.

The quantitative calculation of the dependence of the electrolyte conductivity on concentration begins from expression (A2.4.49) for the potential exerted by a central ion and its associated ionic cloud. As soon as this ionic motion begins, the ion will experience an effective electric field E_{rel} in a direction opposite to that of the applied electric field, whose magnitude will depend on the ionic mobility. In addition, there is a second effect identified by Onsager due to the movement of solvent sheaths associated with the oppositely charged ions encountered during its own migration through the solution. This second term, the electrophoresis term, will depend on the viscosity of the liquid, and combining this with the reduction in conductivity due to relaxation terms we finally emerge with the Debye-Hückel-Onsager equation [8]:

(A2.4.72)

$$\Lambda = \Lambda_0 - \Lambda_0 \frac{z_+ z_- e_0^2}{24\pi \epsilon \epsilon_0 kT} \frac{2q\kappa}{1 + \sqrt{q}} - \frac{L e_0^2 (z_+ + |z_-|) \kappa}{6\pi \eta}$$

-30-

where

$$q = \frac{z_+ z_-}{z_+ + |z_-|} \frac{\lambda_0^+ + \lambda_0^-}{|z_-| \lambda_0^+ + z_+ \lambda_0^-} \quad (\text{A2.4.73})$$

L is Avagadro's constant and κ is defined above. It can be seen that there are indeed two corrections to the conductivity at infinite dilution: the first corresponds to the relaxation effect, and is correct in (A2.4.72) only under the assumption of a zero ionic radius. For a finite ionic radius, a_0 , the first term needs to be modified: Falkenhagen [8] originally showed that simply dividing by a term $(1 + \kappa a_0)$ gives a first-order correction, and more complex corrections have been reviewed by Pitts *et al* [14], who show that, to a second order, the relaxation term in (A2.4.72) should be divided by $(1 + \kappa a_0)(1 + \kappa a_0 \sqrt{q})$. The electrophoretic effect should also be corrected in more concentrated solutions for ionic size; again to a first order, it is sufficient to divide by the correction factor $(1 + \kappa a_0)$. Note that for a completely dissociated 1–1 electrolyte $q = 0.5$, and expression (A2.4.72) can be re-written in terms of the molarity, c/c^0 , remembering that κ can be expressed either in terms of molalities or molarities; in the latter case we have

$$\kappa^2 = \frac{e_0^2 L c^0}{\epsilon \epsilon_0 kT} \frac{\sum_i z_i^2 c_i}{c^0} = \frac{e_0^2 L c^0}{\epsilon \epsilon_0 kT} \sum_i z_i^2 \nu_i (c/c^0) \quad (\text{A2.4.74})$$

where c^0 is the standard concentration of 1 M. Finally, we see

$$\Lambda = \Lambda_0 - (B_1 \Lambda_0 + B_2) \sqrt{c/c^0} \quad (\text{A2.4.75})$$

in which B_1 and B_2 are independent of concentration. This is evidently identical in form to Kohlrausch's empirical law already discussed earlier (equation A2.4.33). Equation A2.4.75 is valid in the same concentration range as the DH limiting law, i.e. for molalities below 0.01 M for symmetrical electrolytes, and for unsymmetrical electrolytes to even lower values. In fact, for symmetrical singly-charged 1–1 electrolytes, useful estimations of the behaviour can be obtained, with a few per cent error, for up to 0.1 mol kg⁻¹ concentrations, but symmetrical multi-charged electrolytes ($z^+ = z^- \neq 1$) usually show deviations, even at 0.01 M.

At higher concentrations, division by the factor $(1 + \kappa a_0)$ gives rise to an expression of the form

$$\Lambda = \Lambda_0 - \frac{(B_1 \Lambda_0 + B_2)}{1 + \kappa a_0} \sqrt{c/c^0} \quad (\text{A2.4.76})$$

which is valid for concentrations up to about 0.1 M

In aqueous solution, the values of B_1 and B_2 can be calculated straightforwardly. If Λ is expressed in m² Ω⁻¹ mol⁻¹ and c^0 is taken as 1 mol dm⁻³, then, for water at 298 K and $z^+ = |z^-| = 1$, $B_1 = 0.229$ and $B_2 = 6.027 \times 10^{-3}$ m² Ω⁻¹ mol⁻¹. At 291 K the corresponding values are 0.229 and 5.15×10^{-3} m² Ω⁻¹ mol⁻¹ respectively. Some data for selected 1-1 electrolytes is given in table A2.4.5 [7]; it can be seen that Onsager's formula is

very well obeyed.

Table A2.4.5. Experimental and theoretical values of $B_1\Lambda_0 + B_2$ for various salts in aqueous solution at 291 K.

Salt	Observed value of	Calculated value of
	$B_1\Lambda_0 + B_2$ ($\text{m}^2 \Omega^{-1} \text{mol}^{-1}$)	$B_1\Lambda_0 + B_2$ ($\text{m}^2 \Omega^{-1} \text{mol}^{-1}$)
LiCl	7.422×10^{-3}	7.343×10^{-3}
NaCl	7.459×10^{-3}	7.569×10^{-3}
KCl	8.054×10^{-3}	8.045×10^{-3}
LiNO ₃	7.236×10^{-3}	7.258×10^{-3}

A 2.4.5 THE ELECTRIFIED DOUBLE LAYER

Once an electrode, which for our purposes may initially be treated as a conducting plane, is introduced into an electrolyte solution, several things change. There is a substantial loss of symmetry, the potential experienced by an ion will now be not only the screened potential of the other ions but will contain a term arising from the field due to the electrode and a term due to the image charge in the electrode. The structure of the solvent is also perturbed: next to the electrode the orientation of the molecules of solvent will be affected by the electric field at the electrode surface and the nett orientation will derive from both the interaction with the electrode and with neighbouring molecules and ions. Finally, there may be a sufficiently strong interaction between the ions and the electrode surface such that the ions lose at least some of their inner solvation sheath and adsorb onto the electrode surface.

The classical model of the electrified interface is shown in [figure A2.4.7 \[15\]](#), and the following features are apparent.

- (1) There is an ordered layer of solvent dipoles next to the electrode surface, the extent of whose orientation is expected to depend on the charge on the electrode.
- (2) There is, or may be, an inner layer of specifically adsorbed *anions* on the surface; these anions have displaced one or more solvent molecules and have lost part of their inner solvation sheath. An imaginary plane can be drawn through the centres of these anions to form the *inner Helmholtz plane* (IHP).
- (3) The layer of solvent molecules not directly adjacent to the metal is the closest distance of approach of solvated *cations*. Since the enthalpy of solvation of cations is normally substantially larger than that of anions, it is normally expected that there will be insufficient energy to strip the cations of their inner solvation sheaths, and a second imaginary plane can be drawn through the centres of the solvated cations. This second plane is termed the *outer Helmholtz plane* (OHP).
- (4) Outside the OHP, there may still be an electric field and hence an imbalance of anions and cations extending in the form of a *diffuse layer* into the solution.

Owing to the various uncompensated charges at the interface there will be associated changes in the potential, but there are subtleties about what can actually be measured that need some attention.

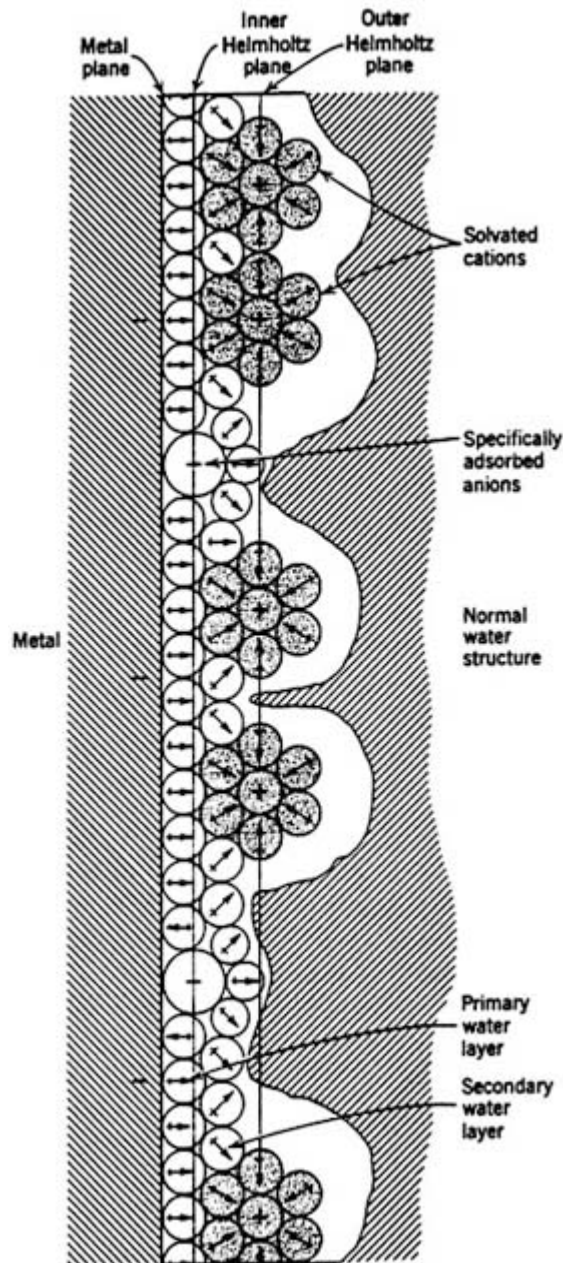


Figure A2.4.7. Hypothetical structure of the electrolyte double layer. From [15].

A2.4.5.1 THE ELECTRODE POTENTIAL

Any measurement of potential must describe a reference point, and we will take as this point the potential of an electron well separated from the metal and at rest *in vacuo*. By reference to figure A2.4.8 [16], we can define the following quantities.

- (1) The Fermi energy ε_F which is the difference in energy between the bottom of the conduction band and the Fermi level; it is positive and in the simple Sommerfeld theory of metals [17],
$$\varepsilon_F = \frac{\hbar^2 k_F^2}{2m} = \frac{\hbar^2 (3\pi^2 n_e)^{2/3}}{2m}$$
 where n_e is the number density of electrons.

- (2) The work function Φ^M , which is the energy required to remove an electron from the inner Fermi level to vacuum.
- (3) The surface potential of the phase, χ^M , due to the presence of surface dipoles. At the metal-vacuum interface these dipoles arise from the fact that the electrons in the metal can relax at the surface to some degree, extending outwards by a distance of the order of 1 Å, and giving rise to a spatial imbalance of charge at the surface.
- (4) The chemical potential of the electrons in the metal, μ_c^M a *negative* quantity.
- (5) The potential energy of the electrons, V , which is a *negative* quantity that can be partitioned into bulk and surface contributions, as shown.

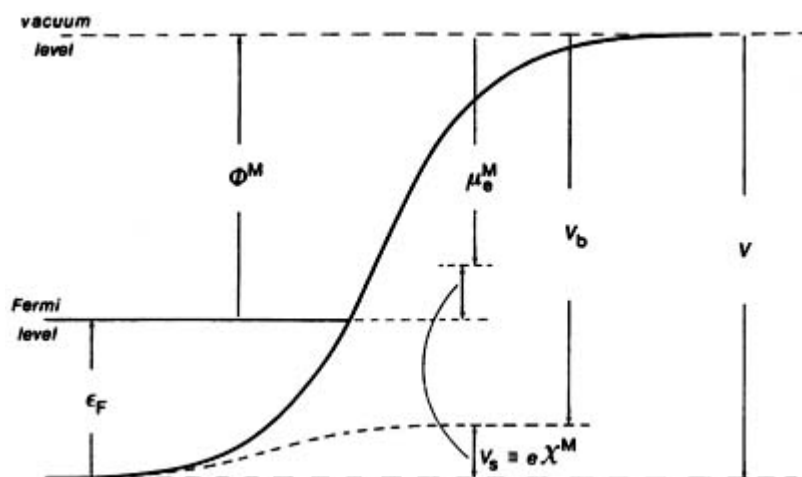


Figure A2.4.8. Potential energy profile at the metal–vacuum interface. Bulk and surface contributions to V are shown separately. From [16].

Of the quantities shown in figure A2.4.8 Φ^M is measurable, as is ϵ_F , but the remainder are not and must be calculated. Values of 1–2 V have been obtained for χ^M , although smaller values are found for the alkali metals.

If two metals with different work functions are placed in contact there will be a flow of electrons from the metal with the *lower* work function to that with the higher work function. This will continue until the *electrochemical* potentials of the electrons in the two phases are equal. This change gives rise to a *measurable* potential difference between the two metals, termed the contact potential or Volta potential difference. Clearly $\Delta_{M_2}^{M_1} \Phi = e_0 \Delta_{M_1}^{M_2} \psi$ where $\Delta_{M_1}^{M_2} \psi$ is the Volta potential difference between a point close to the surface of M_1 and that close to the surface of M_2 . The actual number

-34-

of electrons transferred is very small, so the Fermi energies of the two phases will be unaltered, and only the value of the potential V will have changed. If we assume that the χ^M are unaltered as well, and we define the potential *inside* the metal as ϕ , then the equality of electrochemical potentials also leads to

$$-\mu_c^{M_1} + e_0 \Delta_{M_2}^{M_1} \phi + \mu_c^{M_2} = 0. \quad (\text{A2.4.77})$$

This internal potential, ϕ , is *not* directly measurable; it is termed the Galvani potential, and is the target of most of the modelling discussed below. Clearly we have $\Delta_{M_2}^{M_1} \phi = \Delta_{M_2}^{M_1} \chi + \Delta_{M_2}^{M_1} \psi$.

Once a metal is immersed in a solvent, a second dipolar layer will form at the metal surface due to the

alignment of the solvent dipoles. Again, this contribution to the potential is not directly measurable; in addition, the metal dipole contribution itself will change since the distribution of the electron cloud will be modified by the presence of the solvent. Finally, there will be a contribution from free charges both on the metal and in the electrolyte. The overall contribution to the Galvani potential difference between the metal and solution then consists of these four quantities, as shown in [figure A2.4.9](#) [16]. If the potential due to dipoles at the metal-vacuum interface for the metal is χ^M and for the solvent-vacuum interface is χ^S , then the Galvani potential difference between metal and solvent can be written either as

$$\Delta_S^M \phi = (\chi_M + \delta\chi_M) - (\chi_S + \delta\chi_S) + g(\text{ion}) \equiv g_S^M(\text{dip}) + g(\text{ion}) \quad (\text{A2.4.78})$$

or as

$$\Delta_S^M \phi = \delta_S^M \chi + \Delta_S^M \psi \quad (\text{A2.4.79})$$

where $\delta\chi^M$, $\delta\chi^S$ are the changes in surface dipole for metal and solvent on forming the interface and the g values are local to the interface. In A2.4.78 we pass across the interface, and in A2.4.79 we pass into the vacuum from both the metal and the solvent. As before, the value of $\Delta_S^M \psi$, the Volta potential difference, is measurable experimentally, but it is evident that we cannot associate this potential difference with that due to free charges at the interface, since there are changes in the dipole contribution on both sides as well. Even if there are no free charges at the interface (at the point of zero charge, or PZC), the Volta potential difference is not zero unless $\delta\chi_M = \delta\chi_S$; i.e. the free surfaces of the two phases will still be charged unless the changes in surface dipole of solvent and metal exactly balance. In practice, this is not the case: careful measurements [18] show that $\Delta_{\text{H}_2\text{O}}^{\text{Hg}} \psi \neq 0$ V at the PZC; showing that the dipole changes do not, in fact, compensate. Historically, this discussion is of considerable interest, since a bitter dispute between Galvani and Volta over the origin of the EMF when two different metals are immersed in the same solution could, in principle, be due just to the Volta potential difference between the metals. In fact, it is easy to see that if conditions are such that there are no free charges on either metal, the difference in potential between them, again a measurable quantity, is given by

$$\Delta E_{\sigma=0} = \Delta\Phi + (\Delta_S^{M_1} \psi)_{\sigma=0} - (\Delta_S^{M_2} \psi)_{\sigma=0} \quad (\text{A2.4.80})$$

showing that the difference in work functions would only account for the difference in the electrode potentials if the two Volta terms were actually zero.

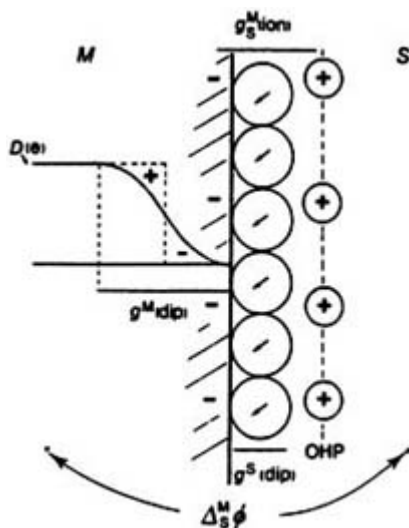


Figure A2.4.9. Components of the Galvani potential difference at a metal–solution interface. From [16].

A2.4.5.2 INTERFACIAL THERMODYNAMICS OF THE DIFFUSE LAYER

Unlike the situation embodied in [section A2.4.1](#), in which the theory was developed in an essentially isotropic manner, the presence of an electrode introduces an essentially non-isotropic element into the equations. Neglecting rotational-dependent interactions, we see that the overall partition function can be written

$$Q(T, V, N) = \frac{q^N}{N! \Lambda^{3N}} \int \cdots \int d\mathbf{r}^N \exp \left[-\beta \left\{ \sum_{i < j}^N u(r_{ij}) + \sum_k^N w(x_k) \right\} \right] \quad (\text{A2.4.81})$$

where $w(x_k)$ is the contribution to the potential energy deriving from the electrode itself, and x_k is the distance between the k th particle and the electrode surface. Clearly, if $w(x_k) \rightarrow 0$, then we recover the partition function for the isotropic fluid. We will work within the McMillan-Mayer theory of solutions, so we will evaluate A2.4.81 for ions in a continuum, recognizing the need to replace $u(r_{ij})$ by a potential of mean force. In a similar way, an exact analytical form for $w(x_k)$ is also expected to prove difficult to derive. A complete account of w must include the following contributions.

-36-

- (1) A short-range contribution, $w^s(x_k)$, which takes into account the nearest distance of approach of the ion to the electrode surface. For ions that do not specifically adsorb this will be the OHP, distance h from the electrode. For ions that do specifically adsorb $w^s(x_k)$ will be more complex, having contributions both from short-range attractive forces and from the energy of de-solvation.
- (2) A contribution from the charge on the surface, $w^{(Q_e)}(x_k)$. If this charge density is written Q_e then elementary electrostatic theory shows that $w^{(Q_e)}(x_k)$ will have the unscreened form

$$w^{(Q_e)}(x_k) = \text{constant} + \frac{z_k e_0 Q_e}{\epsilon \epsilon_0} x_k. \quad (\text{A2.4.82})$$

- (3) An energy of attraction of the ion to its intrinsic image, $w^{(\text{im})}(x_k)$, of unscreened form

$$w^{(\text{im})}(x_k) = \frac{z_k^2 e_0^2}{16\pi \epsilon \epsilon_0 x_k}. \quad (\text{A2.4.83})$$

In addition, the energy of interaction between any two ions will contain a contribution from the mirror potential of the second ion; $u(r_{ij})$ is now given by a short-range term and a term of the form

$$u^{(\text{el})}(r_{ij}) = \frac{z_i z_j e_0^2}{4\pi \epsilon \epsilon_0} \left(\frac{1}{r_{ij}} - \frac{1}{r_{ij}^*} \right) \quad (\text{A2.4.84})$$

where r_{ij}^* is the distance between ion i and the image of ion j .

Note that there are several implicit approximations made in this model: the most important is that we have neglected the effects of the electrode on orientating the solvent molecules at the surface. This is highly significant: image forces arise whenever there is a discontinuity in the dielectric function and the simple model above would suggest that, at least the layer of solvent next to the electrode should have a dielectric

function rather different, especially at low frequencies, from the bulk dielectric function. Implicit in (A2.4.82), (A2.4.83) and A2.4.84 is also the fact that ϵ is assumed independent of x , an assumption again at variance with the simple model presented in A2.4.5.1. In principle, these deficiencies could be overcome by modifying the form of the short-range potentials, but it is not obvious that this will be satisfactory for the description of the image forces, which are intrinsically long range.

The most straightforward development of the above equations has been given by Martynov and Salem [19], who show that to a reasonable approximation in dilute electrolytes:

$$kT \ln(\rho_\alpha^{(1)}(x_\alpha)/\rho_{\alpha 0}^{(1)}) + w_\alpha^s(x_\alpha) + z_\alpha e_0 \phi(x_\alpha) + z_\alpha^2 e_0^2 \delta\phi(x_\alpha) = 0 \quad (\text{A2.4.85})$$

$$kT \ln(g(\mathbf{r}_i, \mathbf{r}_j)) + w_{\alpha\beta}^s(R_{ij}) + z_\alpha e_0 \psi_\alpha(\mathbf{r}_i, \mathbf{r}_j) = 0 \quad (\text{A2.4.86})$$

where $\phi(x)$ is the Galvani potential in the electrolyte at distance x from the electrode, $\delta\phi(x_\alpha)$ is the change in the single-ion mirror-plane potential on moving the ion from infinity to point x_α , $\rho_{\alpha 0}^{(1)}$ is the number density of ions of type α at a

-37-

distance remote from the electrode, and $z_\alpha e_0 \psi_\alpha(\mathbf{r}_i, \mathbf{r}_j)$ is the binary electrostatic potential determined by solution of the relevant Poisson equation:

$$\nabla_j^2 \psi_\alpha(\mathbf{r}_i, \mathbf{r}_j) = -\frac{1}{\epsilon\epsilon_0} \left[z_\alpha e_0 \delta(R_{ij}) + \sum_\beta z_\beta e_0 \rho_\beta^{(1)}(\mathbf{r}_j) \left(\frac{g_{\alpha\beta}^{(2)}(\mathbf{r}_i, \mathbf{r}_j)}{\rho_\alpha^{(1)}(\mathbf{r}_i)/\rho_{\alpha 0}^{(1)}} - \frac{\rho_\beta^{(1)}(\mathbf{r}_j)}{\rho_{\beta 0}^{(1)}} \right) \right]. \quad (\text{A2.4.87})$$

The physical meaning of the second term in A2.4.87 is that the bracket gives the excess concentration of ions β at point \mathbf{r}_j given an ion α at point \mathbf{r}_i . Finally, we need the Poisson equation for the Galvani potential at distance x from the electrode, which is given by

$$\frac{d^2\phi}{dx^2} = -\frac{\sum_\alpha z_\alpha e_0 \rho_\alpha^{(1)}(x)}{\epsilon\epsilon_0}. \quad (\text{A2.4.88})$$

By using the expressions for $\rho_\alpha^{(1)}$ and $g_{\alpha\beta}^{(2)}$ from (A2.4.85) and (A2.4.96) in (A2.4.87) and (A2.4.88), solutions may in principle be obtained for the various potentials and charge distributions in the system. The final equations for a dilute electrolyte are

$$\frac{d^2\phi}{dx^2} = -\frac{\sum_\alpha z_\alpha e_0 n_\alpha^0 \exp\{-\beta[w^s(x) + z_\alpha e_0 \phi(x) + z_\alpha^2 e_0^2 \delta\phi(x)]\}}{\epsilon\epsilon_0} \quad (\text{A2.4.89})$$

where $n_\alpha^0 \equiv \rho_{\alpha 0}^{(1)}$, and which is to be solved under the boundary conditions

$$\begin{aligned} x = h & & \frac{d\phi}{dx} &= -\frac{Q_c}{\epsilon\epsilon_0} \\ x \rightarrow \infty & & \phi &\rightarrow 0 \end{aligned}$$

and

$$\begin{aligned} x_j = 0 & \quad \frac{\partial \psi_\alpha}{\partial x_j} = 0 \\ R_{ij} \rightarrow \infty & \quad \psi_\alpha \rightarrow 0. \end{aligned}$$

$$\begin{aligned} \nabla_j^2 \psi_\alpha(r_i, r_j) = -\frac{1}{\epsilon \epsilon_0} \left\{ z_\alpha e_0 \delta(R_{ij}) + \sum_\beta z_\beta e_0 \rho_\beta^{(1)}(x) [\exp[-(w_{\alpha\beta}^s(R_{ij}) \right. \\ \left. + z_\alpha e_0 \psi_\alpha(r_i, r_j))/kT] - 1] \right\} \end{aligned} \quad (\text{A2.4.90})$$

-38-

with the boundary conditions

equation A2.4.89 and equation A2.4.90 are the most general equations governing the behaviour of an electrolyte near an electrode, and solving them would, in principle, give a combined DH ionic atmosphere and a description of the ionic distribution around each electrode.

The zeroth-order solution to the above equations is the Gouy-Chapman theory dating from the early part of the 20th century [20]. In this solution, the ionic atmosphere is ignored, as is the mirror image potential for the ion. Equation A2.4.90 can therefore be ignored and equation A2.4.89 reduces to

$$\frac{d^2 \phi}{dx^2} = -\frac{\sum_\alpha z_\alpha e_0 n_\alpha^0 \exp\{-\beta[z_\alpha e_0 \phi(x)]\}}{\epsilon \epsilon_0} \quad (\text{A2.4.91})$$

where we have built in the further assumption that $w^s(x) = 0$ for $x > h$ and $w^s(x) = \infty$ for $x < h$. This corresponds to the hard-sphere model introduced above. Whilst A2.4.91 can be solved for general electrolyte solutions, a solution in closed form can be most easily obtained for a 1-1 electrolyte with ionic charges $\pm z$. Under these circumstances, (A2.4.91) reduces to

$$\frac{d^2 \phi}{dx^2} = \frac{2ze_0 n^0}{\epsilon \epsilon_0} \sinh\left(\frac{ze_0}{kT} \phi\right) \quad (\text{A2.4.92})$$

where we have assumed that $n_+^0 = n_-^0 = n^0$. Integration under the boundary conditions above gives:

$$Q_e = (8kT \epsilon \epsilon_0 n^0)^{1/2} \sinh\left(\frac{ze_0 \phi(h)}{2kT}\right) \quad (\text{A2.4.93})$$

$$\phi(h) = \frac{2kT}{ze_0} \sinh^{-1}\left(\frac{Q_e}{(8kT \epsilon \epsilon_0 n^0)^{1/2}}\right) \quad (\text{A2.4.94})$$

$$\phi(x) = \frac{4kT}{ze_0} \tanh^{-1}\left\{e^{-\kappa(x-h)} \tanh\left(\frac{ze_0 \phi(h)}{4kT}\right)\right\} \quad (\text{A2.4.95})$$

and κ has the same meaning as in the DH theory, $\kappa^2 = 2z^2 e_0^2 n^0 / \epsilon \epsilon_0 kT$. These are the central results of the

Gouy-Chapman theory. Clearly, if $ze_0 \phi(h)/4kT$ is small then $\phi(x) \sim \phi(h) e^{-\kappa(x-h)}$ and the potential decays exponentially in the bulk of the electrolyte. The basic physics is similar to the DH analysis, in that the actual field due to the electrode becomes screened by the ionic charges in the electrolyte.

A better approximation may be obtained by expansion of ϕ and ψ_α in powers of the dimensionless variable $\bar{q} = (ze_0^2/\epsilon\epsilon_0kT)\kappa$. If $\phi \approx \phi_0 + \bar{q}\phi_1$ and $\psi \approx \bar{q}\psi_1$, then it is possible to show that

-39-

$$\nabla_j^2 \psi_1(\mathbf{r}_i, \mathbf{r}_j) - \kappa^2 \psi_1(\mathbf{r}_i, \mathbf{r}_j) = -\frac{ze_0}{\epsilon\epsilon_0} \delta(\mathbf{R}_{ij}) \quad (\text{A2.4.96})$$

and

$$\frac{d^2 \phi_1}{dx^2} - \kappa^2 \phi_1 = -\frac{e_0}{kT} \phi_0 \delta\phi \quad (\text{A2.4.97})$$

where $\delta\phi = kT e^{-\kappa z}/4e_0\kappa z$ is the screened image potential. Solutions to equations (A2.4.96 and A2.4.97) have been obtained, and it is found that, for a given Q_e , the value of $\phi(h)$ is always smaller than that predicted by the Gouy-Chapman theory. This theory, both in the zeroth-order and first-order analyses given here, has been the subject of considerable analytical investigation [15], but there has been relatively little progress in devising more accurate theories, since the approximations made even in these simple derivations are very difficult to correct accurately.

A2.4.5.3 SPECIFIC IONIC ADSORPTION AND THE INNER LAYER

Interaction of the water molecules with the electrode surface can be developed through simple statistical models. Clearly for water molecules close to the electrode surface, there will be several opposing effects: the hydrogen bonding, tending to align the water molecules with those in solution; the electric field, tending to align the water molecules with their dipole moments perpendicular to the electrode surface; and dipole-dipole interactions, tending to orient the nearest-neighbour dipoles in opposite directions. Simple estimates [21] based on 20 kJ mol^{-1} for each hydrogen bond suggest that the orientation energy pE becomes comparable to this for $E \sim 5 \times 10^9 \text{ V m}^{-1}$; such field strengths will be associated with surface charges of the order of $0.2\text{-}0.3 \text{ C m}^{-2}$ or $20\text{-}30 \text{ }\mu\text{C cm}^{-2}$ assuming $p = 6.17 \times 10^{-30} \text{ C m}$ and the dielectric function for water at the electrode surface of about six. This corresponds to all molecules being strongly oriented. These are comparable to the fields expected at reasonably high electrode potentials. Similarly, the energy of interaction of two dipoles lying antiparallel to each other is $-p^2/(4\pi\epsilon_0 E^3)$; for $R \sim 4 \text{ \AA}$ the orientational field needs to be in excess of 10^9 V m^{-1} , a comparable number.

The simplest model for water at the electrode surface has just two possible orientations of the water molecules at the surface, and was initially described by Watts-Tobin [22]. The associated potential drop is given by

$$g(\text{dip}) = -\frac{(N_+ - N_-)p}{\epsilon\epsilon_0} \quad (\text{A2.4.98})$$

and if the total potential drop across the inner region of dimension h_i is $\Delta\phi$:

$$N_+/N_- = \exp[-(U_0 - 2p\Delta\phi/h_i)/kT] \quad (\text{A2.4.99})$$

where U_0 is the energy of interaction between neighbouring dipoles. A somewhat more sophisticated model is to assume that water is present in the form of both monomers and dimers, with the dimers so oriented as not to give any nett contribution to the value of $g(\text{dip})$.

A further refinement has come with the work of Parsons [23], building on an analysis by Damaskhin and Frumkin [24]. Parsons suggested that the solvent molecules at the interface could be thought of as being either free or associated as clusters. In a second case the nett dipole moment would be reduced from the value found for perpendicular alignment since the clusters would impose their own alignment. The difficulty with such models is that the structure of the clusters themselves is likely to be a function of the electric field, and simulation methods show, see below, that this is indeed the case.

The experimental data and arguments by Trassatti [25] show that at the PZC, the water dipole contribution to the potential drop across the interface is relatively small, varying from about 0 V for Au to about 0.2 V for In and Cd. For transition metals, values as high as 0.4 V are suggested. The basic idea of water clusters on the electrode surface dissociating as the electric field is increased has also been supported by *in situ* Fourier transform infrared (FTIR) studies [26], and this model also underlies more recent statistical mechanical studies [27].

The model of the inner layer suggests that the interaction energy of water molecules with the metal will be at a minimum somewhere close to the PZC, a result strongly supported by the fact that adsorption of less polar organic molecules often shows a maximum at this same point [18]. However, particularly at anodic potentials, there is now strong evidence that simple anions may lose part of their hydration or solvation sheath and migrate from the OHP to the IHP. There is also evidence that some larger cations, such as $[R_4N]^+$, Tl^+ and Cs^+ also undergo specific adsorption at sufficiently negative potentials. The evidence for specific adsorption comes not only from classical experiments in which the surface tension of mercury is studied as a function of the potential (electrocapillarity), and the coverage derived from rather indirect reasoning [28], but also more direct methods, such as the measurement of the amount of material removed from solution, using radioactive tracers and ellipsometry. A critical problem is much of this work, particularly in those data derived from electrocapillarity, is that the validity of the Gouy-Chapman model must be assumed, an assumption that has been queried. The calculation of the free energy change associated with this process is not simple, and the following effects need to be considered.

- (1) The energy gained on moving from the OHP to the IHP. The electrostatic part of this will have the form $(z_k e_0 Q_e / \epsilon \epsilon_0)(x_{OHP} - x_{IHP})$, but the de-solvation part is much more difficult to estimate.
- (2) The fact that more than one molecule of water may be displaced for each anion adsorbed, and that the adsorption energy of these water molecules will show a complex dependence on the electrode potential.
- (3) The fact that a chemical bond may form between a metal and an anion, leading to, at least, a partial discharge of the ion.
- (4) The necessity to calculate the electrostatic contribution to both the ion-electrode attraction and the ion-ion repulsion energies, bearing in mind that there are at least two dielectric function discontinuities in the simple double-layer model above.
- (5) That short-range contributions to both the ion-ion and ion-electrode interactions must be included.

These calculations have, as their aim, the generation of an *adsorption isotherm*, relating the concentration of ions in the solution to the coverage in the IHP and the potential (or more usually the *charge*) on the electrode. No complete calculations have been carried out incorporating all the above terms. In general, the analytical form for the isotherm is

$$\ln(f(\theta)) = \text{constant} + \ln a_{\pm} + A Q_c + g(\theta) \quad (\text{A2.4.100})$$

where $f(\theta)$ and $g(\theta)$ are functions of the coverage. For models where lateral interactions are dominant, $g(\theta)$

will have a $\theta^{1/2}$ dependence: if multiple electrostatic imaging is important, a term linear in θ will be found. Whereas, if dispersion interactions between ions on the surface are important, then a term in θ^3 becomes significant. The form of $f(\theta)$ is normally taken as $\theta/(1 - \theta)^p$ where p is the number of water molecules displaced by one adsorbed ion. Details of the various isotherms are given elsewhere [28], but modern simulation methods, as reviewed below, are needed to make further progress.

A2.4.5.4 SIMULATION TECHNIQUES

The theoretical complexity of the models discussed above and the relative difficulty of establishing unequivocal models from the experimental data available has led to an increasing interest in Monte Carlo and, particularly, molecular dynamics approaches. Such studies have proved extremely valuable in establishing quite independent models of the interface against which the theories described above can be tested. In particular, these simulation techniques allow a more realistic explicit treatment of the solvent and ions on an equal footing. Typically, the solvent is treated within a rigid multipole model, in which the electrical distribution is modelled by a rigid distribution of charges on the various atoms in the solvent molecule. Dispersion and short-range interactions are modelled using Lennard-Jones or similar model potentials, and the interaction of water with the metal surface is generally modelled with a corrugated potential term to take account of the atomic structure of the metal. Such potentials are particularly marked for metal-oxygen interactions. In the absence of an electrical charge on the electrode the Pt–O interaction energy is usually given by an expression of the form

$$U_{\text{Pt-O}} = [A e^{-\alpha r} - B e^{-\beta r}] f(x, y) + C e^{-\gamma r} [1 - f(x, y)] \quad (\text{A2.4.101})$$

where $f(x, y) = e^{-\lambda(x^2+y^2)}$ and the Pt–H interaction is weakly repulsive, of the form

$$U_{\text{Pt-H}} = D e^{-\mu r}. \quad (\text{A2.4.102})$$

This potential will lead to a single water molecule adsorbing at the PZC on Pt with the dipole pointing *away* from the surface and the oxygen atom pointing directly at a Pt-atom site (on-top configuration).

The main difficulty in these simulations is the long-range nature of the Coulomb interactions, since both mirror-plane images and real charges must be included, and the finite nature of the simulated volume must also be included. A more detailed discussion is given by Benjamin [29], and the following conclusions have been reached.

- (1) Only at extremely high electric fields are the water molecules fully aligned at the electrode surface. For electric fields of the size normally encountered, a distribution of dipole directions is found, whose half-width is strongly dependent on whether specific adsorption of ions takes place. In the absence of such adsorption the distribution function steadily narrows, but in the presence of adsorption the distribution may show little change from that found at the PZC; an example is shown in [figure A2.4.10](#) [30].
- (2) The pair correlation functions g_{OO} , g_{OH} and g_{HH} have been obtained for water on an uncharged electrode surface. For Pt(100), the results are shown in [figure A2.4.11](#) [29], and compared to the correlation functions for the second, much more liquid-like layer. It is clear that the first solvation peak is enhanced by comparison to the liquid, but is in the same position and emphasizing the importance of hydrogen bonding in determining nearest

O–O distances: however, beyond the first peak there are new peaks in the pair correlation function for the water layer immediately adjacent to the electrode that are absent in the liquid, and result from the periodicity of the Pt surface. By contrast, these peaks have disappeared in the second layer, which is

very similar to normal liquid water.

- (3) Simulation results for turning on the electric field at the interface in a system consisting of a water layer between two Pt electrodes 3 nm apart show that the dipole density initially increases fairly slowly, but that between 10 and 20 V nm⁻¹ there is an apparent phase transition from a moderately ordered structure, in which the ordering is confined close to the electrodes only, to a substantially ordered layer over the entire 3 nm thickness. Effectively, at this field, which corresponds to the energy of about four hydrogen bonds, the system loses all the ordering imposed by the hydrogen bonds and reverts to a purely linear array of dipoles.
- (4) For higher concentrations of aqueous electrolyte, the simulations suggest that the ionic densities do not change monotonically near the electrode surface, as might be expected from the Gouy-Chapman analysis above, but oscillate in the region $x < 10$ Å. This oscillation is, in part, associated with the oscillation in the oxygen atom density caused by the layering effect occurring in liquids near a surface.
- (5) At finite positive and negative charge densities on the electrode, the counterion density profiles often exhibit significantly higher maxima, i.e. there is an overshoot, and the derived potential actually shows oscillations itself close to the electrode surface at concentrations above about 1 M.
- (6) Whether the potentials are derived from quantum mechanical calculations or classical image forces, it is quite generally found that there is a stronger barrier to the adsorption of cations at the surface than anions, in agreement with that generally .

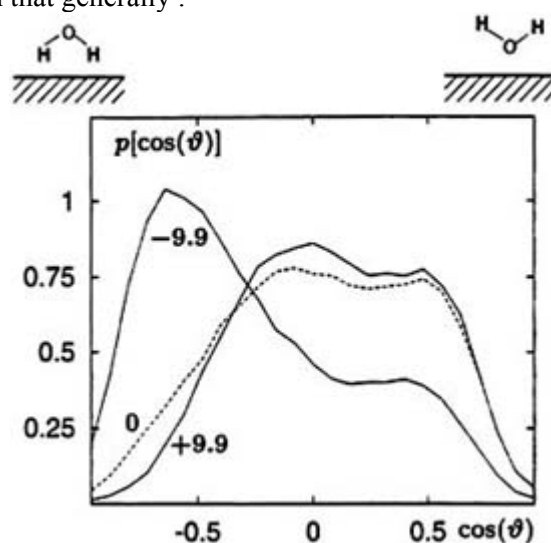


Figure A2.4.10 . Orientational distribution of the water dipole moment in the adsorbate layer for three simulations with different surface charge densities (in units of $\mu\text{C cm}^{-2}$ as indicated). In the figure $\cos \theta$ is the angle between the water dipole vector and the surface normal that points into the aqueous phase. From [30].

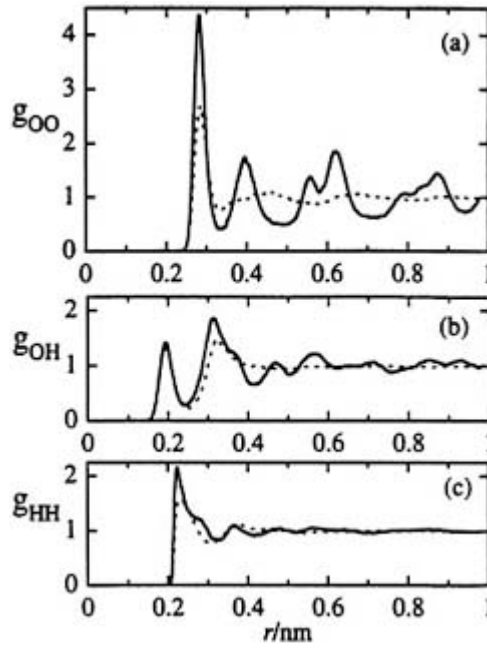


Figure A2.4.11. Water pair correlation functions near the Pt(100) surface. In each panel, the full curve is for water molecules in the first layer, and the broken curve is for water molecules in the second layer. From [30].

A 2.4.6 THERMODYNAMICS OF ELECTRIFIED INTERFACES

If a metal, such as copper, is placed in contact with a solution containing the ions of that metal, such as from aqueous copper sulphate, then we expect an equilibrium to be set up of the following form:



where the subscript M refers to the metal. As indicated above, there will be a potential difference across the interface, between the Galvani potential in the interior of the copper and that in the interior of the electrolyte. The effects of this potential difference must be incorporated into the normal thermodynamic equations describing the interface, which is done, as above, by defining the *electrochemical potential* of a species with charge $z_i e_0$ per ion or $z_i F$ per mole. If one mole of z -valent ions is brought from a remote position to the interior of the solution, in which there exists a potential ϕ , then the work done will be $zF\phi$; this work term must be added to or subtracted from the free energy per mole, μ , depending on the relative signs of the charge and ϕ , and the condition for equilibrium for component i partitioned between two phases with potentials $\phi(\text{I})$ and $\phi(\text{II})$ is

$$\mu_i(\text{I}) + z_i F \phi(\text{I}) = \mu_i(\text{II}) + z_i F \phi(\text{II}) \quad (\text{A2.4.104})$$

-44-

where $\phi(\text{I})$ and $\phi(\text{II})$ are the Galvani or inner potentials in the interior of phases (I) and (II). The expression $\mu_i + z_i F \phi$ is referred to as the electrochemical potential, $\tilde{\mu}_i$. We have

$$\tilde{\mu}_i = \mu_i + z_i F \phi = \mu_i^0 + RT \ln a_i + z_i F \phi \quad (\text{A2.4.105})$$

and the condition for electrochemical equilibrium can be written for our copper system:

$$\tilde{\mu}_{\text{Cu}}(\text{M}) = \tilde{\mu}_{\text{Cu}^{2+}}(\text{S}) + 2\tilde{\mu}_{\text{e}^-}(\text{M}) \quad (\text{A2.4.106})$$

where the labels M and S refer to the metal and to the solution respectively. Assuming the copper atoms in the metal to be neutral, so that $\tilde{\mu}_{\text{Cu}^0} = \mu_{\text{Cu}^0}$, we then have

$$\begin{aligned} \mu_{\text{Cu}^0}^0(\text{M}) + RT \ln(a_{\text{Cu}}(\text{M})) &= \mu_{\text{Cu}^{2+}}^0(\text{S}) + RT \ln(a_{\text{Cu}^{2+}}) + 2F\phi_{\text{S}} + \mu_{\text{e}^-}^0(\text{M}) \\ &+ 2RT \ln(a_{\text{e}^-}) - 2F\phi_{\text{M}} \end{aligned} \quad (\text{A2.4.107})$$

Given that the concentration of both the copper atoms and the electrons in the copper metal will be effectively constant, so that two of the activity terms can be neglected, we finally have, on rearranging A2.4.107,

$$\begin{aligned} \Delta\phi &\equiv \phi_{\text{M}} - \phi_{\text{S}} = \frac{\mu_{\text{Cu}^{2+}}^0(\text{S}) + \mu_{\text{e}^-}^0(\text{M}) - \mu_{\text{Cu}^0}^0(\text{M})}{2F} + \frac{RT}{2F} \ln(a_{\text{Cu}^{2+}}) \\ &\equiv \Delta\phi_0 + \left(\frac{RT}{2F}\right) \ln(a_{\text{Cu}^{2+}}) \end{aligned} \quad (\text{A2.4.108})$$

where $\Delta\phi_0$ is the Galvani potential difference at equilibrium between the electrode and the solution in the case where $a_{\text{Cu}^{2+}}(\text{aq}) = 1$, and is referred to as the standard Galvani potential difference. It can be seen, in general, that the Galvani potential difference will alter by a factor of $(RT/zF) \ln 10 \equiv 0.059/z \text{ V}$ at 298 K, for every order of magnitude change in activity of the metal ion, where z is the valence of the metal ion in solution.

A2.4.6.1 THE NERNST EQUATION FOR REDOX ELECTRODES

In addition to the case of a metal in contact with its ions in solution there are other cases in which a Galvani potential difference between two phases may be found. One case is the immersion of an inert electrode, such as platinum metal, into an electrolyte solution containing a substance 'S' that can exist in either an oxidized or reduced form through the loss or gain of electrons from the electrode. In the simplest case, we have



-45-

an example being



where the physical process described is the exchange of *electrons* (not ions) between the electrolyte and the electrode: at no point is the electron conceived as being free in the solution. The equilibrium properties of the redox reaction [A2.4.109](#) can, in principle, be treated in the same way as above. At equilibrium, once a double layer has formed and a Galvani potential difference set up, we can write

$$\tilde{\mu}_{\text{S}_{\text{ox}}} + n\tilde{\mu}_{\text{e}^-}^{\text{M}} = \mu_{\text{S}_{\text{red}}} \quad (\text{A2.4.111})$$

and, bearing in mind that the positive charge on 'ox' must exceed 'red' by $|ne^-|$ if we are to have

electroneutrality, then A2.4.111 becomes

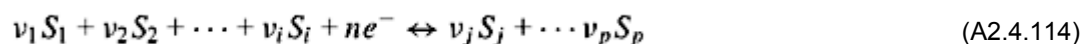
$$\mu_{S_{\text{ox}}}^0 + RT \ln(a_{S_{\text{ox}}}) + nF\phi_S + n\mu_{e^-}^0 - nF\phi_M = \mu_{S_{\text{red}}}^0 + RT \ln(a_{S_{\text{red}}}) \quad (\text{A2.4.112})$$

whence

$$\Delta\phi = \phi_M - \phi_S = \frac{\mu_{S_{\text{ox}}}^0 + n\mu_{e^-}^0 - \mu_{S_{\text{red}}}^0}{nF} + RT \ln\left(\frac{a_{S_{\text{ox}}}}{a_{S_{\text{red}}}}\right) \equiv \Delta\phi^0 + RT \ln\left(\frac{a_{S_{\text{ox}}}}{a_{S_{\text{red}}}}\right) \quad (\text{A2.4.113})$$

where the standard Galvani potential difference is now defined as that for which the activities of S_{ox} and S_{red} are equal. As can be seen, an alteration of this ratio by a factor of ten leads to a change of $0.059/n$ V in $\Delta\phi$ at equilibrium. It can also be seen that $\Delta\phi$ will be independent of the magnitudes of the activities provided that their ratio is a constant.

For more complicated redox reactions, a general form of the Nernst equation may be derived by analogy with A2.4.113. If we consider a stoichiometric reaction of the following type:



which can be written in the abbreviated form



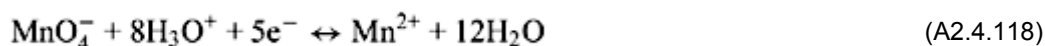
then straightforward manipulation leads to the generalized Nernst equation:

$$\Delta\phi = \Delta\phi^0 + \frac{RT}{nF} \ln\left(\frac{\prod_{\text{ox}} a_{\text{ox}}^{v_{\text{ox}}}}{\prod_{\text{red}} a_{\text{red}}^{v_{\text{red}}}}\right) \quad (\text{A2.4.116})$$

where the notation

$$\prod_{\text{ox}} a_{\text{ox}}^{v_{\text{ox}}} = a_{S_{\text{ox}_1}}^{v_{\text{ox}_1}} a_{S_{\text{ox}_2}}^{v_{\text{ox}_2}} \dots a_{S_{\text{ox}_i}}^{v_{\text{ox}_i}} \quad (\text{A2.4.117})$$

As an example, the reduction of permanganate in acid solution follows the equation



and the potential of a platinum electrode immersed in a solution containing both permanganate and Mn^{2+} is given by

$$\Delta\phi = \Delta\phi^0 + \frac{RT}{nF} \ln\left(\frac{a_{\text{MnO}_4^-} a_{\text{H}_3\text{O}^+}^8}{a_{\text{Mn}^{2+}}}\right) \quad (\text{A2.4.119})$$

assuming that the activity of neutral H₂O can be put equal to unity.

A2.4.6.2 THE NERNST EQUATION FOR GAS ELECTRODES

The Nernst equation above for the dependence of the equilibrium potential of redox electrodes on the activity of solution species is also valid for uncharged species in the gas phase that take part in electron exchange reactions at the electrode-electrolyte interface. For the specific equilibrium process involved in the reduction of chlorine:



the corresponding Nernst equation can easily be shown to be

$$\Delta\phi = \Delta\phi^0 + \frac{RT}{2F} \ln \left(\frac{a_{\text{Cl}_2}(\text{aq})}{a_{\text{Cl}^-}^2} \right) \quad (\text{A2.4.121})$$

where $a_{\text{Cl}_2}(\text{aq})$ is the activity of the chlorine gas dissolved in water. If the Cl₂ solution is in equilibrium with chlorine at pressure p_{Cl_2} in the gas phase, then

$$\mu_{\text{Cl}_2}(\text{gas}) = \mu_{\text{Cl}_2}(\text{aq}). \quad (\text{A2.4.122})$$

Given that $\mu_{\text{Cl}_2}(\text{gas}) = \mu_{\text{Cl}_2}^0(\text{gas}) + RT \ln(p_{\text{Cl}_2}/p^0)$ and $\mu_{\text{Cl}_2}(\text{aq}) = \mu_{\text{Cl}_2}^0(\text{aq}) + RT \ln(a_{\text{Cl}_2}(\text{aq}))$, where p^0 is the standard pressure of 1 atm ($\equiv 101\,325$ Pa), then it is clear that

-47-

$$a_{\text{Cl}_2}(\text{aq}) = \left(\frac{p_{\text{Cl}_2}}{p^0} \right) \exp \left(\frac{\mu_{\text{Cl}_2}^0(\text{gas}) - \mu_{\text{Cl}_2}^0(\text{aq})}{RT} \right) \quad (\text{A2.4.123})$$

and we can write

$$\Delta\phi = \Delta\phi^{0'} + \left(\frac{RT}{2F} \right) \ln \left(\frac{p_{\text{Cl}_2}}{p^0 a_{\text{Cl}^-}^2} \right) \quad (\text{A2.4.124})$$

where $\Delta\phi^{0'}$ is the Galvani potential difference under the standard conditions of $p_{\text{Cl}_2} = p^0$ and $a_{\text{Cl}^-} = 1$.

A2.4.6.3 THE MEASUREMENT OF ELECTRODE POTENTIALS AND CELL VOLTAGES

Although the results quoted above are given in terms of the Galvani potential difference between a metal electrode and an electrolyte solution, direct measurement of this Galvani potential difference between an electrode and an electrolyte is not possible, since any voltmeter or similar device will incorporate unknowable surface potentials into the measurement. In particular, any contact of a measurement probe with the solution phase will have to involve a second phase boundary between the metal and the electrolyte somewhere; at this boundary an electrochemical equilibrium will be set up and with it a second equilibrium Galvani potential difference, and the overall potential difference measured by this instrument will in fact be the difference of two Galvani voltages at the two interfaces. In other words, even at zero current, the actual EMF measured for a galvanic cell will be the difference between the two Galvani voltages $\Delta\phi(\text{I})$ and $\Delta\phi(\text{II})$ for the two interfaces, as shown in [figure A2.4.12](#) [7].

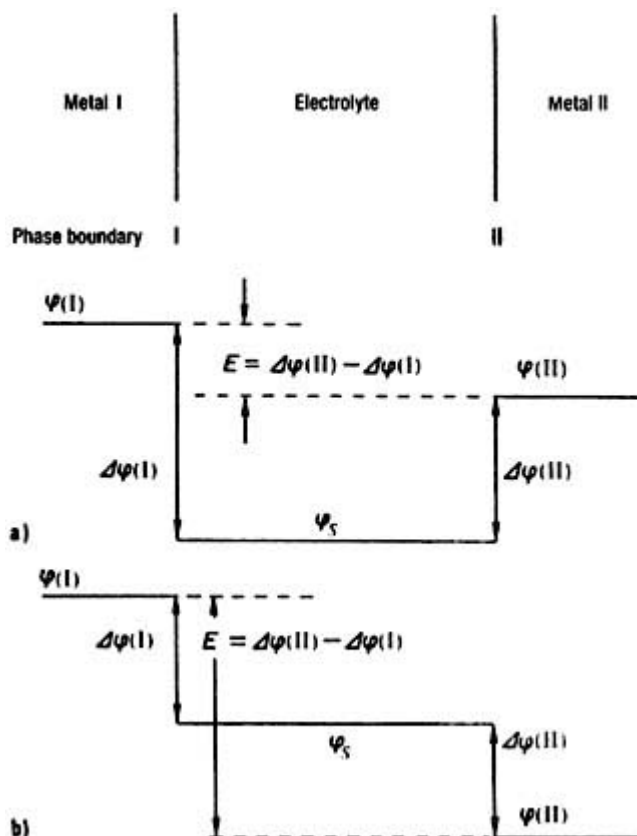


Figure A2.4.12. The EMF of a galvanic cell as the difference between the equilibrium Galvani potentials at the two electrodes: (a) $\Delta\phi(I) > 0$, $\Delta\phi(II) > 0$ and (b) $\Delta\phi(I) > 0$, $\Delta\phi(II) < 0$. From [7].

Figure A2.4.12 shows the two possibilities that can exist, in which the Galvani potential of the solution, ϕ_s , lies between $\phi(I)$ and $\phi(II)$ and in which it lies below (or, equivalently, above) the Galvani potentials of the metals. It should be emphasized that figure A2.4.12 is highly schematic: in reality the potential near the phase boundary in the solution changes initially linearly and then exponentially with distance away from the electrode surface, as we saw above. The other point is that we have assumed that ϕ_s is a constant in the region between the two electrodes. This will only be true provided the two electrodes are immersed in the same solution and that no current is passing.

It is clear from figure A2.4.12 that the EMF or potential difference, E , between the two metals is given by

$$E = \Delta\phi(II) - \Delta\phi(I) = \phi(II) - \phi(I) \tag{A2.4.125}$$

where we adopt the normal electrochemical convention that the EMF is always equal to the potential on the metal on the right of the figure minus the potential of the metal on the left. It follows that once the Galvani potential of any one electrode is known it should be possible, at least in principle, to determine the potentials for all other electrodes.

In practice, since the Galvani potential of no single electrode is known the method adopted is to arbitrarily

choose one *reference* electrode and assign a value for its Galvani potential. The choice actually made is that of the hydrogen electrode, in which hydrogen gas at one atmosphere pressure is bubbled over a platinized platinum electrode immersed in a solution of unit H_3O^+ activity. From the discussion in [section A2.4.6.2](#), it will be clear that provided an equilibrium can be established rapidly for such an electrode, its Galvani potential difference will be a constant, and changes in the measured EMF of the complete cell as conditions are altered at the other electrode will actually reflect the changes in the Galvani potential difference of that electrode.

Cells need not necessarily contain a reference electrode to obtain meaningful results; as an example, if the two electrodes in [figure A2.4.12](#) are made from the same metal, M, but these are now in contact with two solutions of the same metal ions, M^{z+} but with differing ionic activities, which are separated from each other by a glass frit that permits contact, but impedes diffusion, then the EMF of such a cell, termed a concentration cell, is given by

$$E = \Delta\phi(\text{II}) - \Delta\phi(\text{I}) = \frac{RT}{zF} \ln \left(\frac{a_{\text{M}^{z+}}(\text{II})}{a_{\text{M}^{z+}}(\text{I})} \right). \quad (\text{A2.4.126})$$

Equation A2.4.126 shows that the EMF increases by $0.059/z$ V for each decade change in the activity ratio in the two solutions.

A2.4.6.4 CONVENTIONS IN THE DESCRIPTION OF CELLS

In order to describe any electrochemical cell a convention is required for writing down the cells, such as the concentration cell described above. This convention should establish clearly where the boundaries between the different phases exist and, also, what the overall cell reaction is. It is now standard to use vertical lines to delineate phase boundaries, such as those between a solid and a liquid or between two immiscible liquids. The junction between two miscible liquids, which might be maintained by the use of a porous glass frit, is represented by a single vertical dashed line, || , and two dashed lines, ||| , are used to indicate two liquid phases joined by an appropriate electrolyte bridge adjusted to minimize potentials arising from the different diffusion coefficients of the anion and cation (so-called 'junction potentials').

The cell is written such that the cathode is to the right when the cell is acting in the galvanic mode, and electrical energy is being generated from the electrochemical reactions at the two electrodes. From the point of view of external connections, the cathode will appear to be the positive terminal, since electrons will travel in the external circuit from the anode, where they pass from electrolyte to metal, to the cathode, where they pass back into the electrolyte. The EMF of such a cell will then be the difference in the Galvani potentials of the metal electrodes on the right-hand side and the left-hand side. Thus, the concentration cell of [section A2.4.6.3](#) would be represented by $\text{M}|\text{M}^{z+}(\text{I})|\text{M}^{z+}(\text{II})|\text{M}$.

In fact, some care is needed with regard to this type of concentration cell, since the assumption implicit in the derivation of A2.4.126 that the potential in the solution is constant between the two electrodes, cannot be entirely correct. At the phase boundary between the two solutions, which is here a semi-permeable membrane permitting the passage of water molecules but not ions between the two solutions, there will be a potential jump. This so-called liquid-junction potential will increase or decrease the measured EMF of the cell depending on its sign. Potential jumps at liquid-liquid junctions are in general rather small compared to normal cell voltages, and can be minimized further by suitable experimental modifications to the cell.

If two redox electrodes both use an inert electrode material such as platinum, the cell EMF can be written down immediately. Thus, for the hydrogen/chlorine fuel cell, which we represent by the cell $\text{H}_2(\text{g})|\text{Pt}|\text{HCl}(\text{m})|\text{Pt}|\text{Cl}_2(\text{g})$ and for which it is clear that the cathodic reaction is the reduction of Cl_2 as considered in [section](#)

A2.4.6.2:

$$\begin{aligned} E &= \Delta\phi(\text{Cl}_2/\text{Cl}^-) - \Delta\phi(\text{H}_2/\text{H}_3\text{O}^+(\text{aq})) \\ &= E^0 - \left(\frac{RT}{2F}\right) \ln(a_{\text{H}_3\text{O}^+}^2 a_{\text{Cl}^-}^2) + \left(\frac{RT}{2F}\right) \ln\{(p_{\text{H}_2}/p^0)/(p_{\text{Cl}_2}/p^0)\} \end{aligned} \quad (\text{A2.4.127})$$

where E^0 is the standard EMF of the fuel cell, or the EMF at which the activities of H_3O^+ and Cl^- are unity and the pressures of H_2 and Cl_2 are both equal to the standard pressure, p^0 .

A 2.4.7 ELECTRICAL POTENTIALS AND ELECTRICAL CURRENT

The discussion in earlier sections has focussed, by and large, on understanding the equilibrium structures in solution and at the electrode-electrolyte interface. In this last section, some introductory discussion will be given of the situation in which we depart from equilibrium by permitting the flow of electrical current through the cell. Such current flow leads not only to a potential drop across the electrolyte, which affects the cell voltage by virtue of an ohmic drop $I R_i$ (where R_i is the internal resistance of the electrolyte between the electrodes), but each electrode exhibits a characteristic current-voltage behaviour, and the overall cell voltage will, in general, reflect both these effects.

A2.4.7.1 THE CONCEPT OF OVERPOTENTIAL

Once current passes through the interface, the Galvani potential difference will differ from that expected from the Nernst equation above; the magnitude of the difference is termed the *overpotential*, which is defined heuristically as

$$\eta = \Delta\phi - \Delta\phi_r = E - E_r \quad (\text{A2.4.128})$$

where the subscript r refers to the 'rest' situation, i.e. to the potential measured in the absence of any current passing. Provided equilibrium can be established, this rest potential will correspond to that predicted by the Nernst equation. Obviously, the sign of η is determined by whether E is greater than or less than E_r .

At low currents, the rate of change of the electrode potential with current is associated with the limiting rate of electron transfer across the phase boundary between the electronically conducting electrode and the ionically conducting solution, and is termed the electron transfer overpotential. The electron transfer rate at a given overpotential has been found to depend on the nature of the species participating in the reaction, and the properties of the electrolyte and the electrode itself (such as, for example, the chemical nature of the metal). At higher current densities, the primary electron transfer rate is usually no longer limiting; instead, limitations arise through the slow transport of reactants from the solution to the electrode surface or, conversely, the slow transport of the product away from the electrode (diffusion overpotential) or through the inability of chemical reactions coupled to the electron transfer step to keep pace (reaction overpotential).

Examples of the latter include the adsorption or desorption of species participating in the reaction or the participation of chemical reactions before or after the electron transfer step itself. One such process occurs in the evolution of hydrogen from a solution of a weak acid, HA: in this case, the electron transfer from the electrode to the proton in solution must be preceded by the acid dissociation reaction taking place in solution.

A2.4.7.2 THE THEORY OF ELECTRON TRANSFER

The rate of simple chemical reactions can now be calculated with some confidence either within the framework of activated-complex theory or directly from quantum mechanical first principles, and theories that might lead to analogous predictions for simple electron transfer reactions at the electrode-electrolyte interface have been the subject of much recent investigation. Such theories have hitherto been concerned primarily with greatly simplified models for the interaction of an ion in solution with an inert electrode surface. The specific adsorption of electroactive species has been excluded and electron transfer is envisaged only as taking place between the strongly solvated ion in the outer Helmholtz layer and the metal electrode. The electron transfer process itself can only be understood through the formalism of quantum mechanics, since the transfer itself is a tunnelling phenomenon that has no simple analogue in classical mechanics.

Within this framework, by considering the physical situation of the electrode double layer, the free energy of activation of an electron transfer reaction can be identified with the reorganization energy of the solvation sheath around the ion. This idea will be carried through in detail for the simple case of the strongly solvated $\text{Fe}^{3+}/\text{Fe}^{2+}$ couple, following the change in the ligand-ion distance as the critical reaction variable during the transfer process.

In aqueous solution, the oxidation of Fe^{2+} can be conceived as a reaction of two aquo-complexes of the form



The H_2O molecules of these aquo-complexes constitute the inner solvation shell of the ions, which are, in turn, surrounded by an external solvation shell of more or less uncoordinated water molecules forming part of the water continuum, as described in [section A2.4.2](#) above. Owing to the difference in the solvation energies, the radius of the Fe^{3+} aquo-complex is smaller than that of Fe^{2+} , which implies that the mean distance of the vibrating water molecules at their normal equilibrium point must change during the electron transfer. Similarly, changes must take place in the outer solvation shell during electron transfer, all of which implies that the solvation shells themselves inhibit electron transfer. This inhibition by the surrounding solvent molecules in the inner and outer solvation shells can be characterized by an activation free energy ΔG^\ddagger .

Given that the tunnelling process itself requires no activation energy, and that tunnelling will take place at some particular configuration of solvent molecules around the ion, the entire activation energy referred to above must be associated with ligand/solvent movement. Furthermore, from the Franck-Condon principle, the electron tunnelling process will take place on a rapid time scale compared to nuclear motion, so that the ligand and solvent molecules will be essentially stationary during the actual process of electron transfer.

Consider now the aquo-complexes above, and let x be the distance of the centre of mass of the water molecules constituting the inner solvation shell from the central ion. The binding interaction of these molecules leads to vibrations

-52-

of frequency $f = \omega/2\pi$ taking place about an equilibrium point x_0 and, if the harmonic approximation is valid, the potential energy change U_{pot} associated with the ligand vibration can be written in parabolic form as

(A2.4.130)

where M is the mass of the ligands, B is the binding energy of the ligands and U_{el} is the electrical energy of the ion-electrode system. The total energy of the system will also contain the kinetic energy of the ligands, written in the form $p^2/2M$, where p is the momentum of the molecules during vibrations:

It is possible to write two such equations for the initial state, i , (corresponding to the reduced aquo-complex $[\text{Fe}(\text{H}_2\text{O})_6]^{2+}$) and the final state, f , corresponding to the oxidized aquo-complex and the electron now present in the electrode. Clearly

with a corresponding equation for state f , and with the assumption that the frequency of vibration does not alter between the initial and final states of the aquo-complex. During electron transfer, the system moves, as shown in figure A2.4.13 [7], from an equilibrium situation centred at x_0 along the parabolic curve labelled to the point x_s where electron transfer takes place; following this, the system will move along the curve labelled to the new equilibrium situation centred on .

Figure A2.4.13. Potential energy of a redox system as a function of ligand–metal separation. From [7].

-53-

The point at which electron transfer takes place clearly corresponds to the condition $u_{\text{el}}^i = u_{\text{el}}^f$; equating equations (A2.4.132) for the states i and f we find that

$$x_s = \frac{B^f + U_{\text{el}}^f - B^i - U_{\text{el}}^i + (M\omega^2/2)([x_f^0]^2 - [x_i^0]^2)}{M\omega^2(x_0^f - x_0^i)}. \quad (\text{A2.4.133})$$

The activation energy, U_{act} , is defined as the minimum additional energy above the zero-point energy that is needed for a system to pass from the initial to the final state in a chemical reaction. In terms of [equation \(A2.4.132\)](#), the energy of the initial reactants at $x = x_s$ is given by

$$U^i = \frac{p^2}{2M} + \frac{1}{2}M\omega^2(x_s - x_0^i)^2 + B^i + U_{\text{el}}^i \quad (\text{A2.4.134})$$

where $B^i + U_{\text{el}}^i$ is the zero-point energy of the initial state. The minimum energy required to reach the point x_s is clearly that corresponding to the momentum $p = 0$. By substituting for x_s from equation (A2.4.133), we find

$$U_{\text{act}} = \frac{M\omega^2}{2}(x_s - x_0^i)^2 = \frac{(U_s + U_{\text{el}}^f - U_{\text{el}}^i + B^f - B^i)^2}{4U_s}. \quad (\text{A2.4.135})$$

where U_s has the value $(M\omega^2/2)(x_0^f - x_0^i)^2$. U_s is termed the *reorganization energy* since it is the additional energy required to deform the complex from initial to final value of x . It is common to find the symbol λ for U_s , and model calculations suggest that U_s normally has values in the neighbourhood of 1 eV (10^5 J mol^{-1}) for the simplest redox processes.

In our simple model, the expression in A2.4.135 corresponds to the activation energy for a redox process in which only the interaction between the central ion and the ligands in the primary solvation shell is considered, and this only in the form of the totally symmetrical vibration. In reality, the rate of the electron transfer reaction is also influenced by the motion of molecules in the outer solvation shell, as well as by other

vibrational modes of the inner shell. These can be incorporated into the model provided that each type of motion can be treated within the simple harmonic approximation. The total energy of the system will then consist of the kinetic energy of all the atoms in motion together with the potential energy arising from each vibrational degree of freedom. It is no longer possible to picture the motion, as in [figure A2.4.13](#) as a one-dimensional translation over an energy barrier, since the total energy is a function of a large number of normal coordinates describing the motion of the entire system. Instead, we have two potential energy surfaces for the initial and final states of the redox system, whose intersection described the reaction hypersurface. The reaction pathway will proceed now *via* the saddle point, which is the minimum of the total potential energy subject to the condition $v_{\mu}^i = v_{\mu}^f$ as above.

This is a standard problem [31] and essentially the same result is found as in equation (A2.4.135), save that the B^i and B^f now become the sum over all the binding energies of the central ion in the initial and final states and U_s is now given by

-54-

$$U_s = \sum_j \frac{M_j \omega_j^2}{2} (x_{j,0}^f - x_{j,0}^i)^2 \quad (\text{A2.4.136})$$

where M_j is the effective mass of the j th mode and ω_j is the corresponding frequency; we still retain the approximation that these frequencies are all the same for the initial and final states.

With the help of U_s , an expression for the rate constant for the reaction



can be written

$$k_f = k_f^0 \exp\left(-\frac{U_{\text{act}}}{kT}\right) = A \exp\left(-\frac{(U_s + U_{\text{el}}^f - U_{\text{el}}^i + B^f - B^i)^2}{4U_s kT}\right) \quad (\text{A2.4.138})$$

where A is the so-called frequency factor and e_{M}^- refers to an electron in the metal electrode. The rate constant for the back reaction is obtained by interchanging the indices i and f in [equation \(A2.4.74\)](#). It will be observed that under these circumstances U_s remains the same and we obtain

$$k_b = A \exp\left(-\frac{(U_s + U_{\text{el}}^i - U_{\text{el}}^f + B^i - B^f)^2}{4U_s kT}\right) \quad (\text{A2.4.139})$$

A2.4.7.3 THE EXCHANGE CURRENT

It is now possible to derive an expression for the actual current density from A2.4.138 and A2.4.139, assuming reaction A2.4.137, and, for simplicity, assuming that the concentrations, c , of Fe^{2+} and Fe^{3+} are equal. The potential difference between the electrode and the outer Helmholtz layer, $\Delta\phi$, is incorporated into the electronic energy of the $\text{Fe}^{3+} + e_{\text{M}}^-$ system through a potential-dependent term of the form

$$U_{\text{el}}^f = U_{\text{el},0}^f - e_0 \Delta\phi \quad (\text{A2.4.140})$$

where the minus sign in A2.4.140 arises through the negative charge on the electron. Inserting this into A2.4.138 and A2.4.139 and multiplying by concentration and the Faraday to convert from rate constants to current densities, we have

-55-

$$j^+ = FAc \exp \left(- \frac{(U_s + U_{cl,0}^f - U_{cl}^i + B^f - B^i - e_0 \Delta \phi)^2}{4U_s kT} \right) \quad (\text{A2.4.141})$$

$$j^- = -FAc \exp \left(- \frac{(U_s + U_{cl}^i - U_{cl,0}^f + B^i - B^f + e_0 \Delta \phi)^2}{4U_s kT} \right) \quad (\text{A2.4.142})$$

where we adopt the convention that positive current involves *oxidation*. At the rest potential, $\Delta \phi_r$, which is actually the same as the standard Nernst potential $\Delta \phi_0$ when assuming that the activity coefficients of the ions are also equal, the rates of these two reactions are equal, which implies that the terms in brackets in the two equations must also be equal when $\Delta \phi = \Delta \phi_0$. From this it is clear that

$$e_0 \Delta \phi_0 = U_{cl,0}^f - U_{cl}^i + B^f - B^i \quad (\text{A2.4.143})$$

and if we introduce the overpotential, $\eta = \Delta \phi - \Delta \phi_0$, evidently

$$j^+ = FAc \exp \left[- \frac{(U_s - e_0 \eta)^2}{4U_s kT} \right] \quad (\text{A2.4.144})$$

$$j^- = -FAc \exp \left[- \frac{(U_s + e_0 \eta)^2}{4U_s kT} \right] \quad (\text{A2.4.145})$$

from which we obtain the exchange current density as the current at $\eta = 0$:

$$j_0 = FAc \exp \left[- \frac{U_s}{4kT} \right] \quad (\text{A2.4.146})$$

and the activation energy of the exchange current density can be seen to be $U_s/4$. If the overpotential is small, such that $e_0 \eta \ll U_s$ (and recalling that U_s lies in the region of about 1 eV), the quadratic form of A2.4.144 and A2.4.145 can be expanded with neglect of terms in η^2 . Recalling, also, that $e_0/k_B = F/R$, we then finally obtain

$$j = j^+ + j^- = FAc \exp \left(- \frac{U_s}{4kT} \right) \left\{ \exp \left(\frac{F\eta}{2RT} \right) - \exp \left(- \frac{F\eta}{2RT} \right) \right\} \quad (\text{A2.4.147})$$

which is the simplest form of the familiar Butler-Volmer equation with a symmetry factor $\beta = \frac{1}{2}$. This result arises from the strongly simplified molecular model that we have used above and, in particular, the assumption that the values of ω_j are the same for all normal modes. Relaxation of this assumption leads to a more general equation:

-56-

$$j = j^+ + j^- = FAc \exp\left(-\frac{U_s}{4kT}\right) \left\{ \exp\left(\frac{\beta F\eta}{RT}\right) - \exp\left(-\frac{(1-\beta)F\eta}{RT}\right) \right\} \quad (\text{A2.4.148})$$

$$= j^0 \left\{ \exp\left(\frac{\beta F\eta}{RT}\right) - \exp\left(-\frac{(1-\beta)F\eta}{RT}\right) \right\}. \quad (\text{A2.4.149})$$

For a more general reaction of the form $\text{Ox} + ne^- \leftrightarrow \text{Red}$, with differing concentrations of Ox and Red, the exchange current density is given by

$$j^0 = nFA(c_{\text{Ox}}^\beta, c_{\text{Red}}^{1-\beta}) \exp\left(-\frac{U_{\text{act}}}{kT}\right). \quad (\text{A2.4.150})$$

Some values for j^0 and β for electrochemical reactions of importance are given in table A2.4.6, and it can be seen that the exchange currents can be extremely dependent on the electrode material, particularly for more complex processes such as hydrogen oxidation. Many modern electrochemical studies are concerned with understanding the origin of these differences in electrode performance.

Table A2.4.6.

System	Electrolyte	Temperature (°C)	Electrode	j_0 (A cm ⁻²)	β
Fe ³⁺ /Fe ²⁺ (0.005 M)	1 M H ₂ SO ₄	25	Pt	2×10^{-3}	0.58
K ₃ Fe(CN) ₆ /K ₄ Fe(CN) ₆ (0.02 M)	0.5 M K ₂ SO ₄	25	Pt	5×10^{-2}	0.49
Ag/10 ⁻³ M Ag ⁺	1 M HClO ₄	25	Ag	1.5×10^{-1}	0.65
Cd/10 ⁻² M Cd ²⁺	0.4 M K ₂ SO ₄	25	Cd	1.5×10^{-3}	0.55
Cd(Hg)/1.4 × 10 ⁻³ M Cd ²⁺	0.5 M Na ₂ SO ₄	25	Cd(Hg)	2.5×10^{-2}	0.8
Zn(Hg)/2 × 10 ⁻² M Zn ²⁺	1 M HClO ₄	0	Zn(Hg)	5.5×10^{-3}	0.75
Ti ⁴⁺ /Ti ³⁺ (10 ⁻³ M)	1 M acetic acid	25	Pt	9×10^{-4}	0.55
H ₂ /OH ⁻	1 M KOH	25	Pt	10 ⁻³	0.5
H ₂ /H ⁺	1 M H ₂ SO ₄	25	Hg	10 ⁻¹²	0.5
H ₂ /H ⁺	1 M H ₂ SO ₄	25	Pt	10 ⁻³	0.5
O ₂ /H ⁺	1 M H ₂ SO ₄	25	Pt	10 ⁻⁶	0.25
O ₂ /OH ⁻	1 M KOH	25	Pt	10 ⁻⁶	0.3

- [1] Ben-Naim A 1992 *Statistical Thermodynamics for Chemists and Biologists* (London: Plenum)
- [2] Lee L L 1988 *Molecular Thermodynamics of Nonideal Fluids* (Boston: Butterworths)
- [3] Kihara T 1953 *Rev. Mod. Phys.* **25** 831
- [4] Metropolis N A, Rosenbluth A W, Rosenbluth M N, Teller A H and Teller E 1953 *J. Chem. Phys.* **21** 1087
- [5] Richens D T 1997 *The Chemistry of Aqua-ions* (Chichester: Wiley)
- [6] Enderby J E 1983 *Ann. Rev. Phys. Chem.* **34** 155
Enderby J E 1983 *Contemp. Phys.* **24** 561
- [7] Hamann C H, Hamnett A and Vielstich W 1998 *Electrochemistry* (Weinheim: Wiley)
- [8] Robinson R A and Stokes R H 1959 *Electrolyte Solutions* (London: Butterworth)
- [9] Outhwaite C W 1975 *Statistical Mechanics* ed K Singer (London: Chemistry Society)
- [10] Blum L 1975 *Mol. Phys.* **30** 1529
- [11] Gurney R W 1954 *Ionic Processes in Solution* (New York: Dover)
- [12] Desnoyers J E and Jolicoeur C 1969 *Modern Aspects of Electrochemistry* vol 5, ed B E Conway and J O'M Bockris (New York: Plenum)
- [13] Friedman H L 1969 *Modern Aspects of Electrochemistry* vol 6, ed B E Conway and J O'M Bockris (New York: Plenum)
- [14] Pitts E, Tabor B E and Daly J 1969 *Trans. Farad. Soc.* **65** 849
- [15] Barlow C A and MacDonald J R 1967 *Adv. Electrochem. Electrochem. Eng.* **6** 1
- [16] Trassatti S 1980 *Comprehensive Treatise of Electrochemistry* ed J O'M Bockris, B E Conway and E Yeager (New York: Plenum)
- [17] Ashcroft N W and Mermin N D 1976 *Solid-State Physics* (New York: Holt, Rinehart and Winston)
- [18] Frumkin A N, Petrii O A and Damaskin B B 1980 *Comprehensive Treatise of Electrochemistry* ed J O'M Bockris, B E Conway and E Yeager (New York: Plenum)
- [19] Martynov G A and Salem R R 1983 *Electrical Double Layer at a Metal–Dilute Electrolyte Solution Interface* (Berlin: Springer)

- [20] Goüy G 1910 *J. Phys.* **9** 457
Chapman D L 1913 *Phil. Mag.* **25** 475
- [21] Bockris J O'M and Khan S U 1993 *Surface Electrochemistry* (New York: Plenum)
- [22] Watts-Tobin R J 1961 *Phil. Mag.* **6** 133
- [23] Parsons R 1975 *J. Electroanal. Chem.* **53** 229
- [24] Damaskin B B and Frumkin A N 1974 *Electrochim. Acta* **19** 173
- [25] Trassatti S 1986 *Trends in Interfacial Electrochemistry (NATO ASI Series 179)* ed A Fernando Silva (Dordrecht: Reidel)

- [26] Bewick A, Kunitatsu K, Robinson J and Russell J W 1981 *J. Electroanal. Chem.* **276** 175
Habib M A and Backris J O'M 1986 *Langmuir* **2** 388
- [27] Guidelli R 1986 *Trends in Interfacial Electrochemistry (NATO ASI Series 179)* ed A Fernando Silva (Dordrecht: Reidel)
- [28] Habib M A and Bockris J O'M 1980 *Comprehensive Treatise of Electrochemistry* ed J O'M Bockris, B E Conway and E Yeager (New York: Plenum)
- [29] Benjamin I 1997 *Mod. Aspects Electrochem.* **31** 115
- [30] Spohr E 1999 *Electrochim. Acta* **44** 1697
- [31] Schmickler W 1996 *Interfacial Electrochemistry* (Oxford: Oxford University Press)
-

-1-

A2.5 Phase transitions and critical phenomena

Robert L Scott

A2.5.1 ONE-COMPONENT FIRST-ORDER TRANSITIONS

The thermodynamic treatment of simple phase transitions is straightforward and is discussed in A2.1.6 and therefore need not be repeated here. In a one-component two-phase system, the phase rule yields one degree of freedom, so the transition between the two phases can be maintained along a pressure–temperature line. Figure A2.5.1 shows a typical p, T diagram with lines for fusion (solid–liquid), sublimation (solid–gas), and vaporization (liquid–gas) meeting at a triple point (solid–liquid–gas). Each of these lines can, at least in principle, be extended as a metastable line (shown as a dashed line) beyond the triple point. (Supercooling of gases below the condensation point, supercooling of liquids below the freezing point and superheating of liquids above the boiling point are well known; superheating of solids above the melting point is more problematic.) The vaporization line (i.e. the vapour pressure curve) ends at a critical point, with a unique pressure, temperature, and density, features that will be discussed in detail in subsequent sections. Because this line ends it is possible for a system to go around it and move continuously from gas to liquid without a phase transition; above the critical temperature the phase should probably just be called a ‘fluid’.

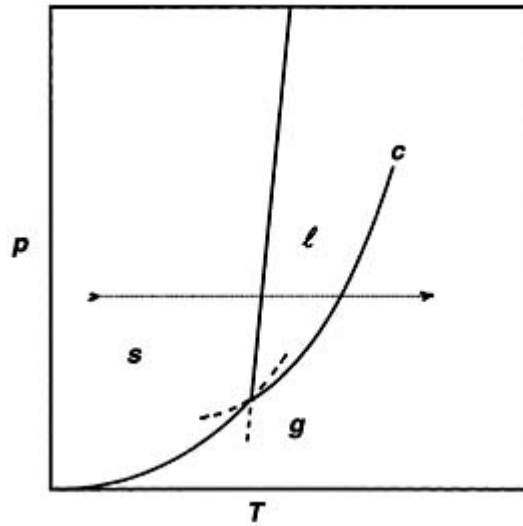


Figure A2.5.1. Schematic phase diagram (pressure p versus temperature T) for a typical one-component substance. The full lines mark the transitions from one phase to another (g, gas; l , liquid; s, solid). The liquid–gas line (the vapour pressure curve) ends at a critical point (c). The dotted line is a constant pressure line. The dashed lines represent metastable extensions of the stable phases.

Figure A2.5.2 shows schematically the behaviour of several thermodynamic functions along a constant-pressure line (shown as a dotted line in Figure A2.5.1)—the molar Gibbs free energy \bar{G} (for a one-component system the same as

the chemical potential μ), the molar enthalpy \bar{H} and the molar heat capacity at constant pressure \bar{C}_p . Again, at least in principle, each of the phases can be extended into a metastable region beyond the equilibrium transition.

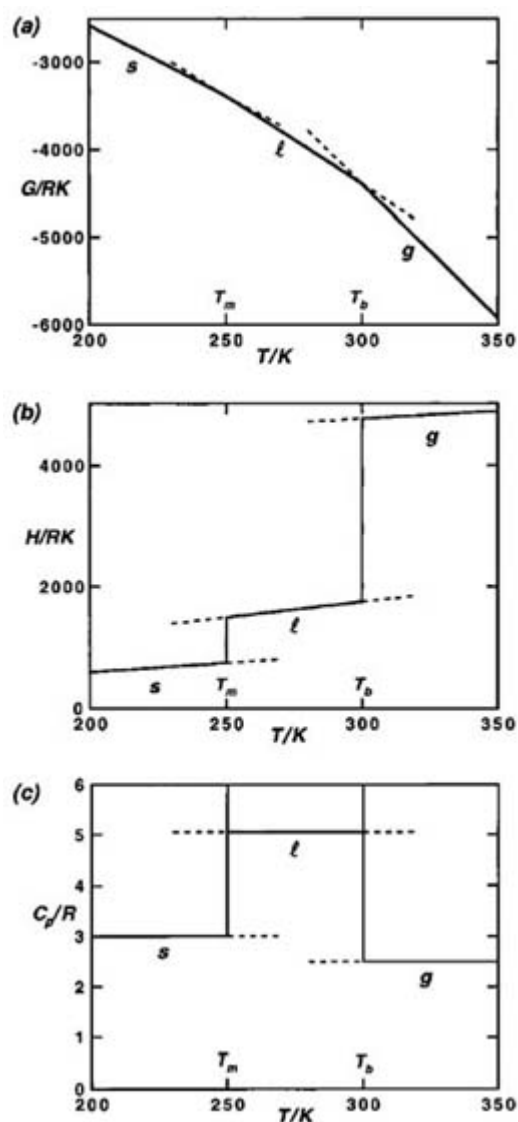


Figure A2.5.2. Schematic representation of the behaviour of several thermodynamic functions as a function of temperature T at constant pressure for the one-component substance shown in figure A2.5.1. (The constant-pressure path is shown as a dotted line in figure A2.5.1.) (a) The molar Gibbs free energy \bar{G} , (b) the molar enthalpy \bar{H} , and (c) the molar heat capacity at constant pressure \bar{C}_p . The functions shown are dimensionless (R is the gas constant per mole, while K is the temperature unit Kelvin). The dashed lines represent metastable extensions of the stable phases beyond the transition temperatures.

-3-

It will be noted that the free energy (figure A2.5.2(a)) is necessarily continuous through the phase transitions, although its first temperature derivative (the negative of the entropy S) is not. The enthalpy $H = G + TS$ (shown in figure A2.5.2(b)) is similarly discontinuous; the vertical discontinuities are of course the enthalpies of transition. The graph for the molar heat capacity \bar{C}_p (figure A2.5.2(c)) looks superficially like that for the enthalpy, but represents something quite different at the transition. The vertical line with an arrow at a transition temperature is a mathematical delta function, representing an ordinate that is infinite and an abscissa (ΔT) that is zero, but whose product is nonzero, an 'area' that is equal to the molar enthalpy of transition $\Delta\bar{H}$.

Phase transitions at which the entropy and enthalpy are discontinuous are called 'first-order transitions' because it is the first derivatives of the free energy that are discontinuous. (The molar volume $\bar{V} = (\partial\bar{G}/\partial p)_T$ is also discontinuous.) Phase transitions at which these derivatives are continuous but second derivatives of G

are discontinuous (e.g. the heat capacity, the isothermal compressibility, the thermal expansivity etc) are called ‘second order’.

The initial classification of phase transitions made by Ehrenfest (1933) was extended and clarified by Pippard [1], who illustrated the distinctions with schematic heat capacity curves. Pippard distinguished different kinds of second- and third-order transitions and examples of some of his second-order transitions will appear in subsequent sections; some of his types are unknown experimentally. Theoretical models exist for third-order transitions, but whether these have ever been found is unclear.

A2.5.2 PHASE TRANSITIONS IN TWO-COMPONENT SYSTEMS

Phase transitions in binary systems, normally measured at constant pressure and composition, usually do not take place entirely at a single temperature, but rather extend over a finite but nonzero temperature range. Figure A2.5.3 shows a temperature–mole fraction (T, x) phase diagram for one of the simplest of such examples, vaporization of an ideal liquid mixture to an ideal gas mixture, all at a fixed pressure, (e.g. 1 atm). Because there is an additional composition variable, the sample path shown in the figure is not only at constant pressure, but also at a constant total mole fraction, here chosen to be $x = 1/2$.

As the temperature of the liquid phase is increased, the system ultimately reaches a phase boundary, the ‘bubble point’ at which the gas phase (vapour) begins to appear, with the composition shown at the left end of the horizontal two-phase ‘tie-line’. As the temperature rises more gas appears and the relative amounts of the two phases are determined by applying a lever-arm principle to the tie-line: the ratio of the fraction f_g of molecules in the gas phase to that f_l in the liquid phase is given by the inverse of the ratio of the distances from the phase boundary to the position of the overall mole fraction x_0 of the system,

$$f_g/f_l = (x_l - x_0)/(x_0 - x_g).$$

-4-

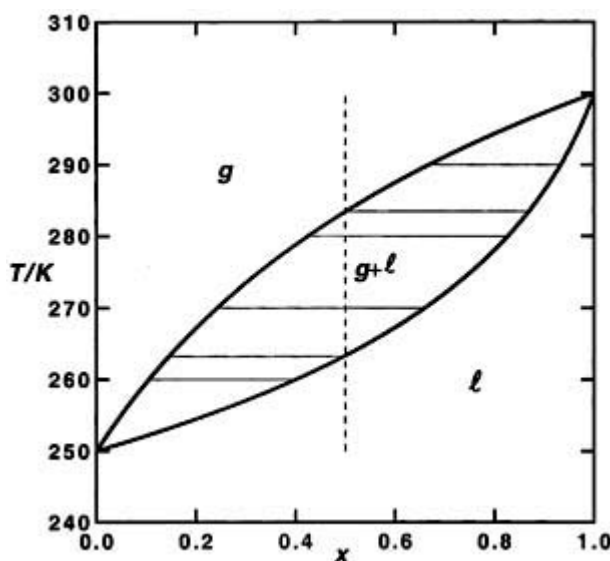


Figure A2.5.3. Typical liquid–gas phase diagram (temperature T versus mole fraction x at constant pressure) for a two-component system in which both the liquid and the gas are ideal mixtures. Note the extent of the two-phase liquid–gas region. The dashed vertical line is the direction ($x = 1/2$) along which the functions in figure A2.5.5 are determined.

With a further increase in the temperature the gas composition moves to the right until it reaches $x = 1/2$ at the phase boundary, at which point all the liquid is gone. (This is called the ‘dew point’ because, when the gas is cooled, this is the first point at which drops of liquid appear.) An important feature of this behaviour is that the transition from liquid to gas occurs gradually over a nonzero range of temperature, unlike the situation shown for a one-component system in [figure A2.5.1](#). Thus the two-phase region is bounded by a dew-point curve and a bubble-point curve.

[Figure A2.5.4](#) shows for this two-component system the same thermodynamic functions as in [figure A2.5.2](#), the molar Gibbs free energy $\bar{G} = x_1\mu_1 + x_2\mu_2$, the molar enthalpy \bar{H} and the molar heat capacity \bar{C}_p , again all at constant pressure, but now also at constant composition, $x = 1/2$. Now the enthalpy is continuous because the vaporization extends over an appreciable temperature range. Moreover, the heat capacity, while discontinuous at the beginning and at the end of the transition, is *not* a delta function. Indeed the graph appears to satisfy the definition of a second-order transition (or rather two, since there are two discontinuities).

-5-

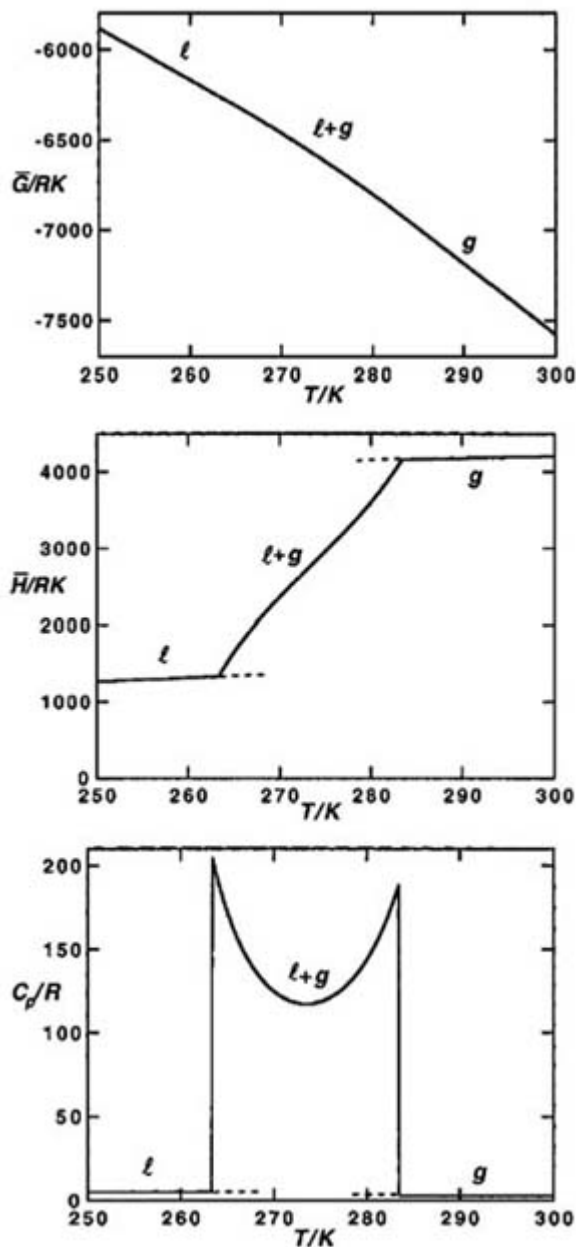


Figure A2.5.4. Thermodynamic functions \bar{G} , \bar{H} , and \bar{C}_p as a function of temperature T at constant pressure and composition ($x = 1/2$) for the two-component system shown in [figure A2.5.3](#). Note the difference between these and those shown for the one-component system shown in [figure A2.5.2](#). The functions shown are dimensionless as in [figure A2.5.2](#). The dashed lines represent metastable extensions (superheating or supercooling) of the one-phase systems.

-6-

However, this behaviour is not restricted to mixtures; it is also found in a one-component fluid system observed along a constant-volume path rather than the constant-pressure path illustrated in [figure A2.5.2](#). Clearly it would be confusing to classify the same transition as first- or second-order depending on the path. Pippard described such one-component constant-volume behaviour (discussed by Gorter) as a ‘simulated second-order transition’ and elected to restrict the term ‘second-order’ to a path along which two phases became more and more nearly alike until at the transition they became identical. As we shall see, that is what is seen when the system is observed along a path through a critical point. Further clarification of this point will be found in subsequent sections.

It is important to note that, in this example, as in ‘real’ second-order transitions, the curves for the two-phase region cannot be extended beyond the transition; to do so would imply that one had more than 100% of one phase and less than 0% of the other phase. Indeed it seems to be a quite general feature of all known second-order transitions (although it does not seem to be a thermodynamic requirement) that some aspect of the system changes gradually until it becomes complete at the transition point.

Three other examples of liquid–gas phase diagrams for a two-component system are illustrated in [figure A2.5.5](#) all a result of deviations from ideal behaviour. Such deviations in the liquid mixture can sometimes produce azeotropic behaviour, in which there are maximum or minimum boiling mixtures (shown in [figure A2.5.5\(a\)](#) and [figure A2.5.5\(b\)](#)). Except at the azeotropic composition (that of the maximum or minimum), a constant-composition path through the vaporization yields the same kind of qualitative behaviour shown in [figure A2.5.4](#). Behavior like that shown in [figure A2.5.2](#) is found only on the special path through the maximum or minimum, where the entire vaporization process occurs at a unique temperature.

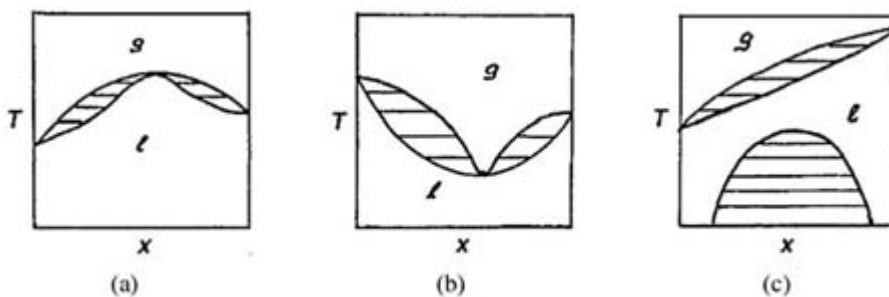


Figure A2.5.5. Phase diagrams for two-component systems with deviations from ideal behaviour (temperature T versus mole fraction x at constant pressure). Liquid–gas phase diagrams with maximum (a) and minimum (b) boiling mixtures (azeotropes). (c) Liquid–liquid phase separation, with a coexistence curve and a critical point.

A third kind of phase diagram in a two-component system (as shown in [figure A2.5.5\(c\)](#)) is one showing liquid–liquid phase separation below a critical-solution point, again at a fixed pressure. (On a T, x diagram, the critical point is always an extremum of the two-phase coexistence curve, but not always a maximum. Some binary systems show a minimum at a lower critical-solution temperature; a few systems show closed-loop two-phase regions with a maximum and a minimum.) As the temperature is increased at any composition other than the critical composition $x = x_c$, the compositions of the two coexisting phases adjust themselves to keep the total mole fraction unchanged until the coexistence curve is reached, above which only one phase

persists. Again, the behaviour of the thermodynamic functions agrees qualitatively with that shown in [figure A2.5.4](#) except that there is now only one transition line, not

-7-

two. However, along any special path leading through the critical point, there are special features in the thermodynamic functions that will be discussed in subsequent sections. First, however, we return to the one-component fluid to consider the features of its critical point.

A2.5.3 ANALYTIC TREATMENT OF CRITICAL PHENOMENA IN FLUID SYSTEMS. THE VAN DER WAALS EQUATION

All simple critical phenomena have similar characteristics, although all the analogies were not always recognized in the beginning. The liquid–vapour transition, the separation of a binary mixture into two phases, the order–disorder transition in binary alloys, and the transition from ferromagnetism to paramagnetism all show striking common features. At a low temperature one has a highly ordered situation (separation into two phases, organization into a superlattice, highly ordered magnetic domains, etc). At a sufficiently high temperature all long-range order is lost, and for all such cases one can construct a phase diagram (not always recognized as such) in which the long-range order is lost gradually until it vanishes at a critical point.

A2.5.3.1 THE VAN DER WAALS FLUID

Although later models for other kinds of systems are symmetrical and thus easier to deal with, the first analytic treatment of critical phenomena is that of van der Waals (1873) for coexisting liquid and gas [2]. The familiar van der Waals equation gives the pressure p as a function of temperature T and molar volume \bar{V} ,

$$p = RT/(\bar{V} - b) - a/\bar{V}^2 \quad (\text{A2.5.1})$$

where R is the gas constant per mole and a and b are constants characteristic of the particular fluid. The constant a is a measure of the strength of molecular attraction, while b is the volume excluded by a mole of molecules considered as hard spheres.

[Figure A2.5.6](#) shows a series of typical p, V isotherms calculated using equation (A2.5.1). (The temperature, pressure and volume are in reduced units to be explained below.) At sufficiently high temperatures the pressure decreases monotonically with increasing volume, but below a critical temperature the isotherm shows a maximum and a minimum.

-8-

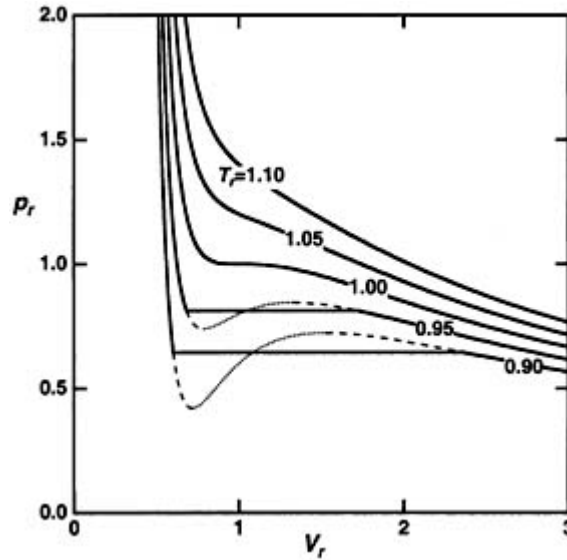


Figure A2.5.6. Constant temperature isotherms of reduced pressure p_r versus reduced volume V_r for a van der Waals fluid. Full curves (including the horizontal two-phase tie-lines) represent stable situations. The dashed parts of the smooth curve are metastable extensions. The dotted curves are unstable regions.

The coexistence lines are determined from the requirement that the potentials (which will subsequently be called ‘fields’), i.e. the pressure, the temperature, and the chemical potential μ , be the same in the conjugate (coexisting) phases, liquid and gas, at opposite ends of the tie-line. The equality of chemical potentials is equivalent to the requirement—for these variables (p, V), not necessarily for other choices—that the two areas between the horizontal coexistence line and the smooth cubic curve be equal. The full two-phase line is the stable situation, but the continuation of the smooth curve represents metastable liquid or gas phases (shown as a dashed curve); these are sometimes accessible experimentally. The part of the curve with a positive slope (shown as a dotted curve) represents a completely unstable situation, never realizable in practice because a negative compressibility implies instant separation into two phases. In analytic treatments like this of van der Waals, the maxima and minima in the isotherms (i.e. the boundary between metastable and unstable regions) define a ‘spinodal’ curve; in principle this line distinguishes between nucleation with an activation energy and ‘spinodal decomposition’. (see A3.3.) It should be noted that, if the free energy is nonanalytic as we shall find necessary in subsequent sections, the concept of the spinodal becomes unclear and can only be maintained by invoking special internal constraints. However, the usefulness of the distinction between activated nucleation and spinodal decomposition can be preserved.

With increasing temperature the two-phase tie-line becomes shorter and shorter until the two conjugate phases become identical at a critical point where the maximum and minimum in the isotherm have coalesced. Thus the critical point is defined as that point at which $(\partial p/\partial \bar{V})_T$ and $(\partial^2 p/\partial \bar{V}^2)_T$ are simultaneously zero, or where the equivalent quantities $(\partial^2 \bar{A}/\partial \bar{V})_T$ and $(\partial^3 \bar{A}/\partial \bar{V}^3)_T$ are simultaneously zero. These requirements yield the critical constants in terms of the constants R, a and b ,

-9-

$$p_c = a/(27b^2) \quad \bar{V}_c = 3b \quad T_c = 8a/(27Rb).$$

Equation (A2.5.1) can then be rewritten in terms of the reduced quantities $p_r = p/p_c$, $T_r = T/T_c$, and $V_r = \bar{V}/\bar{V}_c$

$$p_r = 8T_r/(3V_r - 1) - 3/V_r^2. \quad (\text{A2.5.2})$$

It is this equation with the reduced quantities that appears in [figure A2.5.6](#).

Since the pressure $p = -(\partial \bar{A} / \partial \bar{V})_T$, integration of [equation \(A2.5.1\)](#) yields $\bar{A}(T, \bar{V})$

$$\bar{A}(T, \bar{V}) - \bar{A}^\circ(T, \bar{V}^\circ) = -RT \ln[(\bar{V} - b)/\bar{V}^\circ] - a/\bar{V}$$

where $\bar{A}(T, \bar{V}^\circ)$ is the molar free energy of the ideal gas at T and \bar{V}° . (It is interesting to note that the van der Waals equation involves a simple separation of the free energy into entropy and energy; the first term on the right is just $-T(\bar{S} - \bar{S}^\circ)$, while the second is just $\bar{U} - \bar{U}^\circ$.)

The phase separation shown in [figure A2.5.6](#) can also be illustrated by the entirely equivalent procedure of plotting the molar Helmholtz free energy $\bar{A}(T, \bar{V})$ as a function of the molar volume \bar{V} for a series of constant temperatures, shown in [figure A2.5.7](#). At constant temperature and volume, thermodynamic equilibrium requires that the Helmholtz free energy must be minimized. It is evident for temperatures below the critical point that for certain values of the molar volume the molar free energy $\bar{A}(T, \bar{V})$ can be lowered by separation into two phases. The exact position of the phase separation is found by finding a straight line that is simultaneously tangent to the curve at two points; the slope at any point of the curve is $(\partial \bar{A} / \partial \bar{V}) = -p$, so the pressures are equal at the two tangent points. Similarly the chemical potential $\mu = \bar{A} + p \bar{V}$ is the same at the two points. That the dashed and dotted parts of the curve are metastable or unstable is clear because they represent higher values of \bar{A} than the corresponding points on the two-phase line. (The metastable region is separated from the completely unstable region by a point of inflection on the curve.)

The problem with [figure A2.5.6](#) and [figure A2.5.7](#) is that, because it extends to infinity, volume is not a convenient variable for a graph. A more useful variable is the molar density $\rho = 1 / \bar{V}$ or the reduced density $\rho_r = 1 / V_r$ which have finite ranges, and the familiar van der Waals equation can be transformed into an alternative although relatively unfamiliar form by choosing as independent variables the chemical potential μ and the density ρ .

-10-

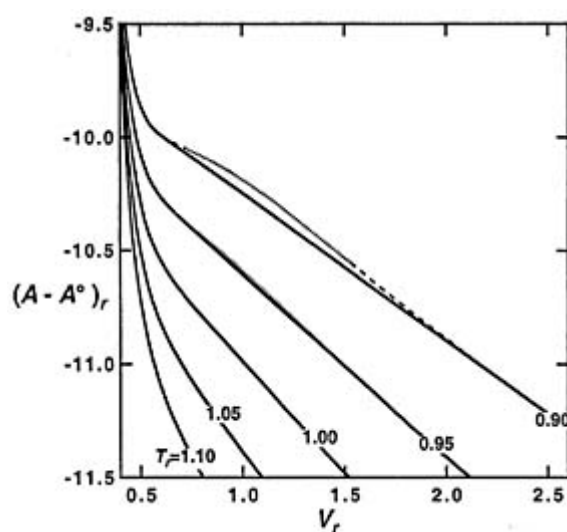


Figure A2.5.7. Constant temperature isotherms of reduced Helmholtz free energy A_r versus reduced volume V_r . The two-phase region is defined by the line simultaneously tangent to two points on the curve. The dashed parts of the smooth curve are metastable one-phase extensions while the dotted curves are unstable regions. (The isotherms are calculated for an unphysical $V_r = 0.1$, the only effect of which is to separate the isotherms

better.)

Unlike the pressure where $p = 0$ has physical meaning, the zero of free energy is arbitrary, so, instead of the ideal gas volume, we can use as a reference the molar volume of the real fluid at its critical point. A reduced Helmholtz free energy A_r in terms of the reduced variables T_r and V_r can be obtained by replacing a and b by their values in terms of the critical constants

$$A_r = [\bar{A}(T, \bar{V}) - \bar{A}(T, \bar{V}_c)] / (p_c \bar{V}_c) = -(8/3)T_r \ln[(3V_r - 1)/2] - 3(1 - V_r)/V_r.$$

Then, since the chemical potential for a one-component system is just $\mu = \bar{G} = \bar{A} + p\bar{V}$, a reduced chemical potential can be written in terms of a reduced density $\rho_r = \rho/\rho_c = \bar{V}_c/\bar{V}$

$$\begin{aligned} \mu_r &= [\mu(T, \rho) - \mu(T, \rho_c)](\rho_c/p_c) \\ &= -(8/3)T_r [\ln((3 - \rho_r)/(2\rho_r)) + (3/2)(\rho_r - 1)/(3 - \rho_r)] - 6(\rho_r - 1). \end{aligned} \quad (\text{A2.5.3})$$

Equation (A2.5.3) is a μ_r, ρ_r equation of state, an alternative to the p_r, V_r equation (A2.5.2).

The van der Waals μ_r, ρ_r isotherms, calculated using equation (A2.5.3), are shown in [figure A2.5.8](#). It is immediately obvious that these are much more nearly antisymmetric around the critical point than are the corresponding p_r, V_r isotherms in [figure A2.5.6](#) (of course, this is mainly due to the finite range of ρ_r from 0 to 3). The symmetry is not exact, however, as a careful examination of the figure will show. This choice of variables also satisfies the equal-area condition for coexistent phases; here the horizontal tie-line makes the chemical potentials equal and the equal-area construction makes the pressures equal.

-11-

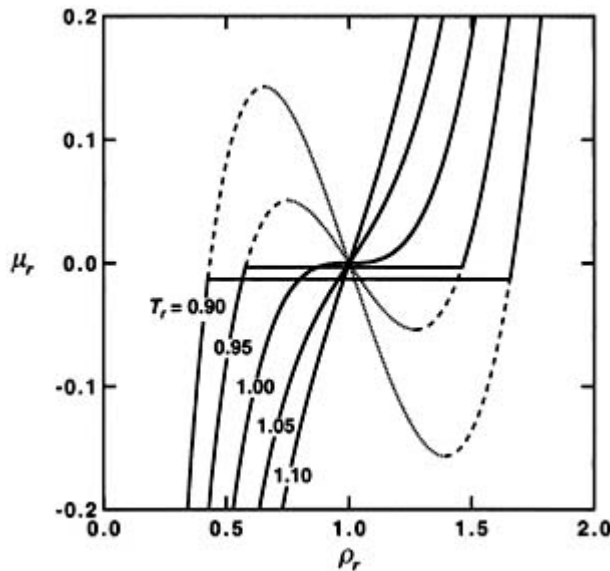


Figure A2.5.8. Constant temperature isotherms of reduced chemical potential μ_r versus reduced density ρ_r for a van der Waals fluid. Full curves (including the horizontal two-phase tie-lines) represent stable situations. The dashed parts of the smooth curve are metastable extensions, while the dotted curves are unstable regions.

For a system in which the total volume remains constant, the same minimization condition that applies to \bar{A} also applies to $\bar{A}/\bar{V} = \bar{A}p$, or to $(Ap)_r$, a quantity that can easily be expressed in terms of T_r and ρ_r ,

$$(A\rho)_T = -(8/3)T_r\rho_r \ln[(3 - \rho_r)/(2\rho_r)] - 3\rho_r(\rho_r - 1) - (4T_r - 3)(\rho_r - 1).$$

It is evident that, for the system shown in [figure A2.5.9](#) $\bar{A}/\bar{V} = \bar{A}\rho$ or $(A\rho)_T$ can be minimized for certain values of ρ on the low-temperature isotherms if the system separates into two phases rather than remaining as a single phase. As in [figure A2.5.7](#) the exact position of the phase separation is found by finding a straight line that is simultaneously tangent to the curve at two points; the slope at any point on the curve is the chemical potential μ , as is easily established by differentiating $\bar{A}\rho$ with respect to ρ ,

$$[\partial(\bar{A}\rho)/\partial\rho]_T = \bar{A} + \rho(\partial\bar{A}/\partial\rho)_T = \bar{A} - \bar{V}(\partial\bar{A}/\partial\bar{V})_T = \bar{A} + p\bar{V} = \mu.$$

(The last term in the equation for $(A\rho)_T$ has been added to avoid adding a constant to μ_r ; doing so does not affect the principle, but makes [figure A2.5.9](#) clearer.) Thus the common tangent satisfies the condition of equal chemical potentials, $\mu_\ell = \mu_g$. (The common tangent also satisfies the condition that $p_\ell = p_g$ because $p = \mu\rho - \rho\bar{A}$.) The two points of tangency determine the densities of liquid and gas, ρ_ℓ and ρ_g , and the relative volumes of the two phases are determined by the lever-arm rule.

-12-

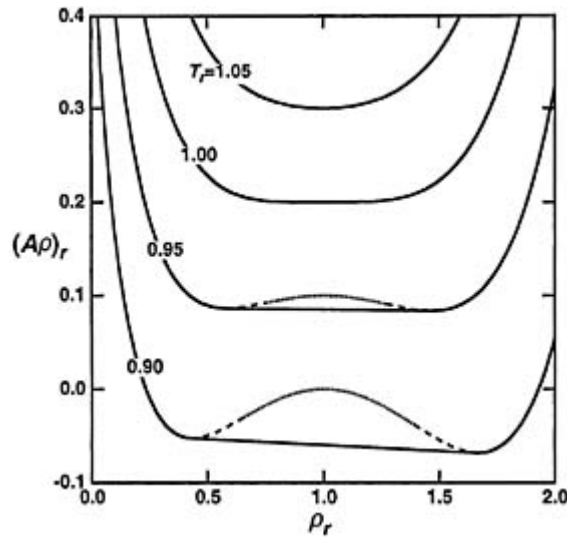


Figure A2.5.9. $(A\rho)_r$, the Helmholtz free energy per unit volume in reduced units, of a van der Waals fluid as a function of the reduced density ρ_r for several constant temperatures above and below the critical temperature. As in the previous figures the full curves (including the tangent two-phase tie-lines) represent stable situations, the dashed parts of the smooth curve are metastable extensions, and the dotted curves are unstable regions. See text for details.

The T_r, ρ_r coexistence curve can be calculated numerically to any desired precision and is shown in [figure A2.5.10](#). The spinodal curve (shown dotted) satisfies the equation

$$(\partial^2 A_r / \partial V_r^2)_{T_r} = -(\partial p_r / \partial V_r)_{T_r} = 6\rho_r^2 [4T_r / (3\rho_r - 1)^2 - \rho_r] = 0.$$

Alternatively, expansion of [equation \(A2.5.1\)](#), [equation \(A2.5.2\)](#) or [equation \(A2.5.3\)](#) into Taylor series leads ultimately to series expressions for the densities of liquid and gas, ρ_ℓ and ρ_g , in terms of their sum (called the ‘diameter’) and their difference:

$$(\rho_\ell + \rho_g) / (2\rho_c) = 1 + (2/5)(1 - T_r) + (128/875)(1 - T_r)^2 + \dots \quad (\text{A2.5.4})$$

$$\begin{aligned}
(\rho_\ell - \rho_g)/(2\rho_c) &= 2(1 - T_r)^{1/2} - (13/25)(1 - T_r)^{3/2} + \dots \\
\text{or } [(\rho_\ell - \rho_g)/(2\rho_c)]^2 &= 4(1 - T_r) - (52/25)(1 - T_r)^2 + \dots
\end{aligned}
\tag{A2.5.5}$$

Note that equation (A2.5.5), like equation (A2.5.4), is just a power series in $(1 - T_r) = (T_c - T)/T_c$, a variable that will appear often and will henceforth be represented by t . All simple equations of state (e.g. the Dieterici and Berthelot equations) yield equations of the same form as equation (A2.5.4) and equation (A2.5.5); only the coefficients differ. There are better expressions for the contribution of the hard-sphere fluid to the pressure than the van der Waals $RT/(\bar{V} - b)$, but the results are similar. Indeed it can be shown that any analytic equation of state, however complex, must necessarily yield similar power series.

-13-

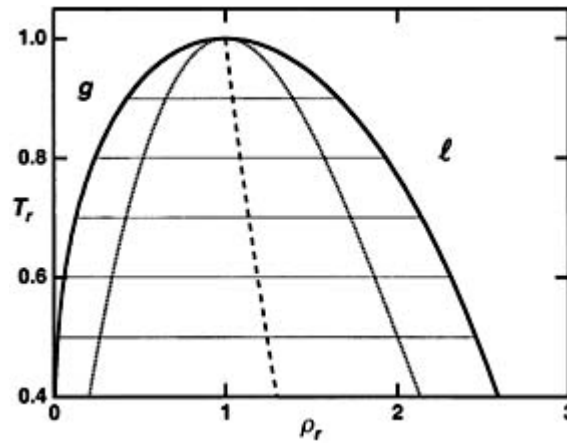


Figure A2.5.10. Phase diagram for the van der Waals fluid, shown as reduced temperature T_r versus reduced density ρ_r . The region under the smooth coexistence curve is a two-phase liquid–gas region as indicated by the horizontal tie-lines. The critical point at the top of the curve has the coordinates (1,1). The dashed line is the diameter, and the dotted curve is the spinodal curve.

If the small terms in t^2 and higher are ignored, equation (A2.5.4) is the ‘law of the rectilinear diameter’ as evidenced by the straight line that extends to the critical point in figure A2.5.10 this prediction is in good qualitative agreement with most experiments. However, equation (A2.5.5), which predicts a parabolic shape for the top of the coexistence curve, is unsatisfactory as we shall see in subsequent sections.

The van der Waals energy $\bar{U} = -a/\bar{V} = -a\rho$. On a path at constant total critical density $\rho = \rho_c$, which is a constant-volume path, the energy of the system will be the sum of contributions from the two conjugate phases, the densities and amounts of which are changing with temperature. With proper attention to the amounts of material in the two phases that maintain the constant volume, this energy can be written relative to the one-phase energy \bar{U}_c at the critical point,

$$\begin{aligned}
(\bar{U} - \bar{U}_c)/(a\rho_c) &= (\bar{U} - \bar{U}_c)/(9RT_c/8) = -(\rho_\ell + \rho_g)/\rho_c + \rho_\ell\rho_g/\rho_c^2 + 1 \\
&= -(\rho_\ell + \rho_g)/\rho_c + (\rho_\ell + \rho_g)^2/(2\rho_c)^2 - (\rho_\ell - \rho_g)^2/(2\rho_c)^2 + 1
\end{aligned}$$

or, substituting from equation (A2.5.4) and equation (A2.5.5),

$$(\bar{U} - \bar{U}_c)/(RT_c) = (9/2)[-t + (14/25)t^2 + \dots].$$

Differentiating this with respect to $T = -T_c t$ yields a heat capacity at constant volume,

$$\bar{C}_V = (\partial \bar{U} / \partial T)_{\bar{V}} = (9/2)R[1 - (28/25)t + \dots]. \quad (\text{A2.5.6})$$

-14-

This is of course an excess heat capacity, an amount in addition to the contributions of the heat capacities \bar{C}_V of the liquid and vapour separately. Note that this excess (as well as the total heat capacity shown later in [figure A2.5.26](#) is always finite, and increases only linearly with temperature in the vicinity of the critical point; at the critical point there is no further change in the excess energy, so this part of the heat capacity drops to zero. This behaviour looks very similar to that of the simple binary system in [section \(A2.5.2\)](#). However, unlike that system, in which there is no critical point, the experimental heat capacity \bar{C}_V along the critical isochore (constant volume) appears to diverge at the critical temperature, contrary to the prediction of [equation \(A2.5.6\)](#).

Finally, we consider the isothermal compressibility $\kappa_T = -(\partial \ln V / \partial p)_T = (\partial \ln \rho / \partial p)_T$ along the coexistence curve. A consideration of [figure A2.5.6](#) shows that the compressibility is finite and positive at every point in the one-phase region except at the critical point. Differentiation of [equation \(A2.5.2\)](#) yields the compressibility along the critical isochore:

$$\rho_c \kappa_T = 1/[6(T_r - 1)] = 1/(-6t) \quad (\rho = \rho_c, T_r \geq 1).$$

At the critical point (and anywhere in the two-phase region because of the horizontal tie-line) the compressibility is infinite. However the compressibility of each conjugate phase can be obtained as a series expansion by evaluating the derivative (as a function of ρ_r) for a particular value of T_r , and then substituting the values of ρ_r for the ends of the coexistence curve. The final result is

$$\begin{aligned} \rho_c \kappa_T &= 1/[12t \pm (216/5)t^{3/2} + \dots] \\ &= [1/(12t)][1 \mp (12/5)t^{1/2} + \dots] \quad (\text{coex}, T_r \leq 1) \end{aligned} \quad (\text{A2.5.7})$$

where in the \pm and the \mp , the upper sign applies to the liquid phase and the lower sign to the gas phase. It is to be noted that although the compressibility becomes infinite as one approaches T_c from either direction, its value at a small δT below T_c is only half that at the same δT above T_c ; this means that there is a discontinuity at the critical point.

Although the previous paragraphs hint at the serious failure of the van der Waals equation to fit the shape of the coexistence curve or the heat capacity, failures to be discussed explicitly in later sections, it is important to recognize that many of the other predictions of analytic theories are reasonably accurate. For example, analytic equations of state, even ones as approximate as that of van der Waals, yield reasonable values (or at least ‘ball park estimates’) of the critical constants p_c , T_c , and \bar{V}_c . Moreover, in two-component systems where the critical point expands to a critical line in p, T space, or in three-component systems where there are critical surfaces, simple models yield many useful predictions. It is only in the vicinity of critical points that analytic theories fail.

A2.5.3.2 THE VAN DER WAALS FLUID MIXTURE

Van der Waals (1890) extended his theory to mixtures of components A and B by introducing mole-fraction-dependent parameters a_m and b_m defined as quadratic averages

$$a_m = (1 - x)^2 a_{AA} + 2x(1 - x)a_{AB} + x^2 a_{BB} \quad (\text{A2.5.8})$$

$$b_m = (1 - x)^2 b_{AA} + 2x(1 - x)b_{AB} + x^2 b_{BB} \quad (\text{A2.5.9})$$

where the a s and b s extend the meanings defined in [section A2.5.3.1](#) for the one-component fluid to the three kinds of pairs in the binary mixture; x is the mole fraction of the second component (B). With these definitions of a_m and b_m , [equation \(A2.5.1\)](#) for the pressure remains unchanged, but an entropy of mixing must be added to the equation for the Helmholtz free energy.

$$\begin{aligned} \bar{A}(T, \bar{V}) &= (1 - x)\bar{A}_A(T, \bar{V}^\circ) + x\bar{A}_B(T, \bar{V}^\circ) \\ &= RT[(1 - x) \ln(1 - x) + x \ln x] + RT \ln[(\bar{V} - b_m)/\bar{V}^\circ] - a_m/\bar{V}. \end{aligned}$$

Van der Waals and especially van Laar simplified these expressions by assuming a geometric mean for a_{AB} and an arithmetic mean for b_{AB} :

$$a_{12} = (a_{11}a_{22})^{1/2} \quad \text{and} \quad b_{12} = (b_{11} + b_{22})/2.$$

Then [equation \(A2.5.8\)](#) and [equation \(A2.5.9\)](#) for a_m and b_m become

$$a_m = [(1 - x)a_{AA}^{1/2} + xa_{BB}^{1/2}]^2 \quad (\text{A2.5.10})$$

$$b_m = (1 - x)b_{AA} + xb_{BB}. \quad (\text{A2.5.11})$$

With these simplifications, and with various values of the a s and b s, van Laar (1906–1910) calculated a wide variety of phase diagrams, determining critical lines, some of which passed continuously from liquid–liquid critical points to liquid–gas critical points. Unfortunately, he could only solve the difficult coupled equations by hand and he restricted his calculations to the geometric mean assumption for a_{12} (i.e. to [equation \(A2.5.10\)](#)). For a variety of reasons, partly due to the eclipse of the van der Waals equation, this extensive work was largely ignored for decades.

A2.5.3.3 GLOBAL PHASE DIAGRAMS

Half a century later Van Konynenburg and Scott (1970, 1980) [3] used the van der Waals equation to derive detailed phase diagrams for two-component systems with various parameters. Unlike van Laar they did not restrict their treatment to the geometric mean for a_{AB} , and for the special case of $b_{AA} = b_{BB} = b_{AB}$ (equal-sized molecules), they defined two reduced variables,

$$\zeta = (a_{BB} - a_{AA})/(a_{BB} + a_{AA}) \quad (\text{A2.5.12})$$

$$\lambda = (a_{BB} - 2a_{AB} + a_{AA})/(a_{BB} + a_{AA}). \quad (\text{A2.5.13})$$

Physically, ζ is a measure of the difference in the energies of vaporization of the two species (roughly a difference in normal boiling point), and λ is a measure of the energy of mixing. With these definitions [equation \(A2.5.8\)](#) can be rewritten as

$$a_m/a_{AA} = [(1 - \zeta) + 2x(\zeta - \lambda) + x^2\lambda]/(1 - \zeta).$$

If a_{AB} is weak in comparison to a_{AA} and a_{BB} , λ is positive and separation into two phases may occur.

With this formulation a large number of very different phase diagrams were calculated using computers that did not exist in 1910. Six principal types of binary fluid phase diagrams can be distinguished by considering where critical lines begin and end. These are presented in [figure A2.5.11](#) as the p, T projections of p, T, x diagrams, which show the vapour pressure curves for the pure substances, critical lines and three-phase lines. To facilitate understanding of these projections, a number of diagrams showing T versus x for a fixed pressure, identified by horizontal lines in [figure A2.5.11](#), are shown in [figure A2.5.12](#). Note that neither of these figures shows any solid state, since the van der Waals equation applies only to fluids. The simple van der Waals equation for mixtures yields five of these six types of phase diagrams. Type VI, with its low-pressure (i.e. below 1 atm) closed loop between a lower critical-solution temperature (LCST) and an upper critical-solution temperature (UCST), cannot be obtained from the van der Waals equation without making λ temperature dependent.

-17-

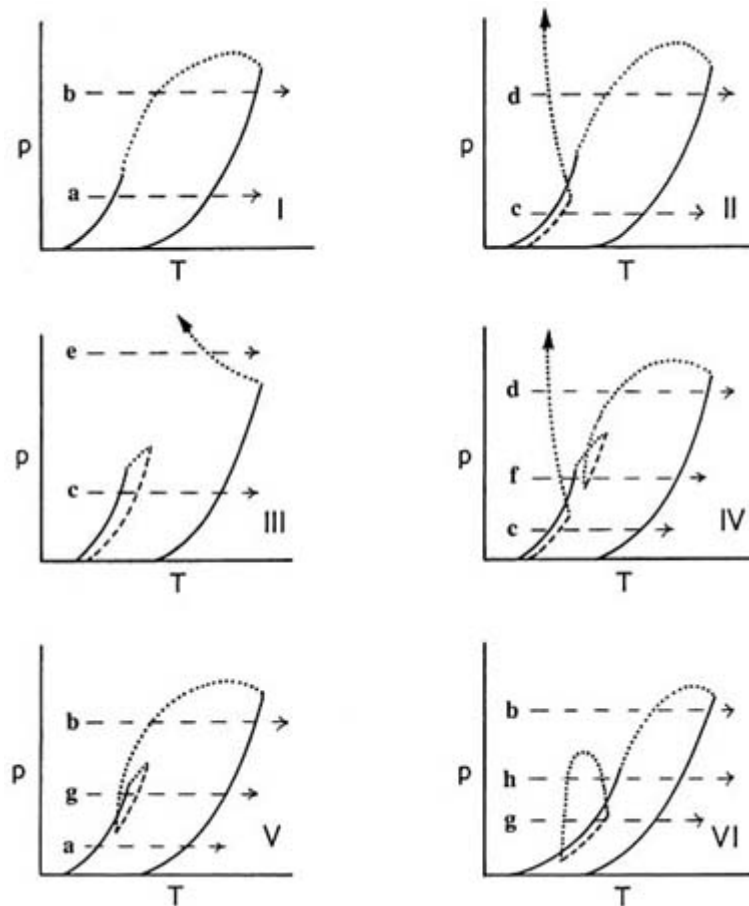


Figure A2.5.11. Typical pressure–temperature phase diagrams for a two-component fluid system. The full curves are vapour pressure lines for the pure fluids, ending at critical points. The dotted curves are critical lines, while the dashed curves are three-phase lines. The dashed horizontal lines are not part of the phase diagram, but indicate constant-pressure paths for the (T, x) diagrams in [figure A2.5.12](#). All but the type VI diagrams are predicted by the van der Waals equation for binary mixtures. Adapted from figures in [3].

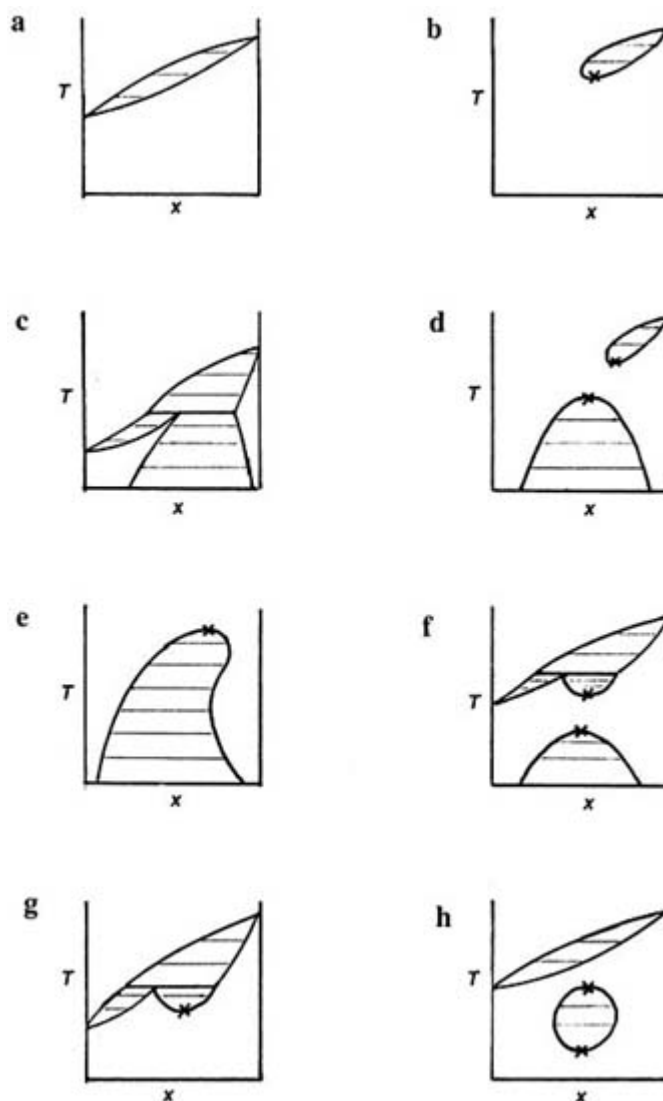


Figure A2.5.12. Typical temperature T versus mole fraction x diagrams for the constant-pressure paths shown in [figure A2.5.11](#). Note the critical points (\times) and the horizontal three-phase lines.

The boundaries separating these principal types of phase behaviour are shown on a λ , ζ diagram (for equal-sized molecules) in [figure A2.5.13](#). For molecules of different size, but with the approximation of [equation \(A2.5.10\)](#), more global phase diagrams were calculated using a third parameter,

$$\xi = (b_{BB} - b_{AA}) / (b_{BB} + b_{AA})$$

and appropriately revised definitions for ζ and λ . For different-sized molecules ($\xi \neq 0$), the global phase diagram is no longer symmetrical, but the topology is the same.

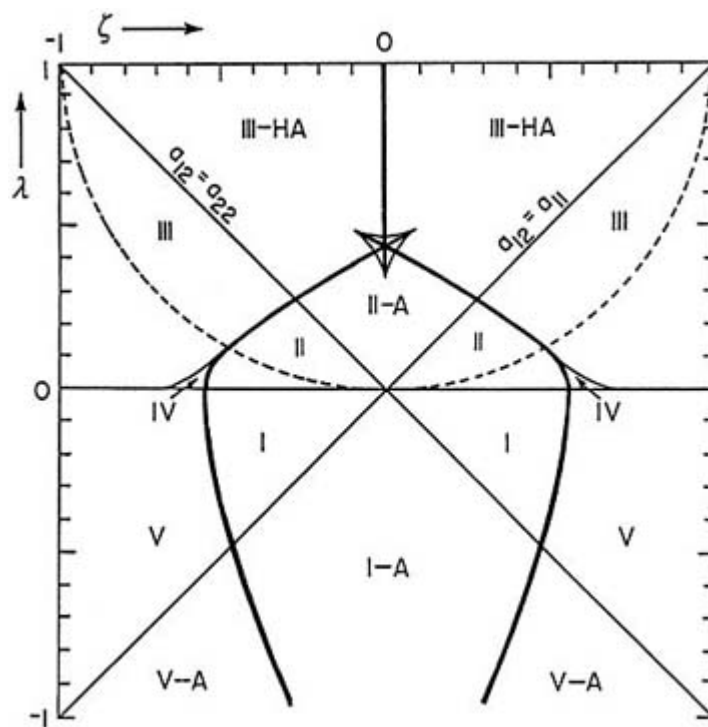


Figure A2.5.13. Global phase diagram for a van der Waals binary mixture for which $b_{AA} = b_{BB}$. The coordinates λ and ζ are explained in the text. The curves separate regions of the different types shown in [figure A2.5.11](#). The heavier curves are tricritical lines (explained in section A2.5.9). The ‘shield region’ in the centre of the diagram where three tricritical lines intersect consists of especially complex phase diagrams not yet found experimentally. Adapted from figures in [3].

In recent years global phase diagrams have been calculated for other equations of state, not only van der Waals-like ones, but others with complex temperature dependences. Some of these have managed to find type VI regions in the overall diagram. Some of the recent work was brought together at a 1999 conference [4].

A2.5.4 ANALYTIC TREATMENTS OF OTHER CRITICAL PHENOMENA

A2.5.4.1 LIQUID–LIQUID PHASE SEPARATION IN A SIMPLE BINARY MIXTURE

The previous section showed how the van der Waals equation was extended to binary mixtures. However, much of the early theoretical treatment of binary mixtures ignored equation-of-state effects (i.e. the contributions of the expansion beyond the volume of a close-packed liquid) and implicitly avoided the distinction between constant pressure and constant volume by putting the molecules, assumed to be equal in size, into a kind of pseudo-lattice. [Figure A2.5.14](#) shows schematically an equimolar mixture of A and B, at a high temperature where the distribution is essentially random, and at a low temperature where the mixture has separated into two virtually one-component phases.

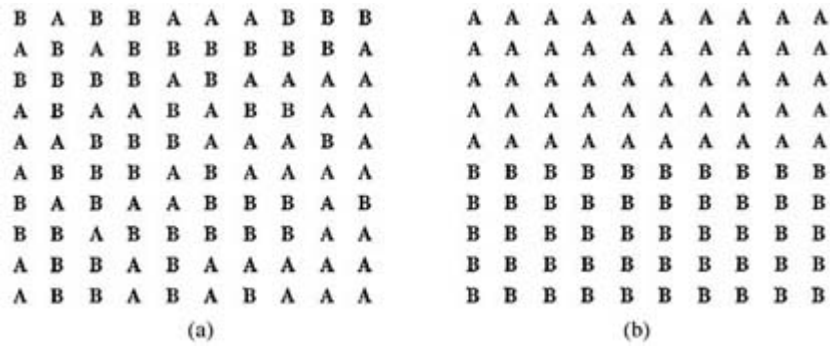


Figure A2.5.14. Quasi-lattice representation of an equimolar binary mixture of A and B (a) randomly mixed at high temperature, and (b) phase separated at low temperature.

The molar Helmholtz free energy of mixing (appropriate at constant volume) for such a symmetrical system of molecules of equal size, usually called a ‘simple mixture’, is written as a function of the mole fraction x of the component B

$$\begin{aligned} \Delta \bar{A}^M &= \bar{A}_m - (1-x)\mu_A^\circ - x\mu_B^\circ \\ &= RT[x \ln x + (1-x) \ln(1-x)] + Kx(1-x) \end{aligned} \tag{A2.5.14}$$

where the μ° 's are the chemical potentials of the pure components. The Gibbs free energy of mixing $\Delta \bar{G}^M$ is (at constant pressure) $\Delta \bar{A}^M + p\Delta \bar{V}^M$ and many theoretical treatments of such a system ignore the volume change on mixing and use the equation above for $\Delta \bar{G}^M$, which is the quantity of interest for experimental measurements at constant pressure. Equation (A2.5.14) is used to plot $\Delta \bar{A}^M$ or equivalently $\Delta \bar{G}^M$ versus x for several temperatures in [figure A2.5.15](#). As in the case of the van der Waals fluid a tangent line determines phase separation, but here the special symmetry requires that it be horizontal and that the mole fractions of the conjugate phases x' and x'' satisfy the condition $x'' = 1 - x'$. The critical-solution point occurs where $(\partial^2 \Delta \bar{G}^M / \partial x^2)_{T,p}$ and $(\partial^3 \Delta \bar{G}^M / \partial x^3)_{T,p}$ are simultaneously zero; for this special case, this point is at $x_c = 1/2$ and $T_c = K/2R$. The reduced temperatures that appear on the isotherms in [figure A2.5.15](#) are then defined as $T_r = T/T_c = 2RT/K$.

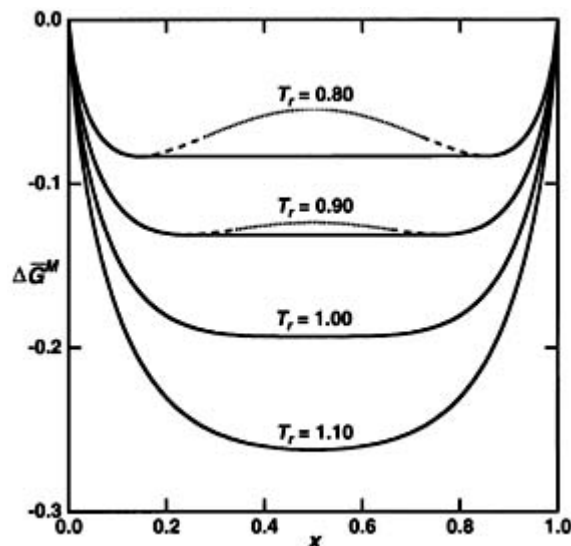


Figure A2.5.15. The molar Gibbs free energy of mixing $\Delta\bar{G}^M$ versus mole fraction x for a simple mixture at several temperatures. Because of the symmetry of equation (A2.5.15) the tangent lines indicating two-phase equilibrium are horizontal. The dashed and dotted curves have the same significance as in previous figures.

In the simplest model the coefficient K depends only on the differences of the attractive energies $-\varepsilon$ of the nearest-neighbour pairs (these energies are negative relative to those of the isolated atoms, but here their magnitudes ε are expressed as positive numbers)

$$K = \bar{N}(z/2)(\varepsilon_{AA} - 2\varepsilon_{AB} + \varepsilon_{BB}) = \bar{N}w.$$

Here \bar{N} is the number of molecules in one mole and z is the coordination number of the pseudo-lattice (4 in [figure A2.5.14](#) ; so $\bar{N}(z/2)$ is the total number of nearest-neighbour pairs in one mole of the mixture. The quantity w is the *interchange energy*, the energy involved in the single exchange of molecules between the two pure components. (Compare the parameter λ for the van der Waals mixture in the section above ([equation \(A2.5.13\)](#))). If w is a constant, independent of temperature, then K is temperature independent and $Kx(1-x)$ is simply the molar energy of mixing $\Delta\bar{U}^M$ (frequently called the molar enthalpy of mixing $\Delta\bar{H}^M$ when the volume change is assumed to be zero). If the chemical potentials μ_1 and μ_2 are derived from the free energy of mixing, μ, x isotherms are obtained that are qualitatively similar to the μ, ρ isotherms shown in [figure A2.5.8](#) an unsurprising result in view of the similarity of the free energy curves for the two cases. For such a symmetrical system one may define a ‘degree of order’ or ‘order parameter’ $s = 2x - 1$ such that s varies from -1 at $x = 0$ to $+1$ at $x = 1$. Then $x = (1 + s)/2$, and $1 - x = (1 - s)/2$ and [equation \(A2.5.14\)](#) can be rewritten as:

$$\Delta\bar{G}^M \cong \Delta\bar{A}^M = RT \left[\frac{(1+s)}{2} \ln \frac{(1+s)}{2} + \frac{(1-s)}{2} \ln \frac{(1-s)}{2} \right] + K \frac{(1-s^2)}{4}. \quad (\text{A2.5.15})$$

It is easy to derive the coexistence curve. Because of the symmetry, the double tangent is horizontal and the coexistent

-22-

phases occur at values of s where $(\partial\Delta^M/\partial s)_T$ equals zero.

Even if K is temperature dependent, the coexistence curve can still be defined in terms of a reduced temperature $T_r = 2RT/K(T)$, although the reduced temperature is then no longer directly proportional to the temperature T .

(A2.5.16)

Figure A2.5.16 shows the coexistence curve obtained from equation (A2.5.16). The logarithms (or the hyperbolic tangent) can be expanded in a power series, yielding

This series can be reverted and the resulting equation is very simple:

(A2.5.17)

The leading term in equation (A2.5.17) is the same kind of parabolic coexistence curve found in [section A2.5.3.1](#) from the van der Waals equation. The similarity between [equation \(A2.5.5\)](#) and equation (A2.5.17) should be obvious; the form is the same even though the coefficients are different.

Figure A2.5.16. The coexistence curve, $T_r = K/(2R)$ versus mole fraction x for a simple mixture. Also shown as an abscissa is the order parameter s , which makes the diagram equally applicable to order–disorder phenomena in solids and to ferromagnetism. The dotted curve is the spinodal.

-23-

The derivative $(\partial\Delta\bar{G}^M/\partial x)_T = \mu_B - \mu_A$, where μ_B and μ_A are the chemical potentials of the two species, is the analogue of the pressure in the one-component system. From [equation \(A2.5.15\)](#) one obtains

$$(\partial\Delta\bar{G}^M/\partial x)_T = 2(\partial\Delta\bar{G}^M/\partial s)_T = RT[\ln(1+s) - \ln(1-s)] - Ks.$$

At $s = 0$ this derivative obviously vanishes for all temperatures, but this is simply a result of the symmetry. The second derivative is another matter:

$$(\partial^2\Delta\bar{G}^M/\partial x^2)_T = 4(\partial^2\Delta\bar{G}^M/\partial s^2)_T = 2[2RT/(1-s^2) - K] = 4RT_c[T_r/(1-s^2) - 1].$$

This vanishes at the critical-solution point as does $(\partial p/\partial\bar{V})_T$ at the one-component fluid critical point. Thus an ‘osmotic compressibility’ or ‘osmotic susceptibility’ can be defined by analogy with the compressibility κ_T of the one-component fluid. Its value along the simple-mixture coexistence curve can be obtained using [equation \(A2.5.17\)](#) and is found to be proportional to t^{-1} . The osmotic compressibility of a binary mixture diverges at the critical point just like the compressibility of a one-component fluid (compare this to [equation \(A2.5.7\)](#)).

For a temperature-independent K , the molar enthalpy of mixing is

$$\Delta\bar{H}^M = Kx(1-x) = 2RT_c(1-s^2)/4 = (RT_c/2)[1 - 3t + (12/5)t^2 - \dots]$$

and the excess mixing contribution to the heat capacity (now at constant pressure) is

$$\Delta\bar{C}_{p,x=x_c}^M = (R/2)[3 - (24/5)t + \dots].$$

Again, as in the case of \bar{C}_V for the van der Waals fluid, there is a linear increase up to a finite value at the critical point and then a sudden drop to the heat capacity of the one-phase system because the liquids are now completely mixed.

Few if any binary mixtures are exactly symmetrical around $x = 1/2$, and phase diagrams like that sketched in [figure A2.5.5\(c\)](#) are typical. In particular one can write for mixtures of molecules of different size (different molar volumes \bar{V}_A° and \bar{V}_B°) the approximate equation

$$\Delta\bar{G}^M = RT[x \ln \phi + (1-x) \ln(1-\phi)] + [(1-x)\bar{V}_A^\circ + x\bar{V}_B^\circ]K'\phi(1-\phi)$$

which is a combination of the Flory entropy of mixing for polymer solutions with an enthalpy of mixing due to Scatchard and Hildebrand. The variable ϕ is the volume fraction of component B, $\phi = x_B \bar{V}_B^\circ / (x_A \bar{V}_A^\circ + x_B \bar{V}_B^\circ)$, and the parameter K' now has the dimensions of energy per unit volume. The condition for a critical-solution point, that the two derivatives cited above must simultaneously equal zero, yields the results

-24-

$$K'(T_c) = (RT_c/2)[(1/\bar{V}_A^\circ)^{1/2} + (1/\bar{V}_B^\circ)^{1/2}]^2$$

$$\phi_c = (\bar{V}_A^\circ)^{1/2} / [(\bar{V}_A^\circ)^{1/2} + (\bar{V}_B^\circ)^{1/2}].$$

This simple model continues to ignore the possibility of volume changes on mixing, so for simplicity the molar volumes \bar{V}_A° and \bar{V}_B° are taken as those of the pure components. It should come as no surprise that in this unsymmetrical system both $\phi' + \phi''$ and $\phi' - \phi''$ must be considered and that the resulting equations have extra terms that look like those in [equation \(A2.5.4\)](#) and [equation \(A2.5.5\)](#) for the van der Waals mixture; as in that case, however, the top of the coexistence curve is still parabolic. Moreover the parameter K' is now surely temperature dependent (especially so for polymer solutions), and the calculation of a coexistence curve will depend on such details.

As in the one-fluid case, the experimental sums are in good agreement with the law of the rectilinear diameter, but the experimental differences fail to give a parabolic shape to the coexistence curve.

It should be noted that a strongly temperature-dependent K (or K') can yield more than one solution to the equation $T_c = K/2R$. [Figure A2.5.17](#) shows three possible examples of a temperature-dependent K for the simple mixture: (a) a constant K as assumed in the discussion above, (b) a K that slowly decreases with T , the most common experimental situation, and (c) a K that is so sharply curved that it produces not only an upper critical-solution temperature (UCST), but also a lower critical-solution temperature (LCST) below which the fluids are completely miscible (i.e. the type VI closed-loop binary diagram of [section A2.5.3.3](#)). The position of the curves can be altered by changing the pressure; if the two-phase region shrinks until the LCST and UCST merge, one has a 'double critical point' where the curve just grazes the critical line. A fourth possibility (known experimentally but not shown in [figure A2.5.17](#)) is an opposite curvature producing a low-temperature UCST and a high-temperature LCST with a one-phase region at intermediate temperatures; if these two critical-solution temperatures coalesce, one has a 'critical double point'.

-25-

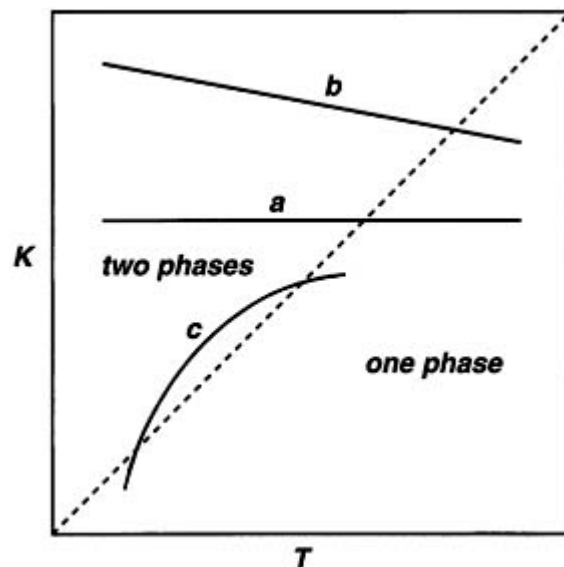


Figure A2.5.17. The coefficient K as a function of temperature T . The line $K = 2RT$ (shown as dashed line) defines the critical point and separates the two-phase region from the one-phase region. (a) A constant K as assumed in the simplest example; (b) a slowly decreasing K , found frequently in experimental systems, and (c) a sharply curved $K(T)$ that produces two critical-solution temperatures with a two-phase region in between.

A2.5.4.2 ORDER-DISORDER IN SOLID MIXTURES

In a liquid mixture with a negative K (negative interchange energy w), the formation of unlike pairs is favoured and there is no phase separation. However, in a crystal there is long-range order and at low temperatures, although there is no physical phase separation, a phase transition from a disordered arrangement to a regular arrangement of alternating atoms is possible. The classic example is that of β -brass (CuZn) which crystallizes in a body-centred cubic lattice. At high temperature the two kinds of atoms are distributed at random, but at low temperature they are arranged on two interpenetrating simple cubic sublattices such that each Cu has eight Zn nearest neighbours, and each Zn eight Cu nearest neighbours, as shown in [figure A2.5.18](#) this is like the arrangement of ions in a crystal of CsCl.

-26-

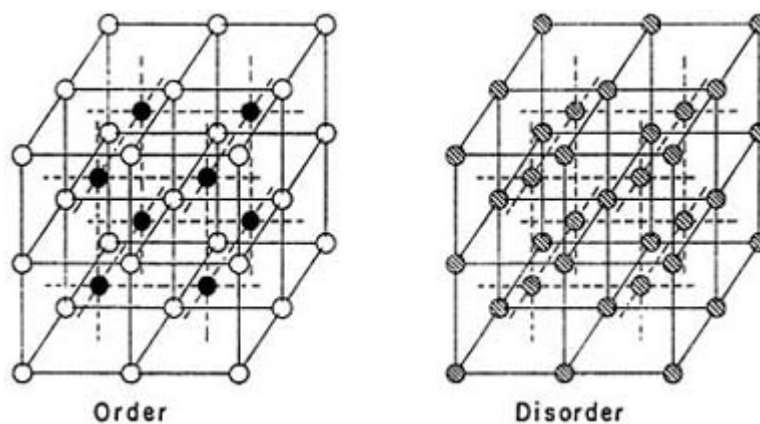


Figure A2.5.18. Body-centred cubic arrangement of β -brass (CuZn) at low temperature showing two interpenetrating simple cubic sublattices, one all Cu, the other all Zn, and a single lattice of randomly distributed atoms at high temperature. Reproduced from Hildebrand J H and Scott R L 1950 *The Solubility of Nonelectrolytes* 3rd edn (New York: Reinhold) p 342.

The treatment of such order–disorder phenomena was initiated by Gorsky (1928) and generalized by Bragg and Williams (1934) [5]. For simplicity we restrict the discussion to the symmetrical situation where there are equal amounts of each component ($x = 1/2$). The lattice is divided into two superlattices α and β , like those in the figure, and a degree of order s is defined such that the mole fraction of component B on superlattice β is $(1 + s)/4$ while that on superlattice α is $(1 - s)/4$. Conservation conditions then yield the mole fraction of A on the two superlattices

$$x_A^\alpha = x_B^\beta = (1 + s)/4 \quad \text{and} \quad x_A^\beta = x_B^\alpha = (1 - s)/4.$$

If the entropy and the enthalpy for the separate mixing in each of the half-mole superlattices are calculated and then combined, the following equation is obtained:

$$\Delta \bar{G}^M = RT \left[\frac{(1 + s)}{2} \ln \frac{(1 + s)}{2} + \frac{(1 - s)}{2} \ln \frac{(1 - s)}{2} \right] - \bar{N}w \frac{(1 + s)^2}{4}. \quad (\text{A2.5.18})$$

Note that equation (A2.5.18) is almost identical with [equation \(A2.5.15\)](#). Only the final term differs and then only by the sign preceding s^2 . Now, however, the interchange energy can be negative if the unlike attraction is stronger than the like attractions; then of course $K = \bar{N}w$ is also negative. If a reduced temperature T_r is defined as $-2RT / K$, a plot of $\Delta \bar{G}^M$ versus s for various T_r 's is identical to that in [figure A2.5.15](#). For all values of $K/(2RT)$ above -1 (i.e. $T_r > 1$), the minimum occurs at $s = 0$, corresponding to complete disorder when each superlattice is filled with equal amounts of A and B. However, for values below -1 , i.e. $T_r < 1$, the minimum occurs at nonzero values of s , values that increase with decreasing temperature. Recall that $K/(2RT) = +1$ defined the critical temperature for phase separation in a symmetrical binary mixture; here a value of -1 defines the limit of long-range ordering. Thus for order–disorder

-27-

behaviour $T_c = -K/2R$ defines a kind of critical temperature, although, by analogy with magnetic phenomena in solids, it is more often called the Curie point.

The free energy minimum is found by differentiating [equation \(A2.5.18\)](#) with respect to s at constant T and setting the derivative equal to zero. In its simplest form the resultant equation is

$$RT \ln[(1 + s)/(1 - s)] = 2RT \tanh^{-1} s = -Ks$$

exactly the same as [equation \(A2.5.13\)](#) for phase separation in simple mixtures except that this has $-Ks$ instead of $+Ks$. However, since it is a negative K that produces superlattice separation, the effect is identical, and [figure A2.5.15](#) and [figure A2.5.16](#) apply to both situations. The physical models are different, but the mathematics are just the same. This ‘disordering curve’, like the coexistence curve, is given by [equation \(A2.5.15\)](#) and is parabolic, and, for a temperature-independent K , the molar heat capacity \bar{C}_p for the equimolar alloy will be exactly the same as that for the simple mixture.

Other examples of order–disorder second-order transitions are found in the alloys CuPd and Fe₃Al. However, not all ordered alloys pass through second-order transitions; frequently the partially ordered structure changes to a disordered structure at a first-order transition.

Nix and Shockley [6] gave a detailed review of the status of order–disorder theory and experiment up to 1938, with emphasis on analytic improvements to the original Bragg–Williams theory, some of which will be

discussed later in [section A2.5.4.4](#).

A2.5.4.3 MAGNETISM

The magnetic case also turns out to be similar to that of fluids, as Curie and Weiss recognized early on, but later for a long period this similarity was overlooked by those working on fluids (mainly chemists) and by those working on magnetism (mainly physicists). In a ferromagnetic material such as iron, the magnetic interactions between adjacent atomic magnetic dipoles causes them to be aligned so that a region (a ‘domain’) has a substantial magnetic dipole. Ordinarily the individual domains are aligned at random, and there is no overall magnetization. However, if the sample is placed in a strong external magnetic field, the domains can be aligned and, if the temperature is sufficiently low, a ‘permanent magnet’ is made, permanent in the sense that the magnetization is retained even though the field is turned off. Above a certain temperature, the Curie temperature T_C , long-range ordering in domains is no longer possible and the material is no longer ferromagnetic, but only paramagnetic. Individual atoms can be aligned in a magnetic field, but all ordering is lost if the field is turned off. (The use of a subscript C for the Curie temperature should pose no serious confusion, since it is a kind of critical temperature too.)

The little atomic magnets are of course quantum mechanical, but Weiss’s original theory of paramagnetism and ferromagnetism (1907) [7] predated even the Bohr atom. He assumed that in addition to the external magnetic field B_0 , there was an additional internal ‘molecular field’ B_i proportional to the overall magnetization M of the sample,

-28-

$$B = B_0 + B_i = B_0 + \lambda M.$$

If this field is then substituted into the Curie law appropriate for independent dipoles one obtains

(A2.5.19)

where C is the Curie constant. The experimental magnetic susceptibility χ is defined as just M/B_0 , since the internal field cannot be measured. Rearrangement of equation (A2.5.19) leads to the result

$$\chi = M/B_0 = C/(T - C\lambda) = C/(T - T_C). \quad (\text{A2.5.20})$$

Equation (A2.5.20) is the Curie–Weiss law, and T_C , the temperature at which the magnetic susceptibility becomes infinite, is the Curie temperature. Below this temperature the substance shows spontaneous magnetization and is ferromagnetic. Normally the Curie temperature lies between 1 and 10 K. However, typical ferromagnetic materials like iron have very much larger values for quantum-mechanical reasons that will not be pursued here.

Equation (A2.5.19) and equation (A2.5.20) are valid only for small values of B_0 and further modelling is really not possible without some assumption, usually quantum mechanical, about the magnitude and orientation of the molecular magnets. This was not known to Weiss, but in the simplest case (half-integral spins), the magnetic dipole has the value of the Bohr magneton β_e , and the maximum possible magnetization M_{\max} when all the dipoles are aligned with the field is $N\beta_e/V$, where N/V is the number of dipoles per unit volume.

If an order parameter s is defined as M/M_{\max} , it can be shown that

$$s = \tanh[(s + \beta_e \mathbf{B}_0 / kT_C)(T_C / T)] = \tanh[(s + \mathbf{B}_r) / T_r]. \quad (\text{A2.5.21})$$

Isotherms of $\beta_e \mathbf{B}_0 / kT_C$, which might be called a reduced variable \mathbf{B}_r , versus s are shown in [figure A2.5.19](#) and look rather similar to the p_r, V_r plots for a fluid ([figure A2.5.6](#)). There are some differences, however, principally the symmetry that the fluid plots lack. At values of $T > T_C$, the curves are smooth and monotonic, but at T_C , as required, the magnetic susceptibility become infinite (i.e. the slope of \mathbf{B}_r versus s becomes horizontal).

-29-

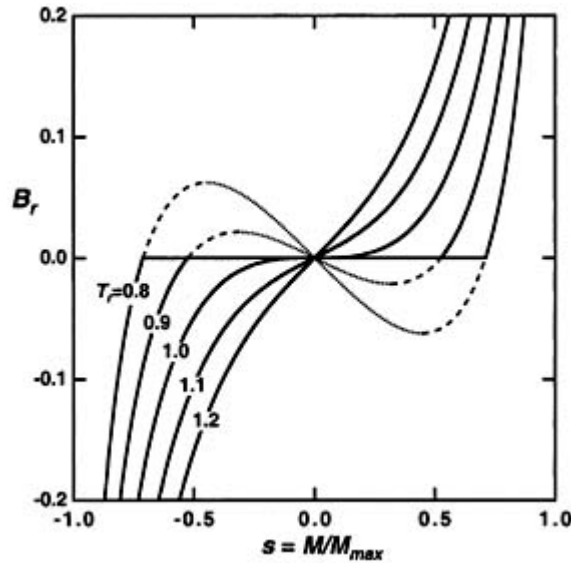


Figure A2.5.19. Isotherms showing the reduced external magnetic field $\mathbf{B}_r = \beta_e \mathbf{B}_0 / kT_C$ versus the order parameter $s = M/M_{\max}$, for various reduced temperatures $T_r = T/T_C$.

For $T < T_C$ ($T_r < 1$), however, the isotherms are S-shaped curves, reminiscent of the p_r, V_r isotherms that the van der Waals equation yields at temperatures below the critical ([figure A2.5.6](#)). As in the van der Waals case, the dashed and dotted portions represent metastable and unstable regions. For zero external field, there are two solutions, corresponding to two spontaneous magnetizations. In effect, these represent two ‘phases’ and the horizontal line is a ‘tie-line’. Note, however, that unlike the fluid case, even as shown in μ_r, ρ_r form ([figure A2.5.8](#)), the symmetry causes all the ‘tie-lines’ to lie on top of one another at $\mathbf{B}_r = 0$ ($\mathbf{B}_0 = 0$).

For $\mathbf{B}_0 = 0$, [equation \(A2.5.21\)](#) reduces to

$$s = \tanh(s / T_r)$$

which, while it looks somewhat different, is exactly the same as [equation \(A2.5.16\)](#) and yields exactly the same parabolic ‘coexistence curve’ as that from [equation \(A2.5.17\)](#). Experimentally, as we shall see in the next section, the curve is not parabolic, but more nearly cubic. More generally, [equation \(A2.5.21\)](#) may be used to plot T_r versus s for fixed values of \mathbf{B}_r as shown in [figure A2.5.20](#). The similarity of this to a typical phase diagram (T, ρ or T, x) is obvious. Note that for nonzero values of the external field \mathbf{B}_r the curves always lie outside the ‘two-phase’ region.

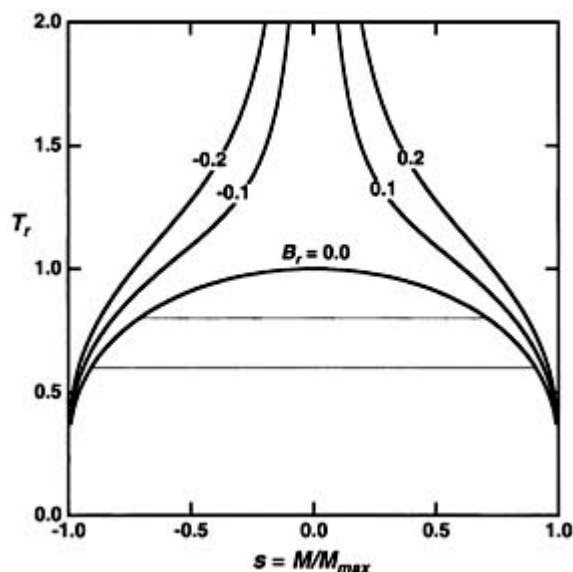


Figure A2.5.20. The reduced temperature $T_r = T/T_C$ versus the order parameter $s = M/M_{\max}$ for various values of the reduced magnetic field B_r . Note that for all nonzero values of the field the curves lie outside the ‘two-phase’ region.

Related to these ferromagnetic materials, but different, are antiferromagnetic substances like certain transition-metal oxides. In these crystals, there is a complicated three-dimensional structure of two interpenetrating superlattices not unlike those in CuZn. Here, at low temperatures, the two superlattices consist primarily of magnetic dipoles of opposite orientation, but above a kind of critical temperature, the Néel temperature T_N , all long-range order is lost and the two superlattices are equivalent. For $B_0 = 0$ the behaviour of an antiferromagnet is exactly analogous to that of a ferromagnet with a similar ‘coexistence curve’ $s(T_r)$, but for nonzero magnetic fields they are different. Unlike a ferromagnet at its Curie temperature, the susceptibility of an antiferromagnet does not diverge at the Néel temperature; extrapolation using the Curie–Weiss law yields a negative Curie temperature. Below the Néel temperature the antiferromagnetic crystal is anisotropic because there is a preferred axis of orientation. The magnetic susceptibility is finite, but varies with the angle between the crystal axis and the external field.

A related phenomenon with electric dipoles is ‘ferroelectricity’ where there is long-range ordering (nonzero values of the polarization P even at zero electric field E) below a second-order transition at a kind of critical temperature.

A2.5.4.4 MEAN FIELD VERSUS ‘MOLECULAR FIELD’

Apparently Weiss believed (although van der Waals did not) that the interactions between molecules were long-range and extended over the entire system; under such conditions, it was reasonable to assume that the energies could be represented as proportional to the populations of the various species. With the development of theories of intermolecular forces in the 1920s that showed that intermolecular interactions were usually very short-range, this view was clearly unrealistic. In the discussions of liquid and solid mixtures in the preceding sections it has been assumed that the principal interactions, or perhaps even the only ones, are between nearest neighbours; this led to energies proportional to the interchange energy w . It was therefore necessary to introduce what is clearly only an

approximation, that the probability of finding a particular molecular species in the nearest-neighbour shell (or

indeed any more distant shell) around a given molecule is simply the probability of finding that species in the entire system. This is the ‘mean-field’ approximation that underlies many of the early analytic theories.

However, one can proceed beyond this zeroth approximation, and this was done independently by Guggenheim (1935) with his ‘quasi-chemical’ approximation for simple mixtures and by Bethe (1935) for the order–disorder solid. These two approximations, which turned out to be identical, yield some enhancement to the probability of finding like or unlike pairs, depending on the sign of w and on the coordination number z of the lattice. (For the unphysical limit of z equal to infinity, they reduce to the mean-field results.)

The integral under the heat capacity curve is an energy (or enthalpy as the case may be) and is more or less independent of the details of the model. The quasi-chemical treatment improved the heat capacity curve, making it sharper and narrower than the mean-field result, but it still remained finite at the critical point. Further improvements were made by Bethe with a second approximation, and by Kirkwood (1938). [Figure A2.5.21](#) compares the various theoretical calculations [6]. These modifications lead to somewhat lower values of the critical temperature, which could be related to a flattening of the coexistence curve. Moreover, and perhaps more important, they show that a short-range order persists to higher temperatures, as it must because of the preference for unlike pairs; the excess heat capacity shows a discontinuity, but it does not drop to zero as mean-field theories predict. Unfortunately these improvements are still analytic and in the vicinity of the critical point still yield a parabolic coexistence curve and a finite heat capacity just as the mean-field treatments do.

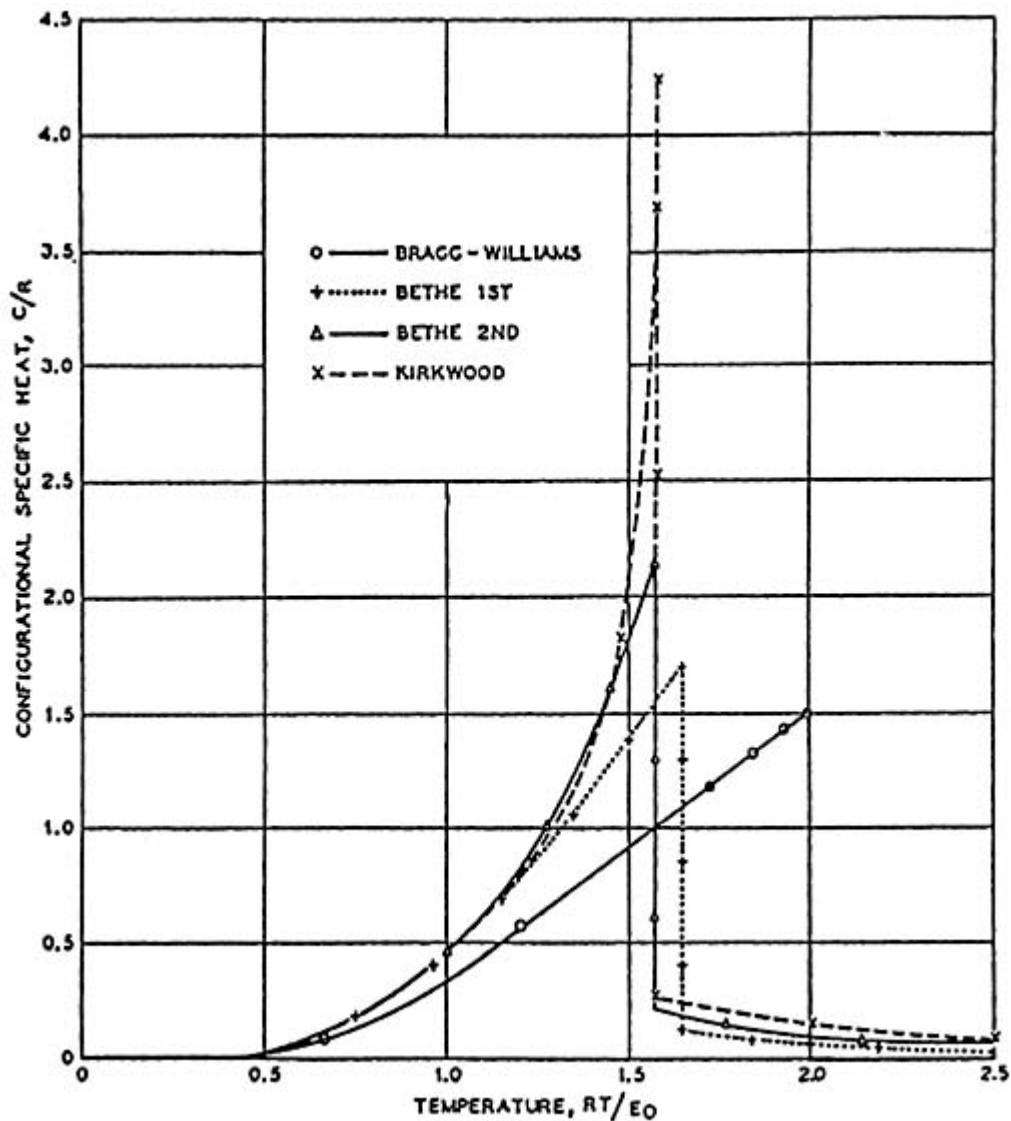


Figure A2.5.21. The heat capacity of an order–disorder alloy like β -brass calculated from various analytic treatments. Bragg–Williams (mean-field or zeroth approximation); Bethe-1 (first approximation also Guggenheim); Bethe-2 (second approximation); Kirkwood. Each approximation makes the heat capacity sharper and higher, but still finite. Reproduced from [6] Nix F C and Shockley W 1938 *Rev. Mod. Phys.* **10** 14, figure 13. Copyright (1938) by the American Physical Society.

Figure A2.5.22 shows [6] the experimental heat capacity of β -brass (CuZn) measured by Moser in 1934. Note that the experimental curve is sharper and goes much higher than any of the theoretical curves in figure A2.5.21 ; however, at that time it was still believed to have a finite limit.

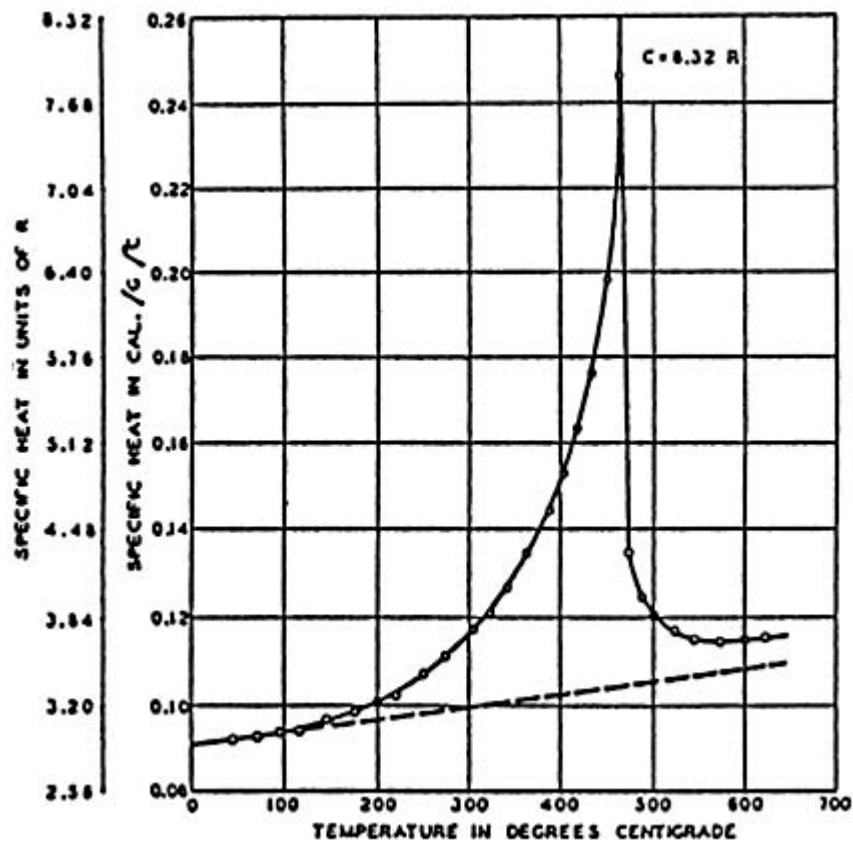


Figure A2.5.22. The experimental heat capacity of a β -brass (CuZn) alloy containing 48.9 atomic percent Zn as measured by Moser (1934). The dashed line is calculated from the specific heats of Cu and Zn assuming an ideal mixture. Reproduced from [6] Nix F C and Shockley W 1938 *Rev. Mod. Phys.* **10** 4, figure 4. Copyright (1938) by the American Physical Society.

A2.5.4.5 THE CRITICAL EXPONENTS

It has become customary to characterize various theories of critical phenomena and the experiments with which they are compared by means of the exponents occurring in certain relations that apply in the limit as the critical point is approached. In general these may be defined by the equation

$$E = \lim_{Y \rightarrow Y_c} \left[\frac{\partial \ln |X - X_c|}{\partial \ln |Y - Y_c|} \right]_{\text{path}}$$

where E is an exponent, X and Y are properties of the system, and the path along which the derivative is evaluated must be specified.

Exponents derived from the analytic theories are frequently called 'classical' as distinct from 'modern' or 'nonclassical' although this has nothing to do with 'classical' versus 'quantum' mechanics or 'classical' versus 'statistical' thermodynamics. The important thermodynamic exponents are defined here, and their classical values noted; the values of the more general nonclassical exponents, determined from experiment and theory, will appear in later sections. The equations are expressed in reduced units in order to compare the amplitude coefficients in subsequent sections.

(A) THE HEAT-CAPACITY EXPONENT A.

An exponent α governs the limiting slope of the molar heat capacity, variously \bar{C}_V , $\bar{C}_{p,x}$, or \bar{C}_M along a line through the critical point,

$$\bar{C}(\rho_c T_c / p_c) = A^\pm t^{-\alpha} + \dots \quad (\text{A2.5.22})$$

where the \pm recognizes that the coefficient A^+ for the function above the critical point will differ from the A^- below the critical point. A similar quantity is the thermal expansivity $\alpha_p = (\partial \ln V / \partial T)_p$. For all these analytic theories, as we have seen on pages 533 and 539, the heat capacity remains finite, so $\alpha = 0$. As we shall see, these properties actually diverge with exponents slightly greater than zero. Such divergences are called ‘weak’.

(B) THE COEXISTENCE-CURVE EXPONENT B .

In general the width of the coexistence line ($\Delta\rho$, Δx , or ΔM) is proportional to an order parameter s , and its absolute value may be written as

$$|(\rho - \rho_c) / \rho_c| = |s| = B t^\beta + \dots \quad (\text{A2.5.23})$$

As we have seen, all the analytic coexistence curves are quadratic in the limit, so for all these analytic theories, the exponent $\beta = 1/2$.

(C) THE SUSCEPTIBILITY EXPONENT Γ .

A third exponent γ , usually called the ‘susceptibility exponent’ from its application to the magnetic susceptibility χ in magnetic systems, governs what in pure-fluid systems is the isothermal compressibility κ_T , and what in mixtures is the osmotic compressibility, and determines how fast these quantities diverge as the critical point is approached (i.e. as $T_T \rightarrow 1$).

$$p_c \kappa_T = p_c (\partial \ln V / \partial p)_T = \Gamma^\pm t^{-\gamma} + \dots \quad (\text{A2.5.24})$$

For analytic theories, γ is simply 1, and we have seen that for the van der Waals fluid Γ^+ / Γ^- equals 2. Divergences with exponents of the order of magnitude of unity are called ‘strong’.

(D) THE CRITICAL-ISOTHERM EXPONENT Δ .

Finally the fourth exponent δ governs the limiting form of the critical isotherm, in the fluid case, simply

$$\quad (\text{A2.5.25})$$

Since all the analytic treatments gave cubic curves, their δ is obviously 3.

Exponent values derived from experiments on fluids, binary alloys, and certain magnets differ substantially from all those derived from analytic (mean-field) theories. However it is surprising that the experimental values appear to be the same from all these experiments, not only for different fluids and fluid mixtures, but indeed the same for the magnets and alloys as well (see section A2.5.5).

(E) THERMODYNAMIC INEQUALITIES.

Without assuming analyticity, but by applying thermodynamics, Rushbrooke (1963) and Griffiths (1964) derived general constraints relating the values of the exponents.

$$\alpha_2^- + 2\beta + \gamma_1^- \geq 2$$

$$\alpha_2^- + \beta(1 + \delta) \geq 2.$$

Here α_2^- is the exponent for the heat capacity measured along the critical isochore (i.e. in the two-phase region) below the critical temperature, while γ_1^- is the exponent for the isothermal compressibility measured in the one-phase region at the edge of the coexistence curve. These inequalities say nothing about the exponents α^+ and γ^+ in the one-phase region above the critical temperature.

Substitution of the classical values of the exponents into these equations shows that they satisfy these conditions as equalities.

A2.5.5 THE EXPERIMENTAL FAILURE OF THE ANALYTIC TREATMENT

Nearly all experimental ‘coexistence’ curves, whether from liquid–gas equilibrium, liquid mixtures, order–disorder in alloys, or in ferromagnetic materials, are far from parabolic, and more nearly cubic, even far below the critical temperature. This was known for fluid systems, at least to some experimentalists, more than one hundred years ago. Verschaffelt (1900), from a careful analysis of data (pressure–volume and densities) on isopentane, concluded that the best fit was with $\beta = 0.34$ and $\delta = 4.26$, far from the classical values. Van Laar apparently rejected this conclusion, believing that, at least very close to the critical temperature, the coexistence curve must become parabolic. Even earlier, van der Waals, who had derived a classical theory of capillarity with a surface-tension exponent of 3/2, found (1893)

-36-

that experimental results on three liquids yielded lower exponents (1.23–1.27); he too apparently expected that the discrepancy would disappear closer to the critical point. Goldhammer (1920) formulated a law of corresponding states for a dozen fluids assuming that the exponent β was 1/3. For reasons that are not entirely clear, this problem seems to have attracted little attention for decades after it was first pointed out. (This interesting history has been detailed by Levelt Sengers [8, 9].)

In 1945 Guggenheim [10], as part of an extensive discussion of the law of corresponding states, showed that, when plotted as reduced temperature T_r versus reduced density ρ_r , all the coexistence-curve measurements on three inert gases (Ar, Kr, Xe) fell on a single curve, and that Ne, N₂, O₂, CO and CH₄ also fit the same curve very closely. Moreover he either rediscovered or re-emphasized the fact that the curve was unequivocally cubic (i.e. $\beta = 1/3$) over the entire range of experimental temperatures, writing for ρ_r

$$\rho_r = 1 + (3/4)t \pm (7/4)t^{1/3}. \tag{A2.5.26}$$

Figure A2.5.23 reproduces Guggenheim’s figure, with experimental results and the fit to [equation \(A2.5.25\)](#).

It is curious that he never commented on the failure to fit the analytic theory even though that treatment—with the quadratic form of the coexistence curve—was presented in great detail in it *Statistical Thermodynamics* (Fowler and Guggenheim, 1939). The paper does not discuss any of the other critical exponents, except to fit the vanishing of the surface tension σ at the critical point to an equation

$$\sigma = \sigma_0 t^{11/9}.$$

This exponent 11/9, now called μ , is almost identical with that found by van der Waals in 1893.

-37-

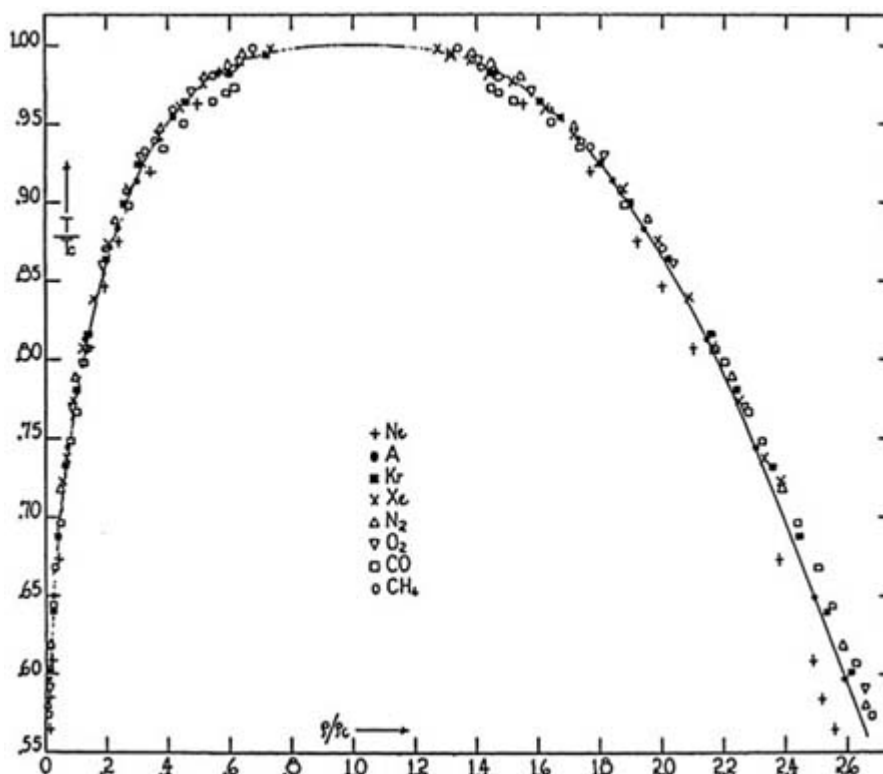


Figure A2.5.23. Reduced temperature $T_r = T/T_c$ versus reduced density $\rho_r = \rho/\rho_c$ for Ne, Ar, Kr, Xe, N_2 , O_2 , CO, and CH_4 . The full curve is the cubic equation (A2.5.26). Reproduced from [10], p 257 by permission of the American Institute of Physics.

In 1953 Scott [11] pointed out that, if the coexistence curve exponent was 1/3, the usual conclusion that the corresponding heat capacity remained finite was invalid. As a result the heat capacity might diverge and he suggested an exponent $\alpha = 1/3$. Although it is now known that the heat capacity does diverge, this suggestion attracted little attention at the time.

However, the discovery in 1962 by Voronel and coworkers [12] that the constant-volume heat capacity of argon showed a weak divergence at the critical point, had a major impact on uniting fluid criticality with that of other systems. They thought the divergence was logarithmic, but it is not quite that weak, satisfying equation (A2.5.21) with an exponent α now known to be about 0.11. The equation applies both above and below the critical point, but with different coefficients; A^- is larger than A^+ . Thus the heat capacity (figure A2.5.24) is quite asymmetrical around T_c and appears like a sharp discontinuity.

-38-

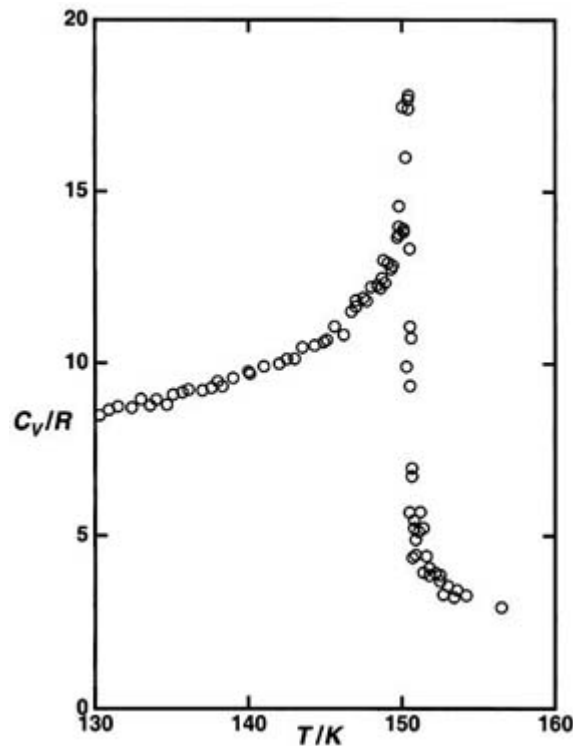


Figure A2.5.24. The heat capacity of argon in the vicinity of the critical point, as measured by Voronel and coworkers. Adapted from figure 1 of [12].

In 1962 Heller and Benedek made accurate measurements of the zero-field magnetization of the antiferromagnet MnF_2 as a function of temperature and reported a β of 0.335 ± 0.005 , a result supporting an experimental parallelism between fluids and magnets.

By 1966 the experimental evidence that the classical exponents were wrong was overwhelming and some significant theoretical advances had been made. In that year an important conference on critical phenomena [13] was held at the US National Bureau of Standards, which brought together physicists and chemists, experimentalists and theoreticians. Much progress had already been made in the preceding several years, and finally the similarity between the various kinds of critical phenomena was clearly recognized. The next decade brought near resolution to the problems.

A2.5.6 THE ISING MODEL AND THE GRADUAL SOLUTION OF THE PROBLEM

A2.5.6.1 THE ISING MODEL

In 1925 Ising [14] suggested (but solved only for the relatively trivial case of one dimension) a lattice model for magnetism in solids that has proved to have applicability to a wide variety of other, but similar, situations. The mathematical solutions, or rather attempts at solution, have made the Ising model one of the most famous problems in classical statistical mechanics.

The model is based on a classical Hamiltonian \mathcal{H} (here shown in script to distinguish it from the enthalpy H)

$$\mathcal{H} = - \sum_{i < j} J_{ij} \sigma_i \sigma_j - h \sum_i \sigma_i$$

where σ_i and σ_j are scalar numbers (+1 or -1) associated with occupancy of the lattice sites. In the magnetic case these are obviously the two orientations of the spin $s = 1/2$, but without any vector significance. The same Hamiltonian can be used for the lattice-solid mixture, where +1 signifies occupancy by molecule A, while -1 signifies a site occupied by molecule B (essentially the model used for the order-disorder transition in section A2.5.4.2). For the ‘lattice gas’, +1 signifies a site occupied by a molecule, while -1 signifies an unoccupied site (a ‘hole’).

The parameter J_{ij} is a measure of the energy of interaction between sites i and j while h is an external potential or field common to the whole system. The term $h \sum_i \sigma_i$ is a generalized work term (i.e. $-pV, \mu N, V\mathbf{B}_0\mathbf{M}$, etc), so \mathcal{H} is a kind of generalized enthalpy. If the interactions J are zero for all but nearest-neighbour sites, there is a single nonzero value for J , and then

$$\mathcal{H} = -J \sum_{nn, i < j} \sigma_i \sigma_j - h \sum_i \sigma_i.$$

Thus any nearest-neighbour pair with the same signs for σ (spins parallel) contributes a term $-J$ to the energy and hence to \mathcal{H} . Conversely any nearest-neighbour pair with opposite signs for σ (spins opposed) contributes $+J$ to the energy. (If this doesn’t seem right when extended to the lattice gas or to the lattice solid, it should be noted that a shift of the zero of energy resolves this problem and yields exactly the same equation. Thus, in the lattice mixture, there is only one relevant energy parameter, the interchange energy w .) What remained to be done was to derive the various thermodynamic functions from this simple Hamiltonian.

The standard analytic treatment of the Ising model is due to Landau (1937). Here we follow the presentation by Landau and Lifschitz [15], which casts the problem in terms of the order-disorder solid, but this is substantially the same as the magnetic problem if the vectors are replaced by scalars (as the Ising model assumes). The thermodynamic

-40-

potential, in this case $G(T, p, s)$, is expanded as a Taylor series in even powers of the order parameter s (because of the symmetry of the problem there are no odd powers)

$$G(T, p, s) = G_0 + G_2 s^2 + G_4 s^4 + G_6 s^6 + \dots$$

Here the coefficients G_2, G_4 , and so on, are functions of p and T , presumably expandable in Taylor series around $p - p_c$ and $T - T_c$. However, it is frequently overlooked that the derivation is accompanied by the comment that ‘since . . . the second-order transition point must be some singular point of the thermodynamic potential, there is every reason to suppose that such an expansion cannot be carried out up to terms of arbitrary order’, but that ‘there are grounds to suppose that its singularity is of higher order than that of the terms of the expansion used’. The theory developed below was based on this assumption.

For the kind of transition above which the order parameter is zero and below which other values are stable, the coefficient A_2 must change sign at the transition point and A_4 must remain positive. As we have seen, the dependence of s on temperature is determined by requiring the free energy to be a minimum (i.e. by setting its derivative with respect to s equal to zero). Thus

$$(\partial G/\partial s)_{T,p} = 2G_2s + 4G_4s^3 + 6G_6s^5 + \dots = 0.$$

If the G coefficients are expanded (at constant pressure p_c) in powers of t , this can be rewritten as

$$(-g_{21}t + g_{22}t^2 + \dots) + (g_{40} - g_{41}t + \dots)s^2 + (g_{60} + \dots)s^4 + \dots = 0.$$

Reverting this series and simplifying yields the final result in powers of t

$$s^2 = (g_{22}/g_{40})t - [(g_{22}g_{44}^2 - g_{21}g_{41}g_{40} + g_{41}^2g_{60})/(g_{40}^2g_{60})]t^2 + \dots \quad (\text{A2.5.27})$$

and we see that, like all the previous cases considered, this curve too is quadratic in the limit. (The derivation here has been carried to higher powers than shown in [15].) These results are more general than the analytic results in previous sections (in the sense that the coefficients are more general), but the basic conclusion is the same; moreover other properties like the heat capacity are also described in the analytic forms discussed in earlier sections. There is no way of explaining the discrepancies without abandoning the assumption of analyticity. (It is an interesting historical note that many Russian scientists were among the last to accept this failure; they were sure that Landau had to have been right, and ignored his stated reservations.)

That analyticity was the source of the problem should have been obvious from the work of Onsager (1944) [16] who obtained an exact solution for the two-dimensional Ising model in zero field and found that the heat capacity goes to infinity at the transition, a logarithmic singularity that yields $\alpha = 0$, but not the $\alpha = 0$ of the analytic theory, which corresponds to a finite discontinuity. (While diverging at the critical point, the heat capacity is symmetrical without an actual discontinuity, so perhaps should be called third-order.) Subsequently Onsager (1948) reported other exponents, and Yang (1952) completed the derivation. The exponents are rational numbers, but not the classical ones.

-41-

The ‘coexistence curve’ is nearly flat at its top, with an exponent $\beta = 1/8$, instead of the mean-field value of $1/2$. The critical isotherm is also nearly flat at T_C ; the exponent δ (determined later) is 15 rather than the 3 of the analytic theories. The susceptibility diverges with an exponent $\gamma = 7/4$, a much stronger divergence than that predicted by the mean-field value of 1.

The classical treatment of the Ising model makes no distinction between systems of different dimensionality, so, if it fails so badly for $d = 2$, one might have expected that it would also fail for $d = 3$. Landau and Lifschitz [15] discussed the Onsager and Yang results, but continued to emphasize the analytic conclusions for $d = 3$.

A2.5.6.2 THE ASSUMPTION OF HOMOGENEITY. THE ‘SCALING’ LAWS

The first clear step away from analyticity was made in 1965 by Widom [17] who suggested that the assumption of analytic functions be replaced by the less severe assumption that the singular part of the appropriate thermodynamic function was a homogeneous function of two variables, $(\rho_T - 1)$ and $(1 - T_T)$. A homogeneous function $f(u, v)$ of two variables is one that satisfies the condition

$$f(\lambda^{a_u}u, \lambda^{a_v}v) = \lambda f(u, v).$$

If one assumes that the singular part A^* of the Helmholtz free energy is such a function

$$A^*[\lambda^{a_\rho}(\rho_r - 1), \lambda^{a_T}(1 - T_r)] = \lambda[(\rho_r - 1), (1 - T_r)]$$

then a great deal follows. In particular, the reduced chemical potential $\mu_r = [\mu(\rho, T) - \mu(\rho_c, T)](\rho_c/p_c)$ of a fluid can be written as

$$\mu_r[\lambda^{a_\rho/(1-a_\rho)}(\rho_r - 1), \lambda^{a_T/(1-a_T)}(1 - T_r)] = \lambda\mu_r[(\rho_r - 1), (1 - T_r)].$$

(The brackets symbolize ‘function of’, not multiplication.) Since there are only two parameters, a_ρ and a_T , in this expression, the homogeneity assumption means that all four exponents α , β , γ and δ must be functions of these two; hence the inequalities in [section A2.5.4.5\(e\)](#) must be equalities. Equations for the various other thermodynamic quantities, in particular the singular part of the heat capacity C_V and the isothermal compressibility κ_T , may be derived from this equation for μ_r . The behaviour of these quantities as the critical point is approached can be satisfied only if

$$a_\rho = 1/(\delta + 1) = \beta/(2 - \alpha) \quad \text{and} \quad a_T = a_\rho/\beta.$$

This implies that μ_r may be written in a scaled form

$$\mu_r = [\mu(\rho, T) - \mu(\rho_c, T)](\rho_c/p_c) = (\rho_r - 1)|\rho_r - 1|^{\delta-1} Dh(x/x_0) \quad (\text{A2.5.28})$$

-42-

where $h(x/x_0)$ is an analytic function of $x = (T_r - 1)/|\rho_r - 1|^{1/\beta}$ and x_0 , the value of x at the critical point, $x_0 = B^{-1/\beta}$. The curve $x = -x_0$ is the coexistence curve, the curve $x = 0$ is the critical isotherm, and the curve $x = \infty$ is the critical isochore. All the rest of the thermodynamic behaviour in the critical region can be derived from this equation, with the appropriate exponents as functions of β and δ . Note that there are now not only just two independent exponents, but also only two independent amplitudes, B and D , the amplitudes in [equation \(A2.5.23\)](#) and [equation \(A2.5.25\)](#). This homogeneity assumption is now known as the ‘principle of two-scale-factor universality’. This principle, proposed as an approximation, seems to have stood the test of time; no further generalization seems to be needed. (We shall return to discuss exponents and amplitudes in [section A2.5.7.1](#)).

An unexpected conclusion from this formulation, shown in various degrees of generality in 1970–71, is that for systems that lack the symmetry of simple lattice models the slope of the diameter of the coexistence curve should have a weak divergence proportional to $t^{-\alpha}$. This is very hard to detect experimentally because it usually produces only a small addition to the classical linear term in the equation for the diameter

$$(\rho_l + \rho_g)/(2\rho_c) = \rho_d = 1 + A_{1-\alpha}t^{1-\alpha} + A_1t + \dots$$

However this effect was shown convincingly first [18] by Jüngst, Knuth and Hensel (1985) for the fluid metals caesium and rubidium (where the effect is surprisingly large) and then by Pestak *et al* (1987) for a series of simple fluids; [figure A2.5.25](#) shows the latter results [19]. Not only is it clear that there is curvature very close to the critical point, but it is also evident that for this reason critical densities determined by extrapolating a linear diameter may be significantly too high. The magnitude of the effect (i.e. the value of the coefficient $A_{1-\alpha}$), seems to increase with the polarizability of the fluid.

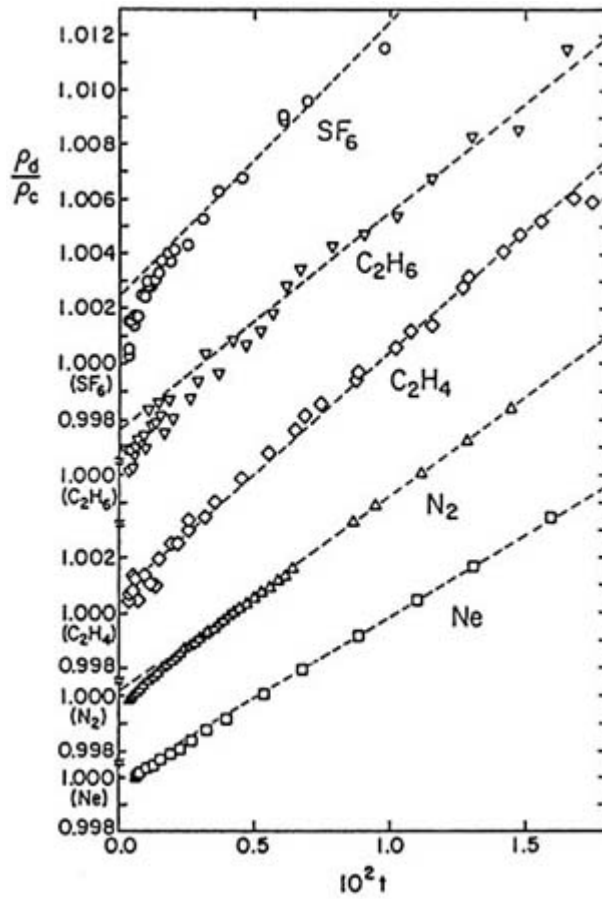


Figure A2.5.25. Coexistence-curve diameters as functions of reduced temperature for Ne, N_2 , C_2H_4 , C_2H_6 , and SF_6 . Dashed lines indicate linear fits to the data far from the critical point. Reproduced from [19] Pestak M W, Goldstein R E, Chan M H W, de Bruyn J R, Balzarini D A and Ashcroft N W 1987 *Phys. Rev. B* **36** 599, figure 3. Copyright (1987) by the American Physical Society.

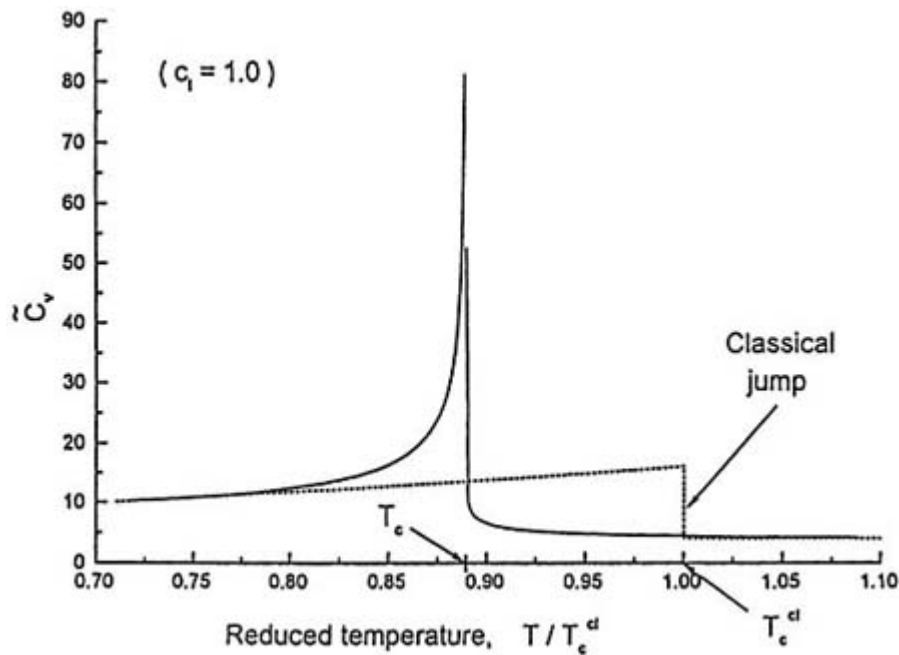


Figure A2.5.26. Molar heat capacity \bar{C}_V of a van der Waals fluid as a function of temperature: from mean-field theory (dotted line); from crossover theory (full curve). Reproduced from [29] Kostrowicka Wyczalkowska A, Anisimov M A and Sengers J V 1999 Global crossover equation of state of a van der Waals fluid *Fluid Phase Equilibria* **158–160** 532, figure 4, by permission of Elsevier Science.

A2.5.6.3 THE 'REASON' FOR THE NONANALYTICITY: FLUCTUATIONS

No system is exactly uniform; even a crystal lattice will have fluctuations in density, and even the Ising model must permit fluctuations in the configuration of spins around a given spin. Moreover, even the classical treatment allows for fluctuations; the statistical mechanics of the grand canonical ensemble yields an exact relation between the isothermal compressibility κ_T and the number of molecules N in volume V :

$$\sigma_N^2 = \langle N^2 \rangle - \langle N \rangle^2 = kT\kappa_T(N^2/V)$$

where σ is the standard deviation of the distribution of N 's, and the brackets indicate averages over the distribution.

If the finite size of the system is ignored (after all, N is probably 10^{20} or greater), the compressibility is essentially infinite at the critical point, and then so are the fluctuations. In reality, however, the compressibility diverges more sharply than classical theory allows (the exponent γ is significantly greater than 1), and thus so do the fluctuations.

Microscopic theory yields an exact relation between the integral of the radial distribution function $g(r)$ and the compressibility

$$RT\rho\kappa_T = 1 + \rho \int (g(r) - 1) dr = 1 + \rho \int h(r) dr$$

where $g(r)$ is the radial distribution function which is the probability density for finding a molecule a distance

r from the centre of a specified molecule, and $h(r)$ is the pair correlation function. At sufficiently long distances $g(r)$ must become unity while $h(r)$ must become zero. Since κ_T diverges at the critical point, so also must the integral. The only way the integral can diverge is for the integrand to develop a very long tail. The range of the fluctuations is measured by the correlation length ξ . Near, but not exactly at, the critical point, the behaviour of $h(r)$ can be represented by the Ornstein–Zernike (1914, 1916) equation

$$h(r) = \exp(-r/\xi)/r$$

while, at the critical point, $h(r) \propto 1/r^{d-2+\eta}$, where d is the dimensionality of the system and η is a very small number (zero classically). The correlation length ξ increases as the critical point is approached and it will ultimately diverge. On the critical isochore, $\rho = \rho_c$, one finds

$$\xi = \xi_0 |t|^{-\nu}$$

where classically $\nu = \gamma/2 = 1/2$. If the hypothesis of homogeneity is extended to the correlation length, what has become known as hyperscaling yields relations between the exponents ν and η and the thermodynamic exponents:

$$\nu = (2 - \alpha)/d \quad \text{and} \quad 2 - \eta = d(1 + \delta)/(1 - \delta).$$

Here d is the dimensionality of the system. (One recovers the analytic values with $d = 4$.)

Fluctuations in density and composition produce opalescence, a recognized feature of the critical region. Since systems very close to a critical point become visibly opaque, the fluctuations must extend over ranges comparable to the wavelength of light (i.e. to distances very much greater than molecular dimensions). Measurements of light scattering can yield quantitative information about the compressibility and thus about the magnitude of the fluctuations. Such measurements in the critical region showed the failure of the analytic predictions and yielded the first good experimental determinations of the exponent γ . As predicted even by classical theory the light scattering (i.e. the compressibility) on the critical isochore at a small temperature δT above the critical temperature is larger than that at the same δT below the critical temperature along the coexistence curve.

What this means is that mean-field (analytic) treatments fail whenever the range of correlations greatly exceeds the range of intermolecular forces. It follows that under these circumstances there should be no difference between the limiting behaviour of an Ising lattice and the nonlattice fluids; they should have the same exponents. Nearly a century after the introduction of the van der Waals equation for fluids, Kac, Uhlenbeck and Hemmer (1963) [20] proved that, in a one-dimensional system, it is exact for an intermolecular interaction that is infinite in range and infinitesimal in magnitude. (It is interesting to note that, in disagreement with van der Waals, Boltzmann insisted that the equation could only be correct if the range of the interactions were infinite.)

Moreover, well away from the critical point, the range of correlations is much smaller, and when this range is of the order of the range of the intermolecular forces, analytic treatments should be appropriate, and the exponents should be ‘classical’. The need to reconcile the nonanalytic region with the classical region has led to attempts to solve the ‘crossover’ problem, to be discussed in [section A2.5.7.2](#).

While there was a general recognition of the similarity of various types of critical phenomena, the situation was greatly clarified in 1970 by a seminal paper by Griffiths and Wheeler [21]. In particular the difference between variables that are ‘fields’ and those that are ‘densities’ was stressed. A ‘field’ is any variable that is the same in two phases at equilibrium, (e.g. pressure, temperature, chemical potential, magnetic field). Conversely a ‘density’ is a variable that is different in the two phases (e.g. molar volume or density, a composition variable like mole fraction or magnetization). The similarity between different kinds of critical phenomena is seen more clearly when the phase diagram is shown exclusively with field variables. (Examples of this are [figure A2.5.1](#) and [figure A2.5.11](#))

The field-density concept is especially useful in recognizing the parallelism of path in different physical situations. The criterion is the number of densities held constant; the number of fields is irrelevant. A path to the critical point that holds only fields constant produces a strong divergence; a path with one density held constant yields a weak divergence; a path with two or more densities held constant is nondivergent. Thus the compressibility κ_T of a one-component fluid shows a strong divergence, while C_V in the one-component fluid is comparable to C_{px} (constant pressure and composition) in the two-component fluid and shows a weak divergence.

The divergences of the heat capacity C_V and of the compressibility κ_T for a one-component fluid are usually defined as along the critical isochore, but if the phase diagram is shown in field space (p versus T as in [figure A2.5.1](#) or [figure A2.5.11](#)), it is evident that this is a ‘special’ direction along the vapour pressure curve. Indeed any direction that lies within the coexistence curve (e.g. constant enthalpy etc) and intersects that curve at the critical point will yield the same exponents. Conversely any path that intersects this special direction, such as the critical isobar, will yield different exponents. These other directions are not unique; there is no such thing as orthogonality in thermodynamics. Along the critical isobar, the compressibility divergence is still strong, but the exponent is reduced by renormalization from γ to $\gamma/\beta\delta$, nearly a 40% reduction. The weak divergence of C_V is reduced by a similar amount from α to $\alpha/\beta\delta$.

Another feature arising from field-density considerations concerns the coexistence curves. For one-component fluids, they are usually shown as temperature T versus density ρ , and for two-component systems, as temperature versus composition (e.g. the mole fraction x); in both cases one field is plotted against one density. However in three-component systems, the usual phase diagram is a triangular one at constant temperature; this involves two densities as independent variables. In such situations exponents may be ‘renormalized’ to higher values; thus the coexistence curve exponent may rise to $\beta/(1 - \alpha)$. (This ‘renormalization’ has nothing to do with the ‘renormalization group’ to be discussed in the next section.)

Finally the concept of fields permits clarification of the definition of the order of transitions [22]. If one considers a space of all fields (e.g. [Figure A2.5.1](#) but not [figure A2.5.3](#)), a first-order transition occurs where there is a discontinuity in the first derivative of one of the fields with respect to another (e.g. $(\partial\mu/\partial T)_p = -\bar{S}$ and $(\partial\mu/\partial p)_T = \bar{V}$), while a second-order transition occurs when the corresponding first derivative is continuous but the second is not and so on. Thus the Ehrenfest–Pippard definitions are preserved if the paths are not defined in terms of any densities.

A feature of a critical point, line, or surface is that it is located where divergences of various properties, in particular correlation lengths, occur. Moreover it is reasonable to assume that at such a point there is always an order parameter that is zero on one side of the transition and that becomes nonzero on the other side. Nothing of this sort occurs at a first-order transition, even the gradual liquid–gas transition shown in [figure A2.5.3](#) and [figure A2.5.4](#).

From 1965 on there was an extensive effort to calculate, or rather to estimate, the exponents for the Ising model. Initially this usually took the form of trying to obtain a low-temperature expansion (i.e. in powers of T) or a high-temperature expansion (i.e. in powers of $1/T$) of the partition function, in the hope of obtaining information about the ultimate form of the series, and hence to learn about the singularities at the critical point. Frequently this effort took the form of converting the finite series (sometimes with as many as 25 terms) into a Padé approximant, the ratio of two finite series. From this procedure, estimates of the various critical exponents (normally as the ratio of two integers) could be obtained. For the two-dimensional Ising model these estimates agreed with the values deduced by Onsager and Yang, which encouraged the belief that those for the three-dimensional model might be nearly correct. Indeed the $d = 3$ exponents estimated from theory were in reasonable agreement with those deduced from experiments close to the critical point. In this period much of the theoretical progress was made by Domb, Fisher, Kadanoff, and their coworkers.

In 1971 Wilson [23] recognized the analogy between quantum-field theory and the statistical mechanics of critical phenomena and developed a renormalization-group (RG) procedure that was quickly recognized as a better approach for dealing with the singularities at the critical point. New calculation methods were developed, one of which, expansion in powers of $\varepsilon = 4 - d$, where d is the dimension taken as a continuous variable, was first proposed by Wilson and Fisher (1972). These new procedures led to theoretical values of the critical exponents with much smaller estimates of uncertainty. The best current values are shown in [table A2.5.1](#) in [section A2.5.7.1](#). The RG method does assume, without proof, the homogeneity hypothesis and thus that the exponent inequalities are equalities. Some might wish that these singularities and exponents could be derived from a truly molecular statistical-mechanical theory; however, since the singular behaviour arises from the approach of the correlations to infinite distance, this does not seem likely in the foreseeable future. This history, including a final chapter on the renormalization group, is discussed in detail in a recent (1996) book by Domb [23].

-48-

Table A2.5.1 Ising model exponents.

Exponent	$d = 2$	$d = 3$	Classical ($d \geq 4$)
α Heat capacity, $\bar{C}_V, \bar{C}_{p, x}, \bar{C}_M$	0 (log)	0.109 ± 0.004	0 (finite jump)
β Coexistence, $\Delta\rho, \Delta x, \Delta M$	1/8	0.3258 ± 0.0014	1/2
γ Compressibility, $\kappa_T, \alpha_p, \kappa_{T, \mu}, \chi_T$	7/4	1.2396 ± 0.0013	1
δ Critical isotherm, $p(V), \mu(x), B_0(M)$	15	4.8047 ± 0.0044	3
ν Correlation length, ξ	1	0.6304 ± 0.0013	1/2
η Critical correlation function	1/4	0.0335 ± 0.0025	0

A2.5.6.6 EXTENDED SCALING. WEGNER CORRECTIONS

In 1972 Wegner [25] derived a power-series expansion for the free energy of a spin system represented by a Hamiltonian roughly equivalent to the scaled [equation \(A2.5.28\)](#), and from this he obtained power-series expansions of various thermodynamic quantities around the critical point. For example the compressibility

can be written as

$$\kappa_T = \kappa_T^0 + \Gamma_0 t^{-\gamma} + \Gamma_1 t^{-\gamma+\Delta_1} + \Gamma_2 t^{-\gamma+\Delta_2} + \dots$$

The new parameters in the exponents, Δ_1 and Δ_2 , are exactly or very nearly 0.50 and 1.00 respectively. Similar equations apply to the ‘extended scaling’ of the heat capacity and the coexistence curve for the determination of α and β .

The Wegner corrections have been useful in analysing experimental results in the critical region. The ‘correct’ exponents are the limiting values as T_r approaches unity, not the average values over a range of temperatures. Unfortunately the Wegner expansions do not converge very quickly (if they converge at all), so the procedure does not help in handling a crossover to the mean-field behaviour at lower temperatures where the correlation length is of the same order of magnitude as the range of intermolecular forces. A consistent method of handling crossover is discussed in [section A2.5.7.2](#).

A2.5.6.7 SOME EXPERIMENTAL PROBLEMS

The scientific studies of the early 1970s are full of concern whether the critical exponents determined experimentally, particularly those for fluids, could be reconciled with the calculated values, and at times it appeared that they could not be. However, not only were the theoretical values more uncertain (before RG calculations) than first believed, but also there were serious problems with the analysis of the experiments, in addition to those associated with the Wegner

-49-

corrections outlined above. Scott [26] has discussed in detail experimental difficulties with binary fluid mixtures, but some of the problems he cited apply to one-component fluids as well.

An experiment in the real world has to deal with gravitational effects. There will be gravity-induced density gradients and concentration gradients such that only at one height in an experimental cell will the system be truly at the critical point. To make matters worse, equilibration in the critical region is very slow. These problems will lead to errors of uncertain magnitude in the determination of all the critical exponents. For example, the observed heat capacity will not display an actual divergence because the total enthalpy is averaged over the whole cell and only one layer is at the critical point.

Another problem can be the choice of an order parameter for the determination of β and of the departure from linearity of the diameter, which should be proportional to $t^{1-\alpha}$. In the symmetrical systems, the choice of the order parameter s is usually obvious, and the symmetry enforces a rectilinear diameter. Moreover, in the one-component fluid, the choice of the reduced density ρ/ρ_c has always seemed the reasonable choice. However, for the two-component fluid, there are two order parameters, density and composition. It is not the density ρ that drives the phase separation, but should the composition order parameter be mole fraction x , volume fraction ϕ , or what? For the coexistence exponent β the choice is ultimately immaterial if one gets close enough to the critical temperature, although some choices are better than others in yielding an essentially cubic curve over a greater range of reduced temperature. (Try plotting the van der Waals coexistence curve against molar volume \bar{V} instead of density ρ .) However this ambiguity can have a very serious effect on any attempt to look for experimental evidence for departures from the rectilinear diameter in binary mixtures; an unwise choice for the order parameter can yield an exponent 2β rather than the theoretical $1 - \alpha$ (previously discussed in [section A2.5.6.2](#)) thus causing a much greater apparent departure from linearity.

A2.5.7 THE CURRENT STATUS OF THE ISING MODEL; THEORY AND EXPERIMENT

Before reviewing the current knowledge about Ising systems, it is important to recognize that there are non-Ising systems as well. A basic feature of the Ising model is that the order parameter is a scalar, even in the magnetic system of spin 1/2. If the order parameter is treated as a vector, it has a dimensionality n , such that $n = 1$ signifies a scalar (the Ising model), $n = 2$ signifies a vector with two components (the XY model), $n = 3$ signifies a three-component vector (the Heisenberg model), $n = \infty$ is an unphysical limit to the vector concept (the so-called spherical model), and $n \rightarrow 0$ is a curious mathematical construct that seems to fit critical phenomena in some polymer equilibria. Some of these models will be discussed in subsequent sections, but first we limit ourselves to the Ising model.

-50-

A2.5.7.1 THE ISING EXPONENTS AND AMPLITUDES

There is now consensus on some questions about which there had been lingering doubts.

- (a) There is now agreement between experiment and theory on the Ising exponents. Indeed it is now reasonable to assume that the theoretical values are better, since their range of uncertainty is less.
- (b) There is no reason to doubt that the inequalities of [section A2.5.4.5\(e\)](#) are other than equalities. The equalities are assumed in most of the theoretical calculations of exponents, but they are confirmed (within experimental error) by the experiments.
- (c) The exponents apply not only to solid systems (e.g. order–disorder phenomena and simple magnetic systems), but also to fluid systems, regardless of the number of components. (As we have seen in [section A2.5.6.4](#) it is necessary in multicomponent systems to choose carefully the variable to which the exponent is appropriate.)
- (d) There is no distinction between the exponents above and below the critical temperature. Thus $\gamma^+ = \gamma^- = \gamma$ and $\nu^+ = \nu^- = \nu$. However, there is usually a significant difference in the coefficients above and below (e.g. A^+ and A^-); this produces the discontinuities at the critical point.

Many of the earlier uncertainties arose from apparent disagreements between the theoretical values and experimental determinations of the critical exponents. These were resolved in part by better calculations, but mainly by measurements closer and closer to the critical point. The analysis of earlier measurements assumed incorrectly that the measurements were close enough. (Van der Waals and van Laar were right that one needed to get closer to the critical point, but were wrong in expecting that the classical exponents would then appear.) As was shown in [section A2.5.6.7](#), there are additional contributions from ‘extended’ scaling.

Moreover, some uncertainty was expressed about the applicability to fluids of exponents obtained for the Ising lattice. Here there seemed to be a serious discrepancy between theory and experiment, only cleared up by later and better experiments. By hindsight one should have realized that long-range fluctuations should be independent of the presence or absence of a lattice.

[Table A2.5.1](#) shows the Ising exponents for two and three dimensions, as well as the classical exponents. The uncertainties are those reported by Guida and Zinn-Justin [27]. These exponent values satisfy the equalities (as they must, considering the scaling assumption) which are here reprised as functions of β and γ :

$$\begin{aligned}\alpha &= 2 - 2\beta - \gamma \\ \delta &= (\beta + \gamma)/\beta \\ \nu &= (2\beta + \gamma)/d \\ \eta &= 2 - d\beta\gamma/(2\beta + \gamma) = 2 - \gamma/\nu.\end{aligned}$$

The small uncertainties in the calculated exponents seem to preclude the possibility that the $d = 3$ exponents are rational numbers (i.e. the ratio of integers). (At an earlier stage this possibility had been suggested, since not only the classical exponents, but also the $d = 2$ exponents are rational numbers; pre-RG calculations had suggested $\beta = 5/16$ and $\gamma = 5/4$.)

-51-

As noted earlier in [section A2.5.6.2](#), the assumption of homogeneity and the resulting principle of two-scale-factor universality requires the amplitude coefficients to be related. In particular the following relations can be derived:

$$\begin{aligned}\alpha A^+ \Gamma^+ / B^2 &= 0.0574 \pm 0.0020 \\ \Gamma^+ D B^{\delta-1} &= 1.669 \pm 0.018 \\ A^+ / A^- &= 0.537 \pm 0.019 \\ \Gamma^+ / \Gamma^- &= 4.79 \pm 0.10.\end{aligned}$$

These numerical values come from theory [26] and are in good agreement with recent experiments.

A2.5.7.2 CROSSOVER FROM MEAN-FIELD TO THE CRITICAL REGION

At temperatures well below the critical region one expects a mean-field treatment (or at least a fully analytic one) to be applicable, since the correlations will be short range. In the critical region, as we have seen, when the correlation length becomes far greater than the range of intermolecular forces, the mean-field treatment fails. Somewhere between these two limits the treatment of the problem has to ‘cross over’. Early attempts to bridge the gap between the two regimes used switching functions, and various other solutions have been proposed. A reasonably successful treatment has been developed during the past few years by Anisimov and Sengers and their collaborators. (Detailed references will be found in a recent review chapter [28].)

As a result of long-range fluctuations, the local density will vary with position; in the classical Landau–Ginzburg theory of fluctuations this introduces a gradient term. A Ginzburg number N_G is defined (for a three-dimensional Ising system) as proportional to a dimensionless parameter ξ_0^6/v_0^2 which may be regarded as the inverse sixth power of a normalized interaction range. (ξ_0 is the coefficient of the correlation length equation in [section A2.5.6.3](#) and v_0 is a molecular volume.) The behaviour of the fluid will be nonanalytic (Ising-like) when $\tau = (T_c - T)/T = t/(1 - t)$ is much smaller than N_G , while it is analytic (van der Waals-like) when τ is much greater than N_G . A significant result of this recent research is that the free energy can be rescaled to produce a continuous function over the whole range of temperatures.

For simple fluids N_G is estimated to be about 0.01, and Kostrowicka Wyczalkowska *et al* [29] have used this to apply crossover theory to the van der Waals equation with interesting results. The critical temperature T_c is reduced by 11% and the coexistence curve is of course flattened to a cubic. The critical density ρ_c is almost unchanged (by 2%), but the critical pressure p_c is reduced greatly by 38%. These changes reduce the critical

compression factor $(p\bar{V}/RT)_c$ from 3.75 to 2.6; the experimental value for argon is 2.9. The molar heat capacity \bar{C}_V for the classical van der Waals fluid and the crossover van der Waals fluid are compared in [figure A2.5.26](#).

Povodyrev *et al* [30] have applied crossover theory to the Flory equation ([section A2.5.4.1](#)) for polymer solutions for various values of N , the number of monomer units in the polymer chain, obtaining the coexistence curve and values of the coefficient β_{eff} from the slope of that curve. [Figure A2.5.27](#) shows their comparison between classical and crossover values of β_{eff} for $N = 1$, which is of course just the simple mixture. As seen in this figure, the crossover to classical behaviour is not complete until far below the critical temperature.

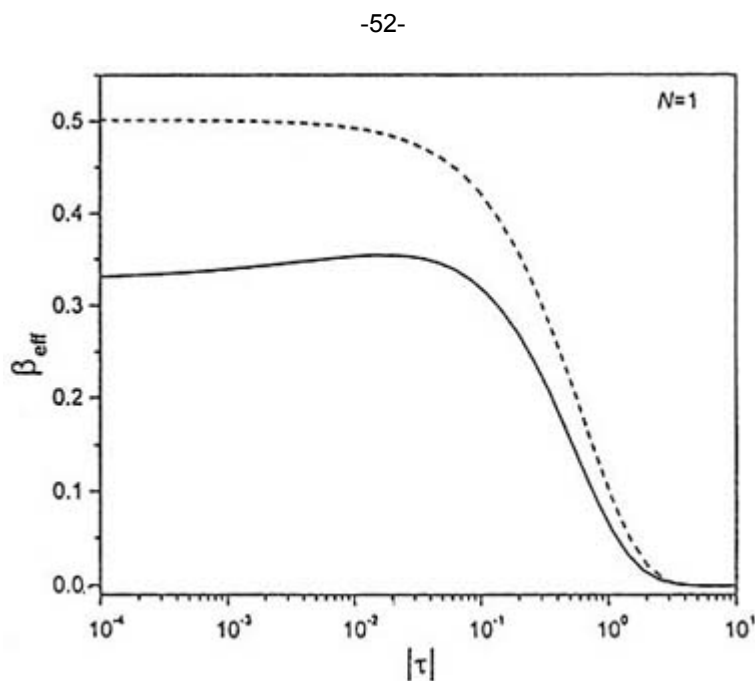


Figure A2.5.27. The effective coexistence curve exponent $\beta_{\text{eff}} = d \ln x / d \ln \tau$ for a simple mixture ($N = 1$) as a function of the temperature parameter $\tau = t / (1 - t)$ calculated from crossover theory and compared with the corresponding curve from mean-field theory (i.e. from [figure A2.5.15](#)). Reproduced from [30], Povodyrev A A, Anisimov M A and Sengers J V 1999 Crossover Flory model for phase separation in polymer solutions *Physica A* **264** 358, figure 3, by permission of Elsevier Science.

Sengers and coworkers (1999) have made calculations for the coexistence curve and the heat capacity of the real fluid SF_6 and the real mixture 3-methylpentane + nitroethane and the agreement with experiment is excellent; their comparison for the mixture [28] is shown in [figure A2.5.28](#).

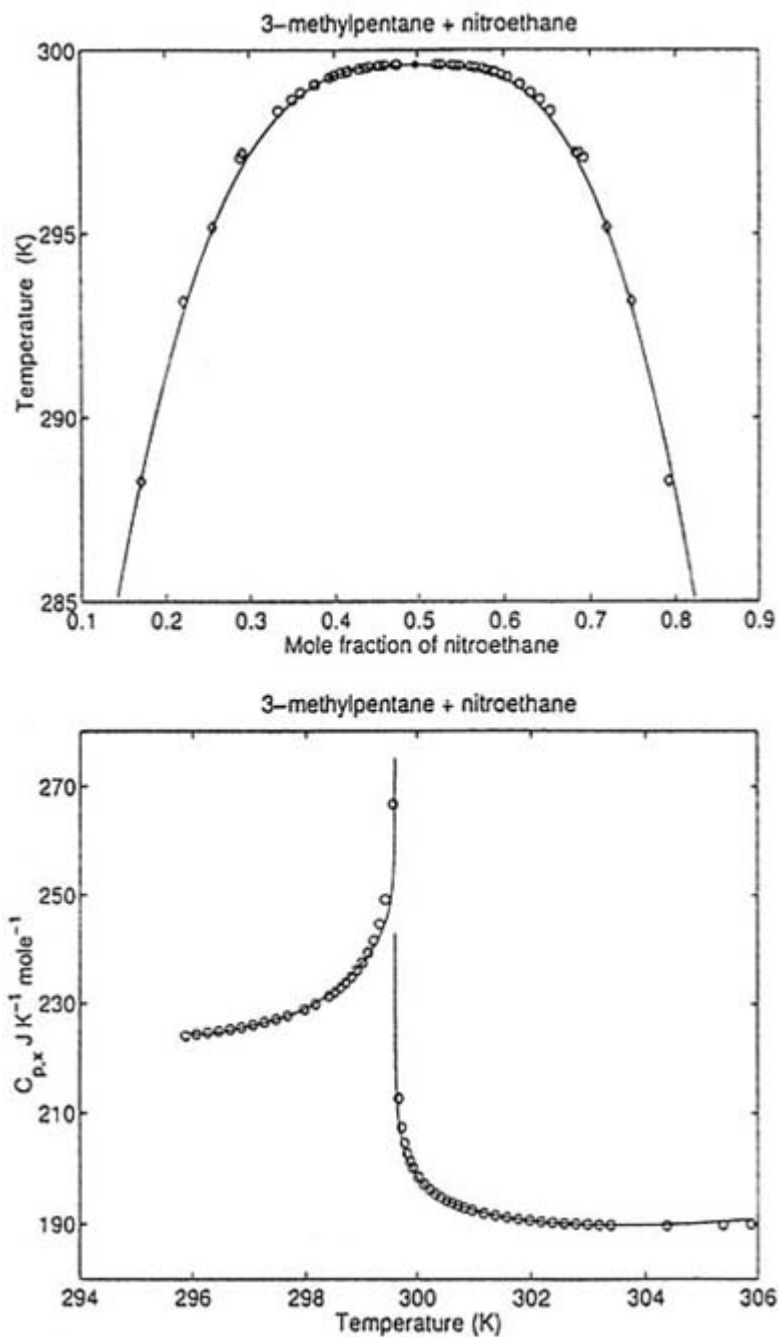


Figure A2.5.28. The coexistence curve and the heat capacity of the binary mixture 3-methylpentane + nitroethane. The circles are the experimental points, and the lines are calculated from the two-term crossover model. Reproduced from [28], 2000 *Supercritical Fluids—Fundamentals and Applications* ed E Kiran, P G Debenedetti and C J Peters (Dordrecht: Kluwer) Anisimov M A and Sengers J V Critical and crossover phenomena in fluids and fluid mixtures, p 16, figure 3, by kind permission from Kluwer Academic Publishers.

However, for more complex fluids such as high-polymer solutions and concentrated ionic solutions, where the range of intermolecular forces is much longer than that for simple fluids and N_G is much smaller, mean-field behaviour is observed much closer to the critical point. Thus the crossover is sharper, and it can also be nonmonotonic.

A2.5.8 OTHER EXAMPLES OF SECOND-ORDER TRANSITIONS

There are many other examples of second-order transitions involving critical phenomena. Only a few can be mentioned here.

A2.5.8.1 TWO-DIMENSIONAL ISING SYSTEMS

No truly two-dimensional systems exist in a three-dimensional world. However monolayers adsorbed on crystalline or fluid surfaces offer an approximation to two-dimensional behaviour. Chan and coworkers [31] have measured the coexistence curve for methane adsorbed on graphite by an ingenious method of determining the maximum in the heat capacity at various coverages. The coexistence curve (figure A2.5.29) is fitted to $\beta = 0.127$, very close to the theoretical $1/8$. A 1992 review [32] summarizes the properties of rare gases on graphite.

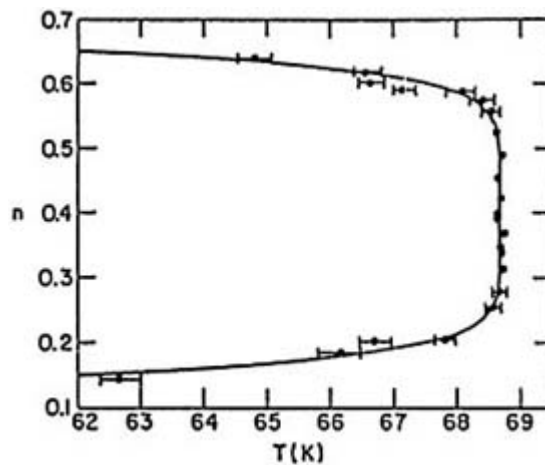


Figure A2.5.29. Peak positions of the liquid–vapour heat capacity as a function of methane coverages on graphite. These points trace out the liquid–vapour coexistence curve. The full curve is drawn for $\beta = 0.127$. Reproduced from [31] Kim H K and Chan M H W *Phys. Rev. Lett.* **53** 171 (1984) figure 2. Copyright (1984) by the American Physical Society.

A2.5.8.2 THE XY MODEL ($N = 2$)

If the scalar order parameter of the Ising model is replaced by a two-component vector ($n = 2$), the XY model results. An important example that satisfies this model is the λ -transition in helium, from superfluid helium-II to ordinary liquid helium, occurring for the isotope ^4He and for mixtures of ^4He with ^3He . (This is the transition at 1.1 K, not the

liquid–gas critical point at 5.2 K, which is Ising.) Calculations indicate that at the $n = 2$ transition, the heat capacity exponent α is very small, but negative. If so, the heat capacity does not diverge, but rather reaches a maximum just at the λ -point, as shown in the following equation:

$$\bar{C}(n = 2, d = 3) = \bar{C}_{\max} - At^{-\alpha}$$

where \bar{C}_{\max} is the value at the λ -transition. At first this prediction was hard to distinguish experimentally from a logarithmic divergence but experiments in space under conditions of microgravity by Lipa and

coworkers (1996) have confirmed it [33] with an $\alpha = -0.01285$, a value within the limits of uncertainty of the theoretical calculations. The results above and below the transition were fitted to the same value of \bar{C}_{\max} and α but with $A^+/A^- = 1.054$. Since the heat capacity is finite and there is no discontinuity, this should perhaps be called a third-order transition.

The liquid-crystal transition between smectic-A and nematic for some systems is an XY transition. Depending on the value of the MacMillan ratio, the ratio of the temperature of the smectic-A-nematic transition to that of the nematic-isotropic transition (which is Ising), the behaviour of such systems varies continuously from a λ -type transition to a tricritical one (see section A2.5.9). Garland and Nounesis [34] reviewed these systems in 1994.

A2.5.8.3 THE HEISENBERG MODEL ($N = 3$)

While the behaviour of some magnetic systems is Ising-like, others require a three-dimensional vector. In the limit of where the value of the quantum number J goes to infinity (i.e. where all values of the magnetic quantum number M are possible), the Heisenberg model ($n = 3$) applies. The exponents β and γ are somewhat larger than the Ising or XY values; the exponent α is substantially negative (about -0.12).

A2.5.8.4 POLYMERIZATION SYSTEMS ($N \rightarrow 0$)

Some equilibrium polymerizations are such that over a range of temperatures only the monomer exists in any significant quantity, but below or above a unique temperature polymers start to form in increasing number. Such a polymerization temperature is a critical point, another kind of second-order transition. The classic example is that of the ring-chain transition in sulfur, but more recently similar behaviour has been found in a number of 'living polymers'. Wheeler and coworkers [35] have shown that these systems can best be treated as examples of the mathematical limit of the n -vector model with $n \rightarrow 0$. The heat capacity in such a system diverges more strongly than that of an Ising system ($\alpha = 0.235$ [27]); the heat capacity of sulfur fits the model qualitatively, but there are chemical complications.

Mixtures of such polymeric substances with solvents show a line of critical points that in theory end at a tricritical point. (See section A2.5.9 for further discussion of tricritical phenomena.)

A2.5.8.5 SUPERCONDUCTIVITY

Alone among all known physical phenomena, the transition in low-temperature ($T_c < 25$ K) superconducting materials (mainly metals and alloys) retains its classical behaviour right up to the critical point; thus the exponents are the analytic ones. Unlike the situation in other systems, such superconducting interactions are truly long range and thus

-56-

mean field. For the newer high-temperature superconducting materials, the situation is different. These substances crystallize in structures that require a two-component order parameter and show XY behaviour, usually three dimensional (i.e. $n = 2, d = 3$). Pasler *et al* [36] have measured the thermal expansivity of $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$ and have found the exponent α to be 0 ± 0.018 , which is consistent with the small negative value calculated for the XY model and found for the λ -transition in helium.

A2.5.9 MULTICRITICAL POINTS

An ordinary critical point such as those discussed in earlier sections occurs when two phases become more

and more nearly alike and finally become one. Because this involves two phases, it is occasionally called a ‘bicritical point’. A point where three phases simultaneously become one is a ‘tricritical point’. There are two kinds of tricritical points, symmetrical and unsymmetrical; there is a mathematical similarity between the two, but the physical situation is so different that they need to be discussed quite separately. One feature that both kinds have in common is that the dimension at and above which modern theory yields agreement between ‘classical’ and ‘nonclassical’ treatments is $d = 3$, so that analytic treatments (e.g. mean-field theories) are applicable to paths leading to tricritical points, unlike the situation with ordinary critical points where the corresponding dimension is $d = 4$. (In principle there are logarithmic corrections to these analytic predictions for $d = 3$, but they have never been observed directly in experiments.)

A 1984 volume reviews in detail theories and experiments [37] on multicritical points; some important papers have appeared since that time.

A2.5.9.1 SYMMETRICAL TRICRITICAL POINTS

In the absence of special symmetry, the phase rule requires a minimum of three components for a tricritical point to occur. Symmetrical tricritical points do have such symmetry, but it is easiest to illustrate such phenomena with a true ternary system with the necessary symmetry. A ternary system comprised of a pair of enantiomers (optically active *d*- and *l*-isomers) together with a third optically inert substance could satisfy this condition. While liquid–liquid phase separation between enantiomers has not yet been found, ternary phase diagrams like those shown in [figure A2.5.30](#) can be imagined; in these diagrams there is a necessary symmetry around a horizontal axis that represents equal amounts of the two enantiomers.

-57-

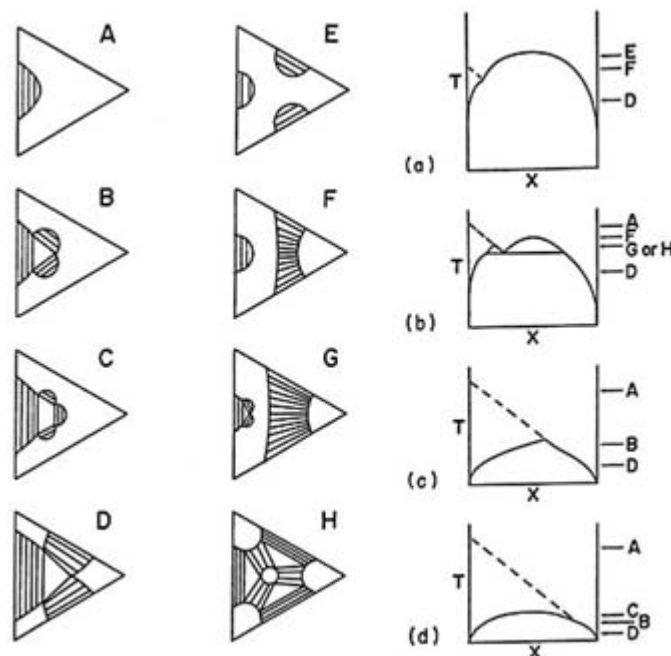


Figure A2.5.30. Left-hand side: Eight hypothetical phase diagrams (A through H) for ternary mixtures of *d*- and *l*-enantiomers with an optically inactive third component. Note the symmetry about a line corresponding to a racemic mixture. Right-hand side: Four T, x diagrams ((a) through (d)) for ‘pseudobinary’ mixtures of a racemic mixture of enantiomers with an optically inactive third component. Reproduced from [37] 1984 *Phase Transitions and Critical Phenomena* ed C Domb and J Lebowitz, vol 9, ch 2, Knobler C M and Scott R L Multicritical points in fluid mixtures. Experimental studies pp 213–14, (Copyright 1984) by permission of the publisher Academic Press.

Now consider such a symmetrical system, that of a racemic mixture of the enantiomers plus the inert third component. A pair of mirror-image conjugate phases will not physically separate or even become turbid, since they have exactly the same density and the same refractive index. Unless we find evidence to the contrary, we might conclude that this is a binary mixture with a T, x phase diagram like one of those on the right-hand side of figure A2.5.30. In particular any symmetrical three-phase region will have to shrink symmetrically, so it may disappear at a tricritical point, as shown in two of the four ‘pseudobinary’ diagrams. The dashed lines in these diagrams are two-phase critical points, and will show the properties of a second-order transition. Indeed, a feature of these diagrams is that with increasing temperature, a first-order transition ends at a tricritical point that is followed by a second-order transition line. (This is even more striking if the phase diagram is shown in field space as a p, T or μ, T diagram.)

These unusual ‘pseudobinary’ phase diagrams were derived initially by Meijering (1950) from a ‘simple mixture’ model for ternary mixtures. Much later, Blume, Emery and Griffiths (1971) deduced the same diagrams from a three-spin model of helium mixtures. The third diagram on the right of figure A2.5.30 is essentially that found experimentally for the fluid mixture $^4\text{He} + ^3\text{He}$; the dashed line (second-order transition) is that of the λ -transition.

Symmetrical tricritical points are predicted for fluid mixtures of sulfur or living polymers in certain solvents. Scott (1965) in a mean-field treatment [38] of sulfur solutions found that a second-order transition line (the critical

-58-

polymerization line) ended where two-phase separation of the polymer and the solvent begins; the theory yields a tricritical point at that point. Later Wheeler and Pfeuty [39] extended their $n \rightarrow 0$ treatment of equilibrium polymerization to sulfur solutions; in mean field their theory reduces to that of Scott, and the predictions from the nonclassical formulation are qualitatively similar. The production of impurities by slow reaction between sulfur and the solvent introduces complications; it can eliminate the predicted three-phase equilibrium, flatten the coexistence curve and even introduce an unsymmetrical tricritical point.

Symmetrical tricritical points are also found in the phase diagrams of some systems forming liquid crystals.

A2.5.9.2 UNSYMMETRICAL TRICRITICAL POINTS

While, in principle, a tricritical point is one where three phases simultaneously coalesce into one, that is not what would be observed in the laboratory if the temperature of a closed system is increased along a path that passes exactly through a tricritical point. Although such a difficult experiment is yet to be performed, it is clear from theory (Kaufman and Griffiths 1982, Pegg *et al* 1990) and from experiments in the vicinity of tricritical points that below the tricritical temperature T_t only two phases coexist and that the volume of one shrinks precipitously to zero at T_t .

While the phase rule requires three components for an unsymmetrical tricritical point, theory can reduce this requirement to two components with a continuous variation of the interaction parameters. Lindh *et al* (1984) calculated a phase diagram from the van der Waals equation for binary mixtures and found (in accord with [figure A2.5.13](#) that a tricritical point occurred at sufficiently large values of the parameter ζ (a measure of the difference between the two components).

One can effectively reduce the three components to two with ‘quasibinary’ mixtures in which the second component is a mixture of very similar higher hydrocarbons. [Figure A2.5.31](#) shows a phase diagram [40] calculated from a generalized van der Waals equation for mixtures of ethane ($n_1 = 2$) with normal hydrocarbons of different carbon number n_2 (treated as continuous). It is evident that, for some values of the parameter n_2 , those to the left of the tricritical point at $n_2 = 16.48$, all that will be observed with increasing

temperature is a two-phase region ($\alpha + \beta$) above which only the β phase exists. Conversely, for larger values of n_2 , those to the right of the tricritical point, increasing the temperature takes the system from the two-phase region ($\alpha + \beta$) through a narrow three-phase region ($\alpha + \beta + \gamma$) to a different two-phase region ($\beta + \gamma$).

-59-

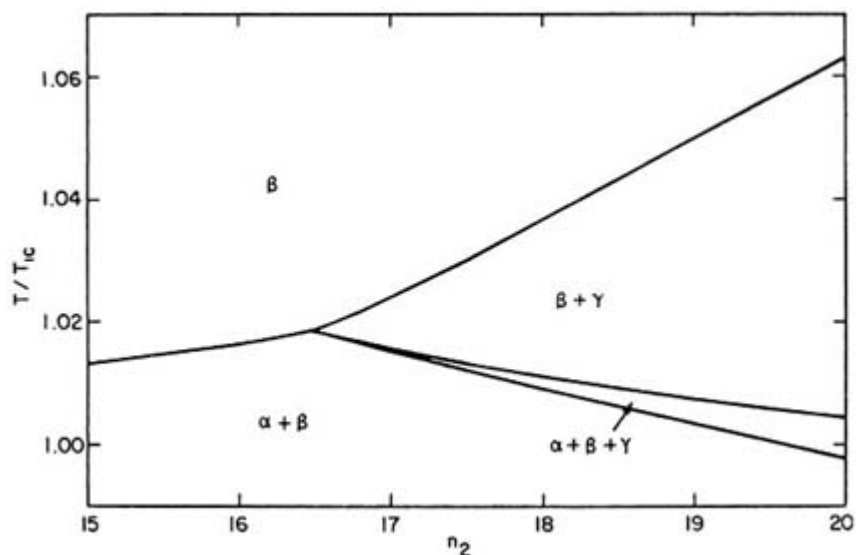


Figure A2.5.31. Calculated $T/T_{1c}, n_2$ phase diagram in the vicinity of the tricritical point for binary mixtures of ethane ($n_1 = 2$) with a higher hydrocarbon of continuous n_2 . The system is in a sealed tube at fixed tricritical density and composition. The tricritical point is at the confluence of the four lines. Because of the fixing of the density and the composition, the system does not pass through critical end points; if the critical end-point lines were shown, the three-phase region would be larger. An experiment increasing the temperature in a closed tube would be represented by a vertical line on this diagram. Reproduced from [40], figure 8, by permission of the American Institute of Physics.

Most of the theoretical predictions have now been substantially verified by a large series of experiments in a number of laboratories. Knobler and Scott and their coworkers (1977–1991) have studied a number of quasibinary mixtures, in particular ethane + (hexadecane + octadecane) for which the experimental $n_2 = 17.6$. Their experimental results essentially confirm the theoretical predictions shown in figure A2.5.31.

A2.5.9.3 HIGHER-ORDER CRITICAL POINTS

Little is known about higher order critical points. Tetracritical points, at least unsymmetrical ones, require four components. However for tetracritical points, the crossover dimension $d = 2$, so any treatment can surely be mean-field, or at least analytic.

A2.5.10 HIGHER-ORDER PHASE TRANSITIONS

We have seen in previous sections that the two-dimensional Ising model yields a symmetrical heat capacity curve that is divergent, but with no discontinuity, and that the experimental heat capacity at the λ -transition of helium is finite without a discontinuity. Thus, according to the Ehrenfest–Pippard criterion these transitions might be called third-order.

-60-

It has long been known from statistical mechanical theory that a Bose–Einstein ideal gas, which at low temperatures would show condensation of molecules into the ground translational state (a condensation in momentum space rather than in position space), should show a third-order phase transition at the temperature at which this condensation starts. Normal helium (^4He) is a Bose–Einstein substance, but is far from ideal at low temperatures, and the very real forces between molecules make the λ -transition to He II very different from that predicted for a Bose–Einstein gas.

Recent research (1995–) has produced at very low temperatures (nanokelvins) a Bose–Einstein condensation of magnetically trapped alkali metal atoms. Measurements [41] of the fraction of molecules in the ground state of ^{87}Rb as a function of temperature show good agreement with the predictions for a finite number of noninteracting bosons in the three-dimensional harmonic potential produced by the magnets; indeed the difference in this occupancy differs only slightly from that predicted for translation in a 3D box. However the variation of the energy as a function of temperature is significantly different from that predicted for a 3D box; the harmonic potential predicts a discontinuity in the heat capacity which is confirmed by experiment; thus this transition is second-order rather than third-order.

ACKNOWLEDGMENTS

I want to thank Anneke and Jan Sengers for supplying me with much information and for critical reading of parts of the manuscript. However any errors, omissions or misplaced emphases are entirely my own.

REFERENCES

- [1] Pippard A B 1957 *The Elements of Classical Thermodynamics* (Cambridge: Cambridge University Press) pp 136–59
- [2] Van der Waals J D 1873 Over de continuïteit van den gas- en vloeistoofstoestand *PhD Thesis* Sijthoff, Leiden (Engl. Transl. 1988 *J. D. van der Waals: On the Continuity of the Gaseous and Liquid States* ed J S Rowlinson, vol. XIV of *Studies in Statistical Mechanics* ed J L Lebowitz (Amsterdam: North-Holland))
- [3] Van Konynenburg P H and Scott R L 1980 Critical lines and phase equilibria in van der Waals mixtures *Phil Trans. R. Soc.* **298** 495–540
- [4] *Workshop on Global Phase Diagrams* 1999 *P C C P* **1** 4225–326 (16 papers from the workshop held at Walberberg, Germany 21–24 March 1999)
- [5] Bragg W L and Williams E J 1934 The effect of thermal agitation on atomic arrangement in alloys *Proc. R. Soc. A* **145** 699–730
- [6] Nix F C and Shockley W 1938 Order and disorder in alloys *Rev. Mod. Phys.* **10** 1–69
- [7] Weiss P 1907 L'Hypothèse du champ moléculaire et la propriété ferromagnétique *J. Phys. Radium Paris* **6** 661–90

- [9] Levelt Sengers J M H 1999 Mean-field theories, their weaknesses and strength *Fluid Phase Equilibria* **158–160** 3–17
- [10] Guggenheim E A 1945 The principle of corresponding states *J. Chem. Phys.* **13** 253–61
- [11] Scott R L 1953 Second-order transitions and critical phenomena *J. Chem. Phys.* **21** 209–11
- [12] Bagatskii M I, Voronel A V and Gusak V G 1962 Determination of heat capacity C_v of argon in the immediate vicinity of the critical point *Zh. Eksp. Teor. Fiz.* **43** 728–9
- [13] Green M S and Sengers J V (eds) 1966 *Critical Phenomena, Proc. Conf. (April, 1965)* (Washington: National Bureau of Standards Miscellaneous Publication 273)
- [14] Ising E 1925 Beitrag sur theorie des ferromagnetismus *Z. Phys.* **31** 253–8
- [15] Landau L D and Lifschitz E M 1969 *Statistical Physics* 2nd English edn (Oxford: Pergamon) chapter XIV. The quotation is from the first English edition (1959). The corresponding statement in the second English edition is slightly more cautious. However even the first edition briefly reports the results of Onsager and Yang for two dimensions, but leaves the reader with the belief (or the hope?) that somehow three dimensions is different.
- [16] Onsager L 1944 Crystal Statistics. I. A two-dimensional model with an order–disorder transition *Phys. Rev.* **65** 117–49
- [17] Widom B 1965 Equation of state in the neighborhood of the critical point *J. Chem. Phys.* **43** 3898–905
- [18] Jüngst S, Knuth B and Hensel F 1985 Observation of singular diameters in the coexistence curves of metals *Phys. Rev. Lett.* **55** 2160–3
- [19] Pestak M W, Goldstein R E, Chan M H W, de Bruyn J R, Balzarini D A and Ashcroft N W 1987 Three-body interactions, scaling variables, and singular diameters in the coexistence curves of fluids *Phys. Rev. B* **36** 599–614
- [20] Kac M, Uhlenbeck G E and Hemmer P C 1963 On the van der Waals theory of the vapor-liquid equilibrium. I. Discussion of a one-dimensional model *J. Math. Phys.* **4** 216–28
- [21] Griffiths R B and Wheeler J C 1970 Critical points in multicomponent systems *Phys. Rev. A* **2** 1047–64
- [22] Wheeler J C 2000 Personal communication
- [23] Wilson K G 1971 Renormalization group and critical phenomena. I. Renormalization group and the Kadanoff scaling picture *Phys. Rev. B* **4** 3174–83
Wilson K G 1971 Renormalization group and critical phenomena. II. Phase space cell analysis of critical behaviour *Phys. Rev. B* **4** 3184–205
- [24] Domb C 1996 *The Critical Point. A Historical Introduction to the Modern Theory of Critical Phenomena* (London and Bristol, PA: Taylor and Francis)

- [25] Wegner F J 1972 Corrections to scaling laws *Phys. Rev. B* **5** 4529–36
- [26] Scott R L 1978 Critical exponents for binary fluid mixtures *Specialist Periodical Reports, Chem. Thermodynam.* **2** 238–74
- [27] Guida R and Zinn-Justin J 1998 Critical exponents of the N -vector model *J. Phys. A Mathematical and General* **31** 8103–21

- [28] Anisimov M A and Sengers J V 2000 Critical and crossover phenomena in fluids and fluid mixtures *Supercritical Fluids-Fundamentals and Applications* ed E Kiran, P G Debenedetti and C J Peters (Dordrecht: Kluwer) pp 1–33
- [29] Kostrowicka Wyczalkowska A, Anisimov M A and Sengers J V 1999 Global crossover equation of state of a van der Waals fluid *Fluid Phase Equilibria* **158–160** 523–35
- [30] Povodyrev A A, Anisimov M A and Sengers J V 1999 Crossover Flory model for phase separation in polymer solutions *Physica A* **264** 345–69
- [31] Kim H K and Chan M H W 1984 Experimental determination of a two-dimensional liquid-vapor critical exponent *Phys. Rev. Lett.* **53** 170–3
- [32] Shrimpton N D, Cole M W, Steele W A and Chan M H W 1992 Rare gases on graphite *Surface Properties of Layered Structures* ed G Benedek, (Dordrecht: Kluwer) pp 219–69
- [33] Lipa J A, Swanson D R, Nissen J A, Chui T C P and Israelsson U E 1996 Heat capacity and thermal relaxation of bulk helium very near the lambda point *Phys. Rev. Lett.* **76** 944–7
- [34] Garland C W and Nounesis G 1994 Critical behavior at nematic-smectic-A phase transitions *Phys. Rev. E* **49** 2964–71
- [35] Wheeler J C, Kennedy S J and Pfeuty P 1980 Equilibrium polymerization as a critical phenomenon *Phys. Rev. Lett.* **45** 1748–52
- [36] Pasler V, Schweiss P, Meingast C, Obst B, Wühl H, Rykov A I and Tajima S 1998 3D-XY critical fluctuations of the thermal expansivity in detwinned $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$ single crystals near optimal doping *Phys. Rev. Lett.* **81** 1094–7
- [37] Domb C and Lebowitz J (eds) 1984 *Phase Transitions and Critical Phenomena* vol 9 (London, New York: Academic) ch 1. Lawrie I D and Sarbach S: Theory of tricritical points; ch 2. Knobler C M and Scott R L Multicritical points in fluid mixtures. Experimental studies.
- [38] Scott R L 1965 Phase equilibria in solutions of liquid sulfur. I. Theory *J. Phys. Chem.* **69** 261–70
- [39] Wheeler J C and Pfeuty P 1981 Critical points and tricritical points in liquid sulfur solutions *J. Chem. Phys.* **74** 6415–30
- [40] Pegg I L, Knobler C M and Scott R L 1990 Tricritical phenomena in quasibinary mixtures. VIII. Calculations from the van der Waals equation for binary mixtures *J. Chem. Phys.* **92** 5442–53
- [41] Ensher J R, Jin D S, Mathews M R, Wieman C E and Cornell E A 1996 Bose–Einstein condensation in a dilute gas: measurement of energy and ground-state occupation *Phys. Rev. Lett.* **77** 4984–7
-

A3.1 Kinetic theory: transport and fluctuations

J R Dorfman

A3.1.1 INTRODUCTION

The kinetic theory of gases has a long history, extending over a period of a century and a half, and is responsible for many central insights into, and results for, the properties of gases, both in and out of thermodynamic equilibrium [1]. Strictly speaking, there are two familiar versions of kinetic theory, an informal version and a formal version. The informal version is based upon very elementary considerations of the collisions suffered by molecules in a gas, and upon elementary probabilistic notions regarding the velocity and free path distributions of the molecules. In the hands of Maxwell, Boltzmann and others, the informal version of kinetic theory led to such important predictions as the independence of the viscosity of a gas on its density at low densities, and to qualitative results for the equilibrium thermodynamic properties, the transport coefficients, and the structure of microscopic boundary layers in a dilute gas. The more formal theory is also due to Maxwell and Boltzmann, and may be said to have had its beginning with the development of the Boltzmann transport equation in 1872 [2]. At that time Boltzmann obtained, by heuristic arguments, an equation for the time dependence of the spatial and velocity distribution function for particles in the gas. This equation provided a formal foundation for the informal methods of kinetic theory. It leads directly to the Maxwell–Boltzmann velocity distribution for the gas in equilibrium. For non-equilibrium systems, the Boltzmann equation leads to a version of the second law of thermodynamics (the Boltzmann H -theorem), as well as to the Navier–Stokes equations of fluid dynamics, with explicit expressions for the transport coefficients in terms of the intermolecular potentials governing the interactions between the particles in the gas [3]. It is not an exaggeration to state that the kinetic theory of gases was one of the great successes of nineteenth century physics. Even now, the Boltzmann equation remains one of the main cornerstones of our understanding of non-equilibrium processes in fluid as well as solid systems, both classical and quantum mechanical. It continues to be a subject of investigation in both the mathematical and physical literature and its predictions often serve as a way of distinguishing different molecular models employed to calculate gas properties. Kinetic theory is typically used to describe the non-equilibrium properties of dilute to moderately dense gases composed of atoms, or diatomic or polyatomic molecules. Such properties include the coefficients of shear and bulk viscosity, thermal conductivity, diffusion, as well as gas phase chemical reaction rates, and other, similar properties.

In this section we will survey both the informal and formal versions of the kinetic theory of gases, starting with the simpler informal version. Here the basic idea is to combine both probabilistic and mechanical arguments to calculate quantities such as the equilibrium pressure of a gas, the mean free distance between collisions for a typical gas particle, and the transport properties of the gas, such as its viscosity and thermal conductivity. The formal version again uses both probabilistic and mechanical arguments to obtain an equation, the Boltzmann transport equation, that determines the distribution function, $f(\mathbf{r}, \mathbf{v}, t)$, that describes the number of gas particles in a small spatial region, $\delta\mathbf{r}$, about a point \mathbf{r} , and in a small region of velocities, $\delta\mathbf{v}$, about a given velocity \mathbf{v} , at some time t . The formal theory forms the basis for almost all applications of kinetic theory to realistic systems.

We will almost always treat the case of a dilute gas, and almost always consider the approximation that the gas particles obey classical, Hamiltonian mechanics. The effects of quantum properties and/or of higher densities will be briefly commented upon. A number of books have been devoted to the kinetic theory of gases. Here we note that some

of the interesting and easily accessible ones are those of Boltzmann [2], Chapman and Cowling [3], Hirshfelder *et al* [4], Hanley [5], Fertziger and Kaper [6], Resibois and de Leener [7], Liboff [8] and Present [9]. Most textbooks on the subject of statistical thermodynamics have one or more chapters on kinetic theory [10, 11, 12 and 13].

A3.1.2 THE INFORMAL KINETIC THEORY FOR THE DILUTE GAS

We begin by considering a gas composed of N particles in a container of volume V . We suppose, first, that the particles are single atoms, interacting with forces of finite range denoted by a . Polyatomic molecules can be incorporated into this informal discussion, to some extent, but atoms and molecules interacting with long-range forces require a separate treatment based upon the Boltzmann transport equation. This equation is capable of treating particles that interact with infinite-range forces, at least if the forces approach zero sufficiently rapidly as the separation of the particles becomes infinite. Typical potential energies describing the interactions between particles in the gas are illustrated in figure A3.1.1 where we describe Lennard–Jones (LJ) and Weeks–Chandler–Anderson (WCA) potentials. The range parameter, a , is usually taken to be a value close to the first point where the potential energy becomes negligible for all greater separations. While choice of the location of this point is largely subjective, it will not be a serious issue in what follows, since the results to be described below are largely qualitative order-of-magnitude results. However we may usefully take the distance a to represent the effective *diameter* of a particle.

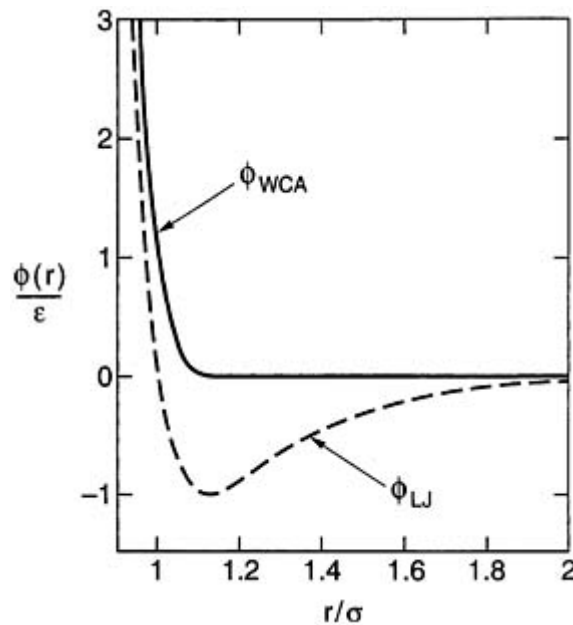


Figure A3.1.1. Typical pair potentials. Illustrated here are the Lennard–Jones potential, ϕ_{LJ} , and the Weeks–Chandler–Anderson potential, ϕ_{WCA} , which gives the same repulsive force as the Lennard–Jones potential. The relative separation is scaled by σ , the distance at which the Lennard–Jones first passes through zero. The energy is scaled by the well depth, ϵ .

The dilute gas condition can be stated as the condition that the available volume per particle in the container is much larger than the volume of the particle itself. In other words

$$\frac{V}{N} \gg a^3 \quad \text{or} \quad na^3 \ll 1 \quad (\text{A3.1.1})$$

where $n = N/V$ is the average number of particles per unit volume. We will see below that this condition is equivalent to the requirement that the mean free path between collisions, which we denote by λ , is much greater than the size of a particle, a . Next we suppose that the state of the gas can be described by a distribution function $f(\mathbf{r}, \mathbf{v}, t)$, such that $f(\mathbf{r}, \mathbf{v}, t) d\mathbf{r} d\mathbf{v}$ is the number of gas particles in $d\mathbf{r}$ about \mathbf{r} , and in $d\mathbf{v}$ about \mathbf{v} at time t . To describe the state of a gas of polyatomic molecules, or of any mixture of different particles, we would need to include additional variables in the argument of f to describe the internal states of the molecules and the various components of the mixture. To keep the discussion simple, we will consider gases of monoatomic particles, for the time being.

At this point it is important to make some clarifying remarks: (1) clearly one cannot regard $d\mathbf{r}$ in the above expression, strictly, as a mathematical differential. It cannot be infinitesimally small, since $d\mathbf{r}$ must be large enough to contain some particles of the gas. We suppose instead that $d\mathbf{r}$ is large enough to contain some particles of the gas but small compared with any important physical length in the problem under consideration, such as a mean free path, or the length scale over which a physical quantity, such as a temperature, might vary. (2) The distribution function $f(\mathbf{r}, \mathbf{v}, t)$ typically does not describe the *exact* state of the gas in the sense that it tells us exactly how many particles are in the designated regions at the given time t . To obtain and use such an exact distribution function one would need to follow the motion of the individual particles in the gas, that is, solve the mechanical equations for the system, and then do the proper counting. Since this is clearly impossible for even a small number of particles in the container, we have to suppose that f is an ensemble average of the microscopic distribution functions for a very large number of identically prepared systems. This, of course, implies that kinetic theory is a branch of the more general area of statistical mechanics. As a result of these two remarks, we should regard any distribution function we use as an ensemble average rather than an exact expression for our particular system, and we should be careful when examining the variation of the distribution with space and time, to make sure that we are not too concerned with variations on spatial scales that are of the order or less than the size of a molecule, or on time scales that are of the order of the duration of a collision of a particle with a wall or of two or more particles with each other.

A3.1.2.1 EQUILIBRIUM PROPERTIES FROM KINETIC THEORY

The equilibrium state for a gas of monoatomic particles is described by a spatially uniform, time independent distribution function whose velocity dependence has the form of the Maxwell–Boltzmann distribution, obtained from equilibrium statistical mechanics. That is, $f(\mathbf{r}, \mathbf{v}, t)$ has the form $f_{\text{eq}}(\mathbf{v})$ given by

$$f_{\text{eq}}(\mathbf{v}) = n\varphi(\mathbf{v}) \quad (\text{A3.1.2})$$

-4-

where

$$\varphi(\mathbf{v}) = \left(\frac{\beta m}{2\pi} \right)^{3/2} e^{-\beta \mathbf{v}^2 / 2m} \quad (\text{A3.1.3})$$

is the usual Maxwell–Boltzmann velocity distribution function. Here m is the mass of the particle, and the quantity $\beta = (k_{\text{B}}T)^{-1}$, where T is the equilibrium thermodynamic temperature of the gas and k_{B} is Boltzmann's

constant, $k_B = 1.380 \times 10^{-23} \text{ J K}^{-1}$.

We are now going to use this distribution function, together with some elementary notions from mechanics and probability theory, to calculate some properties of a dilute gas in equilibrium. We will calculate the pressure that the gas exerts on the walls of the container as well as the rate of effusion of particles from a very small hole in the wall of the container. As a last example, we will calculate the mean free path of a molecule between collisions with other molecules in the gas.

(A) THE PRESSURE

To calculate the pressure, we need to know the force per unit area that the gas exerts on the walls of the vessel. We calculate the force as the negative of the rate of change of the vector momentum of the gas particles as they strike the container. We consider then some small area, A , on the wall of the vessel and look at particles with a particular velocity \mathbf{v} , chosen so that it is physically possible for particles with this velocity to strike the designated area from within the container. We consider a small time interval δt , and look for all particles with velocity \mathbf{v} that will strike this area, A over time interval δt . As illustrated in [figure A3.1.2](#) all such particles must lie in a small ‘cylinder’ of base area A , and height, $|\mathbf{v} \cdot \hat{\mathbf{n}}|\delta t$, where $\hat{\mathbf{n}}$ is a unit normal to the surface of the container at the small area A , and directed toward the interior of the vessel. We will assume that the gas is very dilute and that we can ignore the collisions between particles, and take only collisions of particles with the wall into account. Every time such a particle hits our small area of the wall, its momentum changes, since its momentum after a collision differs from its momentum prior to the collision. Let us suppose that the particles make elastic, specular collisions with the surface, so that the momentum change per particle at each collision is $\Delta\mathbf{p} = -2(\mathbf{p} \cdot \hat{\mathbf{n}})\hat{\mathbf{n}} = -2m(\mathbf{v} \cdot \hat{\mathbf{n}})\hat{\mathbf{n}}$. This vector is directed in toward the container. Now to calculate the total change in the momentum of the gas in time δt due to collisions with the wall at the point of interest, we have to know how many particles with velocity \mathbf{v} collide with the wall, multiply the number of collisions by the change in momentum per collision, and then integrate over all possible values of the velocity \mathbf{v} than can lead to such a collision. To calculate the number of particles striking the small area, A , in time interval δt , we have to invoke probabilistic arguments, since we do not know the actual locations and the velocities of all the particles at the beginning of the time interval. We do know that if we ignore possible collisions amongst the particles themselves, all of the particles with velocity \mathbf{v} colliding with A in time δt will have to reside in the small cylinder illustrated in [figure A3.1.2](#), with volume $A|\mathbf{v} \cdot \hat{\mathbf{n}}|\delta t$. Now, using the distribution function f given by equation [\(A3.1.2\)](#), we find that the number, $\delta\mathcal{N}(\mathbf{v})$, of particles with velocity \mathbf{v} in the range $d\mathbf{v}$, in the collision cylinder is

$$\Delta\mathbf{p}_{\text{total}} = -2A\hat{\mathbf{n}}\delta t n \int_{\mathbf{v} \cdot \hat{\mathbf{n}} \leq 0} d\mathbf{v} |\mathbf{v} \cdot \hat{\mathbf{n}}| (\mathbf{v} \cdot \hat{\mathbf{n}}) \varphi(v). \quad (\text{A3.1.4})$$

-5-

Now each such particle adds its change in momentum, as given above, to the total change of momentum of the gas in time δt . The total change in momentum of the gas is obtained by multiplying $\delta\mathcal{N}$ by the change in momentum per particle and integrating over all allowed values of the velocity vector, namely, those for which $\mathbf{v} \cdot \hat{\mathbf{n}} \leq 0$. That is

$$\Delta\mathbf{p}_{\text{total}} = -2A\hat{\mathbf{n}}\delta t n \int_{\mathbf{v} \cdot \hat{\mathbf{n}} \leq 0} d\mathbf{v} |\mathbf{v} \cdot \hat{\mathbf{n}}| (\mathbf{v} \cdot \hat{\mathbf{n}}) \varphi(v). \quad (\text{A3.1.5})$$

Finally the pressure, P , exerted by the gas on the container, is the negative of the force per unit area that the

wall exerts on the gas. This force is measured by the change in momentum of the gas per unit time. Thus we are led to

$$\begin{aligned}
 P &= 2n \int_{\mathbf{v} \cdot \hat{\mathbf{n}} \leq 0} d\mathbf{v} |\mathbf{v} \cdot \hat{\mathbf{n}}|^2 \varphi(v) \\
 &= \frac{n}{\beta} = nk_B T.
 \end{aligned}
 \tag{A3.1.6}$$

Here we have carried out the velocity integral over the required half-space and used the explicit form of the Maxwell–Boltzmann distribution function, given by equation (A3.1.3).

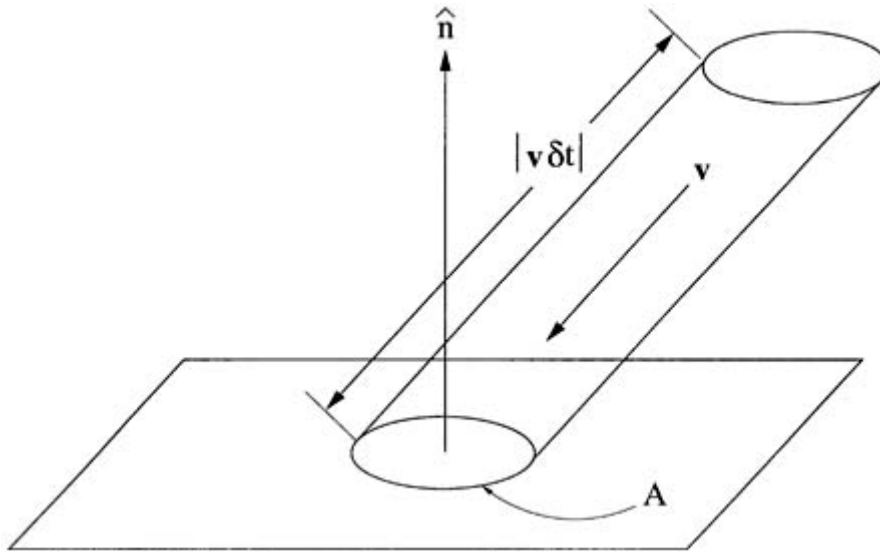


Figure A3.1.2. A collision cylinder for particles with velocity \mathbf{v} striking a small region of area A on the surface of a container within a small time interval δt . Here $\hat{\mathbf{n}}$ is a unit normal to the surface at the small region, and points into the gas.

-6-

(B) THE RATE OF EFFUSION THROUGH A SMALL HOLE

It is a simple matter now to calculate number of particles per unit area, per unit time, that pass through a small hole in the wall of the vessel. This quantity is called the rate of effusion, denoted by n_e , and it governs the loss of particles in a container when there is a small hole in the wall separating the gas from a vacuum, say. This number is in fact obtained by integrating the quantity, $\delta \mathcal{N}(\mathbf{v})$ over all possible velocities having the proper direction, and then dividing this number by $A\delta t$. Thus we find

$$\tag{A3.1.7}$$

where \bar{v} is the average speed of a particle in a gas in equilibrium, given by

$$\bar{v} = \left(\frac{8}{m\pi\beta} \right)^{1/2}.
 \tag{A3.1.8}$$

The result, (A3.1.7), can be viewed also as the number of particles per unit area per unit time colliding from

one side of any small area in the gas, whether real or fictitious. We will use this result in the next section when we consider an elementary kinetic theory for transport coefficients in a gas with some kind of flow taking place.

(C) THE MEAN FREE PATH

The previous calculations, while not altogether trivial, are among the simplest uses one can make of kinetic theory arguments. Next we turn to a somewhat more sophisticated calculation, that for the mean free path of a particle between collisions with other particles in the gas. We will use the general form of the distribution function at first, before restricting ourselves to the equilibrium case, so as to set the stage for discussions in later sections where we describe the formal kinetic theory. Our approach will be first to compute the average frequency with which a particle collides with other particles. The inverse of this frequency is the mean time between collisions. If we then multiply the mean time between collisions by the mean speed, given by equation (A3.1.8), we will obtain the desired result for the mean free path between collisions. It is important to point out that one might choose to define the mean free path somewhat differently, by using the root mean square velocity instead of \bar{v} , for example. The only change will be in a numerical coefficient. The important issue will be to obtain the dependence of the mean free path upon the density and temperature of the gas and on the size of the particles. The numerical factors are not that important.

Let us focus our attention for the moment on a small volume in space, $d\mathbf{r}$, and on particles in the volume with a given velocity \mathbf{v} . Let us sit on such a particle and ask if it might collide in time δt with another particle whose velocity is \mathbf{v}_1 , say. Taking the effective diameter of each particle to be a , as described above, we see that our particle with velocity \mathbf{v} presents a cross sectional area of size πa^2 for collisions with other particles. If we focus on collisions with another

-7-

particle with velocity \mathbf{v}_1 , then, as illustrated in [figure A3.1.3](#) a useful coordinate system to describe this collision is one in which the particle with velocity \mathbf{v} is located at the origin and the z -axis is aligned along the direction of the vector $\mathbf{g} = \mathbf{v}_1 - \mathbf{v}$. In this coordinate system, the centre of the particle with velocity \mathbf{v}_1 must be somewhere in the collision cylinder of volume $\pi a^2 |\mathbf{g}| \delta t$ in order that a collision between the two particles takes place in the time interval δt . Now in the small volume $d\mathbf{r}$ there are $f(\mathbf{r}, \mathbf{v}, t) d\mathbf{r}$ particles with velocity \mathbf{v} at time t , each one with a collision cylinder of the above type attached to it. Thus the total volume, $\delta\mathcal{V}(\mathbf{v}, \mathbf{v}_1)$ of these $(\mathbf{v}, \mathbf{v}_1)$ collision cylinders is

$$\delta\mathcal{V}(\mathbf{v}, \mathbf{v}_1) = \pi a^2 |\mathbf{g}| \delta t f(\mathbf{r}, \mathbf{v}, t). \quad (\text{A3.1.9})$$

Now, again, we use a probabilistic argument to say that the number of particles with velocity \mathbf{v}_1 in this total volume is given by the product of the total volume and the number of particles per unit volume with velocity \mathbf{v}_1 , that is, $\delta\mathcal{V}(\mathbf{v}, \mathbf{v}_1) f(\mathbf{r}, \mathbf{v}_1, t)$. To complete the calculation, we suppose that the gas is so dilute that each of the collision cylinders has either zero or one particle with velocity \mathbf{v}_1 in it, and that each such particle actually collides with the particle with velocity \mathbf{v} . Thus the total number of collisions suffered by particles with velocity \mathbf{v} in time δt is

$$\pi a^2 \delta t f(\mathbf{r}, \mathbf{v}, t) \int d\mathbf{v}_1 |\mathbf{v}_1 - \mathbf{v}| f(\mathbf{r}, \mathbf{v}_1, t).$$

Then it follows that the total number of collisions per unit time suffered by particles with *all* velocities is

$$\pi a^2 \int d\mathbf{v} \int d\mathbf{v}_1 |\mathbf{v}_1 - \mathbf{v}| f(\mathbf{r}, \mathbf{v}, t) f(\mathbf{r}, \mathbf{v}_1, t). \quad (\text{A3.1.10})$$

Notice that each collision is counted twice, once for the particle with velocity \mathbf{v} and once for the particle with velocity \mathbf{v}_1 . We also note that we have assumed that the distribution functions f do not vary over distances which are the lengths of the collision cylinders, as the interval δt approaches some small value, but still large compared with the duration of a binary collision.

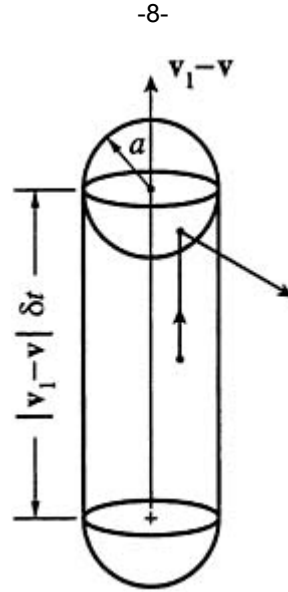


Figure A3.1.3. The collision cylinder for collisions between particles with velocities \mathbf{v} and \mathbf{v}_1 . The origin is placed at the centre of the particle with velocity \mathbf{v} and the z -axis is in the direction of $\mathbf{v}_1 - \mathbf{v}$. The spheres indicate the range, a , of the intermolecular forces.

Our first result is now the average collision frequency obtained from the expression, (A3.1.10), by dividing it by the average number of particles per unit volume. Here it is convenient to consider the equilibrium case, and to use (A3.1.2) for f . Then we find that the average collision frequency, ν , for the particles is

$$\begin{aligned} \nu &= n\pi a^2 \int d\mathbf{v} d\mathbf{v}_1 |\mathbf{v}_1 - \mathbf{v}| \varphi(v) \varphi(v_1) \\ &= n\pi a^2 \left(\frac{16}{\beta\pi m} \right)^{1/2}. \end{aligned} \quad (\text{A3.1.11})$$

The average time between collisions is then ν^{-1} , and in this time the particle will typically travel a distance λ , the mean free path, where

$$\lambda = \bar{v} \nu^{-1} = \frac{1}{2^{1/2} \pi n a^2}. \quad (\text{A3.1.12})$$

This is the desired result. It shows that the mean free path is inversely proportional to the density and the collision cross section. This is a physically sensible result, and could have been obtained by dimensional

arguments alone, except for the unimportant numerical factor.

A3.1.2.2 THE MEAN FREE PATH EXPRESSIONS FOR TRANSPORT COEFFICIENTS

One of the most useful applications of the mean free path concept occurs in the theory of transport processes in systems where there exist gradients of average but local density, local temperature, and/or local velocity. The existence of such gradients causes a transfer of particles, energy or momentum, respectively, from one region of the system to another.

The kinetic theory of transport processes in gases rests upon three basic assumptions.

- (i) The gas is dense enough that the mean free path is small compared with the characteristic size of the container. Consequently, the particles collide with each other much more often than they collide with the walls of the vessel.
- (ii) As stated above, the gas is sufficiently dilute that the mean free path is much larger than the diameter of a particle.
- (iii) The local density, temperature and density vary slowly over distances of the order of a mean free path.

If these assumptions are satisfied then the ideas developed earlier about the mean free path can be used to provide qualitative but useful estimates of the transport properties of a dilute gas. While many varied and complicated processes can take place in fluid systems, such as turbulent flow, pattern formation, and so on, the principles on which these flows are analysed are remarkably simple. The description of both simple and complicated flows in fluids is based on five hydrodynamic equations, the Navier–Stokes equations. These equations, in turn, are based upon the mechanical laws of conservation of particles, momentum and energy in a fluid, together with a set of phenomenological equations, such as Fourier’s law of thermal conduction and Newton’s law of fluid friction. When these phenomenological laws are used in combination with the conservation equations, one obtains the Navier–Stokes equations. Our goal here is to derive the phenomenological laws from elementary mean free path considerations, and to obtain estimates of the associated transport coefficients. Here we will consider thermal conduction and viscous flow as examples.

(A) THERMAL CONDUCTION

We can obtain an understanding of Fourier’s law of thermal conduction by considering a very simple situation, frequently encountered in the laboratory. Imagine a layer of gas, as illustrated in [figure A3.1.4](#) which is small enough to exclude convection, but many orders of magnitude larger than a mean free path. Imagine further that the temperature is maintained at constant values, T_1 and T_2 , $T_2 > T_1$, along two planes separated by a distance L , as illustrated. We suppose that the system has reached a stationary state so that the local temperature at any point in the fluid is constant in time and depends only upon the z -component of the location of the point. Now consider some imaginary plane in the fluid, away from the boundaries, and look at the flow of particles across the plane. We make a major simplification and assume that all particles crossing the plane carry with them the local properties of the system a mean free path above and below the plane. That is, suppose we examine the flow of particles through the plane, coming from above it. Then we can say that the number of particles crossing the plane per unit area and per unit time from above, i.e. the particle current density, \vec{j}_n heading down, is given by

$$j_n^-(z) = \frac{1}{4}n(z + \lambda)\bar{v}(z + \lambda) \quad (\text{A3.1.13})$$

where z is the height of the plane we consider, λ is the mean free path, and we use (A3.1.7) for this current density. Similarly, the upward flux is

$$j_n^+(z) = \frac{1}{4}n(z - \lambda)\bar{v}(z - \lambda). \quad (\text{A3.1.14})$$

In a steady state, with no convection, the two currents must be equal, $j_n^+(z) = j_n^-(z) \equiv j_n(z)$. Now we assume that each particle crossing the plane carries the energy per particle characteristic of the location at a mean free path above or below the plane. Thus the upward and downward energy current densities, j_e^\pm , are

$$j_e^\pm(z) = j_n(z)e(z \mp \lambda) \quad (\text{A3.1.15})$$

where $e(z \mp \lambda)$ is the local energy per particle at a distance λ below and above the plane. The net amount of energy transferred per unit area per unit time in the positive z direction, $q(z)$, is then

$$\begin{aligned} q(z) &= j_e^+ - j_e^- = j_n(z)[e(z - \lambda) - e(z + \lambda)] \\ &= -2j_n(z)\lambda \frac{\partial e(z)}{\partial z} + \mathcal{O}\left(\frac{\partial^2 e}{\partial z^2}\right). \end{aligned} \quad (\text{A3.1.16})$$

Neglecting derivatives of the third order and higher, we obtain Fourier's law of thermal conduction

$$q(z) = -k \frac{\partial T}{\partial z} \quad (\text{A3.1.17})$$

where the coefficient of thermal conductivity, k , is given by

$$k = \frac{1}{2}n\bar{v}\lambda \frac{\partial e}{\partial T}. \quad (\text{A3.1.18})$$

The result is, of course, a case of the more general expression of Fourier's law, namely

$$\mathbf{q} = -k\nabla T \quad (\text{A3.1.19})$$

adjusted to the special situation that the temperature gradient is in the z -direction. Since k is obviously positive, our result is in accord with the second law of thermodynamics, which requires heat to flow from hotter to colder regions.

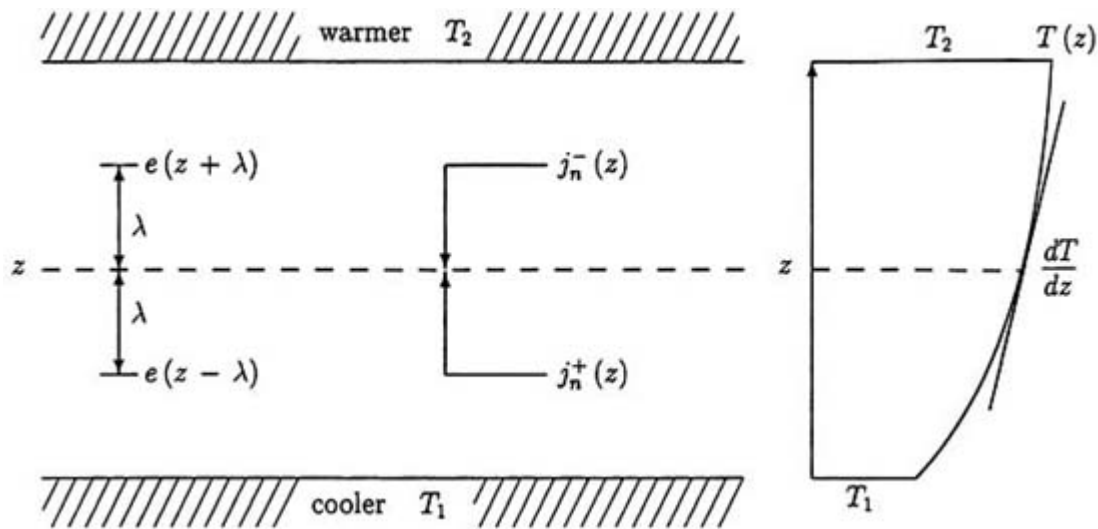


Figure A3.1.4. Steady state heat conduction, illustrating the flow of energy across a plane at a height z .

We can easily obtain an expression for k by using the explicit forms for \bar{v} and λ , given in (A3.1.8) and (A3.1.12). Thus, in this approximation

$$k = \frac{c_v (k_B T)^{1/2}}{a^2 m^{1/2} \pi^{3/2}}. \quad (\text{A3.1.20})$$

where c_v is the specific heat per particle. We have assumed that the gradients are sufficiently small that the local average speed and mean free path can be estimated by their (local) equilibrium values. The most important consequences of this result for the thermal conductivity are its independence of the gas density and its variation with temperature as $T^{1/2}$. The independence of density is well verified at low gas pressures, but the square-root temperature dependence is only verified at high temperatures. Better results for the temperature dependence of κ can be obtained by use of the Boltzmann transport equation, which we discuss in the next section. The temperature dependence turns out to be a useful test of the functional form of the intermolecular potential energy.

(B) THE SHEAR VISCOSITY

A distribution of velocities in a fluid gives rise to a transport of momentum in the fluid in complete analogy with the transport of energy which results from a distribution of temperatures. To analyse this transport of momentum in a fluid with a gradient in the average local velocity, we use the same method as employed in the case of thermal conduction. That is, we consider a layer of fluid contained between two parallel planes, moving with velocities in the x -direction with values U_1 and U_2 , $U_2 > U_1$, as illustrated in figure A3.1.5. We suppose that the width of the layer is very large compared with a mean free path, and that the fluid adjacent to the moving planes moves with the velocity of the adjacent plane. If the velocities are not so large as to develop a turbulent flow, then a steady state can be maintained with an average local velocity, $\mathbf{u}(x,y,z)$, in the fluid of the form, $\mathbf{u}(x,y,z) = u_x(z)\hat{\mathbf{x}}$, where $\hat{\mathbf{x}}$ is a unit vector in the x -direction.

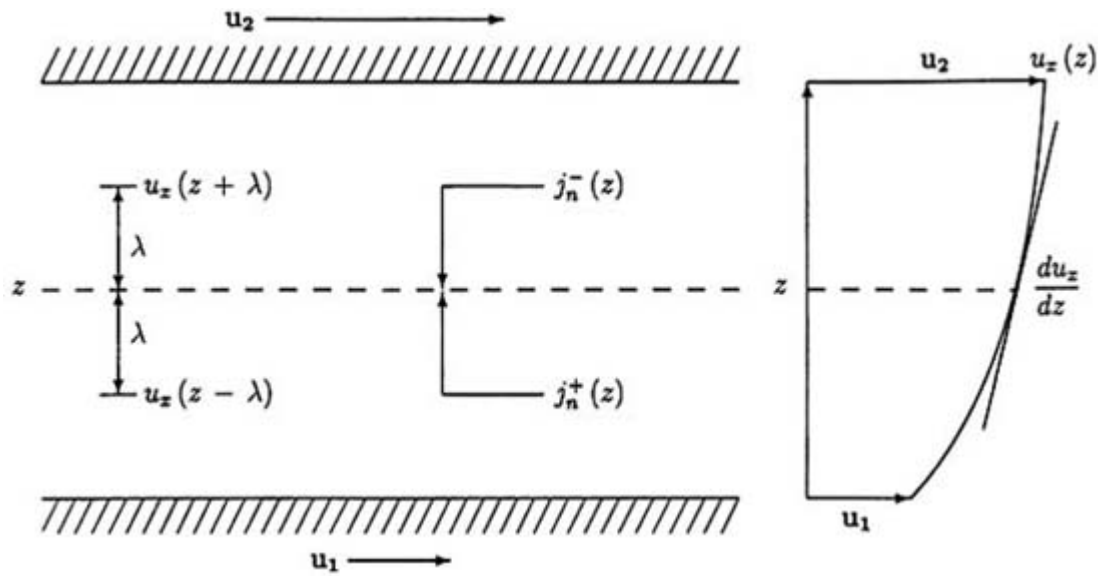


Figure A3.1.5. Steady state shear flow, illustrating the flow of momentum across a plane at a height z .

The molecules of the gas are in constant motion, of course, and there is a transport of particles in all directions in the fluid. If we consider a fictitious plane in the fluid, far from the moving walls, at a height z , then there will be a flow of particles from above and below the plane. The particles coming from above will carry a momentum with them typical of the average flow at a height $z + \lambda$, while those coming from below will carry the typical momentum at height $z - \lambda$, where λ is the mean free path length. Due to the velocity gradient in the fluid there will be a net transport of momentum across the plane, tending to slow down the faster regions and to accelerate the slower regions. This transport of momentum leads to viscous forces (or stresses if measured per unit area) in the fluid, which in our case will be in the x -direction. The analysis of this viscous stress is almost identical to that for thermal conduction.

Following the method used above, we see that there will be an upward flux of momentum in the x -direction, $j_{p_x}^+(z)$, across the plane at z given by

$$j_{p_x}^+(z) = j_n(z) m u_x(z - \lambda) \quad (\text{A3.1.21})$$

and a downward flux

$$j_{p_x}^-(z) = j_n(z) m u_x(z + \lambda). \quad (\text{A3.1.22})$$

The net upward flow in the x -component of the momentum is called the shear stress, σ_{zx} , and by combining (A3.1.21) and (A3.1.22), we see that

$$\begin{aligned} \sigma_{zx} &= j_{p_x}^+(z) - j_{p_x}^-(z) = j_n(z) m u_x(z - \lambda) - j_n(z) m u_x(z + \lambda) \\ &= -\frac{1}{2} m \lambda n(z) \bar{v}(z) \frac{\partial u_x(z)}{\partial z} + \dots \end{aligned} \quad (\text{A3.1.23})$$

Here we have neglected derivatives of the local velocity of third and higher orders. Equation (A3.1.23) has the form of the phenomenological Newton's law of friction

$$\sigma_{zx} = -\eta \frac{\partial u_x(z)}{\partial z} \quad (\text{A3.1.24})$$

if we identify the coefficient of shear viscosity η with the quantity

$$\eta = \frac{1}{2} m \lambda n(z) \bar{v}(z). \quad (\text{A3.1.25})$$

An explicit expression for the coefficient of shear viscosity can be obtained by assuming the system is in local thermodynamic equilibrium and using the previously derived expression for λ and \bar{v} . Thus we obtain

$$\eta = \frac{(mk_B T)^{1/2}}{\pi^{3/2} a^2}. \quad (\text{A3.1.26})$$

As in the case of thermal conductivity, we see that the viscosity is independent of the density at low densities, and grows with the square root of the gas temperature. This latter prediction is modified by a more systematic calculation based upon the Boltzmann equation, but the independence of viscosity on density remains valid in the Boltzmann equation approach as well.

(C) THE EUKEN FACTOR

We notice, using (A3.1.20) and (A3.1.26), that this method leads to a simple relation between the coefficients of shear viscosity and thermal conductivity, given by

$$\frac{k}{m\eta c_v} = 1. \quad (\text{A3.1.27})$$

That is, this ratio should be a universal constant, valid for all dilute gases. A more exact calculation based upon the Boltzmann equation shows that the right-hand side of equation (A3.1.27) should be replaced by 2.5 instead of 1, plus a correction that varies slightly from gas to gas. The value of 2.5 holds with a very high degree of accuracy for dilute monatomic gases [5]. However, when this ratio is computed for diatomic and polyatomic gases, the value of 2.5 is no longer recovered.

-14-

Euken advanced a very simple argument which allowed him to extend the Boltzmann equation formula for $k/(m\eta c_v)$ to diatomic and polyatomic gases. His argument is that when energy is transported in a fluid by particles, the energy associated with each of the internal degrees of freedom of a molecule is transported. However, the internal degrees of freedom play no role in the transport of momentum. Thus we should modify (A3.1.20) to include these internal degrees of freedom. If we also modify it to correct for the factor of 2.5 predicted by the Boltzmann equation, we obtain

$$k = \frac{1}{2} n \bar{v} \lambda (C c_v^{\text{tr}} + c_v^{\text{i}}) \quad (\text{A3.1.28})$$

where $C = 2.5$, c_v^{tr} is the translational specific heat per molecule, and c_v^{i} is the specific heat per molecule associated with the internal degrees of freedom. We can easily obtain a better value for the ratio, $k/(m\eta c_v)$ in terms of the ratio of specific heat at constant pressure per molecule to the specific heat at constant volume, $\gamma = c_p/c_v$, as

$$k/(m\eta c_v) = \frac{1}{4}(9\gamma - 5). \quad (\text{A3.1.29})$$

The right-hand side of (A3.1.29), called the Eucken factor, provides a reasonably good estimate for this ratio [11].

A3.1.3 THE BOLTZMANN TRANSPORT EQUATION

In 1872, Boltzmann introduced the basic equation of transport theory for dilute gases. His equation determines the time-dependent position and velocity distribution function for the molecules in a dilute gas, which we have denoted by $f(\mathbf{r}, \mathbf{v}, t)$. Here we present his derivation and some of its major consequences, particularly the so-called H -theorem, which shows the consistency of the Boltzmann equation with the irreversible form of the second law of thermodynamics. We also briefly discuss some of the famous debates surrounding the mechanical foundations of this equation.

We consider a large vessel of volume V , containing N molecules which interact with central, pairwise additive, repulsive forces. The latter requirement allows us to avoid the complications of long-lived ‘bound’ states of two molecules which, though interesting, are not central to our discussion here. We suppose that the pair potential has a strong repulsive core and a finite range a , such as the WCA potential illustrated in [figure A3.1.1](#). Now, as before, we define a distribution function, $f(\mathbf{r}, \mathbf{v}, t)$, for the gas over a six-dimensional position and velocity space, (\mathbf{r}, \mathbf{v}) , such that

$$f(\mathbf{r}, \mathbf{v}, t) \delta \mathbf{r} \delta \mathbf{v} \equiv \text{the number of particles in } \delta \mathbf{r} \delta \mathbf{v} \text{ around } \mathbf{r} \text{ and } \mathbf{v} \text{ at time } t. \quad (\text{A3.1.30})$$

To get an equation for $f(\mathbf{r}, \mathbf{v}, t)$, we take a region $\delta \mathbf{r} \delta \mathbf{v}$ about a point (\mathbf{r}, \mathbf{v}) , that is large enough to contain a lot of particles, but small compared with the range of variation of f .

There are four mechanisms that change the number of particles in this region. The particles can:

-15-

- (i) flow into or out of $\delta \mathbf{r}$, the *free-streaming term*,
- (ii) leave the $\delta \mathbf{v}$ region as a result of a direct collision, the *loss term*,
- (iii) enter the $\delta \mathbf{v}$ region after a restituting collision, the *gain term*, and
- (iv) collide with the wall of the container (if the region contains part of the walls), the *wall term*.

We again assume that there is a time interval δt which is long compared with the duration of a binary collision but is too short for particles to cross a cell of size $\delta \mathbf{r}$. Then the change in the number of particles in $\delta \mathbf{r} \delta \mathbf{v}$ in time δt can be written as

$$[f(\mathbf{r}, \mathbf{v}, t + \delta t) - f(\mathbf{r}, \mathbf{v}, t)]\delta\mathbf{r}\delta\mathbf{v} = \Gamma_f - \Gamma_- + \Gamma_+ + \Gamma_w \quad (\text{A3.1.31})$$

where Γ_f , Γ_- , Γ_+ , and Γ_w represent the changes in f due to the four mechanisms listed above, respectively. We suppose that each particle in the small region suffers at most one collision during the time interval δt , and calculate the change in f .

The computation of Γ_f is relatively straightforward. We simply consider the free flow of particles into and out of the region in time δt . An expression for this flow in the x -direction, for example, can be obtained by considering two thin layers of size $v_x \delta t \delta r_y \delta r_z$ that contain particles that move into or out of a cell with its centre at (x, y, z) in time δt (see figure A3.1.6).

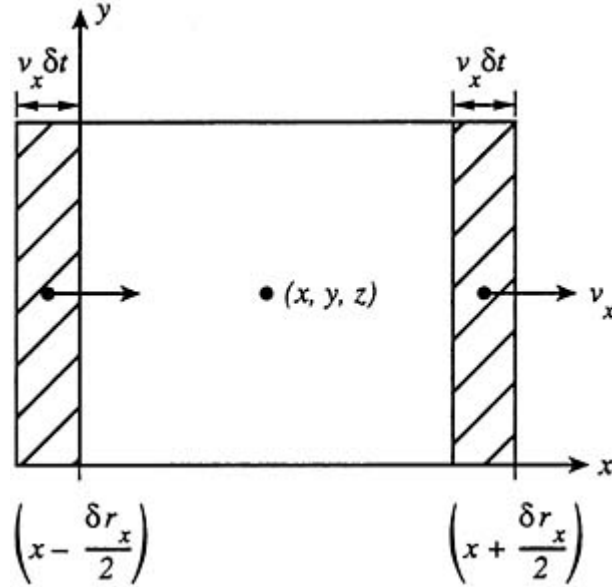


Figure A3.1.6. A schematic illustration of flow into and out of a small region. The hatched areas represent regions where particles enter and leave the region in time δt .

-16-

The free streaming term can be written as the difference between the number of particles entering and leaving the small region in time δt . Consider, for example, a cubic cell and look at the faces perpendicular to the x -axis. The flow of particles across the faces at $x - \frac{1}{2}\delta r_x$ and at $x + \frac{1}{2}\delta r_x$ is

$$\Gamma_f^{(x)} = v_x \delta t \delta r_y \delta r_z \delta \mathbf{v} [f(x - \frac{1}{2}\delta r_x, y, z, \mathbf{v}, t) - f(x + \frac{1}{2}\delta r_x, y, z, \mathbf{v}, t)] \quad (\text{A3.1.32})$$

and similar expressions exist for the y - and z -directions. The function f is supposed to be sufficiently smooth that it can be expanded in a Taylor series around (x, y, z) . The zeroth-order terms between the parentheses cancel and the first-order terms add up. Neglecting terms of order δ^2 and higher and summing over all directions then yields

$$\Gamma_f = -\delta \mathbf{v} \delta t \delta \mathbf{r} (\mathbf{v} \cdot \nabla) f(\mathbf{r}, \mathbf{v}, t). \quad (\text{A3.1.33})$$

Next we consider the computation of the loss term, Γ_- . As in the calculation of the mean free path, we need to calculate the number of collisions suffered by particles with velocity \mathbf{v} in the region $\delta\mathbf{r}\delta\mathbf{v}$ in time δt , assuming that each such collision results in a change of the velocity of the particle. We carry out the calculation in several steps. First, we focus our attention on a particular particle with velocity \mathbf{v} , and suppose that it is going to collide sometime during the interval $[t, t + \delta t]$ with a particle with velocity \mathbf{v}_1 . Now examine again the coordinate system with origin at the center of the particle with velocity \mathbf{v} , and with the z -axis directed along the vector $\mathbf{g} = \mathbf{v}_1 - \mathbf{v}$. By examining [figure A3.1.3](#) one can easily see that if the particle with velocity \mathbf{v}_1 is somewhere at time t within the *collision cylinder* illustrated there, with volume $|\mathbf{v}_1 - \mathbf{v}|\pi a^2\delta t$, this particle will collide sometime during the interval $[t, t + \delta t]$ with the particle with velocity \mathbf{v} , if no other particles interfere, which we assume to be the case. These collision cylinders will be referred to as $(\mathbf{v}_1, \mathbf{v})$ -collision cylinders. We also ignore the possibility that the particle with velocity \mathbf{v}_1 might, at time t , be somewhere within the action sphere of radius a about the centre of the velocity- \mathbf{v} particle, since such events lead to terms that are of higher order in the density than those we are considering here, and such terms do not even exist if the duration of a binary collision is strictly zero, as would be the case for hard spheres, for example.

We now compute Γ_- by noting again the steps involved in calculating the mean free path, but applying them now to the derivation of an expression for Γ_- .

- The number of $(\mathbf{v}_1, \mathbf{v})$ -collision cylinders in the region $\delta\mathbf{r}\delta\mathbf{v}$ is equal to the number of particles with velocity \mathbf{v} in this region, $f(\mathbf{r}, \mathbf{v}, t)\delta\mathbf{r}\delta\mathbf{v}$.
- Each $(\mathbf{v}_1, \mathbf{v})$ -collision cylinder has the volume given above, and the total volume of these cylinders is equal to the product of the volume of each such cylinder with the number of these cylinders, that is $f(\mathbf{r}, \mathbf{v}, t)|\mathbf{v}_1 - \mathbf{v}|\pi a^2\delta\mathbf{r}\delta\mathbf{v}\delta t$.
- If we wish to know the number of $(\mathbf{v}_1, \mathbf{v})$ -collisions that actually take place in this small time interval, we need to know exactly where each particle is located and then follow the motion of *all* the particles from time t to time $t + \delta t$. In fact, this is what is done in computer simulated molecular dynamics. We wish to avoid this exact specification of the particle trajectories, and instead carry out a plausible argument for the computation of Γ_- . To do this, Boltzmann made the following assumption, called the *Stosszahlansatz*, which we encountered already in the calculation of the mean free path:

-17-

Stosszahlansatz. The total number of $(\mathbf{v}_1, \mathbf{v})$ -collisions taking place in δt equals the total volume of the $(\mathbf{v}_1, \mathbf{v})$ -collision cylinders times the number of particles with velocity \mathbf{v}_1 per unit volume.

After integration over \mathbf{v}_1 , we obtain

$$\Gamma_- = \delta\mathbf{r}\delta\mathbf{v}f(\mathbf{r}, \mathbf{v}, t) \int d\mathbf{v}_1 \delta t \pi a^2 |\mathbf{v}_1 - \mathbf{v}| f(\mathbf{r}, \mathbf{v}_1, t). \quad (\text{A3.1.34})$$

The gas has to be dilute because the collision cylinders are assumed not to overlap, and also because collisions between more than two particles are neglected. Also it is assumed that f hardly changes over $\delta\mathbf{r}$ so that the distribution functions for both colliding particles can be taken at the same position \mathbf{r} .

The assumptions that go into the calculation of Γ_- are referred to collectively as the *assumption of molecular*

chaos. In this context, this assumption says that the probability that a pair of particles with given velocities will collide can be calculated by considering each particle separately and ignoring any correlation between the probability for finding one particle with velocity \mathbf{v} and the probability for finding another with velocity \mathbf{v}_1 in the region $\delta\mathbf{r}$.

For the construction of Γ_+ , we need to know how two particles can collide in such a way that one of them has velocity \mathbf{v} after the collision. The answer to this question can be found by a more careful examination of the ‘direct’ collisions which we have just discussed. To proceed with this examination, we note that the factor πa^2 appearing in (A3.1.34) can also be written as an integral over the impact parameters and azimuthal angles of the $(\mathbf{v}_1, \mathbf{v})$ collisions. That is, $\pi a^2 = \int b db \int d\varepsilon$, where b , the impact parameter, is the initial distance between the centre of the incoming \mathbf{v}_1 -particle and the axis of the collision cylinder (z -axis), and ε is the angle between the x -axis and the position of particle 2 in the x - y plane. Here $0 \leq b \leq a$, and $0 \leq \varepsilon \leq 2\pi$. The laws of conservation of linear momentum, angular momentum, and energy require that both the impact parameter b , and $|\mathbf{g}| = |\mathbf{v}_1 - \mathbf{v}|$, the magnitude of the relative velocity, be the same before and after the collision. To see what this means let us follow the two particles through and beyond a direct collision. We denote all quantities after the collision by primes. The conservation of momentum

$$\mathbf{v}_1 + \mathbf{v} = \mathbf{v}'_1 + \mathbf{v}'$$

implies, after squaring and using conservation of energy

$$v_1^2 + v^2 = v_1'^2 + v'^2$$

that

$$\mathbf{v}_1 \cdot \mathbf{v} = \mathbf{v}'_1 \cdot \mathbf{v}'.$$

By multiplying this result by a factor of -2 , and adding the result to the conservation of energy equation, one easily finds $|\mathbf{g}| = |\mathbf{g}'| = |\mathbf{v}'_1 - \mathbf{v}'|$. This result, taken together with conservation of angular momentum, $\mu g b = \mu g' b'$,

where $\mu = \frac{1}{2}m$ is the reduced mass of the two-particle system, shows that b is also conserved, $b = b'$. This is illustrated in figure A3.1.7.

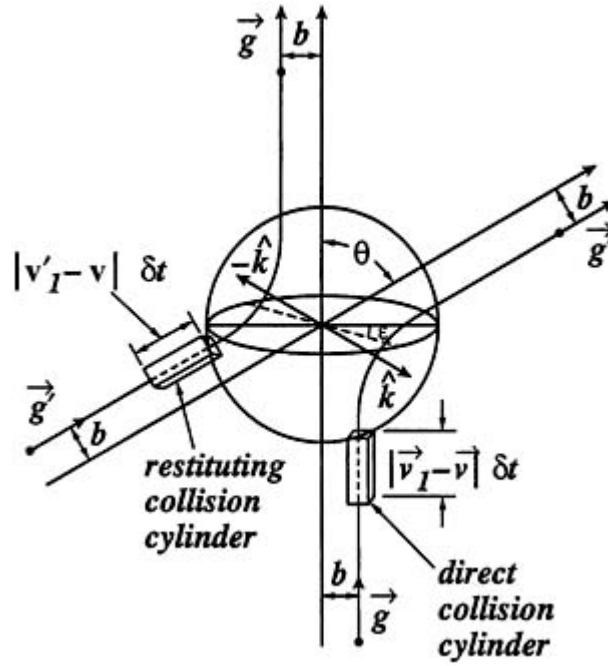


Figure A3.1.7. Direct and restituting collisions in the relative coordinate frame. The collision cylinders as well as the appropriate scattering and azimuthal angles are illustrated.

Next, we denote the line between the centres of the two particles at the point of closest approach by the unit vector $\hat{\mathbf{k}}$. In figure A3.1.7 it can also be seen that the vectors $-\mathbf{g}$ and \mathbf{g}' are each other's mirror images in the direction of $\hat{\mathbf{k}}$ in the plane of the trajectory of particles:

$$\mathbf{g}' = \mathbf{g} - 2(\mathbf{g} \cdot \hat{\mathbf{k}})\hat{\mathbf{k}} \quad (\text{A3.1.35})$$

and thus $(\mathbf{g} \cdot \hat{\mathbf{k}}) = -(\mathbf{g}' \cdot \hat{\mathbf{k}})$. Together with conservation of momentum this gives

$$\begin{aligned} \mathbf{v}'_1 &= \mathbf{v}_1 - (\mathbf{g} \cdot \hat{\mathbf{k}})\hat{\mathbf{k}} \\ \mathbf{v}' &= \mathbf{v} + (\mathbf{g} \cdot \hat{\mathbf{k}})\hat{\mathbf{k}}. \end{aligned} \quad (\text{A3.1.36})$$

The main point of this argument is to show that if particles with velocities \mathbf{v}' and \mathbf{v}'_1 collide in the right geometric configuration with impact parameter b , such a collision will result in one of the particles having the velocity of interest, \mathbf{v} , after the collision. These kinds of collisions which produce particles with velocity \mathbf{v} , contribute to Γ_+ , and are

referred to as 'restituting' collisions. This is illustrated in [figure A3.1.8](#) where particles having velocities \mathbf{v}' and \mathbf{v}'_1 are arranged to collide in such a way that the unit vector of closest approach, $\hat{\mathbf{k}}$, is replaced by $-\hat{\mathbf{k}}$. Consider, then, a collision with initial velocities \mathbf{v}'_1 and \mathbf{v}' and the same impact parameter as in the direct collision, but with $\hat{\mathbf{k}}$ replaced by $-\hat{\mathbf{k}}$. The final velocities are now \mathbf{v}''_1 and \mathbf{v}'' , which are equal to \mathbf{v}_1 and \mathbf{v} , respectively, because

$$\mathbf{v}_1'' = \mathbf{v}_1' - (\mathbf{g}' \cdot \hat{\mathbf{k}})\hat{\mathbf{k}} = \mathbf{v}_1' + (\mathbf{g} \cdot \hat{\mathbf{k}})\hat{\mathbf{k}} = \mathbf{v}_1 \quad (\text{A3.1.37})$$

and

$$\mathbf{v}'' = \mathbf{v}' - (\mathbf{g} \cdot \hat{\mathbf{k}})\hat{\mathbf{k}} = \mathbf{v}. \quad (\text{A3.1.38})$$

Thus the *increase* of particles in our region due to restituting collisions with an impact parameter between b and $b + db$ and azimuthal angle between ε and $\varepsilon + d\varepsilon$ (see [figure A3.1.7](#) can be obtained by adjusting the expression for the *decrease* of particles due to a ‘small’ collision cylinder:

$$\begin{aligned} \text{Loss: } & \delta t b db d\varepsilon |\mathbf{g}| f(\mathbf{r}, \mathbf{v}, t) f(\mathbf{r}, \mathbf{v}_1, t) \delta \mathbf{v} \delta \mathbf{v}_1 \delta \mathbf{r} \\ \text{Gain: } & \delta t b db d\varepsilon |\mathbf{g}'| f(\mathbf{r}, \mathbf{v}', t) f(\mathbf{r}, \mathbf{v}_1', t) \delta \mathbf{v}' \delta \mathbf{v}_1' \delta \mathbf{r} \end{aligned}$$

where b has to be integrated from 0 to a , and ε from 0 to 2π . Also, by considering the Jacobian for the transformation to relative and centre-of-mass velocities, one easily finds that $d\mathbf{v}_1 d\mathbf{v} = d\mathbf{v} d\mathbf{g}$, where \mathbf{v} is the velocity of the centre-of-mass of the two colliding particles with respect to the container. After a collision, \mathbf{g} is rotated in the centre-of-mass frame, so the Jacobian of the transformation $(\mathbf{v}, \mathbf{g}) \rightarrow (\mathbf{v}', \mathbf{g}')$ is unity and $d\mathbf{v} d\mathbf{g} = d\mathbf{v}' d\mathbf{g}'$. So

$$d\mathbf{v}_1 d\mathbf{v} = d\mathbf{v} d\mathbf{g} = d\mathbf{v}' d\mathbf{g}' = d\mathbf{v}'_1 d\mathbf{v}' \quad (\text{A3.1.39})$$

Now we are in the correct position to compute Γ_+ , using exactly the same kinds of arguments as in the computation of Γ_- , namely, the construction of collision cylinders, computing the total volume of the relevant cylinders and again making the *Stosszahlansatz*. Thus, we find that

$$\Gamma_+ = \iiint d\mathbf{v}'_1 b db d\varepsilon |\mathbf{v}'_1 - \mathbf{v}'| f(\mathbf{r}, \mathbf{v}', t) f(\mathbf{r}, \mathbf{v}'_1, t) \delta \mathbf{r} \delta \mathbf{v}' \delta t. \quad (\text{A3.1.40})$$

For every value of the velocity \mathbf{v} , the velocity ranges $d\mathbf{v}'_1 \delta \mathbf{v}'$ in the above expression are only over that range of velocities $\mathbf{v}', \mathbf{v}'_1$ such that particles with velocity in the range $\delta \mathbf{v}$ about \mathbf{v} are produced in the $(\mathbf{v}', \mathbf{v}'_1)$ -collisions. If we now use the equalities, equation (A3.1.39), as well as the fact that $|\mathbf{g}| = |\mathbf{g}'|$, we can write

-20-

$$\Gamma_+ = \iiint d\mathbf{v}_1 b db d\varepsilon |\mathbf{v}_1 - \mathbf{v}| f(\mathbf{r}, \mathbf{v}', t) f(\mathbf{r}, \mathbf{v}'_1, t) \delta \mathbf{r} \delta \mathbf{v} \delta t. \quad (\text{A3.1.41})$$

The term describing the interaction with the walls, Γ_w , is discussed in a paper by Dorfman and van Beijeren [14].

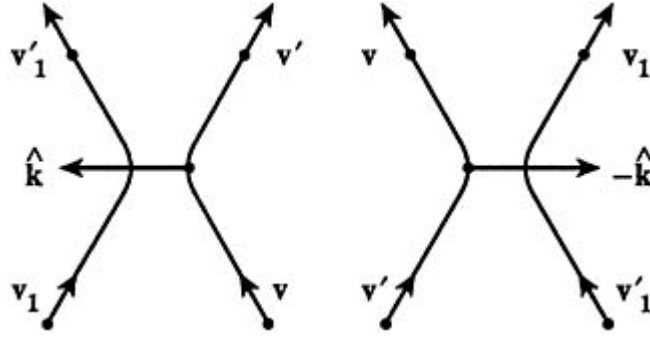


Figure A3.1.8. Schematic illustration of the direct and restituting collisions.

Finally, all of the Γ -terms can be inserted in (A3.1.31), and dividing by $\delta t \delta \mathbf{r} \delta \mathbf{v}$ gives the Boltzmann transport equation

$$\frac{\partial f(\mathbf{r}, \mathbf{v}, t)}{\partial t} + \mathbf{v} \cdot \nabla f(\mathbf{r}, \mathbf{v}, t) = J(f, f) + T_w \quad (\text{A3.1.42})$$

where $J(f, f) = \iiint d\mathbf{v}_1 b db d\epsilon |\mathbf{v}_1 - \mathbf{v}| [f f'_1 - f_1 f]$.

The primes and subscripts on the f s refer to their velocity arguments, and the primed velocities in the gain term should be regarded as functions of the unprimed quantities according to (A3.1.36). It is often convenient to rewrite the integral over the impact parameter and the azimuthal angle as an integral over the unit vector $\hat{\mathbf{k}}$ as

$$g b db d\epsilon = B(\mathbf{g}, \hat{\mathbf{k}}) d\hat{\mathbf{k}} \quad (\text{A3.1.43})$$

where

$$d\hat{\mathbf{k}} = \sin(\pi - \psi) d(\pi - \psi) d\epsilon \quad (\text{A3.1.44})$$

and ψ is the angle between \mathbf{g} and $\hat{\mathbf{k}}$. Then $d\hat{\mathbf{k}} = |\sin \psi d\psi d\epsilon|$, so that

$$B(\mathbf{g}, \hat{\mathbf{k}}) = |\mathbf{v}_1 - \mathbf{v}| \left| \frac{b}{\sin \psi} \right| \left| \frac{db}{d\psi} \right| \quad (\text{A3.1.45})$$

with the restriction for purely repulsive potentials that $\mathbf{g} \cdot \hat{\mathbf{k}} < 0$. As can be seen in figure A3.1.7.

$$B(\mathbf{g}', \hat{\mathbf{k}}) = B(\mathbf{g}, -\hat{\mathbf{k}}). \quad (\text{A3.1.46})$$

Let us apply this to the situation where the molecules are hard spheres of *diameter* a . We have $db/d\psi = d(a \sin$

$\psi)/d\psi = a \cos \psi$ (see figure A3.1.9 , and $B(\mathbf{g}, \hat{\mathbf{k}}) = ga^2 \cos \psi = a^2|(\mathbf{g} \cdot \hat{\mathbf{k}})|$.

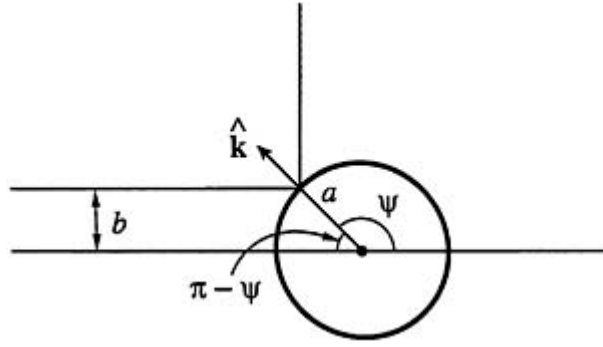


Figure A3.1.9. Hard sphere collision geometry in the plane of the collision. Here a is the diameter of the spheres.

The Boltzmann equation for hard spheres is given then as

$$\frac{\partial f}{\partial t} + (\mathbf{v} \cdot \nabla f) = a^2 \int d\mathbf{v}_1 \int_{\mathbf{g} \cdot \hat{\mathbf{k}} < 0} d\hat{\mathbf{k}} |\mathbf{g} \cdot \hat{\mathbf{k}}| [f'_1 f' - f_1 f]. \quad (\text{A3.1.47})$$

This completes the heuristic derivation of the Boltzmann transport equation. Now we turn to Boltzmann's argument that his equation implies the Clausius form of the second law of thermodynamics, namely, that the entropy of an isolated system will increase as the result of any irreversible process taking place in the system. This result is referred to as *Boltzmann's H-theorem*.

A3.1.3.1 BOLTZMANN'S H-THEOREM

Boltzmann showed that under very general circumstances, there exists a time-dependent quantity, $H(t)$, that never increases in the course of time. This quantity is given by

-22-

$$H(t) \equiv \int d\mathbf{r}_1 \int d\mathbf{v}_1 f(\mathbf{r}_1, \mathbf{v}_1, t) [\ln f(\mathbf{r}_1, \mathbf{v}_1, t) - 1]. \quad (\text{A3.1.48})$$

Here, the spatial integral is to be carried out over the entire volume of the vessel containing the gas, and for convenience we have changed the notation slightly. Now we differentiate H with time,

$$\frac{dH(t)}{dt} = \int d\mathbf{r}_1 \int d\mathbf{v}_1 \frac{\partial f_1}{\partial t} \ln f_1 \quad (\text{A3.1.49})$$

and use the Boltzmann equation to find that

$$\frac{dH}{dt} = \iint d\mathbf{r}_1 d\mathbf{v}_1 [-(\mathbf{v}_1 \cdot \nabla f_1) + J(f_1, f_1) + T_w] \ln f_1. \quad (\text{A3.1.50})$$

We are going to carry out some spatial integrations here. We suppose that the distribution function vanishes at the surface of the container and that there is no flow of energy or momentum into or out of the container. (We mention in passing that it is possible to relax this latter condition and thereby obtain a more general form of the second law than we discuss here. This requires a careful analysis of the wall-collision term T_w . The interested reader is referred to the article by Dorfman and van Beijeren [14]. Here, we will drop the wall operator since for the purposes of this discussion it merely ensures that the distribution function vanishes at the surface of the container.) The first term can be written as

$$- \iint \mathbf{dr}_1 \mathbf{dv}_1 \mathbf{v}_1 \cdot \nabla [f_1 (\ln f_1 - 1)]. \quad (\text{A3.1.51})$$

This can be evaluated easily in terms of the distribution function at the walls of the closed container and therefore it is zero. The second term of (A3.1.50) is based on the *Stosszahlansatz*, and is

$$\frac{dH(t)}{dt} = \iiint \mathbf{dr}_1 \mathbf{dv}_1 \mathbf{dv}_2 \mathbf{d}\hat{\mathbf{k}} B(\mathbf{g}, \hat{\mathbf{k}}) \Psi(\mathbf{v}_1) (f'_1 f'_2 - f_1 f_2) \quad (\text{A3.1.52})$$

with $\Psi(\mathbf{v}_1) = \ln f_1$. The integrand may be symmetrized in \mathbf{v}_1 and \mathbf{v}_2 to give

$$\frac{dH(t)}{dt} = \frac{1}{2} \iiint \mathbf{dr}_1 \mathbf{dv}_1 \mathbf{dv}_2 \mathbf{d}\hat{\mathbf{k}} B(\mathbf{g}, \hat{\mathbf{k}}) [\Psi(\mathbf{v}_1) + \Psi(\mathbf{v}_2)] (f'_1 f'_2 - f_1 f_2).$$

For each collision there is an inverse one, so we can also express the time derivative of the H -function in terms of the inverse collisions as

-23-

$$\begin{aligned} \frac{dH(t)}{dt} &= +\frac{1}{2} \iiint \mathbf{dr}_1 \mathbf{dv}'_1 \mathbf{dv}'_2 \mathbf{d}\hat{\mathbf{k}} B(\mathbf{g}', -\hat{\mathbf{k}}) [\Psi(\mathbf{v}'_1) + \Psi(\mathbf{v}'_2)] (f_1 f_2 - f'_1 f'_2) \\ &= -\frac{1}{2} \iiint \mathbf{dr}_1 \mathbf{dv}_1 \mathbf{dv}_2 \mathbf{d}\hat{\mathbf{k}} B(\mathbf{g}, \hat{\mathbf{k}}) [\Psi(\mathbf{v}'_1) + \Psi(\mathbf{v}'_2)] (f'_1 f'_2 - f_1 f_2). \end{aligned}$$

We obtain the H -theorem by adding these expressions and dividing by two,

$$\frac{dH(t)}{dt} = \frac{1}{4} \iiint \mathbf{dr}_1 \mathbf{dv}_1 \mathbf{dv}_2 \mathbf{d}\hat{\mathbf{k}} B(\mathbf{g}, \hat{\mathbf{k}}) [\Psi_1 + \Psi_2 - \Psi'_1 - \Psi'_2] (f'_1 f'_2 - f_1 f_2).$$

Now, using $\Psi(\mathbf{v}_1) = \ln f_1$, we obtain

$$\frac{dH}{dt} = \frac{1}{4} \iiint \mathbf{dr}_1 \mathbf{dv}_1 \mathbf{dv}_2 \mathbf{d}\hat{\mathbf{k}} B(\mathbf{g}, \hat{\mathbf{k}}) \ln \left(\frac{f_1 f_2}{f'_1 f'_2} \right) (f'_1 f'_2 - f_1 f_2).$$

If $f_1 f_2 \neq f'_1 f'_2$, the integrand is negative;

$$\begin{aligned} f_1 f_2 < f'_1 f'_2 & \quad \text{the second factor is positive, the first is negative;} \\ f_1 f_2 > f'_1 f'_2 & \quad \text{the second is negative, and the first is positive.} \end{aligned}$$

Both cases give a decreasing $H(t)$. That is

$$\frac{dH(t)}{dt} \leq 0. \quad (\text{A3.1.53})$$

The integral is zero only if for all \mathbf{v}_1 and \mathbf{v}_2

$$f_1 f_2 = f'_1 f'_2. \quad (\text{A3.1.54})$$

This is Boltzmann's H -theorem

We now show that when H is constant in time, the gas is in equilibrium. The existence of an equilibrium state requires the rates of the restituting and direct collisions to be equal; that is, that there is a detailed balance of gain and loss processes taking place in the gas.

Taking the natural logarithm of (A3.1.54), we see that $\ln f_1 + \ln f_2$ has to be conserved for an equilibrium solution of the Boltzmann equation. Therefore, $\ln f_1$ can generally be expressed as a linear combination with constant coefficients

-24-

of the $(d + 2)$ quantities conserved by binary collisions, i.e. (i) the number of particles, (ii) the d components of the linear momentum, where d is the number of dimensions, and (iii) the kinetic energy: $\ln f_1 = \mathbf{A} + \mathbf{B} \cdot \mathbf{v}_1 + C v_1^2$. (Adding an angular momentum term to $\ln(f_1 f_2)$ is not independent of conservation of momentum, because the positions of the particles are the same.) The particles are assumed to have no internal degrees of freedom. Then

$$f_1 \propto \exp(\mathbf{B} \cdot \mathbf{v}_1 + C v_1^2) = A \exp[-\frac{1}{2} \beta m (\mathbf{v}_1 - \mathbf{u})^2]. \quad (\text{A3.1.55})$$

When H has reached its minimum value this is the well known Maxwell–Boltzmann distribution for a gas in thermal equilibrium with a uniform motion \mathbf{u} . So, argues Boltzmann, solutions of his equation for an isolated system approach an equilibrium state, just as real gases seem to do. Up to a negative factor ($-k_B$, in fact), differences in H are the same as differences in the thermodynamic entropy between initial and final equilibrium states. Boltzmann thought that his H -theorem gave a foundation of the increase in entropy as a result of the collision integral, whose derivation was based on the *Stosszahlansatz*.

(A) THE REVERSIBILITY AND THE RECURRENCE PARADOXES

Boltzmann's H -theorem raises a number of questions, particularly the central one: how can a gas that is described exactly by the reversible laws of mechanics be characterized by a quantity that always decreases? Perhaps a *non-mechanical* assumption was introduced here. If so, this would suggest, although not imply, that Boltzmann's equation might not be a useful description of nature. In fact, though, this equation is so useful

and accurate a predictor of the properties of dilute gases, that it is now often used as a test of intermolecular potential models.

The question stated above was formulated in two ways, each using an exact result from classical mechanics. One way, associated with the physicist Loschmidt, is fairly obvious. If classical mechanics provides a correct description of the gas, then associated with any physical motion of a gas, there is a time-reversed motion, which is also a solution of Newton's equations. Therefore if H decreases in one of these motions, there ought to be a physical motion of the gas where H increases. This is contrary to the H -theorem. The other objection is based on the recurrence theorem of Poincare [15], and is associated with the mathematician Zermelo. Poincare's theorem states that in a bounded mechanical system with finite energy, any initial state of the gas will eventually recur as a state of the gas, to within any preassigned accuracy. Thus, if H decreases during part of the motion, it must eventually increase so as to approach, arbitrarily closely, its initial value.

The recurrence paradox is easy to refute and was done so by Boltzmann. He pointed out that the recurrence time even for a system of a several particles, much less a system of 10^{23} particles, is so enormously long (orders of magnitude larger than the age of the universe) that one will never live long enough to observe a recurrence. The usual response to Loschmidt is to argue that while the gas is indeed a mechanical system, almost all initial states of the gas one is likely to encounter in the laboratory will show an approach to equilibrium as described by the H -theorem. That is, the Boltzmann equation describes the most typical behaviour of a gas. While an anti-Boltzmann-like behaviour is not ruled out by mechanics, it is very unlikely, in a statistical sense, since such a motion would require a very careful (to put it mildly) preparation of the initial state. Thus, the reversibility paradox is more subtle, and the analysis of it eventually led Boltzmann to the very fruitful idea of an ergodic system [16]. In any case, there is no reason to doubt the validity of the Boltzmann equation for the description of irreversible processes in dilute gases. It describes the typical behaviour of a laboratory system, while any individual system may have small fluctuations about this typical behaviour.

A3.1.3.2 THE CHAPMAN-ENSKOG NORMAL SOLUTIONS OF THE BOLTZMANN EQUATION

The practical value of the Boltzmann equation resides in the utility of the predictions that one can obtain from it. The form of the Boltzmann is such that it can be used to treat systems with long range forces, such as Lennard-Jones particles, as well as systems with finite-range forces. Given a potential energy function, one can calculate the necessary collision cross sections as well as the various restituting velocities well enough to derive practical expressions for transport coefficients from the Boltzmann equation. The method for obtaining solutions of the equation used for fluid dynamics is due to Enskog and Chapman, and proceeds by finding solutions that can be expanded in a series whose first term is a Maxwell-Boltzmann distribution of local equilibrium form. That is, the first takes the form given by (A3.1.55), with the quantities A , β and \mathbf{u} being functions of \mathbf{r} and t . One then assumes that the local temperature, $(k_B\beta)^{-1}$, mean velocity, \mathbf{u} , and local density, n , are slowly varying in space and time, where the distance over which they change, L , say is large compared with a mean free path, λ . The higher terms in the Chapman-Enskog solution are then expressed in a power series in gradients of the five variables, n , β and \mathbf{u} , which can be shown to be an expansion in powers of $l/L \ll 1$. Explicit results are then obtained for the first, and higher, order solution in l/L , which in turn lead to Navier-Stokes as well as higher order hydrodynamic equations. Explicit expressions are obtained for the various transport coefficients, which can then be compared with experimental data. The agreement is sufficiently close that the theoretical results provide a useful way for checking the accuracy of various trial potential energy functions. A complete account of the Chapman-Enskog solution method can be found in the book by Chapman and Cowling [3], and comparisons with experiments, the extension to polyatomic molecules, and to quantum gases, are discussed at some length in the books of Hirshfelder *et al* [4], of Hanley [5] and of Kestin [17] as well as in an enormous literature.

A3.1.3.3 EXTENSION OF THE BOLTZMANN EQUATION TO HIGHER DENSITIES

It took well over three quarters of a century for kinetic theory to develop to the point that a systematic extension of the Boltzmann equation to higher densities could be found. This was due to the work of Bogoliubov, Green and Cohen, who borrowed methods from equilibrium statistical mechanics, particularly the use of cluster expansion techniques, to obtain a virial expansion of the right-hand side of the Boltzmann equation to include the effects of collisions among three, four and higher numbers of particles, successively. However, this virial expansion was soon found to diverge term-by-term in the density, beyond the three-body term in three dimensions, and beyond the two-body term in two dimensions. In order to obtain a well behaved generalized Boltzmann equation one has to sum these divergent terms. This has been accomplished using various methods. One finds that the transport coefficients for moderately dense gases cannot be expressed strictly in a power series of the gas density, but small logarithmic terms in the density also appear. Moreover, one finds that long-range correlations exist in a non-equilibrium gas, that make themselves felt in the effects of light scattering by a dense fluid with gradients in temperature or local velocity. Reviews of the theory of non-equilibrium processes in dense gases can be found in articles by Dorfman and van Beijeren [14], by Cohen [18] and by Ernst [19], as well as in the book of Resibois and de Leener [7].

A3.1.4 FLUCTUATIONS IN GASES

Statistical mechanics and kinetic theory, as we have seen, are typically concerned with the average behaviour of an ensemble of similarly prepared systems. One usually hopes, and occasionally can demonstrate, that the variations of these properties from one system to another in the ensemble, or that the variation with time of the properties of any

-26-

one system, are very small. There is a well developed theory for equilibrium fluctuations of these types. In this theory one can relate, for example, the specific heat at constant volume of a system in contact with a thermal reservoir to the mean square fluctuations of the energy of the system. It is also well known that the scattering of light by a fluid system is determined by the density fluctuations in the system, caused by the motion of the particles in the system. These thermal fluctuations in density, temperature and other equilibrium properties of the system typically scale to zero as the number of degrees of freedom in the system becomes infinite. A good account of these equilibrium fluctuations can be found in the text by Landau and Lifshitz [20].

When a system is not in equilibrium, the mathematical description of fluctuations about some time-dependent ensemble average can become much more complicated than in the equilibrium case. However, starting with the pioneering work of Einstein on Brownian motion in 1905, considerable progress has been made in understanding time-dependent fluctuation phenomena in fluids. Modern treatments of this topic may be found in the texts by Keizer [21] and by van Kampen [22]. Nevertheless, the non-equilibrium theory is not yet at the same level of rigour or development as the equilibrium theory. Here we will discuss the theory of Brownian motion since it illustrates a number of important issues that appear in more general theories.

We consider the motion of a large particle in a fluid composed of lighter, smaller particles. We also suppose that the mean free path of the particles in the fluid, λ , is much smaller than a characteristic size, R , of the large particle. The analysis of the motion of the large particle is based upon a method due to Langevin. Consider the equation of motion of the large particle. We write it in the form

$$M \frac{d\mathbf{v}(t)}{dt} = -\zeta \mathbf{v}(t) + \mathbf{F}(t) \quad (\text{A3.1.56})$$

where M is the mass of the large particle, $\mathbf{v}(t)$ is its velocity at time t , the quantity $-\zeta \mathbf{v}(t)$ represents the hydrodynamic friction exerted by the fluid on the particle, while the term $\mathbf{F}(t)$ represents the fluctuations in the force on the particle produced by the discontinuous nature of the collisions with the fluid particles. If the Brownian particle is spherical, with radius R , then ζ is usually taken to have the form provided by Stokes' law of friction on a slowly moving particle by a continuum fluid,

$$\zeta = 6\pi \eta R \quad (\text{A3.1.57})$$

where η is the shear viscosity of the fluid. The fact that the fluid is not a continuum is incorporated in the fluctuating force $\mathbf{F}(t)$. This fluctuating force is taken to have the following properties

$$\langle \mathbf{F}(t) \rangle = \mathbf{0} \quad (\text{A3.1.58})$$

$$\langle \mathbf{F}_i(t_1) \mathbf{F}_j(t_2) \rangle = A \delta(t_1 - t_2) \delta_{\text{Kr}}(i, j) \quad (\text{A3.1.59})$$

where A is some constant yet to be determined, $\delta(t_1 - t_2)$ is a Dirac delta function in the time interval between t_1 and t_2 , and $\delta_{\text{Kr}}(i, j)$ is a Kronecker delta function in the components of the fluctuating force in the directions denoted by i, j .

-27-

The angular brackets denote an average, but the averaging process is somewhat subtle, and discussed in some detail in the book of van Kampen [22]. For our purposes, we will take the average to be over an ensemble constructed by following a long trajectory of one Brownian particle, cutting the trajectory into a large number of smaller, disjoint pieces, all of the same length, and then taking averages over all of the pieces. Thus the pieces of the long trajectory define the ensemble over which we average. The delta function correlation in the random force assumes that the collisions of the Brownian particle with the particles of the fluid are instantaneous and uncorrelated with each other.

If we now average the Langevin equation, (A3.1.56), we obtain a very simple equation for $\langle \mathbf{v}(t) \rangle$, whose solution is clearly

$$\langle \mathbf{v}(t) \rangle = \langle \mathbf{v}(0) \rangle e^{-\zeta t/M}, \quad (\text{A3.1.60})$$

Thus the average velocity decays exponentially to zero on a time scale determined by the friction coefficient and the mass of the particle. This average behaviour is not very interesting, because it corresponds to the average of a quantity that may take values in all directions, due to the noise and friction, and so the decay of the average value tells us little about the details of the motion of the Brownian particle. A more interesting quantity is the mean square velocity, $\langle \mathbf{v}(t)^2 \rangle$, obtained by solving (A3.1.56), squaring the solution and then averaging. Here the correlation function of the random force plays a central role, for we find

$$\langle \mathbf{v}^2(t) \rangle = \langle \mathbf{v}^2(0) \rangle e^{-2\zeta t/M} + \frac{3A}{2\zeta} (1 - e^{-2\zeta t/M}). \quad (\text{A3.1.61})$$

Notice that this quantity does not decay to zero as time becomes long, but rather it reaches the value

$$\langle \mathbf{v}^2(t) \rangle \rightarrow \frac{3A}{2\zeta} \quad (\text{A3.1.62})$$

as $t \rightarrow \infty$. Here we have the first appearance of A , the coefficient of the correlation of the random force. It is reasonable to suppose that in the infinite-time limit, the Brownian particle becomes equilibrated with the surrounding fluid. This means that the average kinetic energy of the Brownian particle should approach the value $3k_{\text{B}}T/2$ as time gets large. This is consistent with (A3.1.62), if the coefficient A has the value

$$A = \frac{2\zeta k_{\text{B}}T}{M}. \quad (\text{A3.1.63})$$

Thus, the requirement that the Brownian particle becomes equilibrated with the surrounding fluid fixes the unknown value of A , and provides an expression for it in terms of the friction coefficient, the thermodynamic temperature of the fluid, and the mass of the Brownian particle. Equation (A3.1.63) is the simplest and best known example of a *fluctuation–dissipation theorem*, obtained by using an equilibrium condition to relate the strength of the fluctuations to the frictional forces acting on the particle [22].

-28-

Two more important ideas can be illustrated by means of the Langevin approach to Brownian motion. The first result comes from a further integration of the velocity equation to find an expression for the fluctuating displacement of the moving particle, and for the mean square displacement as a function of time. By carrying out the relevant integrals and using the fluctuation–dissipation theorem, we can readily see that the mean square displacement, $\langle (\mathbf{r}(t) - \mathbf{r}(0))^2 \rangle$, grows linearly in time t , for large times, as

$$\langle (\mathbf{r}(t) - \mathbf{r}(0))^2 \rangle = \frac{6k_{\text{B}}TM}{\zeta} t. \quad (\text{A3.1.64})$$

Now for a particle undergoing diffusion, it is also known that its mean square displacement grows linearly in time, for long times, as

$$\langle (\mathbf{r}(t) - \mathbf{r}(0))^2 \rangle = 6Dt \quad (\text{A3.1.65})$$

where D is the diffusion coefficient of the particle. By comparing (A3.1.64) and (A3.1.65) we see that

$$D = \frac{k_{\text{B}}TM}{\zeta}. \quad (\text{A3.1.66})$$

This result is often called the *Stokes–Einstein formula* for the diffusion of a Brownian particle, and the Stokes' law friction coefficient $6\pi\eta R$ is used for ζ .

The final result that we wish to present in this connection is an example of the Green–Kubo time-correlation expressions for transport coefficients. These expressions relate the transport coefficients of a fluid, such as

viscosity, thermal conductivity, etc, in terms of time integrals of some average time-correlation of an appropriate microscopic variable. For example, if we were to compute the time correlation function of one component of the velocity of the Brownian particle, $\langle v_x(t_1)v_x(t_2) \rangle$, we would obtain

$$\langle v_x(t_1)v_x(t_2) \rangle = k_B T e^{-\zeta|t_1-t_2|/M} \quad (\text{A3.1.67})$$

for large times, neglecting factors that decay exponentially in both t_1 and t_2 . The Green–Kubo formula for diffusion relates the diffusion coefficient, D to the time integral of the time-correlation of the velocity through

$$D = \int_0^\infty dt \langle v_x(0)v_x(t) \rangle \quad (\text{A3.1.68})$$

a result which clearly reproduces (A3.1.66). The Green–Kubo formulae are of great interest in kinetic theory and non-equilibrium statistical mechanisms since they provide a new set of functions, the time-correlation functions,

that tell us more about the microscopic properties of the fluid than do the transport coefficients themselves, and that are very useful for analysing fluid behaviour when making computer simulations of the fluid.

REFERENCES

- [1] Brush S 1966–1972 *Kinetic Theory* vols 1–3 (New York; Pergamon)
- [2] Boltzmann L 1995 *Lectures on Gas Theory* translator S Brush (New York: Dover)
- [3] Chapman S and Cowling T G 1970 *The Mathematical Theory of Non-Uniform Gases* 3rd edn (Cambridge: Cambridge University Press)
- [4] Hirschfelder J O, Curtiss C F and Bird R B 1954 *Molecular Theory of Gases and Liquids* (New York: Wiley)
- [5] Hanley H J M 1970 *Transport Phenomena in Fluids* (New York: Marcel Dekker)
- [6] Fertziger J H and Kaper H G 1972 *Mathematical Theory of Transport Processes in Gases* (Amsterdam: North Holland)
- [7] Resibois P and de Leener M 1977 *Classical Kinetic Theory of Fluids* (New York: Wiley)
- [8] Liboff R L 1998 *Kinetic Theory: Classical, Quantum, and Relativistic Descriptions* 2nd edn (New York: Wiley)
- [9] Present R D 1958 *Kinetic Theory of Gases* (New York: McGraw-Hill)
- [10] McQuarrie D A 1976 *Statistical Mechanics* (New York: Harper and Row)
- [11] Kestin J and Dorfman J R 1970 *A Course in Statistical Thermodynamics* (New York: Academic)
- [12] Huang K 1990 *Statistical Mechanics* 2nd edn (New York: Wiley)
- [13] Wannier G 1987 *Statistical Mechanics* (New York: Dover)
- [14] Dorfman J R and van Beijeren H 1977 The kinetic theory of gases *Statistical Mechanics, Part B: Time-Dependent Processes* ed B J Berne (New York: Plenum)

- [15] Uhlenbeck G E and Ford G W 1963 *Lectures in Statistical Mechanics* (Providence, RI: American Mathematical Society)
 - [16] Arnold V I and Avez A 1968 *Ergodic Problems of Classical Mechanics* (New York: Benjamin)
 - [17] Kestin J (ed) 1973 *Transport Phenomena, AIP Conference Proceedings No. 11* (New York: American Institute of Physics)
 - [18] Cohen E G D 1993 Fifty years of kinetic theory *Physica A* **194** 229
 - [19] Ernst M H 1998 Bogoliubov–Choh–Uhlenbeck theory: cradle of modern kinetic theory *Progress in Statistical Physics* ed W Sung *et al* (Singapore: World Scientific)
 - [20] Landau L D and Lifshitz E M 1980 *Statistical Physics, Part I* 3rd edn, translators J B Sykes and M J Kearney (Oxford: Pergamon)
 - [21] Keizer J 1987 *Statistical Thermodynamics of Nonequilibrium Processes* (New York: Springer)
-

-1-

A3.2 Non-equilibrium thermodynamics

Ronald F Fox

A3.2.1 INTRODUCTION

Equilibrium thermodynamics may be developed as an autonomous macroscopic theory or it may be derived from microscopic statistical mechanics. The intrinsic beauty of the macroscopic approach is partially lost with the second treatment. Its beauty lies in its internal consistency. The advantage of the second treatment is that certain quantities are given explicit formulae in terms of fundamental constants, whereas the purely macroscopic approach must use measurements to determine these same quantities. The Stefan–Boltzmann constant is a prime example of this dichotomy. Using purely macroscopic thermodynamic arguments, Boltzmann showed that the energy density emitted per second from a unit surface of a black body is σT^4 where T is the temperature and σ is the Stefan–Boltzmann constant, but it takes statistical mechanics to produce the formula

$$\sigma = \frac{2\pi^5 k^4}{15c^2 h^3}$$

in which k is Boltzmann’s constant, c is the speed of light and h is Planck’s constant. This beautiful formula depends on three fundamental constants, an exhibition of the power of the microscopic viewpoint. Likewise, non-equilibrium thermodynamics may be developed as a purely autonomous macroscopic theory or it may be derived from microscopic kinetic theory, either classically or quantum mechanically. The separation between the macroscopic and microscopic approaches is a little less marked than for the equilibrium theory because the existence of the microscopic underpinning leads to the existence of fluctuations in the macroscopic picture, as well as to the celebrated Onsager reciprocal relations. On purely macroscopic grounds, the fluctuation–dissipation relation that connects the relaxation rates to the strengths of the fluctuations may be established, but it takes the full microscopic theory to compute their quantitative values, at least in principle. In practice, these computations are very difficult. This presentation is primarily about the macroscopic approach, although at the end the microscopic approach, based on linear response theory, is also reviewed.

The foundations for the macroscopic approach to non-equilibrium thermodynamics are found in Einstein's theory of Brownian movement [1] of 1905 and in the equation of Langevin [2] of 1908. Uhlenbeck and Ornstein [3] generalized these ideas in 1930 and Onsager [4, 5] presented his theory of irreversible processes in 1931. Onsager's theory [4] was initially deterministic with little mention of fluctuations. His second paper [5] used fluctuation theory to establish the reciprocal relations. The fundamental role of fluctuations was generalized by the works of Chandrasekhar [6] in 1943, of Wang and Uhlenbeck [7] in 1945, of Casimir [8] in 1945, of Prigogine [9] in 1947 and of Onsager and Machlup [10, 11] in 1953. In 1962 de Groot and Mazur [12] published their definitive treatise *Non-Equilibrium Thermodynamics* which greatly extended the applicability of the theory as well as deepening its foundations. By this time, it was clearly recognized that the mathematical setting for non-equilibrium thermodynamics is the theory of stationary, Gaussian–Markov processes. The Onsager reciprocal relations may be most easily understood within this context. Nevertheless, the issue of the most general form for stationary, Gaussian–Markov processes, although

-2-

broached by de Groot and Mazur [12], was not solved until the work of Fox and Uhlenbeck [13, 14 and 15] in 1969. For example, this work made it possible to rigorously extend the Onsager theory to hydrodynamics and to the Boltzmann equation.

A3.2.2 GENERAL STATIONARY GAUSSIAN–MARKOV PROCESSES

A3.2.2.1 CHARACTERIZATION OF RANDOM PROCESSES

Let $a(t)$ denote a time dependent random process. $a(t)$ is a random process because at time t the value of $a(t)$ is not definitely known but is given instead by a probability distribution function $W_1(a, t)$ where a is the value $a(t)$ can have at time t with probability determined by $W_1(a, t)$. $W_1(a, t)$ is the first of an infinite collection of distribution functions describing the process $a(t)$ [7, 13]. The first two are defined by

$W_1(a, t) da$ = probability at time t that the value of $a(t)$ is between a and $a + da$

$W_2(a_1, t_1; a_2, t_2) da_1 da_2$ = probability that at time t_1 the value of $a(t)$ is between a_1 and $a_1 + da_1$ and that at time t_2 the value of $a(t)$ is between a_2 and $a_2 + da_2$.

The higher order distributions are defined analogously. W_2 contains W_1 through the identity

$$W_1(a_1, t_1) = \int da_2 W_2(a_1, t_1; a_2, t_2).$$

Similar relations hold for the higher distributions. The Gaussian property of the process implies that all statistical information is contained in just W_1 and W_2 .

The condition that the process $a(t)$ is a *stationary* process is equivalent to the requirement that all the distribution functions for $a(t)$ are invariant under time translations. This has as a consequence that $W_1(a, t)$ is independent of t and that $W_2(a_1, t_1; a_2, t_2)$ only depends on $t = t_2 - t_1$. An *even* stationary process [4] has the additional requirement that its distribution functions are invariant under time reflection. For W_2 , this implies $W_2(a_1; a_2, t) = W_2(a_2; a_1, t)$. This is called *microscopic reversibility*. It means that the quantities are even functions of the particle velocities [12]. It is also possible that the variables are odd functions of the particle velocities [8], say, b_1 and b_2 for which $W_2(b_1; b_2, t) = W_2(-b_2; -b_1, t)$. In the general case considered later, the thermodynamic quantities are a mixture of even and odd [14, 15]. For the odd case, the presence of a

magnetic field, \mathbf{B} , or a coriolis force depending on angular velocity ω , requires that \mathbf{B} and ω also change sign during time reversal and microscopic reversibility reads as $W_2(b_1; b_2, \mathbf{B}, \omega t) = W_2(-b_2; -b_1, -\mathbf{B}, -\omega, t)$. Examples of even processes include heat conduction, electrical conduction, diffusion and chemical reactions [4]. Examples of odd processes include the Hall effect [12] and rotating frames of reference [4]. Examples of the general setting that lacks even or odd symmetry include hydrodynamics [14] and the Boltzmann equation [15].

Before defining a *Markov* process $a(t)$, it is necessary to introduce *conditional* probability distribution functions $P_2(a_1, t_1 | a_2, t_2)$ and $P_3(a_1, t_1; a_2, t_2 | a_3, t_3)$ defined by

-3-

$P_2(a_1, t_1 | a_2, t_2) da_2$ = probability at time t_2 that the value of $a(t)$ is between a_2 and $a_2 + da_2$ given that at time $t_1 < t_2$ $a(t)$ had the definite value a_1 .

$P_3(a_1, t_1; a_2, t_2 | a_3, t_3) da_3$ = probability at time t_3 that the value of $a(t)$ is between a_3 and $a_3 + da_3$ given that at time $t_2 < t_3$ $a(t)$ had the definite value a_2 and at time $t_1 < t_2$ $a(t)$ had the definite value a_1 .

These conditional distributions are related to the W_n by

$$\begin{aligned} W_2(a_1, t_1; a_2, t_2) &= W_1(a_1, t_1) P_2(a_1, t_1 | a_2, t_2) \\ W_2(a_1, t_1; a_2, t_2; a_3, t_3) &= W_2(a_1, t_1; a_2, t_2) P_3(a_1, t_1; a_2, t_2 | a_3, t_3) \end{aligned}$$

and so forth for the higher order distributions. The *Markov* property of $a(t)$ is defined by

$$P_2(a_2, t_2 | a_3, t_3) = P_3(a_1, t_1; a_2, t_2 | a_3, t_3) \quad (\text{A3.2.1})$$

which means that knowledge of the value of $a(t)$ at time t_1 does not influence the distribution of values of $a(t)$ at time $t_3 > t_1$ if there is also information giving the value of $a(t)$ at the intermediate time t_2 . Therefore, a *Markov* process is completely characterized by its W_1 and P_2 or equivalently by only W_2 . A stationary *Markov* process has distributions satisfying the Smoluchowski equation (also called the Chapman–Kolmogorov equation)

$$W_2(a_1; a_2 t) = \int da W_2(a_1; a, t - s) P_2(a | a_2, s) \quad \text{for all } s \in [0, t]. \quad (\text{A3.2.2})$$

Proof. For $t_1 < t_3 < t_2$ and using (A3.2.1)

$$\begin{aligned} W_2(a_1; a_2, t_2 - t_1) &= \int da_3 W_3(a_1, t_1; a_3, t_3; a_2, t_2) \\ &= \int da_3 W_2(a_1, t_1; a_3, t_3) P_3(a_1, t_1; a_3, t_3 | a_2, t_2) \\ &= \int da_3 W_2(a_1, t_1; a_3, t_3) P_2(a_3, t_3 | a_2, t_2). \end{aligned}$$

Setting $s = t_2 - t_3$ and $t = t_2 - t_1$ and $a_3 = a$ gives (A3.2.2) for a stationary process $a(t)$. QED

While the Smoluchowski equation is necessary for a Markov process, in general it is not sufficient, but known counter-examples are always non-Gaussian as well.

-4-

A3.2.2.2 THE LANGEVIN EQUATION

The prototype for all physical applications of stationary Gaussian–Markov processes is the treatment of Brownian movement using the Langevin equation [2, 3, 7]. The Langevin equation describes the time change of the velocity of a slowly moving colloidal particle in a fluid. The effect of the interactions between the particle and the fluid molecules produces two forces. One force is an average effect, the *frictional drag* that is proportional to the velocity, whereas the other force is a fluctuating force, $\tilde{F}(t)$, that has mean value zero. Therefore, a particle of mass M obeys the Langevin equation

$$M \frac{du}{dt} = -\alpha u + \tilde{F}(t) \quad (\text{A3.2.3})$$

where α is the frictional drag coefficient and u is the particle's velocity. It is the fluctuating force, $\tilde{F}(t)$, that makes u a random process. For a sphere of radius R in a fluid of viscosity η , $\alpha = 6\pi \eta R$, a result obtained from hydrodynamics by Stokes in 1854. To characterize this process it is necessary to make assumptions about $\tilde{F}(t)$. $\tilde{F}(t)$ is taken to be a stationary Gaussian process that is called *white noise*. This is defined by the correlation formula

$$\langle \tilde{F}(t) \tilde{F}(s) \rangle = 2\lambda \delta(t - s) \quad (\text{A3.2.4})$$

where $\langle \dots \rangle$ denotes averaging over $\tilde{F}(t)$, λ is a constant and the Dirac delta function of time expresses the quality of whiteness for the noise. The linearity of (A3.2.3) is sufficient to guarantee that u is also a stationary Gaussian process, although this claim requires some care as is shown below.

Equation (A3.2.3) must be solved with respect to some initial value for the velocity, $u(0)$. In the conditional distribution for the process $u(t)$, the initial value, $u(0)$, is denoted by u_0 giving $P_2(u_0 | u, t)$. Because $u(t)$ is a Gaussian process, $P_2(u_0 | u, t)$ is completely determined by the mean value of $u(t)$ and by its mean square.

Using (A3.2.4) and recalling that $\langle \tilde{F}(t) \rangle = 0$ it is easy to prove that

$$\langle u(t) \rangle = u(0) \exp \left[-\frac{\alpha}{M} t \right] \quad (\text{A3.2.5})$$

$$\langle u^2(t) \rangle = u^2(0) \exp \left[-\frac{2\alpha}{M} t \right] + \frac{\lambda}{\alpha M} \left(1 - \exp \left[-\frac{2\alpha}{M} t \right] \right) \quad (\text{A3.2.6})$$

$$\langle u(t)u(s) \rangle = u^2(0) \exp \left[-\frac{\alpha}{M} (t+s) \right] + \frac{\lambda}{\alpha M} \left(\exp \left[-\frac{\alpha}{M} |t-s| \right] - \exp \left[-\frac{\alpha}{M} (t+s) \right] \right) \quad (\text{A3.2.7})$$

Using $\sigma^2 = \lambda / \alpha M$ and $\rho(t) = \exp [-(\alpha/M)t]$, $P_2(u_0 | u, t)$ is given by

$$P_2(u_0 | u, t) = \frac{\exp[-(u - u_0\rho(t))^2/2\sigma^2(1 - \rho^2(t))]}{\sqrt{2\pi\sigma^2(1 - \rho^2(t))}} \quad (\text{A3.2.8})$$

which is checked by seeing that it reproduces (A3.2.5) and (A3.2.6). From (A3.2.8), it is easily seen that the $t \rightarrow \infty$ limit eliminates any influence of u_0

$$\lim_{t \rightarrow \infty} P_2(u_0 | u, t) = W_1(u) = \frac{\exp[-\alpha Mu^2/2\lambda]}{\sqrt{2\pi\lambda/\alpha M}}.$$

However, $W_1(u)$ should also be given by the equilibrium Maxwell distribution

$$W_1(u) = \frac{\exp[-Mu^2/2kT]}{\sqrt{2\pi kT/M}} \quad (\text{A3.2.9})$$

in which k is Boltzmann's constant and T is the equilibrium temperature of the fluid. The equality of these two expressions for $W_1(u)$ results in Einstein's relation

$$\lambda = kT\alpha. \quad (\text{A3.2.10})$$

Putting (A3.2.10) into (A3.2.4) gives the prototype example of what is called the *fluctuation-dissipation relation*

$$\langle \tilde{F}(t)\tilde{F}(s) \rangle = 2kT\alpha\delta(t - s).$$

Looking back at (A3.2.7), we see that a second average over $u(0)$ can be performed using $W_1(u_0)$. This second type of averaging is denoted by $\{ \dots \}$. Using (A3.2.9) we obtain

$$\{u^2(0)\} = \frac{kT}{M}$$

which with (A3.2.7) and (A3.2.10), the Einstein relation, implies

$$\{ \langle u(t)u(s) \rangle \} = \frac{kT}{M} \exp \left[-\frac{\alpha}{M}|t - s| \right].$$

This result clearly manifests the stationarity of the process that is not yet evident in (A3.2.7).

Using $W_2 = W_1 P_2$, (A3.2.8) and (A3.2.9) may be used to satisfy the Smoluchowski equation, (A3.2.2), another necessary property for a stationary process. Thus $u(t)$ is an example of a stationary Gaussian-Markov

process. In the form given by (A3.2.3), the process $u(t)$ is also called an Ornstein–Uhlenbeck process (‘OU process’).

Consider an ensemble of Brownian particles. The approach of P_2 to W_1 as $t \rightarrow \infty$ represents a kind of diffusion process in velocity space. The description of Brownian movement in these terms is known as the *Fokker–Planck* method [16]. For the present example, this equation can be shown to be

$$\frac{\partial}{\partial t} P_2(u, t) = \frac{\alpha}{M} \frac{\partial}{\partial u} (u P_2(u, t)) + \frac{kT\alpha}{M^2} \frac{\partial^2}{\partial u^2} P_2(u, t) \quad (\text{A3.2.11})$$

subject to the initial condition $P_2(u, 0) = \delta(u - u_0)$. The solution to (A3.2.11) is given by (A3.2.8). The Langevin equation and the Fokker–Planck equation provide equivalent complementary descriptions [17].

A3.2.3 ONSAGER’S THEORY OF NON-EQUILIBRIUM THERMODYNAMICS

A3.2.3.1 REGRESSION EQUATIONS AND FLUCTUATIONS

For a system which is close to equilibrium, it is assumed that its state is described by a set of extensive thermodynamic variables, $a_1(t), a_2(t), \dots, a_n(t)$ where n is very much less than the total number of degrees of freedom for all of the molecules in the system. The latter may be of order 10^{24} while the former may be fewer than 10. In equilibrium, the a_i are taken to have value zero so that the non-equilibrium entropy is given by

$$S = S_0 - \frac{1}{2} k a_i E_{ij} a_j \quad (\text{A3.2.12})$$

where S_0 is the maximum equilibrium value of the entropy, E_{ij} is a symmetric positive definite time independent entropy matrix and repeated indices are to be summed, a convention used throughout this presentation. *Thermodynamic forces* are defined by

$$X_i = \frac{\partial S}{\partial a_i} = -k E_{ij} a_j. \quad (\text{A3.2.13})$$

Onsager postulates [4, 5] the phenomenological equations for irreversible processes given by

$$R_{ij} \frac{d}{dt} a_j \equiv R_{ij} J_j = X_i = -k E_{ij} a_j \quad (\text{A3.2.14})$$

-7-

in which the J_j are called the *thermodynamic fluxes*, and which is a natural generalization of the linear phenomenological laws such as Fourier’s law of heat conduction, Newton’s law of internal friction etc. The matrix R_{ij} is real with eigenvalues having positive real parts and it is invertible. These equations are *regression* equations whose solutions approach equilibrium asymptotically in time.

Since the a_i are thermodynamic quantities, their values fluctuate with time. Thus, (A3.2.14) is properly interpreted as the averaged regression equation for a random process that is actually driven by random

thermodynamic forces, $\tilde{e}_i(t)$. The completed equations are coupled Langevin-like equations

$$R_{ij} \frac{d}{dt} a_j = R_{ij} J_j = X_i + \tilde{e}_i = -k E_{ij} a_j + \tilde{e}_i. \quad (\text{A3.2.15})$$

The mean values of the $\tilde{e}_i(t)$ are zero and each is assumed to be stationary Gaussian white noise. The linearity of these equations guarantees that the random process described by the a_i is also a stationary Gaussian–Markov process [12]. Denoting the inverse of R_{ij} by L_{ij} and using the definition

$$\tilde{F}_i = L_{ij} \tilde{e}_i$$

(A3.2.15) may be rewritten as

$$\frac{d}{dt} a_i = J_i = L_{ij} X_j + \tilde{F}_i = -k L_{ij} E_{jk} a_k + \tilde{F}_i. \quad (\text{A3.2.16})$$

Since the \tilde{F}_i are linearly related to the $\tilde{e}_i(t)$, they are also stationary Gaussian white noises. This property is explicitly expressed by

$$\langle \tilde{F}_i(t) \tilde{F}_j(t) \rangle = 2 Q_{ij} \delta(t - s) \quad (\text{A3.2.17})$$

in which Q_{ij} , the force–force correlation matrix, is necessarily symmetric and positive definite. While (A3.2.16) suggests that the fluxes may be coupled to any force, symmetry properties may be applied to show that this is not so. By establishing the tensor character of the different flux and force components, it can be shown that only fluxes and forces of like character can be coupled. This result is called Curie’s principle [9, 12].

Let $G_{ij} = k L_{ik} E_{kj}$. The solution to (A3.2.16) is

$$a_i(t) = (\exp[-\mathbf{G}t])_{ij} a_j(0) + \int_0^t ds (\exp[-\mathbf{G}(t - s)])_{ij} \tilde{F}_j(s). \quad (\text{A3.2.18})$$

-8-

The statistics for the initial conditions, $a_j(0)$, are determined by the equilibrium distribution obtained from the entropy in (A3.2.12) and in accordance with the Einstein–Boltzmann–Planck formula

$$W_1(a_1, a_2, \dots, a_n) = \left(\frac{\|\mathbf{E}\|}{(2\pi)^n} \right)^{1/2} \exp \left[-\frac{1}{2} a_i E_{ij} a_j \right] \quad (\text{A3.2.19})$$

where $\|\mathbf{E}\|$ is the determinant of E_{ij} . $\{ \dots \}$ will again denote averaging with respect to W_1 while $\langle \dots \rangle$ continues to denote averaging over the \tilde{F}_i . Notice in (A3.2.19) that W_1 now depends on n variables at a single time, a natural generalization of the situation reviewed above for the one-dimensional Langevin equation. This simply means the process is n dimensional.

A3.2.3.2 TWO TIME CORRELATIONS AND THE ONSAGER RECIPROCAL RELATIONS

From (A3.2.18) it follows that

$$\langle a_i(t) \rangle = (\exp[-\mathbf{G}t])_{ij} a_j(0)$$

and

$$\{\langle a(t) \rangle\} = \mathbf{0}.$$

All the other information needed for this process is contained in the two time correlation matrix because the process is Gaussian. A somewhat involved calculation [18] results (for $t_2 > t_1$) in

$$\begin{aligned} \chi_{ij}(t_2, t_1) \equiv \{\langle a_i(t_2) a_j(t_1) \rangle\} &= (\exp[-\mathbf{G}(t_2 - t_1)])_{ik} (\exp[-\mathbf{G}t_1] \mathbf{E}^{-1} \exp[-\mathbf{G}^\dagger t_1])_{kj} \\ &+ 2(\exp[-\mathbf{G}(t_2 - t_1)])_{ik} \int_0^{t_1} ds (\exp[-\mathbf{G}(t_1 - s)] \mathbf{Q} \exp[-\mathbf{G}^\dagger(t_1 - s)])_{kj}. \end{aligned} \quad (\text{A3.2.20})$$

If we set $t_2 = t_1 = t$, stationarity requires that $\chi_{ij}(t, t) = (\mathbf{E}^{-1})_{ij}$ because (A3.2.19) implies

$$\langle a_i, a_j \rangle = \int d^n a W_1(a_1, a_2, \dots, a_n) a_i a_j = (\mathbf{E}^{-1})_{ij}.$$

By looking at $\mathbf{G}\chi(t, t) + \chi(t, t)\mathbf{G}^\dagger$, the resulting integral implied by (A3.2.20) for $t_2 = t_1 = t$ contains an exact differential [18] and one obtains

$$\mathbf{G}\mathbf{E}^{-1} + \mathbf{E}^{-1}\mathbf{G}^\dagger = \mathbf{G}\chi(t, t) + \chi(t, t)\mathbf{G}^\dagger = 2\mathbf{Q} + \exp[-\mathbf{G}t](\mathbf{G}\mathbf{E}^{-1} + \mathbf{E}^{-1}\mathbf{G}^\dagger - 2\mathbf{Q})\exp[-\mathbf{G}^\dagger t].$$

This is compatible with stationarity if and only if

-9-

$$\mathbf{G}\mathbf{E}^{-1} + \mathbf{E}^{-1}\mathbf{G}^\dagger = 2\mathbf{Q} \quad (\text{A3.2.21})$$

which is the general case fluctuation–dissipation relation [12]. Inserting this identity into (A3.2.17) and using similar techniques reduces $\chi_{ij}(t_2, t_1)$ to the manifestly stationary form

$$\chi_{ij}(t_2 - t_1) = (\exp[-\mathbf{G}|t_2 - t_1])_{ik} (\mathbf{E}^{-1})_{kj}. \quad (\text{A3.2.22})$$

In Onsager's treatment, the additional restriction that the a_i are even functions of the particle velocities is made. As indicated above, this implies microscopic reversibility which in the present n -dimensional case means (for $t = t_2 - t_1 > 0$) $W_2(a_1, a_2, \dots, a_n; a_1', a_2', \dots, a_n', t) = W_2(a_1', a_2', \dots, a_n'; a_1, a_2, \dots, a_n, t)$. This implies

$$\begin{aligned}
\chi_{ij}(t_2 - t_1) &\equiv \int d^n a d^n a' W_2(a_1, a_2, \dots, a_n t_1; a'_1, a'_2, \dots, a'_n, t_2) a'_i a_j \\
&= \int d^n a d^n a' W_2(a'_1, a'_2, \dots, a'_n, t_1; a_1, a_2, \dots, \\
&\quad a_n, t_2) a_j a'_i \equiv \chi_{ji}(t_2 - t_1).
\end{aligned} \tag{A3.2.23}$$

Take the t_2 -derivative of this equation by using (A3.2.22) and set $t_2 = t_1$

$$\frac{d}{dt_2} \chi_{ij}(t_2 - t_1)_{t_2=t_1} = -G_{ik}(\mathbf{E}^{-1})_{kj} = \frac{d}{dt_2} \chi_{ji}(t_2 - t_1)_{t_2=t_1} = -G_{jk}(\mathbf{E}^{-1})_{ki}. \tag{A3.2.24}$$

Inserting the definition of \mathbf{G} gives the celebrated Onsager reciprocal relations [4, 5]

$$L_{ij} = L_{ji}. \tag{A3.2.25}$$

If odd variables, the b , are also included, then a generalization by Casimir [8] results in the Onsager–Casimir relations

$$\begin{aligned}
L_{ij}(\mathbf{B}, \omega) &= L_{ji}(-\mathbf{B}, -\omega) \\
L_{im}(\mathbf{B}, \omega) &= -L_{mi}(-\mathbf{B}, -\omega) \\
L_{nm}(\mathbf{B}, \omega) &= L_{mn}(-\mathbf{B}, -\omega)
\end{aligned} \tag{A3.2.26}$$

wherein the indices i and j are for variables a and the indices m and n are for variables b . These are proved in a similar fashion. For example, when there are mixtures of variables a and b , microscopic reversibility becomes $W_2(a_1, \dots, a_p, b_{p+1}, \dots, b_n; a'_1, \dots, a'_p, b'_{p+1}, \dots, b'_n, \mathbf{B}, \omega, t) = W_2(a'_1, \dots, a'_p, -b'_{p+1}, \dots, -b'_n; a_1, \dots, a_p, -b_{p+1}, \dots, -b_n, -\mathbf{B}, -\omega, t)$. A cross-correlation between an even variable and an odd variable is given by

-10-

$$\begin{aligned}
\chi_{im}(\mathbf{B}, \omega, t_2 - t_1) &\equiv \int d^p a d^{n-p} b d^p a' d^{n-p} b' W_2(a_1, \dots, a_p, b_{p+1}, \dots, b_n, t_1; a'_1, \dots, a'_p, \\
&\quad b'_{p+1}, \dots, b'_n, \mathbf{B}, \omega, t_2) a'_i b_m \\
&= \int d^p a d^{n-p} b d^p a' d^{n-p} b' W_2(a'_1, \dots, a'_p, -b'_{p+1}, \dots, \\
&\quad -b'_n, t_1; a_1, \dots, a_p, \\
&\quad -b_{p+1}, \dots, -b_n, -\mathbf{B}, -\omega, t_2) b_m a'_i \\
&\equiv -\chi_{mi}(-\mathbf{B}, -\omega, t_2 - t_1)
\end{aligned}$$

in which the last identity follows from replacing all b by their negatives. Differentiation of this expression with respect to t_2 followed by setting $t_2 = t_1$ results in the middle result of (A3.2.26).

In the general case, (A3.2.23) cannot hold because it leads to (A3.2.24) which requires $\mathbf{G}\mathbf{E}^{-1} = (\mathbf{G}\mathbf{E}^{-1})^\dagger$ which is in general not true. Indeed, the simple example of the Brownian motion of a harmonic oscillator suffices to make the point [7, 14, 18]. In this case the equations of motion are [3, 7]

$$M \frac{dx}{dt} = p \quad \text{and} \quad \frac{dp}{dt} + M\omega^2 x = -\frac{\alpha}{M} p + \tilde{F} \quad (\text{A3.2.27})$$

where M is the oscillator mass, ω is the oscillator frequency and α is the friction coefficient. The fluctuating force, \tilde{F} , is Gaussian white noise with zero mean and correlation formula

$$\delta a_i^{\text{ext}}(\vec{r}, t) = \delta a_i(\vec{r}) \exp[-(\eta - i\omega)t]$$

Define y by $y = M\omega x$. The identifications

$$a_i = \begin{pmatrix} y \\ p \end{pmatrix} \quad A_{ij} = \begin{pmatrix} 0 & -\omega \\ \omega & 0 \end{pmatrix} \quad S_{ij} = \begin{pmatrix} 0 & 0 \\ 0 & \alpha/M \end{pmatrix} \quad \tilde{F}_i = \begin{pmatrix} 0 \\ \tilde{F} \end{pmatrix}$$

permit writing (A3.2.27) as

$$\frac{d}{dt} a_i = -A_{ij} a_j - S_{ij} a_j + \tilde{F}_i.$$

Clearly, $G = A + S$ in this example. The entropy matrix can be obtained from the Maxwell–Boltzmann distribution

$$W_1(y, p) = W_0 \exp \left[-\frac{p^2 + y^2}{2MkT} \right]$$

-11-

which implies that

$$E_{ij} = \frac{1}{MkT} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The Q of (A3.2.17) is clearly given by

$$Q_{ij} = \begin{pmatrix} 0 & 0 \\ 0 & D \end{pmatrix}.$$

In this case the fluctuation–dissipation relation, (A3.2.21), reduces to $D = kT\alpha$. It is also clear that $\mathbf{GE}^{-1} = (\mathbf{A} + \mathbf{S})/MkT$ which is not self-adjoint.

A3.2.3.3 LEAST DISSIPATION VARIATIONAL PRINCIPLES

Onsager [4, 5] generalized Lord Rayleigh's 'principle of the least dissipation of energy' [19]. In homage to Lord Rayleigh, Onsager retained the name of the principle (i.e. the word *energy*) although he clearly stated [4] that the role of the potential in this principle is played by the *rate of increase of the entropy* [9]. This idea is an attempt to extend the highly fruitful concept of an underlying variational principle for dynamics, such as Hamilton's principle of least action for classical mechanics, to irreversible processes. Because the regression

equations are linear, the parallel with Lord Rayleigh's principle for linear friction in mechanics is easy to make. It has also been extended to velocity dependent forces in electrodynamics [20].

From (A3.2.12), it is seen that the rate of entropy production is given by

$$\frac{d}{dt}S = - \left(\frac{d}{dt} a_i \right) k E_{ij} a_j = J_i X_i.$$

Wherein the definition of the thermodynamic fluxes and forces of (A3.2.13) and (A3.2.14) have been used. Onsager defined [5] the analogue of the Rayleigh dissipation function by

$$\Phi = \frac{1}{2} R_{ij} J_i J_j.$$

When the reciprocal relations are valid in accord with (A3.2.25) then R is also symmetric. The variational principle in this case may be stated as

$$0 = \delta[J_i X_i - \Phi] = [X_i - R_{ij} J_j] \delta J_i$$

-12-

wherein the variation is with respect to the fluxes for fixed forces. The second variation is given by $-R$, which has negative eigenvalues (for symmetric R the eigenvalues are real and positive), which implies that the difference between the entropy production rate and the dissipation function is a maximum for the averaged irreversible process, hence *least dissipation* [5]. By multiplying this by the temperature, the free energy is generated from the entropy and, hence, Onsager's terminology of least dissipation of *energy*. Thus, the principle of least dissipation of entropy for near equilibrium dynamics is already found in the early work of Onsager [5, 9].

Related variational principles for non-equilibrium phenomena have been developed by Chandrasekhar [21]. How far these ideas can be taken remains an open question. Glansdorff and Prigogine [22] attempted to extend Onsager's principle of the least dissipation of entropy [4, 5, 9] to non-linear phenomena far away from equilibrium. Their proposal was ultimately shown to be overly ambitious [23]. A promising approach for the non-linear steady state regime has been proposed by Keizer [23, 24]. This approach focuses on the covariance of the fluctuations rather than on the entropy, although in the linear regime around full equilibrium, the two quantities yield identical principles. In the non-linear regime the distinction between them leads to a novel thermodynamics of steady states that parallels the near equilibrium theory. An experiment for non-equilibrium electromotive force [25] has confirmed this alternative approach and strongly suggests that a fruitful avenue of investigation for far from equilibrium thermodynamics has been opened.

A3.2.4 APPLICATIONS

A3.2.4.1 SORET EFFECT AND DUFOUR EFFECT [12]

Consider an isotropic fluid in which viscous phenomena are neglected. Concentrations and temperature are non-uniform in this system. The rate of entropy production may be written

$$\frac{dS}{dt} = \mathbf{J}_q \cdot \nabla \frac{1}{T} + \sum_{i=1}^n \mathbf{J}_i \cdot \nabla \left(-\frac{\mu_i}{T} \right) \quad (\text{A3.2.28})$$

in which \mathbf{J}_q is the heat flux vector and \mathbf{J}_i is the mass flux vector for species i which has chemical potential μ_i . This is over-simplified for the present discussion because the n mass fluxes, \mathbf{J}_i , are not linearly independent [12]. This fact may be readily accommodated by eliminating one of the fluxes and using the Gibbs–Duhem relation [12]. It is straightforward to identify the thermodynamic forces, X_p , using the generic form for the entropy production in (A3.2.27). The fluxes may be expressed in terms of the forces by

$$\begin{aligned} \mathbf{J}_q &= L_{qq} \nabla \frac{1}{T} - \sum_{i=1}^n L_{qi} \nabla \frac{\mu_i}{T} \\ \mathbf{J}_i &= L_{iq} \nabla \frac{1}{T} - \sum_{j=1}^n L_{ij} \nabla \frac{\mu_j}{T}. \end{aligned}$$

-13-

The Onsager relations in this case are

$$L_{qi} = L_{iq} \quad \text{and} \quad L_{ij} = L_{ji}.$$

The coefficients, L_{iq} , are characteristic of the phenomenon of *thermal diffusion*, i.e. the flow of matter caused by a temperature gradient. In liquids, this is called the Soret effect [12]. A reciprocal effect associated with the coefficient L_{qi} is called the *Dufour effect* [12] and describes heat flow caused by concentration gradients. The Onsager relation implies that measurement of one of these effects is sufficient to determine the coupling for both. The coefficient L_{qq} is proportional to the heat conductivity coefficient and is a single scalar quantity in an isotropic fluid even though its associated flux is a vector. This fact is closely related to the Curie principle. The remaining coefficients, L_{ij} , are proportional to the mutual diffusion coefficients (except for the diagonal ones which are proportional to self-diffusion coefficients).

Chemical reactions may be added to the situation giving an entropy production of

$$\frac{dS}{dt} = \mathbf{J}_q \cdot \nabla \frac{1}{T} + \sum_{i=1}^n \mathbf{J}_i \cdot \nabla \left(-\frac{\mu_i}{T} \right) - \frac{1}{T} \sum_{j=1}^r J_j A_j$$

in which there are r reactions with variable progress rates related to the J_j and with chemical affinities A_j . Once again, these fluxes are not all independent and some care must be taken to rewrite everything so that symmetry is preserved [12]. When this is done, the Curie principle decouples the vectorial forces from the scalar fluxes and *vice versa* [9]. Nevertheless, the reaction terms lead to additional reciprocal relations because

$$J_j = - \sum_{k=1}^r L_{jk} A_k$$

implies that $L_{jk} = L_{kj}$.

These are just a few of the standard examples of explicit applications of the Onsager theory to concrete cases.

There are many more involving acoustical, electrical, gravitational, magnetic, osmotic, thermal and other processes in various combinations. An excellent source for details is the book by DeGroot and Mazur [12], which was published in a Dover edition in 1984, making it readily accessible and inexpensive. There, one will find many specific accounts. For example, in the case of thermal and electric or thermal and electromagnetic couplings: (1) the *Hall effect* is encountered where the Hall coefficient is related to Onsager's relation through the resistivity tensor; (2) the *Peltier effect* is encountered where Onsager's relation implies the *Thompson relation* between the thermo-electric power and the Peltier heat and (3) *galvanomagnetic* and *thermomagnetic* effects are met along with the *Ettinghausen effect*, the *Nernst effect* and the *Bridgman relation*. In the case of so-called *discontinuous systems*, the *thermomolecular pressure effect*, *thermal effusion* and the *mechanocaloric effect* are encountered as well as *electro-osmosis*. Throughout, the entropy production equation plays a central role [12].

-14-

A3.2.4.2 THE FLUCTUATING DIFFUSION EQUATION

A byproduct of the preceding analysis is that the Onsager theory immediately determines the form of the fluctuations that should be added to the diffusion equation. Suppose that a solute is dissolved in a solvent with concentration c . The diffusion equation for this is

(A3.2.29)

in which D is the diffusion constant. This is called *Fick's law* of diffusion [12]. From (A3.2.28), the thermodynamic force is seen to be (at constant T)

$$\frac{dS}{dt} = J_q \cdot \nabla \frac{1}{T} + \sum_{i=1}^n J_i \cdot \nabla \left(-\frac{\mu_i}{T} \right) - \frac{1}{T} \sum_{j=1}^r J_j A_j$$

wherein it has been assumed that the solute is a nonelectrolyte exhibiting ideal behaviour with $\mu = kT \ln(c)$ and c_0 is the equilibrium concentration. Since this is a continuum system, the general results developed above need to be continuously extended as follows. The entropy production in a volume V may be written as

$$\begin{aligned} \frac{d}{dt} S &= \int d^3r (D \nabla c) \cdot \left(\frac{k}{c_0} \nabla c \right) = \frac{Dk}{c_0} \int d^3r (\nabla c) \cdot (\nabla c) \\ &= \frac{Dk}{c_0} \int d^3r \int d^3r' (\nabla c(\mathbf{r})) \cdot \delta(\mathbf{r} - \mathbf{r}') (\nabla c(\mathbf{r}')) \\ &= \frac{Dk}{c_0} \int d^3r \int d^3r' c(\mathbf{r}) (\nabla \cdot \nabla' \delta(\mathbf{r} - \mathbf{r}')) c(\mathbf{r}'). \end{aligned} \tag{A3.2.30}$$

The continuous extension of (A3.2.12) becomes

$$\mathbf{X} = -\nabla \frac{\mu}{T} = -\frac{1}{T} \frac{\partial \mu}{\partial c} \nabla c = -\frac{k}{c_0} \nabla c$$

The time derivative of this expression together with (A3.2.29) implies

$$\begin{aligned}
\frac{d}{dt}S &= -\frac{k}{2} \int d^3r \int d^3r' [(D\nabla^2 c(\mathbf{r}))E(\mathbf{r} - \mathbf{r}') + E(\mathbf{r} - \mathbf{r}')(D\nabla^2 c(\mathbf{r}'))] \\
&= -\frac{kD}{2} \int d^3r \int d^3r' [c(\mathbf{r})(\nabla^2 E(\mathbf{r} - \mathbf{r}') + \nabla'^2 E(\mathbf{r} - \mathbf{r}'))c(\mathbf{r}')].
\end{aligned}
\tag{A3.2.31}$$

-15-

Equations (A3.2.30) and (A3.2.31) imply the identity for the entropy matrix

$$E(\mathbf{r} - \mathbf{r}') = \frac{1}{c_0} \delta(\mathbf{r} - \mathbf{r}').$$

Equation (A3.2.29) also implies that the extension of G is now

$$G(\mathbf{r} - \mathbf{r}') = -D\nabla^2 \delta(\mathbf{r} - \mathbf{r}').$$

The extension of the fluctuation–dissipation relation of (A3.2.21) becomes

$$\begin{aligned}
2Q(\mathbf{r} - \mathbf{r}') &= \int d^3r'' [G(\mathbf{r} - \mathbf{r}'')E^{-1}(\mathbf{r}'' - \mathbf{r}') + E^{-1}(\mathbf{r} - \mathbf{r}'')G(\mathbf{r}'' - \mathbf{r}')] \\
&= -2Dc_0\nabla^2 \delta(\mathbf{r} - \mathbf{r}').
\end{aligned}$$

This means that the fluctuating force can be written as

$$\tilde{F}(\mathbf{r}, t) = \nabla \cdot \tilde{\mathbf{g}}(\mathbf{r}, t) \quad \text{where } \langle \tilde{g}_\alpha(\mathbf{r}, t) \tilde{g}_\beta(\mathbf{r}', s) \rangle = 2Dc_0 \delta_{\alpha\beta} \delta(\mathbf{r} - \mathbf{r}') \delta(t - s).$$

The resulting fluctuating diffusion equation is

$$\frac{\partial}{\partial t}c = D\nabla^2 c + \tilde{F}.$$

The quantity $\tilde{\mathbf{g}}$ can be thought of as a fluctuating mass flux.

Two applications of the fluctuating diffusion equation are made here to illustrate the additional information the fluctuations provide over and beyond the deterministic behaviour. Consider an infinite volume with an initial concentration, c , that is constant, c_0 , everywhere. The solution to the averaged diffusion equation is then simply $\langle c \rangle = c_0$ for all t . However, the two-time correlation function may be shown [26] to be

$$\begin{aligned}
\chi_{cc}(\mathbf{r}, t; \mathbf{r}', s) &= \langle (c(\mathbf{r}, t) - c_0)(c(\mathbf{r}', s) - c_0) \rangle \\
&= c_0 \left[\delta(\mathbf{r} - \mathbf{r}') - \frac{1}{(8\pi D|t - s|)^{3/2}} \exp\left[-\frac{|\mathbf{r} - \mathbf{r}'|^2}{8D|t - s|}\right] \right].
\end{aligned}$$

As the time separation $|t - s|$ approaches ∞ the second term in this correlation vanishes and the remaining term is the equilibrium density–density correlation formula for an ideal solution. The second possibility is to consider a non-equilibrium initial state, $c(\mathbf{r}, t) = c_0 \delta(\mathbf{r})$. The averaged solution is [26]

$$\langle c(\mathbf{r}, t) \rangle = c_0 \frac{1}{(4\pi Dt)^{3/2}} \exp\left[-\frac{r^2}{4Dt}\right]$$

whereas the two-time correlation function may be shown after extensive computation [26] to be

$$\begin{aligned} \chi_{cc}(\mathbf{r}, t; \mathbf{r}', s) &= \langle (c(\mathbf{r}, t) - c_0)(c(\mathbf{r}, s) - c_0) \rangle \\ &= c_0 \left[\delta(\mathbf{r} - \mathbf{r}') \frac{1}{(4\pi D|t-s|)^{3/2}} \exp\left[-\frac{|\mathbf{r} + \mathbf{r}'|^2}{4D|t-s|}\right] \right. \\ &\quad \left. - \frac{1}{(8\pi D|t-s|)^3} \exp\left[-\frac{|\mathbf{r} - \mathbf{r}'|^2}{8D|t-s|}\right] \exp\left[-\frac{|\mathbf{r} + \mathbf{r}'|^2}{8D|t-s|}\right] \right]. \end{aligned}$$

This covariance function vanishes as $|t-s|$ approaches ∞ because the initial density profile has a finite integral, that creates a vanishing density when it spreads out over the infinite volume.

This example illustrates how the Onsager theory may be applied at the macroscopic level in a self-consistent manner. The ingredients are the averaged regression equations and the entropy. Together, these quantities permit the calculation of the fluctuating force correlation matrix, \mathbf{Q} . Diffusion is used here to illustrate the procedure in detail because diffusion is the simplest known case exhibiting continuous variables.

A3.2.4.3 FLUCTUATING HYDRODYNAMICS

A proposal based on Onsager's theory was made by Landau and Lifshitz [27] for the fluctuations that should be added to the Navier–Stokes hydrodynamic equations. Fluctuating stress tensor and heat flux terms were postulated in analogy with the Onsager theory. However, since this is a case where the variables are of mixed time reversal character, the ‘derivation’ was not fully rigorous. This situation was remedied by the derivation by Fox and Uhlenbeck [13, 14, 18] based on general stationary Gaussian–Markov processes [12]. The precise form of the Landau proposal is confirmed by this approach [14].

Let $\Delta\rho$, Δu and ΔT denote the deviations of the mass density, ρ , the velocity field, u , and the temperature, T , from their full equilibrium values. The fluctuating, linearized Navier–Stokes equations are

$$\begin{aligned} \frac{\partial}{\partial t} \Delta\rho + \rho_{\text{eq}} \nabla \cdot \Delta\mathbf{u} &= 0 \\ \rho_{\text{eq}} \frac{\partial}{\partial t} \Delta u_\alpha + A_{\text{eq}} \frac{\partial}{\partial x_\alpha} \Delta\rho + B_{\text{eq}} \frac{\partial}{\partial x_\alpha} \Delta T &= \frac{\partial}{\partial x_\beta} \left[2\eta \Delta D_{\alpha\beta} + \left(\xi - \frac{2}{3}\eta\right) \Delta D_{\gamma\gamma} \delta_{\alpha\beta} \right] \\ &\quad + \frac{\partial}{\partial x_\beta} \tilde{S}_{\alpha\beta} \\ \rho_{\text{eq}} C_{\text{eq}} \frac{\partial}{\partial t} \Delta T &= K \nabla^2 \Delta T - T_{\text{eq}} B_{\text{eq}} \nabla \cdot \Delta\mathbf{u} + \nabla \cdot \tilde{\mathbf{g}} \end{aligned} \tag{A3.2.32}$$

in which η is the shear viscosity, ξ is the bulk viscosity, K is the heat conductivity, the subscript ‘eq’ denotes

equilibrium values and A_{eq} , B_{eq} and C_{eq} are defined [14] by

$$A_{\text{eq}} = \left(\frac{\partial p}{\partial \rho} \right)_{\text{eq}} \quad B_{\text{eq}} = \left(\frac{\partial p}{\partial T} \right)_{\text{eq}} \quad C_{\text{eq}} = \left(\frac{\partial \varepsilon}{\partial T} \right)_{\text{eq}}$$

in which p is the pressure and ε is the energy per unit mass. $D_{\alpha\beta}$ is the strain tensor and $\tilde{S}_{\alpha\beta}$ is the fluctuating stress tensor while \tilde{g}_{α} is the fluctuating heat flux vector. These fluctuating terms are Gaussian white noises with zero mean and correlations given by

$$\begin{aligned} \langle \tilde{S}_{\alpha\beta}(\mathbf{r}, t) \tilde{S}_{\mu\nu}(\mathbf{r}', t') \rangle &= 2kT_{\text{eq}} \delta(\mathbf{r} - \mathbf{r}') \delta(t - t') \\ &\quad \left[\eta(\delta_{\alpha\mu} \delta_{\beta\nu} + \delta_{\alpha\nu} \delta_{\beta\mu}) + \left(\xi - \frac{2}{3} \eta \right) \delta_{\alpha\beta} \delta_{\mu\nu} \right] \\ \langle \tilde{g}_{\alpha}(\mathbf{r}, t) \tilde{g}_{\beta}(\mathbf{r}', t') \rangle &= 2kT_{\text{eq}}^2 \delta(\mathbf{r} - \mathbf{r}') \delta(t - t') K \delta_{\alpha\beta} \\ \langle \tilde{S}_{\alpha\beta}(\mathbf{r}, t) \tilde{g}_{\mu}(\mathbf{r}', t') \rangle &= 0. \end{aligned} \tag{A3.2.33}$$

The lack of correlation between the fluctuating stress tensor and the fluctuating heat flux in the third expression is an example of the Curie principle for the fluctuations. These equations for fluctuating hydrodynamics are arrived at by a procedure very similar to that exhibited in the preceding section for diffusion. A crucial ingredient is the equation for entropy production in a fluid

$$\frac{d}{dt} S(t) = \int d^3r \left[\frac{K}{T^2} (\nabla T) \cdot (\nabla T) + \frac{1}{T} \left(2\eta D_{\alpha\beta} D_{\alpha\beta} + \left(\xi - \frac{2}{3} \eta \right) (D_{\alpha\alpha})^2 \right) \right].$$

This expression determines the entropy matrix needed for the fluctuation–dissipation relation [14] used to obtain (A3.2.33).

Three interesting applications of these equations are made here. The first is one of perspective. A fluid in full equilibrium will exhibit fluctuations. In fact, these fluctuations are responsible for Rayleigh–Brillouin light scattering in fluids [28]. From the light scattering profile of an equilibrium fluid, the viscosities, heat conductivity, speed of sound and sound attenuation coefficient can be determined. This is a remarkable exhibition of how non-equilibrium properties of the fluid reside in the equilibrium fluctuations. Jerry Gollub once posed to the author the question: ‘how does a fluid know to make the transition from steady state conduction to steady state convection at the threshold of instability in the Rayleigh–Benard system [21]?’ The answer is that the fluid fluctuations are incessantly testing the stability and nucleate the transition when threshold conditions exist. Critical opalescence [28] is a manifestation of this macroscopic influence of the fluctuations.

-18-

The second application is to temperature fluctuations in an equilibrium fluid [18]. Using (A3.2.32) and (A3.2.33) the correlation function for temperature deviations is found to be

$$\langle \Delta T(\mathbf{r}, t) \Delta T(\mathbf{r}', t') \rangle = \frac{kT_{\text{eq}}^2}{\rho_{\text{eq}} C_{\text{eq}}} \left(\frac{\rho_{\text{eq}} C_{\text{eq}}}{4\pi K |t - t'|} \right)^{3/2} \exp \left[-\frac{\rho_{\text{eq}} C_{\text{eq}} |\mathbf{r} - \mathbf{r}'|^2}{4K |t - t'|} \right]. \tag{A3.2.34}$$

When the two times are identical, the formula simplifies to

$$\langle \Delta T(\mathbf{r}) \Delta T(\mathbf{r}') \rangle = \frac{kT_{\text{eq}}^2}{\rho_{\text{eq}} C_{\text{eq}}} \delta(\mathbf{r} - \mathbf{r}').$$

Define the temperature fluctuations in a volume V by

$$\Delta T_V = \frac{1}{V} \int d^3r \Delta T(\mathbf{r}).$$

This leads to the well known formula

$$\langle \Delta T_V \Delta T_V \rangle = \frac{kT_{\text{eq}}^2}{\rho_{\text{eq}} V C_{\text{eq}}} = \frac{kT_{\text{eq}}^2}{C_V}$$

in which C_V is the ordinary heat capacity since C_{eq} is the heat capacity per unit mass. This formula can be obtained by purely macroscopic thermodynamic arguments [29]. However, the dynamical information in (A3.2.34) cannot be obtained from equilibrium thermodynamics alone.

The third application is to velocity field fluctuations. For an equilibrium fluid the velocity field is, on average, zero everywhere but it does fluctuate. The correlations turn out to be

$$\begin{aligned} \langle u_\alpha(\mathbf{r}, t) u_\beta(\mathbf{r}', t') \rangle &= \frac{kT_{\text{eq}}}{\rho_{\text{eq}}} \left[(4\pi\nu|t-t'|)^{-3/2} \exp\left[-\frac{|\mathbf{r}-\mathbf{r}'|^2}{4\nu|t-t'|}\right] \right. \\ &\quad \left. + \frac{\partial^2}{\partial x_\alpha \partial x_\beta} \left[(4\pi|\mathbf{r}-\mathbf{r}'|)^{-1} \Phi\left(\frac{|\mathbf{r}-\mathbf{r}'|}{2\nu^{1/2}|t-t'|^{1/2}}\right) \right] \right] \end{aligned} \quad (\text{A3.2.35})$$

in which $\nu = \eta / \rho_{\text{eq}}$, the kinematic viscosity and $\Phi(x)$ is defined by

$$\Phi(x) = \frac{2}{\sqrt{\pi}} \int_0^x dy \exp(-y^2).$$

-19-

When $\mathbf{r} = \mathbf{r}'$ or for $\nu^{1/2}|t-t'|^{1/2} \gg |\mathbf{r}-\mathbf{r}'|$, (A3.2.35) simplifies greatly, yielding

$$\langle u_\alpha(\mathbf{r}, t) u_\beta(\mathbf{r}, t') \rangle = \frac{3}{2} \frac{kT_{\text{eq}}}{\rho_{\text{eq}}} (4\pi\nu|t-t'|)^{-3/2} \delta_{\alpha\beta}$$

which is indistinguishable from the famous long-time-tail result [30].

A3.2.4.4 FLUCTUATING BOLTZMANN EQUATION

Onsager's theory can also be used to determine the form of the fluctuations for the Boltzmann equation [15]. Since hydrodynamics can be derived from the Boltzmann equation as a *contracted description*, a contraction of the fluctuating Boltzmann equation determines fluctuations for hydrodynamics. In general, a contraction of the description creates a new description which is non-Markovian, i.e. has memory. The Markov

approximation to the contraction of the fluctuating Boltzmann equation is identical with fluctuating hydrodynamics [15]. This is an example of the internal consistency of the Onsager approach. Similarly, it is possible to consider the hydrodynamic problem of the motion of a sphere in a fluctuating fluid described by fluctuating hydrodynamics (with appropriate boundary conditions). A contraction of this description [14] produces Langevin's equation for Brownian movement. Thus, three levels of description exist in this hierarchy: fluctuating Boltzmann equation, fluctuating hydrodynamic equations and Langevin's equation. The general theory for such hierarchies of description and their contractions can be found in the book by Keizer [31].

A3.2.5 LINEAR RESPONSE THEORY

Linear response theory is an example of a microscopic approach to the foundations of non-equilibrium thermodynamics. It requires knowledge of the Hamiltonian for the underlying microscopic description. In principle, it produces explicit formulae for the relaxation parameters that make up the Onsager coefficients. In reality, these expressions are extremely difficult to evaluate and approximation methods are necessary. Nevertheless, they provide a deeper insight into the physics.

The linear response of a system is determined by the lowest order effect of a perturbation on a dynamical system. Formally, this effect can be computed either classically or quantum mechanically in essentially the same way. The connection is made by converting quantum mechanical commutators into classical Poisson brackets, or *vice versa*. Suppose that the system is described by Hamiltonian $H + H_{\text{ex}}$ where H_{ex} denotes an external perturbation that may depend on time and generally does not commute with H . The density matrix equation for this situation is given by the Bloch equation [32]

$$\frac{\partial}{\partial t} \rho = -\frac{i}{\hbar} [H + H_{\text{ex}}, \rho] \quad (\text{A3.2.36})$$

-20-

where ρ denotes the density matrix and the square brackets containing two quantities separated by a comma denotes a commutator. In the classical limit, the density matrix becomes a phase space distribution, f , of the coordinates and conjugate momenta and the Bloch equation becomes Liouville's equation [32]

$$\frac{\partial}{\partial t} f = \sum_i \left(\frac{\partial(H + H_{\text{ex}})}{\partial q_i} \frac{\partial f}{\partial p_i} - \frac{\partial(H + H_{\text{ex}})}{\partial p_i} \frac{\partial f}{\partial q_i} \right) \equiv \{H + H_{\text{ex}}, f\} \quad (\text{A3.2.37})$$

in which the index i labels the different degrees of freedom and the second equality defines the Poisson bracket. Both of these equations may be expressed in terms of Liouville operators in the form

$$\frac{\partial}{\partial t} \rho = i(L + L_{\text{ex}})\rho \quad (\text{A3.2.38})$$

where quantum mechanically these operators are defined by (A3.2.36), and classically ρ means f and the operators are defined by (A3.2.37) [32].

Assuming explicit time dependence in L_{ex} , (A3.2.38) is equivalent to the integral equation

$$\rho(t) = \exp[i(t - t_0)L]\rho(t_0) + \int_{t_0}^t ds \exp[i(t - t_0)L]iL_{\text{ex}}(s)\rho(s)$$

as is easily proved by t -differentiation. Note that the exponential of the quantum mechanical Liouville operator may be shown to have the action

$$\exp[itL]A = \exp\left[-\frac{i}{\hbar}H\right]A\exp\left[\frac{i}{\hbar}H\right]$$

in which A denotes an arbitrary operator. This identity is also easily proved by t -differentiation.

The usual context for linear response theory is that the system is prepared in the infinite past, $t_0 \rightarrow -\infty$, to be in equilibrium with Hamiltonian H and then H_{ex} is turned on. This means that $\rho(t_0)$ is given by the canonical density matrix

$$\rho(t_0) = \rho_{\text{eq}} = \frac{1}{Z} \exp[-\beta H]$$

where $\beta = 1/kT$ and $Z = \text{Trace} \exp [-\beta H]$. Clearly

$$\exp[itL]\rho_{\text{eq}} = \rho_{\text{eq}}.$$

-21-

Thus, to first order in L_{ex} , $\rho(t)$ is given by

$$\rho(t) = \rho_{\text{eq}} - \frac{i}{\hbar} \int_{-\infty}^t ds \exp\left[-\frac{i}{\hbar}(t-s)H\right] [H_{\text{ex}}(s), \rho_{\text{eq}}] \exp\left[\frac{i}{\hbar}(t-s)H\right] + \dots \quad (\text{A3.2.39})$$

The classical analogue is

$$\exp[itL]\rho_{\text{eq}} = \rho_{\text{eq}}.$$

Let B denote an observable value. Its expectation value at time t is given by

$$f(t) = f_{\text{eq}} + \int_{-\infty}^t ds \exp[i(t-s)L]\{H_{\text{ex}}(s), f_{\text{eq}}\} + \dots$$

Denote the deviation of B from its equilibrium expectation value by $\Delta B = B - \text{Trace}(B\rho_{\text{eq}})$. From (A3.2.39), the deviation of the expectation value of B from its equilibrium expectation value, $\delta \langle B \rangle$, is

$$\langle B \rangle \equiv \text{Ex}(B) \equiv \text{Trace}(B\rho(t)).$$

where $B(t-s)$ is the Heisenberg operator solution to the Heisenberg equation of motion

$$\frac{d}{dt} \Delta B = \frac{i}{\hbar} [H, \Delta B]. \quad (\text{A3.2.40})$$

The second equality follows from the fact that going from B to ΔB involves subtracting a c -number from B and that c -number can be taken outside the trace. The resulting trace is the trace of a commutator, which vanishes. The invariance of the trace to cyclic permutations of the order of the operators is used in the third equality. It is straightforward to write

$$\begin{aligned} \delta\langle B \rangle &= -\frac{i}{\hbar} \int_{-\infty}^t ds \text{Trace} \left(B \exp \left[-\frac{i}{\hbar} (t-s) H \right] [H_{\text{ex}}(s), \rho_{\text{eq}}] \exp \left[\frac{i}{\hbar} (t-s) H \right] \right) \\ &= -\frac{i}{\hbar} \int_{-\infty}^t ds \text{Trace} \left(\Delta B \exp \left[-\frac{i}{\hbar} (t-s) H \right] [H_{\text{ex}}(s), \rho_{\text{eq}}] \exp \left[\frac{i}{\hbar} (t-s) H \right] \right) \\ &= -\frac{i}{\hbar} \int_{-\infty}^t ds \text{Trace} \left(\exp \left[\frac{i}{\hbar} (t-s) H \right] \Delta B \exp \left[-\frac{i}{\hbar} (t-s) H \right] [H_{\text{ex}}(s), \rho_{\text{eq}}] \right) \\ &= -\frac{i}{\hbar} \int_{-\infty}^t ds \text{Trace}(\Delta B(t-s)[H_{\text{ex}}(s), \rho_{\text{eq}}]) \end{aligned}$$

-22-

The transition from H_{ex} to ΔH_{ex} inside the commutator is allowed since $\text{Trace}(H_{\text{ex}} \rho_{\text{eq}})$ is a c -number and commutes with any operator. Thus, the final expression is [32]

$$\text{Trace}(\Delta B(t-s)[H_{\text{ex}}(s), \rho_{\text{eq}}]) = \text{Trace}([\Delta B(t-s), H_{\text{ex}}(s)] \rho_{\text{eq}}).$$

If the external perturbation is turned on with a time dependent function $F(t)$ and H_{ex} takes the form $AF(t)$ where A is a time independent operator (or H_{ex} is the sum of such terms), then

$$\delta\langle B(t) \rangle = -\frac{i}{\hbar} \int_{-\infty}^t ds \text{Trace}([\Delta B(t-s), \Delta H_{\text{ex}}(s)] \rho_{\text{eq}}).$$

which defines the linear response function $\Phi_{BA}(t-s)$. This quantity may be written compactly as

$$\Phi_{BA}(t) = -\frac{i}{\hbar} \langle [\Delta B(t), \Delta A] \rangle_{\text{eq}}.$$

An identical expression holds classically [32] if $-i/\hbar$ times the commutator is replaced by the classical Poisson bracket.

The Heisenberg equation of motion, (A3.2.40), may be recast for imaginary times $t = -i\hbar\lambda$ as

$$\frac{d}{d\lambda} A = [H, A]$$

with the solution

$$A(\lambda) = \exp[\lambda H] A \exp[-\lambda H].$$

Therefore, the Kubo identity [32] follows

-23-

$$\begin{aligned}
& \exp[-\beta H] \int_0^\beta d\lambda \exp[\lambda H] [A, H] \exp[-\lambda H] \\
&= \exp[-\beta H] \int_0^\beta d\lambda \exp[\lambda H] \left(-\frac{d}{d\lambda} A \right) \exp[-\lambda H] \\
&= \exp[-\beta H] \int_0^\beta d\lambda (H \exp[\lambda H] A \exp[-\lambda H] - \exp[\lambda H] A \exp[-\lambda H] H) \\
&= \exp[-\beta H] \int_0^\beta d\lambda [H, \exp[\lambda H] A \exp[-\lambda H]] \\
&= \exp[-\beta H] \int_0^\beta d\lambda \frac{d}{d\lambda} (\exp[\lambda H] A \exp[-\lambda H]) \\
&= \exp[-\beta H] (\exp[\beta H] A \exp[-\beta H] - A) \\
&= A \exp[-\beta H] - \exp[-\beta H] A = [A, \exp[-\beta H]].
\end{aligned}$$

Therefore,

(A3.2.41)

Using the Heisenberg equation of motion, (A3.2.40), the commutator in the last expression may be replaced by the time-derivative operator

$$i\hbar \frac{d}{dt} \Delta A = [\Delta A, H].$$

This converts (A3.2.41) into

$$\Phi_{BA}(t) = \left\langle \Delta B(t) \int_0^\beta d\lambda \exp[\lambda H] \frac{d}{dt} \Delta A \exp[-\lambda H] \right\rangle_{\text{eq}}$$

-24-

where the time derivative of ΔA is evaluated at $t = 0$. The quantity $kT\Phi_{BA}(t)$ is called the *canonical correlation* of $d/dt \Delta A$ and ΔB [32]. It is invariant under time translation by $\exp[-iH\tau/\hbar]$ because both ρ_{eq} and $\exp[\pm\lambda H]$ commute with this time evolution operator and the trace operation is invariant to cyclic permutations of the product of operators upon which it acts. Thus

(A3.2.42)

$$\begin{aligned}
\Phi_{BA}(t+\tau) &= \text{Trace} \left(\rho_{\text{eq}} \exp \left[-\frac{i}{\hbar} H \tau \right] \Delta B(t) \exp \left[\frac{i}{\hbar} H \tau \right] \int_0^\beta d\lambda \exp \left[H \left(\lambda - \frac{i}{\hbar} t \right) \right] \left(\frac{d}{dt} \Delta A \right) \right. \\
&\quad \left. \times \exp \left[H \left(-\lambda + \frac{i}{\hbar} t \right) \right] \right) \\
&= \text{Trace} \left(\exp \left[\frac{i}{\hbar} H \tau \right] \rho_{\text{eq}} \exp \left[-\frac{i}{\hbar} H \tau \right] \Delta B(t) \int_0^\beta d\lambda \exp \left[H \frac{i}{\hbar} t \right] \exp \left[H \left(\lambda - \frac{i}{\hbar} t \right) \right] \right. \\
&\quad \left. \times \left(\frac{d}{dt} \Delta A \right) \exp[-H\lambda] \right) = \Phi_{BA}(t).
\end{aligned}$$

Consider the canonical correlation of ΔA and ΔB , $C(\Delta A, \Delta B)$, defined by

$$C(\Delta A(0), \Delta B(t)) = kT \langle \Delta B(t) \int_0^\beta d\lambda \exp[\lambda H] \Delta A(0) \exp[-\lambda H] \rangle_{\text{eq}}.$$

The analysis used for (A3.2.42) implies

$$C(\Delta A(\tau), \Delta B(t+\tau)) = C(\Delta A(0), \Delta B(t))$$

which means that this correlation is independent of τ , i.e. *stationary*. Taking the τ -derivative implies

$$C \left(\frac{d}{dt} \Delta A(0), \Delta B(t) \right) + C \left(\Delta A(0), \frac{d}{dt} \Delta B(t) \right) = 0.$$

This is equivalent to

$$\begin{aligned}
\Phi_{BA}(t) &= \left\langle \Delta B(t) \int_0^\beta d\lambda \exp[\lambda H] \left(\frac{d}{dt} \Delta A(0) \right) \exp[-\lambda H] \right\rangle_{\text{eq}} \\
&= - \left\langle \left(\frac{d}{dt} \Delta B(t) \right) \int_0^\beta d\lambda \exp[\lambda H] \Delta A(0) \exp[-\lambda H] \right\rangle_{\text{eq}}.
\end{aligned} \tag{A3.2.43}$$

In different applications, one or the other of these two equivalent expressions may prove useful.

-25-

As an example, let B be the current J_i corresponding to the displacement A_i appearing in H_{ex} . Clearly

$$J_i(t) = \frac{d}{dt} A_i(t) = \frac{i}{\hbar} [H, A_i].$$

Because this current is given by a commutator, its equilibrium expectation value is zero. Using the first expression in (A3.2.43), the response function is given by

$$\Phi_{ij}(t) = \left\langle J_i(t) \int_0^\beta d\lambda \exp[\lambda H] J_j(0) \exp[-\lambda H] \right\rangle_{\text{eq}}. \tag{A3.2.44}$$

For a periodic perturbation, $\delta \langle \Delta B(t) \rangle$ is also periodic. The *complex admittance* [30] is given by

$$\chi_{BA}(t) = \int_0^\infty dt \Phi_{BA}(t) e^{i\omega t}.$$

For the case of a current as in (A3.2.44) the result is the Kubo formula [32] for the complex conductivity

$$\sigma_{ij}(\omega) = \int_0^\infty dt e^{i\omega t} \left\langle J_i(t) \int_0^\beta d\lambda \exp[\lambda H] J_j(0) \exp[-\lambda H] \right\rangle_{\text{eq}}.$$

Several explicit applications of these relations may be found in the books by Kubo *et al* [32] and by McLennan [33].

There are other techniques leading to results closely related to Kubo's formula for the conductivity coefficient. Notable among them is the Mori–Zwanzig theory [34, 35] based on projection operator techniques and yielding the generalized Langevin equation [18]. The formula for the conductivity coefficient is an example of the general formula for relaxation parameters, the Green–Kubo formula [36, 37]. The examples of Green–Kubo formulae for viscosity, thermal conduction and diffusion are in the book by McLennan [33].

A3.2.6 PROSPECTS

The current frontiers for the subject of non-equilibrium thermodynamics are rich and active. Two areas dominate interest: non-linear effects and molecular bioenergetics. The linearization step used in the near equilibrium regime is inappropriate far from equilibrium. Progress with a microscopic kinetic theory [38] for non-linear fluctuation phenomena has been made. Careful experiments [39] confirm this theory. Non-equilibrium long range correlations play an important role in some of the light scattering effects in fluids in far from equilibrium states [38, 39].

The role of non-equilibrium thermodynamics in molecular bioenergetics has experienced an experimental revolution during the last 35 years. Membrane energetics is now understood in terms of chemiosmosis [40]. In chemiosmosis, a trans-membrane electrochemical potential energetically couples the oxidation–reduction energy generated during catabolism to the adenosine triphosphate (ATP) energy needed for chemosynthesis during anabolism. Numerous advances in experimental technology have opened up whole new areas of exploration [41]. Quantitative analysis using non-equilibrium thermodynamics to account for the free energy and entropy changes works accurately in a variety of settings. There is a rich diversity of problems to be worked on in this area. Another biological application brings the subject back to its foundations. Rectified Brownian movement (involving a Brownian ratchet) is being invoked as the mechanism behind many macromolecular processes [42]. It may even explain the dynamics of actin and myosin interactions in muscle fibres [43]. In rectified Brownian movement, metabolic free energy generated during catabolism is used to bias boundary conditions for ordinary diffusion, thereby producing a non-zero flux. In this way, thermal fluctuations give the molecular mechanisms of cellular processes their vitality [44].

REFERENCES

- [1] Einstein A 1956 *Investigations on the Theory of Brownian Movement* (New York: Dover). This book is based on a series of papers Einstein published from 1905 until 1908
 - [2] Langevin P 1908 Sur la theorie du mouvement brownien *C. R. Acad. Sci. Paris* **146** 530
 - [3] Uhlenbeck G E and Ornstein L S 1930 On the theory of the Brownian motion *Phys. Rev.* **36** 823
 - [4] Onsager L 1931 Reciprocal relations in irreversible processes. I *Phys. Rev.* **37** 405
 - [5] Onsager L 1931 Reciprocal relations in irreversible processes. II *Phys. Rev.* **38** 2265
 - [6] Chandrasekhar S 1943 Stochastic problems in physics and astronomy *Rev. Mod. Phys.* **15** 1
 - [7] Wang M C and Uhlenbeck G E 1945 On the theory of Brownian motion II *Rev. Mod. Phys.* **17** 323
 - [8] Casimir H B G On Onsager's principle of microscopic reversibility *Rev. Mod. Phys.* **17** 343
 - [9] Prigogine I 1947 *Etude Thermodynamique des Phenomenes Irreversibles* (Liege: Desoer)
 - [10] Onsager L and Machlup S 1953 Fluctuations and irreversible processes *Phys. Rev.* **91** 1505
 - [11] Machlup S and Onsager L 1953 Fluctuations and irreversible processes. II. Systems with kinetic energy *Phys. Rev.* **91** 1512
 - [12] de Groot S R and Mazur P 1962 *Non-Equilibrium Thermodynamics* (Amsterdam: North-Holland)
 - [13] Fox R F 1969 Contributions to the theory of non-equilibrium thermodynamics *PhD Thesis* Rockefeller University, New York
 - [14] Fox R F and Uhlenbeck G E 1970 Contributions to non-equilibrium thermodynamics. I. Theory of hydrodynamical fluctuations *Phys. Fluids* **13** 1893
-

- [15] Fox R F and Uhlenbeck G E 1970 Contributions to non-equilibrium thermodynamics. II. Fluctuation theory for the Boltzmann equation *Phys. Fluids* **13** 2881
- [16] Risken H 1984 *The Fokker–Planck Equation, Methods of Solution and Application* (Berlin: Springer)
- [17] Arnold L 1974 *Stochastic Differential Equations* (New York: Wiley–Interscience)
- [18] Fox R F 1978 Gaussian stochastic processes in physics *Phys. Rev.* **48** 179
- [19] Rayleigh J W S 1945 *The Theory of Sound* vol 1 (New York: Dover) ch 4
- [20] Goldstein H 1980 *Classical Mechanics* 2nd edn (Reading, MA: Addison-Wesley) ch 1
- [21] Chandrasekhar S 1961 *Hydrodynamic and Hydromagnetic Stability* (London: Oxford University Press)
- [22] Glansdorff P and Prigogine I 1971 *Thermodynamic Theory of Structure, Stability and Fluctuations* (London: Wiley–Interscience)
- [23] Lavenda B H 1985 *Nonequilibrium Statistical Thermodynamics* (New York: Wiley) ch 3
- [24] Keizer J E 1987 *Statistical Thermodynamics of Nonequilibrium Processes* (New York: Springer) ch 8
- [25] Keizer J and Chang O K 1987 *J. Chem. Phys.* **87** 4064
- [26] Keizer J E 1987 *Statistical Thermodynamics of Nonequilibrium Processes* (New York: Springer) ch 6

- [27] Landau L D and Lifshitz E M 1959 *Fluid Mechanics* (London: Pergamon) ch 17
- [28] Berne B J and Pecora R 1976 *Dynamic Light Scattering* (New York: Wiley) ch 10
- [29] Landau L D and Lifshitz E M 1958 *Statistical Physics* (London: Pergamon) ch 12, equation (111.6)
- [30] Fox R F 1983 Long-time tails and diffusion *Phys. Rev. A* **27** 3216
- [31] Keizer J E 1987 *Statistical Thermodynamics of Nonequilibrium Processes* (New York: Springer) ch 9
- [32] Kubo R, Toda M and Hashitsume N 1985 *Statistical Physics II* (Berlin: Springer) ch 4
- [33] McLennan J A 1989 *Introduction to Non-Equilibrium Statistical Mechanics* (Englewood Cliffs, NJ: Prentice-Hall) ch 9
- [34] Zwanzig R 1961 Memory effects in irreversible thermodynamics *Phys. Rev.* **124** 983
- [35] Mori H 1965 Transport, collective motion and Brownian motion *Prog. Theor. Phys.* **33** 423
- [36] Green M S 1954 Markov random processes and the statistical mechanics of time-dependent phenomena. II. Irreversible processes in fluids *J. Chem. Phys.* **22** 398
- [37] Kubo R, Yokota M and Nakajima S 1957 Statistical-mechanical theory of irreversible processes. II. Response to thermal disturbance *J. Phys. Soc. Japan* **12** 1203

-28-

- [38] Kirkpatrick T R, Cohen E G D and Dorfman J R 1982 Light scattering by a fluid in a nonequilibrium steady state. II. Large gradients *Phys. Rev. A* **26** 995
- [39] Segre P N, Gammon R W, Sengers J V and Law B M 1992 Rayleigh scattering in a liquid far from thermal equilibrium *Phys. Rev. A* **45** 714
- [40] Harold F M 1986 *The Vital Force: A Study of Bioenergetics* (New York: Freeman)
- [41] de Duve C 1984 *A Guided Tour of the Living Cell* vols 1 and 2 (New York: Scientific American)
- [42] Peskin C S, Odell G M and Oster G F 1993 Cellular motions and thermal fluctuations: the Brownian ratchet *Biophys. J.* **65** 316
- [43] Huxley A F 1957 Muscle structure and theories of contraction *Prog. Biophys. Biophys. Chem.* **7** 255
- [44] Fox R F 1998 Rectified Brownian movement in molecular and cell biology *Phys. Rev. E* **57** 2177

FURTHER READING

Wax N (ed) 1954 *Selected Papers on Noise and Stochastic Processes* (New York: Dover)

van Kampen N G 1981 *Stochastic Processes in Physics and Chemistry* (Amsterdam: North-Holland)

Katchalsky A and Curran P F 1965 *Nonequilibrium Thermodynamics in Biophysics* (Cambridge, MA: Harvard University Press)

Lavenda B H 1985 *Nonequilibrium Statistical Thermodynamics* (New York: Wiley)

Keizer J E 1987 *Statistical Thermodynamics of Nonequilibrium Processes* (New York: Springer)

McLennan J A 1989 *Introduction to Non-equilibrium Statistical Mechanics* (Englewood Cliffs, NJ: Prentice-Hall)

-1-

A 3.3 Dynamics in condensed phase (including nucleation)

Rashmi C Desai

A3.3.1 INTRODUCTION

Radiation probes such as neutrons, x-rays and visible light are used to ‘see’ the structure of physical systems through elastic scattering experiments. Inelastic scattering experiments measure both the structural and dynamical correlations that exist in a physical system. For a system which is in thermodynamic equilibrium, the molecular dynamics create spatio-temporal correlations which are the manifestation of thermal fluctuations around the equilibrium state. For a condensed phase system, dynamical correlations are intimately linked to its structure. For systems in equilibrium, linear response theory is an appropriate framework to use to inquire on the spatio-temporal correlations resulting from thermodynamic fluctuations. Appropriate response and correlation functions emerge naturally in this framework, and the role of theory is to understand these correlation functions from first principles. This is the subject of [section A3.3.2](#).

A system of interest may be macroscopically homogeneous or inhomogeneous. The inhomogeneity may arise on account of interfaces between coexisting phases in a system or due to the system’s finite size and proximity to its external surface. Near the surfaces and interfaces, the system’s translational symmetry is broken; this has important consequences. The spatial structure of an inhomogeneous system is its average equilibrium property and has to be incorporated in the overall theoretical structure, in order to study spatio-temporal correlations due to thermal fluctuations around an inhomogeneous spatial profile. This is also illustrated in [section A3.3.2](#).

Another possibility is that a system may be held in a constrained equilibrium by external forces and thus be in a non-equilibrium steady state (NESS). In this case, the spatio-temporal correlations contain new ingredients, which are also exemplified in [section A3.3.2](#).

There are also important instances when the system is neither in equilibrium nor in a steady state, but is actually evolving in time. This happens, for example, when a binary homogeneous mixture at high temperature is suddenly quenched to a low-temperature non-equilibrium state in the middle of the coexistence region of the mixture. Following the quench, the mixture may be in a metastable state or in an unstable state, as defined within a mean field description. The subsequent dynamical evolution of the system, which follows an initial thermodynamic metastability or instability, and the associated kinetics, is a rich subject involving many fundamental questions, some of which are yet to be fully answered. The kinetics of thermodynamically unstable systems and phenomena like spinodal decomposition are treated in [section A3.3.3](#), after some introductory remarks. The late-stage kinetics of domain growth is discussed in [section A3.3.4](#). The discussion in this section is applicable to late-stage growth regardless of whether the initial post-quench state was thermodynamically unstable or metastable. The study of metastable states is connected with the subject of

nucleation and subsequent growth kinetics (treated in [section A3.3.4](#)). Homogeneous nucleation is the subject of [section A3.3.5](#). As will be clear from [section A3.3.1](#), the distinction between the spinodal decomposition and nucleation is not sharp. Growth morphology with apparent nucleation characteristics can occur when post-quench states are within the classical spinodal, except when the binary mixture is symmetric.

-2-

The specific examples chosen in this section, to illustrate the dynamics in condensed phases for the variety of system-specific situations outlined above, correspond to long-wavelength and low-frequency phenomena. In such cases, conservation laws and broken symmetry play important roles in the dynamics, and a macroscopic hydrodynamic description is either adequate or is amenable to an appropriate generalization. There are other examples where short-wavelength and/or high-frequency behaviour is evident. If this is the case, one would require a more microscopic description. For fluid systems which are the focus of this section, such descriptions may involve a kinetic theory of dense fluids or generalized hydrodynamics which may be linear or may involve nonlinear mode coupling. Such microscopic descriptions are not considered in this section.

A3.3.2 EQUILIBRIUM SYSTEMS: THERMAL FLUCTUATIONS AND SPATIO-TEMPORAL CORRELATIONS

In this section, we consider systems in thermodynamic equilibrium. Even though the system is in equilibrium, molecular constituents are in constant motion. We inquire into the nature of the thermodynamic fluctuations which have at their root the molecular dynamics. The space-time correlations that occur on account of thermodynamic fluctuations can be probed through inelastic scattering experiments, and the range of space and time scales explored depends on the wavenumber and frequency of the probe radiation. We illustrate this by using inelastic light scattering from dense fluids. Electromagnetic radiation couples to matter through its dielectric fluctuations. Consider a non-magnetic, non-conducting and non-absorbing medium with the average dielectric constant ϵ_0 . Let the incident electric field be a plane wave of the form

$$\vec{E}_i(\vec{r}, t) = \vec{n}_i E_0 \exp[i(\vec{k}_i \cdot \vec{r} - \omega_i t)]$$

where \vec{n}_i is a unit vector in the direction of the incident field, E_0 is the field amplitude, \vec{k}_i is the incident wavevector and ω_i is the incident angular frequency. The plane wave is incident upon a medium with local dielectric function $\epsilon(\vec{r}, t) = \epsilon_0 \mathbf{I} + \delta\epsilon(\vec{r}, t)$, where $\delta\epsilon(\vec{r}, t)$ is the dielectric tensor fluctuation at position \vec{r} and time t , and \mathbf{I} is a unit second-rank tensor. Basic light scattering theory can be used to find the inelastically scattered light spectrum. If the scattered field at the detector is also in the direction \vec{n}_i (i.e. $\vec{n}_f = \vec{n}_i$ for a polarized light scattering experiment), the scattered wavevector is $\vec{k}_f = \vec{k}_i - \vec{k}$ and the scattered frequency is $\omega_f = \omega_i - \omega$, then, apart from some known constant factors that depend on the geometry of the experiment and the incident field, the inelastically scattered light intensity is proportional to the spectral density of the local dielectric fluctuations. If the medium is isotropic and made up of spherically symmetrical molecules, then the dielectric tensor is proportional to the unit tensor \mathbf{I} : $\epsilon(\vec{r}, t) = [\epsilon_0 + \delta\epsilon(\vec{r}, t)]\mathbf{I}$. From the dielectric equation of state $\epsilon_0 = \epsilon(\rho_0, T_0)$, one can proceed to obtain the local dielectric fluctuation as $\delta\epsilon(\vec{r}, t) = (\partial\epsilon/\partial\rho)_T \delta\rho(\vec{r}, t) + (\partial\epsilon/\partial T)_\rho \delta T(\vec{r}, t)$. In many simple fluids, it is experimentally found that the thermodynamic derivative $(\partial\epsilon/\partial T)_\rho$ is approximately zero. One then has a simple result that

$$I_\epsilon(\vec{k}, \omega) \sim \left(\frac{\partial\epsilon}{\partial\rho} \right)_T^2 S_{\rho\rho}(\vec{k}, \omega) \quad (\text{A3.3.1})$$

where $S_{\rho\rho}(\vec{k}, \omega)$ is the spectrum of density fluctuations in the simple fluid system. $S_{\rho\rho}(\vec{k}, \omega)$ is the space–time Fourier transform of the density–density correlation function $S_{\rho\rho}(\vec{r}, t; \vec{r}', t') = \langle \delta\rho(\vec{r}, t) \delta\rho(\vec{r}', t') \rangle$.

Depending on the type of scattering probe and the scattering geometry, other experiments can probe other similar correlation functions. Elastic scattering experiments effectively measure frequency integrated spectra and, hence, probe only the space-dependent static structure of a system. Electron scattering experiments probe charge density correlations, and magnetic neutron scattering experiments the spin density correlations. Inelastic thermal neutron scattering from a non-magnetic system is a sharper probe of density–density correlations in a system but, due to the shorter wavelengths and higher frequencies involved, these results are complementary to those obtained from inelastic polarized light scattering experiments. The latter provide space–time correlations in the long-wavelength hydrodynamic regime.

In order to analyse results from such experiments, it is appropriate to consider a general framework, linear response theory, which is useful whenever the probe radiation weakly couples to the system. The linear response framework is also convenient for utilizing various symmetry and analyticity properties of correlation functions and response functions, thereby reducing the general problem to determining quantities which are amenable to approximations in such a way that the symmetry and analyticity properties are left intact. Such approximations are necessary in order to avoid the full complexity of many-body dynamics. The central quantity in the linear response theory is the response function. It is related to the corresponding correlation function (typically obtained from experimental measurements) through a fluctuation dissipation theorem. In the next section, section A3.3.2.1, we discuss only the subset of necessary results from the linear response theory, which is described in detail in the book by Forster (see [Further Reading](#)).

A3.3.2.1 LINEAR RESPONSE THEORY

Consider a set of physical observables $\{A_i(\vec{r}, t)\}$. If a small external field $\delta a_i^{\text{ext}}(\vec{r}, t)$ couples to the observable A_i , then in presence of a set of small external fields $\{\delta a_i\}$, the Hamiltonian H of a system is perturbed to

$$\mathcal{H}(t) = H - \sum_i \int d\vec{r} A_i(\vec{r}) \delta a_i^{\text{ext}}(\vec{r}, t) \quad (\text{A3.3.2})$$

in a Schrödinger representation. One can use time-dependent perturbation theory to find the linear response of the system to the small external fields. If the system is in equilibrium at time $t = -\infty$, and is evolved under $\mathcal{H}(t)$, the effect on $A_i(\vec{r}, t)$ which is $\delta\langle A_i \rangle = \langle A_i \rangle_{\text{noneq}} - \langle A_i \rangle_{\text{eq}}$ can be calculated to first order in external fields. The result is (causality dictates the upper limit of time integration to t)

$$\delta\langle A_i(\vec{r}, t) \rangle = \sum_j \int_{-\infty}^t dt' \int d\vec{r}' 2i \chi_{ij}''(\vec{r}, t; \vec{r}', t') \delta a_j^{\text{ext}}(\vec{r}', t') \quad (\text{A3.3.3})$$

where the response function (matrix) is given by

$$\chi_{ij}''(\vec{r}, t; \vec{r}', t') = \chi_{ij}''(\vec{r} - \vec{r}', t - t') = \left\langle \frac{1}{2\hbar} [A_i(\vec{r}, t), A_j(\vec{r}', t')] \right\rangle \quad (\text{A3.3.4})$$

in a translationally invariant system. Note that the response function is an equilibrium property of the system with Hamiltonian H , independent of the small external fields $\{\delta a_i\}$. In the classical limit (see [section A2.2.3](#)) the quantum mechanical commutator becomes the classical Poisson bracket and the response function reduces to $\langle (i/2)[A_i(\vec{r}, t), A_j(\vec{r}', t')]_{\text{P.B.}} \rangle$.

Since typical scattering experiments probe the system fluctuations in the frequency–wavenumber space, the Fourier transform $\chi''_{ij}(\vec{k}, \omega)$ is closer to measurements, which is in fact the imaginary (dissipative) part of the response function (matrix) defined as

$$\chi_{ij}(\vec{k}, z) = \int \frac{d\omega}{\pi} \frac{\chi''_{ij}(\vec{k}, \omega)}{\omega - z} \quad (\text{Im } z \neq 0). \quad (\text{A3.3.5})$$

The real part of $\chi_{ij}(\vec{k}, \omega)$, χ'_{ij} , is the dispersive (reactive) part of χ_{ij} , and the definition of χ_{ij} implies a relation between χ'_{ij} and χ''_{ij} which is known as the Kramers–Kronig relation.

The response function $\chi''_{ij}(\vec{r}, t)$, which is defined in equation (A3.3.4), is related to the corresponding correlation function, $S_{ij}(\vec{r}, t)$ through the fluctuation dissipation theorem:

$$\chi''_{ij}(\vec{k}, \omega) = \frac{1}{2\hbar} (1 - e^{-\beta\hbar\omega}) S_{ij}(\vec{k}, \omega). \quad (\text{A3.3.6})$$

The fluctuation dissipation theorem relates the dissipative part of the response function (χ'') to the correlation of fluctuations (A_i), for any system in thermal equilibrium. The left-hand side describes the dissipative behaviour of a many-body system: all or part of the work done by the external forces is irreversibly distributed into the infinitely many degrees of freedom of the thermal system. The correlation function on the right-hand side describes the manner in which a fluctuation arising spontaneously in a system in thermal equilibrium, even in the absence of external forces, may dissipate in time. In the classical limit, the fluctuation dissipation theorem becomes $\chi''_{ij}(\vec{k}, \omega) = (\beta/2)\omega S_{ij}(\vec{k}, \omega)$.

There are two generic types of external fields that are of general interest. In one of these, which relates to the scattering experiments, the external fields are to be taken as periodic perturbations

$$\delta a_i^{\text{ext}}(\vec{r}, t) = \delta a_i(\vec{r}) \exp[-(\eta - i\omega)t]$$

where η is an infinitesimally small negative constant, and $\delta a_i(\vec{r})$ can also be a periodic variation in \vec{r} , as in the case for incident plane wave electromagnetic radiation considered earlier.

In the other class of experiments, the system, in equilibrium at $t = -\infty$, is adiabatically perturbed to a non-equilibrium state which gets fully switched on by $t = 0$, through the field, $\delta a_i^{\text{ext}}(\vec{r}, t) = \delta a_i(\vec{r}) e^{\varepsilon t}$, $t \leq 0$, with ε an infinitesimally small positive constant. At $t = 0$ the external field is turned off, and the system so prepared in a non-equilibrium state will, if left to itself, relax back to equilibrium. This is the generic relaxation experiment during which the decay of the initial ($t = 0$) value is measured. Such an external field will produce, at $t = 0$, spatially varying initial values $\delta \langle A_i(\vec{r}, t = 0) \rangle$ whose spatial Fourier transforms are given by

$$\delta\langle A_i(\vec{k}, t = 0) \rangle = \sum \chi_{ij}(\vec{k}) \delta a_j(\vec{k}) \quad (\text{A3.3.7})$$

where

$$\chi_{ij}(\vec{k}) = \int \frac{dw}{\pi} \frac{\chi''_{ij}(\vec{k}, w)}{w}. \quad (\text{A3.3.8})$$

If $\delta a_i(\vec{r})$ is slowly varying in space, the long-wavelength limit $\chi_{ij}(\vec{k} \rightarrow 0)$ reduces to a set of static susceptibilities or thermodynamic derivatives. Now, since for $t > 0$ the external fields are zero, it is useful to evaluate the one-sided transform

$$\begin{aligned} \delta\langle A_i \rangle(\vec{k}, z) &= \int_0^\infty dt e^{izt} \delta\langle A_i(\vec{k}, t) \rangle \\ &= \sum_j \int \frac{dw}{i\pi} \frac{\chi''_{ij}(\vec{k}, w)}{w(w-z)} \cdot \delta a_j(\vec{k}) \\ &= \frac{1}{iz} \sum_j [\chi(\vec{k}, z) \chi^{-1}(\vec{k}) - 1]_{ij} \cdot \delta\langle A_j(\vec{k}, 0) \rangle. \end{aligned} \quad (\text{A3.3.9})$$

The second equality is obtained using the form of the external field $\delta a_i^{\text{ext}}(\vec{r}, t)$ specific to the relaxation experiments. The last equality is to be read as a matrix equation. The system stability leads to the positivity of all susceptibilities $\chi(\vec{k})$, so that its inverse exists. This last equality is superior to the second one, since the external fields δa_i have been eliminated in favour of the initial values $\delta\langle A_j(\vec{k}, t=0) \rangle$, which are directly measurable in a relaxation experiment. It is then possible to analyse the relaxation experiments by obtaining the measurements for positive times and comparing them to $\delta\langle A_i(\vec{k}, t) \rangle$ as evaluated in terms of the initial values using some approximate model for the dynamics of the system's evolution. One such model is a linear hydrodynamic description.

-6-

A3.3.2.2 FLUCTUATIONS IN THE HYDRODYNAMIC DOMAIN

We start with a simple example: the decay of concentration fluctuations in a binary mixture which is in equilibrium. Let $\delta C(\vec{r}, t) = C(\vec{r}, t) - C_0$ be the concentration fluctuation field in the system where C_0 is the mean concentration. C is a conserved variable and thus satisfies a continuity equation:

$$\quad (\text{A3.3.10})$$

where a phenomenological linear constitutive relation relates the concentration flux \vec{j}_c to the gradient of the local chemical potential $\mu(\vec{r}, t)$ as follows:

$$\vec{j}_c(\vec{r}, t) = -L \vec{\nabla} \mu(\vec{r}, t). \quad (\text{A3.3.11})$$

Here L is the Onsager coefficient and the minus sign ($-$) indicates that the concentration flow occurs from regions of high μ to low μ in order that the system irreversibly flows towards the equilibrium state of a

uniform chemical potential. In a system slightly away from equilibrium, the dependence of μ on the thermodynamic state variables, concentration, pressure and temperature (C, p, T) , would, in general, relate changes in the chemical potential like $\bar{\nabla}\mu$ to $\bar{\nabla}C$, $\bar{\nabla}p$ and $\bar{\nabla}T$. However, for most systems the thermodynamic derivatives $\partial\mu/\partial p$ and $\partial\mu/\partial T$ are small, and one has, to a good approximation, $\bar{\nabla}\mu = (\partial\mu/\partial C)_{p,T} \bar{\nabla}\delta C$. This linear approximation is not always valid; however, it is valid for the thermodynamic fluctuations in a binary mixture at equilibrium. It enables us to thus construct a closed linear equation for the concentration fluctuations, the diffusion equation:

$$\frac{\partial\delta C}{\partial t} = D\nabla^2\delta C \quad (\text{A3.3.12})$$

where $D=L(\partial\mu/\partial C)_{p,T}$ is the diffusion coefficient, which we assume to be a constant. The diffusion equation is an example of a hydrodynamic equation. The characteristic ingredients of a hydrodynamic equation are a conservation law and a linear transport law.

The solutions of such partial differential equations require information on the spatial boundary conditions and initial conditions. Suppose we have an infinite system in which the concentration fluctuations vanish at the infinite boundary. If, at $t = 0$ we have a fluctuation at origin $\delta C(\vec{r}, 0) = \Delta C_0 \delta(\vec{r})$, then the diffusion equation can be solved using the spatial Fourier transforms. The solution in Fourier space is $C(k, t) = \exp(-Dk^2 t) \Delta C_0$, which can be inverted analytically since the Fourier transform of a Gaussian is a Gaussian. In real space, the initial fluctuation decays in time in a manner such that the initial delta function fluctuation broadens to a Gaussian whose width increases in time as $(Dt)^{d/2}$ for a d -dimensional system, while the area under the Gaussian remains equal to ΔC_0 due to the conservation law. Linear hydrodynamics are not always valid. For this example, near the consolute (critical) point of the mixture, the concentration fluctuations nonlinearly couple to transverse velocity modes and qualitatively change the result. Away from the critical point, however, the above, simple analysis illustrates the manner in which the thermodynamic

-7-

fluctuations decay in the hydrodynamic (i.e. long-wavelength) regime. The diffusion equation and its solutions constitute a rich subject with deep connections to brownian motion theory [1]: both form a paradigm for many other models of dynamics in which diffusion-like decay and damping play important roles.

In dense systems like liquids, the molecular description has a large number of degrees of freedom. There are, however, a few collective degrees of freedom, collective modes, which when perturbed through a fluctuation, relax to equilibrium very slowly, i.e. with a characteristic decay time that is long compared to the molecular interaction time. These modes involve a large number of particles and their relaxation time is proportional to the square of their characteristic wavelength, which is large compared to the intermolecular separation. Hydrodynamics is suitable to describe the dynamics of such long-wavelength, slowly-relaxing modes.

In a hydrodynamic description, the fluid is considered as a continuous medium which is locally homogeneous and isotropic, with dissipation occurring through viscous friction and thermal conduction. For a one-component system, the hydrodynamic (collective) variables are deduced from conservation laws and broken symmetry. We first consider (section A3.3.2.3) the example of a Rayleigh–Brillouin spectrum of a one-component monatomic fluid. Here conservation laws play the important role. In the next example (section A3.3.2.4), we use a fluctuating hydrodynamic description for capillary waves at a liquid–vapour interface where broken symmetry plays an important role. A significant understanding of underlying phenomena for each of these examples has been obtained using linear hydrodynamics [2], even though, in principle, nonlinear dynamical aspects are within the exact dynamics of these systems.

In the next section we discuss linear hydrodynamics and its role in understanding the inelastic light scattering experiments from liquids, by calculating the density–density correlation function, $S_{\rho\rho}$.

A3.3.2.3 RAYLEIGH–BRILLOUIN SPECTRUM

The three conservation laws of mass, momentum and energy play a central role in the hydrodynamic description. For a one-component system, these are the only hydrodynamic variables. The mass density has an interesting feature in the associated continuity equation: the mass current (flux) is the momentum density and thus itself is conserved, in the absence of external forces. The mass density $\rho(\vec{r}, t)$ satisfies a continuity equation which can be expressed in the form (see, for example, the book on fluid mechanics by Landau and Lifshitz, cited in the Further Reading)

$$\left(\frac{\partial}{\partial t} + \vec{v} \cdot \vec{\nabla}\right) \rho = -\rho \vec{\nabla} \cdot \vec{v}. \quad (\text{A3.3.13})$$

The equation of momentum conservation, along with the linear transport law due to Newton, which relates the dissipative stress tensor to the rate of strain tensor $e_{ik} = \frac{1}{2}(\nabla_i v_k + \nabla_k v_i)$, and which introduces two transport coefficients, shear viscosity η and bulk viscosity η_b , lead to the equation of motion for a Newtonian fluid:

$$\left(\frac{\partial}{\partial t} + \vec{v} \cdot \vec{\nabla}\right) \vec{v} = -\frac{1}{\rho} \vec{\nabla} p + \nu \nabla^2 \vec{v} + (v_l - \nu) \vec{\nabla}(\vec{\nabla} \cdot \vec{v}) \quad (\text{A3.3.14})$$

-8-

where the kinematic viscosity $\nu = \eta/\rho$ and the kinematic longitudinal viscosity $v_l = (\frac{4}{3}\eta + \eta_b)/\rho$.

The energy conservation law also leads to an associated continuity equation for the total energy density. The total energy density contains both the kinetic energy density per unit volume and the internal energy density. The energy flux is made up of four terms: a kinematic term, the rates of work done by reversible pressure and dissipative viscous stress, and a dissipative heat flux. It is the dissipative heat flux that is assumed to be proportional to the temperature gradient and this linear transport law, Fourier's law, introduces as a proportionality coefficient, the coefficient of thermal conductivity, κ . From the resulting energy equation, one can obtain the equation for the rate of entropy balance in the system, which on account of the irreversibility and the arrow of time implied by the Second Law of Thermodynamics leads to the result that each of the transport coefficients η , η_b and κ is a positive definite quantity. Using the mass conservation equation (A3.3.13), and thermodynamic relations which relate entropy change to changes in density and temperature, the entropy balance equation can be transformed to the hydrodynamic equation for the local temperature $T(\vec{r}, t)$:

$$\begin{aligned} \left(\frac{\partial}{\partial t} + \vec{v} \cdot \vec{\nabla}\right) T = & -\alpha^{-1}(\gamma - 1) \vec{\nabla} \cdot \vec{v} + (\rho C_V)^{-1} [\vec{\nabla} \cdot (\kappa \vec{\nabla} T) \\ & + 2\eta[e_{ik} - \frac{1}{3}e_{jj}\delta_{ik}]^2 + \eta_b[e_{jj}]^2] \end{aligned} \quad (\text{A3.3.15})$$

where α is the thermal expansion coefficient, $\gamma = C_p/C_V$, C_p the heat capacity per unit mass at constant pressure and C_V the same at constant volume. e_{ik} is the rate of the strain tensor defined above, and a repeated

subscript implies a summation over that subscript, here and below.

The three [equation \(A3.3.13\)](#), [equation \(A3.3.14\)](#) and [equation \(A3.3.15\)](#) are a useful starting point in many hydrodynamic problems. We now apply them to compute the density–density correlation function

$$S_{\rho\rho}(\vec{r}, t; \vec{r}', t') = \langle \delta\rho(\vec{r}, t) \delta\rho(\vec{r}', t') \rangle. \quad (\text{A3.3.16})$$

Since the fluctuations are small, it is appropriate to linearize the three equations in $\delta\rho(\vec{r}, t) = \rho - \rho_0$, $\delta T(\vec{r}, t) = T - T_0$ and $\vec{v}(\vec{r}, t) = \vec{v}$, by expanding around their respective equilibrium values ρ_0 , T_0 and zero, where we assume that the mean fluid velocity is zero. The linearization eliminates the advective term $\vec{v} \cdot \vec{\nabla}$ etc from each of the three equations, and also removes the bilinear viscous dissipation terms from the temperature equation (A3.3.15). The $\vec{\nabla} p$ term in the velocity [equation \(A3.3.14\)](#) can be expressed in terms of density and temperature gradients using thermodynamic derivative identities:

$$\vec{\nabla} p = \left(\frac{\partial p}{\partial \rho} \right)_T \left[\vec{\nabla} \delta\rho - \left(\frac{\partial \rho}{\partial T} \right)_p \vec{\nabla} \delta T \right] = \frac{c^2}{\gamma} \left[\vec{\nabla} \delta\rho + \rho_0 \alpha \vec{\nabla} \delta T \right]$$

where $(\partial p / \partial \rho)_T = c_T^2 = c^2 / \gamma$ with c_T and c the isothermal and adiabatic speeds of sound, respectively. The momentum equation then linearizes to

-9-

$$\frac{\partial \vec{v}}{\partial t} + \frac{c^2}{\gamma \rho_0} \left[\vec{\nabla} \delta\rho + \rho_0 \alpha \vec{\nabla} \delta T \right] - \nu \nabla^2 \vec{v} - (\nu_l - \nu) \vec{\nabla} (\vec{\nabla} \cdot \vec{v}) = 0. \quad (\text{A3.3.17})$$

The linearized equations for density and temperature are:

$$\frac{\partial \delta\rho}{\partial t} + \rho_0 \vec{\nabla} \cdot \vec{v} = 0 \quad (\text{A3.3.18})$$

and

$$\frac{\partial \delta T}{\partial t} + \alpha^{-1} (\gamma - 1) \vec{\nabla} \cdot \vec{v} - \gamma D_T \nabla^2 \delta T = 0. \quad (\text{A3.3.19})$$

Here the thermal diffusivity $D_T \equiv \kappa / (\rho_0 C_p)$. These two equations couple only to the longitudinal part $\Psi \equiv \vec{\nabla} \cdot \vec{v}$ of the fluid velocity. From [equation \(A3.3.17\)](#) it is easy to see that Ψ satisfies

$$\frac{\partial \Psi}{\partial t} + \frac{c^2}{\gamma \rho_0} \left[\nabla^2 \delta\rho + \rho_0 \alpha \nabla^2 \delta T \right] - \nu_l \nabla^2 \Psi. \quad (\text{A3.3.20})$$

Out of the five hydrodynamic modes, the polarized inelastic light scattering experiment can probe only the three modes represented by [equation \(A3.3.18\)](#), [equation \(A3.3.19\)](#) and [equation \(A3.3.20\)](#). The other two modes, which are in [equation \(A3.3.17\)](#), decouple from the density fluctuations; these are due to transverse

velocity components which is the vorticity $\vec{\omega} \equiv \vec{\nabla} \times \vec{v}$. Vorticity fluctuations decay in a manner analogous to that of the concentration fluctuations discussed in [section A3.3.2.2](#), if one considers the vorticity fluctuation Fourier mode of the wavevector \vec{k} . Then the correlations of the k th Fourier mode of vorticity also decays in an exponential manner with the form $\exp(-vk^2t)$.

The density fluctuation spectrum can be obtained by taking a spatial Fourier transform and a temporal Laplace transform of the three coupled equation (A3.3.18), equation (A3.3.19) and equation (A3.3.20), and then solving the resulting linear coupled algebraic set for the density fluctuation spectrum. (See details in the books by Berne–Pecora and Boon–Yip.) The result for $S_{\rho\rho}(\vec{k}, w)$ given below is proportional to its frequency integral $S_{\rho\rho}(\vec{k})$ which is the liquid structure factor discussed earlier in [section A2.2.5.2](#). The density fluctuation spectrum is

$$\begin{aligned} \frac{S_{\rho\rho}(k, w)}{S_{\rho\rho}(k)} &= 2 \operatorname{Re} \lim_{\epsilon \rightarrow 0} \frac{\langle \delta\rho^*(k, t = 0) \delta\rho(k, s = \epsilon + iw) \rangle}{\langle \delta\rho^*(k, t = 0) \delta\rho(k, t = 0) \rangle} \\ &= \frac{\gamma - 1}{\gamma} \frac{2D_T k^2}{w^2 + (D_T k^2)^2} + \frac{1}{\gamma} \left[\frac{\Gamma k^2}{(w + ck)^2 + (\Gamma k^2)^2} + \frac{\Gamma k^2}{(w - ck)^2 + (\Gamma k^2)^2} \right] \\ &\quad + \frac{1}{\gamma} [\Gamma + (\gamma - 1)D_T] \frac{k}{c} \left[\frac{(w + ck)}{(w + ck)^2 + (\Gamma k^2)^2} - \frac{(w - ck)}{(w - ck)^2 + (\Gamma k^2)^2} \right] \end{aligned} \quad (\text{A3.3.21})$$

-10-

where $\Gamma = \frac{1}{2}[v_1 + (\gamma - 1)D_T]$.

This is the result for monatomic fluids and is well approximated by a sum of three Lorentzians, as given by the first three terms on the right-hand side. The physics of these three Lorentzians can be understood by thinking about a local density fluctuation as made up of thermodynamically independent entropy and pressure fluctuations: $\rho = \rho(s, p)$. The first term is a consequence of the thermal processes quantified by the entropy fluctuations at constant pressure, which lead to the decaying mode $[(\gamma - 1)/\gamma] \exp[-D_T k^2 |t|]$ and the associated Lorentzian known as the *Rayleigh* peak is centred at zero frequency with a half-width at half-maximum of $D_T k^2$. The next two terms (Lorentzians) arise from the mechanical part of the density fluctuations, the pressure fluctuations at constant entropy. These are the adiabatic sound modes $(1/\gamma) \exp[-\Gamma k^2 |t|] \cos[\omega(k)|t|]$ with $\omega(k) = \pm ck$, and lead to the two spectral lines (Lorentzians) which are shifted in frequency by $-ck$ (Stokes line) and $+ck$ (anti-Stokes line). These are known as the *Brillouin–Mandelstam* doublet. The half-width at half-maximum of this pair is Γk^2 which gives the attenuation of acoustic modes. In dense liquids, the last two terms in the density fluctuation spectrum above are smaller by orders of magnitude compared to the three Lorentzians, and lead to s-shaped curves centred at $w = \pm ck$. They cause a weak asymmetry in the Brillouin peaks which induces a slight pulling of their position towards the central Rayleigh peak. The Rayleigh–Brillouin spectrum from liquid argon, as measured by an inelastic polarized light scattering experiment, is shown in figure A3.3.1. An accurate measurement of the Rayleigh–Brillouin lineshape can be used to measure many of the thermodynamic and transport properties of a fluid. The ratio of the integrated intensity of the Rayleigh peak to those of the Brillouin peaks, known as the Landau–Placzek ratio, is $(I_R)/(2I_B) = (\gamma - 1)$, and directly measures the ratio of specific heats γ . From the position of the Brillouin peaks one can obtain the adiabatic speed of sound c , and knowing γ and c one can infer isothermal compressibility. From the width of the Rayleigh peak, one can obtain thermal diffusivity (and if C_p is known, the thermal conductivity κ). Then from the width of the Brillouin peaks, one can obtain the longitudinal viscosity (and, if shear viscosity is known, the bulk viscosity).

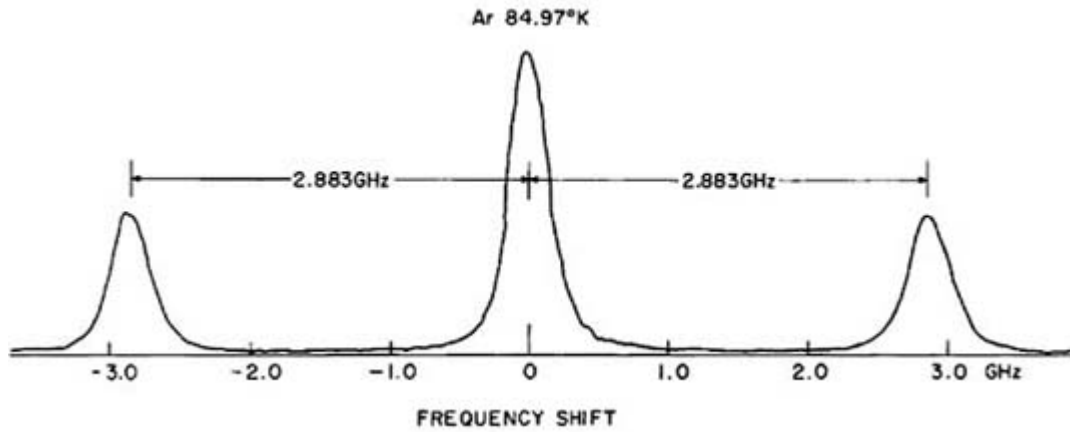


Figure A3.3.1 Rayleigh–Brillouin spectrum from liquid argon, taken from [4].

A large variety of scattering experiments (inelastic light scattering using polarized and depolarized set ups, Raman scattering, inelastic neutron scattering) have been used over the past four decades to probe and understand the spatio–temporal correlations and molecular dynamics in monatomic and polyatomic fluids in equilibrium, spanning the density range from low-density gases to dense liquids. In the same fashion, concentration fluctuations in binary mixtures have also been probed. See [3, 4, 5, 6, 7 and 8] for further reading for these topics.

In the next section, we consider thermal fluctuations in an inhomogeneous system.

A3.3.2.4 CAPILLARY WAVES

In this section we discuss the frequency spectrum of excitations on a liquid surface. While we used linearized equations of hydrodynamics in the last section to obtain the density fluctuation spectrum in the bulk of a homogeneous fluid, here we use linear *fluctuating* hydrodynamics to derive an equation of motion for the instantaneous position of the interface. We then use this equation to analyse the fluctuations in such an inhomogeneous system, around equilibrium and around a NESS characterized by a small temperature gradient. More details can be found in [9, 10].

Surface waves at an interface between two immiscible fluids involve effects due to gravity (g) and surface tension (σ) forces. (In this section, σ denotes surface tension and σ_{ik} denotes the stress tensor. The two should not be confused with one another.) In a hydrodynamic approach, the interface is treated as a sharp boundary and the two bulk phases as incompressible. The Navier–Stokes equations for the two bulk phases (balance of macroscopic forces is the ingredient) along with the boundary condition at the interface (surface tension σ enters here) are solved for possible harmonic oscillations of the interface of the form, $\exp[-(i\omega + \varepsilon)t + i\vec{q}\cdot\vec{x}]$, where ω is the frequency, ε is the damping coefficient, \vec{q} is the $2-d$ wavevector of the periodic oscillation and \vec{x} a $2-d$ vector parallel to the surface. For a liquid–vapour interface which we consider, away from the critical point, the vapour density is negligible compared to the liquid density and one obtains the hydrodynamic dispersion relation for surface waves $\omega_s^2 = (\sigma/\rho_0)q^3 + gq$. The term gq in the dispersion relation arises from the gravity waves, and dominates for macroscopic wavelengths, but becomes negligible for wavelengths shorter than the capillary constant $(2\sigma/g\rho_0)^{1/2}$, which is of the order of a few millimetres for water. In what follows we discuss phenomena at a planar interface (for which g is essential), but restrict ourselves to the capillary waves regime and set $g = 0^+$. Capillary wave dispersion is then $\omega_c(q) = (\frac{\sigma}{\rho_0})^{1/2}q^{3/2}$, and the

damping coefficient $\varepsilon(q) = (2\eta/\rho_0)q^2$. Consider a system of coexisting liquid and vapour contained in a cubical box of volume L^3 . An external, infinitesimal gravitational field locates the liquid of density ρ_l in the region $z < -\xi$, while the vapour of lower density ρ_v is in the region $z > \xi$. A flat surface, of thickness 2ξ , is located about $z = 0$ in the $\vec{x} = (x, y)$ plane. The origin of the z axis is defined in accord with Gibbs' prescription:

$$\int_{-L/2}^0 [\rho(z) - \rho_l] dz + \int_0^{L/2} [\rho(z) - \rho_v] dz = 0 \quad (\text{A3.3.22})$$

where $\rho(z)$ is the equilibrium density profile of the inhomogeneous system. Let us first consider the system in equilibrium. Let it also be away from the critical point. Then $\rho_v \ll \rho_l$ and the interface thickness is only a few nanometres, and a model with zero interfacial width and a step function profile (Fowler model) is appropriate. Also, since the speed of sound is much larger than the capillary wave speed we can assume the liquid to be incompressible, which implies a constant ρ in the liquid and, due to the mass continuity equation (equation (A3.3.13)), also implies $\vec{\nabla} \cdot \vec{v} = 0$. Furthermore, if the amplitude of the capillary waves is small, the nonlinear convective (advective) term $(\vec{v} \cdot \vec{\nabla} \vec{v})$ can also be ignored in (A3.3.14). The approach of fluctuating hydrodynamics corresponds to having additional Gaussian random stress-tensor fluctuations in the Newtonian transport law and analogous heat flux fluctuations in the Fourier transport law. These fluctuations arise from those short lifetime degrees of freedom that are not included in a hydrodynamic description, a description based only on long-lifetime conserved hydrodynamic variables.

-12-

The equations of motion for the bulk fluid for $z < 0$ are:

$$\frac{\partial v_x}{\partial x} + \frac{\partial v_z}{\partial z} = 0 \quad (\text{A3.3.23})$$

which is the continuity equation, and

$$\begin{aligned} \frac{\partial v_x}{\partial t} &= -\frac{\partial p}{\partial x} + \frac{\eta}{\rho} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial z^2} \right) v_x + \frac{\partial}{\partial x} \left(\frac{s_{xx}}{\rho} \right) + \frac{\partial}{\partial z} \left(\frac{s_{xz}}{\rho} \right) \\ \frac{\partial v_z}{\partial t} &= -\frac{\partial p}{\partial z} + \frac{\eta}{\rho} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial z^2} \right) v_z + \frac{\partial}{\partial z} \left(\frac{s_{zz}}{\rho} \right) + \frac{\partial}{\partial x} \left(\frac{s_{xz}}{\rho} \right). \end{aligned}$$

These are the two components of the Navier–Stokes equation including fluctuations s_{ij} , which obey the fluctuation dissipation theorem, valid for incompressible, classical fluids:

$$\begin{aligned} \langle s_{ik}(\vec{x}, z, t) s_{lm}(\vec{x}', z', t') \rangle_{\text{eq}} &= 2kT\eta [\delta_{il}\delta_{km} + \delta_{im}\delta_{kl} - \frac{2}{3}\delta_{ik}\delta_{lm}] \\ &\delta(\vec{x} - \vec{x}')\delta(z - z')\delta(t - t'). \end{aligned} \quad (\text{A3.3.24})$$

This second moment of the fluctuations around equilibrium also defines the form of ensemble $\langle \dots \rangle_{\text{eq}}$ for the equilibrium average at temperature T .

Surface properties enter through the Young–Laplace equation of state for the ‘surface pressure’ P_{sur} :

$$P_{\text{sur}} = -\sigma \frac{\partial^2}{\partial x^2} \zeta \quad \text{at } z = 0. \quad (\text{A3.3.25})$$

The *non-conserved* variable $\zeta(\vec{x}, t)$ is a *broken symmetry variable*; it is the instantaneous position of the Gibbs' surface, and it is the translational symmetry in z direction that is broken by the inhomogeneity due to the liquid–vapour interface. In a more microscopic statistical mechanical approach [9], it is related to the number density fluctuation $\delta\rho(\vec{x}, z, t)$ as

$$\zeta(\vec{x}, t) \simeq (\rho_l - \rho_v)^{-1} \int_{-\xi}^{\xi} dz \delta\rho(\vec{x}, z, t) \quad (\text{A3.3.26})$$

but in the present hydrodynamic approach it is defined by

$$\frac{\partial \zeta}{\partial t} = v_z \quad \text{at } z = 0. \quad (\text{A3.3.27})$$

-13-

The boundary conditions at the $z=0$ surface arise from the mechanical equilibrium, which implies that both the normal and tangential forces are balanced there. This leads to

$$p = -\sigma \frac{\partial^2 \zeta}{\partial x^2} + 2\eta \frac{\partial v_z}{\partial z} + s_{zz} \quad \text{at } z = 0 \quad (\text{A3.3.28})$$

$$\eta \left(\frac{\partial v_x}{\partial z} + \frac{\partial v_z}{\partial x} \right) = -s_{xz} \quad \text{at } z = 0. \quad (\text{A3.3.29})$$

If the surface tension is a function of position, then there is an additional term, $\partial\sigma/\partial x$, to the right-hand side in the last equation. From the above description it can be shown that the equation of motion for the Fourier component $\zeta(\vec{q}, t)$ of the broken symmetry variable ζ is

$$\frac{\partial^2 \zeta(\vec{q}, t)}{\partial t^2} + 2\epsilon(q) \frac{\partial \zeta(\vec{q}, t)}{\partial t} + w_c^2(q) \zeta(\vec{q}, t) = -\frac{q^2}{\rho} \int_{-\infty}^0 dz e^{qz} [s_{zz}(\vec{q}, z, t) - s_{xx}(\vec{q}, z, t) - 2is_{xz}(\vec{q}, z, t)] \quad (\text{A3.3.30})$$

where $\epsilon(q)$ and $w_c(q)$ are the damping coefficient and dispersion relation for the capillary waves defined earlier. This damped driven harmonic oscillator equation is driven by spontaneous thermal fluctuations and is valid in the small viscosity limit. It does not have any special capillary wave fluctuations. The thermal random force fluctuations s_{ij} are in the bulk and are coupled to the surface by the e^{qz} factor. This surface–bulk coupling is an essential ingredient of any hydrodynamic theory of the liquid surface: the surface is not a separable phase.

We now evaluate the spectrum of interfacial fluctuations $S(\vec{q}, \omega)$. It is the space–time Fourier transform of the correlation function $\langle \zeta(\vec{x}, t) \zeta(\vec{x}', t') \rangle$. It is convenient to do this calculation first for the fluctuations around a NESS which has a small constant temperature gradient, no convection and constant pressure. The corresponding results for the system in equilibrium are obtained by setting the temperature gradient to zero.

There are three steps in the calculation: first, solve the full nonlinear set of hydrodynamic equations in the steady state, where the time derivatives of all quantities are zero; second, linearize about the steady-state solutions; third, postulate a non-equilibrium ensemble through a generalized fluctuation dissipation relation.

A steady-state solution of the full nonlinear hydrodynamic equations is $\vec{v}=0$, $p = \text{constant}$ and $\partial T/\partial x = \text{constant}$, where the yz walls perpendicular to the xy plane of the interface are kept at different temperatures. This steady-state solution for a *small* temperature gradient means that the characteristic length scale of the temperature gradient $(\partial \ln T/\partial x)^{-1} \ll L$. The solution also implicitly means that the thermal expansion coefficient and surface excess entropy are negligible, i.e. $(\partial \ln \rho/\partial \ln T)$ and $(\partial \ln \sigma/\partial \ln T)$ are both approximately zero, which in turn ensures that there is no convection in the bulk or at the surface. We again assume that the fluid is incompressible and away from the critical point. Then, linearizing around the steady-state solution once again leads, for ζ , to the equation of motion identical to (A3.3.30), which in Fourier space (\vec{q}, w) can be written as (assuming $\varepsilon \ll w_c$)

-14-

$$[w^2 + 2\epsilon(q)w + w_c^2(q)]\zeta(\vec{q}, w) = -\frac{q^2}{\rho} \int_{-\infty}^0 dz e^{qz} [s_{zz}(\vec{q}, z, w) - s_{xx}(\vec{q}, z, w) - 2is_{xz}(\vec{q}, z, w)]. \quad (\text{A3.3.31})$$

The shear viscosity η is, to a good approximation, independent of T . So the only way the temperature gradient can enter the analysis is through the form of the non-equilibrium ensemble, i.e. through the random forces s_{ik} . Now we assume that the short-ranged random forces have the same form of the real-space correlation as in the thermal equilibrium case above (equation (A3.3.24)), but with T replaced by $T(\vec{x}) = T_0 + (\partial T/\partial \vec{x})_0 \cdot \vec{x}$. Thus the generalized fluctuation dissipation relation for a NESS, which determines the NESS ensemble, is

$$\begin{aligned} \langle s_{ik}(\vec{q}, z, w) s_{lm}(\vec{q}', z', w') \rangle_{\text{NESS}} &= 2kT\eta \left[\delta_{il}\delta_{km} + \delta_{im}\delta_{kl} - \frac{2}{3}\delta_{ik}\delta_{lm} \right] \\ &\quad \delta(z - z')\delta(w - w')(2\pi)^3 \\ &\quad \times \left[\delta(\vec{q} - \vec{q}') - \left(\frac{\partial \ln T}{\partial \vec{x}} \right)_0 \cdot \left(\frac{\partial}{\partial i(\vec{q} - \vec{q}')} \right) \delta(\vec{q} - \vec{q}') \right]. \end{aligned} \quad (\text{A3.3.32})$$

Then, from equation (A3.3.31) and equation (A3.3.32), we obtain the spectrum of interfacial fluctuations:

$$\langle \zeta(\vec{x}, w) \zeta(\vec{x}', w') \rangle = 2\pi \delta(w - w') \int \frac{d^2q}{(2\pi)^2} e^{i\vec{q} \cdot (\vec{x} - \vec{x}')} S(\vec{q}, w). \quad (\text{A3.3.33})$$

In the absence of a temperature gradient, i.e. in thermal equilibrium, the dynamic structure factor $S(\vec{q}, w)$ is

$$S(\vec{q}, w) = \frac{8kT\eta q^3/\rho^2}{[w^2 - w_c^2(q)]^2 + 4\epsilon^2(q)w^2} \quad (\text{A3.3.34})$$

which is sharply and symmetrically peaked at the capillary wave frequencies $w_c(q) = \pm(\sigma q^3/\rho)^{1/2}$. In the NESS, the result has asymmetry and is given by

$$S_{\text{NESS}}(\vec{q}, w) = S(\vec{q}, w)(1 - \frac{1}{2}\Delta(\vec{q}, w)) \quad (\text{A3.3.35})$$

where

$$\Delta(\vec{q}, w) = \frac{[2w^2 + w_c(q)^2]4w(\eta/\rho)\vec{q} \cdot (\partial \ln T / \partial \vec{x})_o}{[w^2 - w_c^2(q)]^2 + 4\epsilon^2(q)w^2}. \quad (\text{A3.3.36})$$

-15-

Since ζ is the ‘surface-averaged’ part of $\delta\rho$ from equation (A3.3.36), $S(\vec{q}, w)$ is the appropriately ‘surface-averaged’ density fluctuation spectrum near an interface, and is thus experimentally accessible. The correction term $\Delta(\vec{q}, w)$ is an odd function of frequency which creates an asymmetry in the heights of the two ripplon peaks. This is on account of the small temperature gradient breaking the time reversal symmetry: there are more ripples travelling from the hot side to the cold side than from cold to hot. One can also calculate the zeroth and first frequency moments of $S(\vec{q}, w)$:

$$S^0(q) = \frac{kT_o}{\sigma q^2} \quad (\text{A3.3.37})$$

and

$$S^1(q) = -\left[\frac{3}{8\eta} \hat{q} \cdot \left(\frac{\partial \ln T}{\partial \vec{x}} \right)_o \right] \frac{kT_o}{q^2}. \quad (\text{A3.3.38})$$

These are both long ranged in the long-wavelength limit $q \rightarrow 0$: $S^0(q)$ due to broken translational symmetry and $S^1(q)$ due to broken time reversal symmetry. $S^1(q)$ vanishes for fluctuations around equilibrium, and $S^0(q)$ is the same for both NESS and equilibrium. The results above are valid only for $L_\nabla \gg L \gg l_c$ where the two bounding length scales are respectively characteristic of the temperature gradient

$$L_\nabla = \left[\hat{q} \cdot \left(\frac{\partial \ln T}{\partial \vec{x}} \right)_o \right]^{-1}$$

and of the capillary wave mean free path

$$l_c \equiv \frac{w_c(q)}{q\epsilon(q)}.$$

The correction due to the temperature gradient in the capillary wave peak heights is the corresponding fractional difference, which can be obtained by evaluating $\Delta(\vec{q}, w = w_c)$. The result is simple:

$$\Delta(\vec{q}, w = w_c) = 3 \frac{l_c(q)}{L_\nabla}.$$

For the system in thermal equilibrium, one can compute the time-dependent mean square displacement $\langle |\zeta|^2 \rangle$

(q, t) , from the damped forced harmonic oscillator equation for ζ , [equation \(A3.3.30\)](#). The result is

$$\langle |\zeta|^2 \rangle(q, t) = \frac{kT_0}{\sigma q^2} \left[1 - \exp\left(-\frac{4\eta}{\rho} q^2 t\right) \right] \quad (\text{A3.3.39})$$

-16-

which goes to $S^0(q)$ as $t \rightarrow \infty$ as required. By integrating it over the two-dimensional wavevector \vec{q} , one can find the mean square displacement of the interface:

$$\langle \zeta^2 \rangle(t) = \frac{kT_0}{4\sigma} [\ln \tau + E_1(\tau) + C]$$

where $\tau = (4\eta q_{\max}^2 / \rho)t$ with $q_{\max} = 2\pi/a$, with a being a typical molecular size (diameter), E_1 is Euler's integral and $C \sim 0.577$ is Euler's number. Thus, as $t \rightarrow \infty$, $\langle \zeta^2 \rangle(t)$ diverges as $\ln t$, which is the dynamic analogue of the well known infrared divergence of the interfacial thickness. Numerically, the effect is small: at $t \sim 10^{18}$ s we find using typical values of T , σ , η and ρ such that $[\langle \zeta^2 \rangle(t)]^{1/2} \sim 9 \text{ \AA}$. From [equation \(A3.3.39\)](#) one can also proceed by first taking the $t \rightarrow \infty$ limit and then integrating over \vec{q} , with the result

$$\langle \zeta^2 \rangle = \frac{kT_0}{2\pi\sigma} \ln \frac{q_{\max}}{q_{\min}}$$

where $q_{\min} = 2\pi/L$. Again $\langle \zeta^2 \rangle$ shows a logarithmic infrared divergence as $2\pi/L \rightarrow 0$ which is the conventional result obtained from equilibrium statistical mechanics. This method hides the fact, which is transparent in the dynamic treatment, that the source of the divergence is the spontaneous random force fluctuations in the bulk, which drive the oscillations of the surface ζ . The equilibrium description of capillary wave excitations of a surface are often introduced in the framework of the so-called capillary wave model: the true free energy is functionally expanded around a 'bare' free energy in terms of the suppressed density fluctuations $\delta\rho$, and these 'capillary wave fluctuations' are assumed to be of the form

$$\delta\rho_{\text{cwf}} = -\zeta(x) \frac{d\rho(z)}{dz}. \quad (\text{A3.3.40})$$

It can be shown that this form leads to an unphysical dispersion relation for capillary waves: $w_c^2 \sim q^4$, rather than $\sim q^3$. This is precisely because of the neglect of the surface-bulk coupling in the above assumed form. One can show that a fluctuation *consistent* with capillary wave dispersion is

$$\delta\rho(\vec{x}, z, t) = -\zeta(x, t) \left[\frac{d\rho(z)}{dz} - \frac{q^2\sigma}{c^2} e^{qz} \right] \quad (\text{A3.3.41})$$

for $z > \xi$, where one neglects the vapour density, and where c is the speed of sound in the bulk phase coupled to the surface. Thus, if one wants to introduce density fluctuations into the description, the entire fluid has to be self-consistently treated as compressible. Physically, the first term $\zeta(d\rho(z)/dz)$ corresponds to a perturbation, or kick, of the interface, and the second term self-consistently accounts for the pressure fluctuations in the bulk due to that kick. Neglecting the second term amounts to violating momentum conservation, resulting in an incorrect 'energy-momentum relation' for the capillary wave excitations.

A3.3.3 NON-EQUILIBRIUM TIME-EVOLVING SYSTEMS

There are many examples in nature where a system is not in equilibrium and is evolving in time towards a thermodynamic equilibrium state. (There are also instances where non-equilibrium and time variation appear to be a persistent feature. These include chaos, oscillations and strange attractors. Such phenomena are not considered here.)

A pervasive natural phenomenon is the growth of order from disorder which occurs in a variety of systems. As a result, an interdisciplinary area rich in problems involving the formation and evolution of spatial structures has developed, which combines non-equilibrium dynamics and nonlinear analysis. An important class of such problems deals with the kinetics of phase ordering and phase separation, which are characteristics of any first-order phase transition. Examples of such growth processes occur in many diverse systems, such as chemically reacting systems, biological structures, simple and binary fluids, crystals, polymer melts and metallic alloys. It is interesting that such a variety of systems, which display growth processes, have common characteristics. In the remainder of chapter A3.3 we focus our attention on such common features of kinetics, and on the models which attempt to explain them. Substantial progress has occurred over the past few decades in our understanding of the kinetics of domain growth during first-order phase transitions.

Consider an example of phase separation. It is typically initiated by a rapid change (quench) in a thermodynamic variable (often temperature, and sometimes pressure) which places a disordered system in a post-quench initial non-equilibrium state. The system then evolves towards an inhomogeneous ordered state of coexisting phases, which is its final equilibrium state. Depending on the nature of the quench, the system can be placed in a post-quench state which is thermodynamically unstable or metastable (see [figure A3.3.2](#)). In the former case, the onset of separation is spontaneous, and the kinetics that follows is known as *spinodal decomposition*. For the metastable case, the nonlinear fluctuations are required to initiate the separation process; the system is said to undergo phase separation through *homogeneous nucleation* if the system is pure and through *heterogeneous nucleation* if system has impurities or surfaces which help initiate nucleation. The phase transformation kinetics of supercooled substances via homogeneous nucleation is a fundamental topic. It is also important in science and technology: gases can be compressed way beyond their equilibrium pressures without forming liquids, and liquids can be supercooled several decades below their freezing temperature without crystallizing.

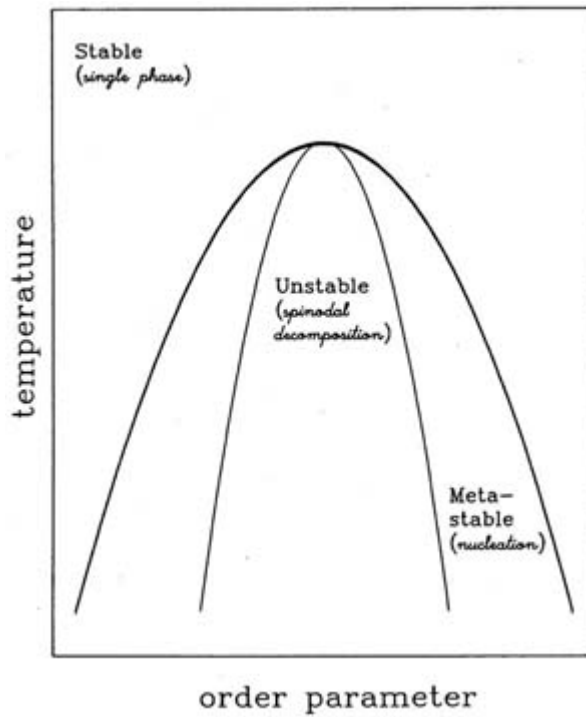


Figure A3.3.2 A schematic phase diagram for a typical binary mixture showing stable, unstable and metastable regions according to a van der Waals mean field description. The coexistence curve (outer curve) and the spinodal curve (inner curve) meet at the (upper) critical point. A critical quench corresponds to a sudden decrease in temperature along a constant order parameter (concentration) path passing through the critical point. Other constant order parameter paths ending within the coexistence curve are called off-critical quenches.

In both cases the late stages of kinetics show power law domain growth, the nature of which does not depend on the initial state; it depends on the nature of the fluctuating variable(s) which is (are) driving the phase separation process. Such a fluctuating variable is called the order parameter; for a binary mixture, the order parameter $\phi(\vec{r}, t)$ is the relative concentration of one of the two species and its fluctuation around the mean value is $\delta c(\vec{r}, t) = c(\vec{r}, t) - c_0$. In the disordered phase, the system's concentration is homogeneous and the order parameter fluctuations are microscopic. In the ordered phase, the inhomogeneity created by two coexisting phases leads to a macroscopic spatial variation in the order parameter field near the interfacial region. In a magnetic system, the average magnetization characterises the para-ferro magnetic transition and is the order parameter. Depending on the system and the nature of the phase transition, the order parameter may be scalar, vector or complex, and may be conserved or non-conserved.

Here we shall consider two simple cases: one in which the order parameter is a non-conserved scalar variable and another in which it is a conserved scalar variable. The latter is exemplified by the binary mixture phase separation, and is treated here at much greater length. The former occurs in a variety of examples, including some order-disorder transitions and antiferromagnets. The example of the para-ferro transition is one in which the magnetization is a conserved quantity in the absence of an external magnetic field, but becomes non-conserved in its presence.

For a one-component fluid, the vapour-liquid transition is characterized by density fluctuations; here the order parameter, mass density ρ , is also conserved. The equilibrium structure factor $S(\vec{k})$ of a one component fluid is

discussed in [section A2.2.5.2](#) and is the Fourier transform of the density–density correlation function. For each of the examples above one can construct the analogous order parameter correlation function. Its spatial Fourier transform (often also denoted by $S(\vec{k})$) is, in most instances, measurable through an appropriate elastic scattering experiment. In a quench experiment which monitors the kinetics of phase transition, the relevant structure evolves in time. That is, the *equal-time* correlation function of the order parameter fluctuations $\langle \delta\phi(\vec{r}, t)\delta\phi(0, t) \rangle_{\text{noneq}}$, which would be time independent in equilibrium, acquires time dependence associated with the growth of order in the non-equilibrium system. Its spatial Fourier transform, $S(\vec{k}, t)$ is called the time-dependent structure factor and is experimentally measured.

The evolution of the system following the quench contains different stages. The early stage involves the emergence of macroscopic domains from the initial post-quench state, and is characterized by the formation of interfaces (domain walls) separating regions of space where the system approaches one of its final coexisting states (domains). Late stages are dominated by the motion of these interfaces as the system acts to minimize its surface free energy. During this stage the mean size of the domains grows with time while the total amount of interface decreases. Substantial progress in the understanding of late stage domain growth kinetics has been inspired by the discovery of *dynamical scaling*, which arises when a single length dominates the time evolution. Then various measures of the morphology depend on time only through this length (an instantaneous snapshot of the order parameter’s space dependence is referred to as the system’s morphology at that time). The evolution of the system then acquires self-similarity in the sense that the spatial patterns formed by the domains at two different times are statistically identical apart from a global change of the length scale.

The time-dependent structure factor $S(\vec{k}, t)$, which is proportional to the intensity $I(k, t)$ measured in an elastic scattering experiment, is a measure of the strength of the spatial correlations in the ordering system with wavenumber k at time t . It exhibits a peak whose position is inversely proportional to the average domain size. As the system phase separates (orders) the peak moves towards increasingly smaller wavenumbers (see [figure A3.3.3](#)).

-20-

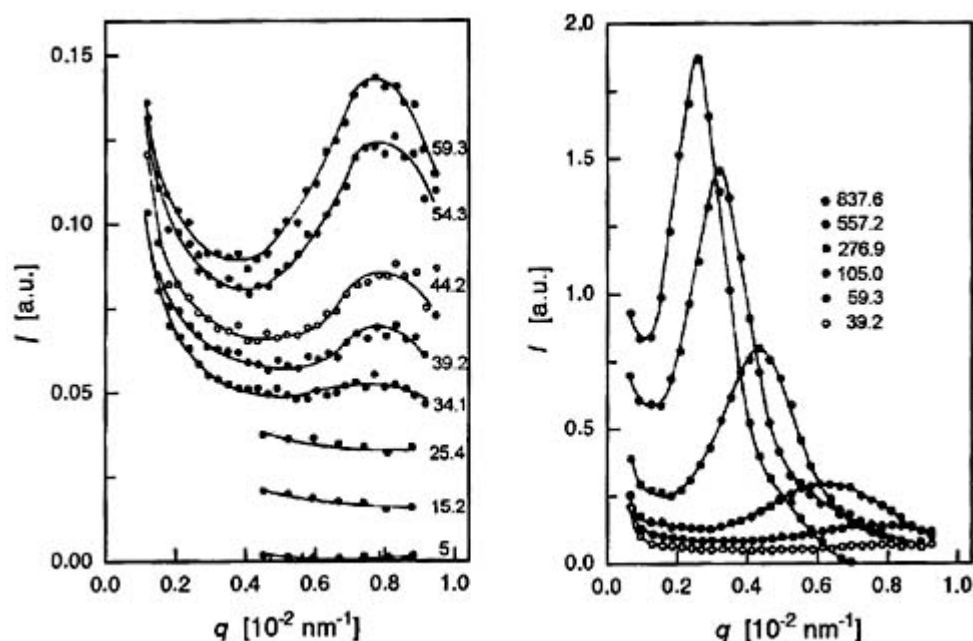


Figure A3.3.3 Time-dependent structure factor as measured through light scattering experiments from a phase separating mixture of polystyrene (PS) ($M = 1.5 \times 10^5$) and poly(vinylmethylether) (PVME) ($M = 4.6 \times 10^4$) following a fast quench from a homogeneous state to $T = 101$ °C located in the two-phase region. The time in

seconds following the quench is indicated for each structure factor curve. Taken from [11].

A signature of the dynamical scaling is evidenced by the collapse of the experimental data to a scaled form, for a d -dimensional system:

$$S(k, t) = (R(t))^d S_0(kR(t)) \tag{A3.3.42}$$

where S_0 is a time-independent function and $R(t)$ is a characteristic length (such as the average domain size) (see figure A3.3.4). To the extent that other lengths in the system, such as the interfacial width, play important roles in the kinetics, the dynamical scaling may be valid only asymptotically at very late times.

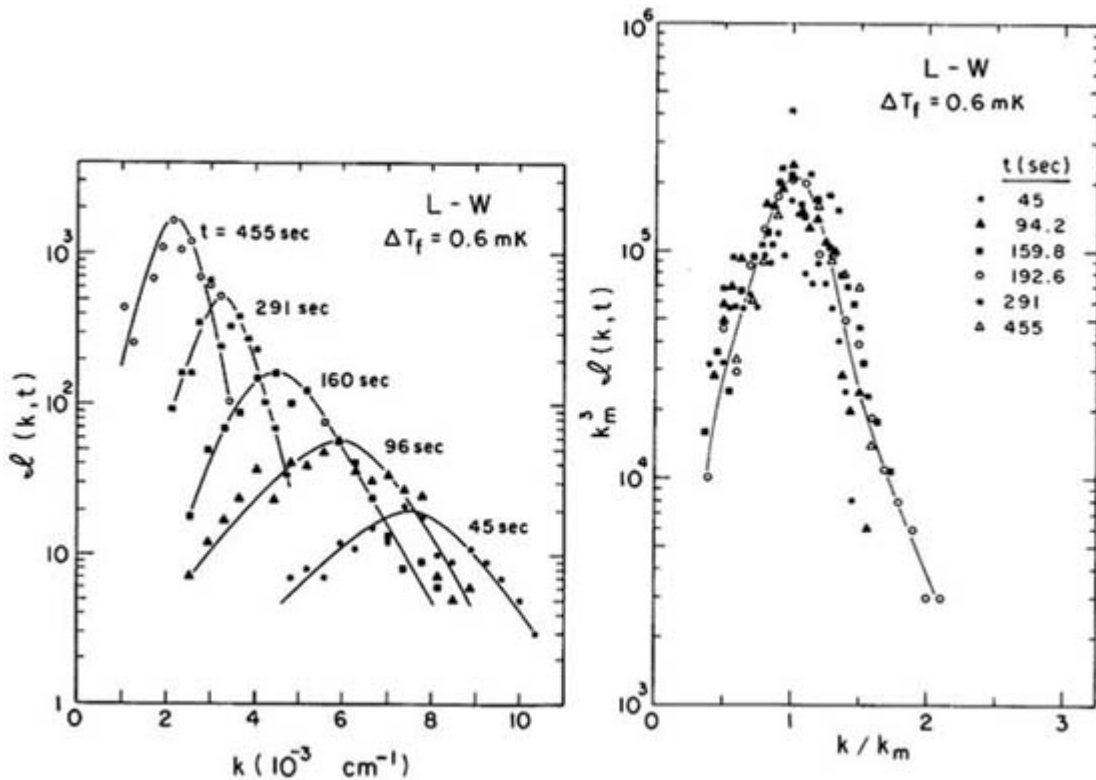


Figure A3.3.4 Time-dependent structure factor as measured through light scattering experiments from a phase separating mixture of 2,6-lutidine and water, following a fast quench from a homogeneous state through the critical point to a temperature 0.6 mK below the critical temperature. The time in seconds, following the quench is indicated for each structure factor curve. In the figure on the right-hand side the data collapse indicates dynamic scaling. Taken from [12].

Another important characteristic of the late stages of phase separation kinetics, for asymmetric mixtures, is the cluster size distribution function of the minority phase clusters: $n(R, \tau) dR$ is the number of clusters of minority phase per unit volume with radii between R and $R + dR$. Its zeroth moment gives the mean number of clusters at time τ and the first moment is proportional to the mean cluster size.

A3.3.3.1 LANGEVIN MODELS FOR PHASE TRANSITION KINETICS

Considerable amount of research effort has been devoted, especially over the last three decades, on various issues in domain growth and dynamical scaling. See the reviews [13, 14, 15, 16 and 17].

Although in principle the microscopic Hamiltonian contains the information necessary to describe the phase separation kinetics, in practice the large number of degrees of freedom in the system makes it necessary to construct a reduced description. Generally, a subset of slowly varying macrovariables, such as the hydrodynamic modes, is a useful starting point. The equation of motion of the macrovariables can, in principle, be derived from the microscopic

Hamiltonian, but in practice one often begins with a phenomenological set of equations. The set of macrovariables are chosen to include the order parameter and all other slow variables to which it couples. Such slow variables are typically obtained from the consideration of the conservation laws and broken symmetries of the system. The remaining degrees of freedom are assumed to vary on a much faster timescale and enter the phenomenological description as random thermal noise. The resulting coupled nonlinear stochastic differential equations for such a chosen ‘relevant’ set of macrovariables are collectively referred to as the Langevin field theory description.

In two of the simplest Langevin models, the order parameter ϕ is the only relevant macrovariable; in model A it is non-conserved and in model B it is conserved. (The labels A, B, etc have historical origin from the Langevin models of critical dynamics; the scheme is often referred to as the Hohenberg–Halperin classification scheme.) For model A, the Langevin description assumes that, on average, the time rate of change of the order parameter is proportional to (the negative of) the thermodynamic force that drives the phase transition. For this single variable case, the thermodynamic force is canonically conjugate to the order parameter: i.e. in a thermodynamic description, if ϕ is a state variable, then its canonically conjugate force is $\partial f/\partial\phi$ (see figure A3.3.5), where f is the free energy.

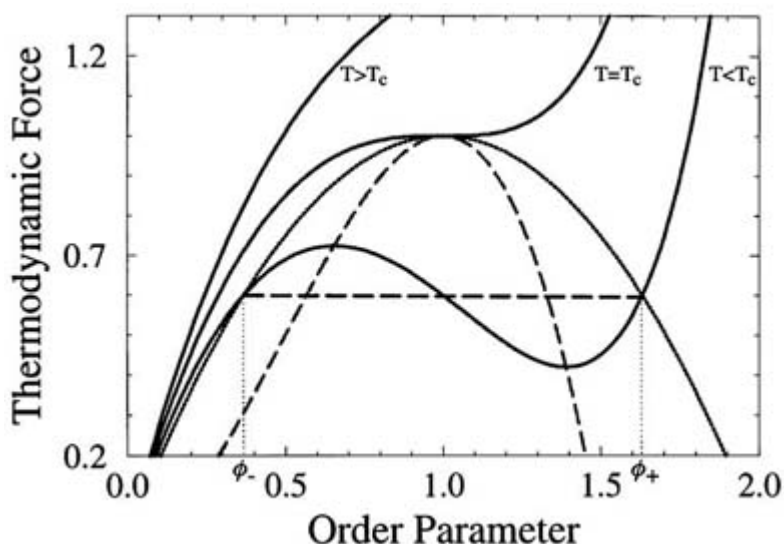


Figure A3.3.5 Thermodynamic force as a function of the order parameter. Three equilibrium isotherms (full curves) are shown according to a mean field description. For $T < T_c$, the isotherm has a van der Waals loop, from which the use of the Maxwell equal area construction leads to the horizontal dashed line for the equilibrium isotherm. Associated coexistence curve (dotted curve) and spinodal curve (dashed line) are also shown. The spinodal curve is the locus of extrema of the various van der Waals loops for $T < T_c$. The states within the spinodal curve are thermodynamically unstable, and those between the spinodal and coexistence

curves are metastable according to the mean field description.

In a field theory description, the thermodynamic free energy f is generalized to a free energy functional $\mathcal{F}[\phi(\vec{r}, t)]$, leading to the thermodynamic force as the analogous functional derivative. The Langevin equation for model A is then

$$\frac{\partial \phi}{\partial t} = -M \frac{\delta \mathcal{F}}{\delta \phi} + \eta(\vec{r}, t) \quad (\text{A3.3.43})$$

-23-

where the proportionality coefficient M is the mobility coefficient, which is related to the random thermal noise η through the fluctuation dissipation theorem:

$$\langle \eta(\vec{r}, t) \eta(\vec{r}', t') \rangle = kT M \delta(\vec{r} - \vec{r}') \delta(t - t'). \quad (\text{A3.3.44})$$

The phenomenology of model B, where ϕ is conserved, can also be outlined simply. Since ϕ is conserved, it obeys a conservation law (continuity equation):

$$\frac{\partial \phi}{\partial t} = -\vec{\nabla} \cdot \vec{j}(\vec{r}, t) \quad (\text{A3.3.45})$$

where (provided \vec{j} itself is not a conserved variable) one can write the transport law

$$\vec{j}(\vec{r}, t) = -[M \vec{\nabla} \mu(\vec{r}, t) + \vec{\zeta}^*] \quad (\text{A3.3.46})$$

with $\vec{\zeta}^*$ being the order parameter current arising from thermal noise, and $\mu(\vec{r}, t)$, which is the local chemical potential, being synonymous with the thermodynamic force discussed above. It is related to the free energy functional as

$$\mu(\vec{r}, t) = \frac{\delta \mathcal{F}}{\delta \phi} \quad (\text{A3.3.47})$$

Putting it all together, one has the Langevin equation for model B:

$$\frac{\partial \phi}{\partial t} = +M \nabla^2 \left(\frac{\delta \mathcal{F}}{\delta \phi} \right) + \zeta \quad (\text{A3.3.48})$$

where $\zeta = \vec{\nabla} \cdot \vec{\zeta}^*$ is the random thermal noise which satisfies the fluctuation dissipation theorem:

$$\langle \zeta(\vec{r}, t) \zeta(\vec{r}', t') \rangle = -2kT M \nabla^2 \delta(\vec{r} - \vec{r}') \delta(t - t'). \quad (\text{A3.3.49})$$

As is evident, the free energy functional \mathcal{F} plays a crucial role in the model A/B kinetics. It contains a number of terms. One of these is the local free energy term $f(\phi)$ which can be thought of as a straightforward generalization of the thermodynamic free energy function in which the global thermodynamic variable ϕ is

replaced by its local field value $\phi(\vec{r}, t)$. Many universal features of kinetics are insensitive to the detailed shape of $f(\phi)$. Following Landau, one often uses for it a form obtained by expanding around the value of ϕ at the critical point, ϕ_c . If the mean value of ϕ is $\bar{\phi}$, then

$$\delta\phi \equiv (\phi - \bar{\phi}) = (\phi - \phi_c) + (\phi_c - \bar{\phi}) \equiv \phi^* + \phi_0 \quad (\text{A3.3.50})$$

-24-

with $\phi^* = (\phi - \phi_c)$ and $\phi_0 = (\phi_c - \bar{\phi})$. The Landau expansion is written in terms of ϕ^* as

$$f(\phi) = \frac{1}{2}a_0(T - T_c)\phi^{*2} + \frac{1}{4}u\phi^{*4} - \mathcal{H}\phi \quad (\text{A3.3.51})$$

where an external field \mathcal{H} is assumed to couple linearly to ϕ . In the absence of \mathcal{H} , $f(\phi)$ has a single minimum for temperatures above T_c at $\phi^* = 0$, and two minima below T_c at $\phi^* = \pm[a_0(T_c - T)/u]^{1/2} \equiv \pm\phi_{\min}^*$ corresponding to the two coexisting ordered phases in equilibrium (see figure A3.3.6).

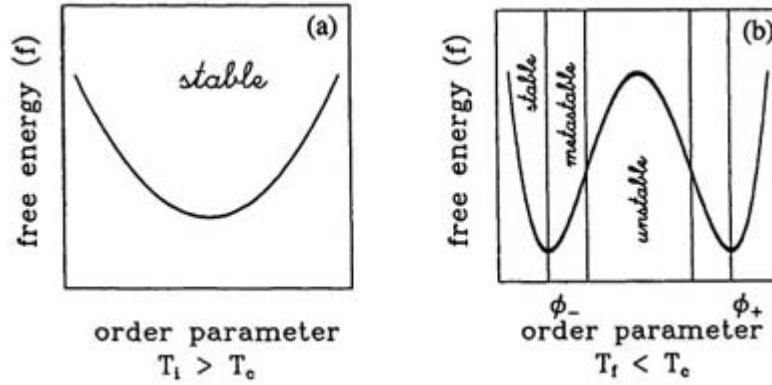


Figure A3.3.6 Free energy as a function of the order parameter ϕ^* for the homogeneous single phase (a) and for the two-phase regions (b), $\mathcal{H}=0$.

The free energy functional also contains a square gradient term which is the cost of the inhomogeneity of the order parameter at each point in the system. Such a surface energy cost term occurs in many different contexts; it was made explicit for binary mixtures by Cahn and Hilliard [18], for superconductors by Ginzburg, and is now commonplace. It is often referred to as the Ginzburg term. This Landau–Ginzburg free energy functional is

$$\mathcal{F}[\phi(\vec{r}, t)] = \int d^d r \left[f(\phi) + \frac{\kappa}{2}(\nabla\phi)^2 \right] \quad (\text{A3.3.52})$$

where the coefficient κ of the square gradient term is related to the interfacial (domain wall) tension σ through the mean field expression

$$\sigma = \kappa \int_{-\infty}^{\infty} dz \left(\frac{d\phi}{dz} \right)^2 \quad (\text{A3.3.53})$$

where z is the direction of the local normal to the interface, and $\phi(z)$ is the equilibrium order parameter profile.

-25-

There is a class of systems where, in addition to the above two terms, there is a non-local free energy functional term in \mathcal{F} . This can arise due to elastic fields, or due to other long-range repulsive interactions (LRRI) originating from the coherent action of molecular dipoles (electric or magnetic). The square gradient (Ginzburg) term is a short-range attractive term which then competes with the LRRI, resulting in a rich variety of structures often termed supercrystals. For such systems, which include Langmuir monolayers and uniaxial magnetic garnet films, the kinetics is much richer. It will not be considered here. See [19, 20] for reviews.

With the form of free energy functional prescribed in equation (A3.3.52), equation (A3.3.43) and equation (A3.3.48) respectively define the problem of kinetics in models A and B. The Langevin equation for model A is also referred to as the time-dependent Ginzburg–Landau equation (if the noise term is ignored); the model B equation is often referred to as the Cahn–Hilliard–Cook equation, and as the Cahn–Hilliard equation in the absence of the noise term.

For deep quenches, where the post-quench T is far below T_c , the equations are conveniently written in terms of scaled (dimensionless) variables: $\psi(\vec{x}, \tau) = \phi^*/\phi_{\min}^*$ and $\vec{x} = \vec{r}/\xi$, where the correlation length $\xi = (\kappa/(a_0|T_c - T|))^{1/2}$, and the dimensionless time τ is defined as equal to $[(2Ma_0|T_c - T)t]$ for model A and equal to $[(2Ma_0^2(T_c - T)^2/\kappa)t]$ for model B. In terms of these variables the model B Langevin equation can be written as

$$\frac{\partial \psi}{\partial \tau} = -\frac{1}{2} \nabla_x^2 (\nabla_x^2 \psi + \psi - \psi^3) + \epsilon^{1/2} \mu(\vec{x}, \tau) \quad (\text{A3.3.54})$$

where

$$\epsilon = \frac{kTu}{a_0^2(T_c - T)^2} \left(\frac{a_0|T_c - T|}{\kappa} \right)^{d/2} \quad (\text{A3.3.55})$$

is the strength of the random thermal noise μ which satisfies

$$\langle \mu(\vec{x}, \tau) \mu(\vec{x}', \tau') \rangle = -\nabla_x^2 \delta(\vec{x} - \vec{x}') \delta(\tau - \tau'). \quad (\text{A3.3.56})$$

Similarly, the dimensionless model A Langevin equation can also be obtained. The result is recovered by replacing the outermost ∇_x^2 by (-1) in equation (A3.3.54) and by $(-\frac{1}{2})$ in equation (A3.3.56).

Using the renormalization group techniques, it has been shown, by Bray [16], that the thermal noise is irrelevant in the deep-quench kinetics. This is because the free energy has two stable fixed points to which the system can flow: for $T > T_c$ it is the infinite temperature fixed point, and for $T < T_c$ it is the zero-temperature strong coupling fixed point. Since at $T = 0$, the strength of the noise ϵ vanishes, the thermal noise term μ can be neglected in model B phase separation kinetics during which $T < T_c$. The same conclusion was also obtained, earlier, [21] from a numerical simulation of equation (A3.3.54) and equation (A3.3.56). In what follows, we ignore the thermal noise term. One must note, however, that there are many examples of kinetics

where the thermal noise can play an important role. See for example a recent monograph [22].

-26-

For critical quench experiments there is a symmetry $\phi_0 = 0$ and from equation (A3.3.50) $\delta\phi = \phi^*$, leading to a symmetric local free energy (figure A3.3.6) and a scaled order parameter whose average is zero, $\delta\psi = \psi$. For off-critical quenches this symmetry is lost. One has $\delta\phi = \phi^* + \phi_0$ which scales to $\delta\psi = \psi + \psi_0$ with $\psi_0 = \phi_0/\phi_{\min}^*$. ψ_0 is a measure of how far off-critical the system is. For $\psi_0 = \pm 1$ the system will be quenched to the coexistence curve and $\psi_0 = 0$ corresponds to a quench through the critical point. In general, one has to interpret the dimensionless order parameter ψ in (A3.3.54) as a mean value plus the fluctuations. If one replaces ψ in (A3.3.54) by $\psi + \psi_0$, the mean value of the order parameter ψ_0 becomes explicit and the average of such a replaced ψ becomes zero, so that now ψ is the order parameter fluctuation. The conservation law dictates that the average value of the order parameter remains equal to ψ_0 throughout the time evolution. Since the final equilibrium phase corresponds to $\psi_{\pm} = \pm 1$, non-zero ψ_0 reflects an asymmetry in the spatial extent of these two phases. The degree of asymmetry is given by the lever rule. A substitution of ψ by $\psi + \psi_0$ in (A3.3.54) yields the following nonlinear partial differential equation (we ignore the noise term):

$$\frac{\partial\psi}{\partial\tau} = -\frac{1}{2}\nabla_x^2([q_c^2 + \nabla_x^2]\psi - 3\psi_0\psi^2 - \psi^3) \quad (\text{A3.3.57})$$

where $q_c^2 = (1 - 3\psi_0^2)$. For a critical quench, when $\psi_0 = 0$, the bilinear term vanishes and q_c^2 becomes one, so that the equation reduces to the symmetric equation (A3.3.54). In terms of the scaled variables, it can be shown that the equation of the classical spinodal, shown in figure A3.3.2 and figure A3.3.5 is $q_c^2 = 0$ or $|\psi_0| = 1/\sqrt{3}$. For states within the classical mean field spinodal, $q_c^2 > 0$.

Equation (A3.3.57) must be supplied with appropriate initial conditions describing the system prior to the onset of phase separation. The initial post-quench state is characterized by the order parameter fluctuations characteristic of the pre-quench initial temperature T_i . The role of these fluctuations has been described in detail in [23]. However, again using the renormalization group arguments, any initial short-range correlations should be irrelevant, and one can take the initial conditions to represent a completely disordered state at $T = \infty$. For example, one can choose the white noise form $\langle\psi(\vec{x}, 0)\psi(\vec{x}', 0)\rangle = \varepsilon_0\delta(\vec{x} - \vec{x}')$, where $\langle\cdots\rangle$ represents an average over an ensemble of initial conditions, and ε_0 controls the size of the initial fluctuations in ψ ; $\varepsilon_0 \ll 1$.

The fundamental problem of understanding phase separation kinetics is then posed as finding the nature of late-time solutions of deterministic equations such as (A3.3.57) subject to random initial conditions.

A linear stability analysis of (A3.3.57) can provide some insight into the structure of solutions to model B. The linear approximation to (A3.3.57) can be easily solved by taking a spatial Fourier transform. The result for the k th Fourier mode is

$$\psi(\vec{k}, \tau) = e^{\gamma_k\tau}\psi(\vec{k}, 0) \quad (\text{A3.3.58})$$

where the exponential growth exponent γ_k is given by

$$\gamma_k = \frac{1}{2}k^2(q_c^2 - k^2). \quad (\text{A3.3.59})$$

For $0 < k < q_c$, γ_k is positive, and the corresponding Fourier mode fluctuations grow in time, i.e. these are the *linearly* unstable modes of the system. The maximally unstable mode occurs at $k_m = q_c/\sqrt{2}$ and overwhelms all other growing modes due to exponential growth in the linear approximation. The structure factor can also be computed analytically in this linear approximation, and has a time invariant maximum at $k = k_m$. In binary polymer mixtures (polymer melts), the early time experimental observations can be fitted to a structure factor form obtained from a linear theory on account of its slow dynamics.

The limitations and range of validity of the linear theory have been discussed in [17, 23, 24]. The linear approximation to [equation \(A3.3.54\)](#) and [equation \(A3.3.57\)](#) assumes that the nonlinear terms are small compared to the linear terms. As $\psi(\vec{k}, \tau)$ increases with time, at some crossover time t_{cr} the linear approximation becomes invalid. This occurs roughly when $\langle \psi^2 \rangle$ becomes comparable to $(\psi_{sp} - \psi_o)^2 = \psi_{sp}^2 q_c^2$. One can obtain t_{cr} using [equation \(A3.3.58\)](#), in which k can be replaced by k_m , since the maximally unstable mode grows exponentially faster than other modes. Then the dimensionless crossover time $\tau_{cr} \equiv t_{cr}(2M\kappa/\xi_o^4)$ is obtained from

$$(\psi_{sp} - \psi_o)^2 = \langle |\psi(\vec{k}_m, \tau_{cr})|^2 \rangle = e^{2\gamma_{k_m} \tau_{cr}} \langle |\psi(\vec{k}_m, 0)|^2 \rangle$$

where the initial fluctuation spectrum is to be determined from the Ornstein–Zernicke theory, at the pre-quench temperature T_o :

$$\langle |\psi(\vec{k}, 0)|^2 \rangle = \frac{\epsilon_o}{(k_m^2 + q_c^2)}.$$

Here ϵ_o is given by [equation \(A3.3.55\)](#) evaluated at T_o , and can be written as $\epsilon_o = kT_o u \kappa^{-2} \xi_o^{4-d/2}$. Using the values $k_m^2 = q_c^2/2$, $\psi_{sp}^2 = 1/3$, and $\kappa = \kappa/[a_o(T_o - T_c)]$, one obtains

$$2\gamma_{k_m} \tau_{cr} = \frac{d}{4} \ln(\kappa) + \ln(q_c^4) + \ln\left(\frac{[a_o(T_o - T_c)]^{(2-d/4)}}{2kT_o u}\right).$$

As is evident from the form of the square gradient term in the free energy functional, [equation \(A3.3.52\)](#), κ is like the square of the effective range of interaction. Thus, the dimensionless crossover time depends only weakly on the range of interaction as $\ln(\kappa)$. For polymer chains of length N , $\kappa \sim N$. Thus for practical purposes, the dimensionless crossover time τ_{cr} is not very different for polymeric systems as compared to the small molecule case. On the other hand, the scaling of t_{cr} to τ_{cr} is through a characteristic time which itself increases linearly with κ , and one has

$$t_{cr} = \frac{2\kappa[(d/4) \ln(\kappa) + \ln(q_c^4) + \ln([a_o(T_o - T_c)]^{(2-d/4)})/2kT_o u]}{q_c^4 M a_o^2 (T_o - T_c)^2}$$

which behaves like $\kappa \ln(\kappa) \sim N \ln(N)$ for polymeric systems. It is clear that the longer time for the validity of linear theory for polymer systems is essentially a longer characteristic time phenomenon.

For initial post-quench states in the metastable region between the classical spinodal and coexistence curves, q_c^2

is negative and so is γ_k for all values of k . Linear stability analysis is not adequate for the metastable region, since it predicts that all modes are stable. Nonlinear terms are important and cannot be ignored in the kinetics leading to either nucleation or spinodal decomposition. The transition from spinodal decomposition to nucleation is also not well defined because nonlinear instabilities play an increasingly more important role as the ‘classical spinodal’ is approached from within.

A3.3.3.2 UNSTABLE STATES AND KINETICS OF SPINODAL DECOMPOSITION

Equation (A3.3.57) is an interesting nonlinear partial differential equation, but it is mathematically intractable. It contains quadratic and cubic nonlinear terms. The cubic term treats both phases in a symmetric manner; for a symmetric binary mixture only this term survives and leads to a labyrinthian morphology in which both phases have an equal share of the system volume. This term is the source of spinodal decomposition. For the symmetric case the partial differential equation is parameter free and there is no convenient small expansion parameter, especially during early times (the linear approximation loses its validity around $\tau \sim 10$). At late times, the ratio of the interfacial width to the time-dependent domain size $\xi/R(\tau)$ was used as a small parameter by Pego [25] in a matched asymptotic expansion method. It leads to useful connections of this nonlinear problem to the Mullins–Sekerka instability for the slowest timescale and to the classic Stefan problem on a faster timescale. The quadratic term treats the two phases in an asymmetric manner and is the source of nucleation-like morphology. As the off criticality ψ_0 increases, the quadratic term gradually assumes a greater role compared to the cubic nonlinear term. Nucleation-like features in the kinetics occur even for a 49–51 mixture in principle, and are evident at long enough times, since the minority phase will form clusters within the majority background phase for any asymmetric mixture.

While approximate analytical methods have played a role in advancing our understanding of the model B kinetics, complimentary information from laboratory experiments and numerical simulations have also played an important role. Figure A3.3.3 and figure A3.3.4 show the time-dependent structure factors from laboratory experiments on a binary polymer melt and a small molecule binary mixture, respectively. Compared to the conceptual model B Langevin equation discussed above, real binary mixtures have additional physical effects: for a binary polymer melt, hydrodynamic interactions play a role at late times [17]; for a small molecule binary fluid mixture, hydrodynamic flow effects become important at late times [26]; and for a binary alloy, the elastic effects play a subsidiary, but important, role [37]. In each of these systems, however, there is a broad range of times when model B kinetics are applicable. Comparing the approximate theory of model B kinetics with the experimental results from such systems may not be very revealing, since the differences may be due to effects not contained in model B. Comparing an approximate theory to computer simulation results provides a good test for the theory, provided a good estimate of the numerical errors in the simulation can be made.

In the literature there are numerical simulations of equation (A3.3.57) for both two- and three-dimensional systems [21, 23, 28, 29, 30 and 31]. For a two-dimensional system, morphology snapshots of the order parameter field ψ are shown in figure A3.3.7 for late times, as obtained from the numerical simulations of (A3.3.57). The light regions correspond to positive values of ψ and the dark regions to negative values. For the critical quench case (figure A3.3.7(a) and (b)), the (statistical) symmetry of ψ between the two phases is apparent. The topological difference between the critical and off-critical quench evolutions at late times is also clear: bicontinuous for critical quench and isolated closed cluster topology for asymmetric off-critical quench. Domain coarsening is also evident from these snapshots for each of the two topologies. For the off-critical quench, from such snapshots one can obtain the time evolution of the cluster size distribution.

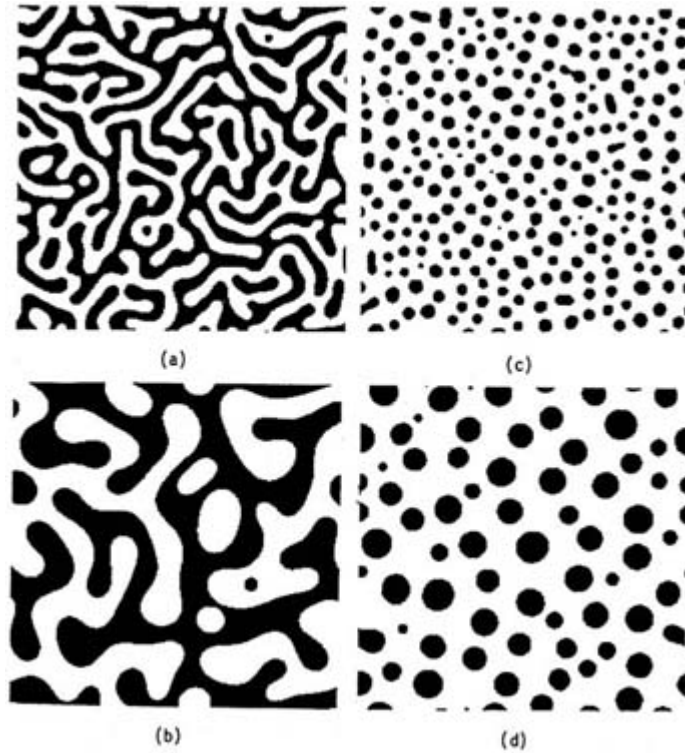


Figure A3.3.7 The order parameter field morphology, for $\psi_0 = 0.0$ at (a) $\tau = 500$ and (b) $\tau = 5000$; and for $\psi_0 = 0.4$ at (c) $\tau = 500$ and (d) $\tau = 5000$. The dark regions have $\psi < 0$. From [28].

From a snapshot at time τ , the spatial correlation function $G(x, \tau) \equiv \langle \psi(\vec{x}, \tau) \psi(0, \tau) \rangle$ can be computed, where $\langle \dots \rangle$ includes the angular average assuming the system to be spatially isotropic. Repeating this for various snapshots yields the full space-and-time-dependent correlation function $G(x, \tau)$. Its spatial Fourier transform is essentially the time-dependent structure factor $S(k, t)$ measured in light scattering experiments (see [figure A3.3.3](#) and [figure A3.3.4](#)). There are a number of ways to obtain the time-dependent domain size, $R(\tau)$: (i) first zero of $G(x, \tau)$, (ii) first moment of $S(k, t)$, (iii) value k_m where $S(k, t)$ is a maximum. The result that is now firmly established from experiments and simulations, is that

$$R(\tau) \sim \tau^{1/3} \quad (\text{A3.3.60})$$

independent of the system dimensionality d . In the next section ([section A3.3.4](#)) we describe the classic theory of Lifshitz, Slyozov and Wagner, which is one of the cornerstone for understanding the $\tau^{1/3}$ growth law and asymptotic cluster size distribution for quenches to the coexistence curve.

As in the experiments, the simulation results also show dynamic scaling at late times. The scaling function $S_0(kR(\tau))$ at late times has the large k behaviour $S_0(y) \sim y^{-(d+1)}$ known as Porod's law [13, 16]. This result is understood to be the consequence of the sharp interfaces at late times. The small k behaviour, $S_0(y) \sim y^4$ was independently predicted in [32, 33], and was put on a firm basis in [34].

Interfaces play a central role in phase transition kinetics of both models A and B. [Figure A3.3.8](#) shows the interfacial structure corresponding to [Figure A3.3.7](#) (b). One can see the relationship between the interfacial width and the domain size for a late-stage configuration. The upper part of the figure demarks the interfacial

regions of the system where $0.75\psi_- > \psi > 0.75\psi_+$. The lower plot gives a cross sectional variation of ψ as the system is traversed. The steep gradients in ψ in the lower plot clearly indicates the sharpness of interfaces at late times.

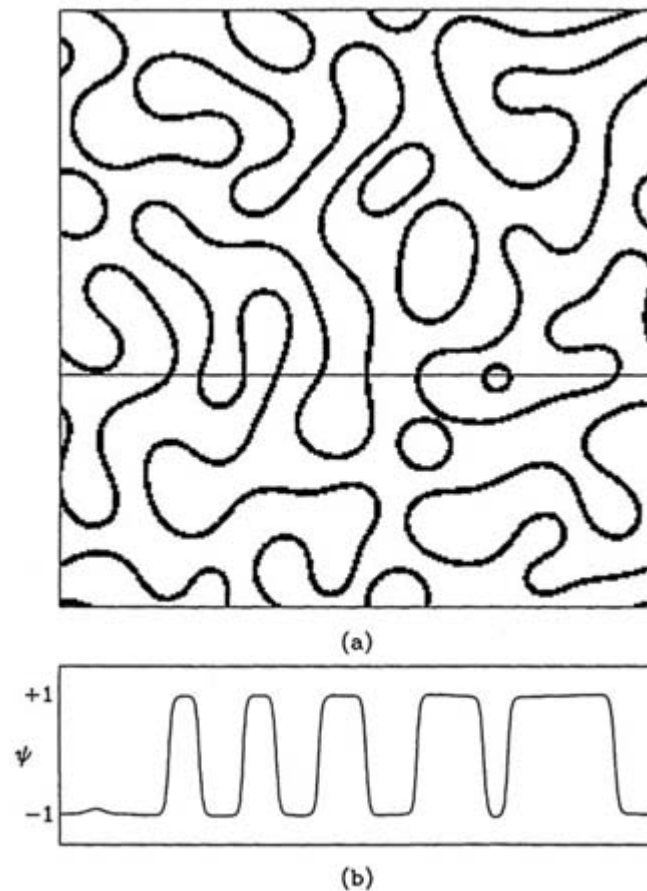


Figure A3.3.8 Interface structure for $\tau = 5000$, $\psi_0 = 0$. In (a) the shaded regions correspond to interfaces separating the domains. In (b) a cross sectional view of the order parameter ψ is given. The location of the cross section is denoted by the horizontal line in (a). From [35].

In [figure A3.3.9](#) the early-time results of the interface formation are shown for $\psi_0 = 0.48$. The classical spinodal corresponds to $\psi_0 \sim 0.58$. Interface motion can be simply monitored by defining the domain boundary as the location where $\psi = 0$. Surface tension smooths the domain boundaries as time increases. Large interconnected clusters begin to break apart into small circular droplets around $\tau = 160$. This is because the quadratic nonlinearity eventually outpaces the cubic one when off-criticality is large, as is the case here.

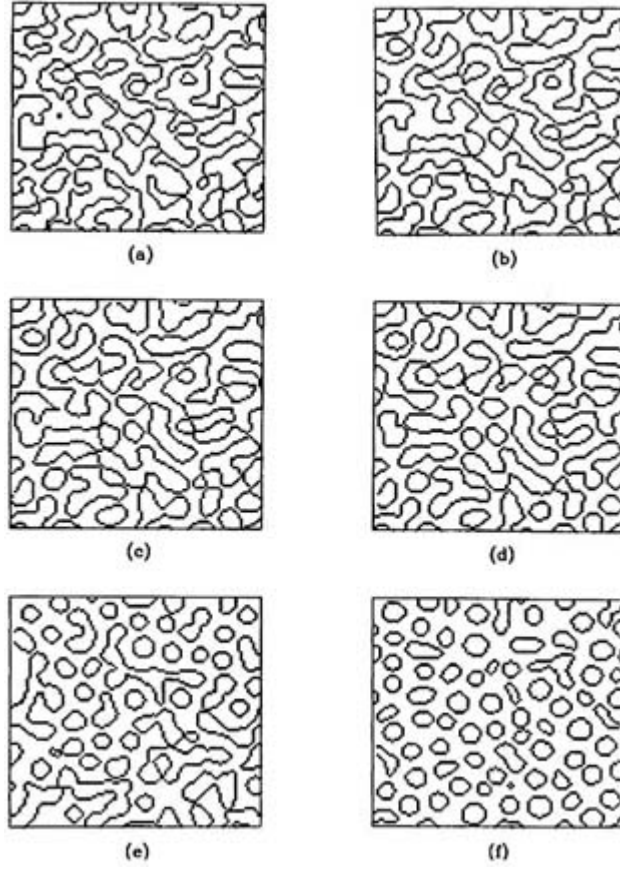


Figure A3.3.9 Time dependence of the domain boundary morphology for $\psi_0 = 0.48$. Here the domain boundary is the location where $\psi = 0$. The evolution is shown for early-time τ values of (a) 50, (b) 100, (c) 150, (d) 200, (e) 250 and (f) 300. From [29].

Some features of late-stage interface dynamics are understood for model B and also for model A. We now proceed to discuss essential aspects of this interface dynamics. Consider the Langevin equations without noise. Equation (A3.3.57) can be written in a more general form:

$$\frac{\partial \psi}{\partial \tau} = -\nabla_x^2 (\xi^2 \nabla_x^2 \psi - f'(\psi)) \quad (\text{A3.3.61})$$

where we have absorbed the factor $\frac{1}{2}$ in the time units of τ , introduced ξ even though it is one, in order to keep track of characteristic lengths, and denoted the thermodynamic force (chemical potential) by f' . At late times the domain size $R(\tau)$ is much bigger than the interfacial width ξ . Locally, therefore, the interface appears to be planar. Let its normal be in direction u , and let $u = u_0$ at the point within the interface where $\psi = 0$. Then $\psi = \psi(u, \vec{s}, \tau)$ where \vec{s} refers to the $(d-1)$ coordinates parallel to the interface at point \vec{x} . In essence, we have used the interface specific coordinates: $\vec{x} = \vec{R}(\vec{s}) + u\hat{n}(s)$, where $\hat{n}(s)$ is a unit normal at the interface, pointing from the ψ_- phase into the ψ_+ phase.

The stationary solution of (A3.3.61), when $\xi/R(\tau)$ is very small, satisfies

$$\xi^2 \frac{d^2 \psi_0}{du^2} = \frac{\partial f}{\partial \psi_0} \quad (\text{A3.3.62})$$

which has a kink profile solution $\psi_0(u) = \pm \tanh((u - u_0)/(\sqrt{2}\xi))$ for the double well free energy $f(\psi_0) = \psi_0^4/4 - \psi_0^2/2$. By linearizing around such a kink solution, one obtains a linear eigenvalue problem. Its lowest eigenmode is $\zeta_0(u) \equiv d\psi_0(u)/du$ and the corresponding eigenvalue is zero. It is localized within the interface and is called the Goldstone mode arising out of the broken translational symmetry at the interface. The higher eigenmodes, which are constructed to be orthogonal to the Goldstone mode, are the capillary wave fluctuation modes with the dispersion relation that is a generalised version [35] of that discussed in [section A3.3.2.4](#). The orthogonality to the Goldstone mode leads to a constraint which is used to show an important relation between the local interface velocity $v(\vec{x})$ and local curvature $K(\vec{s})$. It is, to the lowest order in ξ K ,

$$\sigma \xi^2 K(s) = \int du d\vec{x}' G(\vec{x}, \vec{x}') \zeta_0(u) \zeta_0(u') v(\vec{x}') \quad (\text{A3.3.63})$$

where $G(\vec{x}, \vec{x}')$ is the diffusion Green's function satisfying $\nabla^2 G(\vec{x}, \vec{x}') = \delta(\vec{x} - \vec{x}')$. The mean field surface tension σ , defined in [equation \(A3.3.53\)](#), is the driving force for the interface dynamics. The diffusion Green's function couples the interface motion, at two points (\vec{x}, \vec{x}') on the interface, inextricably to the bulk dynamics. For a conserved order parameter, the interface dynamics and late-stage domain growth involve the *evaporation–diffusion–condensation mechanism* whereby large droplets (small curvature) grow at the expense of small droplets (large curvature). This is also the basis for the Lifshitz–Slyozov analysis which is discussed in [section A3.3.4](#).

If the order parameter is not conserved, the results are much simpler and were discussed by Lifshitz and by Allen and Cahn [36]. For model A, [equation \(A3.3.61\)](#) is to be replaced by the time-dependent Ginzburg–Landau equation which is obtained by removing the overall factor of $(-\nabla_x^2)$ from the right-hand side. This has the consequence that, in the constraint, [equation \(A3.3.63\)](#), the diffusion Green's function is replaced by $\delta(\vec{x} - \vec{x}')$ and the integrals can be performed with the right-hand side reducing to $-\sigma v(s)$. The surface tension then cancels from both sides and one gets the Allen–Cahn result:

$$-\xi^2 K(s) = v(s). \quad (\text{A3.3.64})$$

For model A, the interfaces decouple from the bulk dynamics and their motion is driven entirely by the local curvature, and the surface tension plays only a background, but still an important, role. From this model A interface dynamics result, one can also simply deduce that the domains grow as $R(\tau) \sim \tau^{1/2}$: at some late time, a spherical cluster of radius R grows; since $K \sim (d-1)/R$ and $v \sim -dR/d\tau$, one has $R^2 \sim (d-1)\tau$.

A3.3.4 LATE-STAGE GROWTH KINETICS AND OSTWALD RIPENING

Late stages of model B dynamics for asymmetric quenches may be described by the Lifshitz–Slyozov–Wagner (LSW) theory of coarsening. When the scalar order parameter is conserved, the late-stage coarsening is referred to as Ostwald ripening. The LSW analysis is valid for late-stage domain growth following either spinodal decomposition or nucleation. A recent paper [37] has combined the steady-state homogeneous nucleation theory described in the next [section, A3.3.5](#), with the LSW analysis in a new model for the entire process of phase separation. If the initial condition places the post-quench system just inside and quite near

the coexistence curve, the conservation law dictates that one (minority) phase will occupy a much smaller ‘volume’ fraction than the other (majority) phase in the final equilibrium state.

The dynamics is governed by interactions between different domains of the minority phase. At late times these will have attained spherical (circular) shape for a three (two)-dimensional system. For model B systems, the classic work of Lifshitz and Slyozov [38] and the independent work by Wagner [39] form the theoretical cornerstone. The late-stage dynamics is mapped onto a diffusion equation with sources and sinks (i.e. domains) whose boundaries are time dependent. The Lifshitz–Slyozov (LS) treatment of coarsening is based on a mean field treatment of the diffusive interaction between the domains and on the assumption of an infinitely sharp interface with well defined boundary conditions. The analysis predicts the onset of dynamical scaling. As in [section A3.3.3.1](#) we shall denote the extent of the off-criticality by $\psi_0 > 0$. The majority phase equilibrates at $\psi_+ = +1$ and the minority phase at $\psi_- = -1$. At late times, the minority clusters have radius $R(\tau)$ which is much larger than the interface width ξ . An important coupling exists between the interface and the majority phase through the surface tension σ . This coupling is manifested through a Gibbs–Thomson boundary condition, which is given later.

The LS analysis is based on the premise that the clusters of the minority phase compete for growth through an evaporation–condensation mechanism, whereby larger clusters grow at the expense of smaller ones. (Material (of the minority phase) evaporates from a smaller cluster, diffuses through the majority phase background matrix and condenses on a larger cluster.) That is, the dominant growth mechanism is the transport of the order parameter from interfaces of high curvature to regions of low curvature by diffusion through the intervening bulk phases. The basic model B equations, [\(A3.3.48\)](#) and [\(A3.3.52\)](#), can be linearized around the majority phase bulk equilibrium value of the order parameter, $\psi_+ = 1$ (which corresponds to the off-criticality $\psi_0 = -1$), by using $\psi = 1 + \delta\psi$ and keeping only up to first-order terms in $\delta\psi$. The result in dimensionless form is

$$\frac{\partial}{\partial \tau} \delta\psi = -\xi^2 \nabla^4 \delta\psi + \left(\frac{\delta^2 f}{\delta \psi^2} \right)_{\psi=1} \nabla^2 \delta\psi \quad (\text{A3.3.65})$$

where we have kept the interfacial width ξ as a parameter to be thought of as one; we retain it in order to keep track of the length scales in the problem. Since at late times the characteristic length scales are large compared to ξ , the ∇^4 term is negligible and $\delta\psi$ satisfies a diffusion equation,

$$\frac{\partial}{\partial \tau} \delta\psi = f''(1) \nabla^2 \delta\psi. \quad (\text{A3.3.66})$$

-34-

Due to the conservation law, the diffusion field $\delta\psi$ relaxes in a time much shorter than the time taken by significant interface motion. If the domain size is $R(\tau)$, the diffusion field relaxes over a time scale $\tau_D \sim R^2$. However a typical interface velocity is shown below to be $\sim R^{-2}$. Thus in time τ_D , interfaces move a distance of about one, much smaller compared to R . This implies that the diffusion field $\delta\psi$ is essentially always in equilibrium with the interfaces and, thus, obeys Laplace’s equation

$$\nabla^2 \delta\psi = 0 \quad (\text{A3.3.67})$$

in the bulk.

A3.3.4.1 GIBBS–THOMSON BOUNDARY CONDITION

To derive the boundary condition, it is better to work with the chemical potential instead of the diffusion field. We have

$$\frac{\partial \psi}{\partial \tau} = -\nabla \cdot \vec{j} \quad (\text{A3.3.68})$$

$$\vec{j} = -\nabla \mu \quad (\text{A3.3.69})$$

and

$$\mu = f'(\psi) - \xi^2 \nabla^2 \psi. \quad (\text{A3.3.70})$$

In the bulk, linearizing μ leads to $\mu = f''(\psi_*)\delta\psi - \xi^2 \nabla^2 \delta\psi$, where the ∇^2 term is again negligible, so that μ is proportional to $\delta\psi$. Thus μ also obeys Laplace's equation

$$\nabla^2 \mu = 0. \quad (\text{A3.3.71})$$

Let us analyse μ near an interface. The Laplacian in the curvilinear coordinates (u, \vec{s}) can be written such that (A3.3.71) becomes (near the interface)

$$\mu = f'(\psi) - \xi^2 \left(\frac{\partial \psi}{\partial u} \right)_\tau K - \xi^2 \left(\frac{\partial^2 \psi}{\partial u^2} \right)_\tau \quad (\text{A3.3.72})$$

where $K = \vec{\nabla} \cdot \hat{n}$ is the total curvature. The value of μ at the interface can be obtained from (A3.3.72) by multiplying it with $(\partial\psi/\partial u)_\tau$ (which is sharply peaked at the interface) and integrating over u across the interface.

Since μ and K vary smoothly through the interface, one obtains a general result that, at the interface,

$$\mu \Delta \psi = \Delta f - \xi^2 \sigma K \quad (\text{A3.3.73})$$

where $\Delta \psi$ is the change in ψ across the interface and Δf is the difference in the minima of the free energy f for the two bulk phases. For the symmetric double well, $\Delta f = 0$ and $\Delta \psi = 2$. Thus

$$\mu = -\frac{1}{2} \xi^2 \sigma K. \quad (\text{A3.3.74})$$

We make two side remarks.

- (1) If the free energy minima have unequal depths, then this calculation can also be done. See [14].

(2) Far away from the interface, $\mu = f'(\psi_{\pm})\delta\psi = 2\delta\psi$ for the ψ^4 form. Then one also has for the supersaturation

$$\delta\psi(\infty) = - \lim_{u \rightarrow \infty} \delta\psi(u) = -\mu/2 = +\xi^2\sigma K/4. \quad (\text{A3.3.75})$$

The supersaturation $\varepsilon \equiv \delta\psi(\infty)$ is the mean value of $\delta\psi$, which reflects the presence of other subcritical clusters in the system.

Equation (A3.3.73) is referred to as the Gibbs–Thomson boundary condition. equation (A3.3.74) determines μ on the interfaces in terms of the curvature, and between the interfaces μ satisfies Laplace's equation, [equation \(A3.3.71\)](#). Now, since $\vec{j} = -\vec{\nabla}\mu$, an interface moves due to the imbalance between the current flowing into and out of it. The interface velocity is therefore given by

$$j_{\text{out}} - j_{\text{in}} = v\Delta\psi \quad (\text{A3.3.76})$$

and also from [equation \(A3.3.69\)](#),

$$j_{\text{out}} - j_{\text{in}} = - \left[\frac{\partial\mu}{\partial u} \right] = -[\hat{n} \cdot \nabla\mu]. \quad (\text{A3.3.77})$$

Here [...] denotes the discontinuity in ... across the interface. [Equation \(A3.3.71\)](#), [equation \(A3.3.74\)](#), [equation \(A3.3.76\)](#) and [equation \(A3.3.77\)](#) together determine the interface motion.

Consider a single spherical domain of minority phase ($\psi_- = -1$) in an infinite sea of majority phase ($\psi_+ = +1$). From the definition of μ in [\(A3.3.70\)](#), $\mu = 0$ at infinity. Let $R(\tau)$ be the domain radius. The solution of Laplace's equation, [\(A3.3.71\)](#), for $d > 2$, with a boundary condition at ∞ and [equation \(A3.3.74\)](#) at $r = R$, is spherically symmetric and is, using $K = (d - 1)/R$,

$$\mu = - \frac{(d - 1)\sigma\xi^2}{2r} \quad \text{for } r \geq R \quad (\text{A3.3.78})$$

and

$$\mu = - \frac{(d - 1)\sigma\xi^2}{2R} \quad \text{for } r \leq R. \quad (\text{A3.3.79})$$

Then, using [equation \(A3.3.76\)](#) and [equation \(A3.3.77\)](#), we obtain, since $\Delta\psi = 2$,

$$\frac{dR}{d\tau} = v = - \frac{1}{2} \left[\frac{\partial\mu}{\partial r} \right]_{R-\epsilon}^{R+\epsilon} = - \frac{(d - 1)\xi^2\sigma}{4R^2}. \quad (\text{A3.3.80})$$

Integrating equation (A3.3.80), we get (setting $\xi = 1$)

$$R^3(\tau) = R^3(0) - \frac{3}{4}(d-1)\sigma\tau \quad (\text{A3.3.81})$$

which leads to a ' R^3 proportional to τ ' time dependence of the evaporating domain: the domain evaporates in time τ proportional to $R^3(0)$.

A3.3.4.2 LS ANALYSIS FOR GROWING DROPLETS

Again consider a single spherical droplet of minority phase ($\psi_- = -1$) of radius R immersed in a sea of majority phase. But now let the majority phase have an order parameter at infinity that is (slightly) smaller than +1, i.e. $\psi(\infty) \equiv \psi_0 < 1$. The majority phase is now 'supersaturated' with the dissolved minority species, and if the minority droplet is large enough it will grow by absorbing material from the majority phase. Otherwise it will evaporate as above. The two regimes are separated by a critical radius R_c .

Let $f(\pm 1) = 0$ by convention, then the Gibbs–Thomson boundary condition, [equation \(A3.3.73\)](#), becomes at $r = R$,

$$(1 + \psi_0)\mu = f(\psi_0) - \frac{(d-1)\sigma}{R}. \quad (\text{A3.3.82})$$

At $r = \infty$, from [equation \(A3.3.70\)](#),

$$\mu = f'(\psi_0). \quad (\text{A3.3.83})$$

-37-

The solution of Laplace's equation, [\(A3.3.71\)](#), with these boundary conditions is, for $d = 3$,

$$\mu = \begin{cases} f'(\psi_0) + \left(\frac{f(\psi_0)}{1 + \psi_0} - f'(\psi_0) \right) \frac{R}{r} - \frac{2\sigma}{(1 + \psi_0)} \frac{1}{r} & r \geq R \\ \frac{f(\psi_0)}{1 + \psi_0} - \frac{2\sigma}{(1 + \psi_0)} \frac{1}{R} & r \leq R. \end{cases} \quad (\text{A3.3.84})$$

(A3.3.85)

Using [equation \(A3.3.76\)](#), [equation \(A3.3.77\)](#) and [equation \(A3.3.84\)](#), one finds the interface velocity $v \equiv dR/d\tau$ as

$$\frac{dR}{d\tau} = \left(\frac{f(\psi_0)}{(1 + \psi_0)^2} - \frac{f'(\psi_0)}{(1 + \psi_0)} \right) \frac{1}{R} - \frac{2\sigma}{(1 + \psi_0)^2} \frac{1}{R^2}. \quad (\text{A3.3.86})$$

For a small supersaturation, $\psi_0 = 1 - \varepsilon$ with $\varepsilon \ll 1$. To leading (non-trivial) order in ε , [equation \(A3.3.86\)](#) reduces to

$$v(R) \equiv \frac{dR}{d\tau} = \frac{\sigma}{2R} \left(\frac{1}{R_c} - \frac{1}{R} \right) \quad (\text{A3.3.87})$$

with $R_c = \sigma/(f''(1)\epsilon)$ as the critical radius.

The form of $v(R)$ in (A3.3.87) is valid only for $d = 3$. If we write it as

$$\frac{dR}{d\tau} = \frac{\alpha_d}{R} \left(\frac{1}{R_c} - \frac{1}{R} \right) \quad (\text{A3.3.88})$$

then the general expression (see [40]) for α_d is $\alpha_d = (d-1)(d-2)\sigma/4$. For $d = 2$, α_d vanishes due to the singular nature of the Laplacian in two-dimensional systems. For $d = 2$ and in the limit of a small (zero) volume fraction of the minority phase, equation (A3.3.87) is modified to (see the appendix of [28]),

$$\frac{dR}{d\tau} = \frac{\sigma}{4R \ln(4\tau)} \left(\frac{1}{R_c} - \frac{1}{R} \right) \quad (\text{A3.3.89})$$

with $R_c = \sigma/(2f''(1)\epsilon)$. A change of variable $\tau^* = \tau/\ln(4\tau)$ converts (A3.3.89) into the same form as (A3.3.87), but now the time-like variable has a logarithmic modification.

In the LS analysis, an assembly of drops is considered. Growth proceeds by evaporation from drops with $R < R_c$ and condensation onto drops $R > R_c$. The supersaturation ϵ changes in time, so that $\epsilon(\tau)$ becomes a sort of mean field due to all the other droplets and also implies a time-dependent critical radius $R_c(\tau) = \sigma/[f''(1)\epsilon(\tau)]$. One of the starting equations in the LS analysis is equation (A3.3.87) with $R_c(\tau)$.

-38-

For a general dimension d , the cluster size distribution function $n(R, \tau)$ is defined such that $n(R, \tau)dR$ equals the number of clusters per unit 'volume' with a radius between R and $R + dR$. Assuming no nucleation of new clusters and no coalescence, $n(R, \tau)$ satisfies a continuity equation

$$\frac{\partial n}{\partial \tau} + \frac{\partial}{\partial R}(vn) = 0 \quad (\text{A3.3.90})$$

where $v \equiv dR/d\tau$ is given by equation (A3.3.87). Finally, the conservation law is imposed on the entire system as follows. Let the spatial average of the conserved order parameter be $(1 - \epsilon_0)$. At late times the supersaturation $\epsilon(\tau)$ tends to zero giving the constraint

$$\epsilon_0 = \epsilon(\tau) + V_d \int_0^\infty dR R^d n(R, \tau) \sim V_d \int_0^\infty dR R^d n(R, \tau) \quad (\text{A3.3.91})$$

where V_d is the volume of the d -dimensional unit sphere. Equation (A3.3.88), equation (A3.3.90) and equation (A3.3.91) constitute the LS problem for the cluster size distribution function $n(R, \tau)$. The LS analysis of these equations starts by introducing a scaling distribution of droplet sizes. For a d -dimensional system, one writes

$$n(R, \tau) = R_c^{-(d+1)} f\left(\frac{R}{R_c}\right). \quad (\text{A3.3.92})$$

Equation (A3.3.91) becomes, denoting R/R_c by x ,

$$\epsilon_0 = 2V_d \int_0^\infty dx x^d f(x) \quad (\text{A3.3.93})$$

and fixes the normalization of $f(x)$. If equation (A3.3.92) is substituted into equation (A3.3.90) we obtain, using the velocity [equation \(A3.3.88\)](#),

$$\frac{\dot{R}_c}{R_c^{d+2}} \left[(d+1)f(x) + x \frac{df}{dx} \right] = \frac{\alpha_d}{R_c^{d+4}} \left[\left(\frac{2}{x^3} - \frac{1}{x^2} \right) f(x) + \left(\frac{1}{x} - \frac{1}{x^2} \right) \frac{df}{dx} \right]. \quad (\text{A3.3.94})$$

For the consistency of the scaling form, equation (A3.3.92), R_c dependence should drop out from equation (A3.3.94); i.e.

$$R_c^2 \dot{R}_c = \alpha_d \gamma \quad (\text{A3.3.95})$$

-39-

which integrates to

$$R_c(\tau) = (3\alpha_d \gamma \tau)^{1/3}. \quad (\text{A3.3.96})$$

[Equation \(A3.3.94\)](#) simplifies to

$$\left[\frac{2}{x^3} - \frac{1}{x^2} - \gamma(d+1) \right] f(x) = \left[\gamma x - \frac{1}{x} + \frac{1}{x^2} \right] \frac{df}{dx} \quad (\text{A3.3.97})$$

which integrates to

$$\ln f(x) = \int^x \frac{dy}{y} \frac{(2-y-\gamma(d+1)y^3)}{(\gamma y^3 - y + 1)}. \quad (\text{A3.3.98})$$

Due to the normalization integral, [equation \(A3.3.93\)](#), $f(x)$ cannot be non-zero for arbitrarily large x ; $f(x)$ must vanish for x greater than some cut-off value x_0 , which must be a pole of the integrand in [equation \(A3.3.98\)](#) on the positive real axis. For this to occur $\gamma \leq \frac{4}{27} \equiv \gamma_0$. [Equation \(A3.3.88\)](#) and [equation \(A3.3.96\)](#) together yield an equation for $x = R/R_c$:

$$\frac{dx}{d\tau} = \frac{1}{3\gamma\tau} \left(\frac{1}{x} - \frac{1}{x^2} - \gamma x \right) \quad (\text{A3.3.99})$$

$$(\text{A3.3.100})$$

The form of $g(x)$ is shown in [figure A3.3.10](#). For $\gamma < \gamma_0$, all drops with $x > x_1$ will asymptotically approach the size $x_2 R_c(\tau)$, which tends to infinity with τ as $\tau^{1/3}$ from equation (A3.3.96). For $\gamma > \gamma_0$, all points move to the origin and the conservation condition again cannot be satisfied. The only allowed solution consistent with conservation condition, [equation \(A3.3.93\)](#), is that γ asymptotically approaches γ_0 from above. In doing this it takes an infinite time. (If it reaches γ_0 in finite time, all drops with $x > \frac{3}{2}$ would eventually arrive at $x = \frac{3}{2}$ and become stuck and one has a repeat of the $\gamma > \gamma_0$ case.) $\gamma = \gamma_0 = \frac{4}{27}$ then corresponds to a double pole in the integrand in (A3.3.98). LS show that $\gamma(\tau) = \gamma_0 [1 - \tilde{\epsilon}^2(\tau)]$ with $\tilde{\epsilon}(\tau) \rightarrow 0$ as $\tau \rightarrow \infty$ as $\tau \rightarrow \infty$. For an asymptotic scaled distribution, one uses $\gamma = \gamma_0 = \frac{4}{27}$ and evaluates the integral in (A3.3.98) to obtain

$$f(x) = \begin{cases} \text{constant } x^2 (3+x)^{-(1+\frac{d}{\gamma})} \left(\frac{3}{2}-x\right)^{-(2+\frac{d}{\gamma})} \exp\left(-\frac{d}{3-2x}\right) & \text{for } x < \frac{3}{2} \\ 0 & \text{for } x \geq \frac{3}{2} \end{cases} \quad (\text{A3.3.101})$$

where the normalization constraint, [equation \(A3.3.93\)](#), can be used to determine the constant. $f(x)$ is the scaled LS cluster distribution and is shown in [figure A3.3.11](#) for $d = 3$.

-40-

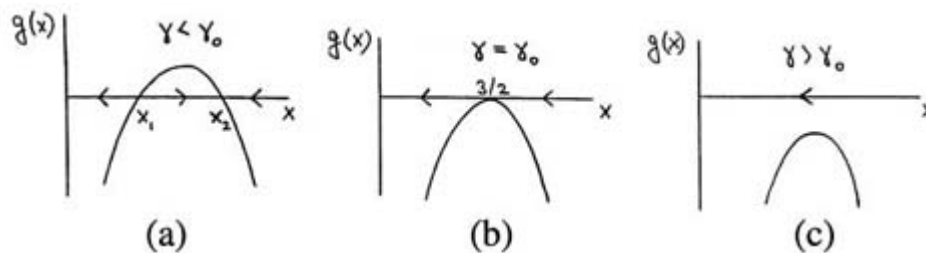


Figure A3.3.10 $g(x)$ as a function of x for the three possible classes of γ .

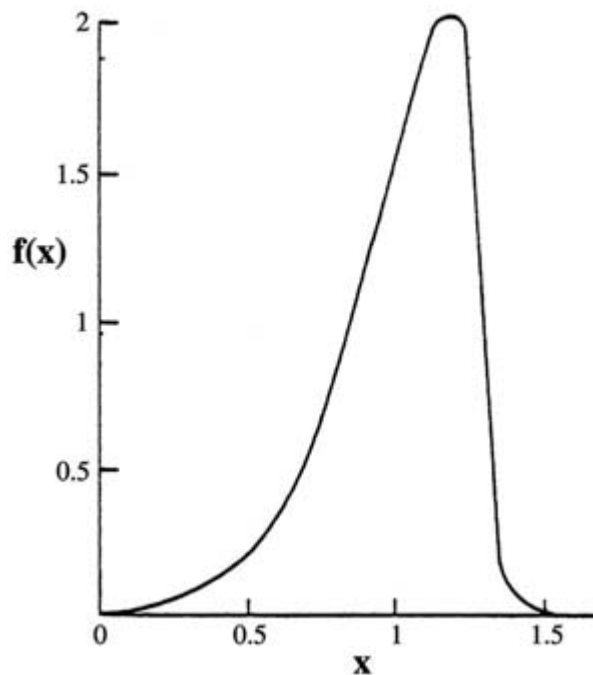


Figure A3.3.11 The asymptotic cluster size distribution $f(x)$ from LS analysis for $d = 3$.

In [section A3.3.3](#) the Langevin models that were introduced for phase transition kinetics utilized the Landau mean field expansion of the free energy, [equation \(A3.3.50\)](#) and [equation \(A3.3.51\)](#). In spite of this, many of the subsequent results based on [\(A3.3.57\)](#) as a starting point are more broadly valid and are dependent only on the existence of a double-well nature of the free energy functional as shown in [figure A3.3.6](#). Also, as the renormalization group analysis shows, the role of thermal noise is irrelevant for the evolution of the initially unstable state. Thus, apart from the random fluctuations in the initial state, which are essential for the subsequent growth of unstable modes, the mean field description is a good theoretical starting point in understanding spinodal decomposition and the ensuing growth. The late-stage growth analysis given in this section is a qualitatively valid starting point for quenches with sufficient off-criticality, and becomes correct asymptotically as the off-criticality ψ_0 increases, bringing the initial post-quench state closer to the coexistence curve, where it is one. In general, one has to add to the LS analysis the cluster–cluster interactions. The current states of such extensions (which are non-trivial) are reviewed in [[37](#), [40](#), [41](#)].

-41-

The main results are that the universal scaling form of the LS cluster distribution function $f(x)$ given above acquires a dependence on $\delta\psi_0 \equiv (1-\psi_0)$ which measures the proximity to the coexistence curve (it is essentially the volume fraction for the vapour–liquid nucleation); also, the $\tau^{1/3}$ growth law for the domain size has the form: $R(\tau) = [K(\delta\psi_0)\tau]^{1/3}$, where $K(\delta\psi_0)$ is a monotonically increasing function of $\delta\psi_0$.

A3.3.5 NUCLEATION KINETICS—METASTABLE SYSTEMS

In this section, we restrict our discussion to homogeneous nucleation, which has a illustrious history spanning at least six decades (see [[37](#), [42](#)] and references therein). Heterogeneous nucleation occurs more commonly in nature, since suspended impurities or imperfectly wetted surfaces provide the interface on which the growth of the new phase is initiated. Heterogeneous nucleation is treated in a recent book by Debenedetti (see [section 3.4](#) of this book, which is listed in Further Reading); interesting phenomena of breath figures and dew formation are related to heterogeneous nucleation, and are discussed in [[43](#), [44](#) and [45](#)].

In contrast to spinodal decomposition, where small-amplitude, long-wavelength fluctuations initiate the growth, the kinetics following an initial metastable state requires large-amplitude (nonlinear) fluctuations for the stable phase to nucleate. A qualitative picture for the nucleation event is as follows. For the initial metastable state, the two minima of the local free energy functional are not degenerate, in contrast to the initially unstable case shown in [figure A3.3.6](#). The metastable system is initially in the higher of the two minimum energy states and has to overcome a free energy barrier (provided by the third extremum which is a maximum) in order to go over to the absolute minimum, which is the system's ground state. For this, it requires an activation energy which is obtained through rarely occurring large-amplitude fluctuations of the order parameter, in the form of a critical droplet. Physically, the rarity of the nonlinear fluctuation introduces large characteristic times for nucleation to occur.

The central quantity of interest in homogeneous nucleation is the nucleation rate J , which gives the number of droplets nucleated per unit volume per unit time for a given supersaturation. The free energy barrier is the dominant factor in determining J ; J depends on it exponentially. Thus, a small difference in the different model predictions for the barrier can lead to orders of magnitude differences in J . Similarly, experimental measurements of J are sensitive to the purity of the sample and to experimental conditions such as temperature. In modern field theories, J has a general form

(A3.3.102)

$$J = \frac{1}{\tau^*} \Omega e^{(E_c/kT)}$$

where τ^* is the time scale for the macroscopic fluctuations and Ω is the volume of phase space accessible for fluctuations. The barrier height to nucleation E_c is described below.

A homogeneous metastable phase is always stable with respect to the formation of infinitesimal droplets, provided the surface tension σ is positive. Between this extreme and the other thermodynamic equilibrium state, which is inhomogeneous and consists of two coexisting phases, a critical size droplet state exists, which is in unstable equilibrium. In the ‘classical’ theory, one makes the capillarity approximation: the critical droplet is assumed homogeneous up to the boundary separating it from the metastable background and is assumed to be the same as the new phase in the bulk. Then the work of formation $W(R)$ of such a droplet of arbitrary radius R is the sum of the

-42-

free energy gain of the new stable phase droplet and the free energy cost due to the new interface that is formed:

$$W(R) = 4\pi R^2 \sigma - \frac{4}{3}\pi R^3 \Delta f \quad (\text{A3.3.103})$$

where Δf is the positive bulk free energy difference per unit volume between the stable and metastable phases. From this, by maximizing $W(R)$ with respect to R , one obtains the barrier height to nucleation $W(R_c) \equiv E_c$ and the critical radius R_c : $R_c = 2\sigma/(\Delta f)$ and $E_c = (16\pi/3)\sigma^3/(\Delta f)^2$. For a supercooled vapour nucleating into liquid drops, Δf is given by $kT\rho_1 \ln(\varepsilon)$, where ρ_1 is the bulk liquid density and $\varepsilon = P/P_e$ is the supersaturation ratio, which is the ratio of the actual pressure P to the equilibrium vapour pressure P_e of the liquid at the same T . For the case of the nucleation of a crystal from a supercooled liquid, $\Delta f = \Delta\mu/v$, where v is the volume per particle of the solid and $\Delta\mu$ is the chemical potential difference between the bulk solid and the bulk liquid. These results, given here for three-dimensional systems, can be easily generalized to an arbitrary d -dimensional case [37]. Often, it is useful to use the capillary length $l_c = (2\sigma v)/(kT)$ as the unit of length: the critical radius is $R_c = l_c/\varepsilon(t)$, where the supersaturation $\varepsilon(t) = \Delta\mu/(kT)$, and the nucleation barrier is $(E_c/kT) = (\varepsilon_0/\varepsilon(t))^2$, where the dimensionless quantity $\varepsilon_0 = l_c[(4\pi\sigma)/(3kT)]^{1/2}$.

Early (classical) theories of homogeneous nucleation are based on a microscopic description of cluster dynamics (see reference (1) in [37]). A kinetic equation for the droplet number density $n_i(t)$ of a given size i at time t is written, in which its time rate of change is the difference between J_{i-1} and J_i , where J_i is the rate at which droplets of size i grow to size $i+1$ by gaining a single molecule [13]. By providing a model for the forward and backward rates at which a cluster gains or loses a particle, J_i is related to $\{n_i(t)\}$, and a set of coupled rate equations for $\{n_i(t)\}$ is obtained. The nucleation rate is obtained from the steady-state solution in which $J_i = J$ for large i . The result is in the form of equation (A3.3.102), with specific expressions for $J_0 \equiv \Omega/\tau^*$ obtained for various cases such as vapour–liquid and liquid–solid transitions. Classical theories give nucleation rates that are low compared to experimental measurements. Considerable effort has gone into attempts to understand ‘classical’ theories, and compare their results to experiments (see references in 42).

In two classic papers [18, 46], Cahn and Hilliard developed a field theoretic extension of early theories of nucleation by considering a spatially inhomogeneous system. Their free energy functional, equations (A3.3.52), has already been discussed at length in section A3.3.3. They considered a two-component incompressible fluid. The square gradient approximation implied a slow variation of the concentration on the

coarse-graining length scale ξ (i.e. a diffuse interface). In their 1959 paper [46], they determined the saddle point of this free energy functional and analysed the properties of a critical nucleus of the minority phase within the metastable binary mixture. While the results agree with those of the early theories for low supersaturation, the properties of the critical droplet change as the supersaturation is increased: (i) the work required to form a critical droplet becomes progressively less compared to ‘classical’ theory result, and approaches zero continuously as spinodal is approached; (ii) the interface with the exterior phase becomes more diffuse and the interior of the droplet becomes inhomogeneous in its entirety; (iii) the concentration at the droplet centre approaches that of the exterior phase; and (iv) the radius and excess concentration in the droplet at first decrease, pass through a minimum and become infinite at the spinodal. These papers provide a description of the spatially inhomogeneous critical droplet, which is not restricted to planar interfaces, and yields, for $W(R)$, an expression that goes to zero at the mean field spinodal. The Cahn–Hilliard theory has been a useful starting point in the development of modern nucleation theories.

A full theory of nucleation requires a dynamical description. In the late 1960s, the early theories of homogeneous nucleation were generalized and made rigorous by Langer [47]. Here one starts with an appropriate Fokker–Planck

-43-

(or its equivalent Langevin) equation for the probability distribution function $P(\{\psi_i\}, t)$ for the set of relevant field variables $\{\psi_i\}$ which are semi-macroscopic and slowly varying:

$$\frac{\partial P}{\partial t} = - \sum_i \frac{\delta J_i}{\delta \psi_i} \quad (\text{A3.3.104})$$

where the probability current J_i is given by

$$J_i = \sum_j M_{ij} \left[\frac{\delta \mathcal{F}}{\delta \psi_j} + kT \frac{\delta P}{\delta \psi_j} \right]. \quad (\text{A3.3.105})$$

\mathcal{F} is the free energy functional, for which one can use equation (A3.3.52). The summation above corresponds to both the sum over the semi-macroscopic variables and an integration over the spatial variable \vec{r} . The mobility matrix M_{ij} consists of a symmetric dissipative part and an antisymmetric non-dissipative part. The symmetric part corresponds to a set of generalized Onsager coefficients.

The decay of a metastable state corresponds to passing from a local minimum of \mathcal{F} to another minimum of lower free energy which occurs only through improbable free energy fluctuations. The most probable path for this passage to occur when the nucleation barrier is high is via the saddle point. The saddle point corresponds to a critical droplet of the stable phase in a metastable phase background. The nucleation rate is given by the steady-state solution of the Fokker–Planck equation that describes a finite probability current across the saddle point. The result is of the form given in (A3.3.102). The quantity $1/\tau^*$ is also referred to as a dynamical prefactor and Ω as a statistical prefactor.

Within this general framework there have been many different systems modelled and the dynamical, statistical prefactors have been calculated. These are detailed in [42]. For a binary mixture, phase separating from an initially metastable state, the work of Langer and Schwartz [48] using the Langer theory [47] gives the nucleation rate as

$$J(t) = (l_c^3 t_c)^{-1} \frac{3\epsilon_0^6}{4\pi} \left(\frac{\epsilon(t)}{\epsilon_0} \right)^{2/3} \left(1 + \frac{\epsilon(t)}{\epsilon_0} \right)^{3.55} \exp \left[- \left(\frac{\epsilon_0}{\epsilon(t)} \right)^2 \right] \quad (\text{A3.3.106})$$

where l_c is the capillary length defined above and the characteristic time $t_c = l_c^2/[DvC_{\text{eq}}(\infty)]$ with D as the diffusion coefficient and $C_{\text{eq}}(\infty)$ the solute concentration in the background matrix at a planar interface in the phase separated system [37].

One can introduce a distributed nucleation rate $j(R, t)dR$ for nucleating clusters of radius between R and $R + dR$. Its integral over R is the total nucleation rate $J(t)$. Equation (A3.3.103) can be viewed as a radius-dependent droplet energy which has a maximum at $R = R_c$. If one assumes $j(R, t)$ to be a Gaussian function, then

$$j(R, t) = \frac{J(t)}{\sqrt{2\pi}(\delta R)} \exp \left[- \frac{(R - R_c)^2}{2(\delta R)^2} \right] \quad (\text{A3.3.107})$$

where $(\delta R)^2 = 2[E_c - W(R)]/|E_c''|$, with E_c and E_c'' being, respectively, the values of $W(R)$ and its second derivative evaluated at $R = R_c$. Langer [47] showed that the droplet energy is not only a function of R but can also depend on the capillary wavelength fluctuations w ; i.e. $W(R) \rightarrow E(R, w)$. Then, the droplets appear at the saddle point in the surface of $E(R, w)$. The $2 - d$ surface area of the droplet is given by $4\pi(R^2 + w^2)$, which gives the change in the droplet energy due to non-zero w , as $\Delta E(R) = 4\pi\sigma w^2$. Both approaches lead to the same Gaussian form of the distributed nucleation rate with $w \equiv (\delta R)$ estimated from an uncertainty in the required activation energy of the order of $kT/2$.

Just as is the case for the LSW theory of Ostwald ripening, the Langer–Schwartz theory is also valid for quenches close to the coexistence curve. Its extension to non-zero volume fractions requires that such a theory take into account cluster–cluster correlations. A framework for such a theory has been developed [37] using a multi-droplet diffusion equation for the concentration field. This equation has been solved analytically using (i) a truncated multipole expansion and (ii) a mean field Thomas–Fermi approximation. The equation has also been numerically simulated. Such studies are among the first attempts to construct a unified model for the entire process of phase separation that combines steady-state homogeneous nucleation theory with the LSW mechanism for ripening, modified to account for the inter-cluster correlations.

A3.3.6 SUMMARY

In this brief review of dynamics in condensed phases, we have considered dense systems in various situations. First, we considered systems in equilibrium and gave an overview of how the space–time correlations, arising from the thermal fluctuations of slowly varying physical variables like density, can be computed and experimentally probed. We also considered capillary waves in an inhomogeneous system with a planar interface for two cases: an equilibrium system and a NESS system under a small temperature gradient. Finally, we considered time evolving non-equilibrium systems in which a quench brings a homogeneous system to an initially unstable (spinodal decomposition) or metastable state (nucleation) from which it evolves to a final inhomogeneous state of two coexisting equilibrium phases. The kinetics of the associated processes provides rich physics involving nonlinearities and inhomogeneities. The early-stage kinetics associated with the formation of interfaces and the late-stage interface dynamics in such systems continues to provide

challenging unsolved problems that have emerged from the experimental observations on real systems and from the numerical simulations of model systems.

REFERENCES

- [1] Chandrasekhar S 1943 *Rev. Mod. Phys.* **15** 1
 - [2] Koch S W, Desai R C and Abraham F F 1982 *Phys. Mod. A* **26** 1015
 - [3] Mountain R D 1966 *Rev. Mod. Phys.* **38** 205
 - [4] Fleury P A and Boon J P 1969 *Phys. Mod.* **186** 244 Fleury P A and Boon J P 1973 *Adv. Chem. Phys.* **24** 1
-
- 45-
- [5] Tong E and Desai R C 1970 *Phys. Mod. A* **2** 2129
 - [6] Desai R C and Kapral R 1972 *Phys. Mod. A* **6** 2377
 - [7] Weinberg M, Kapral R and Desai R C 1973 *Phys. Mod. A* **7** 1413
 - [8] Kapral R and Desai R C 1974 *Chem. Phys.* **3** 141
 - [9] Grant M and Desai R C 1983 *Phys. Mod. A* **27** 2577
 - [10] Jhon M, Dahler J S and Desai R C 1981 *Adv. Chem. Phys.* **46** 279
 - [11] Hashimoto T, Kumaki J and Kawai H 1983 *Macromolecules* **16** 641
 - [12] Chou Y C and Goldberg W I 1981 *Phys. Rev. A* **23** 858 see also Wong N-C and Knobler C M 1978 *J. Chem. Phys.* **69** 725
 - [13] Gunton J D, San Miguel M and Sahni P S 1983 *Phase Transitions and Critical Phenomena* vol 8, ed C Domb and J L Lebowitz (New York: Academic)
 - [14] Furukawa H 1985 *Adv. Phys.* **34** 703
 - [15] Binder K 1987 *Rep. Prog. Phys.* **50** 783
 - [16] Bray A J 1994 *Adv. Phys.* **43** 357
 - [17] Glotzer S C 1995 *Ann. Rev. Comput. Phys.* **II** 1–46
 - [18] Cahn J W and Hilliard J E 1958 *J. Chem. Phys.* **28** 258
 - [19] Seul M and Andelman D 1995 *Science* **267** 476
 - [20] Desai R C 1997 *Phys. Can.* **53** 210
 - [21] Rogers T M, Elder K R and Desai R C 1988 *Phys. Mod. B* **37** 9638
 - [22] Garcia-Ojalvo J and Sancho J M 1999 *Noise in Spatially Extended Systems* (Berlin: Springer)
 - [23] Elder K R, Rogers T M and Desai R C 1988 *Phys. Mod. B* **38** 4725
 - [24] Binder K 1984 *Phys. Rev. A* **29** 341
 - [25] Pego R L 1989 *Phys. Rev. S A* **422** 261
 - [26] Siggia E D 1979 *Phys. Rev. A* **20** 595
 - [27] Cahn J W 1961 *Acta. Metall.* **9** 795 Cahn J W 1966 *Acta. Metall.* **14** 1685 Cahn J W 1968 *Trans. Metall. Soc. AIME* **242** 166
 - [28] Rogers T M and Desai R C 1989 *Phys. Mod. B* **39** 11 956
 - [29] Elder K R and Desai R C 1989 *Phys. Mod. B* **40** 243

- [30] Toral R, Chakrabarti A and Gunton J D 1989 *Phys. Mod. B* **39** 901
[31] Shinozaki A and Oono Y 1993 *Phys. Mod. E* **48** 2622 Shinozaki A and Oono Y 1991 *Phys. Mod. A* **66** 173
[32] Yeung C 1988 *Phys. Mod. L* **61** 1135
[33] Furukawa H 1989 *Phys. Mod. B* **40** 2341
-

-46-

- [34] Fratzl P, Lebowitz J L, Penrose O and Amar J 1991 *Phys. Rev. B* **44** 4794, see appendix B
[35] Rogers T M 1989 *PhD Thesis* University of Toronto
[36] Allen S M and Cahn J W 1979 *Acta. Metall.* **27** 1085 see also Ohta T, Jasnow D and Kawasaki K 1982 *Phys. Rev. Lett.* **49** 1223 for the model A scaled structure factor
[37] Sagui C and Grant M 1999 *Phys. Mod. E* **59** 4175 and references therein
[38] Lifshitz I M and Slyozov V V 1961 *J. Phys. Chem. Solids* **19** 35
[39] Wagner C 1961 *Z. Elektrochem.* **65** 581
[40] Yao J H, Elder K R, Guo H and Grant M 1993 *Phys. Mod. B* **47** 1410
[41] Akaiwa N and Voorhees P W 1994 *Phys. Mod. E* **49** 3860
Akaiwa N and Voorhees P W 1996 *Phys. Mod. E* **54** R13
[42] Gunton J D 1999 *J. Stat. Phys.* **95** 903 and references therein
[43] Fritter D, Knobler C M and Beysens D 1991 *Phys. Mod. A* **43** 2858
[44] Beysens D, Steyer A, Guenoun P, Fritter D and Knobler C M 1991 *Phase Trans.* **31** 219
[45] Rogers T M, Elder K R and Desai R C 1988 *Phys. Mod. B* **38** 5303
[46] Cahn J W and Hilliard J E 1959 *J. Chem. Phys.* **31** 688
[47] Langer J S 1967 *Ann. Phys.* **41** 108 Langer J S 1969 *Ann. Phys.* **54** 258
[48] Langer J S and Schwartz A J 1980 *Phys. Mod. A* **21** 948
-

FURTHER READING

- Balucani U and Zoppi M 1994 *Dynamics of the Liquid State* (Oxford: Oxford University Press)
- Berne B J and Pecora R 1976 *Dynamic Light Scattering* (New York: Wiley)
- Boon J P and Yip S 1980 *Molecular Hydrodynamics* (New York: McGraw-Hill)
- Forster D 1975 *Hydrodynamic Fluctuations, Broken Symmetry, and Correlation Functions* (New York: Benjamin)
- Debenedetti P G 1996 *Metastable Liquids* (Princeton, NJ: Princeton University Press)
- Gunton J D and Droz M 1983 *Introduction to the Theory of Metastable and Unstable States* (Berlin: Springer)
- Landau L D and Lifshitz E M 1959 *Fluid Mechanics* (Reading, MA: Addison-Wesley) ch 2, 7, 16, 17. (More recent editions do not have chapter 17.)
- Rowlinson J S and Widom B 1982 *Molecular Theory of Capillarity* (Oxford: Clarendon)

A 3.4 Gas-phase kinetics

David Luckhaus and Martin Quack

A3.4.1 INTRODUCTION

Gas-phase reactions play a fundamental role in nature, for example atmospheric chemistry [1, 2, 3, 4 and 5] and interstellar chemistry [6], as well as in many technical processes, for example combustion and exhaust fume cleansing [7, 8 and 9]. Apart from such practical aspects the study of gas-phase reactions has provided the basis for our understanding of chemical reaction mechanisms on a microscopic level. The typically small particle densities in the gas phase mean that reactions occur in well defined elementary steps, usually not involving more than three particles.

At the limit of extremely low particle densities, for example under the conditions prevalent in interstellar space, ion–molecule reactions become important (see [chapter A3.5](#)). At very high pressures gas-phase kinetics approach the limit of condensed phase kinetics where elementary reactions are less clearly defined due to the large number of particles involved (see [chapter A3.6](#)).

Here, we mainly discuss homogeneous gas-phase reactions at intermediate densities where ideal gas behaviour can frequently be assumed to be a good approximation and diffusion is sufficiently fast that transport processes are not rate determining. The focus is on thermally activated reactions induced by collisions at well defined temperatures, although laser induced processes are widely used for the experimental study of such gas-phase reactions (see [chapter B2.1](#)). The aim of the present chapter is to introduce the basic concepts at our current level of understanding. It is not our goal to cover the vast original literature on the general topic of gas reactions. We refer to the books and reviews cited as well as to [chapter B2.1](#) for specific applications.

Photochemical reactions ([chapter A3.13](#)) and heterogeneous reactions on surfaces ([chapter A3.10](#)) are discussed in separate chapters.

A3.4.2 DEFINITIONS OF THE REACTION RATE

There are many ways to define the rate of a chemical reaction. The most general definition uses the rate of change of a thermodynamic state function. Following the second law of thermodynamics, for example, the change of entropy S with time t would be an appropriate definition under reaction conditions at constant energy U and volume V :

$$v_S(t) = \left(\frac{\partial S}{\partial t} \right)_{U,V} \geq 0. \quad (\text{A3.4.1})$$

An alternative rate quantity under conditions of constant temperature T and volume, frequently realized in gas kinetics, would be

$$v_A(t) = - \left(\frac{\partial A}{\partial t} \right)_{T,V} \geq 0, \quad (\text{A3.4.2})$$

where A is the Helmholtz free energy.

For non-zero v_S and v_A the problem of defining the thermodynamic state functions under non-equilibrium conditions arises (see [chapter A3.2](#)). The definition of rate of change implied by [equation \(A3.4.1\)](#) and [equation \(A3.4.2\)](#) includes changes that are not due to chemical reactions.

In *reaction kinetics* it is conventional to define *reaction rates* in the context of chemical reactions with a well defined stoichiometric equation

$$0 = \sum_i \nu_i B_i \quad (\text{A3.4.3})$$

where ν_i are the stoichiometric coefficients of species B_i ($\nu_i < 0$ for reactants and $\nu_i > 0$ for products, by convention). This leads to the conventional definition of the ‘rate of conversion’:

$$v_\xi(t) = \frac{d\xi}{dt} = \nu_i^{-1} \frac{dn_i}{dt}, \quad (\text{A3.4.4})$$

The ‘extent of reaction’ ξ is defined in terms of the amount n_i of species B_i (i.e. the amount of substance or enplethy n_p , usually expressed in moles [10]):

$$\xi(t) = \frac{n_i(t) - n_i(t=0)}{\nu_i}; \quad (\text{A3.4.5})$$

v_ξ is an extensive quantity, i.e. for two independent subsystems I and II we have $v_\xi(I + II) = v_\xi(I) + v_\xi(II)$. For homogeneous reactions we obtain the conventional definition of the ‘reaction rate’ v_c as rate of conversion per volume

$$v_c(t) = V^{-1} v_\xi(t) = \nu_i^{-1} \frac{dc_i}{dt} \quad (\text{A3.4.6})$$

where c_i is the concentration of species B_p , for which we shall equivalently use the notation $[B_i]$ (with the common unit mol dm^{-3} and the unit of v_c being $\text{mol dm}^{-3} \text{s}^{-1}$). In gas kinetics it is particularly common to use the quantity particle

density for concentration, for which we shall use C_i (capital letter) with

$$v_c(t) = N_A V^{-1} v_\xi(t) = v_i^{-1} \frac{dc_i}{dt}.$$

N_A is Avogadro's constant. The most commonly used unit then is $\text{cm}^{-3} \text{s}^{-1}$, sometimes inconsistently written ($\text{molecule cm}^{-3} \text{s}^{-1}$). v_c is an intensive quantity. Table A3.4.1 summarizes the definitions.

Table A3.4.1 Definitions of the reaction rate.

Constraint	Extensive quantity	Intensive quantity	Reaction rate
$U, V = \text{constant}$ adiabatic	Entropy S (thermodynamics of irreversible processes)	Local entropy $S_V = \frac{\delta S}{\delta V}$	$v_S = \frac{dS}{dt} \geq 0$ unit: $\text{J K}^{-1} \text{s}^{-1}$ $v_{S_V} = \frac{dS_V}{dt} \geq 0$ unit: $\text{J K}^{-1} \text{s}^{-1} \text{cm}^{-3}$
$T, V = \text{constant}$ isothermal	Helmholtz energy $A = U - TS$	Local A $A_V = \frac{\delta A}{\delta V}$	$v_A = -\frac{dA}{dt} \geq 0$ unit: J s^{-1} $v_{A_V} = \frac{dA_V}{dt} \geq 0$ unit: $\text{J s}^{-1} \text{cm}^{-3}$
$V = \text{constant}$ isothermal or adiabatic, fixed stoichiometry	Amount of substance n_i , number of particles N_i $0 = \sum_i \nu_i B_i$ extent of reaction ξ $d\xi = \nu_i^{-1} dn_i$	Concentration $c_i = n_i / V$ $\frac{\delta \xi}{\delta V} \simeq \frac{\xi}{V}$ $c_i = \frac{N_i}{V}$	$\frac{dn_i}{dt}$ or $\frac{dc_i}{dt}$ unit: mol s^{-1} or $\text{mol cm}^{-3} \text{s}^{-1}$ $\frac{d\xi}{dt}$ or $\frac{1}{V} \frac{d\xi}{dt} = \frac{1}{\nu_i} \frac{dc_i}{dt}$ unit: mol s^{-1} , $\text{mol cm}^{-3} \text{s}^{-1}$ or $\text{molecule cm}^{-3} \text{s}^{-1}$ $v_c = \frac{1}{\nu_i} \frac{dc_i}{dt}$

Figure A3.4.1 shows as an example the time dependent concentrations and entropy for the simple decomposition reaction of chloroethane:



The slopes of the functions shown provide the reaction rates according to the various definitions under the reaction conditions specified in the figure caption. These slopes are similar, but not identical (nor exactly proportional), in this simple case. In more complex cases, such as oscillatory reactions ([chapter A3.14](#) and [chapter C3.6](#)), the simple definition of an overall rate law through [equation \(A3.4.6\)](#) loses its usefulness, whereas [equation \(A3.4.1\)](#) could still be used for an isolated system.

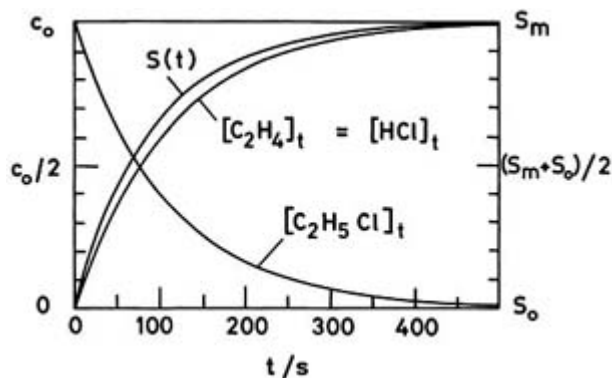


Figure A3.4.1. Concentration and entropy as functions of time for reaction equation (A3.4.8). S_m is the maximum value of the entropy ([20]).

A3.4.3 EMPIRICAL RATE LAWS AND REACTION ORDER

A general form of the ‘rate law’, i.e. the differential equation for the concentrations is given by

$$v_c(t) = v_i^{-1} \frac{dc_i}{dt} = f(c_1, c_2, \dots). \quad (\text{A3.4.9})$$

The functional dependence of the reaction rate on concentrations may be arbitrarily complicated and include species not appearing in the stoichiometric equation, for example, catalysts, inhibitors, etc. Sometimes, however, it takes a particularly simple form, for example, under certain conditions for elementary reactions and for other relatively simple reactions:

$$v_c(t) = k \prod_i c_i^{m_i} \quad (\text{A3.4.10})$$

with a concentration-independent and frequently time-independent ‘rate coefficient’ or ‘rate constant’ k . m_i is the order of the reaction with respect to the species B_i and the total order of the reaction m is given by

$$m = \sum_i m_i \quad (\text{A3.4.11})$$

where m and m_i are real numbers. Table A3.4.2 summarizes a few examples of such rate laws. In general, one may allow for rate coefficients that depend on time (but not on concentration) [11].

Table A3.4.2 Rate laws, reaction order, and rate constants.

Reaction	Rate law	Reaction order	Dimension of rate constant [k]
Isomerization $\text{CH}_3\text{NC} = \text{CH}_3\text{CN}$ excess of inert gas	$-\frac{d[\text{CH}_3\text{NC}]}{dt} = k[\text{CH}_3\text{NC}]$	First order in CH_3NC . First-order total	$[\text{s}^{-1}]$
Atom transfer $\text{F} + \text{CHF}_3 = \text{HF} + \text{CF}_3$	$-\frac{d[\text{F}]}{dt} = k[\text{F}][\text{CHF}_3]$	First order in F. First order in CHF_3 . Second-order total	$[\text{cm}^3 \text{mol}^{-1} \text{s}^{-1}]$ or $[\text{cm}^3 \text{s}^{-1}]$
Radical recombination $2\text{CH}_3 = \text{C}_2\text{H}_6$	$-\frac{1}{2} \frac{d[\text{CH}_3]}{dt} = k[\text{CH}_3]^2$	Second order in CH_3 . Second-order total	$[\text{cm}^3 \text{mol}^{-1} \text{s}^{-1}]$ or $[\text{cm}^3 \text{s}^{-1}]$
Decomposition $\text{CH}_3\text{CHO} = \text{CH}_4 + \text{CO}$	$-\frac{d[\text{CH}_3\text{CHO}]}{dt} = k[\text{CH}_3\text{CHO}]^{3/2}$	Order 3/2 in CH_3CHO and total	$[\text{cm}^{3/2} \text{mol}^{-1/2} \text{s}^{-1}]$

If certain species are present in large excess, their concentration stays approximately constant during the course of a reaction. In this case the dependence of the reaction rate on the concentration of these species can be included in an effective rate constant k_{eff} . The dependence on the concentrations of the remaining species then defines the *apparent order of the reaction*. Take for example [equation \(A3.4.10\)](#) with $c_{i>1} \gg c_1$. The result would be a *pseudo m_1 th order* effective rate law:

$$v_c(t) = k_{\text{eff}} c_1^{m_1} \quad (\text{A3.4.12})$$

$$k_{\text{eff}} = k \prod_{i>1} c_i^{m_i}. \quad (\text{A3.4.13})$$

This is the situation exploited by the so-called *isolation* method to determine the order of the reaction with respect to each species (see [chapter B2.1](#)). It should be stressed that the rate coefficient k in [\(A3.4.10\)](#) depends upon the definition of the v_i in the stoichiometric equation. It is a conventionally defined quantity to within multiplication of the stoichiometric equation by an arbitrary factor (similar to reaction enthalpy).

The definitions of the empirical rate laws given above do not exclude empirical rate laws of another form. Examples are reactions, where a reverse reaction is important, such as in the *cis-trans* isomerization of 1,2-dichloroethene:



$$-\frac{d[\text{cis}]}{dt} = k_a[\text{cis}] - k_b[\text{trans}] \quad (\text{A3.4.15})$$

or the classic example of hydrogen bromide formation:



$$-\frac{d[\text{HBr}]}{dt} = k_a[\text{H}_2][\text{Br}_2]^{1/2} \left(1 + k_b \frac{[\text{HBr}]}{[\text{Br}_2]} \right)^{-1} \quad (\text{A3.4.17})$$

Neither (A3.4.15) nor (A3.4.17) is of the form (A3.4.10) and thus neither reaction order nor a unique rate coefficient can be defined. Indeed, the number of possible rate laws that are *not* of the form of (A3.4.10) greatly exceeds those cases following (A3.4.10). However, certain particularly simple reactions necessarily follow a law of type of (A3.4.10). They are particularly important from a mechanistic point of view and are discussed in the next section.

A3.4.4 ELEMENTARY REACTIONS AND MOLECULARITY

Sometimes the reaction orders m_i take on integer values. This is generally the case, if a chemical reaction



or



takes place on a microscopic scale through direct interactions between particles as implied by equation (A3.4.18) or equation (A3.4.19). Thus, the coefficients of the substances in (A3.4.18) and (A3.4.19) represent the actual number of particles involved in the reaction, rather than just the stoichiometric coefficients. To keep the distinction clear we shall reserve the reaction arrow ‘ \rightarrow ’ for such *elementary reactions*. Sometimes the inclusion of the reverse elementary reaction will be signified by a double arrow ‘ \rightleftharpoons ’. Other, *compound reactions* can always be decomposed into a set of—not necessarily consecutive—elementary steps representing the *reaction mechanism*.

-7-

Elementary reactions are characterized by their *molecularity*, to be clearly distinguished from the reaction order. We distinguish *uni-* (or *mono-*), *bi-*, and *trimolecular* reactions depending on the number of particles involved in the ‘essential’ step of the reaction. There is some looseness in what is to be considered ‘essential’, but in gas kinetics the definitions usually are clearcut through the number of particles involved in a reactive collision; plus, perhaps, an additional convention as is customary in unimolecular reactions.

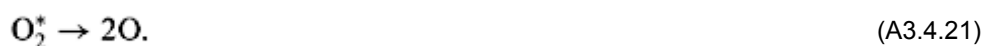
A3.4.4.1 UNIMOLECULAR REACTIONS

Strictly unimolecular processes—sometimes also called *monomolecular*—involve only a single particle:

(A3.4.20)

A → products.

Classic examples are the spontaneous emission of light or spontaneous radioactive decay. In chemistry, an important class of monomolecular reactions is the predissociation of metastable (excited) species. An example is the formation of oxygen atoms in the upper atmosphere by predissociation of electronically excited O₂ molecules [12, 13 and 14]:



Excited O₂^{*} molecules are formed by UV light absorption. Monomolecular reactions (e.g., $c = [\text{O}_2^*]$) show a first-order rate law:

$$-\frac{dc}{dt} = kc(t). \quad (\text{A3.4.22})$$

Integration of the differential equation with time-independent k leads to the familiar exponential decay:

$$c(t) = c(0) \exp\{-kt\}. \quad (\text{A3.4.23})$$

The rate constant in this case is of the order of 10¹¹ s⁻¹ depending on the rovibronic level considered.

Another example of current interest is the *vibrational predissociation* of hydrogen bonded complexes such as (HF)₂:



-8-

With one quantum of non-bonded (HF)-stretching excitation (*) the internal energy (~50 kJ mol⁻¹) is about four times in excess of the hydrogen bond dissociation energy (12.7 kJ mol⁻¹). At this energy the rate constant is about $k \approx 5 \times 10^7 \text{ s}^{-1}$ [15]. With two quanta of (HF)-stretching (at about seven times the dissociation energy) the rate constant is $k \approx 7.5 \times 10^8 \text{ s}^{-1}$ in all cases, depending on the rovibrational level considered [16, 17].

While monomolecular collision-free predissociation excludes the preparation process from explicit consideration, thermal unimolecular reactions involve collisional excitation as part of the unimolecular mechanism. The simple mechanism for a thermal chemical reaction may be formally decomposed into three (possibly reversible) steps (with rovibronically excited (CH₃NC)*):





The inert collision partner M is assumed to be present in large excess:

$$[\text{M}] \gg [\text{CH}_3\text{NC}] \quad (\text{A3.4.28})$$

$$[\text{M}] \approx \text{constant}. \quad (\text{A3.4.29})$$

This mechanism as a whole is called ‘unimolecular’ since the essential isomerization step equation (A3.4.26) only involves a single particle, *viz.* CH_3NC . Therefore it is often simply written as follows:



Experimentally, one finds the same first-order rate law as for monomolecular reactions, but with an effective rate constant k that now depends on $[\text{M}]$.

$$-\frac{dc}{dt} = k([\text{M}]c(t)). \quad (\text{A3.4.31})$$

The correct treatment of the mechanism (equation (A3.4.25), equation (A3.4.26) and equation (A3.4.27), which goes back to Lindemann [18] and Hinshelwood [19], also describes the pressure dependence of the effective rate constant in the low-pressure limit ($[\text{M}] \leq [\text{CH}_3\text{NC}]$, see [section A3.4.8.2](#)).

The unimolecular rate law can be justified by a probabilistic argument. The number ($N_{\text{A}} Vdc \propto dc$) of particles which react in a time dt is proportional both to this same time interval dt and to the number of particles present ($N_{\text{A}} Vc \propto c$). However, this probabilistic argument need not always be valid, as illustrated in figure A3.4.2 for a simple model [20]:

A number of particles perform periodic rotations in a ring-shaped container with a small opening, through which some particles can escape. Two situations can now be distinguished.

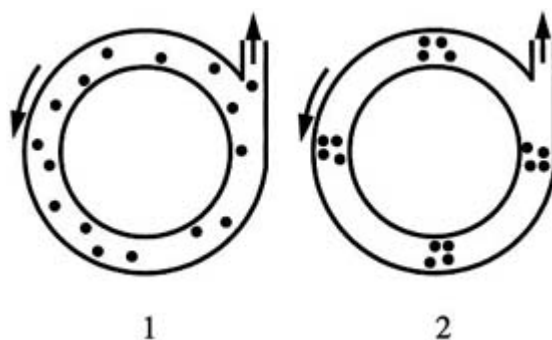


Figure A3.4.2. A simple illustration of limiting dynamical behaviour: case 1 statistical, case 2 coherent (after [20]).

Case 1. The particles are statistically distributed around the ring. Then, the number of escaping particles will be proportional both to the time interval (opening time) dt and to the total number of particles in the container. The result is a first-order rate law.

Case 2. The particles rotate in small packets ('coherently' or 'in phase'). Obviously, the first-order rate law no longer holds. In [chapter B2.1](#) we shall see that this simple consideration has found a deeper meaning in some of the most recent kinetic investigations [21].

A3.4.4.2 BIMOLECULAR REACTIONS

Bimolecular reactions involve two particles in their essential step. In the so-called *self-reactions* they are of the same species:



with the stoichiometric equation



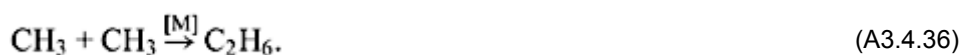
Typical examples are radical recombinations:



-10-



Here the initially formed excited species $(\text{C}_2\text{H}_6)^*$ is sufficiently long lived that the deactivation step (equation (A3.4.35)) is not essential and one writes



The rate is given by the second-order law ($c \equiv [\text{CH}_3]$ or $c \equiv [\text{A}]$)

$$-\frac{1}{2} \frac{dc}{dt} = kc^2. \quad (\text{A3.4.37})$$

Integration leads to

$$\frac{1}{c(t)} = 2kt + \frac{1}{c(0)}. \quad (\text{A3.4.38})$$

Bimolecular reactions between different species



lead to the second-order rate law

$$-\frac{dc_A}{dt} = kc_Ac_B. \quad (\text{A3.4.40})$$

For $c_B(0) \neq c_A(0)$ the solution of this differential equation is

$$\ln \left(\frac{c_B(t)}{c_A(t)} \right) - \ln \left(\frac{c_B(0)}{c_A(0)} \right) = (c_B(0) - c_A(0))kt. \quad (\text{A3.4.41})$$

The case of equal concentrations, $c_B = c_A = c(t)$, is similar to the case A + A in equation (A3.4.37), except for the stoichiometric factor of two. The result thus is

$$\frac{1}{c(t)} = kt + \frac{1}{c(0)}. \quad (\text{A3.4.42})$$

If one of the reactants is present in large excess $c_B \gg c_A$ its concentration will essentially remain constant throughout the reaction. Equation (A3.4.41) then simplifies to

$$c_A(t) = c_A(0) \exp\{-k_{\text{eff}}t\} \quad (\text{A3.4.43})$$

with the effective *pseudo first-order* rate constant $k_{\text{eff}} = kc_B$.

One may justify the differential equation (A3.4.37) and equation (A3.4.40) again by a probability argument. The number of reacting particles $N_A Vdc \propto dc$ is proportional to the frequency of encounters between two particles and to the time interval dt . Since not every encounter leads to reaction, an additional reaction probability P_R has to be introduced. The frequency of encounters is obtained by the following simple argument. Assuming a statistical distribution of particles, the probability for a given particle to occupy a

volume element δV is proportional to the concentration c . If the particles move independently from each other (ideal behaviour) the same is true for a second particle. Therefore the probability for two particles to occupy the same volume element (an encounter) is proportional to c^2 . This leads to the number of particles reacting in the time interval dt :

$$N_A V dc \propto P_R c^2 dt. \quad (\text{A3.4.44})$$

In the case of bimolecular gas-phase reactions, ‘encounters’ are simply collisions between two molecules in the framework of the general collision theory of gas-phase reactions (section A3.4.5.2). For a random thermal distribution of positions and momenta in an ideal gas reaction, the probabilistic reasoning has an exact foundation. However, as noted in the case of unimolecular reactions, in principle one must allow for deviations from this ideal behaviour and, thus, from the simple rate law, although in practice such deviations are rarely taken into account theoretically or established empirically.

The second-order rate law for bimolecular reactions is empirically well confirmed. Figure A3.4.3 shows the example of methyl radical recombination (equation (A3.4.36)) in a graphical representation following equation (A3.4.38) [22, 23 and 24]. For this example the bimolecular rate constant is

$$k = 4.4 \times 10^{-11} \text{ cm}^3 \text{ s}^{-1} \quad (\text{A3.4.45})$$

or

$$k = 2.6 \times 10^{13} \text{ cm}^3 \text{ mol}^{-1} \text{ s}^{-1}. \quad (\text{A3.4.46})$$

It is clear from figure A3.4.3 that the second-order law is well followed. However, in particular for recombination reactions at low pressures, a transition to a third-order rate law (second order in the recombining species and first order in some collision partner) must be considered. If the non-reactive collision partner M is present in excess and its concentration $[M]$ is time-independent, the rate law still is pseudo-second order with an effective second-order rate coefficient proportional to $[M]$.

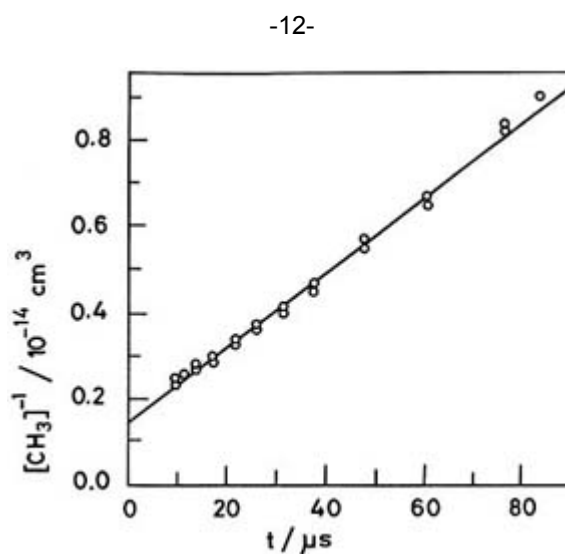


Figure A3.4.3. Methyl radical recombination as a second-order reaction (after [22, 23]).

A3.4.4.3 TRIMOLECULAR REACTIONS

Trimolecular reactions require the simultaneous encounter of three particles. At the usually low particle densities of gas phase reactions they are relatively unlikely. Examples for trimolecular reactions are atom recombination reactions



with the stoichiometric equation



In contrast to the bimolecular recombination of polyatomic radicals (equation (A3.4.34)) there is no long-lived intermediate AB^* since there are no extra intramolecular vibrational degrees of freedom to accommodate the excess energy. Therefore, the formation of the bond and the deactivation through collision with the inert collision partner M have to occur simultaneously (within 10–100 fs). The rate law for trimolecular recombination reactions of the type in equation (A3.4.47) is given by

$$-\frac{dc_A}{dt} = k[M]c_Ac_B \quad (\text{A3.4.49})$$

as can be derived by a probability argument similar to bimolecular reactions (and with similar limitations). Generally, collisions with different collision partners M_i may have quite different efficiencies. The rate law actually observed is therefore given by

$$-\frac{dc_A}{dt} = \sum_i k_i[M_i]c_Ac_B. \quad (\text{A3.4.50})$$

If the dominant contributions $k_i[M_i]$ are approximately constant, this leads to pseudo second-order kinetics with an effective rate constant

$$k_{\text{eff}} = \sum_i k_i[M_i]. \quad (\text{A3.4.51})$$

The recombination of oxygen atoms affords an instructive example:



with the common stoichiometric equation



Here $k_{\text{O}} \gg k_{\text{O}_2}$ because (A3.4.52) proceeds through a highly-excited molecular complex O_3^* with particularly efficient redistribution pathways for the excess energy. As long as $[\text{O}] \geq [\text{O}_2]$ the rate law for this trimolecular reaction is given by $(c(t) = [\text{O}], k = k_{\text{O}})$:

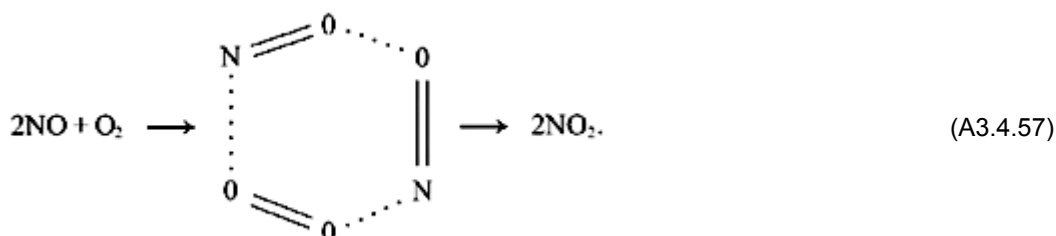
$$-\frac{1}{2} \frac{dc}{dt} = kc^3. \quad (\text{A3.4.55})$$

Integration leads to

$$\frac{1}{c(t)^2} = 4kt + \frac{1}{c(0)^2}. \quad (\text{A3.4.56})$$

Trimolecular reactions have also been discussed for molecular reactions postulating concerted reactions via cyclic intermediate complexes, for example

-14-



Empirically, one indeed finds a third-order rate law

$$-\frac{1}{2} \frac{d[\text{NO}]}{dt} = k[\text{NO}]^2[\text{O}_2]. \quad (\text{A3.4.58})$$

However, the postulated trimolecular mechanism is highly questionable. The third-order rate law would also be consistent with mechanisms arising from consecutive bimolecular elementary reactions, such as



or





In fact, the bimolecular mechanisms are generally more likely. Even the atom recombination reactions sometimes follow a mechanism consisting of a sequence of bimolecular reactions



This so-called complex mechanism has occasionally been proven to apply [25, 26].

A3.4.5 THEORY OF ELEMENTARY GAS-PHASE REACTIONS

A3.4.5.1 GENERAL THEORY

The foundations of the modern theory of elementary gas-phase reactions lie in the time-dependent molecular quantum dynamics and molecular scattering theory, which provides the link between time-dependent quantum dynamics and chemical kinetics (see also [chapter A3.11](#)). A brief outline of the steps in the development is as follows [27].

We start from the time-dependent Schrödinger equation for the state function (wave function $\Psi(t)$) of the reactive molecular system with Hamiltonian operator \hat{H} :

$$i\hbar \frac{\partial \Psi(t)}{\partial t} = \hat{H} \Psi(t). \quad (\text{A3.4.65})$$

Its solution can be written in terms of the time evolution operator \hat{U}

$$\Psi(t) = \hat{U}(t, t_0) \Psi(t_0) \quad (\text{A3.4.66})$$

which satisfies a similar differential equation

$$i\hbar \frac{\partial \hat{U}}{\partial t} = \hat{H} \hat{U}. \quad (\text{A3.4.67})$$

For time-independent Hamiltonians we have

$$\hat{U}(t, t_0) = \exp[-i\hat{H}(t - t_0)/\hbar]. \quad (\text{A3.4.68})$$

For strictly monomolecular processes the general theory would now proceed by analysing the time-dependent

wavefunction as a function of space (and perhaps spin) coordinates $\{q_i\}$ of the particles in terms of time-dependent probability densities.

$$P(\{q_i\}, t) = |\Psi(\{q_i\}, t)|^2 \quad (\text{A3.4.69})$$

which are integrated over appropriate regions of coordinate space assigned to reactants and products. These time-dependent probabilities can be associated with time-dependent concentrations, reaction rates and, if applicable, rate coefficients.

-16-

For thermal unimolecular reactions with bimolecular collisional activation steps and for bimolecular reactions, more specifically one takes the limit of the time evolution operator for $t_0 \rightarrow -\infty$ and $t \rightarrow +\infty$ to describe isolated binary collision events. The corresponding matrix representation of \hat{U} is called the scattering matrix or S-matrix with matrix elements

$$S_{fi} = U_{fi}(t \rightarrow +\infty, t_0 \rightarrow -\infty). \quad (\text{A3.4.70})$$

The physical interpretation of the scattering matrix elements is best understood in terms of its square modulus

$$P_{fi} = |S_{fi}|^2 \quad (\text{A3.4.71})$$

which is the transition probability between an initial fully specified quantum state $|i\rangle$ before the collision and a final quantum state $|f\rangle$ after the collision.

In a third step the S-matrix is related to state-selected reaction cross sections σ_{fi} , in principle observable in beam scattering experiments [28, 29, 30, 31, 32, 33, 34 and 35], by the fundamental equation of scattering theory

$$\sigma_{fi} = \frac{\pi}{k_i^2} |\delta_{fi} - S_{fi}|^2. \quad (\text{A3.4.72})$$

Here $\delta_{fi} = 1(0)$ is the Kronecker delta for $f=i$ ($f \neq i$) and k_i is the wavenumber for the collision, related to the initial relative centre of mass translational energy $E_{t,i}$ before the collision

$$k_i = \hbar^{-1} \sqrt{2\mu E_{t,i}} \quad (\text{A3.4.73})$$

with reduced mass μ for the collision partners of mass m_A and m_B :

$$\mu = \frac{m_A m_B}{m_A + m_B}. \quad (\text{A3.4.74})$$

Actually equation (A3.4.72) for σ_{fi} is still formal, as practically observable cross sections, even at the highest quantum state resolution usually available in molecular scattering, correspond to certain sums and averages of

the individual σ_{fi} . We use capital indices for such coarse-grained state-selected cross sections

$$\sigma_{FI} = \langle \sigma_{fi} \rangle. \quad (\text{A3.4.75})$$

-17-

In a fourth step the cross section is related to a state-selected specific bimolecular rate coefficient

$$k_{FI}(E_I) = \sigma_{FI}(E_{t,I}) \sqrt{2E_{t,I}/\mu}. \quad (\text{A3.4.76})$$

This rate coefficient can be averaged in a fifth step over a translational energy distribution $P(E_I)$ appropriate for the bulk experiment. In principle, any distribution $P(E_I)$ as applicable in the experiment can be introduced at this point. If this distribution is a thermal Maxwell–Boltzmann distribution one obtains a partially state-selected thermal rate coefficient

$$k_{FI}(T) = \left(\frac{8k_{\text{B}}T}{\pi\mu} \right)^{1/2} \int_0^\infty \left(\frac{E_{t,I}}{k_{\text{B}}T} \right) \sigma_{FI}(E_{t,I}) \exp \left\{ -\frac{E_{t,I}}{k_{\text{B}}T} \right\} \left(\frac{dE_{t,I}}{k_{\text{B}}T} \right). \quad (\text{A3.4.77})$$

In a final, sixth step one may also average (sum) over a thermal (or other) quantum state distribution I (and F) and obtain the usual thermal rate coefficient

$$k(T) = \langle k_{FI}(T) \rangle. \quad (\text{A3.4.78})$$

[Figure A3.4.4](#) summarizes these steps in one scheme. Different theories of elementary reactions represent different degrees of approximations to certain averages, which are observed in experiments.

There are two different aspects to these approximations. One consists in the approximate treatment of the underlying many-body quantum dynamics; the other, in the statistical approach to observable average quantities. An exhaustive discussion of different approaches would go beyond the scope of this introduction. Some of the most important aspects are discussed in separate chapters (see [chapter A3.7](#), [chapter A3.11](#), [chapter A3.12](#), [chapter A3.13](#)).

Here, we shall concentrate on basic approaches which lie at the foundations of the most widely used models. Simplified collision theories for bimolecular reactions are frequently used for the interpretation of experimental gas-phase kinetic data. The general transition state theory of elementary reactions forms the starting point of many more elaborate versions of quasi-equilibrium theories of chemical reaction kinetics [[27](#), [36](#), [37](#) and [38](#)].

-18-

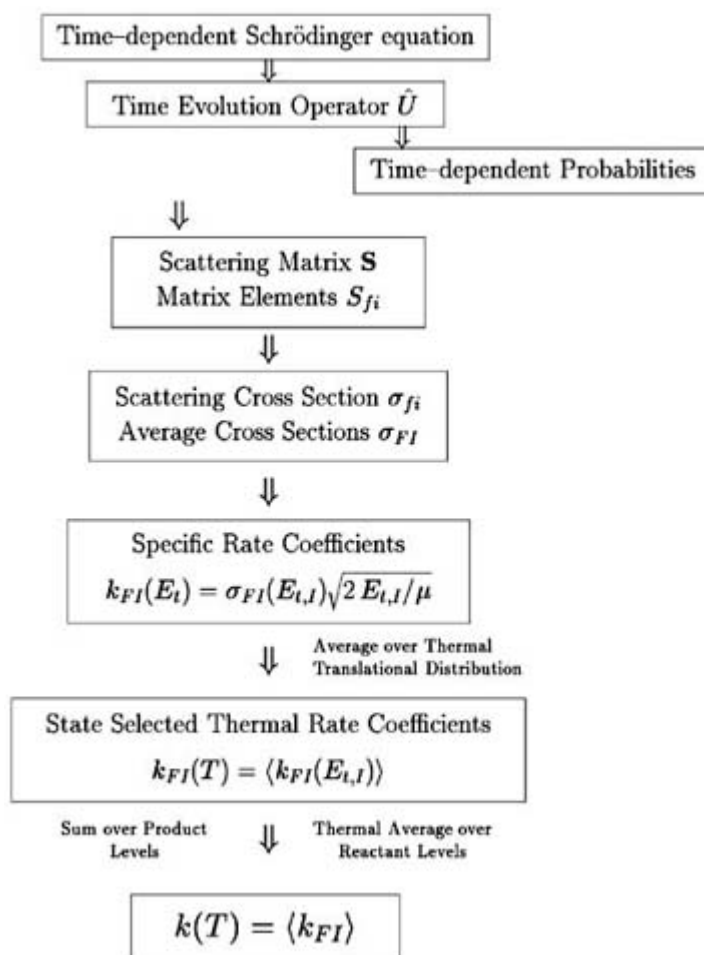


Figure A3.4.4. Steps in the general theory of chemical reactions.

In practice, one of the most important aspects of interpreting experimental kinetic data in terms of model parameters concerns the temperature dependence of rate constants. It can often be described phenomenologically by the Arrhenius equation [39, 40 and 41]

$$k(T) = A(T) \exp\{-E_A(T)/RT\} \quad (\text{A3.4.79})$$

where the *pre-exponential Arrhenius factor* A and the *Arrhenius activation energy* E_A generally depend on the temperature. R is the gas constant. This leads to the *definition* of the Arrhenius parameters:

$$E_A(T) \stackrel{\text{def}}{=} RT^2 \frac{d \ln(k(T))}{dT} \quad (\text{A3.4.80})$$

-19-

$$A(T) \stackrel{\text{def}}{=} k(T) \exp\left\{\frac{E_A(T)}{RT}\right\}. \quad (\text{A3.4.81})$$

The usefulness of these definitions is related to the usually weak temperature dependence of E_A and A . In the simplest models they are constant, whereas $k(T)$ shows a very strong temperature dependence.

A3.4.5.2 SIMPLE COLLISION THEORIES OF BIMOLECULAR REACTIONS

A bimolecular reaction can be regarded as a reactive collision with a reaction cross section σ that depends on the relative translational energy E_t of the reactant molecules A and B (masses m_A and m_B). The *specific rate constant* $k(E_t)$ can thus formally be written in terms of an effective reaction cross section σ , multiplied by the relative centre of mass velocity v_{rel}

$$k(E_t) = \sigma(E_t)v_{\text{rel}} = \sigma(E_t)\sqrt{2E_t/\mu}. \quad (\text{A3.4.82})$$

Simple collision theories neglect the internal quantum state dependence of σ . The rate constant as a function of temperature T results as a thermal average over the Maxwell–Boltzmann velocity distribution $p(E_t)$:

$$k(T) = \int_0^\infty p(E_t)k(E_t) dE_t = \langle v_{\text{rel}} \rangle \langle \sigma \rangle. \quad (\text{A3.4.83})$$

Here one has the thermal average centre of mass velocity

$$\langle v_{\text{rel}} \rangle = \sqrt{\frac{8k_B T}{\pi \mu}} \quad (\text{A3.4.84})$$

and the thermally averaged reaction cross section

$$\langle \sigma \rangle \stackrel{\text{def}}{=} \int_0^\infty \left(\frac{E_t}{k_B T} \right) \sigma(E_t) \exp \left\{ -\frac{E_t}{k_B T} \right\} \left(\frac{dE_t}{k_B T} \right). \quad (\text{A3.4.85})$$

We use the symbol k_B for Boltzmann's constant to distinguish it from the rate constant k . Equation (A3.4.85) defines the thermal average reaction cross section $\langle \sigma \rangle$.

In principle, the reaction cross section not only depends on the relative translational energy, but also on individual reactant and product quantum states. Its sole dependence on E_t in the simplified effective expression (equation (A3.4.82)) already implies unspecified averages over reactant states and sums over product states. For practical purposes it is therefore appropriate to consider simplified models for the energy dependence of the effective reaction cross section. They often form the basis for the interpretation of the temperature dependence of thermal cross sections. Figure A3.4.5 illustrates several cross section models.

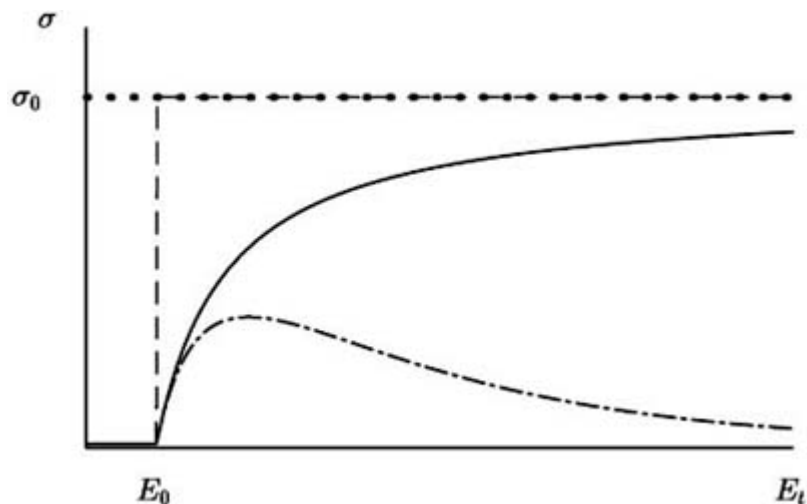


Figure A3.4.5. Simple models for effective collision cross sections σ : hard sphere without threshold (dotted line) hard sphere with threshold (dashed line) and hyperbolic threshold (full curve). E_t is the (translational) collision energy and E_0 is the threshold energy. σ_0 is the hard sphere collision cross section. The dashed--dotted curve is of the generalized type $\sigma_R(E_t > E_0) = \sigma_0 (1 - E_0/E_t) \exp[(1 - E_t/E_0)/(aE_0)]$ with the parameter $a = 3 E_0$.

(A) HARD SPHERE COLLISIONS

The reactants are considered as hard spheres with radii r_A and r_B , respectively. A (reactive) collision occurs on contact yielding a constant cross section σ_0 independent of the energy:

$$\sigma_0 = \pi(r_A + r_B)^2 \quad (\text{A3.4.86})$$

$$k(T) = \sigma_0 \sqrt{\frac{8k_B T}{\pi \mu}} \quad (\text{A3.4.87})$$

(B) CONSTANT CROSS SECTION WITH A THRESHOLD

The reaction can only occur once the collision energy reaches at least a value E_0 . The reaction cross section remains constant above this threshold:

$$\sigma = \begin{cases} 0 & \text{for } E_t < E_0 \\ \sigma_0 & \text{for } E_t \geq E_0 \end{cases} \quad (\text{A3.4.88})$$

$$k(T) = \sigma_0 \sqrt{\frac{8k_B T}{\pi \mu}} \left(1 + \frac{E_0}{k_B T}\right) \exp\left\{-\frac{E_0}{k_B T}\right\}. \quad (\text{A3.4.89})$$

(C) CROSS SECTION WITH A HYPERBOLIC THRESHOLD

Again, the reaction requires a minimum collision energy E_0 , but increases only gradually above the threshold towards a finite, high-energy limit σ_0 :

$$\sigma = \begin{cases} 0 & \text{for } E_t < E_0 \\ \sigma_0 \left(1 - \frac{E_0}{E_t}\right) & \text{for } E_t \geq E_0 \end{cases} \quad (\text{A3.4.90})$$

$$k(T) = \sigma_0 \sqrt{\frac{8k_B T}{\pi \mu}} \exp\left\{-\frac{E_0}{k_B T}\right\}. \quad (\text{A3.4.91})$$

(D) GENERALIZED COLLISION MODEL

The hyperbolic cross section model can be generalized further by introducing a function $f(\Delta E)$ ($\Delta E = E_t - E_0$) to describe the reaction cross section above a threshold:

$$\sigma = \begin{cases} 0 & \text{for } E_t < E_0 \\ \sigma_0 \left(1 - \frac{E_0}{E_t}\right) f(\Delta E) & \text{for } E_t \geq E_0 \end{cases} \quad (\text{A3.4.92})$$

$$k(T) = \sigma_0 \sqrt{\frac{8k_B T}{\pi \mu}} g(T) \exp\left\{-\frac{E_0}{k_B T}\right\} \quad (\text{A3.4.93})$$

$$g(T) = \int_0^\infty \frac{\Delta E}{k_B T} f(\Delta E) \exp\left\{-\frac{\Delta E}{k_B T}\right\} d\left(\frac{\Delta E}{k_B T}\right). \quad (\text{A3.4.94})$$

A3.4.6 TRANSITION STATE THEORY

Transition state theory or ‘activated complex theory’ has been one of the most fruitful approximations in reaction kinetics and has had a long and complex history [42, 43 and 44]. Transition state theory is originally based on the idea that reactant molecules have to pass a bottleneck to reach the product side and that they stay in quasi-equilibrium on the reactant side until this bottleneck is reached. The progress of a chemical reaction can often be described by the motion along a *reaction path* in the multidimensional molecular configuration space. Figure A3.4.6 shows typical potential energy profiles along such paths for uni- and bimolecular reactions. The effective potential energy $V(r_q)$ includes the zero point energy due to the motion orthogonal to the reaction coordinate r_q . The bottleneck is located at r_q^\ddagger , usually coinciding with an effective potential barrier, i.e. a first-order saddle point of the multidimensional potential hypersurface. Its height with respect to the reactants’ zero point level is E_0 . In its canonical form the transition state theory assumes a thermal

equilibrium between the reactant molecules A and molecules X moving in some infinitesimal range δ over the barrier towards the product side. For the unimolecular case this yields the equilibrium concentration:

$$[X] = \frac{q_X}{q_A} \exp\{-E_0/k_B T\}[A] \quad (\text{A3.4.95})$$

$$q_X = \frac{1}{2} q^\ddagger \delta \sqrt{2\pi\mu k_B T/h^2} \quad (\text{A3.4.96})$$

where h is Planck's constant. q stands for molecular partition functions referred to the corresponding zero point level. Thus q_A is the partition function for the reactant A. q^\ddagger is the restricted partition function for fixed reaction coordinate $r = r^\ddagger$ referring to the top of the effective (i.e. zero point corrected) barrier. It is often called the 'partition function of the transition state' bearing in mind that—in contrast to the X molecules—it does not correspond to any observable species. Rather, it defines the meaning of the purely technical term 'transition state'. Classically it corresponds to a $(3N - 7)$ -dimensional hypersurface in the $(3N - 6)$ -dimensional internal coordinate space of an N atomic system. The remainder of (A3.4.96) derives from the classical partition function for the motion in a one-dimensional box of length δ with an associated reduced mass μ . The factor of one half accounts for the fact that, in equilibrium, only half of the molecules located within $r^\ddagger \pm \delta/2$ move towards the product side.

-23-

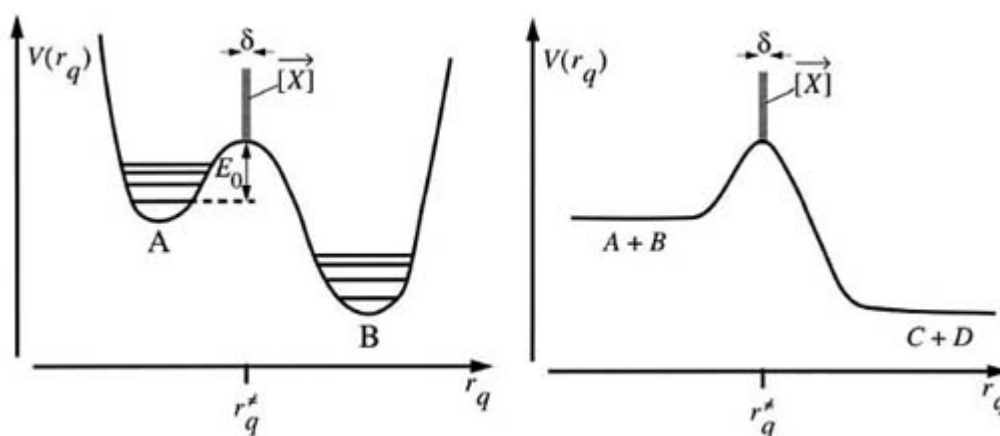


Figure A3.4.6. Potential energy along the reaction coordinate r_q for an unimolecular isomerization (left) and a bimolecular reaction (right). r_q^\ddagger is the location of the transition state at the saddle point. $[X]$ is the concentration of molecules located within δ of r_q^\ddagger moving from the reactant side to the product side (indicated by the arrow, which is omitted in the text).

Assuming a thermal one-dimensional velocity (Maxwell–Boltzmann) distribution with average velocity $\sqrt{2k_B T/\pi\mu}$ the reaction rate is given by the equilibrium flux if (1) the flux from the product side is neglected and (2) the thermal equilibrium is retained throughout the reaction:

$$-\frac{d[A]}{dt} = [X] \frac{\sqrt{2k_B T/\pi\mu}}{\delta} \quad (\text{A3.4.97})$$

Combining [equation \(A3.4.95\)](#), [equation \(A3.4.96\)](#) and [equation \(A3.4.97\)](#) one obtains the *first Eyring equation* for unimolecular rate constants:

$$k_{\text{uni}}(T) = \frac{k_{\text{B}}T}{h} \frac{q^{\ddagger}}{q_{\text{A}}} \exp\{-E_0/k_{\text{B}}T\}. \quad (\text{A3.4.98})$$

A completely analogous derivation leads to the rate coefficient for bimolecular reactions, where \tilde{q} are partition functions *per unit volume*:

$$k_{\text{bi}}(T) = \frac{k_{\text{B}}T}{h} \frac{\tilde{q}^{\ddagger}}{\tilde{q}_{\text{A}}\tilde{q}_{\text{B}}} \exp\{-E_0/k_{\text{B}}T\}. \quad (\text{A3.4.99})$$

In the high barrier limit, $E_0 \gg k_{\text{B}}T$, E_0 is approximately equal to the Arrhenius activation energy. The ratio of the partition functions is sometimes called the ‘statistical’ or ‘entropic’ factor. Its product with the ‘universal frequency factor’ $k_{\text{B}}T/h$ corresponds approximately to Arrhenius’ pre-exponential factor $A(T)$.

-24-

The quasi-equilibrium assumption in the above canonical form of the transition state theory usually gives an upper bound to the real rate constant. This is sometimes corrected for by multiplying [\(A3.4.98\)](#) and [\(A3.4.99\)](#) with a *transmission coefficient* $0 \leq \kappa \leq 1$.

In a *formal* analogy to the expressions for the thermodynamical quantities one can now *define* the standard *enthalpy* $\Delta^{\ddagger} H^{\ominus}$ and *entropy* $\Delta^{\ddagger} S^{\ominus}$ of activation. This leads to the *second Eyring equation*:

$$k(T) = \frac{k_{\text{B}}T}{h} \exp\{\Delta^{\ddagger} S^{\ominus}/R\} \exp\{-\Delta^{\ddagger} H^{\ominus}/RT\} \left(\frac{k_{\text{B}}T}{p^{\ominus}}\right)^j \quad (\text{A3.4.100})$$

where p^{\ominus} is the standard pressure of the ideal gas ($j=0$ for unimolecular and $j=1$ for bimolecular reactions). As a definition [\(A3.4.100\)](#) is strictly identical to [\(A3.4.98\)](#) and [\(A3.4.99\)](#) if considered as a theoretical equation. Since neither $\Delta^{\ddagger} S^{\ominus}$ nor $\Delta^{\ddagger} H^{\ominus}$ are connected to observable species, [equation \(A3.4.100\)](#) may also be taken as an empirical equation, *viz.* an alternative representation of Arrhenius’ equation ([equation \(A3.4.79\)](#)). In the field of thermochemical kinetics [43] one tries, however, to estimate $\Delta^{\ddagger} H^{\ominus}$ and $\Delta^{\ddagger} S^{\ominus}$ on the basis of molecular properties.

There is an immediate connection to the collision theory of bimolecular reactions. Introducing internal partition functions q_{int} , excluding the (separable) degrees of freedom for overall translation,

$$q = q_{\text{int}}q_{\text{trans}} \quad (\text{A3.4.101})$$

with

$$(\text{A3.4.102})$$

$$q_{\text{trans}} = V \left(\frac{2\pi M k_B T}{h^2} \right)^{3/2}$$

and comparing with [equation \(A3.4.83\)](#) the transition state theory expression for the effective thermal cross section of reaction becomes

$$\langle \sigma \rangle = \frac{h^2}{8\pi \mu_{AB} k_B T} \frac{q_{\text{int}}^\ddagger}{q_{\text{int},A} q_{\text{int},B}} \exp\{-E_0/k_B T\} \quad (\text{A3.4.103})$$

where V is the volume, $M = m_A + m_B$ is the total mass, and μ_{AB} is the reduced mass for the relative translation of A and B. One may interpret [equation \(A3.4.103\)](#) as the transition state version of the collision theory of bimolecular reactions: Transition state theory is used to calculate the thermally averaged reaction cross section to be inserted into [equation \(A3.4.83\)](#).

A3.4.7 STATISTICAL THEORIES BEYOND CANONICAL TRANSITION STATE THEORY

Transition state theory may be embedded in the more general framework of statistical theories of chemical reactions, as summarized in [figure A3.4.7](#) [27, 36]. Such theories have aimed at going beyond canonical transition state theory in several ways. The first extension concerns reaction systems with potential energy schemes depicted in [figure A3.4.8](#) (in analogy to [figure A3.4.6](#)), where one cannot identify a saddle point on the potential hypersurface to be related to a transition state. The left-hand diagram corresponds to a complex forming bimolecular reaction, and the right-hand to a direct barrierless bimolecular reaction. The individual sections (the left- and right-hand parts) of the left diagram correspond to the two unimolecular dissociation channels for the intermediate characterized by the potential minimum. These unimolecular dissociation channels correspond to simple bond fissions. The general types of reactions shown in [figure A3.4.8](#) are quite abundant in gas kinetics. Most ion molecule reactions as well as many radical-radical reactions are of this type. Thus, most of the very fast reactions in interstellar chemistry, atmospheric and combustion chemistry belong to this class of reaction, where standard canonical transition state theory cannot be applied and extension is clearly necessary. A second extension of interest would apply the fundamental ideas of transition state theory to state-selected reaction cross sections (see [section A3.4.5.1](#)). This theoretical program is carried out in the framework of phase space theory [45, 46] and of the statistical adiabatic channel model [27, 47], the latter being more general and containing phase space theory as a special case. In essence, the statistical adiabatic channel model is a completely state-selected version of the transition state theory. Here, the starting point is the \mathbf{S} -matrix element ([equation \(A3.4.104\)](#)), which in the statistical limit takes the statistically averaged form

$$\langle |S_{fi}|^2 \rangle_{F,I,\Delta E} = \begin{cases} W(E, J)^{-1} & \text{for strongly coupled channels} \\ \delta_{f,i} & \text{for weakly coupled channels} \end{cases} \quad (\text{A3.4.104})$$

where $W(E, J)$ is the total number of adiabatically open reaction channels for a given total angular momentum quantum number J (or any other good quantum number). $\langle \rangle_{F, I, \Delta E}$ refers to the averaging over groups of final and initial states ('coarse graining') and over suitably chosen collision energy intervals ΔE . Following the

lines of the general theory of reaction cross sections, [section A3.4.5.1](#), and starting from equation (A3.4.104) one can derive all the relevant kinetic specific reaction cross sections, specific rate constants and lifetimes in unimolecular reactions and the thermal rate constants analogous to transition state theory.

-26-

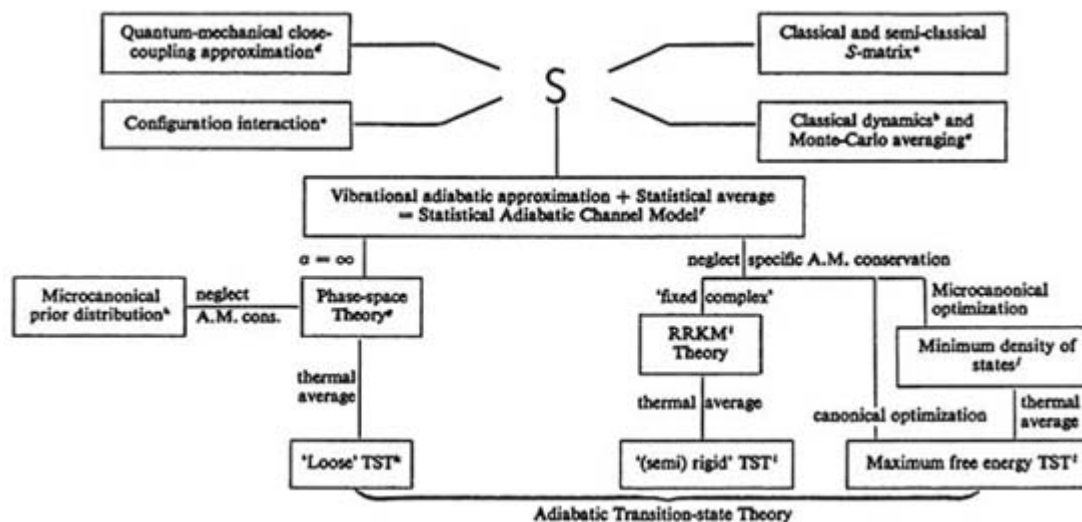


Figure A3.4.7. Summary of statistical theories of gas kinetics with emphasis on complex forming reactions (in the figure A.M. is the angular momentum, after Quack and Troe [27, 36, 74]). The indices refer to the following references: (a) [75, 76 and 77]; (b) [78]; (c) [79, 80 and 81]; (d) [82, 83, 84 and 85]; (e) [86, 87 and 88]; (f) [36, 37, 47, 89 and 90]; (g) [45, 46, 91]; (h) [92, 93, 94 and 95]; (i) [96, 97, 98, 99, 100, 101, 102, 103, 104 and 105]; (j) [106, 107, 108 and 109]; (k) [88, 94, 98, 99]; and (l) [94, 106, 107, 108, 109, 110, 111, 112].

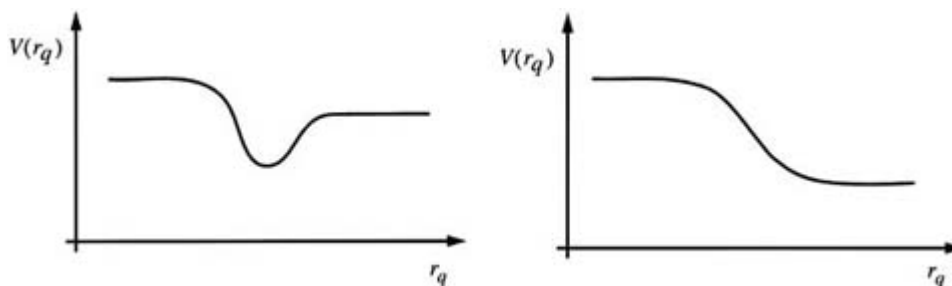


Figure A3.4.8. Potential energy profiles for reactions without barrier. Complex forming bimolecular reaction (left) and direct barrierless bimolecular reaction (right).

We summarize here only the main results of the theory and refer to a recent review [27] for details. The total number of adiabatically open channels is computed by searching for channel potential maxima $V_{a,\max}$. The channel potentials $V_a(r_q)$ are obtained by following the quantum energy levels of the reaction system along the reaction path r_q . An individual adiabatic channel connects an asymptotic scattering channel (corresponding to a reactant or to a product quantum level) with the reaction complex. One has the total number of open channels as a function of energy E , angular momentum J and other good quantum numbers:

$$W(E, J, \dots) = \sum_{a(J\dots)} h(E - V_{a,\max}). \quad (\text{A3.4.105})$$

Here $h(x)$ is the Heaviside step function with $h(x > 0) = 1$ and $h(x \geq 0) = 0$ (not to be confused with Planck's constant). The limit $a(J \dots)$ indicates that the summation is restricted to channel potentials with a given set of good quantum numbers ($J \dots$).

A state-to-state integral reaction cross section from reactant level a to product level b takes the form

$$\sigma_{ba} = \frac{\pi}{g_a k_a^2} \sum_{J=0}^{\infty} (2J+1) \frac{W(E, J, a)W(E, J, b)}{W(E, J)}. \quad (\text{A3.4.106})$$

Here the levels consist of several states. g_a is the reactant level degeneracy and k_a is the collision wavenumber (see [equation \(A3.4.73\)](#)).

A specific unimolecular rate constant for the decay of a highly excited molecule at energy E and angular momentum J takes the form

$$k(E, J, \dots) = \gamma \frac{W(E, J, \dots)}{h\rho(E, J, \dots)} \quad (\text{A3.4.107})$$

where γ is a dimensionless transmission coefficient (usually $0 \leq \gamma \leq 1$) and $\rho(E, J, \dots)$ is the density of molecular states. These expressions are relevant in the theory of thermal and non-thermal unimolecular reactions and are generalizations of the Rice–Ramsperger–Kassel–Marcus (RRKM) theory (see [chapter A3.12](#)).

Finally, the generalization of the partition function q^\ddagger in transition state theory ([equation \(A3.4.96\)](#)) is given by

$$Q_{\text{int}}^* = \sum_a \exp(-V_{a,\text{max}}/k_B T) = \int_0^\infty W(E) \exp(-E/k_B T) \left(\frac{dE}{k_B T} \right) \quad (\text{A3.4.108})$$

with the total number of open channels

$$W(E) = \sum_a \sum_{J=0}^{\infty} (2J+1)W(E, J, a). \quad (\text{A3.4.109})$$

These equations lead to forms for the thermal rate constants that are perfectly similar to transition state theory, although the computations of the partition functions are different in detail. As described in [figure A3.4.7](#) various levels of the theory can be derived by successive approximations in this general state-selected form of the transition state theory in the framework of the statistical adiabatic channel model. We refer to the literature cited in the diagram for details.

It may be useful to mention here one currently widely applied approximation for barrierless reactions, which is now frequently called microcanonical and canonical variational transition state theory (equivalent to the ‘minimum density of states’ and ‘maximum free energy’ transition state theory in [figure A3.4.7](#) . This type of theory can be understood by considering the partition functions $Q(r_q)$ as functions of r_q similar to [equation \(A3.4.108\)](#) but with $V_a(r_q)$ instead of $V_{a,\max}$. Obviously $Q(r_q) \geq Q^*$ so that the best possible choice for a transition state results from minimizing the partition function along the reaction coordinate r_q :

$$Q^\ddagger(T) = \min_{r_q} Q(r_q, T) = Q(r_q^\ddagger, T). \quad (\text{A3.4.110})$$

Equation (A3.4.110) represents the canonical form ($T = \text{constant}$) of the ‘variational’ theory. Minimization at constant energy yields the analogous microcanonical version. It is clear that, in general, this is only an approximation to the general theory, although this point has sometimes been overlooked. One may also define a free energy

$$A(r_q) = -k_B T \ln Q(r_q) \quad (\text{A3.4.111})$$

which leads to a maximum free energy condition

$$A^\ddagger(T) = \max_{r_q} A(r_q, T) = A(r_q^\ddagger, T). \quad (\text{A3.4.112})$$

The free energy as a function of reaction coordinates has been explicitly represented by Quack and Troe [[36, 112](#)] for the reaction



but the general concept goes back to Eyring (see [[27, 36](#)]).

A3.4.8 GAS-PHASE REACTION MECHANISMS

The kinetics of a system of elementary reactions forming a reaction mechanism are described by a system of coupled differential equations. Disregarding transport processes there is one differential equation for each species involved. Few examples for these systems of coupled differential equations can be solved exactly in closed form. The accurate solution more generally requires integration by numerical methods. In the simplest case of reversible elementary reactions the stoichiometry is sufficient to decouple the differential equations leading to simple rate laws. For more complicated compound reaction mechanisms this can only be achieved with more or less far reaching approximations, usually concerning reactive intermediates. The most important are *quasi-equilibrium* (or partial equilibrium) and the *quasi-stationarity* (or quasi-steady-state), whose practical importance goes far beyond gas-phase kinetics.

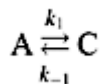
A3.4.8.1 ELEMENTARY REACTIONS WITH BACK-REACTION

The simplest possible gas-phase reaction mechanisms consist of an elementary reaction and its back reaction.

Here we consider uni- and bimolecular reactions yielding three different combinations. The resulting rate laws can all be integrated in closed form.

(A) UNIMOLECULAR REACTIONS WITH UNIMOLECULAR BACK REACTION

The equation



is the elementary mechanism of reversible isomerization reactions, for example



The rate law is given by

$$-\frac{dc_A}{dt} = k_1 c_A - k_{-1} c_C. \quad (\text{A3.4.115})$$

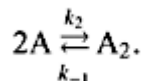
Exploiting the stoichiometric equation one can eliminate c_C . Integration yields the simple relaxation of the initial concentrations into the equilibrium, $c_A^{\text{eq}} = (c_A(\infty))$, with a relaxation time τ :

$$c_A(t) - c_A^{\text{eq}} = (c_A(t) - c_A^{\text{eq}}) \exp\{-t/\tau\} \quad (\text{A3.4.116})$$

$$\tau = \frac{1}{k_1 + k_{-1}}. \quad (\text{A3.4.117})$$

(B) BIMOLECULAR REACTIONS WITH UNIMOLECULAR BACK REACTION

For example



The rate law is given by

$$-\frac{1}{2} \frac{dc_A}{dt} = k_2 c_A^2 - k_{-1} c_{A_2}. \quad (\text{A3.4.118})$$

After transformation to the turnover variable $x = (c_A(0) - c_A(t))/2 = c_{A_2}(t) - c_{A_2}(0)$, integration yields

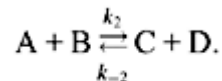
$$\ln\left(\frac{x+x_e-c_A(0)-K/4}{x-x_e}\right) - \ln\left(\frac{c_A(0)+K/4-x_e}{x_e}\right) = k_2(4c_A(0)+K-8x_e)t \quad (\text{A3.4.119})$$

$$x_e = \frac{c_A(0)}{2} + \frac{K}{8} - \left(\frac{K}{4}\left(c_{A_2}(0) + \frac{c_A(0)}{2}\right) + \left(\frac{K}{8}\right)^2\right)^{1/2} \quad (\text{A3.4.120})$$

where $K = k_2/k_{-1}$ is the equilibrium constant.

(C) BIMOLECULAR REACTIONS WITH BIMOLECULAR BACK-REACTION

For example



The rate law is given by

$$-\frac{dc_A}{dt} = k_2c_Ac_B - k_{-2}c_Cc_D \quad (\text{A3.4.121})$$

After transformation to the turnover variable $x = c_A(0) - c_A(t)$, integration yields

$$\ln\left(\frac{1 - [x/(a+b)]}{1 - [x/(a-b)]}\right) = 2k_2(1 - K^{-1})bt \quad (\text{A3.4.122})$$

$$a = \frac{c_A(0) + c_B(0) + K^{-1}[c_C(0) + c_D(0)]}{2(1 - K^{-1})} \quad (\text{A3.4.123})$$

$$b = \left(a^2 - \frac{c_A(0)c_B(0) - K^{-1}c_C(0)c_D(0)}{1 - K^{-1}}\right)^{1/2} \quad (\text{A3.4.124})$$

where $K = k_2/k_{-2}$ is the equilibrium constant.

Bimolecular steps involving identical species yield correspondingly simpler expressions.

A3.4.8.2 THE LINDEMANN-HINSHELWOOD MECHANISM FOR UNIMOLECULAR REACTIONS

The system of coupled differential equations that result from a compound reaction mechanism consists of several different (reversible) elementary steps. The kinetics are described by a system of coupled differential equations rather than a single rate law. This system can sometimes be decoupled by assuming that the concentrations of the intermediate species are small and quasi-stationary. The *Lindemann mechanism of thermal unimolecular reactions* [18, 19] affords an instructive example for the application of such approximations. This mechanism is based on the idea that a molecule A has to pick up sufficient energy

before it can undergo a monomolecular reaction, for example, bond breaking or isomerization. In thermal reactions this energy is provided by collisions with other molecules M in the gas to produce excited species A*:



Two important points must be noted here.

- (1) The collision partners may be any molecule present in the reaction mixture, i.e., inert bath gas molecules, but also reactant or product species. The activation (k_a) and deactivation (k_d) rate constants in equation (A3.4.125) therefore represent the effective average rate constants.
- (1) The collision (k_a, k_d) and reaction (k_r) efficiencies may significantly differ between different excited reactant states. This is essentially neglected in the Lindemann–Hinshelwood mechanism. In particular, the *strong collision* assumption implies that so much energy is transferred in a collision that the collision efficiency can be regarded as effectively independent of the energy.

With $k_1 = k_a[M]$ and $k_{-1} = k_d[M]$ the resulting system of differential equations is

$$-\frac{d[A]}{dt} = k_1[A] - k_{-1}[A^*] \quad (\text{A3.4.127})$$

$$-\frac{d[A^*]}{dt} = -k_1[A] + k_r[A^*] + k_{-1}[A^*] \quad (\text{A3.4.128})$$

$$\frac{d[\text{products}]}{dt} = k_r[A^*]. \quad (\text{A3.4.129})$$

If the excitation energy required to form activated species A* is much larger than $k_B T$ its concentration will remain small. This is fulfilled if $k_a \ll k_d$. Following Bodenstein, [A*] is then assumed to be quasi-stationary, i.e. after some initialization phase the concentration of activated species remains approximately constant (strictly speaking the ratio [A*]/[A] remains approximately constant (see [section A3.4.8.3](#))):

$$-\frac{d[A^*]}{dt} \approx 0 \quad (\text{A3.4.130})$$

$$\Rightarrow [A^*]_{\text{QS}} = \frac{k_1[A]}{k_{-1} + k_r} \quad (\text{A3.4.131})$$

This yields the quasi-stationary reaction rate with an effective unimolecular rate constant

$$v_c = \frac{d[\text{products}]}{dt} = k_{\text{eff}}[A] = k_r[A^*]_{\text{QS}} \quad (\text{A3.4.132})$$

$$k_{\text{eff}} = \frac{k_1 k_r}{k_{-1} + k_r} = \frac{k_a[M]k_r}{k_d[M] + k_r}. \quad (\text{A3.4.133})$$

The effective rate law correctly describes the pressure dependence of unimolecular reaction rates at least qualitatively. This is illustrated in [figure A3.4.9](#). In the limit of high pressures, i.e. large $[M]$, k_{eff} becomes independent of $[M]$ yielding the high-pressure rate constant k_{∞} of an effective first-order rate law. At very low pressures, product formation becomes much faster than deactivation. k_{eff} now depends linearly on $[M]$. This corresponds to an effective second-order rate law with the pseudo first-order rate constant k_0 :

$$k_{\infty} = \frac{k_a}{k_d} k_r \quad (\text{A3.4.134})$$

$$k_0 = k_a[M]. \quad (\text{A3.4.135})$$

In addition to $[A^*]$ being quasi-stationary the *quasi-equilibrium* approximation assumes a virtually unperturbed equilibrium between activation and deactivation ([equation \(A3.4.125\)](#)):

$$\frac{[A^*]}{[A]} = \frac{k_a}{k_d}. \quad (\text{A3.4.136})$$

-33-

This approximation is generally valid if $k_r \ll k_{-1}$. For the Lindemann mechanism of unimolecular reactions this corresponds to the high-pressure limit $k_{\text{eff}} = k_{\infty}$.

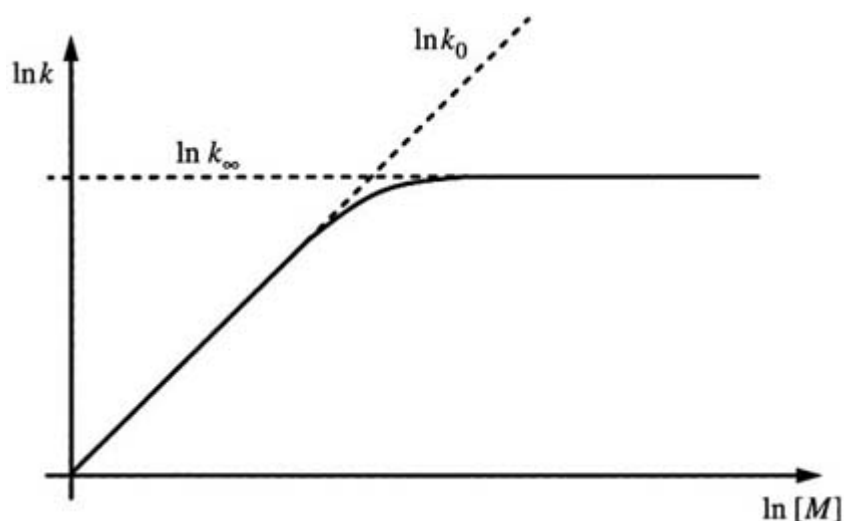


Figure A3.4.9. Pressure dependence of the effective unimolecular rate constant. Schematic fall-off curve for the Lindemann–Hinshelwood mechanism. k_{∞} is the (constant) high-pressure limit of the effective rate constant

k_{eff} and k_0 is the low-pressure limit, which depends linearly on the concentration of the inert collision partner [M].

The approximate results can be compared with the long time limit of the exact stationary state solution derived in section A3.4.8.3:

$$k_{\text{eff}} = \frac{1}{2} \{k_1 + k_{-1} + k_r - [(k_1 + k_{-1} + k_r)^2 - 4k_1k_r]^{1/2}\}. \quad (\text{A3.4.137})$$

This leads to the quasi-stationary rate constant of equation (A3.4.133) if $4k_1k_r \ll (k_1 + k_{-1} + k_r)^2$, which is more general than the Bodenstein condition $k_a \ll k_d$.

A3.4.8.3 GENERALIZED FIRST-ORDER KINETICS

The Lindemann mechanism for thermally activated unimolecular reactions is a simple example of a particular class of compound reaction mechanisms. They are mechanisms whose constituent reactions individually follow first-order rate laws [11, 20, 36, 48, 49, 50, 51, 52, 53, 54, 55 and 56]:



where N is the number of different species involved. With $c_i = [A_i]$ this leads to the following system of N coupled differential equations called *generalized first-order kinetics*:

-34-

$$-\frac{dc_i}{dt} = \sum_{j=1}^N K_{ij}c_j \quad i = 1, \dots, N. \quad (\text{A3.4.139})$$

The individual reactions need not be unimolecular. It can be shown that the relaxation kinetics after small perturbations of the equilibrium can always be reduced to the form of (A3.4.138) in terms of extension variables from equilibrium, even if the underlying reaction system is not of first order [51, 52, 57, 58].

Generalized first-order kinetics have been extensively reviewed in relation to technical chemical applications [59] and have been discussed in the context of copolymerization [53]. From a theoretical point of view, the general class of coupled kinetic equation (A3.4.138) and equation (A3.4.139) is important, because it allows for a general closed-form solution (in matrix form) [49]. Important applications include the Pauli master equation for statistical mechanical systems (in particular gas-phase statistical mechanical kinetics) [48] and the investigation of certain simple reaction systems [49, 50, 55]. It is the basis of the many-level treatment of thermal unimolecular reactions in terms of the appropriate master equations for energy transfer [36, 55, 60, 61, 62 and 63]. Generalized first-order kinetics also form the basis for certain statistical limiting cases of multiphoton induced chemical reactions and laser chemistry [54, 56].

Written in matrix notation, the system of first-order differential equations, (A3.4.139) takes the form

$$-\frac{d\mathbf{c}(t)}{dt} = \mathbf{K}\mathbf{c}(t) \quad (\text{A3.4.140})$$

With time independent matrix \mathbf{K} it has the general solution

$$\mathbf{c}(t) = \exp\{-\mathbf{K}t\}\mathbf{c}(0). \quad (\text{A3.4.141})$$

The exponential function of the matrix can be evaluated through the power series expansion of $\exp()$. \mathbf{c} is the column vector whose elements are the concentrations c_i . The matrix elements of the *rate coefficient matrix* \mathbf{K} are the first-order rate constants K_{ij} . The system is called *closed* if all reactions and back reactions are included. Then \mathbf{K} is of rank $N - 1$ with positive eigenvalues, of which exactly one is zero. It corresponds to the equilibrium state, with concentrations c_i^{eq} determined by the principle of microscopic reversibility:

$$\frac{K_{ij}}{K_{ji}} = \frac{c_i^{\text{eq}}}{c_j^{\text{eq}}}. \quad (\text{A3.4.142})$$

In this case \mathbf{K} is similar to a real symmetric matrix and equation (A3.4.141) can easily be solved by diagonalization of \mathbf{K} .

-35-

If some of the reactions of (A3.4.138) are neglected in (A3.4.139), the system is called open. This generally complicates the solution of (A3.4.141). In particular, the system no longer has a well defined equilibrium. However, as long as the eigenvalues of \mathbf{K} remain positive, the kinetics at long times will be dominated by the smallest eigenvalue. This corresponds to a stationary state solution.

As an example we take again the Lindemann mechanism of unimolecular reactions. The system of differential equations is given by equation (A3.4.127), equation (A3.4.128) and equation (A3.4.129). The rate coefficient matrix is

$$\mathbf{K} = \begin{pmatrix} k_1 & -k_{-1} & \vdots & 0 \\ -k_1 & k_{-1} + k_r & \vdots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & -k_r & & 0 \end{pmatrix}. \quad (\text{A3.4.143})$$

Since the back reaction, products $\rightarrow A^*$, has been neglected this is an open system. Still \mathbf{K} has a trivial zero eigenvalue corresponding to complete reaction, i.e. pure products. Therefore we only need to consider (A3.4.127) and (A3.4.128) and the corresponding (2×2) submatrix indicated in equation (A3.4.143).

The eigenvalues $\lambda_1 < \lambda_2$ of \mathbf{K} are both positive

$$\lambda_{1,2} = \frac{1}{2} \left\{ k_1 + k_{-1} + k_r \pm [(k_1 + k_{-1} + k_r)^2 - 4k_1k_r]^{1/2} \right\} > 0. \quad (\text{A3.4.144})$$

For long times, the smaller eigenvalue λ_1 will dominate (A3.4.141), yielding the stationary solution

$$\begin{pmatrix} c_A(t) \\ c_{A^*}(t) \end{pmatrix} = \exp\{-\lambda_1 t\} \begin{pmatrix} a \\ b \end{pmatrix} \quad (\text{A3.4.145})$$

where a and b are time-independent functions of the initial concentrations. With the condition $\lambda_1 \ll \lambda_2$ one obtains the effective unimolecular rate constant

$$k_{\text{eff}} = -\frac{d \ln(c_A + c_{A^*})}{dt} = \lambda_1 \stackrel{\lambda_1 \ll \lambda_2}{\approx} \frac{k_1 k_r}{k_1 + k_{-1} + k_r}. \quad (\text{A3.4.146})$$

For $k_a \ll k_d$ this is identical to the quasi-stationary result, [equation \(A3.4.133\)](#), although only the ratio $[A^*]/[A] = b/a$ ([equation \(A3.4.145\)](#)) is stationary and not $[A^*]$ itself. This suggests $d[A^*]/dt \ll d[A]/dt$ as a more appropriate formulation of quasi-stationarity. Furthermore, the general stationary state solution ([equation \(A3.4.144\)](#)) for the Lindemann mechanism contains cases that are not usually retained in the Bodenstein quasi-steady-state solution.

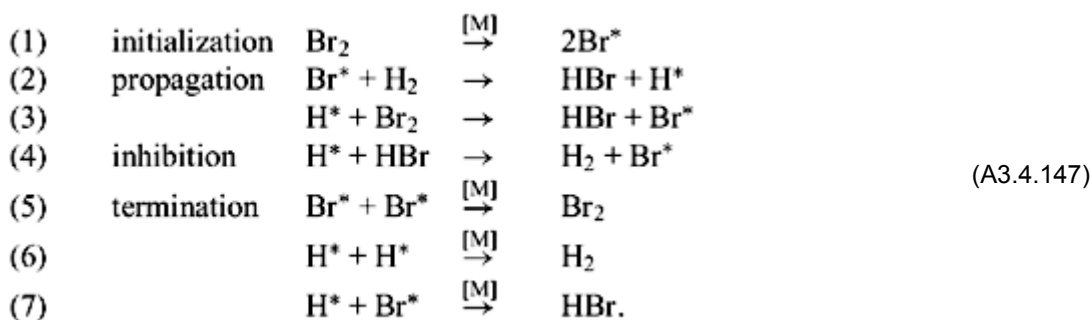
An important example for the application of general first-order kinetics in gas-phase reactions is the master equation treatment of the fall-off range of thermal unimolecular reactions to describe non-equilibrium effects in the weak collision limit when activation and deactivation cross sections ([equation \(A3.4.125\)](#)) are to be retained in detail [[60](#)].

General first-order kinetics also play an important role for the so-called *local eigenvalue analysis* of more complicated reaction mechanisms, which are usually described by nonlinear systems of differential equations. Linearization leads to effective general first-order kinetics whose analysis reveals information on the time scales of chemical reactions, species in steady states (quasi-stationarity), or partial equilibria (quasi-equilibrium) [[64](#), [65](#) and [66](#)].

A3.4.8.4 GENERAL COMPOUND REACTION MECHANISMS

More general compound reaction mechanisms lead to systems of differential equations of different orders. They can sometimes be treated by applying a quasi-stationarity or a quasi-equilibrium approximation. Often, this may even work for simple chain reactions. Chain reactions generally consist of four types of reaction steps: In the *chain initiation* steps, reactive species (radicals) are produced from stable species (reactants or catalysts). They react with stable species to form other reactive species in the *chain propagation*. Reactive species recovered in the chain propagation steps are called *chain carriers*. Propagation steps where one reactive species is replaced by another less-reactive species are sometimes termed *inhibiting*. *Chain branching* occurs if more than one reactive species are formed. Finally, the chain is *terminated* by reactions of reactive species, which yield stable species, for example through recombination in the gas phase or at the surface of the reaction vessel.

The assumption of quasi-stationarity can sometimes be justified if there is no significant chain branching, for example in HBr formation at 200–300°C:



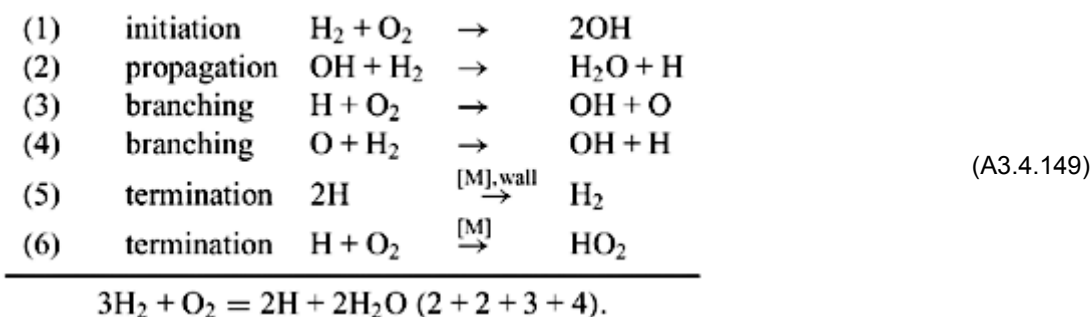
Chain carriers are indicated by an asterisk. Assuming quasi-stationarity for $[\text{H}^*]$ and $[\text{Br}^*]$ and neglecting (6) and (7) (because $[\text{H}^*] \ll [\text{Br}^*]$) yields

$$\frac{d[\text{HBr}]}{dt} = \frac{2k_2(k_1/k_5)^{1/2}[\text{H}_2][\text{Br}_2]^{1/2}}{1 + (k_4[\text{HBr}]/k_3[\text{Br}_2])}. \quad (\text{A3.4.148})$$

The resulting rate law agrees with the form found experimentally. Of course the postulated mechanism can only be proven by measuring the rate constants of the individual elementary steps separately and comparing calculated rates of equation (A3.4.148) with observed rates of HBr formation.

-37-

In general, the assumption of quasi-stationarity is difficult to justify *a priori*. There may be several possible choices of intermediates for which the assumption of quasi-stationary concentrations appears justified. It is possible to check for consistency, for example $[\text{A}^*]_{\text{QS}} \ll [\text{A}]$, but the final justification can only come from a comparison with the exact solution. These have usually to be obtained with numerical solvers for systems of differential equations [7, 8 and 9]. In particular, if transport phenomena with complicated boundary conditions must be taken into account this is the only viable solution. Modern fields of application include atmospheric chemistry and combustion chemistry [67, 68]. A classic example is the H_2/O_2 reaction. The mechanism includes more than 60 elementary steps and has been discussed in great detail [69]. A recent analysis of the explosion limits of this system in the range of 0.3–15.7 atm and 850–1040 K included 19 reversible elementary reactions [67]. Table A3.4.3 summarizes some of the major reactions for the hydrogen–oxygen reaction. A simplified mechanism involves only six reactions:



Reaction (5) proceeds mostly heterogeneously, reaction (6) mostly homogeneously. This mechanism can be integrated with simplifying assumptions to demonstrate the main features of gas-phase explosion kinetics [8].

The importance of numerical treatments, however, cannot be overemphasized in this context. Over the decades enormous progress has been made in the numerical treatment of differential equations of complex gas-phase reactions [8, 70, 71]. Complex reaction systems can also be seen in the context of nonlinear and self-organizing reactions, which are separate subjects in this encyclopedia (see [chapter A3.14](#), [chapter C3.6](#)).

-38-

Table A3.4.3. Rate constants for the reaction of H₂ with O₂ [73]. The rate constants are given in terms of the following expression: $k(T) = A(T/K)^b \exp(-E/RT)$.

Reaction	A (cm ³ mol ⁻¹ s ⁻¹)	b	E (kJ mol ⁻¹)
OH + H ₂ → H ₂ O + H	1.2 × 10 ⁹	1.3	15.2
H + H ₂ O → OH + H ₂	4.5 × 10 ⁹	1.3	78.7
H + O ₂ → OH + O	2.2 × 10 ¹⁴	0	70.4
OH + O → H + O ₂	1.0 × 10 ¹³	0	0
O + H ₂ → OH + H	1.8 × 10 ¹⁰	1.0	37.3
OH + H → O + H ₂	8.3 × 10 ⁹	1.0	29.1
OH + OH → H ₂ O + O	1.5 × 10 ⁹	1.14	0
O + H ₂ O → OH + OH	1.6 × 10 ¹⁰	1.14	72.4
H + HO ₂ → OH + OH	1.5 × 10 ¹⁴	0	4.2
H + HO ₂ → H ₂ + O ₂	2.5 × 10 ¹³	0	2.9
OH + HO ₂ → H ₂ O + O	1.5 × 10 ¹³	0	0
O + HO ₂ → OH + O ₂	2.0 × 10 ¹³	0	0
	cm ⁶ mol ⁻² s ⁻¹		
H + H + M → H ₂ + M	9.0 × 10 ¹⁶	-0.6	0
O + OH + M → HO ₂ + M	2.2 × 10 ²²	-2.0	0
H + O ₂ + M → HO ₂ + M	2.3 × 10 ¹⁸	-0.8	0

-39-

A3.4.9 SUMMARIZING OVERVIEW

Although the field of gas-phase kinetics remains full of challenges it has reached a certain degree of maturity. Many of the fundamental concepts of kinetics, in general take a particularly clear and rigorous form in gas-phase kinetics. The relation between fundamental quantum dynamical theory, empirical kinetic treatments, and experimental measurements, for example of combustion processes [72], is most clearly established in gas-phase kinetics. It is the aim of this article to review some of these most basic aspects. Details can be found in the sections on applications as well as in the literature cited.

REFERENCES

- [1] Molina M J and Rowland F S 1974 *Nature* **249** 810
- [2] Barker J R (ed) 1995 *Progress and Problems in Atmospheric Chemistry (Advanced Series in Physical Chemistry)* vol 3 (Singapore: World Scientific)
- [3] Crutzen P J 1995 Overview of tropospheric chemistry: developments during the past quarter century and a look ahead *Faraday Discuss.* **100** 1–21
- [4] Molina M J, Molina L T and Golden D M 1996 Environmental chemistry (gas and gas–solid interactions): the role of physical chemistry *J. Phys. Chem.* **100** 12 888
- [5] Crutzen P J 1996 Mein leben mit O₃, NO_x, etc. *Angew. Chem.* **108** 1878–98 (*Angew. Chem. Int. Ed. Engl.* **35** 1758–77)
- [6] Herbst E 1987 Gas phase chemical processes in molecular clouds *Interstellar Processes* ed D J Hollenbach and H A Tronson (Dordrecht: Reidel) pp 611–29
- [7] Gardiner W C Jr (ed) 1984 *Combustion Chemistry* (New York: Springer)
- [8] Warnatz J, Maas U and Dibble R W 1999 *Combustion: Physical and Chemical Fundamentals, Modelling and Simulation, Experiments, Pollutant Formation* 2nd edn (Heidelberg: Springer)
- [9] Gardiner W C Jr (ed) 2000 *Gas-Phase Combustion Chemistry* 2nd edn (Heidelberg: Springer)
- [10] Mills I, Cvitaš T, Homann K, Kallay N and Kuchitsu K 1993 *Quantities, Units and Symbols in Physical Chemistry* 2nd edn (Oxford: Blackwell) (3rd edn in preparation)
- [11] Quack M 1984 On the mechanism of reversible unimolecular reactions and the canonical ('high pressure') limit of the rate coefficient at low pressures *Ber. Bunsenges. Phys. Chem.* **88** 94–100
- [12] Herzberg G 1989 *Molecular Spectra and Molecular Structure. I. Spectra of Diatomic Molecules* (Malabar, FL: Krieger)
- [13] Ackermann M and Biauwe F 1979 *J. Mol. Spectrosc.* **35** 73
- [14] Cheung A S C, Yoshino K, Freeman D E, Friedman R S, Dalgarno A and Parkinson W H 1989 The Schumann–Runge absorption-bands of ¹⁶O¹⁸O in the wavelength region 175–205 nm and spectroscopic constants of isotopic oxygen molecules *J. Mol. Spectrosc.* **134** 362–89
- [15] Pine A S, Lafferty W J and Howard B J 1984 Vibrational predissociation, tunneling, and rotational saturation in the HF and DF dimers *J. Chem. Phys.* **81** 2939–50
- [16] Quack M and Suhm M A 1998 Spectroscopy and quantum dynamics of hydrogen fluoride clusters *Adv. in Mol. Vibr. Coll. Dyn.* vol 3 (JAI) pp 205–48

-40-

- [17] He Y, Müller H B, Quack M and Suhm M A 2000 *J. Chem. Phys.*
- [18] Lindemann F A 1922 Discussion on 'the radiation theory of chemical reactions' *Trans. Faraday Soc.* **17** 598–9
- [19] Hinshelwood C N 1933 *The Kinetics of Chemical Change in Gaseous Systems* 3rd edn (Oxford: Clarendon)
- [20] Quack M and Jans-Bürli S 1986 *Molekulare Thermodynamik und Kinetik. Teil 1: Chemische Reaktionskinetik* (Zürich: Fachvereine) (New English edition in preparation by D Luckhaus and M Quack)
- [21] Rosker M J, Rose T S and Zewail A 1988 Femtosecond real-time dynamics of photofragment–trapping resonances on dissociative potential-energy surfaces *Chem. Phys. Lett.* **146** 175–9
- [22] van den Bergh H E, Callear A B and Norström R J 1969 *Chem. Phys. Lett.* **4** 101–2
- [23] Callear A B and Metcalfe M P 1976 *Chem. Phys.* **14** 275

- [24] Glänzer K, Quack M and Troe J 1977 High temperature UV absorption and recombination of methyl radicals in shock waves *16th Int. Symp. on Combustion* (Pittsburgh: The Combustion Institute) pp 949–60
- [25] Hippler H, Luther K and Troe J 1974 On the role of complexes in the recombination of halogen atoms *Ber. Bunsenges. Phys. Chem.* **78** 178–9
- [26] van den Bergh H and Troe J 1975 NO-catalyzed recombination of iodine atoms. Elementary steps of the complex mechanism *Chem. Phys. Lett.* **31** 351–4
- [27] Quack M and Troe J 1998 Statistical adiabatic channel models *Encyclopedia of Computational Chemistry* ed P v R Schleyer *et al* (New York: Wiley) pp 2708–26
- [28] Faubel M and Toennies J P 1978 *Adv. Atom. Mol. Phys.* **13** 229
- [29] Zare R N 1979 Kinetics of state selected species *Faraday Discuss. Chem. Soc.* **67** 7–15
- [30] Bernstein R B and Zare R N 1980 State to state reaction dynamics *Phys. Today* **33** 43
- [31] Bernstein R B (ed) 1982 *Chemical Dynamics via Molecular Beam and Laser Techniques (The Hinshelwood Lectures, Oxford, 1980)* (Oxford: Oxford University Press)
- [32] Faubel M 1983 *Adv. Atom. Mol. Phys.* **19** 345
- [33] Polanyi J C 1987 *Science* **236** 680
- [34] Lee Y T 1987 *Science* **236** 793
- [35] Herschbach D R 1987 *Angew. Chem. Int. Ed. Engl.* **26** 1221
- [36] Quack M and Troe J 1977 Unimolecular reactions and energy transfer of highly excited molecules *Gas Kinetics and Energy Transfer* vol 2 (London: The Chemical Society)
- [37] Quack M and Troe J 1981 Statistical methods in scattering *Theoretical Chemistry: Advances and Perspectives* vol 6B (New York: Academic) pp 199–276
- [38] Truhlar D G, Garrett B C and Klippenstein S J 1996 Current status of transition-state theory *J. Phys. Chem.* **100** 12 771–800
- [39] Arrhenius S 1889 Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren *Z. Physik. Chem.* **4** 226–48
- [40] Arrhenius S 1899 Zur Theorie der chemischen Reaktionsgeschwindigkeiten *Z. Physik. Chem.* **28** 7–35
- [41] van't Hoff J H 1884 *Études de Dynamique Chimique* (Amsterdam: Müller)
- [42] Denbigh K 1981 *Principles of Chemical Equilibrium* 4th edn (London: Cambridge University Press)
- [43] Benson S W 1976 *Thermochemical Kinetics* 2nd edn (New York: Wiley)

-41-

- [44] Laidler K J and King M C 1983 The development of transition state theory *J. Phys. Chem.* **87** 2657–64
- [45] Pechukas P and Light J C 1965 *J. Chem. Phys.* **42** 3281–91
- [46] Nikitin E E 1965 *Theor. Exp. Chem.* **1** 144 (Engl. Transl.)
- [47] Quack M and Troe J 1974 Specific rate constants of unimolecular processes ii. Adiabatic channel model *Ber. Bunsenges. Phys. Chem.* **78** 240–52
- [48] Pauli W Jr 1928 Über das H-Theorem vom Anwachsen der Entropie vom Standpunkt der neuen Quantenmechanik *Probleme der modernen Physik* ed P Debye (Leipzig: Hirzel) pp 30–45
- [49] Jost W 1947 *Z. Naturf. a* **2** 159
- [50] Jost W 1950 *Z. Phys. Chem.* **195** 317
- [51] Eigen M 1954 *Faraday Discuss. Chem. Soc.* **17** 194
- [52] Eigen M and Schoen J 1955 *Ber. Bunsenges. Phys. Chem.* **59** 483
- [53] Horn F 1971 General first order kinetics *Ber. Bunsenges. Phys. Chem.* **75** 1191–201
- [54] Quack M 1979 Master equations for photochemistry with intense infrared light *Ber. Bunsenges. Phys. Chem.* **83** 757–75
- [55] Quack M and Troe J 1981 Current aspects of unimolecular reactions *Int. Rev. Phys. Chem.* **1** 97–147

- [56] Quack M 1998 Multiphoton excitation *Encyclopedia of Computational Chemistry* vol 3, ed P v R Schleyer et al (New York: Wiley) pp 1775–91
- [57] Castellan G W 1963 *Ber. Bunsenges. Phys. Chem.* **67** 898
- [58] Bernasconi C F (ed) 1976 *Relaxation Kinetics* (New York: Academic)
- [59] Wei J and Prater C D 1962 The structure and analysis of complex reaction systems *Advances in Catalysis* (New York: Academic) pp 203–392
- [60] Gilbert R G, Luther K and Troe J 1983 Theory of thermal unimolecular reactions in the fall-off range. II. Weak collision rate constants *Ber. Bunsenges. Phys. Chem.* **87** 169–77
- [61] Pilling M J 1996 *Ann. Rev. Phys. Chem.* **47** 81
- [62] Venkatesh P K, Dean A M, Cohen M H and Carr R W 1997 *J. Chem. Phys.* **107** 8904
- [63] Venkatesh P K, Dean A M, Cohen M H and Carr R W 1999 *J. Chem. Phys.* **111** 8313–29
- [64] Lam S H and Goussis D A 1988 Understanding complex chemical kinetics with computational singular perturbation *22nd Int. Symp. on Combustion* ed M C Salomony (Pittsburgh, PA: The Combustion Institute) pp 931–41
- [65] Maas U and Pope S B 1992 Simplifying chemical kinetics: intrinsic low-dimensional manifolds in composition space *Comb. Flame* **88** 239
- [66] Warnatz J, Maas U and Dibble R W 1999 *Combustion: Physical and Chemical Fundamentals, Modelling and Simulation, Experiments, Pollutant Formation* (Heidelberg: Springer)
- [67] Mueller M A, Yetter R A and Dryer F L 1999 Flow reactor studies and kinetic modelling of the H₂/O₂/NO_x reaction *Int. J. Chem. Kinet.* **31** 113–25
- [68] Mueller M A, Yetter R A and Dryer D L 1999 Flow reactor studies and kinetic modelling of the H₂/O₂/NO_x and CO/H₂O/O₂/NO_x reactions *Int. J. Chem. Kinet.* **31** 705–24
- [69] Dougherty E P and Rabitz H 1980 Computational kinetics and sensitivity analysis of hydrogen–oxygen combustion *J. Chem. Phys.* **72** 6571
-

- [70] Gear C W 1971 *Numerical Initial Value Problems in Ordinary Differential Equations* (Englewood Cliffs, NJ: Prentice-Hall)
- [71] Deuflhard P and Wulkow M 1989 Computational treatment of polyreaction kinetics by orthogonal polynomials of a discrete variable *Impact of Computing in Science and Engineering* vol 1
- [72] Ebert V, Schulz C, Volpp H R, Wolfrum J and Monkhouse P 1999 Laser diagnostics of combustion processes: from chemical dynamics to technical devices *Israel J. Chem.* **39** 1–24
- [73] Warnatz J 1979 *Ber. Bunsenges. Phys. Chem.* **83** 950
- [74] Quack M 1977 Detailed symmetry selection rules for reactive collisions *Mol. Phys.* **34** 477–504
- [75] Miller W H 1970 *J. Chem. Phys.* **53** 1949
- [76] Marcus R A 1971 *J. Chem. Phys.* **54** 3965
- [77] Miller W H 1975 *Adv. Chem. Phys.* **30** 77–136
- [78] Slater N B 1959 *Theory of Unimolecular Reactions* (Ithaca, NY: Cornell University Press)
- [79] Bunker D L 1971 *Methods in Computational Physics* vol 10 (New York: Academic) pp 287–325
- [80] Porter R N 1974 *Ann. Rev. Phys. Chem.* **25** 317–55
- [81] Polanyi J C and Schreiber J L 1973 *Physical Chemistry—An Advanced Treatise* vol 6 (New York: Academic)
- [82] Gordon R G 1971 *Methods in Computational Physics* vol 10 (London: Academic) p 82
- [83] Redmon M J and Micha D A 1974 *Chem. Phys. Lett.* **28** 341
- [84] Shapiro M 1972 *J. Chem. Phys.* **56** 2582
- [85] Micha D A 1973 *Acc. Chem. Res.* **6** 138–44
- [86] Mies F H and Krauss M 1966 *J. Chem. Phys.* **45** 4455–68

- [87] Mies F H 1968 *Phys. Rev.* **175** 164–75
- [88] Mies F H 1969 *J. Chem. Phys.* **51** 787–97
Mies F H 1969 *J. Chem. Phys.* **51** 798–807
- [89] Quack M and Troe J 1975 *Ber. Bunsenges. Phys. Chem.* **79** 170–83
- [90] Quack M and Troe J 1975 *Ber. Bunsenges. Phys. Chem.* **79** 469–75
- [91] White R A and Light J C 1971 Statistical theory of bimolecular exchange reactions: angular distribution *J. Chem. Phys.* **55** 379–87
- [92] Kinsey J L 1971 *J. Chem. Phys.* **54** 1206
- [93] Ben-Shaul A, Levine R D and Bernstein R B 1974 *J. Chem. Phys.* **61** 4937
- [94] Quack M and Troe J 1976 *Ber. Bunsenges. Phys. Chem.* **80** 1140
- [95] Levine R D and Bernstein R B (eds) 1989 *Molecular Reaction Dynamics and Chemical Reactivity* (Oxford: Oxford University Press)
- [96] Robinson P J and Holbrook K A 1972 *Unimolecular Reactions* (London: Wiley)
- [97] Forst W 1973 *Theory of Unimolecular Reactions* (New York: Academic)
- [98] Marcus R A and Rice O K J 1951 *Phys. Colloid Chem.* **55** 894–908
- [99] Marcus R A 1952 *J. Chem. Phys.* **20** 359–64
-

-43-

- [100] Rosenstock H M, Wallenstein M B, Wahrhaftig A L and Eyring H 1952 *Proc. Natl Acad. Sci. USA* **38** 667–78
- [101] Marcus R A 1965 *J. Chem. Phys.* **43** 2658
- [102] Pilling M J and Smith I W M (eds) 1987 *Modern Gas Kinetics. Theory, Experiment and Application* (Oxford: Blackwell)
- [103] Gilbert R G and Smith S C (eds) 1990 *Theory of Unimolecular and Recombination Reactions* (Oxford: Blackwell)
- [104] Holbrook K A, Pilling M J and Robertson S H (eds) 1996 *Unimolecular Reactions* 2nd edn (Chichester: Wiley)
- [105] Baer T and Hase W L (eds) 1996 *Unimolecular Reaction Dynamics* (Oxford: Oxford University Press)
- [106] Bunker D L and Pattengill M 1968 *J. Chem. Phys.* **48** 772–6
- [107] Wong W A and Marcus R A 1971 *J. Chem. Phys.* **55** 5625–9
- [108] Gaedtke H and Troe J 1973 *Ber. Bunsenges. Phys. Chem.* **77** 24–9
- [109] Garret D C and Truhlar D G 1979 *J. Chem. Phys.* **70** 1592–8
- [110] Glasstone S, Laidler K J and Eyring H 1941 *The Theory of Rate Processes* (New York: McGraw-Hill)
- [111] Wardlaw D M and Marcus R A 1988 *Adv. Chem. Phys.* **70** 231–63
- [112] Quack M and Troe J 1977 *Ber. Bunsenges. Phys. Chem.* **81** 329–37
-

FURTHER READING

Johnston H S 1966 *Gas Phase Reaction Rate Theory* (Ronald)

Nikitin E E 1974 *Theory of Elementary Atomic and Molecular Processes in Gases* (Oxford: Clarendon)

Quack M and Jans-Bürli S 1986 *Molekulare Thermodynamik und Kinetik. Teil 1. Chemische Reaktionskinetik* (Zürich: Fachvereine)

Pilling M J and Smith I W M (eds) 1987 *Modern Gas Kinetics. Theory, Experiment and Application* (Oxford: Blackwell)

Laidler K J 1987 *Chemical Kinetics* (New York: Harper Collins)

Gilbert R G and Smith S C 1990 *Theory of Unimolecular and Recombination Reactions* (Oxford: Blackwell)

Baer T and Hase W L 1996 *Unimolecular Reaction Dynamics* (Oxford: Oxford University Press)

Holbrook K A, Pilling M J and Robertson S H 1996 *Unimolecular Reactions* 2nd edn (Chichester: Wiley)

Warnatz J, Maas U and Dibble R W 1999 *Combustion: Physical and Chemical Fundamentals, Modelling and Simulation, Experiments, Pollutant Formation* (Heidelberg: Springer)

-1-

Ion chemistry A3.5

A A Viggiano and Thomas M Miller

A3.5.1 INTRODUCTION

Ion chemistry is a product of the 20th century. J J Thomson discovered the electron in 1897 and identified it as a constituent of all matter. Free positive ions (as distinct from ions deduced to exist in solids or electrolytes) were first produced by Thomson just before the turn of the century. He produced beams of light ions, and measured their mass-to-charge ratios, in the early 1900s, culminating in the discovery of two isotopes of neon in 1912 [1]. This year also marked Thomson's discovery of H_3^+ , which turns out to be the single most important astrophysical ion and which may be said to mark the beginning of the study of the *chemistry* of ions. Thomson noted that 'the existence of this substance is interesting from a chemical point of view', and the problem of its structure soon attracted the distinguished theorist Niels Bohr [2]. (In 1925, the specific reaction producing H_3^+ was recognized [2].) The mobilities of electrons and ions drifting in weak electric fields were first measured by Thomson, Rutherford and Townsend at the Cavendish Laboratory of Cambridge University in the closing years of the 19th century. The average mobility of the negative charge carrier was observed to increase dramatically in some gases, while the positive charge carrier mobility was unchanged—the *anomalous mobility problem*—which led to the hypothesis of electron attachment to molecules to form negative ions [3]. In 1936, Eyring, Hirschfelder and Taylor calculated the rate constant for an ion–molecule reaction (the production of H_3^+ !), showing it to be 100 times greater than for a typical neutral reaction, but it was not until 20 years later that any ion–molecule rate constant was measured experimentally [4]. Negative ion–molecule reactions were not studied at all until 1957 [5].

In this section, the wide diversity of techniques used to explore ion chemistry and ion structure will be outlined and a sampling of the applications of ion chemistry will be given in studies of lamps, lasers, plasma processing, ionospheres and interstellar clouds.

Note that chemists tend to refer to positive ions as *cations* (attracted to the cathode in electrolysis) and negative ions as *anions* (attracted to an anode). In this section of the encyclopedia, the terms *positive ion* and *negative ion* will be used for the sake of clarity.

A3.5.2 METHODOLOGIES

A3.5.2.1 SPECTROSCOPY

(A) ACTION SPECTROSCOPY

The term *action spectroscopy* refers to how a particular ‘action’, or process, depends on photon energy. For example, the photodissociation of O_2^- with UV light leads to energetic $\text{O}^- + \text{O}$ fragments; the kinetic energy released has been

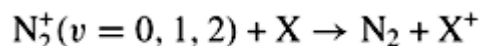
-2-

studied as a function of photon energy by Lavrich *et al* [6, 7]. Many of the processes discussed in this section may yield such an action spectrum and we will deal with the processes individually.

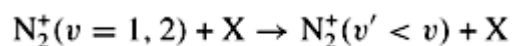
(B) LASER INDUCED FLUORESCENCE

Laser induced fluorescence (LIF) detection of molecules has served as a valuable tool in the study of gas-phase processes in combustion, in the atmosphere, and in plasmas [8, 9, 10, 11 and 12]. In the LIF technique, laser light is used to excite a particular level of an atom or molecule which then radiates (fluoresces) to some lower excited state or back to the ground state. It is the fluorescence photon which signifies detection of the target. Detection may be by measurement of the total fluorescence signal or by resolved fluorescence, in which the various rovibrational populations are separated. LIF is highly selective and may be used to detect molecules with densities as low as 10^5 cm^{-3} in low pressure situations ($<0.1 \text{ Pa}$) where collisional quenching is negligible. In the presence of an atmosphere of air, the detection limit is about 10^{10} cm^{-3} . The use of LIF for ions is more difficult than for neutrals because a typical ion number density may be orders of magnitude lower than for neutrals. Nevertheless, important LIF work with ions has been reported.

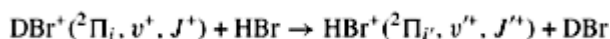
LIF has been used to study state-selected ion–atom and ion–molecule collisions in gas cells. Ar^+ reactions with N_2 and CO were investigated by Leone and colleagues in the 1980s [13, 14] and that group has continued to contribute new understanding of the drifting and reaction of ions in gases, including studies of velocity distributions and rotational alignment [15, 16, 17, 18 and 19]. The vibrational state dependence of the charge transfer reaction



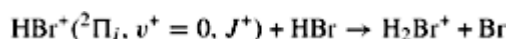
where $\text{X} = \text{Ar}$ or O_2 and the collisional deactivation reaction



were studied by this group using LIF [20]. They showed that charge transfer is enhanced by vibrational excitation and that vibrational deactivation is much more likely with O_2 than with Ar . We also consider here a reaction that displays both electron-transfer and proton-transfer channels,



and



studied via LIF of the $\text{A}^2\Sigma - \text{X}^2\Pi_{1/2,3/2}$ (0,0) bands of HBr^+ , using photons in the range 358–378 nm [21, 22]. For the electron transfer reaction, it was found that any excess energy in the process was statistically partitioned among all degrees of freedom of the complex and was manifested in the LIF spectra as rotational heating. Flow tube experiments tuned to different Br isotopes also showed a hydrogen-atom transfer channel in the $\text{HBr}^+ + \text{HBr}$ reaction.

LIF is also used with liquid and solid samples. For example, LIF is used to detect UO_2^{2+} ions in minerals; the uranyl ion is responsible for the bright green fluorescence given off by minerals such as autunite and opal upon exposure to UV light [23].

(C) PHOTODISSOCIATION OF IONS

Photodissociation of molecular ions occurs when a photon is absorbed by the ion and the energy is released (at least partly) by the breaking of one of the molecular bonds. The photodissociation of a molecular ion is conceptually similar to that for neutral molecules, but the experimental techniques differ. Photodissociation events are divided into two categories: *direct dissociation*, in which the photoexcitation is from a bound state to a repulsive state and *predissociation*, in which a quasi-bound state is accessed in the excitation. Direct dissociation takes place rapidly (fs to ps timescale). The shape of the direct dissociation cross section curve against photon energy is governed by the (Franck–Condon) overlap of wavefunctions of the initial state (usually the ground state) and the final, repulsive state. It will normally consist of peaks corresponding to vibrational structure in the initial level of the target ion with shapes skewed by the overlap with the repulsive state. One can model these shapes to obtain the potential curves of both the initial and repulsive states. Predissociation, in contrast, may take place over a much longer timescale; the lifetime of a particular predissociating state may be determined from the width of the resonance observed. Measurements of the lifetime for a series of predissociating states gives a picture of the predissociation mechanism. Photodissociation cross sections tend to peak in the 10^{-18} cm² range and hence are often given in Mb (megabarn) units.

There are many experimental methods by which photodissociation of ions have been studied. The earliest were crossed-beams experiments on H_2^+ beginning in the late 1960s [24, 25 and 26] and experiments on a variety of ions in the 1970s using drift tubes [27, 28 and 29]. Later techniques allowed more detailed information to be obtained on state symmetries and kinetic energy releases [30, 31 and 32]. [Figure A3.5.1](#) shows the fast ion beam photofragment spectrometer at SRI International; similar apparatus is in use at other institutions [33]. The apparatus consists of an ion source and mass selector, two electrostatic quadrupole benders that allow a laser beam to interact coaxially with the ion beam, a product-ion (photofragment) energy analyser and a particle detector. An interesting feature of the coaxial beam technique, aside from the long interaction region, is that sub-Doppler line widths can be obtained because of a thousandfold or more narrowing of the ion velocity distribution in the centre-of-mass reference frame for typical keV ion energies. By the same token, photofragment ions that differ in energy by a tenth of an electron volt in the centre-of-mass frame will be separated by typically 10–20 eV in the laboratory frame. This simplifies the job of the photofragment energy analyser. An example of a photofragment kinetic energy spectrum is shown in [figure A3.5.2](#). If the laser beam is sent at right angles to the ion beam (instead of coaxially), the optical polarization vector can be rotated to map out angular distributions of photofragments. (In the coaxial arrangement, the optical polarization is necessarily always perpendicular to the ion beam direction.)

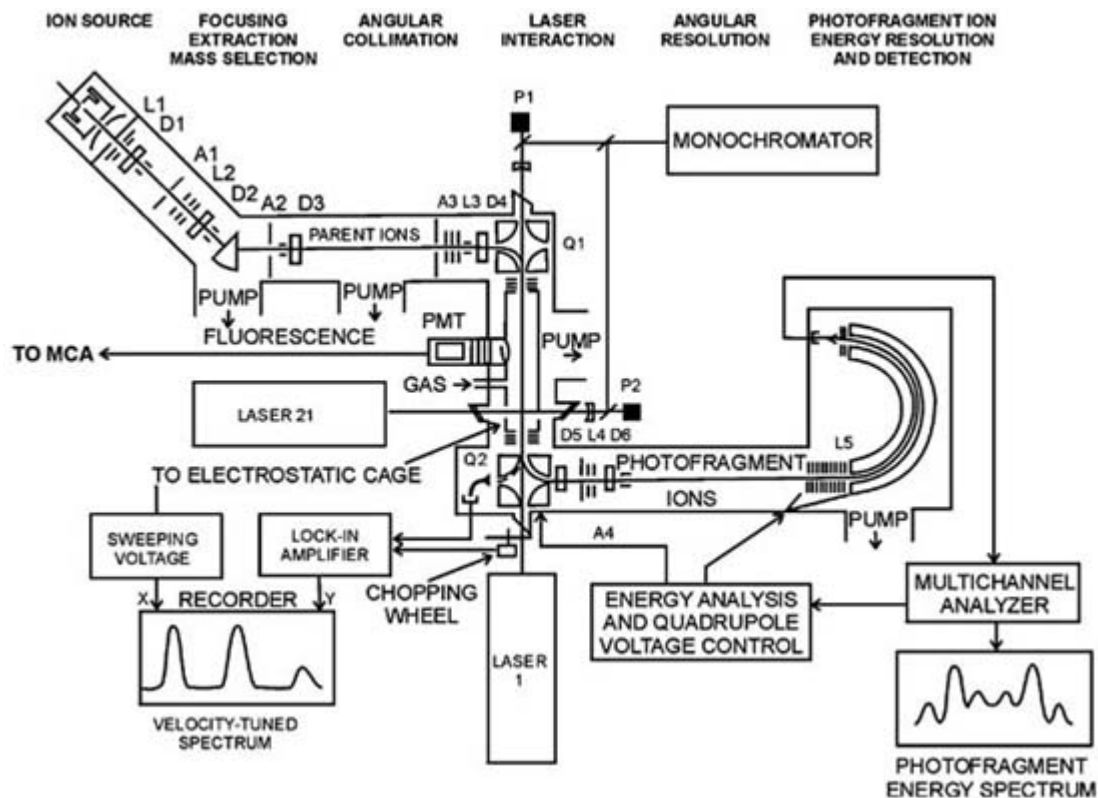


Figure A3.5.1. The fast ion beam photofragment spectrometer at SRI International. ‘L’ labels electrostatic lenses, ‘D’ labels deflectors and ‘A’ labels apertures.

-5-

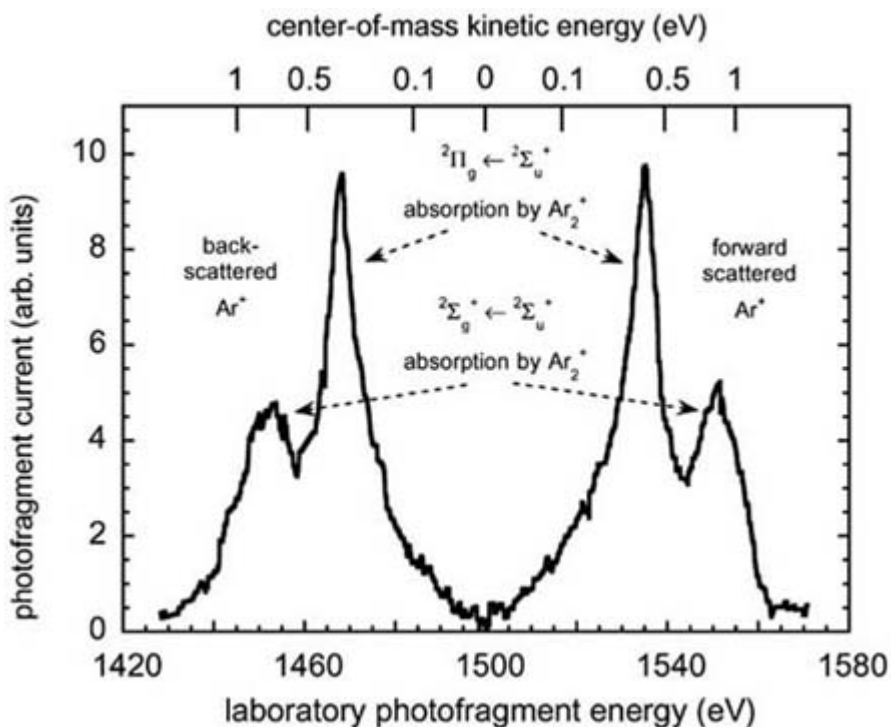


Figure A3.5.2. The Ar^+ photofragment energy spectrum for the dissociation of $3 \text{ keV } \text{Ar}_2^+$ ions at 752.5 nm . The upper scale gives the kinetic energy release in the centre-of-mass reference frame, both parallel and antiparallel to the ion beam velocity vector in the laboratory.

In the past decade there has been photodissociation work on doubly charged positive ions, e.g., N_2^{2+} , NO_2^{2+} , CF_3^{2+} , CCl_3^{2+} , SiF_2^{2+} and SiF_3^{2+} [34]. Interestingly, the result of photoexcitation of the latter four molecular ions is the loss of neutrals, as a consequence of the electronic structures. There are two possible scenarios, illustrated by



There has been much activity in the study of photodissociation of cluster ions, dating back to the 1970s when it was realized that most ions in the earth's lower atmosphere were heavily clustered [7, 35, 36].

-6-

(D) PHOTOELECTRON SPECTROSCOPY

Photoelectron spectroscopy (PES) of negative ions involves irradiation of an ion beam with laser light and energy analysis of the electrons liberated when the photon energy exceeds the binding energy of the electron. The kinetic energy of the detached electron is the difference between the photon energy and the binding energy of the electron [37, 38, 39, 40, 41, 42, 43, 44, 45, 46 and 47]. Analysis of the electron energy thus gives a direct measurement of the electron affinity of the corresponding neutral atom or molecule, a very important thermochemical quantity. Generally speaking, PES yields more information about the *neutral* atom or molecule than the corresponding negative ion, because the target ion is ideally in its ground state and the electron kinetic energy is then dependent on the final state of the neutral product. The energy resolution of a PES experiment is usually adequate (often 5–10 meV) to resolve vibrational structure due to the neutral molecule, certainly for low-mass systems of few atoms and likewise electronic structure, including singlet–triplet splittings and fine structure separations. In a few cases, rotational energy levels have been resolved. Features may appear in a photoelectron spectrum due to excited levels of the target negative ion and give valuable information about the structure of the negative ion, but at the cost of complicating the spectrum.

An example of a PES apparatus is shown in [figure A3.5.3](#). A PES apparatus consists of (a) an ion source, (b) a fixed-frequency laser, (c) an interaction region, (d) an electron energy analyser and (e) an electron detector. Ion sources include gas discharge, sputtering, electron-impact and flowing afterglow. The laser may be cw (the argon-ion laser operated at 488 nm is common) or pulsed (which allows frequency doubling etc.). Recent trends have been toward UV laser light because the negative ions of importance in practical chemistry (e.g., atmospheric chemistry and biochemistry) tend to be strongly bound and because the more energetic light allows one to access more electronic and vibrational states. The interaction region may include a magnetic field that routes detached electrons toward the energy analyser. The energy analyser is either a hemispherical electrostatic device or a time-of-flight energy analyser; the latter is especially suited to a pulsed-laser system.

-7-

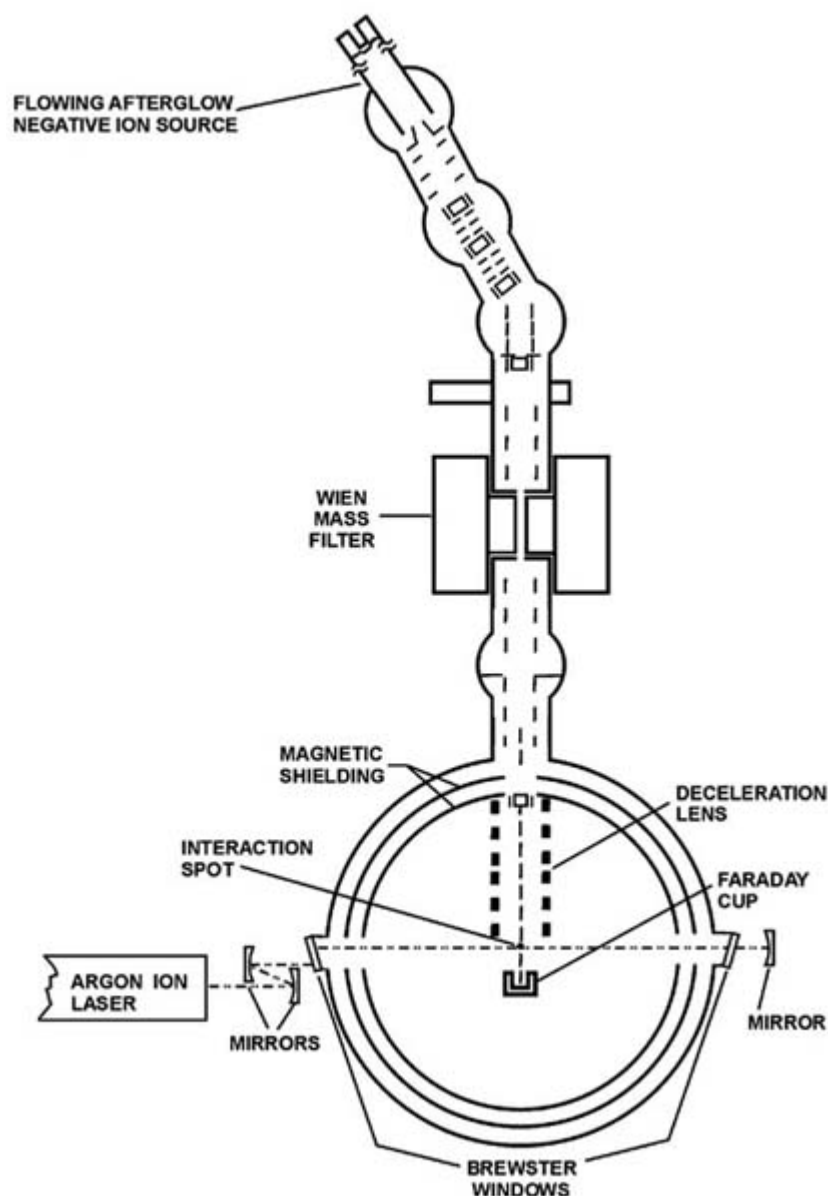


Figure A3.5.3. The negative ion photoelectron spectrometer used at the University of Colorado. The apparatus now contains a UV-buildup cavity inside the vacuum system (not shown in this sketch).

Data obtained with a PES apparatus are shown in [figure A3.5.4](#) [48]. Interpretation of the spectrum for a diatomic molecule is particularly straightforward. Peaks to the left of the origin band (each band containing unresolved rotational structure) are spaced by the vibrational separation in the neutral molecule, and their relative intensities are determined by the amount of spatial overlap between wavefunctions for the negative ion and the neutral molecule (Franck–Condon factors). Peaks to the right of the origin band are spaced by the vibrational separation in the negative ion and their relative intensities give the effective temperature of the ion source. Subtracting the electron kinetic energy

corresponding to the origin band from the photon energy yields the electron affinity of the molecule. The energy of the maximum of the envelope of neutral molecule peaks is referred to as the *vertical detachment energy*, i.e., the energy required from the ground vibrational state of the negative ion to the neutral molecule with no change in nuclear geometry. Photoelectron spectra are often far more complicated than the example shown, especially for polyatomic molecules. The origin band may have zero intensity, making the electron

affinity difficult to determine directly. Fine structure at least doubles the number of peaks in the spectrum. PES experiments have been carried out for doubly charged negative ions [49] and using multiphoton detachment [37].

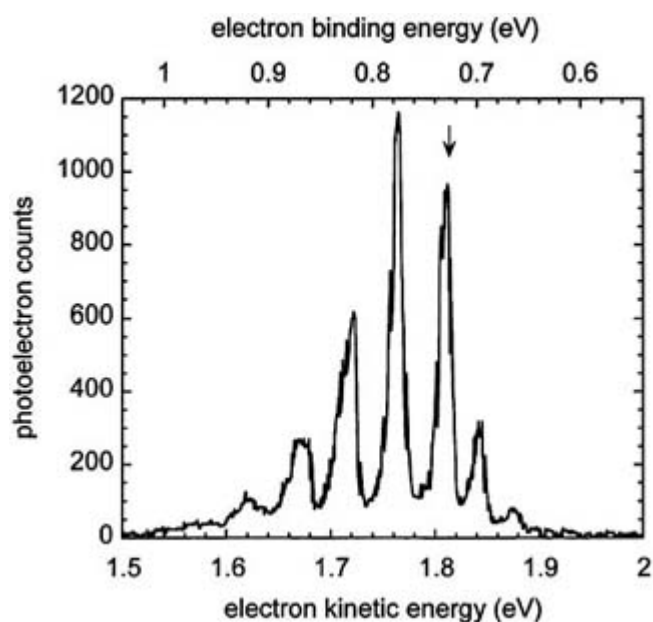


Figure A3.5.4. The 488 nm photoelectron spectrum of NaCl^- . The arrow marks the origin band, for transitions from $\text{NaCl}^- (v=0)$ to $\text{NaCl} (v=0)$, from which the electron affinity of NaCl is obtained.

It is advantageous if the laser system permits rotation of the optical polarization. Detached electrons correlated with different final electronic states of the neutral molecule will generally be emitted with different angular distributions about the direction of polarization. Measurement of the angular distribution helps in the interpretation of complex photoelectron spectra. The angular distribution $f(\theta)$ of photoelectrons is [50]

$$f(\theta) = [1 + \beta P_2(\cos \theta)]/4\pi$$

where θ is the angle between the optical polarization vector and the direction of emission of a photoelectron, β is the asymmetry parameter (β is in the range -1 to 2 and in general depends on photon energy) and $P_2(\cos \theta)$ is the second-order Legendre polynomial, $(3 \cos^2 \theta - 1)/2$. $P_2(\cos \theta)$ is zero if $\theta = 54.7^\circ$, in which case the detected electron signal is proportional to the angle-averaged detachment cross section. The distribution function is independent of the azimuthal angle.

ZEKE (zero kinetic energy) photoelectron spectroscopy has also been applied to negative ions [51]. In ZEKE work, the laser wavelength is swept through photodetachment thresholds and only electrons with near-zero kinetic energy are

allowed into a detector, resulting in narrow threshold peaks. The resulting resolution ($2\text{--}3 \text{ cm}^{-1}$) is superior to that commonly encountered with PES ($40\text{--}300 \text{ cm}^{-1}$).

PES of neutral molecules to give positive ions is a much older field [52]. The information is valuable to chemists because it tells one about unoccupied orbitals in the neutral that may become occupied in chemical reactions. Since UV light is needed to ionize neutrals, UV lamps and synchrotron radiation have been used as well as UV laser light. With suitable electron-energy resolution, vibrational states of the positive ions can be

resolved, as with the negative-ion PES described above. The angular distribution of photoelectrons can be also determined as described above.

(E) ABSORPTION

In absorption spectroscopy, the attenuation of light as it passes through a sample is measured as a function of wavelength. The attenuation is due to rovibrational or electronic transitions occurring in the sample. Mapping out the attenuation versus photon frequency gives a description of the molecule or molecules responsible for the absorption. The attenuation at a particular frequency follows the Beer–Lambert law,

$$I = I_0 \exp(-\sigma nL)$$

where I and I_0 are the attenuated and unattenuated intensities, σ is the cross section, n is the number density of target molecules and L is the path length. Broadening of spectral lines may be observed, and is classed as *homogeneous broadening* (e.g., collisions with other molecules and laser power effects) or *inhomogeneous broadening* (e.g., Doppler broadening) [53, 54]. So-called ‘UV/vis’ absorption spectroscopy is a standard tool for analysis of chemical samples. In organic samples, absorption by functional groups in the sample aids in identification of the species because it is strongly dependent upon the relative number of single, double, and triple bonds.

Microwave spectra (giving pure rotational spectra) are especially useful for the detection of interstellar molecular ions (in some cases the microwave spectrum has first been observed in interstellar spectra!).

Typically a DC glow discharge tube is used to produce the target ion (e.g., HCO^+) [55, 56]. If the photons travel parallel or antiparallel to the electric field direction, there is a small but measurable Doppler shift in frequencies. This is due to the drift velocity of ions in the electric field, which may aid in distinguishing ion spectra from neutral spectra, but in any case must be accounted for [55, 58]. Infrared spectroscopy has also been carried out on ions in a glow discharge tube using the beat frequency between a fixed-frequency visible laser and a tunable dye laser. The difference frequency laser, in the IR, irradiates a long discharge tube. The method was first used to study the important astrophysical ion H_3^+ [59]. *Velocity modulation spectroscopy* utilizes an audio frequency glow discharge coupled with phase synchronous demodulation of the absorbed IR laser radiation to take advantage of the Doppler shift occurring for ions drifting in glow discharge tubes. Many important positive ions, such as H_3O^+ , NH_4^+ and H_3^+ , have been studied with this technique with the high precision common to IR spectroscopy [58, 60].

Far-infrared spectra of great sensitivity may be obtained with *laser magnetic resonance* (LMR). The sensitivity comes about because the gas sample is located inside the long cavity of the laser where the circulating power is typically 100 times that used in extracavity work. A discharge in the gas cell produces the radical species to be studied, and an axial magnetic field is varied to bring energy levels into resonance with one of many laser lines. HBr^+ was the first ion to be observed with LMR, in 1979. OH^- was one of the first negative ions to have been detected by direct absorption spectroscopy [61].

Strictly speaking, the term absorption spectroscopy refers to measurements of light intensity. In practice, the absorption may be deduced from the detection of electrons or ions produced in the process, such as in absorption of light leading to photodetachment or photodissociation, i.e., action spectra. In the absorption spectroscopy of ions, this is a natural tack to take as the charged-particle production can be detected with greater precision than is possible for a measurement of a small change in the light intensity. The most important of such experiment types is the coaxial beams spectrometer, one of which is described in detail in the section on dissociation of ions. In these experiments, the ions are identified by mass and collimated, and interact with the laser beam over a long distance (0.25–1 m). The method was first used with ions in 1976 for HD^+ , with the absorption events detected via enhanced production of buffer gas ions as a result of charge

transfer reactions [62]. Since this time many small ions have been studied in great detail, notably H_3^+ , O_2^+ , CH^+ , H_2O^+ , and CO^+ [31, 32, 63]. A few negative ions have been studied using coaxial fast-ion/laser beams. The high-resolution IR spectrum of NH^- , for example, was studied in this manner. The negative ion was excited to an autodetaching state with a photon energy greater than the electron affinity of NH . Detection of autodetached electrons signified an absorption event. Aside from determination of spectroscopic constants for NH^- , information on autodetachment dynamics was obtained [64].

A3.5.2.2 KINETICS AND DYNAMICS

In principle the study of ion–molecule kinetics and dynamics is no different from studies of the same processes in neutral species; however, there are additional forces that govern reactivity, often leading to behaviour that is fundamentally different from neutral processes. An important factor in determining ion–molecule rate constants and cross sections is the rate at which the reactants collide, i.e. the collision rate. In contrast to neutral kinetics, the collision rate at low energies or temperatures is determined not by the size of the molecule but by electrical forces. The ion–molecule collision rate is determined by the classical capture cross section for a point charge interacting with a structureless multipole. This was first described analytically for a point charge interacting with a polarizable species with no other multipole. In this case, the collisional value of the rate constant is independent of temperature [65]. The only other force of any significance is from the ion–permanent dipole interaction. Other forces, such as those between the ion and the quadrupole moment of the neutral, and between the neutral dipole and the induced dipole of the ion, have been shown to be of minor importance [58, 59, 60, 61, 62 and 63]. If the physical size of the reactants is greater than the capture radius, e.g. at translational energies of several tenths of an electron volt and greater, more conventional notions apply. Except for species with very small polarizabilities and systems of large mass, ion–molecule collision rates are above 10^{-9} molecules $\text{cm}^{-3} \text{s}^{-1}$, or about a factor of ten larger than neutral collision rates.

Several processes are unique to ions. A common reaction type in which no chemical rearrangement occurs but rather an electron is transferred to a positive ion or from a negative ion is termed charge transfer or electron transfer. Proton transfer is also common in both positive and negative ion reactions. Many proton- and electron-transfer reactions occur at or near the collision rate [72]. A reaction pertaining only to negative ions is associative detachment [73, 74],



Associative detachment reactions are important in controlling the electron concentration in the earth's mesosphere [75]. Reactions in which more than one neutral product are formed also occur and are sometimes referred to as reactive detachment [76].

Several reactivity trends are worth noting. Reactions that are rapid frequently stay rapid as the temperature or centre-of-mass kinetic energy of the reactants is varied. Slow exothermic reactions almost always show behaviour such that

the rate constant decreases with increasing temperature at low temperature or kinetic energies and then increases at higher temperature. As an example, figure A3.5.5 shows rate constants for the charge-transfer reaction of Ar^+ with O_2 as a function of temperature. The data are from five separate experiments and four experimental techniques [77, 78, 79, 80 and 81] and cover the extremely wide temperature range of 0.8 K to 1400 K. The extremely low temperature data are relatively flat. At approximately 20 K, the rate constants decrease. The decrease is described by a power law. A minimum is found at 800–900 K and a steep increase is found above 1000 K. The position of the minimum varies considerably for other reactions.

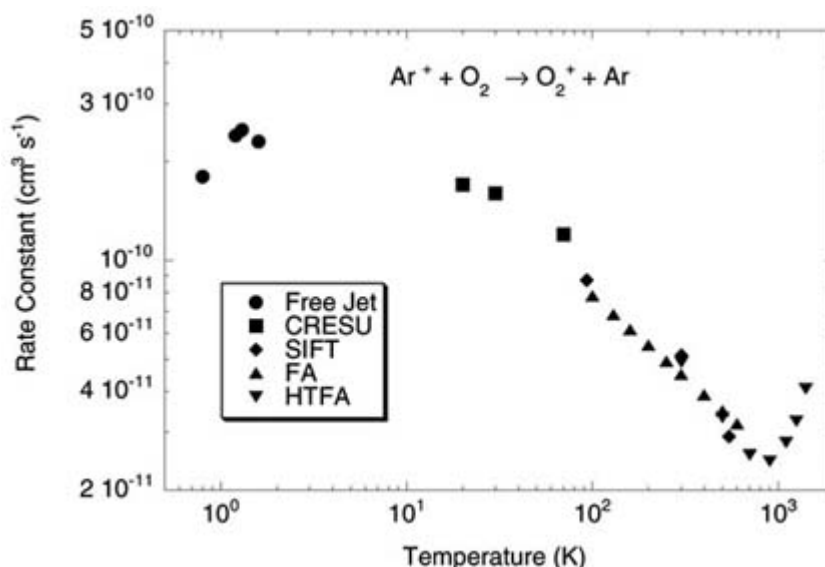
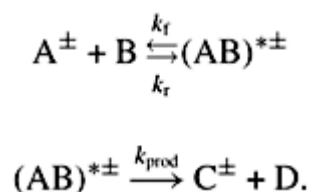


Figure A3.5.5. Rate constants for the reaction of Ar^+ with O_2 as a function of temperature. CRESU stands for the French translation of reaction kinetics at supersonic conditions, SIFT is selected ion flow tube, FA is flowing afterglow and HTFA is high temperature flowing afterglow.

The decrease in reactivity with increasing temperature is due to the fact that many low-energy ion–molecule reactions proceed through a double-well potential with the following mechanism [82]:



The minimum energy pathway for the reaction of Cl^- with CH_3Br is shown in [figure A3.5.6](#) [83]. As the reactants approach they are attracted by the ion–dipole and ion–induced-dipole forces and enter the entrance channel complex. As the reaction proceeds along the minimum energy path, the potential energy increases due to the forces necessary for rearrangement. The species then enter the product well and finally separate into products. The two wells are separated by a barrier that is often below the energy of the reactants but still plays an important role in controlling reactivity. The decrease in the overall rate constant with increasing temperature is due to the rate constant for collision complex

dissociation to reactants, k_b , increasing more rapidly with temperature than the rate constant for the complex going to products, k_{prod} [68]. The increase at higher energies and temperatures is often due to new channels opening, including new vibrational and electronic states as well as new chemical channels.

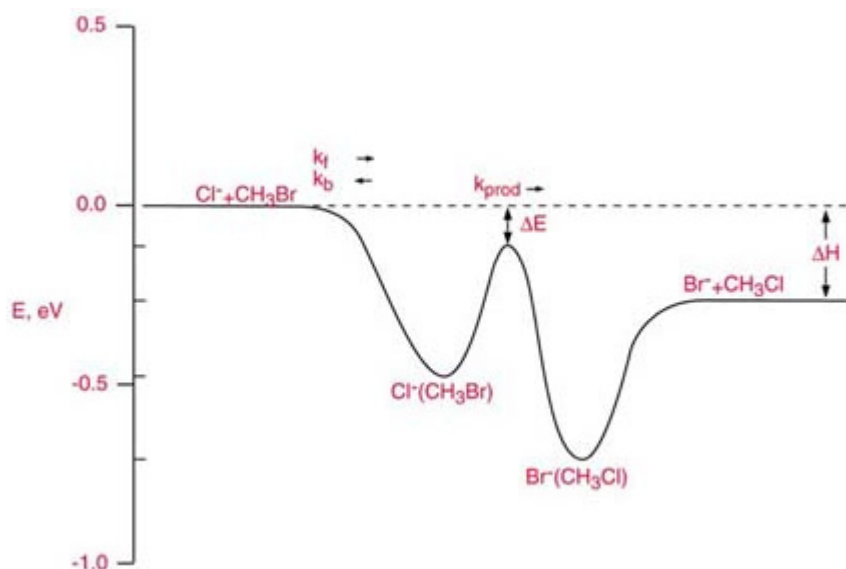


Figure A3.5.6. Minimum energy pathway for the reaction of Cl^- with CH_3Br .

Rotational and translational energy have been shown to be equivalent in controlling most reactions [84]. Vibrational energy often increases reactivity; however, sometimes it does not affect reactivity, or occasionally decreases reactivity. The following sections describe several of the more common techniques used to measure ion–molecule rate constants or cross sections.

(A) FLOW TUBES

Flow tube studies of ion–molecule reactions date back to the early 1960s, when the flowing afterglow was adapted to study ion kinetics [85]. This represented a major advance since the flowing afterglow is a thermal device under most situations and previous instruments were not. Since that time, many iterations of the ion–molecule flow tube have been developed and it is an extremely flexible method for studying ion–molecule reactions [86, 87, 88, 89, 90, 91 and 92].

The basic flow system is conceptually straightforward. A carrier gas, often helium, flows into the upstream end of a tube approximately 1 m long with a radius of several centimetres. This buffer gas pressure is approximately 100 Pa. Ions are created either in the flow tube or injected from an external source at the upstream end of the pipe. The carrier gas transports the ions downstream at approximately 100 m s^{-1} . Part way down the tube, a neutral reactant is added and the ions created in the source region are transformed into products. Conditions are chosen so that all ion chemistry leading to the reactant ion is complete and the ions are thermalized before they encounter the neutral reagent gas. The rate constant is determined by sampling a small portion of the gas with a quadrupole mass spectrometer and monitoring the disappearance of the primary ion and the appearance of product ions. For the reaction of $\text{A}^+ + \text{B} \rightarrow \text{products}$, the rate constant is given by

-13-

$$k = 1/[\text{B}]\tau \ln([\text{A}_0^+]/[\text{A}^+])$$

A3.5.1

where $[\text{B}]$ is the reactant neutral concentration in the flow tube, τ is the reaction time and $[\text{A}_0^+]$ and $[\text{A}^+]$ are the ion concentrations with and without reactant neutrals in the flow tube. This equation assumes that the

concentration of B is much greater than that of A^+ , i.e. first order kinetics apply. This situation applies to all ion kinetics since it is difficult to make large quantities of ions. Fortunately, the derivation of the rate constant depends only on the relative ion concentration, which is much easier to measure than the absolute concentration. The reaction time is determined from the flow velocity of the carrier gas. The average ion flow velocity is approximately 1.6 times the average neutral flow velocity [87], a result of ion diffusion, ions being neutralized on the flow tube walls and the carrier gas having a parabolic flow profile characteristic of laminar flow.

The basic system described above can be easily modified to study many processes. Figure A3.5.7 shows an example of a modern ion–molecule flow tube [93] with a number of interesting features. First, ions are created external to the flow tube. Any suitable ion source can be used, including high- and low-pressure electron-impact ion sources, a supersonic-expansion source (shown) or a flow-tube source. Once created the ions are injected into a quadrupole mass spectrometer and only ions with the proper mass are injected into the flow tube through a Venturi inlet. Under favourable circumstances, only one ion species enters the flow tube. This configuration is called the selected-ion flow tube (SIFT) [89, 90 and 91]. Alternatively, ions can be created in the carrier-gas flow by a filament or discharge. Neutral reagents are added through a variety of inlets. Unstable species such as O, H and N atoms, molecular radicals and vibrationally excited diatomics can be injected by passing the appropriate gas through a microwave discharge. In a SIFT, the chemistry is usually straightforward since there is only one reactant ion and one neutral present in the flow tube.

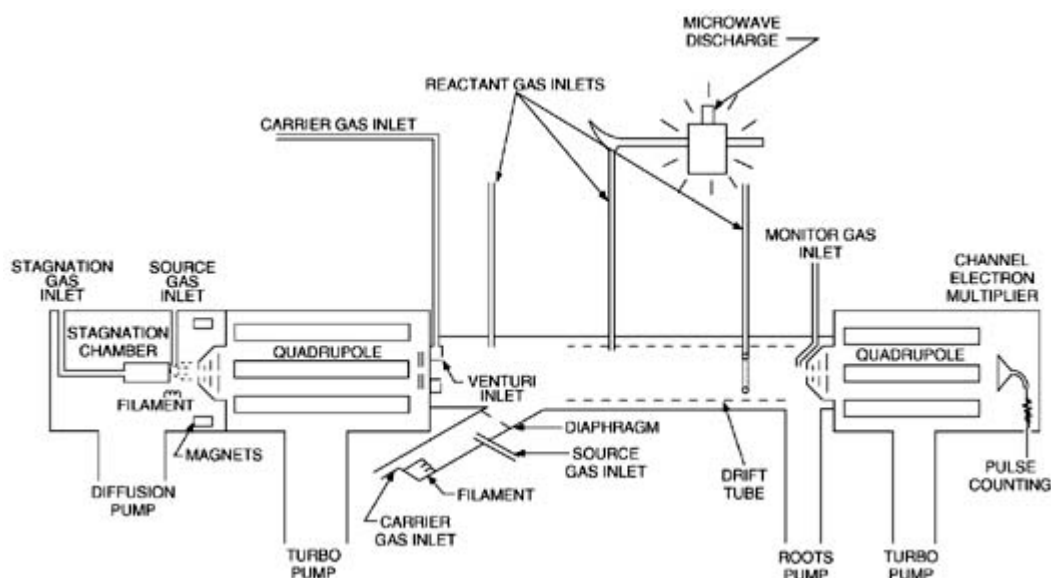


Figure A3.5.7. Schematic diagram of a selected ion flow drift tube with supersonic expansion ion source.

Flowing afterglows and SIFTs have been operated between 80 K and 1800 K. In addition, the ion kinetic energy can be varied by adding a drift tube either at room temperature [94, 95, 96 and 97] or with temperature variability [84]. A drift tube is a series of rings electrically connected by uniform resistors. Applying a voltage to the resistor chain forms a uniform electric field in the flow tube. Ions are accelerated by the electric field and decelerated by collisions so that a steady-state velocity results. The ion kinetic energy in the centre of mass, $\langle KE \rangle_{cm}$, is given by the Wannier expression [98]

$$\langle KE \rangle_{cm} = \frac{(m_i + m_b)m_n}{2(m_i + m_n)} v_d^2 + 3/2kT$$

where m_i , m_n and m_b are the mass of the ion, neutral and buffer respectively, v_d is the velocity of the ion due to the electric field and T is the temperature. Equation (A3.5.2) shows that the kinetic energy in a drift tube is the sum of a thermal component ($\frac{3}{2}kT$) and a drift field component. At a fixed kinetic energy, varying the contribution of each term yields information on rotational and vibrational effects [84]. Excited-state effects can be studied in a number of other ways. Electronic excitation often occurs in SIFT studies of atomic ions. Vibrational effects can be studied by exciting neutral diatomics in a microwave discharge. Ion vibrations can be excited in the source and monitored by LIF or judicious choice of a reactant neutral, i.e. one that reacts differently with excited states than for ground states. Often one looks for a reaction that is endothermic with respect to the ground state and energetically allowed for the excited state. Product-state information can be obtained by the monitor method or through optical spectroscopy. This list of possibilities is not exhaustive but it does give a sample of the type of information that can be obtained.

(B) TRAPS

Another powerful class of instrumentation used to study ion–molecule reactivity is trapping devices. Traps use electric and magnetic fields to store ions for an appreciable length of time, ranging from milliseconds to thousands of seconds. Generally, these devices run at low pressure and thus can be used to obtain data at pressures well below the range in which flow tubes operate.

The most widely used type of trap for the study of ion–molecule reactivity is the ion-cyclotron-resonance (ICR) [99] mass spectrometer and its successor, the Fourier-transform mass spectrometer (FTMS) [100, 101]. Figure A3.5.8 shows the cubic trapping cell used in many FTMS instruments [101]. Ions are created in or injected into a cubic cell in a vacuum of 10^{-2} Pa or lower. A magnetic field, B , confines the motion in the x – y plane through ion-cyclotron motion. The frequency of motion, ω , is given as $1.537 \times 10^7 B e/m$ where B is the magnetic field in tesla (typically 1–7 T), e is the charge of an electron and m the mass in atomic mass units. To trap ions in the z direction, a potential is placed on the two end electrodes. The ions oscillate in the z direction until their motion is damped by collisions. The magnetic field adds little energy to the motion, and the ions can be described as thermal. Ions are detected by applying a radio frequency (RF) pulse (a chirp) to the transmitter plates. The RF pulse causes ions with the matching cyclotron frequency to absorb energy. The ions are not only energized but quickly move coherently. Image currents on the receiver plates are detected. By putting a rapidly varying RF pulse (0–2 MHz) on the transmitter plate one obtains image currents as a function of time. A fast Fourier transform yields the frequency spectrum that is directly related to the mass spectrum by the equation described above.

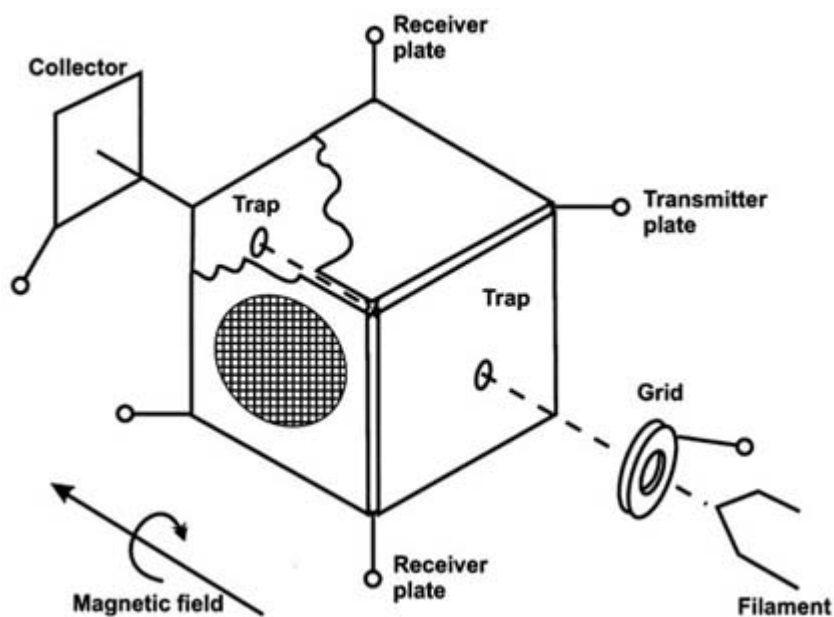


Figure A3.5.8. Schematic diagram of the cell used in a Fourier transform mass spectrometer.

Kinetics measurements are made by adding a known concentration of reactant gas to the cell and monitoring the time evolution of the ion intensities. Rate constants are derived from [equation \(A3.5.1\)](#). Many useful tricks can be employed. The most useful is to chirp the cell so that all ions except those with the correct mass are excited out of the cell. In this manner the kinetics are simplified; only one ion and one neutral exist at time zero, and product information is easily obtained. A mass-specific excitation pulse adds energy such that ions may acquire enough energy to dissociate upon collision with background gas. The pattern of the dissociation often yields structural information. The ICR is particularly suited to the study of radiative association [\[101\]](#) and radiative cooling [\[102\]](#) of ions since the pressure is low and the trapping time can be long.

Another class of trapping device that is gaining importance is the radiofrequency trap [\[103\]](#). Quadrupole ion traps (also called Paul traps) are three-dimensional traps with rotationally symmetric ring and endcap electrodes. An RF voltage of opposite phase is applied to the ring and endcaps, respectively, to create a quadrupolar RF field. This type of trap suffers from electric field heating of the ions and can be classified as a nonthermal device. More innovative traps in limited use are the ring electrode trap and 22-pole trap. Both of these devices trap ions in a large field-free region and produce thermal ions. Reactions at very low temperatures have been studied with these types of trap [\[81, 103\]](#).

(C) BEAMS

The guided-ion beam has become the instrument of choice for studying ion–molecule reactions at elevated kinetic energies [\[103\]](#). In many guided-ion beam systems the lowest energy obtainable is slightly above thermal energy ($\sim 0.1\text{--}0.2$ eV), although it can be as low as the thermal energy of the target gas. The upper range varies but is generally in the tens of electron volts.

In essence, a guided-ion beam is a double mass spectrometer. [Figure A3.5.9](#) shows a schematic diagram of a guided-ion beam apparatus [\[104\]](#). Ions are created and extracted from an ion source. Many types of source have been used and the choice depends upon the application. Combining a flow tube such as that described in this chapter has proven to be versatile and it ensures the ions are thermalized [\[105\]](#). After extraction, the ions are mass selected. Many types of mass spectrometer can be used; a Wien ExB filter is shown. The ions are then injected into an octopole ion trap. The octopole consists of eight parallel rods arranged on a circle. An RF

voltage is applied to alternating sets of rods to trap ions in the centre of the octopole in an approximately square well potential. Little energy is transferred to the ions. The surrounding part of the octopole is a chamber where reactive gas is added. Typical pressures of added gas are of the order of 10^{-2} Pa. Pressure is kept low so single-collision conditions apply; the primary ions collide at most once with the reactant gas. The collision cell is generally run at room temperature although cooled and heated versions have been used. The main advantage of the octopole collision cell arrangement, over the arrangement used in early beam apparatuses, is greater collection efficiency of the product ions since products scattering in all directions are collected. The primary ions react with the reactant neutral and the resulting mix of ions exits the octopole to be mass analysed and detected. A quadrupole mass filter is often used for mass analysis although other mass spectrometers can be used. The reaction cross section is derived from the Beer–Lambert law, $I = I_0 \exp(-\sigma nL)$ where I and I_0 are the reactant ion signals with and without the reactant gas, n is the number density in the gas, L is the length of the collision cell. In the single-collision limit, I is taken as the product ion signal and I_0 as the primary ion signal.

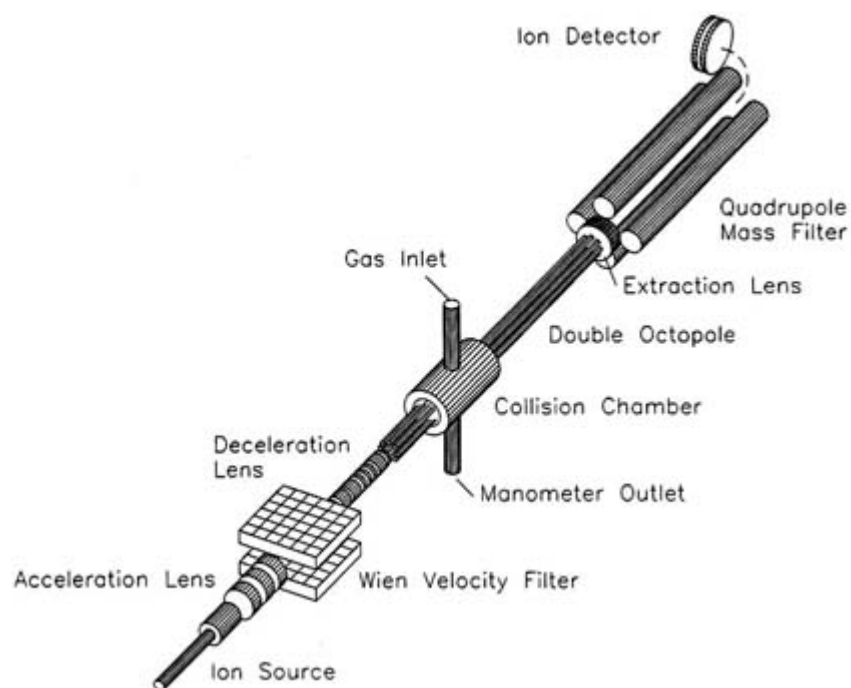


Figure A3.5.9. Schematic diagram of a guided-ion beam.

As with most methods for studying ion–molecule kinetics and dynamics, numerous variations exist. For low-energy processes, the collision cell can be replaced with a molecular beam perpendicular to the ion beam [106]. This greatly reduces the thermal energy spread of the reactant neutral. Another approach for low energies is to use a merged beam [103]. In this system the supersonic expansion is aimed at the throat of the octopole, and the ions are passed through

a quadrupole bender and merged with the neutral beam. Exceedingly low collision energies can be obtained with this arrangement. Another important modification is obtained by adding a second octopole between the collision cell and the mass analyser. This allows the product-ion flight times to be measured, thus yielding the kinetic-energy release in the reaction to provide important dynamical information. Laser radiation can be introduced into the octopole to measure product distributions or to study dissociative processes. One valuable use of guided-ion beams has been the study of thresholds for endothermic processes in order to measure bond strengths and other thermodynamic quantities.

Two techniques exist for measuring the angular distribution of products. In the crossed-beam setup, the

octopole/collision cell is replaced with an interaction zone defined by the overlap of an ion beam and a supersonic neutral beam [106]. The angular distribution is measured by moving a mass spectrometer to detect ions at various angles. A simpler approach is to measure the product transmission as a function of trapping potential on the octopole [103]. The derivative of the signal yields the angular information but with limited resolution. Angular distributions are often used to determine the extent of collision complex formation.

(D) OTHER TECHNIQUES

While the techniques described above are the most common and versatile for measuring ion–molecule kinetics, several other techniques are worth mentioning. An important technique for measuring ion energetics is the pulsed, high-pressure mass spectrometer (PHPMS) [107]. In PHPMS, a pulsed beam of 2 keV electrons enters a small chamber containing reactants and a buffer gas. The chamber is maintained at ~500 Pa. Ion signals are then recorded as a function of time until equilibrium is established. Knowledge of the ion signals and partial pressures of the reactant neutral(s) yields the equilibrium constant. Temperature variation allows the enthalpy and entropy of reaction to be derived. Important thermodynamic information obtained by this technique includes ligand bond strengths, proton affinities, gas phase basicities, electron affinities and ionization energies. Information on kinetics can also be obtained.

Several instruments have been developed for measuring kinetics at temperatures below that of liquid nitrogen [81]. Liquid helium cooled drift tubes and ion traps have been employed, but this apparatus is of limited use since most gases freeze at temperatures below about 80 K. Molecules can be maintained in the gas phase at low temperatures in a free jet expansion. The CRESU apparatus (acronym for the French translation of reaction kinetics at supersonic conditions) uses a Laval nozzle expansion to obtain temperatures of 8–160 K. The merged ion beam and molecular beam apparatus are described above. These techniques have provided important information on reactions pertinent to interstellar-cloud chemistry as well as the temperature dependence of reactions in a regime not otherwise accessible. In particular, information on ion–molecule collision rates as a function of temperature has proven valuable in refining theoretical calculations.

Most ion–molecule techniques study reactivity at pressures below 1000 Pa; however, several techniques now exist for studying reactions above this pressure range. These include time-resolved, atmospheric-pressure, mass spectrometry; optical spectroscopy in a pulsed discharge; ion-mobility spectrometry [108] and the turbulent flow reactor [109].

A3.5.3 APPLICATIONS

A3.5.3.1 ION STRUCTURE AND ENERGETICS

The molecular constants that describe the structure of a molecule can be measured using many optical techniques described in [section A3.5.1](#) as long as the resolution is sufficient to separate the rovibrational states [110, 111 and 112]. Absorption spectroscopy is difficult with ions in the gas phase, hence many ion species have been first studied by matrix isolation methods [113], in which the IR spectrum is observed for ions trapped within a frozen noble gas on a liquid-helium cooled surface. The measured frequencies may be shifted as much as 1% from gas phase values because of the weak interaction with the matrix.

These days, remarkably high-resolution spectra are obtained for positive and negative ions using coaxial-beam spectrometers and various microwave and IR absorption techniques as described earlier. Information on molecular bond strengths, isomeric forms and energetics may also be obtained from the techniques discussed earlier. The kinetics of cluster-ion formation, as studied in a selected-ion flow tube (SIFT) or by high-pressure

mass spectrometry, may be interpreted in terms of cluster bond strengths [114]. In addition, the chemistry of ions may be used to identify the structure. For example, the ionic product of reaction between O_2^+ and CH_4 at 300 K has been identified as CH_2OOH^+ from its chemistry; the reaction mechanism is insertion [115]. Collision-induced dissociation (in a SIFT apparatus, a triple-quadrupole apparatus, a guided-ion beam apparatus, an ICR or a beam-gas collision apparatus) may be used to determine ligand-bond energies, isomeric forms of ions and gas-phase acidities.

Photoelectron spectra of cluster ions yields cluster-bond strengths, because each added ligand increases the binding energy of the extra electron in the negative ion by the amount of the ligand bond strength (provided the bond is electrostatic and does not appreciably affect the chromophore ion) [116].

One example of the determination of molecular constants can be taken from the photoelectron spectrum for NaCl^- shown in figure A3.5.4 [48]. The peak spacing to the left of the origin band is 45 meV: the nominal vibrational frequency ω_e in neutral NaCl . The spectral resolution is not good enough to specify the small anharmonic correction, $\omega_e x_e$, or the rotational constant, B_e , but these, along with the equilibrium separation, r_e ($=2.361 \text{ \AA}$), are accurately known from optical spectra. The spectrum in figure A3.5.4 also provides new information about the negative ion: the peaks to the right of the origin band are spaced by 33 meV, the nominal vibrational spacing in NaCl^- . The distribution of peak heights everywhere is determined by the Franck–Condon overlap of wavefunctions for NaCl and NaCl^- vibrational states, so the data give the ion temperature and the magnitude of the change in r_e between the neutral and negative ion (0.136 \AA in this case). Vibrational frequencies and bond-energy considerations imply that $r_e(\text{NaCl}^-) > r_e(\text{NaCl})$. Therefore, $r_e(\text{NaCl}^-) = 2.497 \text{ \AA}$, and $B_e = 0.195 \text{ cm}^{-1}$. Finally, the position of the origin peak gives the electron binding energy (the electron affinity of NaCl , 0.727 eV) and a thermochemical cycle allows one to calculate the bond energy of NaCl^- (all other quantities being known):

$$D_0(\text{Na-Cl}^-) = D_0(\text{Na-Cl}) + \text{EA}(\text{NaCl}) - \text{EA}(\text{Cl})$$

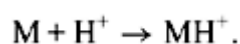
-19-

yielding $D_0(\text{Na-Cl}^-) = 1.34 \text{ eV}$. Admittedly, this is a simple spectrum to interpret. Had the system involved a Π state instead of Σ states, additional peaks would have complicated the spectrum (but yielded additional information if resolved). Had the molecules been polyatomic, where vibrations may be bending modes or stretches and combinations, the spectra and interpretation become more complex—the systems must be described in terms of normal modes instead of the more intuitive, but coupled, stretches and bends.

A.3.5.3.2 THERMOCHEMISTRY

The principles of ion thermochemistry are the same as those for neutral systems; however, there are several important quantities pertinent only to ions. For positive ions, the most fundamental quantity is the adiabatic ionization potential (IP), defined as the energy required at 0 K to remove an electron from a neutral molecule [117, 118 and 119].

Positive ions also form readily by adding a proton to a neutral atom or molecule [120]



The proton affinity, PA, is defined (at 298 K) as [117]

$$\text{PA} = \Delta H_f^0(\text{M}) + \Delta H_f(\text{H}^+) - \Delta H_f(\text{MH}^+).$$

Negative ions also have two unique thermodynamic quantities associated with them: the electron affinity, EA, defined as the negative of the enthalpy change for addition of an electron to a molecule at 0 K [117, 121, 122]



and the gas-phase acidity of a molecule, defined as the Gibbs energy change at 298 K, $\Delta G_{\text{acid}}(\text{AH})$, for the process [117, 121]

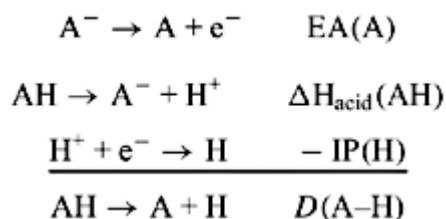


The enthalpy for this process is the proton affinity of the negative ion.

Much effort has gone into determining these quantities since they are fundamental to ionic reactivity. Examples include thermodynamic equilibrium measurements for all quantities and photoelectron studies for determination of EAs and IPs. The most up-to-date tabulation on ion thermochemistry is the *NIST Chemistry WebBook* (webbook.nist.gov/chemistry) [123].

Neutral thermochemistry can be determined by studying ion thermochemistry. For example, the following cycle can be used to determine a neutral bond strength,

-20-

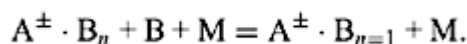


where $D(\text{A-H})$ is the bond dissociation energy or enthalpy for dissociating a hydrogen atom from AH. Often it is easier to determine the EA and anionic proton affinity than it is to determine the bond strength directly, especially when AH is a radical. The IP of hydrogen is well known. As an example, this technique has been used to determine all bond strengths in ethylene and acetylene [124].

A3.5.3.3 CLUSTER PROPERTIES

A gas phase ionic cluster can be described as a core ion solvated by one or more neutral atoms or molecules and it is often represented as $\text{A}^{\pm}(\text{B})_n$ or $\text{A}^{\pm} \cdot \text{B}_n$, where A^{\pm} is the core ion and B are the ligand molecules. Of course, the core and the ligand can be the same species, e.g. the hydrated electron. The interactions governing the properties of these species are often similar to those governing liquid-phase ionic solvation. Modern techniques allow clusters with a specific number of neutral molecules to be studied, providing information on the evolution of properties as a function of solvation number. This leads to insights into the fundamental properties of solutions and has made this field an active area of research.

The most fundamental of cluster properties are the bond strengths and entropy changes for the process [125]



The thermodynamic quantities are derived from equilibrium measurements as a function of temperature. The measurements are frequently made in a high-pressure mass spectrometer [107]. The pertinent equation is ln

$(K_{n,n+1}) = -\Delta G^0/RT = -\Delta H^0/RT + \Delta S^0/T$. Another important method to determine bond strengths is from threshold measurements in collisional dissociation experiments [126]. Typically, ΔH^0 changes for $n = 0$ are 1.5 to several times the solution value [127]. The value usually drops monotonically with increasing n . The step size can have discontinuities as solvent shells are filled. The discontinuities in the thermodynamic properties appear as magic numbers in the mass spectra, i.e. ions of particular stability such as those with a closed solvation shell tend to be more abundant than those with one more or less ligand. A graph of bond strengths for H_2O bonding to several ions against cluster size is shown in figure A3.5.10 [125].

-21-

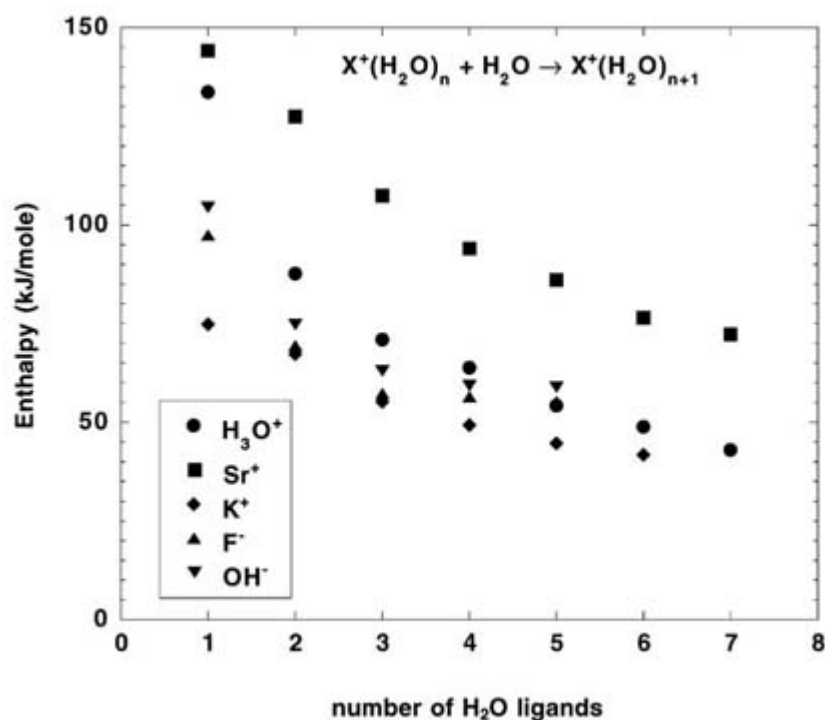


Figure A3.5.10. Bond strengths of water clustering to various core ions as a function of the number of water molecules.

Clusters can undergo a variety of chemical reactions, some relevant only to clusters. The simplest reaction involving clusters is association, namely the sticking of ligands to an ionic core. For association to occur, the ion-neutral complex must release energy either by radiating or by collision with an inert third body. The latter is an important process in the earth's atmosphere [128, 129, 130 and 131] and the former in interstellar clouds [101]. Cluster ions can be formed by photon or electron interaction with a neutral cluster produced in a supersonic expansion [132]. Another process restricted to clusters is ligand-switching or the replacement of one ligand for another. Often exothermic ligand-switching reactions take place at rates near the gas kinetic limit, especially for small values of n [72, 133]. Chemical-reactivity studies as a function of cluster size show a variety of trends [93, 127, 133]. Proton-transfer reactions are often unaffected by solvation, while nucleophilic-displacement reactions are often shut down by as few as one or two solvent molecules.

Increasing solvation number can also change the type of reactivity. A good example is the reaction of $NO^+(H_2O)_n$ with H_2O . These associate for small n but react to form $H_3O^+(H_2O)_n$ ions for $n = 3$. This is an important process in much of the earth's atmosphere. Neutral reactions have been shown to proceed up to 30 orders of magnitude faster when clustered to inert alkali ions than in the absence of the ionic clustering [134].

Caging is an important property in solution and insight into this phenomenon has been obtained by studying photodestruction of $Br_2^-(M)_n$ and $I_2^-(M)_n$ clusters, where M is a ligand such as Ar or CO_2 . When the X_2^- core is

photoexcited above the dissociation threshold of X_2^- , the competition between the two processes forming $X^-(M)_m$ and $X_2^-(M)_m$ indicates when caging is occurring. For $I_2^-(CO_2)_n$ the caging is complete at $n = 16$ [127, 135].

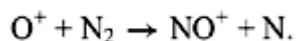
-22-

An important class of molecule often described as clusters may better be referred to as micro-particles. This class includes metal, semiconductor and carbon clusters. Particularly interesting are the carbon clusters, C_n^+ . Mass spectra from a carbon cluster ion source show strong magic numbers at C_{60}^+ and C_{70}^+ [136]. This led to the discovery of the class of molecules called buckminsterfullerenes. Since that time, other polyhedra have been discovered, most notably metallocarbohedrenes [137]. The first species of this type discovered was $Ti_8C_{12}^+$. Much of the work done on metal clusters has been focused on the transition from cluster properties to bulk properties as the clusters become larger, e.g. the transition from quantum chemistry to band theory [127].

A3.5.3.4 ATMOSPHERIC CHEMISTRY

Atmospheric ions are important in controlling atmospheric electrical properties and communications and, in certain circumstances, aerosol formation [128, 130, 131, 138, 139, 140, 141, 142, 143, 144 and 145]. In addition, ion composition measurements can be used to derive trace neutral concentrations of the species involved in the chemistry. Figure A3.5.11 shows the total-charged-particle concentration as a function of altitude [146]. The total density varies between 10^3 and 10^6 ions cm^{-3} . The highest densities occur above 100 km. Below 100 km the total ion density is roughly constant even though the neutral density changes by a factor of approximately 4×10^6 . Most negative charge is in the form of electrons above 80 km, while negative ions dominate below this altitude.

Above approximately 80 km, the prominent bulge in electron concentration is called the ionosphere. In this region ions are created from UV photoionization of the major constituents—O, NO, N_2 and O_2 . The ionosphere has a profound effect on radio communications since electrons reflect radio waves with the same frequency as the plasma frequency, $f = 8.98 \times 10^3 n_e^{1/3}$, where n_e is the electron density in cm^{-3} [147]. The large gradient in electron density ensures that a wide variety of frequencies are absorbed. It is this phenomenon that allows one to hear distant radio signals. Ion chemistry plays a major role in determining the electron density. Diatomic ions recombine rapidly with electrons while monatomic ions do not. Monatomic positive ions do not destroy electrons until they are converted to diatomic ions. The most important reaction in the ionosphere is the reaction of O^+ with N_2 ,



Although this reaction is exothermic, the reaction has a small rate constant. This is one of the most studied ion-molecule reactions, and dependences on many parameters have been measured [148].

-23-

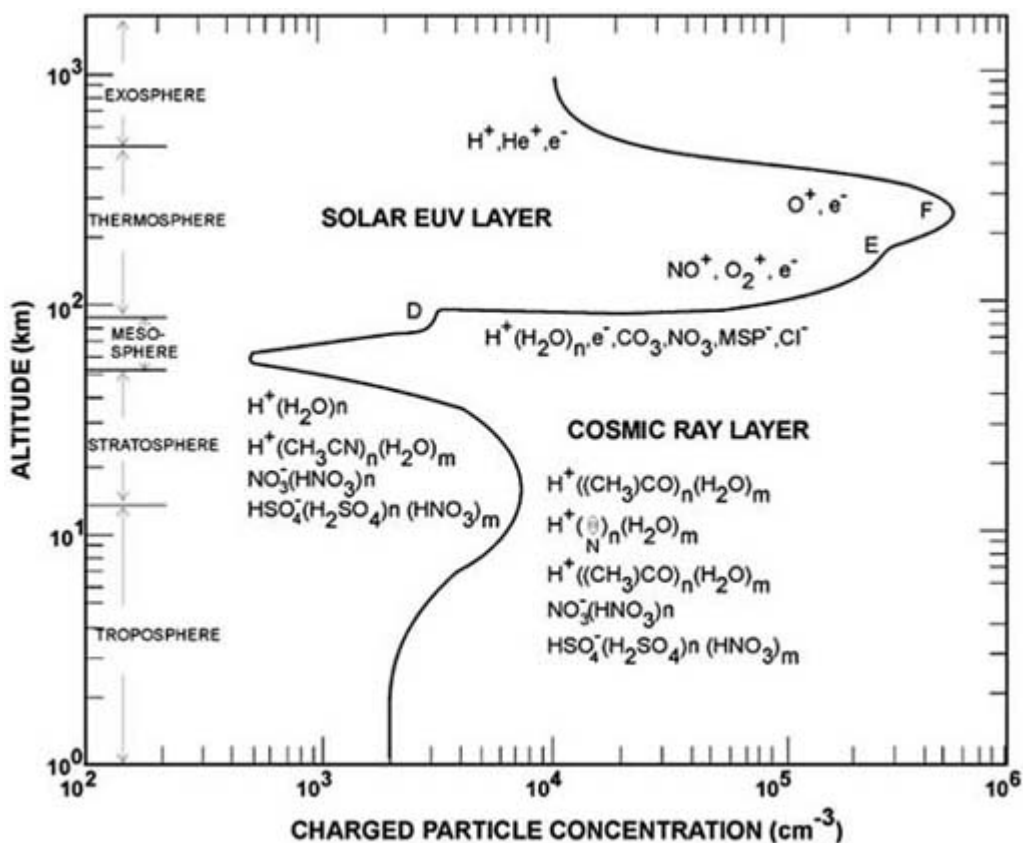


Figure A3.5.11. Charged particle concentrations in the atmosphere.

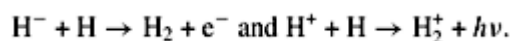
More complex ions are created lower in the atmosphere. Almost all ions below 70–80 km are cluster ions. Below this altitude range free electrons disappear and negative ions form. Three-body reactions become important. Even though the complexity of the ions increases, the determination of the final species follows a rather simple scheme. For positive ions, formation of $\text{H}^+(\text{H}_2\text{O})_n$ is rapid, occurring in times of the order of milliseconds or shorter in the stratosphere and troposphere. After formation of $\text{H}^+(\text{H}_2\text{O})_n$, the chemistry involves reaction with species that have a higher proton affinity than that of H_2O . The resulting species can be written as $\text{H}^+(\text{X})_m(\text{H}_2\text{O})_n$. The main chemical processes include ligand exchange and proton transfer as well as association and dissociation of H_2O ligands. Examples of species X include NH_3 [149], CH_3COCH_3 [150] and CH_3CN [151]. The rate constants are large, so the proton hydrates are transformed even when the concentration of X is low.

The negative ion chemistry is equally clear. $\text{NO}_3^-(\text{HNO}_3)_m(\text{H}_2\text{O})_n$ ions are formed rapidly. Only acids, HX, stronger than HNO_3 react with this series of ions producing $\text{X}^-(\text{HX})_m(\text{H}_2\text{O})_n$. Most regions of the atmosphere have low concentrations of such acids. The two exceptions are a layer of H_2SO_4 in the 30–40 km region [152, 153] and H_2SO_4 and $\text{CH}_3\text{SO}_3\text{H}$ which play an important role near the ground under some circumstances [154].

Ion-composition measurements can be used to derive the concentrations of X and HX involved in the chemistry. This remains the only practical method of monitoring abundances of H_2SO_4 and CH_3CN in the upper atmosphere. Concentrations as low as 10^4 molecules cm^{-3} have been measured.

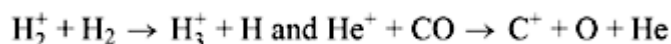
A3.5.3.5 ASTROCHEMISTRY

The astrochemistry of ions may be divided into topics of interstellar clouds, stellar atmospheres, planetary atmospheres and comets. There are many areas of astrophysics (stars, planetary nebulae, novae, supernovae) where highly ionized species are important, but beyond the scope of ‘ion chemistry’. (Still, molecules, including H₂O, are observed in solar spectra [155] and a surprise in the study of Supernova 1987A was the identification of molecular species, CO, SiO and possibly H₃⁺[156, 157].) In the early universe, after expansion had cooled matter to the point that molecules could form, the small fraction of positive and negative ions that remained was crucial to the formation of molecules, for example [156]



The formation of molecules was the first step toward local gravitational collapses which led to, among other things, the production of this encyclopedia.

Interstellar clouds of gases contain mostly H, H₂ and He, but the minority species are responsible for the interesting chemistry that takes place, just as in the earth’s atmosphere. Interstellar clouds are divided into two types: *diffuse*, with atomic or molecular concentrations in the neighbourhood of 100 cm⁻³ and temperatures of 100–200 K, in which ionization is accomplished primarily by stellar UV light, and *dense* (or dark) clouds, with densities of 10⁴–10⁶ cm⁻³ and temperatures of 10–20 K, in which ionization is a result of galactic cosmic rays since visible and UV light cannot penetrate the dense clouds [156, 158]. The dense clouds also contain particulate matter (referred to as dust or grains). Close to 100 molecular species, as large as 13-atomic, have been detected in interstellar clouds by RF and MM spectroscopy; among these are nine types of molecular positive ion. It is assumed that the neutral molecular species (except H₂) are mainly synthesized through ion–molecule reactions, followed by electron–ion recombination, since neutral–neutral chemical reactions proceed very slowly in the cold temperatures of the clouds, except on grain surfaces. Ion–molecule reactions are typically even faster at lower temperatures. Extensive laboratory studies of ion–molecule reactions, including work at very low temperatures, have mapped out the reaction schemes that take place in interstellar clouds. In dense clouds the reactions



are of paramount importance. These reactions are followed by reactions with C and H₂ to produce CH₃⁺, that subsequently undergoes reaction with many neutral molecules to give ion products such as CH₅⁺, C₂H₅OH₂⁺ and CH₃CNH⁺. Many of the reactions involve radiative association. Dissociative electron–ion recombination then yields neutrals such as CH₄ (methane), C₂H₅OH (ethanol) and CH₃CN (acetonitrile) [158]. It is often joked that diffuse interstellar clouds contain enough grain alcohol to keep space travellers happy on their long journeys. In diffuse clouds, the reaction scheme is more varied and leads to smaller molecules in general.

solute surface which, under the assumption of slip boundary conditions, gives for the correction factor *C* in equation (A3.6.35):

$$C_{\text{size}} = \frac{f_{\text{slip}} V_h}{f_{\text{slip}} V_h + BkT\kappa_T\eta(4/\sigma_r^2 + 1)} \quad (\text{A3.6.36})$$

with isothermal compressibility κ_T , ratio of radii of solvent to solute σ_r and a temperature-dependent parameter *B*. If one compares equation (A3.6.36) with the empirical friction model mentioned above, one

realizes that both contain a factor of the form $C = 1/1 + a\eta$, suggesting that these models might be physically related.

Another, purely experimental possibility to obtain a better estimate of the friction coefficient for rotational motion γ_{rot} in chemical reactions consists of measuring rotational relaxation times τ_{rot} of reactants and calculating it according to equation (A3.6.35) as $\gamma_{\text{rot}} = 6kT\tau_{\text{rot}}$.

A3.6.4 SELECTED REACTIONS

A3.6.4.1 PHOTOISOMERIZATION

According to Kramers' model, for flat barrier tops associated with predominantly small barriers, the transition from the low- to the high-damping regime is expected to occur in low-density fluids. This expectation is borne out by an extensively studied model reaction, the photoisomerization of *trans*-stilbene and similar compounds [70, 71] involving a small energy barrier in the first excited singlet state whose decay after photoexcitation is directly related to the rate coefficient of *trans-cis*-photoisomerization and can be conveniently measured by ultrafast laser spectroscopic techniques.

(A) PRESSURE DEPENDENCE OF PHOTOISOMERIZATION RATE CONSTANTS

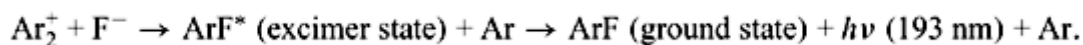
The results of pressure-dependent measurements for *trans*-stilbene in supercritical *n*-pentane [46] (figure A3.6.5) and the prediction from the model described by equation (A3.6.29), using experimentally determined microcanonical rate coefficients in jet-cooled *trans*-stilbene to calculate k_{∞} , show two marked discrepancies between model calculation and measurement: (1) experimental values of k are an order of magnitude higher already at low pressure and (2) the decrease of k due to friction is much less pronounced than predicted. As interpretations for the first observation, several ideas have been put forward that will not be further discussed here, such as a decrease of the effective potential barrier height due to electrostatic solute–solvent interactions enhanced by cluster formation at relatively low pressures [72, 73], or incomplete intramolecular vibrational energy redistribution in the isolated molecule [74, 75, 76, 77, 78, 79 and 80], or Franck–Condon cooling in the excitation process [79, 80]. The second effect, the weak viscosity dependence, which was first observed in solvent series experiments in liquid solution [81, 82 and 83], has also led to controversial interpretations: (i) the macroscopic solvent viscosity is an inadequate measure for microscopic friction acting along the reaction path [84, 85], (ii) the multidimensional character of the barrier crossing process leads to a fractional power dependence of k on $1/\eta$ [54, 81, 86, 87], (iii) as the reaction is very fast, one has to take into account the finite response time of the solvent, i.e. consider frequency-dependent friction [81, 87] and (iv) the effective barrier

Depending on the electron and ion temperatures in a plasma, all of the processes mentioned in this section of the encyclopedia may be taking place simultaneously in the plasma [163]. Understanding, or modelling, the plasma may be quite complicated [164]. Flame chemistry involves charged particles [165]. Most of the early investigations and classifications of electron and ion interactions came about in attempts to understand electric discharges, and these continue today in regard to electric power devices, such as switches and high-intensity lamps [166]. Often the goal is to prevent discharges in the face of high voltages. Military applications involving the earth's ionosphere funded refined work during and following the Second World War. Newer applications such as gas discharge lasers have driven recent studies of plasma chemistry. The rare-gas halide excimer laser is a marvellous example of plasma chemistry, because the lasing molecule may be formed in

recombination between a positive and a negative ion, for example [167, 168 and 169]



or



The Ar_2^+ is formed from $\text{Ar}^+ + \text{Ar}$, where the metastable Ar^* is a product of electron-impact or charge-transfer collisions. The F^- is formed by dissociative electron attachment to F_2 or NF_3 . The population inversion required for light amplification is simple to obtain in the ArF laser since the ground state of the lasing molecule is not bound, except by van der Waals forces and quickly dissociates upon emission of the laser light.

REFERENCES

- [1] Evans R D 1955 *The Atomic Nucleus* (New York: McGraw-Hill)
- [2] Oka T 1992 The infrared spectrum of H_3^+ in laboratory and space plasmas *Rev. Mod. Phys.* **64** 1141–9
- [3] Loeb L B 1960 *Basic Processes of Gaseous Electronics* 2nd edn (Berkeley, CA: University of California)
- [4] Franklin J L (ed) 1979 *Ion–Molecule Reactions, Part I, Kinetics and Dynamics* (Stroudsburg, PA: Dowden, Hutchinson and Ross)
- [5] Muschlitz E E 1957 Formation of negative ions in gases by secondary collision processes *J. Appl. Phys.* **28** 1414–18
- [6] Lavrich D J, Buntine M A, Serxner D and Johnson M A 1993 Excess energy-dependent photodissociation probabilities for O_2^- in water clusters: $\text{O}_2^-(\text{H}_2\text{O})_n$, $1 \leq n \leq 33$ *J. Chem. Phys.* **99** 5910–16
- [7] Lavrich D J, Buntine M A, Serxner D and Johnson M A 1995 Excess energy-dependent photodissociation probabilities for O_2^- in water clusters: $\text{O}_2^-(\text{H}_2\text{O})_n$, $1 \leq n \leq 33$ *J. Phys. C: Solid State Phys.* **99** 8453–7
- [8] Zare R N and Dagdigian P J 1974 Tunable laser fluorescence method for product state analysis *Science* **185** 739–46

- [9] Greene C H and Zare R N 1983 Determination of product population and alignment using laser-induced fluorescence *J. Chem. Phys.* **78** 6741–53
- [10] Crosley D R 1981 Collisional effects on laser-induced fluorescence *Opt. Eng.* **20** 511–21
- [11] Crosley D R 1996 Applications to combustion *Atomic, Molecular, and Optical Physics Handbook* ed G W F Drake (Woodbury, NY: AIP)
- [12] Altkorn R and Zare R N 1984 Effects of saturation on laser-induced fluorescence measurements of population and polarization *Annual Review of Physical Chemistry* ed B S Rabinovitch, J M Schurr and H L Strauss (Palo Alto, CA: Annual Reviews)
- [13] Hamilton C E, Bierbaum V M and Leone S R 1985 Product vibrational state distributions of thermal energy charge transfer reactions determined by laser-induced fluorescence in a flowing afterglow: $\text{Ar}^+ + \text{CO} \rightarrow \text{CO}^+(v = 0-6) + \text{Ar}$ *J. Chem. Phys.* **83** 2284–92

- [14] Sonnenfroh D M and Leone S R 1989 A laser-induced fluorescence study of product rotational state distributions in the charge transfer reaction: $\text{Ar}^+(\text{}^2\text{P}_{3/2}) + \text{N}_2 \rightarrow \text{Ar} + \text{N}_2^+(\text{X})$ at 0.28 and 0.40 eV *J. Chem. Phys.* **90** 1677–85
- [15] Dressler R A, Meyer H and Leone S R 1987 Laser probing of the rotational alignment of N_2^+ drifted in He *J. Chem. Phys.* **87** 6029–39
- [16] Anthony E B, Schade W, Bastian M J, Bierbaum V M and Leone S R 1997 Laser probing of velocity-subgroup dependent rotational alignment of N_2^+ drifted in He *J. Chem. Phys.* **106** 5413–22
- [17] Duncan M A, Bierbaum V M, Ellison G B and Leone S R 1983 Laser-induced fluorescence studies of ion collisional excitation in a drift field: rotational excitation of N_2^+ in He *J. Chem. Phys.* **79** 5448–56
- [18] Leone S R 1989 Laser probing of ion collisions in drift fields: state excitation, velocity distributions, and alignment effects *Gas Phase Bimolecular Collisions* ed M N R Ashford and J E Baggott (London: Royal Society of Chemistry)
- [19] de Gouw J A, Krishnamurthy M and Leone S R 1997 The mobilities of ions and cluster ions drifting in polar gases *J. Chem. Phys.* **106** 5937–42
- [20] Kato S, Frost M J, Bierbaum V M and Leone S R 1994 Vibrational specificity for charge transfer versus deactivation in $\text{N}_2^+(\nu = 0, 1, 2) + \text{Ar}$ and O_2 reactions *Can. J. Chem.* **72** 625–36
- [21] Xie J and Zare R N 1992 Determination of the absolute thermal rate constants for the charge-transfer reaction $\text{DBr}^+(\text{}^2\Pi, \nu^+) + \text{HBr} \rightarrow \text{HBr}^+(\text{}^2\Pi^-, \nu^+) + \text{DBr}$ *J. Chem. Phys.* **96** 4293–302
- [22] Green R J, Xie J, Zare R N, Viggiano A A and Morris R A 1997 Rate constants and products for the reaction of HBr^+ with HBr and DBr *Chem. Phys. Lett.* **277** 1–5
- [23] deNeufville J P, Kasden A and Chimenti R J L 1981 Selective detection of uranium by laser-induced fluorescence: a potential remote-sensing technique. 1: Optical characteristics of uranyl geologic targets *Appl. Opt.* **20** 1279–96
- [24] von Busch F and Dunn G H 1972 Photodissociation of H_2^+ and D_2^+ : experimental *Phys. Rev. A* **5** 1726–43
- [25] Ozenne J-B, Pham D and Durup J 1972 Photodissociation of H_2^+ by monochromatic light with energy analysis of the ejected H^+ ions *Chem. Phys. Lett.* **17** 422–4
- [26] van Asselt N P F B, Maas J G and Los L 1974 Laser induced photodissociation of H_2^+ ions *Chem. Phys. Lett.* **24** 555–8

-28-

- [27] Lee L C, Smith G P, Miller T M and Cosby P C 1978 Photodissociation cross sections of Ar_2^+ , Kr_2^+ , and Xe_2^+ from 6200 to 8600 Å *Phys. Rev. A* **17** 2005–11
- [28] Lee L C and Smith G P 1979 Photodissociation cross sections of Ne_2^+ , Ar_2^+ , Kr_2^+ , and Xe_2^+ , from 3500 to 5400 Å *Phys. Rev. A* **19** 2329–34
- [29] Lee L C, Smith G P, Moseley J T, Cosby P C and Guest J A 1979 Photodissociation and photodetachment of Cl_2^- , ClO^- , Cl_3^- and BrCl_2^- *J. Chem. Phys.* **70** 3237–46
- [30] Moseley J T 1984 Determination of ion molecular potential curves using photodissociative processes *Applied Atomic Collision Physics* ed H S W Massey, E W McDaniel and B Bederson (New York: Academic)
- [31] Moseley J and Durup J 1981 Fast ion beam photofragment spectroscopy *Annual Review of Physical Chemistry* ed B S Rabinovitch, J M Schurr and H L Strauss (Palo Alto, CA: Annual Reviews)
- [32] Moseley J T 1985 Ion photofragment spectroscopy *Photodissociation and Photoionization* ed K P Lawley (New York: Wiley)
- [33] Huber B A, Miller T M, Cosby P C, Zeman H D, Leon R L, Moseley J T and Peterson J R 1977 Laser-ion coaxial beams spectrometer *Rev. Sci. Instrum.* **48** 1306–13

- [34] Lee Y-Y, Leone S R, Champkin P, Kaltoyannis N and Price S D 1997 Laser photofragmentation and collision-induced reactions of SiF_2^+ and SiF_3^+ *J. Chem. Phys.* **106** 7981–94
- [35] Farrar J M 1993 Electronic photodissociation spectroscopy of mass-selected clusters: solvation and the approach to the bulk *Cluster Ions* ed C Y Ng and I Provis (New York: Wiley)
- [36] McDaniel E W 1989 *Atomic Collisions: Electron and Photon Projectiles* (New York: Wiley)
- [37] Pegg D J 1996 Photodetachment *Atomic, Molecular, and Optical Physics Handbook* ed G W F Drake (Woodbury, NY: AIP)
- [38] Dessent C E H and Johnson M A 1998 Fundamentals of negative ion photoelectron spectroscopy *Fundamentals and Applications of Gas Phase Ion Chemistry* ed K R Jennings (Berlin: Kluwer)
- [39] Lineberger W C 1982 Negative ion photoelectron spectroscopy *Applied Atomic Collision Physics, Vol 5, Special Topics* ed H S W Massey, E W McDaniel and B Bederson (New York: Academic)
- [40] Mead R D, Stevens A E and Lineberger W C 1984 Photodetachment in negative ion beams *Gas Phase Ion Chemistry* ed M T Bowers (New York: Academic)
- [41] Drzaic P S, Marks J and Brauman J I 1984 Electron photodetachment from gas phase molecular anions *Gas Phase Ion Chemistry: Ions and Light* ed M T Bowers (New York: Academic)
- [42] Cordermann R R and Lineberger W C 1979 Negative ion spectroscopy *Annual Review of Physical Chemistry* ed B S Rabinovitch, J M Schurr and H L Strauss (Palo Alto, CA: Annual Reviews)
- [43] Miller T M 1981 Photodetachment and photodissociation of ions *Advances in Electronics and Electron Physics* ed L Marton and C Marton (New York: Academic)
- [44] Esaulov A V 1986 Electron detachment from atomic negative ions *Ann. Phys., Paris* **11** 493–592
- [45] Dunbar R C 1979 Ion photodissociation *Gas Phase Ion Chemistry* ed M T Bowers (New York: Academic)

- [46] Wang L, Lee Y T and Shirley D A 1987 Molecular beam photoelectron spectroscopy of SO_2 : geometry, spectroscopy, and dynamics of *J. Chem. Phys.* **87** 2489–97
- [47] Pollard J E, Trevor D J, Lee Y T and Shirley D A 1981 Photoelectron spectroscopy of supersonic molecular beams *Rev. Sci. Instrum.* **52** 1837–46
- [48] Miller T M, Leopold D G, Murray K K and Lineberger W C 1986 Electron affinities of the alkali halides and the structure of their negative ions *J. Chem. Phys.* **85** 2368–75
- [49] Wang L-S, Ding C-F, Wang X-B and Nicholas J B 1998 Probing the potential barriers in intramolecular electrostatic interactions in free doubly charged anions *Phys. Rev. Lett.* at press
- [50] Cooper J and Zare R N 1968 Angular distributions of photoelectrons *J. Chem. Phys.* **48** 942–3
- [51] Yourshaw I, Zhao Y and Neumark D M 1996 Many-body effects in weakly bound anion and neutral clusters: zero electron kinetic energy spectroscopy and threshold photodetachment spectroscopy of Ar_nBr^- ($n = 2-9$) and Ar_nI^- ($n = 2-19$) *J. Chem. Phys.* **105** 351–73
- [52] Wang K and McKoy V 1995 High-resolution photoelectron spectroscopy of molecules *Annual Review of Physical Chemistry* ed H L Strauss, G T Babcock and S R Leone (Palo Alto, CA: Annual Reviews)
- [53] Stenholm S 1996 Absorption and gain spectra *Atomic, Molecular, and Optical Physics Handbook* ed G F W Drake (New York: AIP)
- [54] Miller T A 1982 Light and radical ions *Annual Review of Physical Chemistry* ed B S Rabinovitch, J M Schurr and H L Strauss (Palo Alto, CA: Annual Reviews)
- [55] Woods R C, Saykally R J, Anderson T G, Dixon T A and Szanto P G 1981 The molecular structure of HCO^+ by the

microwave substitution method *J. Chem. Phys.* **75** 4256–60

- [56] Saykally R J and Woods R C 1981 High resolution spectroscopy of molecular ions *Annual Reviews of Physical Chemistry* ed B S Rabinovitch, J M Schurr and H L Strauss (Palo Alto, CA: Annual Reviews)
- [57] Haese N N, Pan F-S and Oka T 1983 Doppler shift and ion mobility measurements of ArH^+ in a He DC glow discharge by infrared laser spectroscopy *Phys. Rev. Lett.* **50** 1575–8
- [58] Gudeman C S and Saykally R J 1984 Velocity modulation infrared laser spectroscopy of molecular ions *Annual Review of Physical Chemistry* ed B S Rabinovitch, J M Schurr and H L Strauss (Palo Alto, CA: Annual Reviews)
- [59] Huet T R, Kabbadj Y, Gabrys C M and Oka T 1994 The $\nu_2 + \nu_3 - \nu_2$ band of NH_3^+ *J. Mol. Spectrosc.* **163** 206–13
- [60] Kabbadj Y, Huet T R, Uy D and Oka T 1996 Infrared spectroscopy of the amidogen ion, NH_2^+ *J. Mol. Spectrosc.* **175** 277–88
- [61] Rosenbaum N H, Owrutsky J C, Tack L M and Saykally R J 1986 Velocity modulation laser spectroscopy of negative ions: the infrared spectrum of hydroxide (OH^-) *J. Chem. Phys.* **84** 5308–13
- [62] Wing W H, Ruff G A, Lamb W E and Spezeski J J 1976 Observation of the infrared spectrum of the hydrogen molecular ion HD^+ *Phys. Rev. Lett.* **36** 1488–91
- [63] Carrington A and McNab I R 1989 The infrared predissociation spectrum of H_3^+ *Accounts Chem. Res.* **22** 218–22
-

-30-

- [64] Neumark D M, Lykke K R, Andersen T and Lineberger W C 1985 Infrared spectrum and autodetachment dynamics of NH^- *J. Chem. Phys.* **83** 4364–73
- [65] Gioumousis G and Stevenson D P 1958 Reactions of gaseous molecule ions with gaseous molecules. V. Theory *J. Chem. Phys.* **29** 294–9
- [66] Su T and Chesnavich W J 1982 Parametrization of the ion–polar molecule collision rate constant by trajectory calculations *J. Chem. Phys.* **76** 5183–5
- [67] Su T 1985 Kinetic energy dependences of ion polar molecule collision rate constants by trajectory calculations *J. Chem. Phys.* **82** 2164–6
- [68] Troe J 1992 Statistical aspects of ion–molecule reactions *State-Selected and State-to-State Ion–Molecule Reaction Dynamics: Theory* ed M Baer M and C-Y Ng (New York: Wiley)
- [69] Clary D C, Smith D and Adams N G 1985 Temperature dependence of rate coefficients for reactions of ions with dipolar molecules *Chem. Phys. Lett.* **119** 320–6
- [70] Bhowmik P K and Su T 1986 Trajectory calculations of ion–quadrupolar molecule collision rate constants *J. Chem. Phys.* **84** 1432–4
- [71] Su T, Viggiano A A and Paulson J F 1992 The effect of the dipole-induced dipole potential on ion–polar molecule collision rate constants *J. Chem. Phys.* **96** 5550–1
- [72] Ikezoe Y, Matsuoka S, Takebe M and Viggiano A A 1987 *Gas Phase Ion–Molecule Reaction Rate Constants Through 1986* (Tokyo: Maruzen)
- [73] Fehsenfeld F C 1975 Associative Detachment *Interactions Between Ions and Molecules* ed P Ausloos (New York: Plenum)
- [74] Viggiano A A and Paulson J F 1983 Temperature dependence of associative detachment reactions *J. Chem. Phys.* **79** 2241–5
- [75] Ferguson E E 1972 Review of laboratory measurements of aeronomic ion–neutral reactions *Ann. Geophys.* **28** 389
- [76] Van Doren J M, Miller T M, Miller A E S, Viggiano A A, Morris R A and Paulson J F 1993 Reactivity of the radical anion

- [77] Adams N G, Bohme D K, Dunkin D B and Fehsenfeld F C 1970 Temperature dependence of the rate coefficients for the reactions of Ar with O₂, H₂, and D₂ *J. Chem. Phys.* **52** 1951
- [78] Rebrion C, Rowe B R and Marquette J B 1989 Reactions of Ar⁺ with H₂N₂O₂ and CO at 20, 30, and 70 K *J. Chem. Phys.* **91** 6142–7
- [79] Midey A J and Viggiano A A 1998 Rate constants for the reaction of Ar⁺ with O₂ and CO as a function of temperature from 300 to 1400 K: derivation of rotational and vibrational energy effects *J. Chem. Phys.* at press
- [80] Dotan I and Viggiano A A 1993 Temperature, kinetic energy, and rotational temperature dependences for the reactions of Ar⁺(²P_{3/2}) with O₂ and CO *Chem. Phys. Lett.* **209** 67–71
- [81] Smith M A 1994 Ion–molecule reaction dynamics at very low temperatures *Unimolecular and Bimolecular Ion–Molecule Reaction Dynamics* ed C-Y Ng, T Baer and I Powis (New York: Wiley)

- [82] Olmstead W N and Brauman J I 1977 Gas phase nucleophilic displacement reactions *J. Am. Chem. Soc.* **99** 4219–28
- [83] Seeley J V, Morris R A, Viggiano A A, Wang H and Hase W L 1997 Temperature dependencies of the rate constants and branching ratios for the reactions of Cl⁻(H₂O)_{0–3} with CH₃Br and thermal dissociation rates for Cl⁻(CH₃Br) *J. Am. Chem. Soc.* **119** 577–84
- [84] Viggiano A A and Morris R A 1996 Rotational and vibrational energy effects on ion–molecule reactivity as studied by the VT-SIFDT technique *J. Phys. Chem.* **100** 19 227–40
- [85] Ferguson E E, Fehsenfeld F C, Dunkin D B, Schmeltekopf A L and Schiff H I 1964 Laboratory studies of helium ion loss processes of interest in the ionosphere *Planet. Space Sci.* **12** 1169–71
- [86] Graul S T and Squires R R 1988 Advances in flow reactor techniques for the study of gas-phase ion chemistry *Mass Spectrom. Rev.* **7** 263–358
- [87] Ferguson E E, Fehsenfeld F C and Schmeltekopf A L 1969 Flowing afterglow measurements of ion–neutral reactions *Adv. At. Mol. Phys.* **5** 1–56
- [88] Ferguson E E 1992 A personal history of the early development of the flowing afterglow technique for ion molecule reactions studies *J. Am. Soc. Mass Spectrom.* **3** 479–86
- [89] Adams N G and Smith D 1976 The selected ion flow tube (SIFT): a technique for studying ion–neutral reactions *Int. J. Mass Spectrom. Ion Phys.* **21** 349
- [90] Smith D and Adams N G 1988 The selected ion flow tube (SIFT): studies of ion–neutral reactions *Adv. At. Mol. Phys.* **24** 1–49
- [91] Adams N G and Smith D 1988 Flowing afterglow and SIFT *Techniques for the Study of Ion–Molecule Reactions* ed J M Farrar and W H Saunders Jr (New York: Wiley)
- [92] Hierl P M *et al* 1996 Flowing afterglow apparatus for the study of ion–molecule reactions at high temperatures *Rev. Sci. Instrum.* **67** 2142–8
- [93] Viggiano A A, Arnold S T and Morris R A 1998 Reactions of mass selected cluster ions in a thermal bath gas *Int. Rev. Phys. Chem.* **17** 147–84
- [94] McFarland M, Albritton D L, Fehsenfeld F C, Ferguson E E and Schmeltekopf A L 1973 Flow-drift technique for ion mobility and ion–molecule reaction rate constant measurements. I. Apparatus and mobility measurements *J. Chem. Phys.* **59** 6610–19
- [95] McFarland M, Albritton D L, Fehsenfeld F C, Ferguson E E and Schmeltekopf A L 1973 Flow-drift technique for ion mobility and ion–molecule reaction rate constant measurements. II. Positive ion reactions of N⁺, O⁺, and N₂⁺ with O₂

and O⁺ with N₂ from thermal to ≈2 eV *J. Chem. Phys.* **59** 6620–8

- [96] McFarland M, Albritton D L, Fehsenfeld F C, Ferguson E E and Schmeltekopf A L 1973 Flow-drift technique for ion mobility and ion–molecule reaction rate constant measurements. III. Negative ion reactions of O[−] + CO, NO, H₂, and D₂ *J. Chem. Phys.* **59** 6629–35
- [97] Lindinger W and Smith D 1983 Influence of translational and internal energy on ion–neutral reactions *Reactions of Small Transient Species* ed A Fontijn and M A A Clyne (New York: Academic)
- [98] Viehland L A and Robson R E 1989 Mean energies of ion swarms drifting and diffusing through neutral gases *Int. J. Mass Spectrom. Ion Processes* **90** 167–86
-

-32-

- [99] Kemper P R and Bowers M T 1988 Ion cyclotron resonance spectrometry *Techniques for the Study of Ion–Molecule Reactions* ed J M Farrar and W H Saunders Jr (New York: Wiley)
- [100] Freiser B S 1988 Fourier Transform Mass Spectrometry *Techniques for the Study of Ion–Molecule Reactions* ed J M Farrar and W H Saunders Jr (New York: Wiley)
- [101] Dunbar R C 1994 Ion–molecule radiative association *Unimolecular and Bimolecular Ion–Molecule Reaction Dynamics* ed C-Y Ng, T Baer and I Powis (New York: Wiley)
- [102] Heninger M, Fenistein S, Durup-Ferguson M, Ferguson E E, Marx R and Mauclaire G 1986 Radiative lifetime for $v = 1$ and $v = 2$ ground state NO⁺ ions *Chem. Phys. Lett.* **131** 439–43
- [103] Gerlich D 1992 Inhomogeneous RF fields: a versatile tool for the study of processes with slow ions *State-Selected and State-to-State Ion–Molecule Reaction Dynamics: Part 1. Experiment* ed C Ng and M Baer (New York: Wiley)
- [104] Watson L R, Thiem T L, Dressler R A, Salter R H and Murad E 1993 High temperature mass spectrometric studies of the bond energies of gas-phase ZnO, NiO, and CuO *J. Phys. Chem.* **97** 5577–80
- [105] Schultz R H and Armentrout P B 1991 Reactions of N₂⁺ with rare gases from thermal to 10 eV center of mass energy: collision induced dissociation, charge transfer, and ligand exchange *Int. J. Mass Spectrom. Ion Proc.* **107** 29–48
- [106] Futrell J H 1992 Crossed-molecular beam studies of state-to-state reaction dynamics *State-Selected and State-to-State Ion–Molecule Reaction Dynamics: Part 1. Experiment* ed C Ng and M Baer (New York: Wiley)
- [107] Kebarle P 1988 Pulsed electron high pressure mass spectrometer *Techniques for the Study of Ion–Molecule Reactions* ed J M Farrar and W H Saunders Jr (New York: Wiley)
- [108] Knighton W B and Grimsrud E P 1996 Gas phase ion chemistry under conditions of very high pressure *Advances in Gas Phase Ion Chemistry* ed N G Adams and L M Babcock (JAI)
- [109] Seeley J V, Jayne J T and Molina M J 1993 High pressure fast-flow technique for gas phase kinetics studies *Int. J. Chem. Kinet.* **25** 571–94
- [110] Huber K P and Herzberg G 1979 *Molecular Spectra and Molecular Structure. IV. Constants of Diatomic Molecules* (New York: Van Nostrand Reinhold)
- [111] Mallard W G and Linstrom P J (eds) 1988 *NIST Standard Reference Database 69*:<http://webbook.nist.gov/chemistry/>
- [112] Bates D R 1991 Negative ions: structure and spectra *Advances in Atomic, Molecular and Optical Physics* ed D R Bates and B Bederson (New York: Academic)
- [113] Shida T 1991 Photochemistry and spectroscopy of organic ions and radicals *Annual Review of Physical Chemistry* ed H L Strauss, G T Babcock and S R Leone (Palo Alto, CA: Annual Reviews)
- [114] Castleman A W and Märk T D 1986 Cluster ions: their formation, properties, and role in elucidating the properties of matter in the condensed state *Gaseous Ion Chemistry and Mass Spectrometry* ed J H Futrell (New York: Wiley)
- [115] Van Doren J M, Barlow S E, DePuy C H, Bierbaum V M, Dotan I and Ferguson E E 1986 Chemistry and structure of

the CH_3O_2^+ product of the $\text{O}_2^+ + \text{CH}_4$ reaction *J. Phys. Chem.* **90** 2772–7

- [116] Papanikolas J M, Gord J R, Levinger N E, Ray D, Vorsa V and Lineberger W C 1991 Photodissociation and geminate recombination dynamics of I_2^- in mass-selected $\text{I}_2^-(\text{CO}_2)_n$ cluster ions *J. Phys. Chem.* **90** 8028–40
-

-33-

- [117] Lias S G, Bartmess J E, Liebman J F, Holmes J L, Levin R D and Mallard W G 1988 Gas-phase ion and neutral thermochemistry *J. Phys. Chem. Ref. Data* **17**, **Supplement 1** 1-861
- [118] Lias S G 1998 Ionization energy evaluation *NIST Chemistry WebBook, NIST Standard Reference Database Number 69* ed W G Mallard and P J Linstrom (Gaithersburg, MD: National Institute of Standards and Technology)
- [119] Lias S G 1997 Ionization energies of gas phase molecules *Handbook of Chemistry and Physics* ed D R Lide (Boca Raton, FL: CRC Press)
- [120] Hunter E P and Lias S G 1998 Proton affinity evaluation (WebBook) *NIST Standard Reference Database Number 69* ed W G Mallard and P J Linstrom (Gaithersburg, MD: National Institute of Standards and Technology)
- [121] Bartmess J E 1998 Negative ion energetics data *NIST Chemistry WebBook, NIST Standard Reference Database Number 69* ed W G Mallard and P J Linstrom (Gaithersburg, MD: National Institute of Standards and Technology)
- [122] Miller T M 1997 Electron affinities *Handbook of Chemistry and Physics* ed D R Lide (Boca Raton, FL: CRC)
- [123] Mallard W G and Linstrom P J (eds) 1998 *NIST Standard Reference Database No 69 (Gaithersburg)*
<http://webbook.nist.gov>
- [124] Ervin K M, Gronert S, Barlow S E, Gilles M K, Harrison A G, Bierbaum V M, DePuy C H, Lineberger W C and Ellison G B 1990 Bond strengths of ethylene and acetylene 1990 *J. Am. Chem. Soc.* **112** 5750–9
- [125] Keesee R G and Castleman A W Jr 1986 Thermochemical data on gas-phase ion–molecule association and clustering reactions *J. Phys. Chem. Ref. Data* **15** 1011
- [126] Armentrout P B and Baer T 1996 Gas phase ion dynamics and chemistry *J. Phys. Chem.* **100** 12 866–77
- [127] Castleman A W Jr and Bowen K H Jr 1996 Clusters: structure, energetics, and dynamics of intermediate states of matter *J. Phys. Chem.* **100** 12 911–44
- [128] Viggiano A A and Arnold F 1995 Ion chemistry and composition of the atmosphere *Atmospheric Electrodynamics* ed H Volland (Boca Raton, FL: CRC Press)
- [129] Viggiano A A 1993 *In-situ* mass spectrometry and ion chemistry in the stratosphere and troposphere *Mass Spectrom. Rev.* **12** 115–37
- [130] Ferguson E E, Fehsenfeld F C and Albritton D L 1979 Ion chemistry of the earth's atmosphere *Gas Phase Ion Chemistry* ed M T Bowers (San Diego, CA: Academic)
- [131] Ferguson E E 1979 Ion–molecule reactions in the atmosphere *Kinetics of Ion–Molecule Reactions* ed P Ausloos (New York: Plenum)
- [132] Johnson M A and Lineberger W C 1988 Pulsed methods for cluster ion spectroscopy *Techniques for the Study of Ion–Molecule Reactions* ed J M Farrar and W H Saunders Jr (New York: Wiley)
- [133] Bohme D K and Raksit A B 1984 Gas phase measurements of the influence of stepwise solvation on the kinetics of nucleophilic displacement reactions with CH_3Cl and CH_3Br at room temperature *J. Am. Chem. Soc.* **106** 3447–52
- [134] Viggiano A A, Deakyne C A, Dale F and Paulson J F 1987 Neutral reactions in the presence of alkali ions *J. Chem. Phys.* **87** 6544–52
- [135] Vorsa V, Campagnola P J, Nandi S, Larsson M and Lineberger W C 1996 Protofragments of $\text{I}_2^- \cdot \text{Ar}_n$ clusters: observation of metastable isomeric ionic fragments *J. Chem. Phys.* **105** 2298–308
-

- [136] Kroto H W, Heath J R, O'Brian S C, Curl R F and Smalley R E 1985 C_{60} : Buckminsterfullerene *Nature* **318** 162–3
- [137] Guo B C, Kerns K P and Castleman A W Jr 1992 $Ti_8C_{12}^+$ -metallo-carbohedrenes: a new class of molecular clusters? *Science* **255** 1411–13
- [138] Reid G C 1976 Ion chemistry of the D-region *Advances in Atomic and Molecular Physics* ed D R Bates and B Bederson (Orlando, FL: Academic)
- [139] Smith D and Adams N G 1980 Elementary plasma reactions of environmental interest *Topics in Current Chemistry* ed F L Boschke (Berlin: Springer)
- [140] Ferguson E E and Arnold F 1981 Ion chemistry of the stratosphere *Accounts. of Chem. Res.* **14** 327–34
- [141] Thomas L 1983 Modelling of the ion composition of the middle atmosphere *Ann. Geophys.* **1** 61–73
- [142] Arnold F and Viggiano A A 1986 Review of rocket-borne ion mass spectrometry in the middle atmosphere *Middle Atmosphere Program Handbook, Vol. 19* ed R A Goldberg (Urbana, IL: SCOSTEP)
- [143] Brasseur G and Solomon S 1986 *Aeronomy of the Middle Atmosphere* 2nd edn (Boston, MA: D Reidel)
- [144] Brasseur G and De Baets P 1986 Ions in the mesosphere and lower thermosphere: a two-dimensional model *J. Geophys. Res.* **91** 4025–46
- [145] Viggiano A A, Morris R A and Paulson J F 1994 Effects of O_2^- and SF_6 vibrational energy on the rate constant for charge transfer between O_2^- and SF_6 *Int. J. Mass Spectrom. Ion Processes* **135** 31–7
- [146] Arnold F 1980 The middle atmosphere ionized component *Vth ESA-PAC Symposium on European Rocket and Balloon Programmes and Related Research* (Bournemouth, UK: ESA) pp 479–95
- [147] Book D L 1987 *NRL Plasma Formulary* (Washington, DC: Naval Research Laboratory)
- [148] Hierl P M, Dotan I, Seeley J V, Van Doren J M, Morris R A and Viggiano A A 1997 Rate constants for the reactions of O^+ with N_2 and O_2 as a function of temperature (300–1800 K) *J. Chem. Phys.* **106** 3540–4
- [149] Eisele F L 1986 Identification of tropospheric ions *J. Geophys. Res.* **91** 7897–906
- [150] Hauck G and Arnold F 1984 Improved positive-ion composition measurements in the upper troposphere and lower stratosphere and the detection of acetone *Nature* **311** 547–50
- [151] Schlager H and Arnold F 1985 Balloon-borne fragment ion mass spectrometry studies of stratospheric positive ions: unambiguous detection of $H^+(CH_3CN)_1(H_2O)$ -clusters *Planet. Space Sci.* **33** 1363–6
- [152] Viggiano A A and Arnold F 1981 The first height measurements of the negative ion composition of the stratosphere *Planet. Space Sci.* **29** 895–906
- [153] Arnold F and Henschen G 1978 First mass analysis of stratospheric negative ions *Nature* **257** 521–2
- [154] Eisele F L 1989 Natural and anthropogenic negative ions in the troposphere *J. Geophys. Res.* **94** 2183–96
- [155] Oka T 1997 Water on the sun—molecules everywhere *Science* **277** 328–9

- [156] Dalgarno A and Lepp S 1996 Applications of atomic and molecular physics to astrophysics *Atomic, Molecular, and Optical Physics Handbook* ed G W F Drake (Woodbury, NY: AIP)

- [157] Dalgarno A and Fox J 1994 Ion chemistry in atmospheric and astrophysical plasmas *Unimolecular and Bimolecular Ion–Molecule Reaction Dynamics* ed C-Y Ng, T Baer and I Powis (New York: Wiley)
- [158] Smith D and Spanel P 1995 Ions in the terrestrial atmosphere and in interstellar clouds *Mass Spectrom. Rev.* **14** 255–78
- [159] Dalgarno A 1994 Terrestrial and extraterrestrial H_3^+ *Advances in Atomic, Molecular and Optical Physics* ed B Bederson and A Dalgarno (New York: Academic)
- [160] Fox J L 1996 Aeronomy *Atomic, Molecular, and Optical Physics Handbook* ed G W F Drake (Woodbury, NY: AIP)
- [161] Geballe T R and Oka T 1996 Detection of H_3^+ in interstellar space *Nature* **384** 334–5
- [162] Haeberli R M, Altwegg K, Balsiger H and Geiss J 1995 Physics and chemistry of ions in the pile-up region of comet P/Halley *Astron. Astrophys.* **297** 881–91
- [163] Capitelli M, Celiberto R and Cacciatore M 1994 Needs for cross sections in plasma chemistry *Advances in Atomic, Molecular and Optical Physics* ed B Bederson, H Walther and M Inokuti (New York: Academic)
- [164] Garscadden A 1996 Conduction of electricity in gases *Atomic, Molecular, and Optical Handbook* ed G W F Drake (Woodbury, NY: AIP)
- [165] Fontijn A 1982 Combustion and flames *Applied Atomic Collision Physics, Vol 5, Special Topics* ed H S W Massey, E W McDaniel and B Bederson (New York: Academic)
- [166] Weymouth J F 1982 Collision phenomena in electrical discharge lamps *Applied Atomic Collision Physics, Vol 5, Special Topics* ed H S W Massey, E W McDaniel and B Bederson (New York: Academic)
- [167] Huestis D L 1982 Introduction and overview *Applied Atomic Collision Physics, Vol 3, Gas Lasers* ed H S W Massey, E W McDaniel, B Bederson and W L Nighan (New York: Academic)
- [168] Chantry P J 1982 Negative ion formation in gas lasers *Applied Atomic Collision Physics Vol 3, Gas Lasers* ed H S W Massey, E W McDaniel, B Bederson and W L Nighan (New York: Academic)
- [169] Rokni M and Jacob J H 1982 Rare-gas halide lasers *Applied Atomic Collision Physics, Vol 3, Gas Lasers* ed H S W Massey, E W McDaniel, B Bederson and W L Nighan (New York: Academic)

FURTHER READING

Farrar J M and Saunders W H Jr (eds) 1988 *Techniques for the Study of Ion–Molecule Reactions* (New York: Wiley)

The best place to start for a detailed look at the instrumentation for the study of ion–molecule chemistry.

Ng C-Y, Baer T and Powis I (eds) 1994 *Unimolecular and Bimolecular Ion–Molecule Reaction Dynamics* (New York: Wiley)

An excellent reference for recent work on ion chemistry.

Ng C-Y and Baer T (eds) 1992 *State-Selected and State-to-State Ion–Molecule Reaction Dynamics* vols 1 and 2 (New York: Wiley)

A comprehensive look at the effect of state selection on ion–molecule reactions from both experimental and theoretical viewpoints.

Bowers M T (ed) 1979 and 1984 *Gas Phase Ion Chemistry* vols 1–3 (New York: Academic)

An older look at the field of ion chemistry. Most of the concepts are still valid and this series is a good foundation for a beginner in the field.

On ongoing series about current topics in ion chemistry.

Drake G W F (ed) 1996 *Atomic, Molecular, and Optical Physics Handbook* (Woodbury, NY: AIP)

Fundamental chemical physics descriptions of both ion and neutral processes.

-1-

A3.6 Chemical kinetics in condensed phases

Jorg Schroeder

A3.6.1 INTRODUCTION

The transition from the low-pressure gas to the condensed phase is accompanied by qualitative changes in the kinetics of many reactions caused by the presence of a dense solvent environment permanently interacting with reactants (also during their motion along the reaction path). Though this solvent influence in general may be a complex phenomenon as contributions of different origin tend to overlap, it is convenient to single out aspects that dominate the kinetics of certain types of reactions under different physical conditions.

Basic features of solvent effects can be illustrated by considering the variation of the rate constant k_{uni} of a unimolecular reaction as one gradually passes from the low-pressure gas phase into the regime of liquid-like densities [1] (see [figure A3.6.1](#).) At low pressures, where the rate is controlled by thermal activation in isolated binary collisions with bath gas molecules, k_{uni} is proportional to pressure, i.e. it is in the low-pressure limit k_0 . Raising the pressure further, one reaches the fall-off region where the pressure dependence of k_{uni} becomes increasingly weaker until, eventually, it attains the constant so-called high-pressure limit k_∞ . At this stage, collisions with bath gas molecules, which can still be considered as isolated binary events, are sufficiently frequent to sustain an equilibrium distribution over rotational and vibrational degrees of freedom of the reactant molecule, and k_∞ is determined entirely by the intramolecular motion along the reaction path. k_∞ may be calculated by statistical theories (see [chapter A3.4](#)) if the potential-energy (hyper)surface (PES) for the reaction is known. What kind of additional effects can be expected, if the density of the compressed bath gas approaches that of a dense fluid? Ideally, there will be little further change, as equilibration becomes even more effective because of permanent energy exchange with the dense heat bath. So, even with more confidence than in the gas phase, one could predict the rate constant using statistical reaction rate theories such as, for example, transition state theory (TST). However, this ideal picture may break down if (i) there is an appreciable change in charge distribution or molar volume as the system moves along the reaction path from reactant to product state, (ii) the reaction entails large-amplitude structural changes that are subject to solvent frictional forces retarding the motion along the reaction path or (iii) motion along the reaction path is sufficiently fast that thermal equilibrium over all degrees of freedom of the solute and the bath cannot be maintained.

- (i) This situation can still be handled by quasi-equilibrium models such as TST, because the solvent only influences the equilibrium energetics of the system. The ensuing phenomena may be loosely referred to as ‘static’ solvent effects. These may be caused by electronic solute–solvent interactions that change the effective PES by shifting intersection regions of different electronic states or by lowering or raising potential-energy barriers, but also by solvent structural effects that influence the

free-energy change along the reaction path associated with variations in molar volume.

- (ii) The decrease of the rate constant due to the viscous drag exerted by the solvent medium requires an extension of statistical rate models to include diffusive barrier crossing, because the no-recrossing postulate of TST is obviously violated. In the so-called Smoluchowski limit, one would expect an inverse dependence of k_{uni} on solvent viscosity η at sufficiently high pressure. A reaction rate constant is still well defined and kinetic rate equations may be used to describe the course of the reaction.

-2-

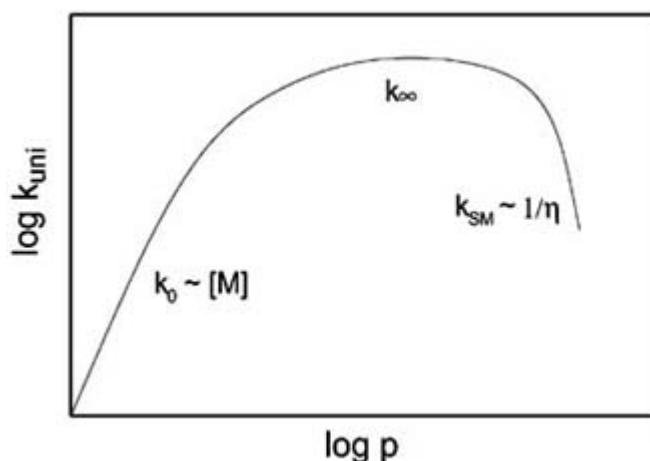


Figure A3.6.1. Pressure dependence of unimolecular rate constant k_{uni} .

This is no longer the case when (iii) motion along the reaction path occurs on a time scale comparable to other relaxation times of the solute or the solvent, i.e. the system is partially non-relaxed. In this situation dynamic effects have to be taken into account explicitly, such as solvent-assisted intramolecular vibrational energy redistribution (IVR) in the solute, solvent-induced electronic surface hopping, dephasing, solute–solvent energy transfer, dynamic caging, rotational relaxation, or solvent dielectric and momentum relaxation.

The introductory remarks about unimolecular reactions apply equivalently to bimolecular reactions in condensed phase. An essential additional phenomenon is the effect the solvent has on the rate of approach of reactants and the lifetime of the collision complex. In a dense fluid the rate of approach evidently is determined by the mutual diffusion coefficient of reactants under the given physical conditions. Once reactants have met, they are temporarily trapped in a solvent cage until they either diffusively separate again or react. It is common to refer to the pair of reactants trapped in the solvent cage as an encounter complex. If the ‘unimolecular’ reaction of this encounter complex is much faster than diffusive separation: i.e., if the effective reaction barrier is sufficiently small or negligible, the rate of the overall bimolecular reaction is diffusion controlled.

As it has appeared in recent years that many fundamental aspects of elementary chemical reactions in solution can be understood on the basis of the dependence of reaction rate coefficients on solvent density [2, 3, 4 and 5], increasing attention is paid to reaction kinetics in the gas-to-liquid transition range and supercritical fluids under varying pressure. In this way, the essential differences between the regime of binary collisions in the low-pressure gas phase and that of a dense environment with typical many-body interactions become apparent. An extremely useful approach in this respect is the investigation of rate coefficients, reaction yields and concentration–time profiles of some typical model reactions over as wide a pressure range as possible, which permits the continuous and well controlled variation of the physical properties of the solvent. Among these the most important are density, polarity and viscosity in a continuum description or collision frequency,

Progress in the theoretical description of reaction rates in solution of course correlates strongly with that in other theoretical disciplines, in particular those which have profited most from the enormous advances in computing power such as quantum chemistry and equilibrium as well as non-equilibrium statistical mechanics of liquid solutions where Monte Carlo and molecular dynamics simulations in many cases have taken on the traditional role of experiments, as they allow the detailed investigation of the influence of intra- and intermolecular potential parameters on the microscopic dynamics not accessible to measurements in the laboratory. No attempt, however, will be made here to address these areas in more than a cursory way, and the interested reader is referred to the corresponding chapters of the encyclopedia.

In the sections below a brief overview of static solvent influences is given in A3.6.2, while in A3.6.3 the focus is on the effect of transport phenomena on reaction rates, i.e. diffusion control and the influence of friction on intramolecular motion. In A3.6.4 some special topics are addressed that involve the superposition of static and transport contributions as well as some aspects of dynamic solvent effects that seem relevant to understanding the solvent influence on reaction rate coefficients observed in homologous solvent series and compressed solution. More comprehensive accounts of dynamics of condensed-phase reactions can be found in [chapter A3.8](#), [chapter A3.13](#), [chapter B3.3](#), [chapter C3.1](#), [chapter C3.2](#) and [chapter C3.5](#).

A3.6.2 STATIC SOLVENT EFFECTS

The treatment of equilibrium solvation effects in condensed-phase kinetics on the basis of TST has a long history and the literature on this topic is extensive. As the basic ideas can be found in most physical chemistry textbooks and excellent reviews and monographs on more advanced aspects are available (see, for example, the recent review article by Truhlar *et al* [6] and references therein), the following presentation will be brief and far from providing a complete picture.

A3.6.2.1 SEPARATION OF TIME SCALES

A reactive species in liquid solution is subject to permanent random collisions with solvent molecules that lead to statistical fluctuations of position, momentum and internal energy of the solute. The situation can be described by a reaction coordinate X coupled to a huge number of solvent bath modes. If there is a reaction barrier E_0^\pm ('+' refers to the forward direction and '-' to the reverse reaction), in a way similar to what is common in gas phase reaction kinetics, one may separate the reaction into the elementary steps of activation of A or B, barrier crossing, and equilibration of B or A, respectively (see [figure A3.6.2](#).) The time scale τ_r^\pm for mounting and crossing the barrier is determined by the magnitude of statistical fluctuations $X(t) = \langle X(t) \rangle$ at temperature T , where $\langle \rangle$ indicates ensemble average. In a canonical ensemble this is mainly the Boltzmann factor $\tau_r^\pm \sim e^{E_0^\pm/kT}$, where k denotes Boltzmann's constant. Obviously, the reaction is a rare event if the barrier is large. On the other hand, the time scale for energy relaxation τ_s in a potential well is inversely proportional to the curvature of the potential V along X ,

$$\tau_s \sim \sqrt{\mu / \left. \frac{\partial^2 V}{\partial X^2} \right|_{X=A,B}}$$

where μ denotes reduced mass. So the overall time scale for the reaction

$$\tau_r^\pm \approx \tau_{s(A,B)} \exp(E_0^\pm/kT) \gg \tau_{s(A,B)}$$

for $E_0^\pm \gg kT$. If at the same time τ_r^\pm is also significantly larger than all other relevant time constants of the solute–bath system (correlation time of the bath, energy and momentum relaxation time, barrier passage time), X may be considered to be a random variable and the motion of the reacting species along this reaction coordinate a stochastic Markov process under the influence of a statistically fluctuating force. This simply means that before, during and after the reaction all degrees of freedom of the solute–solvent system but X are in thermodynamic equilibrium. In this case quasi-equilibrium models of reaction rates are applicable. If the additional requirements are met that (i) each trajectory crossing the transition state at the barrier top never recrosses and (ii) the Born–Oppenheimer approximation is fulfilled, TST can be used to calculate the reaction rate and provide an upper limit to the real rate coefficient (see [chapter A3.12](#)).

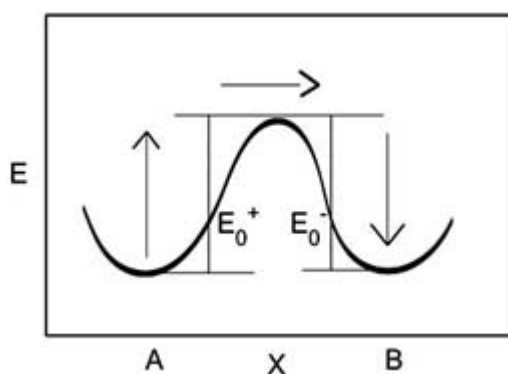


Figure A3.6.2. Activation and barrier crossing.

A3.6.2.2 THERMODYNAMIC FORMULATION OF TST AND REFERENCE STATES

For analysing equilibrium solvent effects on reaction rates it is common to use the thermodynamic formulation of TST and to relate observed solvent-induced changes in the rate coefficient to variations in Gibbs free-energy differences between solvated reactant and transition states with respect to some reference state. Starting from the simple one-dimensional expression for the TST rate coefficient of a unimolecular reaction $A \xrightarrow{k_{\text{TST}}} P$

$$k_{\text{TST}} = \frac{kT}{h} \frac{Q^\ddagger}{Q_A} \exp(-E_0/kT) = \frac{kT}{h} \frac{[A^\ddagger]}{[A]} \quad (\text{A3.6.1})$$

where Q_A and Q^\ddagger denote the partition functions of reactant and transition state per volume, respectively, E_0 is the barrier height, and $[A]$, $[A^\ddagger]$ stand for equilibrium concentration of reactant and, in a formal not physical sense, the transition state, respectively. Defining an equilibrium constant in terms of activities a

-5-

$$K_a^\ddagger \equiv \frac{a^\ddagger}{a_A} = \frac{\gamma^\ddagger [A^\ddagger]}{\gamma_A [A]} = \frac{Q^\ddagger}{Q_A} \exp(-E_0/kT) \quad (\text{A3.6.2})$$

with corresponding activity coefficients denoted by γ , one obtains for the rate coefficient from [equation](#)

(A3.6.1) and equation (A3.6.2)

$$k_{\text{TST}} = \frac{kT}{h} \frac{[A^\ddagger]}{[A]} = \frac{kT}{h} \frac{Q^\ddagger}{Q_A} \exp(-E_0/kT) \frac{\gamma_A}{\gamma^\ddagger} \equiv k_{\text{TST}}^0 \frac{\gamma_A}{\gamma^\ddagger} \quad (\text{A3.6.3})$$

where k_{TST}^0 is a standard rate coefficient which depends on the reference state chosen. If one uses the dilute-gas phase as reference, i.e. $k_{\text{TST}}^0 = k_{\text{gas}}$, all equilibrium solvation effects according to equation (A3.6.3) are included in the ratio of activity coefficients γ_A/γ^\ddagger which is related to the Gibbs free energy of activation for the reaction in the gas $\Delta G_{\text{gas}}^\ddagger$ and in solution $\Delta G_{\text{solution}}^\ddagger$;

$$kT \ln \left(\frac{k_{\text{solution}}}{k_{\text{gas}}} \right) = kT \ln \left(\frac{\gamma_A}{\gamma^\ddagger} \right) = \Delta G_{\text{gas}}^\ddagger - \Delta G_{\text{solution}}^\ddagger. \quad (\text{A3.6.4})$$

Since $\Delta G_{\text{gas}}^\ddagger - \Delta G_{\text{solution}}^\ddagger$ in equation (A3.6.4) is equal to the difference between the Gibbs free energy of solvation of reactant and transition state, $\Delta G_{\text{sol}}(A) - \Delta G_{\text{sol}}(A^\ddagger)$, one has a direct correlation between equilibrium solvation free enthalpies and rate coefficient ratios. It is common practice in physical organic chemistry to use as a reference state not the gas phase, but a suitable reference solvent M, such that one correlates measured rate coefficient ratios of equation (A3.6.4) to relative changes in Gibbs free energy of solvation

$$[\Delta G_{\text{sol,S}}(A) - \Delta G_{\text{sol,S}}(A^\ddagger)] - [\Delta G_{\text{sol,M}}(A) - \Delta G_{\text{sol,M}}(A^\ddagger)] \equiv \delta_M \Delta G^\ddagger. \quad (\text{A3.6.5})$$

The shorthand notation in the rhs of equation (A3.6.5) is frequently referred to as the Leffler–Grunwald operator [7].

Considering a bimolecular reaction $A+B \xrightarrow{k_{\text{TST}}} P$, one correspondingly obtains for the rate constant ratio

$$k_{\text{solution}}/k_{\text{gas}} = \gamma_A \gamma_B / \gamma^\ddagger. \quad (\text{A3.6.6})$$

In the TST limit, the remaining task strictly speaking does not belong to the field of reaction kinetics: it is a matter of obtaining sufficiently accurate reactant and transition state structures and charge distributions from quantum chemical calculations, constructing sufficiently realistic models of the solvent and the solute–solvent interaction potential, and calculating from these ingredients values of Gibbs free energies of solvation and activity coefficients. In many cases, a microscopic description may prove a task too complex, and one rather has to use simplifying approximations to characterize influences of different solvents on the kinetics of a reaction in terms of some macroscopic physical or empirical solvent parameters. In many cases, however, this approach is sufficient to capture the kinetically significant contribution of the solvent–solute interactions.

A3.6.2.3 EQUILIBRIUM SOLVATION—MACROSCOPIC DESCRIPTION

(A) NAÏVE VIEW OF SOLVENT CAVITY EFFECTS

Considering equation (A3.6.3), if activity coefficients of reactant and transition state are approximately equal, for a unimolecular reaction one should observe $k_{\text{solution}} \approx k_{\text{gas}}$. This in fact is observed for many unimolecular reactions where the reactant is very similar to the transition state, i.e. only a few bond lengths and angles change by a small amount and there is an essentially constant charge distribution. There are, however, also large deviations from this simplistic prediction, in particular for dissociation reactions that require separation

of fragments initially formed inside a common solvent cavity into individually solvated products.

For a bimolecular reaction in such a case one obtains from equation (A3.6.6) $k_{\text{solution}} \approx \gamma \cdot k_{\text{gas}}$, so one has to estimate the activity coefficient of a reactant to qualitatively predict the solvent effect. Using *ad hoc* models of solvation based on the free-volume theory of liquids or the cohesive energy density of a solvent cavity, purely thermodynamic arguments yield $\gamma \sim 10^2 - 10^3$ [8, 9 and 10].

The reason for this enhancement is intuitively obvious: once the two reactants have met, they temporarily are trapped in a common solvent shell and form a short-lived so-called encounter complex. During the lifetime of the encounter complex they can undergo multiple collisions, which give them a much bigger chance to react before they separate again, than in the gas phase. So this effect is due to the microscopic solvent structure in the vicinity of the reactant pair. Its description in the framework of equilibrium statistical mechanics requires the specification of an appropriate interaction potential.

(B) ELECTROSTATIC EFFECTS—ONSAGER AND BORN MODELS

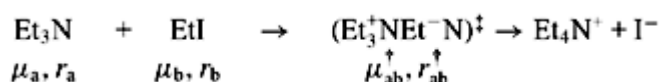
If the charge distribution changes appreciably during the reaction, solvent polarity effects become dominant and in liquid solution often mask the structural influences mentioned above. The calculation of solvation energy differences between reactant and transition state mainly consists of estimating the Gibbs free energies of solvation ΔG_{sol} of charges, dipoles, quadrupoles etc in a polarizable medium. If the solute itself is considered non-polarizable and the solvent a continuous linear dielectric medium without internal structure, then $\Delta G_{\text{sol}} = \frac{1}{2} E_{\text{int}}$, where E_{int} is the solute-solvent interaction energy [11]. Reactant and transition state are modelled as point charges or point dipoles situated at the centre of a spherical solvent cavity. The point charge or the point dipole will polarize the surrounding dielectric continuum giving rise to an electric field which in turn will act on the charge distribution inside the cavity. The energy of the solute in this so-called reaction field may be calculated by a method originally developed by Onsager. Using his reaction field theory [12, 13], one obtains the molar Gibbs free energy of solvation (with respect to vacuum) of an electric point dipole μ_{el} in a spherical cavity of radius r embedded in a homogeneous dielectric of dielectric constant ϵ as

$$\Delta G_{\text{sol,dip}} = -N_{\text{A}} \frac{\epsilon - 1}{2\epsilon + 1} \frac{\mu_{\text{el}}^2}{4\pi \epsilon_0 r^3} \quad (\text{A3.6.7})$$

with ϵ_0 and N_{A} denoting vacuum permittivity and Avogadro's constant, respectively. The dielectric constant inside the cavity in this approximation is assumed to be unity. Applying this expression to a solvent series study of a reaction

-7-

involving large charge separation, such as the Menshutkin reaction of triethylamine with ethyliodide



one obtains

$$\delta_M \Delta G^\ddagger = -N_A \frac{\epsilon - 1}{2\epsilon + 1} \frac{1}{4\pi \epsilon_0} \left(\frac{(\mu_{ab}^\ddagger)^2}{(r_{ab}^\ddagger)^3} - \frac{\mu_a^2}{r_a^3} - \frac{\mu_b^2}{r_b^3} \right)$$

predicting a linear relationship between $\ln(k_{\text{solvent}}/k_{\text{reference}})$ and $(\epsilon - 1)/(2\epsilon + 1)$ which is only approximately reflected in the experimental data covering a wide range of solvents [14] (see figure A3.6.3). This is not surprising, in view of the approximate character of the model and, also, because a change of solvent does not only lead to a variation in the dielectric constant, but at the same time may be accompanied by a change in other kinetically relevant properties of the medium, demonstrating a general weakness of this type of experimental approach.

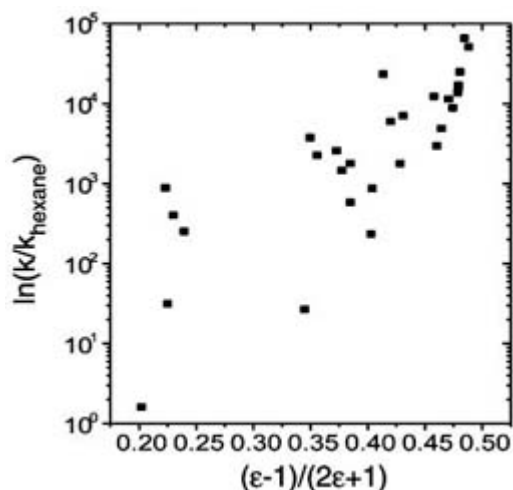


Figure A3.6.3. Solvent polarity dependence of the rate constant for the Menshutkin reaction (data from [14]).

Within the framework of the same dielectric continuum model for the solvent, the Gibbs free energy of solvation of an ion of radius r_{ion} and charge $z_{\text{ion}}e$ may be estimated by calculating the electrostatic work done when hypothetically charging a sphere at constant radius r_{ion} from $q = 0 \rightarrow q = z_{\text{ion}}e$. This yields the Born equation [13]

$$\Delta G_{\text{sol,ion}} = -N_A \frac{z_{\text{ion}}^2 e^2}{8\pi \epsilon_0 r_{\text{ion}}} \left(1 - \frac{1}{\epsilon} \right) \quad (\text{A3.6.8})$$

-8-

such that for a reaction of the type



the change in effective barrier height (difference of Gibbs free energy of solvation changes between transition state and reactants) according to equation (A3.6.7) and equation (A3.6.8) equals

$$\delta_M \Delta G^\ddagger = -\frac{N_A}{8\pi \epsilon_0} \left[(ze)^2 \left(1 - \frac{1}{\epsilon} \right) \left(\frac{1}{r^\ddagger} - \frac{1}{r_A} \right) - \frac{\mu_B^2}{r_B^3} \frac{2(\epsilon - 1)}{2\epsilon + 1} \right].$$

This formula does not include the charge–dipole interaction between reactants A and B. The correlation between measured rate constants in different solvents and their dielectric parameters in general is of a similar quality as illustrated for neutral reactants. This is not, however, due to the approximate nature of the Born model itself which, in spite of its simplicity, leads to remarkably accurate values of ion solvation energies, if the ionic radii can be reliably estimated [15].

Onsager’s reaction field model in its original form offers a description of major aspects of equilibrium solvation effects on reaction rates in solution that includes the basic physical ideas, but the inherent simplifications seriously limit its practical use for quantitative predictions. It since has been extended along several lines, some of which are briefly summarized in the next section.

(C) IMPROVED DIELECTRIC CONTINUUM MODELS

Onsager’s original reaction field method imposes some serious limitations: the description of the solute as a point dipole located at the centre of a cavity, the spherical form of the cavity and the assumption that cavity size and solute dipole moment are independent of the solvent dielectric constant.

Kirkwood generalized the Onsager reaction field method to arbitrary charge distributions and, for a spherical cavity, obtained the Gibbs free energy of solvation in terms of a multipole expansion of the electrostatic field generated by the charge distribution [12, 13]

$$\Delta G_{\text{sol,K}} = \frac{N_A e^2}{8\pi \epsilon_0} \sum_{n=0}^{\infty} \frac{(n+1)(1-\epsilon)}{(n+1)\epsilon+n} \frac{1}{r^{2n+1}} \sum_{k=1}^N \sum_{l=1}^N z_k z_l \mathbf{r}_k^n \mathbf{r}_l^n P_n(\cos \vartheta_{kl}) \quad (\text{A3.6.9})$$

where N is the number of point charges in the cavity, vectors \mathbf{r}_i denote their position, ϑ_{ij} the angle between respective vectors, and P_n are the Legendre polynomials. This expression reduces to [equation \(A3.6.8\)](#) and [equation \(A3.6.7\)](#) for $n = 0$ and $n = 1$, respectively. It turns out that usually it is sufficient to consider terms up to $n \approx 4$ to achieve convergence of the expansion. The absolute value of the solvation energy calculated from [equation \(A3.6.9\)](#), however, critically depends on the size and the shape of the cavity. Even when the charge distribution of reactants and transition state can be calculated to sufficient accuracy by advanced quantum chemical methods, this approach only will give

useful quantitative information about the solvent dependence of reaction rates, if the cavity does not change appreciably along the reaction path from reactants to transition state and if it is largely solvent independent.

As this condition usually is not met, considerable effort has gone into developing methods to calculate solvation energies for cavities of arbitrary shape that as closely as possible mimic the topology of the interface between solute molecule and solvent continuum. Among these are various implementations of boundary element methods [16], in which a cavity surface of arbitrary shape is divided into surface elements carrying a specified surface charge. In one of the more simple variants, a virtual charge scheme as proposed by Miertuš [17], the charge distribution of the solute $\rho^0(\mathbf{r}_i)$ reflects itself in corresponding polarization surface charge densities at the cavity interface $\sigma(\mathbf{s}_i)$ that are assigned to each of m surface elements \mathbf{s}_i and assumed to be constant across the respective surface areas ΔS_i . The electric potential generated by these virtual charges is

$$V_\sigma = \sum_{i=1}^m \Delta S_i \frac{\sigma(\mathbf{s}_i)}{4\pi \epsilon_0 r_i} \quad (\text{A3.6.10})$$

The surface charge density on each surface element is determined by the boundary condition

$$\sigma(\mathbf{s}_i) = \frac{\epsilon_0(\epsilon - 1)}{\epsilon} \left(\frac{dV}{dn_i} \right)_{\mathbf{s}_i} \quad (\text{A3.6.11})$$

where ϵ denotes the static dielectric constant of the solvent, and the derivative of the total electrical potential V at the interface is taken with respect to the normal vector \mathbf{n}_i of each surface element \mathbf{s}_i . V is the sum of contributions from the solute charges ρ^0 and the induced polarization surface charges $\sigma(\mathbf{s}_i)$. Using equation (A3.6.10) and equation (A3.6.11), the virtual surface charge densities $\sigma(\mathbf{s}_i)$ can be calculated iteratively, and the Gibbs free energy of solvation is then half the electrostatic interaction energy of the solute charge distribution in the electric potential generated by the induced polarization surface charges

$$\Delta G_{\text{sol}} = \frac{1}{2} \int V_{\sigma}(\mathbf{r}) \rho^0(\mathbf{r}) d\mathbf{r}. \quad (\text{A3.6.12})$$

Of course, one has to fix the actual shape and size of the cavity, before one can apply equation (A3.6.12). Since taking simply ionic or van der Waals radii is too crude an approximation, one often uses basis-set-dependent *ab initio* atomic radii and constructs the cavity from a set of intersecting spheres centred on the atoms [18, 19]. An alternative approach, which is comparatively easy to implement, consists of using an electrical equipotential surface to define the solute–solvent interface shape [20].

The most serious limitation remaining after modifying the reaction field method as mentioned above is the neglect of solute polarizability. The reaction field that acts back on the solute will affect its charge distribution as well as the cavity shape as the equipotential surface changes. To solve this problem while still using the polarizable continuum model (PCM) for the solvent, one has to calculate the surface charges on the solute by quantum chemical methods and represent their interaction with the solvent continuum as in classical electrostatics. The Hamiltonian of the system thus is written as the sum of the Hamilton operator for the isolated solute molecule and its interaction with the macroscopic

electrostatic reaction field. The coupled equations of the solute subject to the reaction field induced in the solvent are then solved self-consistently to obtain the electron density of the solute in the presence of the polarizable dielectric—the basis of self-consistent reaction field (SCRF) models [21]. Whether this is done in the framework of, for example, Hartree–Fock theory or density functional theory, is a question of optimizing quantum chemical techniques outside the topics addressed here.

If reliable quantum mechanical calculations of reactant and transition state structures in vacuum are feasible, treating electrostatic solvent effects on the basis of SRCF-PCM using cavity shapes derived from methods mentioned above is now sufficiently accurate to predict variations of Gibbs free energies of activation $\delta\Delta G^{\ddagger}$ with solvent polarity reliably, at least in the absence of specific solute–solvent interactions. For instance, considering again a Menshutkin reaction, in this case of pyridine with methylbromide, $\text{Pyr} + \text{MeBr} \rightarrow \text{MePyr}^+ + \text{Br}^-$, in cyclohexane and di-*n*-butyl ether, the difference between calculated and experimental values of ΔG^{\ddagger} is only about 2% and 4%, respectively [22, 23].

As with SCRF-PCM only macroscopic electrostatic contributions to the Gibbs free energy of solvation are taken into account, short-range effects which are limited predominantly to the first solvation shell have to be considered by adding additional terms. These correct for the neglect of effects caused by solute–solvent electron correlation including dispersion forces, hydrophobic interactions, dielectric saturation in the case of

multiply charged ions and solvent structural influences on cavitation. In many cases, however, the electrostatic contribution dominates and dielectric continuum models provide a satisfactory description.

A3.6.2.4 EQUILIBRIUM SOLVENT EFFECTS—MICROSCOPIC VIEW

Specific solute–solvent interactions involving the first solvation shell only can be treated in detail by discrete solvent models. The various approaches like point charge models, supermolecular calculations, quantum theories of reactions in solution, and their implementations in Monte Carlo methods and molecular dynamics simulations like the Car–Parrinello method are discussed elsewhere in this encyclopedia. Here only some points will be briefly mentioned that seem of relevance for later sections.

(A) POINT CHARGE DISTRIBUTION MODEL [11]

Considering, for simplicity, only electrostatic interactions, one may write the solute–solvent interaction term of the Hamiltonian for a solute molecule surrounded by S solvent molecules as

$$\hat{H}_{\text{elstat}} = \sum_{s=1}^S \left[\sum_{\lambda=1}^N \sum_{\alpha=1}^{N_s} \frac{1}{r_{\lambda\alpha s}} - \sum_{\lambda=1}^N \sum_{a=1}^{M_s} \frac{Z_a}{r_{\lambda\alpha s}} + \sum_{l=1}^M \sum_{a=1}^{M_s} \frac{Z_l Z_a}{r_{l\alpha s}} - \sum_{l=1}^M \sum_{\alpha=1}^{N_s} \frac{Z_l}{r_{l\alpha s}} \right] \quad (\text{A3.6.13})$$

where the solute contains N electrons and M nuclei with charges Z_l and the solvent molecules N_s electrons and M_s nuclei with charge Z_a . In the point charge method equation (A3.6.13) reduces to

$$\hat{H}_{\text{pc}} = \sum_{p=1}^P \sum_{\lambda=1}^N \frac{q_p}{r_{\lambda p}} - \sum_{p=1}^P \sum_{l=1}^M \frac{q_p Z_l}{r_{lp}}. \quad (\text{A3.6.14})$$

-11-

Here the position r_{ip} of the point charges located on the solvent molecules q_p is determined by the structure of the solvent shell and the electron density distribution within the solvent molecule. In this type of model, the latter is assumed to be fixed, i.e. the solvent molecules are considered non-polarizable while solving the Schrödinger equation for the coupled system.

Instead of using point charges one may also approximate the interaction Hamiltonian in terms of solute electrons and nuclei interacting with solvent point dipoles μ_d

$$\hat{H}_{\text{pd}} = \sum_{d=1}^D \sum_{\lambda=1}^N \frac{\mu_d r_{d\lambda}}{r_{d\lambda}^3} - \sum_{d=1}^D \sum_{l=1}^M \frac{\mu_d r_{dl} Z_l}{r_{dl}^3}. \quad (\text{A3.6.15})$$

In either case, the structure of the solvation shell has to be calculated by other methods supplied or introduced *ad hoc* by some further model assumptions, while charge distributions of the solute and within solvent molecules are obtained from quantum chemistry.

(B) SOLVATION SHELL STRUCTURE

The quality of the results that can be obtained with point charge or dipole models depends critically on the input solvation shell structure. In view of the computer power available today, taking the most rigorous route

is feasible in many cases, i.e. using statistical methods to calculate distribution functions in solution. In this way the average structure of solvation shells is accessible, that is, to be used in equilibrium solvation calculations required to obtain, for example, TST rate constants.

Assuming that additive pair potentials are sufficient to describe the inter-particle interactions in solution, the local equilibrium solvent shell structure can be described using the pair correlation function $g^{(2)}(\mathbf{r}_1, \mathbf{r}_2)$. If the potential only depends on inter-particle distance, $g^{(2)}(\mathbf{r}_1, \mathbf{r}_2)$ reduces to the radial distribution function $g(r) \equiv g^{(2)}(|\mathbf{r}_1 - \mathbf{r}_2|)$ such that $\rho \cdot 4\pi r^2 dr g(r)$ gives the number of particles in a spherical shell of thickness dr at distance r from a reference particle (ρ denotes average particle density). The local particle density is then simply $\rho \cdot g(r)$. The radial distribution function can be obtained experimentally in neutron scattering experiments by measuring the angular dependence of the scattering amplitude, or by numerical simulation using Monte Carlo methods.

(C) POTENTIAL OF MEAN FORCE

At low solvent density, where isolated binary collisions prevail, the radial distribution function $g(r)$ is simply related to the pair potential $u(r)$ via $g_0(r) = \exp[-u(r)/kT]$. Correspondingly, at higher density one defines a function $w(r) \equiv -kT \ln[g(r)]$. It can be shown that the gradient of this function is equivalent to the mean force between two particles obtained by holding them at fixed distance r and averaging over the remaining $N - 2$ particles of the system. Hence $w(r)$ is called the potential of mean force. Choosing the low-density system as a reference state one has the relation

$$\lim_{\rho \rightarrow 0} g(r) = g_0(r) \Rightarrow \lim_{\rho \rightarrow 0} w(r) = u(r)$$

-12-

and $\Delta w(r) \equiv w(r) - u(r)$ describes the average many-body contribution such as, for example, effects due to solvation shell structure. In the language of the thermodynamic formulation of TST, the ratio of rate constants in solution and dilute-gas phase consequently may be written as

$$kT \ln \frac{k_{\text{solution}}}{k_{\text{gas}}} = -\delta \Delta w^\ddagger \equiv -[\Delta w(r^\ddagger) - \Delta w(r_{\text{react}})]. \quad (\text{A3.6.16})$$

A3.6.2.5 PRESSURE EFFECTS

The inherent difficulties in interpreting the effects observed in solvent series studies of chemical reaction rates, which offer little control over the multitude of parameters that may influence the reaction, suggest rather using a single liquid solvent and varying the pressure instead, thereby changing solvent density and polarity in a well known way. One also may have to consider, of course, variations in the local solvent shell structure with increasing pressure.

(A) ACTIVATION VOLUME

In the thermodynamic formulation of TST the pressure dependence of the reaction rate coefficient defines a volume of activation [24, 25 and 26]

$$\left(\frac{\partial \ln k}{\partial p} \right)_T = -\frac{1}{RT} \left(\frac{\partial \Delta G^\ddagger}{\partial p} \right)_T \equiv -\frac{\Delta V^\ddagger}{RT} \quad (\text{A3.6.17})$$

with $\Delta V^\ddagger = \bar{V}^\ddagger - \sum_i^{\text{reactants}} \bar{V}_i$, the difference of the molar volume of transition state and the sum over molar volumes of reactants. Experimental evidence shows that $|\Delta V^\ddagger|$ is of the order of $10^0 - 10^1 \text{ cm}^3 \text{ mol}^{-1}$ and usually pressure dependent [27]. It is common practice to interpret it using geometric arguments considering reactant and transition state structures and by differences in solvation effects between reactant and transition state. If one uses a molar concentration scale (standard state 1 mol dm^{-3}), an additional term $+\kappa_{\text{solv}} \Delta v^\ddagger$ appears in the rhs of equation (A3.6.16), the product of isothermal solvent compressibility and change in sum over stoichiometric coefficients between reactants and transition state.

There is one important *caveat* to consider before one starts to interpret activation volumes in terms of changes of structure and solvation during the reaction: the pressure dependence of the rate coefficient may also be caused by transport or dynamic effects, as solvent viscosity, diffusion coefficients and relaxation times may also change with pressure [2]. Examples will be given in subsequent sections.

(B) ACTIVATION VOLUME IN A DIELECTRIC CONTINUUM

If, in analogy to equation (A3.6.5), one denotes the change of activation volume with respect to some reference solvent as $\delta_M \Delta V^\ddagger$ and considers only electrostatic interactions of reactant and transition state with a dielectric continuum solvent, one can calculate it directly from

-13-

$$\delta_M \Delta V^\ddagger = \left(\frac{\partial(\delta_M \Delta G^\ddagger)}{\partial p} \right)_T \quad (\text{A3.6.18})$$

by using any of the models mentioned above. If the amount of charge redistribution is significant and the solvent is polar, the dielectric contribution to ΔV^\ddagger by far dominates any so-called intrinsic effects connected with structural changes between reactant and transition state. For the Menshutkin reaction, for example, equation (A3.6.17) gives

$$\delta_M \Delta V^\ddagger = \frac{-N_A}{4\pi \epsilon_0} \left[\frac{3}{(2\epsilon(p) + 1)^2} \left(\frac{\partial \epsilon}{\partial p} \right)_T \right] \left(\frac{(\mu_{ab}^\ddagger)^2}{(r_{ab}^\ddagger)^3} - \frac{\mu_a^2}{r_a^3} - \frac{\mu_b^2}{r_b^3} \right)$$

which includes a positive term resulting from the pressure dependence of the dielectric constant (in square brackets) and represents the experimentally observed pressure dependence of the activation volume quite satisfactorily [25]. For the Menshutkin reaction, only the large dipole moment of the transition state needs to be considered, resulting in a negative activation volume, a typical example of electrostriction. If one assumes that the neglect of solute polarizability is justified and, in addition, the cavity radius is constant, one may use this kind of expression to estimate transition state dipole moments. Improved continuum models as outlined in the preceding sections may, of course, also be applied to analyse activation volumes.

(C) ACTIVATION VOLUME AND LOCAL SOLVENT STRUCTURE

In a microscopic equilibrium description the pressure-dependent local solvent shell structure enters through variations of the potential of mean force, $(\partial \delta \Delta w^\ddagger / \partial p)_T$, such that the volume of activation contains a contribution related to the pressure dependence of radial distribution functions for reactants and transition state, i.e.

$$\Delta V_{\text{local}}^{\ddagger} = -kT \left\{ \frac{\partial}{\partial p} \left[\ln \left(\frac{g(r^{\ddagger})g_0(r_{\text{react.}})}{g_0(r^{\ddagger})g(r_{\text{react.}})} \right) \right] \right\}_{\text{T}}$$

This contribution of local solvent structure to ΔV^{\ddagger} may be quite significant and, even in nonpolar solvents, in many cases outweigh the intrinsic part. It essentially describes a caging phenomenon, as with increasing pressure the local solvent density or packing fraction of solvent molecules around reactants and transition state increases, thereby enhancing the stability of the solvent cage. This constitutes an equilibrium view of caging in contrast to descriptions of the cage effect in, for example, photodissociation where solvent friction is assumed to play a central role.

How large the magnitude of this packing effect can be was demonstrated in simple calculations for the atom transfer reaction $\text{CH}_3 + \text{CH}_4 \rightarrow \text{CH}_4 + \text{CH}_3$ using a binary solution of hard spheres at infinite dilution as the model system [28]. Allowing spheres to partially overlap in the transition state, i.e. assuming a common cavity, reaction rates were calculated by variational TST for different solute-to-solvent hard-sphere ratios $r_{\text{G}} = \sigma_{\text{M}}/\sigma_{\text{S}}$ and solvent densities $\rho_{\text{S}}\sigma_{\text{S}}^3$. Increasing the latter from 0.70 to 0.95 led to an enhancement of the relative rate constant $k_{\text{solution}}/k_{\text{gas}}$ by factors of 8.5, 15.5 and 53 for r_{G} equal to 0.93, 1.07 and 1.41, respectively, thus clearly showing the effect of local packing density. With respect to the calculated gas phase value the rate constants at the highest density were 95, 280 and

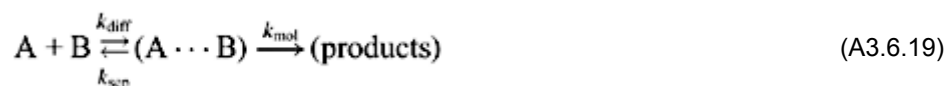
-14-

2670 times larger, respectively. This behaviour is typical for ‘tight’ transition states, whereas for loose transition states as they appear, for example, in isomerization reactions, this caging effect is orders of magnitude smaller.

A3.6.3 TRANSPORT EFFECTS

If reactant motion along the reaction path in the condensed phase involves significant displacement with respect to the surrounding solvent medium and there is non-negligible solute–solvent coupling, frictional forces arise that oppose the reactive motion. The overall rate of intrinsically fast reactions for which, for example, the TST rate constant is sufficiently large, therefore, may be influenced by the viscous drag that the molecules experience on their way from reactants to products. As mentioned in the introduction, dynamic effects due to other partially non-relaxed degrees of freedom will not be considered in this section.

For a bimolecular reaction, this situation is easily illustrated by simply writing the reaction as a sequence of two steps



where brackets denote common solvent cage (encounter complex), k_{diff} is the rate constant of diffusive approach of reactants, sometimes called the ‘encounter rate’, k_{sep} is that of diffusive separation of the unreacted encounter pair and k_{mol} that of the reactive step in the encounter complex. If $k_{\text{mol}} \gg k_{\text{sep}}$, the overall reaction rate constant k essentially equals k_{diff} and the reaction is said to be diffusion controlled. One important implicit assumption of this phenomenological description is that diffusive approach and separation

are statistically independent processes, i.e. the lifetime of the encounter pair is sufficiently long to erase any memory about its formation history. Examples of processes that often become diffusion controlled in solution are atom and radical recombination, electron and proton transfer, fluorescence quenching and electronic energy transfer.

In a similar phenomenological approach to unimolecular reactions involving large-amplitude motion, the effect of friction on the rate constant can be described by a simple transition formula between the high-pressure limit k_∞ of the rate constant at negligible solvent viscosity and the so-called Smoluchowski limit of the rate constant, k_{SM} , approached in the high-damping regime at large solvent viscosity [2]:

$$\frac{1}{k} = \frac{1}{k_\infty} + \frac{1}{k_{SM}}. \quad (\text{A3.6.20})$$

As k_{SM} is inversely proportional to solvent viscosity, in sufficiently viscous solvents the rate constant k becomes equal to k_{SM} . This concerns, for example, reactions such as isomerizations involving significant rotation around single or double bonds, or dissociations requiring separation of fragments, although it may be difficult to experimentally distinguish between effects due to local solvent structure and solvent friction.

Systematic experimental investigations of these transport effects on reaction rates can either be done by varying solvents in a homologous series to change viscosity without affecting other physicochemical or chemical properties

-15-

(or as little as possible) or, much more elegantly and experimentally demanding, by varying pressure and temperature in a single solvent, maintaining control over viscosity, polarity and density at the same time. As detailed physical insight is gained by the latter approach, the few examples shown all will be from pressure-dependent experimental studies. Computer experiments involving stochastic trajectory simulations or classical molecular dynamics simulations have also been extremely useful for understanding details of transport effects on chemical reaction rates, though they have mostly addressed dynamic effects and been less successful in actually providing a quantitative connection with experimentally determined solvent or pressure dependences of rate constants or quantum yields of reactions.

A3.6.3.1 DIFFUSION AND BIMOLECULAR REACTIONS

(A) DIFFUSION-CONTROLLED RATE CONSTANT

Smoluchowski theory [29, 30] and its modifications form the basis of most approaches used to interpret bimolecular rate constants obtained from chemical kinetics experiments in terms of diffusion effects [31]. The Smoluchowski model is based on Brownian motion theory underlying the phenomenological diffusion equation in the absence of external forces. In the standard picture, one considers a dilute fluid solution of reactants A and B with $[A] \ll [B]$ and asks for the time evolution of $[B]$ in the vicinity of A, i.e. of the density distribution $\rho(r,t) \equiv [B](r,t)/[B]_{t=0} \simeq [B](r(t))/[B]_{t=0}$ ($[B]$ is assumed not to change appreciably during the reaction). The initial distribution and the outer and inner boundary conditions are chosen, respectively, as

$$\rho(r, 0) = \begin{cases} 0 & \text{for } r \leq R \\ 1 & \text{for } r > R \end{cases}$$

$$\rho(r \rightarrow \infty, t) = 1 \quad \text{for } t \geq 0 \quad (\text{A3.6.21})$$

$$k_{\text{mol}}\rho(R) = 4\pi R^2 D_{\text{AB}} \left. \frac{\partial \rho}{\partial r} \right|_R$$

where R is the encounter radius and D_{AB} the mutual diffusion coefficient of reactants. The reflecting boundary condition [32] at the encounter distance R ensures that, once a stationary concentration of encounter pairs is established, the intrinsic reaction rate in the encounter pair, $k_{\text{mol}}\rho(R)$, equals the rate of diffusive formation of encounter pairs. In this formulation k_{mol} is a second-order rate constant. Solving the diffusion equation

$$\frac{\partial \rho}{\partial t} = D_{\text{AB}} \left[\frac{\partial^2 \rho}{\partial r^2} + \frac{2}{r} \frac{\partial \rho}{\partial r} \right] \quad (\text{A3.6.22})$$

subject to conditions (A3.6.21) and realizing that the observed reaction rate coefficient $k(t)$ equals $k_{\text{mol}}\rho(R, t)$, one obtains

$$k(t) = \frac{k_{\text{mol}}}{1+x} \{1 + x \exp[y^2(1+x)^2 t] \text{erfc}[y(1+x)\sqrt{t}]\} \quad (\text{A3.6.23})$$

-16-

using the abbreviations $x \equiv k_{\text{mol}}/4\pi R D_{\text{AB}}$ and $y \equiv \sqrt{D_{\text{AB}}}R$. The time-dependent terms reflect the transition from the initial to the stationary distribution. After this transient term has decayed to zero, the reaction rate attains its stationary value

$$k = \frac{k_{\text{mol}}}{1+x} = \frac{4\pi R D_{\text{AB}} k_{\text{mol}}}{4\pi R D_{\text{AB}} + k_{\text{mol}}} = \frac{k_{\text{diff}} k_{\text{mol}}}{k_{\text{diff}} + k_{\text{mol}}} \quad (\text{A3.6.24})$$

such that for $k_{\text{mol}} \gg k_{\text{diff}}$ one reaches the diffusion limit $k \simeq k_{\text{diff}}$. Comparing equation (A3.6.24) with the simple kinetic scheme (A3.6.19), one realizes that at this level of Smoluchowski theory one has $k_{\text{sep}} = k_{\text{diff}}\rho(R)$, i.e. there is no effect due to caging of the encounter complex in the common solvation shell. There exist numerous modifications and extensions of this basic theory that not only involve different initial and boundary conditions, but also the inclusion of microscopic structural aspects [31]. Among these are hydrodynamic repulsion at short distances that may be modelled, for example, by a distance-dependent diffusion coefficient

$$D_{\text{AB}}(r) \simeq D_{\text{AB}} \left[1 - \frac{1}{2} \exp\left(1 - \frac{r}{R}\right) \right]$$

or the potential of mean force *via* the radial distribution function $g(r)$, which leads to a significant reduction of the steady-state rate constant by about one-third with respect to the Smoluchowski value [33, 34]:

$$k = k_{\text{mol}} g(R) \left[1 + k_{\text{mol}} g(R) \int \frac{dr}{4\pi r^2 D_{AB}(r) g(r)} \right]^{-1}.$$

Diffusion-controlled reactions between ions in solution are strongly influenced by the Coulomb interaction accelerating or retarding ion diffusion. In this case, the diffusion equation for ρ concerning motion of one reactant about the other stationary reactant, the Debye–Smoluchowski equation,

$$\frac{\partial \rho}{\partial t} = D_{AB} \nabla \cdot \left[\nabla \rho + \frac{\rho}{kT} \nabla V(r) \right] \quad (\text{A3.6.25})$$

includes the gradient of the potential energy $V(r)$ of the ions in the Coulomb field. Using boundary conditions equivalent to equation (A3.6.21) and an initial condition corresponding to a Boltzmann distribution of interionic distances

$$\rho(r, 0) = e^{-V(r)/kT} = e^{-R_C/r} \quad R_C = \frac{z_A z_B e^2}{4\pi \epsilon \epsilon_0 kT}$$

and solving equation (A3.6.25), one obtains the steady-state solution

-17-

$$k = 4\pi R_C D_{AB} \left[\left(1 + \frac{4\pi R_C D_{AB}}{k_{\text{mol}}} \right) e^{R_C/R} - 1 \right]^{-1}.$$

Many additional refinements have been made, primarily to take into account more aspects of the microscopic solvent structure, within the framework of diffusion models of bimolecular chemical reactions that encompass also many-body and dynamic effects, such as, for example, treatments based on kinetic theory [35]. One should keep in mind, however, that in many cases the practical value of these advanced theoretical models for a quantitative analysis or prediction of reaction rate data in solution may be limited.

(B) TRANSITION FROM GASEOUS TO LIQUID SOLVENT—ONSET OF DIFFUSION CONTROL

Instead of concentrating on the diffusion limit of reaction rates in liquid solution, it can be instructive to consider the dependence of bimolecular rate coefficients of elementary chemical reactions on pressure over a wide solvent density range covering gas and liquid phase alike. Particularly amenable to such studies are atom recombination reactions whose rate coefficients can be easily investigated over a wide range of physical conditions from the dilute-gas phase to compressed liquid solution [3, 4].

As discussed above, one may try to represent the density dependence of atom recombination rate coefficients k in the spirit of equation (A3.6.24) as

$$\frac{1}{k} \approx \frac{1}{k_{\text{rec}}^g} + \frac{1}{k_{\text{diff}}} \quad (\text{A3.6.26})$$

where k_{rec}^g denotes the low-pressure second-order rate coefficient proportional to bath gas density, and k_{diff} is

the second-order rate coefficient of diffusion-controlled atom recombination as discussed in the previous section. In order to apply equation (A3.6.26), a number of items require answers specific to the reaction under study: (i) the density dependence of the diffusion coefficient D_{AA} , (ii) the magnitude of the encounter radius R , (iii) the possible participation of excited electronic states and (iv) the density dependence of $k_{\text{rec}}^{\text{g}}$. After these have been dealt with adequately, it can be shown that for many solvent bath gases, the phenomenon of the turnover from a molecular reaction into a diffusion-controlled recombination follows equation (A3.6.26) without any apparent discontinuity in the rate coefficient k at the gas–liquid phase transition, as illustrated for iodine atom recombination in argon [36, 37]. For this particular case, D_{AA} is based on and extrapolated from experimental data, R is taken to be one-half the sum of the Lennard-Jones radii of iodine atom and solvent molecule, and the density-dependent contribution of excited electronic states is implicitly considered by making the transition from the measured $k_{\text{rec}}^{\text{g}}$ in dilute ethane gas to k_{diff} in dense liquid ethane.

A more subtle point concerns scaling of $k_{\text{rec}}^{\text{g}}$ with density. Among the various possibilities that exist, either employing local densities obtained from numerically calculated radial distribution functions [38]

$$k_{\text{rec}}^{\text{g}}(\rho) = k_{\text{rec}}^{\text{g}}(\rho_0) \frac{\rho g(r; \rho)}{\rho_0 g(r; \rho_0)}$$

-18-

or taking into account that in the gas phase the reaction is controlled to a large extent by the energy transfer mechanism, such that $k_{\text{rec}}^{\text{g}} \approx \beta_c Z_{\text{LJ}}$ where β_c is a collision efficiency and Z_{LJ} the Lennard-Jones collision frequency, are probably the most practical. As $Z_{\text{LJ}} \sim 1/D_{\text{AM}}$ throughout the whole density range, $k_{\text{rec}}^{\text{g}}(\rho)$ in the latter case may be estimated by scaling with the diffusion coefficient [37]

$$k_{\text{rec}}^{\text{g}}(\rho) = k_{\text{rec}}^{\text{g}}(\rho_0) \frac{D_{\text{AM}}(\rho_0)}{D_{\text{AM}}(\rho)}$$

Although the transition to diffusion control is satisfactorily described in such an approach, even for these apparently simple elementary reactions the situation in reality appears to be more complex due to the participation of weakly bonding or repulsive electronic states which may become increasingly coupled as the bath gas density increases. These processes manifest themselves in iodine atom and bromine atom recombination in some bath gases at high densities where marked deviations from ‘normal’ behaviour are observed [3, 4]. In particular, it is found that the transition from $k_{\text{rec}}^{\text{g}}$ to k_{diff} is significantly broader than predicted by equation (A3.6.26), the reaction order of iodine recombination in propane is higher than 3, and S-shaped curves are observed with He as a bath gas [36] (see figure A3.6.4). This is in contrast to the recombination of the methyl radicals in Ar which can be satisfactorily described by a theory of particle encounter kinetics using appropriate interaction potentials and a modified friction for relative motion [39]. The only phenomena that cannot be reproduced by such treatments were observed at moderate gas pressures between 1 and 100 bar. This indicates that the kinetics of the reaction in this density regime may be influenced to a large extent by reactant–solute clustering or even chemical association of atoms or radicals with solvent molecules.

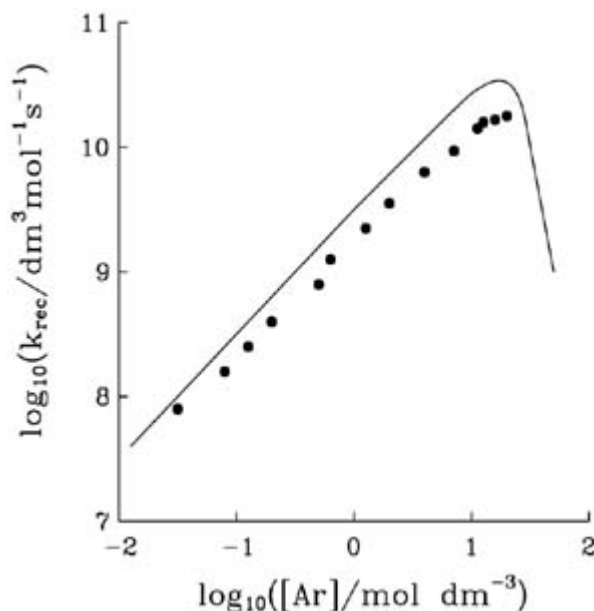


Figure A3.6.4. Pressure dependence of atom recombination rate constant of iodine in argon: experiment (points) [36] and theory (full line) [120].

-19-

This problem is related to the question of appropriate electronic degeneracy factors in chemical kinetics. Whereas the general belief is that, at very low gas pressures, only the electronic ground state participates in atom recombination and that, in the liquid phase, at least most of the accessible states are coupled somewhere ‘far out’ on the reaction coordinate, the transition between these two limits as a function of solvent density is by no means understood. Direct evidence for the participation of different electronic states in iodine geminate recombination in the liquid phase comes from picosecond time-resolved transient absorption experiments in solution [40, 41] that demonstrate the participation of the low-lying, weakly bound iodine A and A' states, which is also taken into account in recent mixed classical–quantum molecular dynamics simulations [42, 43].

A3.6.3.2 UNIMOLECULAR REACTIONS AND FRICTION

So far the influence of the dense solvent environment on the barrier crossing process in a chemical reaction has been ignored. It is evident from the typical pressure dependence of the rate coefficient k of a unimolecular reaction from the low-pressure gas phase to the compressed-liquid phase that the prerequisites of TST are only met, if at all, in a narrow density regime corresponding to the plateau region of the curve. At low pressures, where the rate is controlled by thermal activation in binary collisions with the solvent molecules, k is proportional to pressure. This regime is followed by a plateau region where k is pressure independent and controlled by intramolecular motion along the reaction coordinate. Here k attains the so-called high-pressure limit k_∞ which can be calculated by statistical theories if the PES for the reaction is known. If the reaction entails large-amplitude structural changes, further increasing the pressure can lead to a decrease of k as a result of frictional forces retarding the barrier crossing process. In the simplest approach, k eventually approaches an inverse dependence on solvent friction, the so-called Smoluchowski limit k_{SM} of the reaction rate.

The transition from k_0 to k_∞ on the low-pressure side can be constructed using multidimensional unimolecular rate theory [1, 44], if one knows the barrier height for the reaction and the vibrational frequencies of the reactant and transition state. The transition from k_∞ to k_{SM} can be described in terms of Kramers’ theory [45]

which, in addition, requires knowledge of the pressure dependence of the solvent friction acting on the molecule during the particular barrier crossing process. The result can be compared with rate coefficients measured over a wide pressure range in selected solvents to test the theoretical models that are used to describe this so-called Kramers' turnover of the rate coefficient.

(A) KRAMERS' THEORY

Kramers' solution of the barrier crossing problem [45] is discussed at length in [chapter A3.8](#) dealing with condensed-phase reaction dynamics. As the starting point to derive its simplest version one may use the Langevin equation, a stochastic differential equation for the time evolution of a slow variable, the reaction coordinate \mathbf{r} , subject to a rapidly statistically fluctuating force \mathbf{F} caused by microscopic solute–solvent interactions under the influence of an external force field generated by the PES V for the reaction

$$M\dot{\mathbf{u}} = -\gamma\mathbf{u} + \mathbf{F}(t) - \nabla_{\mathbf{r}}V \quad (\text{A3.6.27})$$

where dots denote time derivative, M is the mass moving with velocity \mathbf{u} along the reaction path and γ is the constant friction coefficient for motion along that path. The assumption is that there are no memory effects in the solvent bath,

-20-

i.e. one considers a Markov process such that for the ensemble average $\langle \mathbf{F}(t) \cdot \mathbf{F}(t') \rangle \sim \delta(t - t')$. The corresponding two-dimensional Fokker–Planck equation for the probability distribution in phase space can be solved for the potential-barrier problem involving a harmonic well and a parabolic barrier in the limit of low and large friction. Since the low-friction limit, corresponding to the reaction in the gas phase, is correctly described by multidimensional unimolecular rate theory, only the solution in the large-friction limit is of interest in this context. One obtains a correction factor F_{Kr} to the high-pressure limit of the reaction rate constant k_{∞}

$$F_{\text{Kr}} = \left[\left(\left(\frac{\gamma/M}{2\omega_{\text{B}}} \right)^2 - 1 \right)^{1/2} - \frac{\gamma/M}{2\omega_{\text{B}}} \right] \quad (\text{A3.6.28})$$

which contains as an additional parameter the curvature of the parabolic barrier top, the so-called imaginary barrier frequency ω_{B} . F_{Kr} is less than unity and represents the dynamic effect of trajectories recrossing the barrier top, in contrast to the central assumption of canonical and microcanonical statistical theories, like TST or RRKM theory. In the high-damping limit, when $\gamma/M \gg \omega_{\text{B}}$, F_{Kr} reduces to $\omega_{\text{B}}M/\gamma$ which simply represents the Smoluchowski limit where velocities relax much faster than the barrier is crossed. As γ approaches zero, F_{Kr} goes to unity and the rate coefficient becomes equal to the high-pressure limit k_{∞} . In contrast to the situation in the Smoluchowski limit, the velocities do not obey a Maxwell–Boltzmann distribution.

(B) PRESSURE DEPENDENCE OF REACTION RATES

If other fall-off broadening factors arising in unimolecular rate theory can be neglected, the overall dependence of the rate coefficient on pressure or, equivalently, solvent density may be represented by the expression [1, 2]

$$k(\rho) = \frac{k_0 \rho k_\infty}{k_0 \rho + k_\infty} F_{\text{Kr}}(\rho). \quad (\text{A3.6.29})$$

This ensures the correct connection between the one-dimensional Kramers model in the regime of large friction and multidimensional unimolecular rate theory in that of low friction, where Kramers' model is known to be incorrect as it is restricted to the energy diffusion limit. For low damping, equation (A3.6.29) reduces to the Lindemann–Hinshelwood expression, while in the case of very large damping, it attains the Smoluchowski limit

$$k_{\text{SM}} = k_\infty \frac{\omega_{\text{B}}}{\gamma/M}. \quad (\text{A3.6.30})$$

Sometimes it may be convenient to use an even simpler interpolation formula that connects the different rate coefficient limits [4]

$$\frac{1}{k} \approx \frac{1}{k_0 \rho} + \frac{1}{k_\infty} + \frac{1}{k_{\text{SM}}} \Rightarrow k \approx \frac{k_0 \rho k_\infty}{k_\infty + k_0(1 + \gamma/M\omega_{\text{B}})} \quad (\text{A3.6.31})$$

for which numerical simulations have shown that it is accurate to within 10–20%.

-21-

Predicting the solvent or density dependence of rate constants by [equation \(A3.6.29\)](#) or [equation \(A3.6.31\)](#) requires the same ingredients as the calculation of TST rate constants plus an estimate of ω_{B} and a suitable model for the friction coefficient γ and its density dependence. While in the framework of molecular dynamics simulations it may be worthwhile to numerically calculate friction coefficients from the average of the relevant time correlation functions, for practical purposes in the analysis of kinetic data it is much more convenient and instructive to use experimentally determined macroscopic solvent parameters.

As in the case of atom recombination, a convenient ‘pressure scale’ to use across the entire range is the inverse of the binary diffusion coefficient, D_{AM}^{-1} , of reactant A in solvent M, as compared to density ρ in the low-pressure gas and the inverse of solvent viscosity η^{-1} in liquid solution [46]. According to kinetic theory the diffusion coefficient in a dilute Lennard-Jones gas is given by

$$D_{\text{AM}} = \frac{3}{2\sqrt{2}} \frac{kT}{\mu_{\text{AM}}} \frac{\Omega^{(2,2)*}}{\Omega^{(1,1)*}} \frac{1}{Z_{\text{LJ}}\rho} \equiv \frac{A_{\text{D}}}{Z_{\text{LJ}}\rho}$$

with reduced collision integrals $\Omega^{(l,j)*}$ for Lennard-Jones well depths $\epsilon_{\text{AM}} = \sqrt{\epsilon_{\text{A}}\epsilon_{\text{M}}}$ and reduced mass μ_{AM} , such that the low-pressure rate coefficient is

$$k_0 = \frac{A_{\text{D}}}{D_{\text{AM}}} \int_{E_0}^{\infty} f(E) dE \equiv \frac{A_{\text{D}}}{D_{\text{AM}} k_{00}}.$$

In liquid solution, Brownian motion theory provides the relation between diffusion and friction coefficient

$D_{AM} = kT/\gamma$. Substituting correspondingly in equation (A3.6.31), one arrives at an expression representing the pressure dependence of the rate constant in terms of the pressure-dependent diffusion coefficient:

$$k \approx \frac{k_{00} A_D k_{\infty} D_{AM}}{k_{\infty} + k_{00} A_D (D_{AM} + kT/\mu_{AM} \omega_B)}. \quad (\text{A3.6.32})$$

As data of the binary diffusion coefficient $D_{AM}(\rho, T)$ are not available in many cases, one has to resort to taking the solvent self-diffusion coefficient $D_M(\rho, T)$ which requires rescaling in the low-pressure regime according to

$$\frac{D_M}{D_{AM}} = \left[\frac{2\mu_{AM}}{M} \right]^{1/2} \left[\frac{\sigma_M}{\sigma_{AM}} \right]^2 \frac{\Omega_M^{(1,1)*}}{\Omega_{AM}^{(1,1)*}}.$$

In the Smoluchowski limit, one usually assumes that the Stokes–Einstein relation $(D\eta/kT)\sigma = C$ holds, which forms the basis of taking the solvent viscosity as a measure for the zero-frequency friction coefficient appearing in Kramers' expressions. Here C is a constant whose exact value depends on the type of boundary conditions used in deriving Stokes' law. It follows that the diffusion coefficient ratio is given by $D_M/D_{AM} = C_M \sigma_{AM}/C_{AM} \sigma_M$, which may be considered as approximately pressure independent.

(C) EXTENSIONS OF KRAMERS' BASIC MODEL

As extensions of the Kramers theory [47] are essentially a topic of condensed-phase reaction dynamics, only a few remarks are in place here. These concern the barrier shape and the dimensionality in the high-damping regime. The curvature at the parabolic barrier top obviously determines the magnitude of the friction coefficient at which the rate constant starts to decrease below the upper limit defined by the high-pressure limit: for relatively sharp barriers this 'turnover' will occur at comparatively high solvent density corresponding almost to liquid phase densities, whereas reactions involving flat barriers will show this phenomenon in the moderately dense gas, maybe even in the unimolecular fall-off regime before they reach k_{∞} .

Non-parabolic barrier tops cause the prefactor to become temperature dependent [48]. In the Smoluchowski limit, $k_{SM} \propto T^n$, $|n| \sim 1$, with $n > 0$ and $n < 0$ for curvatures smaller and larger than parabolic, respectively. For a cusp-shaped barrier top, i.e. in the limit $\omega_B \rightarrow \infty$ as might be applicable to electron transfer reactions, one obtains [45]

$$k_{SM} = k_{\infty} \frac{\omega_A M \sqrt{\pi}}{\gamma} \sqrt{\frac{E_0}{kT}}$$

where ω_A is the harmonic frequency of the potential well in this one-dimensional model. In the other limit, for an almost completely flat barrier top, the transition curve is extremely broad and the maximum of k is far below k_{∞} [49]. A qualitatively different situation arises when reactant and product well are no longer separated by a barrier, but one considers escape out of a Lennard-Jones potential well. In this case, dynamics inside the well and outside on top of the 'barrier' plateau are no longer separable and, in a strict sense, the Smoluchowski limit is not reached any more. The stationary rate coefficient in the high-damping limit turns

out to be [50]

$$\left(\frac{k}{k_\infty}\right)^{\text{LJ}} = \frac{1}{\gamma\sigma_{\text{LJ}}} \sqrt{\frac{\pi M k T}{2}} \frac{1}{L/\sigma_{\text{LJ}} - 2^{1/6}} \left[1 - \frac{2\varepsilon_{\text{LJ}} M}{(L\gamma)^2} + \dots \right].$$

The original Kramers model is restricted to one-dimensional barriers and cannot describe effects due to the multidimensional barrier topology that may become important in cases where the system does not follow the minimum energy path on the PES but takes a detour across a higher effective potential energy barrier which is compensated by a gain in entropy. Considering a two-dimensional circular reaction path, the Smoluchowski limit of the rate coefficient obtained by solving the two-dimensional Fokker–Planck equation in coordinate space was shown to be [51]

$$k_{\text{SM}}^{2\text{D}} = k_{\text{SM}} \left[1 + \frac{kT}{M(\omega_\perp r_c)^2} \right]$$

where ω_\perp is the harmonic frequency of the transverse potential well and r_c the radius of curvature of the reaction path. This result is in good agreement with corresponding Langevin simulations [52]. A related concept is based on the picture that with increasing excitation of modes transverse to the reaction path the effective barrier curvature may increase according to $\omega_{\text{B}}^{\text{eff}}(E_\perp) \propto (E_\perp/b)^a$, where a and b are dimensionless parameters [53]. Approximating the

topology of the saddle point region by a combination of a parabolic barrier top and a transverse parabolic potential, one arrives at a rate constant in the Smoluchowski limit given by

$$k_{\text{SM}}^{2\text{D}} = \frac{k_\infty}{\gamma/M} \omega_{\text{B}}(T) \quad \text{with } \omega_{\text{B}}(T) \propto \left(\frac{T}{a}\right)^b \Gamma\left(b + \frac{1}{2}\right).$$

Multidimensionality may also manifest itself in the rate coefficient as a consequence of anisotropy of the friction coefficient [54]. Weak friction transverse to the minimum energy reaction path causes a significant reduction of the effective friction and leads to a much weaker dependence of the rate constant on solvent viscosity. These conclusions based on two-dimensional models also have been shown to hold for the general multidimensional case [55, 56, 57, 58, 59, 60 and 61].

To conclude this section it should be pointed out again that the friction coefficient has been considered to be frequency independent as implied in assuming a Markov process, and that zero-frequency friction as represented by solvent viscosity is an adequate parameter to describe the effect of friction on observed reaction rates.

(D) FREQUENCY-DEPENDENT FRICTION

For very fast reactions, as they are accessible to investigation by pico- and femtosecond laser spectroscopy, the separation of time scales into slow motion along the reaction path and fast relaxation of other degrees of freedom in most cases is no longer possible and it is necessary to consider dynamical models, which are not the topic of this section. But often the temperature, solvent or pressure dependence of reaction rate

coefficients determined in chemical kinetics studies exhibit a signature of underlying dynamic effects, which may justify the inclusion of some remarks at this point.

The key quantity in barrier crossing processes in this respect is the barrier curvature ω_B which sets the time window for possible influences of the dynamic solvent response. A sharp barrier entails short barrier passage times during which the memory of the solvent environment may be partially maintained. This non-Markov situation may be expressed by a generalized Langevin equation including a time-dependent friction kernel $\gamma(t)$ [62]

$$M\ddot{u} = - \int_0^t \gamma(t - \tau) \mathbf{u}(\tau) d\tau + \mathbf{F}(t) = \nabla_{\mathbf{r}} V$$

in which case the autocorrelation function of the randomly fluctuating force is no longer a δ -function but obeys $\langle \mathbf{F}(t) \cdot \mathbf{F}(t') \rangle = kT\gamma(t - t')$. This ensures that a Maxwell–Boltzmann distribution is re-established after decay of the solvent response. Adding the assumption of a Gaussian friction kernel, a generalized Fokker–Planck equation with time-dependent friction may be set up, and for a piecewise parabolic potential one obtains an expression for the rate coefficient, the so-called Grote–Hynes formula [63]:

$$k_{GH} = \frac{k_{\infty}}{\omega_B} \lambda_r. \quad (\text{A3.6.33})$$

-24-

λ_r is the reactive frequency or unstable mode which is related to the friction coefficient by the implicit equation

$$\lambda_r = \frac{\omega_B^2}{\lambda_r + (\gamma(\lambda_r)/M)} \quad (\text{A3.6.34})$$

with $\gamma(\lambda_r)$ being the Laplace transform of the time-dependent friction, $\hat{\gamma}(\lambda_r) = \int_0^{\infty} \exp(-\lambda_r t) \gamma(t) dt$. It is obvious that calculation of k_{GH} requires knowledge of potential barrier parameters and the complete viscoelastic response of the solvent, demonstrating the fundamental intimate link between condensed-phase reaction dynamics and solvation dynamics. This kind of description may be equivalently transferred to the dielectric response of the solvent causing dielectric friction effects in reactions with significant and fast charge rearrangement [64, 65 and 66].

In the Smoluchowski limit the reaction is by definition the slow coordinate, such that $\hat{\gamma}(\lambda_r) \approx \hat{\gamma}(0) = \int_0^{\infty} \gamma(t) dt$, $\hat{\gamma}(0) \gg \lambda_r$ and $k_{GH} \approx k_{SM} = k_{\infty} \omega_B M / \hat{\gamma}(0)$. Though the time-dependent friction in principle is accessible *via* molecular dynamics simulations, for practical purposes in chemical kinetics in most cases analytical friction models have to be used including a short-time Gaussian ‘inertial’ component and a hydrodynamic tail at longer times. In the Grote–Hynes description the latter term only comes into play when the barrier top is sufficiently flat. As has been pointed out, the reactive mode frequency λ_r can be interpreted as an effective barrier curvature such that coupling of the reaction coordinate to the solvent changes position and shape of the barrier in phase space.

Because of the general difficulty encountered in generating reliable potentials energy surfaces and estimating reasonable friction kernels, it still remains an open question whether by analysis of experimental rate constants one can decide whether non-Markovian bath effects or other influences cause a particular solvent or pressure dependence of reaction rate coefficients in condensed phase. From that point of view, a purely

empirical friction model might be a viable alternative, in which the frequency-dependent friction is replaced by a state-dependent friction $\hat{\gamma}(\omega) \rightarrow K^2 = \omega_B^2/(A/\eta + B)$ that is described in terms of properties of PES and solute–solvent interaction, depicting the reaction as occurring in a frozen environment of fixed microscopic viscosity [67, 68].

(E) MICROSCOPIC FRICTION

The relation between the microscopic friction acting on a molecule during its motion in a solvent environment and macroscopic bulk solvent viscosity is a key problem affecting the rates of many reactions in condensed phase. The sequence of steps leading from friction to diffusion coefficient to viscosity is based on the general validity of the Stokes–Einstein relation and the concept of describing friction by hydrodynamic as opposed to microscopic models involving local solvent structure. In the hydrodynamic limit the effect of solvent friction on, for example, rotational relaxation times of a solute molecule is [69]

$$\tau_{\text{rot}} = 1/6D_{\text{rot}} = (V_h/kT)\eta f_{\text{bc}}C + \tau_0 \quad (\text{A3.6.35})$$

where V_h is the hydrodynamic volume of the solute in the particular solvent, whereas f_{bc} and C are parameters describing hydrodynamic boundary conditions and correcting for aspherical shape, respectively. τ_0 in turn may be related to the relaxation time of the free rotor. Though in many cases this equation correctly reproduces the viscosity dependence of τ_{rot} , in particular when solute and solvent molecules are comparable in size there are quite a number of significant deviations. One may incorporate this size effect by explicitly considering the first solvation shell on the

-25-

solute surface which, under the assumption of slip boundary conditions, gives for the correction factor C in equation (A3.6.35):

$$C_{\text{size}} = \frac{f_{\text{slip}}V_h}{f_{\text{slip}}V_h + BkT\kappa_T\eta(4/\sigma_r^2 + 1)} \quad (\text{A3.6.36})$$

with isothermal compressibility κ_T , ratio of radii of solvent to solute σ_r and a temperature-dependent parameter B . If one compares equation (A3.6.36) with the empirical friction model mentioned above, one realizes that both contain a factor of the form $C = 1/1 + a\eta$, suggesting that these models might be physically related.

Another, purely experimental possibility to obtain a better estimate of the friction coefficient for rotational motion γ_{rot} in chemical reactions consists of measuring rotational relaxation times τ_{rot} of reactants and calculating it according to equation (A3.6.35) as $\gamma_{\text{rot}} = 6kT\tau_{\text{rot}}$.

A3.6.4 SELECTED REACTIONS

A3.6.4.1 PHOTOISOMERIZATION

According to Kramers' model, for flat barrier tops associated with predominantly small barriers, the transition from the low- to the high-damping regime is expected to occur in low-density fluids. This expectation is borne

out by an extensively studied model reaction, the photoisomerization of *trans*-stilbene and similar compounds [70, 71] involving a small energy barrier in the first excited singlet state whose decay after photoexcitation is directly related to the rate coefficient of *trans-cis*-photoisomerization and can be conveniently measured by ultrafast laser spectroscopic techniques.

(A) PRESSURE DEPENDENCE OF PHOTOISOMERIZATION RATE CONSTANTS

The results of pressure-dependent measurements for *trans*-stilbene in supercritical *n*-pentane [46] (figure A3.6.5) and the prediction from the model described by equation (A3.6.29), using experimentally determined microcanonical rate coefficients in jet-cooled *trans*-stilbene to calculate k_∞ , show two marked discrepancies between model calculation and measurement: (1) experimental values of k are an order of magnitude higher already at low pressure and (2) the decrease of k due to friction is much less pronounced than predicted. As interpretations for the first observation, several ideas have been put forward that will not be further discussed here, such as a decrease of the effective potential barrier height due to electrostatic solute–solvent interactions enhanced by cluster formation at relatively low pressures [72, 73], or incomplete intramolecular vibrational energy redistribution in the isolated molecule [74, 75, 76, 77, 78, 79 and 80], or Franck–Condon cooling in the excitation process [79, 80]. The second effect, the weak viscosity dependence, which was first observed in solvent series experiments in liquid solution [81, 82 and 83], has also led to controversial interpretations: (i) the macroscopic solvent viscosity is an inadequate measure for microscopic friction acting along the reaction path [84, 85], (ii) the multidimensional character of the barrier crossing process leads to a fractional power dependence of k on $1/\eta$ [54, 81, 86, 87], (iii) as the reaction is very fast, one has to take into account the finite response time of the solvent, i.e. consider frequency-dependent friction [81, 87] and (iv) the effective barrier

-26-

height decreases further with increasing electronic polarizability and polarity of the solvent, and the observed phenomenon is a manifestation of the superposition of a static solvent effect and hydrodynamic solvent friction correctly described by η [88]. One may test these hypotheses by studying molecular rotational motion and reaction independently in compressed sample solutions. A few examples will serve here to illustrate the main conclusions one can draw from the experimental results.

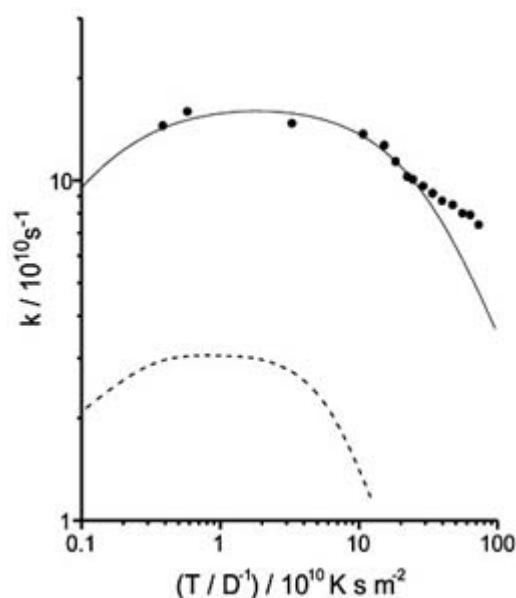


Figure A3.6.5. Photoisomerization rate constant of *trans*-stilbene in *n*-pentane versus inverse of the self-diffusion coefficient. Points represent experimental data, the dashed curve is a model calculation based on an

RRKM fit to microcanonical rate constants of isolated *trans*-stilbene and the solid curve a fit that uses a reaction barrier height reduced by solute–solvent interaction [46].

(B) MICROSCOPIC AND FREQUENCY-DEPENDENT FRICTION

Rotational relaxation times τ_{rot} of *trans*-stilbene and E,E-diphenylbutadiene (DPB) in liquid solvents like subcritical ethane and *n*-octane show a perfectly linear viscosity dependence with a slope that depends on the solvent [89] (figure A3.6.6), showing that microscopic friction acting during molecular rotational diffusion is proportional to the macroscopic solvent viscosity and that the relevant solute–solvent coupling changes with solvent. It seems reasonable to assume, therefore, that a corresponding relation also holds for microscopic friction governing diffusive motion along the reaction path.

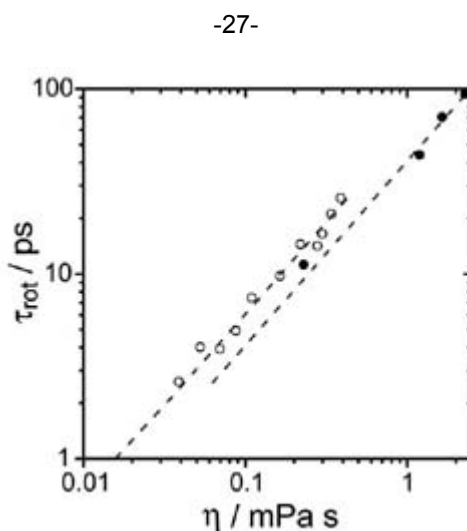


Figure A3.6.6. Viscosity dependence of rotational relaxation times of *trans*-stilbene in ethane (open circles) and *n*-octane (full circles) [89].

The validity of this assumption is apparent in the viscosity dependence of rate coefficients for S_1 -photoisomerization reactions in a number of related molecules such as *cis*-stilbene [90] (see figure A3.6.7), tetraphenylethylene (TPE) [91], DPB [92] and ‘stiff’ *trans*-stilbene [93] (where the phenyl ring is fixed by a five-membered ring to the ethylenic carbon atom). In all these cases a study of the pressure dependence reveals a linear correlation between k and $1/\eta$ in *n*-alkane and *n*-alkanol solvents, again with a solvent-dependent slope. The time scale for motion along the reaction path extends from several hundred picoseconds in DPB to a couple of hundred femtoseconds in *cis*-stilbene. There is no evidence for a frequency dependence of the friction coefficient in these reactions. As the time scale for the similar reaction in *trans*-stilbene is between 30 and 300 ps, one may conclude that also in this case the dynamics is mainly controlled by the zero-frequency friction which, in turn, is adequately represented by the macroscopic solvent viscosity. Therefore, the discrepancy between experiment and model calculation observed for *trans*-stilbene in compressed-liquid *n*-alkanes does not indicate a breakdown of the simple friction model in the Kramers–Smoluchowski theory. This result is in contrast to the analysis of solvent series study in linear alkanes, in which a solvent size effect of the microviscosity was made responsible for weak viscosity dependence [94]. Surprisingly, in a different type of non-polar solvent like methylcyclohexane, an equally weak viscosity dependence was found when the pressure was varied [95]. So the details of the viscosity influence are still posing puzzling questions.

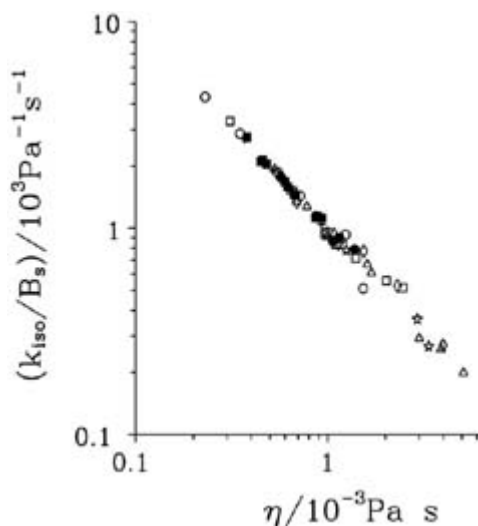


Figure A3.6.7. Viscosity dependence of reduced S_1 -decay rate constants of *cis*-stilbene in various solvents [90]. The rate constants are divided by the slope of a linear regression to the measured rate constants in the respective solvent.

(C) EFFECTIVE BARRIER HEIGHT

Measuring the pressure dependence of k at different temperatures shows that the apparent activation energy at constant viscosity decreases with increasing viscosity [46, 89] (figure A3.6.8). From a detailed analysis one can extract an effective barrier height E_0 along the reaction path that decreases linearly with increasing density of the solvent. The magnitude of this barrier shift effect is more than a factor of two in nonpolar solvents like *n*-hexane or *n*-pentane [46]. It is interesting to note that in compressed-liquid *n*-propanol one almost reaches the regime of barrierless dynamics [96]. This is also evident in the room-temperature $k(\eta)$ isotherm measured in *n*-butanol (figure A3.6.9) which turns into linear k versus $1/\eta$ dependence at higher pressures, indicating that there is no further decrease of the effective barrier height. Thus the unexpected dependence of the reaction rate on solvent viscosity is connected with specific properties of the PES of *trans*-stilbene in its first excited singlet state, because corresponding measurements for, for example, DPB or TPE in *n*-alkanes and *n*-alkanols do not show any evidence for deviations from standard Kramers–Smoluchowski behaviour.

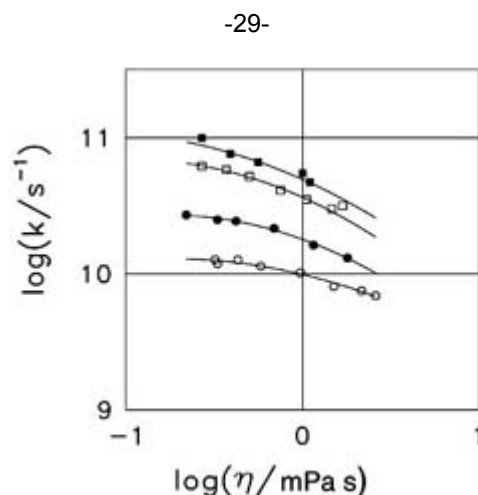


Figure A3.6.8. Isotherms of $k(\eta)$ for *trans*-stilbene photoisomerization in *n*-hexane at temperatures between

300 K (bottom) and 480 K (top). The curvature of the isotherms is interpreted as a temperature-dependent barrier shape [89].

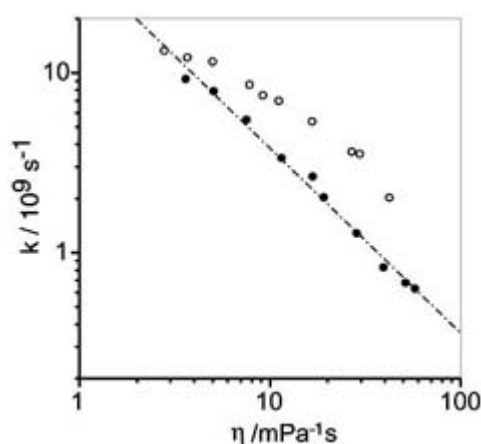


Figure A3.6.9. Viscosity dependence of photoisomerization rate constants of *trans*-stilbene (open circles) and *E,E*-diphenylbutadiene (full circles) in *n*-butanol. The broken line indicates a η^{-1} -dependence of k [96].

As a multidimensional PES for the reaction from quantum chemical calculations is not available at present, one does not know the reason for the surprising barrier effect in excited *trans*-stilbene. One could suspect that *trans*-stilbene possesses already a significant amount of zwitterionic character in the conformation at the barrier top, implying a fairly ‘late’ barrier along the reaction path towards the twisted perpendicular structure. On the other hand, it could also be possible that the effective barrier changes with viscosity as a result of a multidimensional barrier crossing process along a curved reaction path.

(D) SOLVATION DYNAMICS

The dependence of k on viscosity becomes even more puzzling when the time scale of motion along the reaction coordinate becomes comparable to that of solvent dipole reorientation around the changing charge distribution

within the reacting molecule—in addition to mechanical, one also has to consider dielectric friction. For *trans*-stilbene in ethanol, the $k(\eta)$ curve exhibits a turning point which is caused by a crossover of competing solvation and reaction time scales [97] (figure A3.6.10): as the viscosity increases the dielectric relaxation time of the solvent increases more rapidly than the typical time necessary for barrier crossing. Gradually, the solvation dynamics starts to freeze out on the time scale of reactive motion, the polar barrier is no longer decreased by solvent dipole reorientation and the rate coefficient drops more rapidly with increasing viscosity. As soon as the solvent dipoles are completely ‘frozen’, one has the same situation as in a non-polar solvent: i.e. only the electronic polarizability of the solvent causes further decrease of the barrier height.

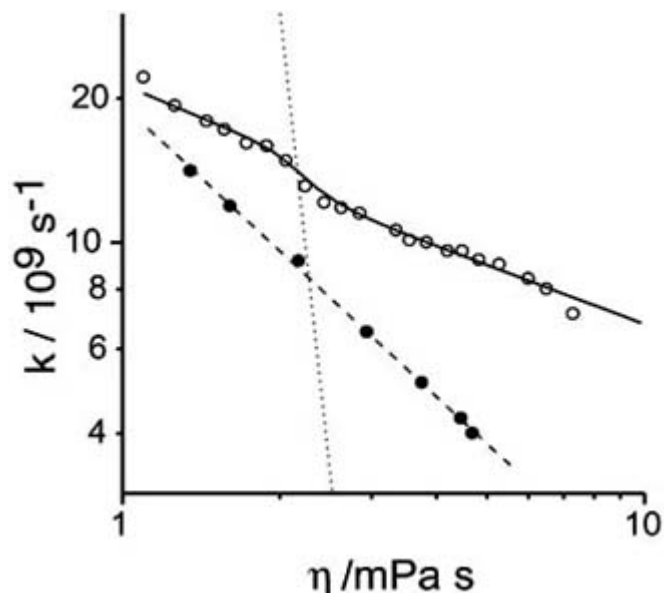


Figure A3.6.10. Viscosity dependence of photoisomerization rate constants of *trans*-stilbene (open circles) and *E,E*-diphenylbutadiene (full circles) in ethanol. The dashed line indicates a η^{-1} -dependence of k , the dotted line indicates the viscosity dependence of the dielectric relaxation time of ethanol and the solid curve is the result of a kinetic model describing the parallel processes of reaction and solvent relaxation [97].

A3.6.4.2 CHAIR-BOAT INVERSION OF CYCLOHEXANE

As mentioned above, in liquid solution most reactions are expected to have passed beyond the low-damping regime where the dynamics is dominated by activating and deactivating collisions between reactants' solvent molecules. In general, this expectation is met, as long as there is a sufficiently strong intramolecular coupling of the reaction coordinate to a large number of the remaining modes of the reactant at the transition state which leads to fast IVR within the reactant. In this case, the high-pressure limit of unimolecular rate theory is reached, and additional coupling to the liquid solvent environment leads to a decrease of the rate coefficient through the factor F_{KR} . From this point of view, the observation of rate coefficient maxima in liquid solution would appear to signal a breakdown of RRKM theory. In particular it has been argued that, for the case of weak *intramolecular* coupling, a strong coupling of the reaction coordinate to the solvent could effectively decrease the volume of phase space accessible to the reactant in

the liquid with respect to the gas phase [98, 99]. As the relative strength of *intra*- and *intermolecular* coupling may change with solvent properties, the breakdown of the RRKM model might be accompanied by the appearance of a rate coefficient maximum in liquid solution as a function of solvent friction.

Among the few reactions for which an increase of a reaction rate coefficient in liquid solution with increasing reactant–solvent coupling strength has been observed, the most notable is the thermal chair-boat isomerization reaction of cyclohexane (figure A3.6.11) and 1,1-difluorocyclohexane [100, 101, 102 and 103]. The observed pressure dependence of the rate coefficients along different isotherms was analysed in terms of one-dimensional transition state theory by introducing a transmission coefficient κ describing the effect of solvent friction $k_{\text{obs}} = \kappa k_{\text{TST}}$. In the intermediate- to high-damping regime, κ can be identified with the Kramers term F_{KR} . The observed pressure-dependent activation volumes $\Delta V_{\text{OBS}}^{\ddagger}$ were considered to represent the sum of a

pressure-independent intrinsic activation volume $\Delta V_{\text{TST}}^\ddagger$ and a pressure-dependent formal collisional activation volume $\Delta V_{\text{COLL}}^\ddagger$ arising from the increase of that reactant–solvent coupling with pressure which corresponds to viscous effects

$$RT \left(\frac{\partial \ln k_{\text{TST}}}{\partial p} \right)_T = -\Delta V_{\text{TST}}^\ddagger$$

$$RT \left(\frac{\partial \ln \kappa}{\partial p} \right)_T = -\Delta V_{\text{COLL}}^\ddagger.$$

The intrinsic volume of activation was estimated to correspond to the molar volume difference between cyclohexene and cyclohexane, adding the molar volume difference between ethane and ethene to account for the two missing protons and shortened double bond in cyclohexane. This yields a value of $\Delta V_{\text{TST}}^\ddagger = -1.5 \text{ cm}^3 \text{ mol}^{-1}$. Then, knowing the pressure dependence of the solvent viscosity, the viscosity dependence of the relative transmission coefficient κ was estimated from

$$\frac{\kappa(\eta)}{\kappa(1.5 \text{ cP})} = \frac{k_{\text{obs}}(\eta)}{k_{\text{obs}}(1.5 \text{ cP})} \exp \left[\frac{p \Delta V_{\text{TST}}^\ddagger}{RT} \right].$$

The experimental values of $\kappa(\eta)$ have a maximum at a viscosity close to 3 cP and varies by about 15% over the entire viscosity range studied. As discussed above, this unexpected dependence of κ on solvent friction in liquid CS_2 is thought to be caused by a relatively weak intramolecular coupling of the reaction coordinate to the remaining modes in cyclohexane. At viscosities below the maximum, motion along the reaction coordinate due to the reduction of the accessible phase space region is fast. The barrier passage is still in the inertial regime, and the strong coupling to the solvent leads to increasingly rapid stabilization in the product well. With increasing solvent friction, the barrier crossing enters the diffusive regime and begins to show a slowdown with further increasing solvent viscosity.

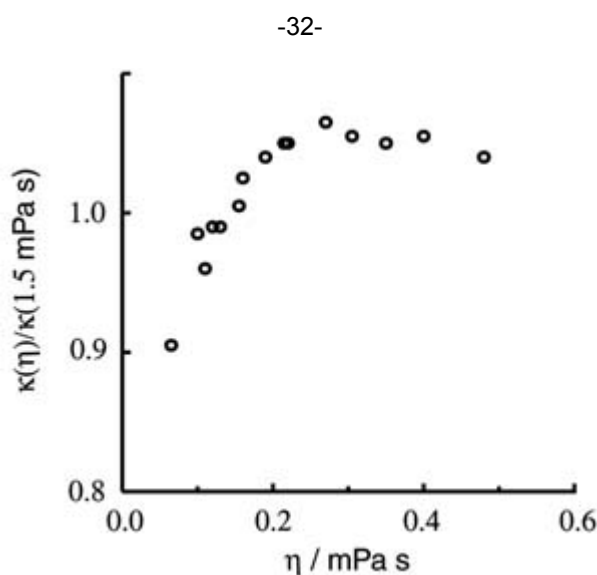


Figure A3.6.11. Viscosity dependence of transmission coefficient of the rate of cyclohexane chair-boat inversion in liquid solution (data from [100]).

This interpretation of the experimentally determined pressure dependence of the isomerization rate rests on

the assumptions that (i) the barrier height for the reaction is independent of pressure and (ii) the estimate of the intrinsic volume of activation is reliable to within a factor of two and $\Delta V_{\text{TST}}^\ddagger$ does not change with pressure. As pointed out previously, due to the differences in the pressure dependences of solvent viscosity and density, a change of the barrier height with solvent density can give rise also to an apparent maximum of the rate coefficient as a function of viscosity. In particular, a decrease of E_0 with pressure by about 1 kJ mol^{-1} could explain the observed non-monotonic viscosity dependence. Therefore, the constancy to within 0.05 kJ mol^{-1} of the isoviscous activation energy over a limited viscosity range from 1.34 to 2.0 cP lends some support to the first assumption.

From stochastic molecular dynamics calculations on the same system, in the viscosity regime covered by the experiment, it appears that *intra*- and *intermolecular* energy flow occur on comparable time scales, which leads to the conclusion that cyclohexane isomerization in liquid CS_2 is an activated process [99]. Classical molecular dynamics calculations [104] also reproduce the observed non-monotonic viscosity dependence of κ . Furthermore, they also yield a solvent contribution to the free energy of activation for the isomerization reaction which in liquid CS_2 *increases* by about 0.4 kJ mol^{-1} , when the solvent density is increased from 1.3 to 1.5 g cm^{-3} . Thus the molecular dynamics calculations support the conclusion that the high-pressure limit of this unimolecular reaction is not attained in liquid solution at ambient pressure. It has to be remembered, though, that the analysis of the measured isomerization rates depends critically on the estimated value of $\Delta V_{\text{TST}}^\ddagger$. What is still needed is a reliable calculation of this quantity in CS_2 .

A3.6.4.3 PHOTOLYTIC CAGE EFFECT AND GEMINATE RECOMBINATION

For very fast reactions, the competition between geminate recombination of a pair of initially formed reactants and its escape from the common solvent cage is an important phenomenon in condensed-phase kinetics that has received considerable attention both theoretically and experimentally. An extremely well studied example is the

-33-

photodissociation of iodine for which the quantum yield Φ_d decreases from unity in the dilute-gas phase by up to a factor of ten or more in compressed-liquid solution. An intuitively appealing interpretation of this so-called photolytic cage effect, predicted by Franck and Rabinovitch in the 1930s [105], is based on models describing it as diffusive escape of the pair [106], formed instantaneously at t_0 with initial separation \mathbf{r}_0 , from the solvent cage under the influence of Stokes friction subject to inner boundary conditions similar to equation (A3.6.21) [31],

$$\frac{\partial \rho}{\partial t} = D_{AA} \nabla^2 \rho + \delta(\mathbf{r} - \mathbf{r}_0) \delta(t - t_0) \quad k_{\text{mol}} \rho(\mathbf{r}, t) = 4\pi R D_{AA} \left. \frac{\partial \rho}{\partial r} \right|_R.$$

Solving this diffusion problem yields an analytical expression for the time-dependent escape probability $q(t)$:

$$q(t) = 1 - \frac{x}{z(1+x)} \left\{ \text{erfc} \left(\frac{z-1}{2y\sqrt{t}} \right) - \exp[(z-1)(1+x) + y^2 t (1+x)^2] \text{erfc} \left[y\sqrt{t}(1+x) + \frac{z-1}{2y\sqrt{t}} \right] \right\}$$

where x and y are as defined above and $z = r_0/R$. This equation can be compared with time-resolved measurements of geminate recombination dynamics in liquid solution [107, 108] if the parameters r_0 , R , D_{AA} and k_{mol} are known or can be reliably estimated. This simple diffusion model, however, does not satisfactorily represent the observed dynamics, which is in part due to the participation of different electronic states. Direct evidence for this comes from picosecond time-resolved transient absorption experiments in solution that demonstrate the involvement of the low-lying, weakly bound iodine A and A' states. In these experiments it was possible to separate geminate pair dynamics and vibrational energy relaxation of the initially formed hot iodine molecules [40, 41, 109]. The details of the complex steps of recombination dynamics are still only partially understood and the subject of mixed quantum–classical molecular dynamics simulations [110].

In order to probe the importance of van der Waals interactions between reactants and solvent, experiments in the gas–liquid transition range appear to be mandatory. Time-resolved studies of the density dependence of the cage and cluster dynamics in halogen photodissociation are needed to extend earlier quantum yield studies which clearly demonstrated the importance of van der Waals clustering at moderate gas densities [37, 111] (see figure A3.6.12). The pressure dependence of the quantum yield established the existence of two different regimes for the cage effect: (i) at low solvent densities, excitation of solvent-clustered halogen molecules leads to predissociation of the van der Waals bond and thereby to stabilization of the halogen molecule, whereas (ii), at high liquid phase densities, the hard-sphere repulsive caging takes over which leads to a strong reduction in the photodissociation quantum yield.

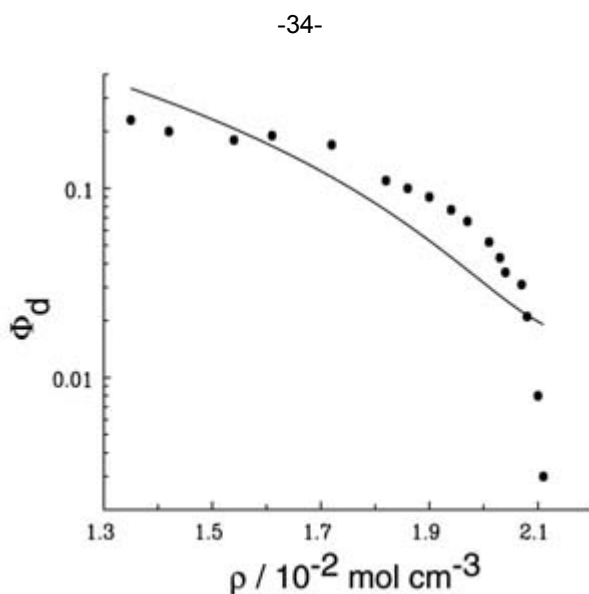


Figure A3.6.12. Photolytic cage effect of iodine in supercritical ethane. Points represent measured photodissociation quantum yields [37] and the solid curve is the result of a numerical simulation [111].

Attractive long-range and repulsive short-range forces both play a role in the cage effect, though each type dominates in a different density range. Whereas the second component has traditionally been recognized as being responsible for caging in liquid solution and solids, theoretical models and molecular dynamics calculations [112, 113] have confirmed the idea that complex formation between halogen and solvent molecules in supercritical solvents is important in photodissociation dynamics and responsible for the lowering of quantum yields at moderate gas densities [114, 115].

The traditional diffusion model permits estimation of the magnitude of the cage effect in solution according to [37]

$$\lim_{t \rightarrow \infty} q(t) = 1 - \frac{x}{z(1+x)}$$

which should directly represent overall photodissociation quantum yields measured in dense solvents, as in this quantity dynamical effects are averaged out as a consequence of multiple collisions in the cage and effective collision-induced hopping between different electronic states at large interatomic distances. The initial separation of the iodine atom pair in the solvent cage may be calculated by assuming that immediately after excitation, the atoms are spherical particles subject to Stokes friction undergoing a damped motion on a repulsive potential represented by a parabolic branch. This leads to an excitation energy dependence of the initial separation [37]

$$z - 1 = \frac{\sqrt{1 - c^2}}{2c} \exp\left(-\frac{\pi c}{\sqrt{1 - c^2}}\right) \quad \text{with } c = \frac{6\pi\eta\sigma_1 R}{\sqrt{m_1(h\nu - D_0)}} \text{ for } c < 1$$

where σ_1 and m_1 are radius and mass of the iodine atom, respectively, $h\nu$ is the photon energy and D_0 the dissociation energy of iodine molecules. Obviously, $c \geq 1$ corresponds to the overdamped case for which $r_0 = R$ irrespective of

-35-

initial energy. As in experiments a fairly weak dependence of Φ_d on excitation wavelength was found, it seems that, at least at liquid phase densities, separation of the iodine pair is overdamped, a finding corroborated by recent classical molecular dynamics simulations using simple model potentials [38].

The simple diffusion model of the cage effect again can be improved by taking effects of the local solvent structure, i.e. hydrodynamic repulsion, into account in the same way as discussed above for bimolecular reactions. The consequence is that the potential of mean force tends to favour escape at larger distances ($r_0 > 1.5R$) more than it enhances caging at small distances, leading to larger overall photodissociation quantum yields [116, 117].

The analysis of recent measurements of the density dependence of Φ_d has shown, however, that considering only the variation of solvent structure in the vicinity of the atom pair as a function of density is entirely sufficient to understand the observed changes in Φ_d with pressure and also with size of the solvent molecules [38]. Assuming that iodine atoms colliding with a solvent molecule of the first solvation shell under an angle α less than α_{\max} (the value of α_{\max} is solvent dependent and has to be found by simulations) are reflected back onto each other in the solvent cage, Φ_d is given by

$$\Phi_d = 1 - 4\pi\rho(1 - \cos\alpha_{\max}) \int_0^{r_{\text{shell}}} r^2 g(r) dr$$

where the solvation shell radius shell is obtained from Lennard-Jones radii (figure A3.6.13).

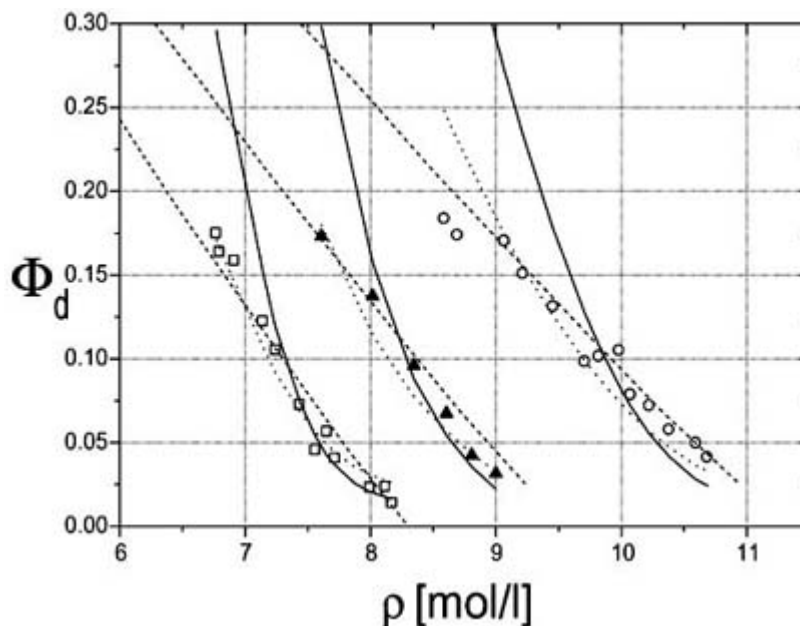


Figure A3.6.13. Density dependence of the photolytic cage effect of iodine in compressed liquid *n*-pentane (circles), *n*-hexane (triangles), and *n*-heptane (squares) [38]. The solid curves represent calculations using the diffusion model [37], the dotted and dashed curves are from ‘static’ caging models using Carnahan–Starling packing fractions and calculated radial distribution functions, respectively [38].

-36-

As these examples have demonstrated, in particular for fast reactions, chemical kinetics can only be appropriately described if one takes into account dynamic effects, though in practice it may prove extremely difficult to separate and identify different phenomena. It seems that more experiments under systematically controlled variation of solvent environment parameters are needed, in conjunction with numerical simulations that as closely as possible mimic the experimental conditions to improve our understanding of condensed-phase reaction kinetics. The theoretical tools that are available to do so are covered in more depth in other chapters of this encyclopedia and also in comprehensive reviews [6, 118, 119].

REFERENCES

- [1] Troe J 1975 Unimolecular reactions: experiments and theories *Kinetics of Gas Reactions* ed W Jost (New York: Academic) p 835
- [2] Troe J 1978 Kinetic phenomena in gases at high pressure *High Pressure Chemistry* ed H Kelm (Amsterdam: Reidel) pp 489–520
- [3] Schroeder J and Troe J 1987 Elementary reactions in the gas–liquid transition range *Ann. Rev. Phys. Chem.* **38** 163
- [4] Schroeder J and Troe J 1993 Solvent effects in the dynamics of dissociation, recombination and isomerization reactions *Activated Barrier Crossing* ed G R Fleming and P Hänggi (Singapore: World Scientific) p 206
- [5] Kajimoto O 1999 Solvation in supercritical fluids: its effects on energy transfer and chemical reactions *Chem. Rev.* **99** 355–89
- [6] Truhlar D G, Garrett B C and Klippenstein S J 1996 Current status of transition state theory *J. Phys. Chem. A* **100** 12,771–800

- [7] Leffler J E and Grunwald E 1963 *Rates and Equilibria in Organic Reactions* (New York: Wiley)
- [8] Hildebrand J H, Prausnitz J M and Scott R L 1970 *Regular and Related Solutions* (New York: Van Nostrand)
- [9] Reichardt C 1988 *Solvents and Solvent Effects in Organic Chemistry* (Weinheim: VCH)
- [10] Steinfeld J I, Francisco J S and Hase W L 1989 *Chemical Kinetics and Dynamics* (Englewood Cliffs, NJ: Prentice-Hall)
- [11] Simkin B Ya and Sheikhet I I 1995 *Quantum Chemical and Statistical Theory of Solutions* (London: Ellis Horwood)
- [12] Fröhlich H 1958 *Theory of Dielectrics* (New York: Plenum)
- [13] Böttcher C J F 1973 *Theory of Dielectric Polarization* (Amsterdam: Elsevier)
- [14] Abraham M H 1974 Solvent effects on transition states and reaction rates *Prog. Phys. Org. Chem.* **11** 1–87
- [15] Popvykh O and Tomkins R P T 1981 *Nonaqueous Solution Chemistry* (New York: Wiley)
- [16] Brebbia C A and Walker S 1980 *Boundary Element Technique in Engineering* (London: Newnes-Butterworth)
- [17] Miertuš S, Scrocco E and Tomasi J 1981 Electrostatic interactions of a solute with a continuum. A direct utilization of *ab initio* molecular potentials for the provision of solvent effects *Chem. Phys.* **55** 117–25
- [18] Aguilar M A and Olivares del Valle F J 1989 Solute–solvent interactions. A simple procedure for constructing the solvent capacity for retaining a molecular solute *Chem. Phys.* **129** 439–50
-

-37-

- [19] Aguilar M A and Olivares del Valle F J 1989 A computation procedure for the dispersion component of the interaction energy in continuum solute solvent models *Chem. Phys.* **138** 327–36
- [20] Rivail J L 1989 *New Theoretical Concepts for Understanding Organic Reactions* ed J Bertran and I G Cizmada (Amsterdam: Kluwer) p 219
- [21] Cramer C J and Truhlar D G 1996 Continuum solvation models *Solvent Effects and Chemical Reactivity* ed O Tapia and J Bertran (Dordrecht: Kluwer) pp 1–80
- [22] Mineva T, Russo N and Sicilia E 1998 Solvation effects on reaction profiles by the polarizable continuum model coupled with Gaussian density functional method *J. Comp. Chem.* **19** 290–9
- [23] Castejon H and Wiberg K B 1999 Solvent effects on methyl transfer reactions. 1. The Menshutkin reaction *J. Am. Chem. Soc.* **121** 2139–46
- [24] Evans M G and Polanyi M 1935 Some applications of the transition state method to the calculation of reaction velocities, especially in solution *Trans. Faraday Soc.* **31** 875–94
- [25] Isaacs N S 1981 *Liquid Phase High Pressure Chemistry* (Chichester: Wiley-Interscience)
- [26] Schmidt R 1998 Interpretation of reaction and activation volumes in solution *J. Phys. Chem. A* **102** 9082–6
- [27] Basilevsky M V, Weinberg N N and Zhulin V M 1985 Pressure dependence of activation and reaction volumes *J. Chem. Soc. Faraday Trans. 1* **81** 875–84
- [28] Ladanyi B M and Hynes J T 1986 Transition state solvent effects on atom transfer rates in solution *J. Am. Chem. Soc.* **108** 585–93
- [29] Smoluchowski Mv 1918 Versuch einer mathematischen Theorie der Koagulationskinetik kolloider Lösungen *Z. Phys. Chem.* **92** 129–39
- [30] Smoluchowski Mv 1915 Über Brownsche Molekularbewegung unter Einwirkung äußerer Kräfte und deren Zusammenhang mit der verallgemeinerten Diffusionsgleichung *Ann. Phys.* **48** 1103–12
- [31] Rice S A 1985 Diffusion-limited reactions *Comprehensive Chemical Kinetics* vol 25, ed C H Bamford, C F H Tipper and R G Compton (Amsterdam: Elsevier)

- [32] Collins F C and Kimball G E 1949 Diffusion-controlled rate processes *J. Colloid Sci.* **4** 425
- [33] Northrup S H and Hynes J T 1978 On the description of reactions in solution *Chem. Phys. Lett.* **54** 244
- [34] Northrup S H and Hynes J T 1980 The stable states picture of chemical reactions. I. Formulation for rate constants and initial condition effects *J. Chem. Phys.* **73** 2700–14
- [35] Kapral R 1981 Kinetic theory of chemical reactions in liquids *Adv. Chem. Phys.* **48** 71
- [36] Hippler H, Luther K and Troe J 1973 Untersuchung der Rekombination von Jodatomen in stark komprimierten Gasen und in Flüssigkeiten *Ber. Bunsenges Phys. Chem.* **77** 1104–14
- [37] Otto B, Schroeder J and Troe J 1984 Photolytic cage effect and atom recombination of iodine in compressed gases and liquids: experiments and simple models *J. Chem. Phys.* **81** 202
- [38] Schwarzer D, Schroeder J and Schröder Ch 2000 Quantum yields for the photodissociation of iodine in compressed liquids and supercritical fluids *Z. Phys. Chem.* **214**
-

-38-

- [39] Sceats M G 1988 *Chem. Phys. Lett.* **143** 123
- [40] Harris A L, Berg M and Harris C B 1986 Studies of chemical reactivity in the condensed phase. I. The dynamics of iodine photodissociation and recombination on a picosecond time scale and comparison to theories for chemical reactions in solution *J. Chem. Phys.* **84** 788
- [41] Paige M E, Russell D J and Harris C B 1986 Studies of chemical reactivity in the condensed phase. II. Vibrational relaxation of iodine in liquid xenon following geminate recombination *J. Chem. Phys.* **85** 3699–700
- [42] Wang W, Nelson K A, Xiao L and Coker D F 1994 Molecular dynamics simulation studies of solvent cage effects on photodissociation in condensed phases *J. Chem. Phys.* **101** 9663–71
- [43] Batista V S and Coker D F 1996 Nonadiabatic molecular dynamics simulation of photodissociation and geminate recombination of I₂ liquid xenon *J. Chem. Phys.* **105** 4033–54
- [44] Gilbert R G and Smith S C 1990 *Theory of Unimolecular and Recombination Reactions* (Oxford: Blackwell)
- [45] Kramers H A 1940 Brownian motion in a field of force and the diffusion model of chemical reactions *Physica* **7** 284–304
- [46] Schroeder J, Troe J and Vöhringer P 1995 Photoisomerization of *trans*-stilbene in compressed solvents: Kramers turnover and solvent induced barrier shift *Z. Phys. Chem.* **188** 287
- [47] Hänggi P, Talkner P and Borkovec M 1990 Reaction-rate theory: fifty years after Kramers *Rev. Mod. Phys.* **62** 251–341
- [48] Brinkman H C 1956 Brownian motion in a field of force and the diffusion theory of chemical reactions *Physica* **12** 149–55
- [49] Garrity D K and Skinner J L 1983 Effect of potential shape on isomerization rate constants for the BGK model *Chem. Phys. Lett.* **95** 46–51
- [50] Larson R S and Lightfoot E J 1988 Thermally activated escape from a Lennard-Jones potential well *Physica A* **149** 296–312
- [51] Larson R S and Kostin M D 1982 Kramers' theory of chemical kinetics: curvilinear reaction coordinates *J. Chem. Phys.* **77** 5017–25
- [52] Larson R S 1986 Simulation of two-dimensional diffusive barrier crossing with a curved reaction path *Physica A* **137** 295–305
- [53] Gehrke C, Schroeder J, Schwarzer D, Troe J and Voss F 1990 Photoisomerization of diphenylbutadiene in low-viscosity nonpolar solvents: experimental manifestations of multidimensional Kramers behavior and cluster effects *J. Chem. Phys.* **92** 4805–16

- [54] Agmon N and Kosloff R 1987 Dynamics of two-dimensional diffusional barrier crossing *J. Phys. Chem.* **91** 1988–96
- [55] Berezhkovskii A M, Berezhkovskii L M and Zitserman V Yu 1989 The rate constant in the Kramers multidimensional theory and th *Chem. Phys.* **130** 55–63
- [56] Berezhkovskii A M and Zitserman V Yu 1990 Activated rate processes in a multidimensional case *Physica A* **166** 585–621
- [57] Berezhkovskii A M and Zitserman V Yu 1991 Activated rate processes in the multidimensional case. Consideration of recrossings in the multidimensional Kramers problem with anisotropic friction *Chem. Phys.* **157** 141–55
-

-39-

- [58] Berezhkovskii A M and Zitserman V Yu 1991 Comment on: diffusion theory of multidimensional activated rate processes: the role of anisotropy *J. Chem. Phys.* **95** 1424
- [59] Berezhkovskii A M and Zitserman V Yu 1992 Generalization of the Kramers–Langer theory: decay of the metastable state in the case of strongly anisotropic friction *J. Phys. A: Math. Gen.* **25** 2077–92
- [60] Berezhkovskii A M and Zitserman V Yu 1992 Multidimensional activated rate processes with slowly relaxing mode *Physica A* **187** 519–50
- [61] Berezhkovskii A M and Zitserman V Yu 1993 Multi-dimensional Kramers theory of the reaction rate with highly anisotropic friction. Energy diffusion for the fast coordinate versus overdamped regime for the slow coordinate *Chem. Phys. Lett.* **212** 413–19
- [62] Zwanzig R 1973 Nonlinear generalized langevin equations *J. Stat. Phys.* **9** 215–20
- [63] Grote R F and Hynes J T 1980 The stable states picture of chemical reactions. II. Rate constants for condensed and gas phase reaction models *J. Chem. Phys.* **73** 2715–32
- [64] Van der Zwan G and Hynes J T 1982 Dynamical polar solvent effects on solution reactions: A simple continuum model *J. Chem. Phys.* **76** 2993–3001
- [65] Van der Zwan G and Hynes J T 1983 Nonequilibrium solvation dynamics in solution reaction *J. Chem. Phys.* **78** 4174–85
- [66] Van der Zwan G and Hynes J T 1984 A simple dipole isomerization model for non-equilibrium solvation dynamics in reactions in polar solvents *Chem. Phys.* **90** 21–35
- [67] Zhu S-B, Lee J, Robinson G W and Lin S H 1988 A microscopic form of the extended Kramers equation. A simple friction model for cis-trans isomerization reactions *Chem. Phys. Lett.* **148** 164–8
- [68] Zhu S-B, Lee J, Robinson G W and Lin S H 1989 Theoretical study of memory kernel and velocity correlation function for condensed phase isomerization. I. Memory kernel *J. Chem. Phys.* **90** 6335–9
- [69] Dote J L, Kivelson D and Schwartz R N 1981 A molecular quasi-hydrodynamic free-space model for molecular rotational relaxation *J. Phys. Chem.* **85** 2169–80
- [70] Waldeck D H 1991 *Chem. Rev.* **91** 415
- [71] Waldeck D H 1993 Photoisomerization dynamics of stilbenes in polar solvents *J. Mol. Liq.* **57** 127–48
- [72] Schroeder J, Schwarzer D, Troe J and Voss F 1990 Cluster and barrier effects in the temperature and pressure dependence of the photoisomerization of trans-stilbene *J. Chem. Phys.* **93** 2393–404
- [73] Meyer A, Schroeder J and Troe J 1999 Photoisomerization of *trans*-stilbene in moderately compressed gases: pressure-dependent effective barriers *J. Phys. Chem. A* **103** 10 528–39
- [74] Khundkar L R, Marcus R A and Zewail A H 1983 Unimolecular reactions at low energies and RRKM-behaviour: isomerization and dissociation *J. Phys. Chem.* **87** 2473–6
- [75] Syage J A, Felker P M and Zewail A H 1984 Picosecond dynamics and photoisomerization of stilbene in supersonic beams. I. Spectra and mode assignments *J. Chem. Phys.* **81** 4685–705

- [76] Syage J A, Felker P M and Zewail A H 1984 Picosecond dynamics and photoisomerization of stilbene in supersonic beams. II. Reaction rates and potential energy surface *J. Chem. Phys.* **81** 4706–23
- [77] Nordholm S 1989 Photoisomerization of stilbene—a theoretical study of deuteration shifts and limited internal vibrational redistribution *Chem. Phys.* **137** 109–20
-

-40-

- [78] Bolton K and Nordholm S 1996 A classical molecular dynamics study of the intramolecular energy transfer of model trans-stilbene *Chem. Phys.* **203** 101–26
- [79] Leitner D M and Wolynes P G 1997 Quantum energy flow during molecular isomerization *Chem. Phys. Lett.* **280** 411–18
- [80] Leitner D M 1999 Influence of quantum energy flow and localization on molecular isomerization in gas and condensed phases *Int. J. Quant. Chem.* **75** 523–31
- [81] Rothenberger G, Negus D K and Hochstrasser R M 1983 Solvent influence on photoisomerization dynamics *J. Chem. Phys.* **79** 5360–7
- [82] Sundström V and Gillbro T 1984 Dynamics of the isomerization of trans-stilbene in n-alcohols studied by ultraviolet picosecond absorption recovery *Chem. Phys. Lett.* **109** 538–43
- [83] Sundström V and Gillbro T 1985 Dynamics of trans-cis photoisomerization of stilbene in hydrocarbon solutions *Ber. Bunsenges Phys. Chem.* **89** 222–6
- [84] Courtney S H, Kim S K, Canonica S and Fleming G R 1986 Rotational diffusion of stilbene in alkane and alcohol solutions *J. Chem. Soc. Faraday Trans. 2* **82** 2065–72
- [85] Lee M, Haseltine J N, Smith A B III and Hochstrasser R M 1989 Isomerization processes of electronically excited stilbene and diphenylbutadiene in liquids: Are they one-dimensional? *J. Am. Chem. Soc.* **111** 5044–51
- [86] Park N S and Waldeck D H 1989 Implications for multidimensional effects on isomerization dynamics: photoisomerization study of 4,4'-dimethylstilbene in n-alkane solvents *J. Chem. Phys.* **91** 943–52
- Park N S and Waldeck D H 1990 On the dimensionality of stilbene isomerization *Chem. Phys. Lett.* **168** 379–84
- [87] Velsko S P and Fleming G R 1982 Photochemical isomerization in solution. Photophysics of diphenylbutadiene *J. Chem. Phys.* **76** 3553–62
- [88] Schroeder J and Troe J 1985 Solvent shift and transport contributions in reactions in dense media *Chem. Phys. Lett.* **116** 453
- [89] Schroeder J 1997 Picosecond kinetics of trans-cis photoisomerisations: from jet-cooled molecules to compressed solutions *Ber. Bunsenges Phys. Chem.* **101** 643
- [90] Nikowa L, Schwarzer D, Troe J and Schroeder J 1992 Viscosity and solvent dependence of low barrier processes: photoisomerization of cis-stilbene in compressed liquid solvents *J. Chem. Phys.* **97** 4827
- [91] Schroeder J 1996 The role of solute-solvent interactions in the dynamics of unimolecular reactions in compressed solvents *J. Phys.: Condens. Matter* **8** 9379
- [92] Gehrke C, Mohrschladt R, Schroeder J, Troe J and Vöhringer P 1991 Photoisomerization dynamics of diphenylbutadiene in compressed liquid alkanes and in solid environment *Chem. Phys.* **152** 45
- [93] Mohrschladt R, Schroeder J, Troe J, Vöhringer P and Votsmeier M 1994 Solvent influence on barrier crossing in the S₁-state of cis- and trans-'stiff' stilbene *Ultrafast Phenomena IX* ed P F Barbara *et al* (New York: Springer) pp 499–503
- [94] Saltiel J and Sun Y-P 1989 Intrinsic potential energy barrier for twisting in the trans-stilbene S₁ State in hydrocarbon solvents *J. Phys. Chem.* **93** 6246–50
- Sun Y-P and Saltiel J 1989 Application of the Kramers equation to stilbene photoisomerization in n-alkanes using translational diffusion coefficients to define microviscosity *J. Phys. Chem.* **93** 8310–16
-

- [95] Vöhringer P 1993 Photoisomerisierung in komprimierten Lösungen. Dissertation, Göttingen University
- [96] Schroeder J, Schwarzer D, Troe J and Vöhringer P 1994 From barrier crossing to barrierless relaxation dynamics: photoisomerization of *trans*-stilbene in compressed alkanols *Chem. Phys. Lett.* **218** 43
- [97] Mohrschladt R, Schroeder J, Schwarzer D, Troe J and Vöhringer P 1994 Barrier crossing and solvation dynamics in polar solvents: photoisomerization of *trans*-stilbene and *E,E*-diphenylbutadiene in compressed alkanols *J. Chem. Phys.* **101** 7566
- [98] Borkovec M and Berne B J 1985 Reaction dynamics in the low pressure regime: the Kramers model and collision models of molecules with many degrees of freedom *J. Chem. Phys.* **82** 794–9
- Borkovec M, Straub J E and Berne B J 1986 The influence of intramolecular vibrational relaxation on the pressure dependence of unimolecular rate constants *J. Chem. Phys.* **85** 146–9
- Straub J E and Berne B J 1986 Energy diffusion in many-dimensional Markovian systems: the consequences of competition between inter- and intramolecular vibrational energy transfer *J. Chem. Phys.* **85** 2999–3006
- [99] Kuharski R A, Chandler D, Montgomery J, Rabii F and Singer S J 1988 Stochastic molecular dynamics study of cyclohexane isomerization *J. Phys. Chem.* **92** 3261
- [100] Hasha D L, Eguchi T and Jonas J 1982 High pressure NMR study of dynamical effects on conformational isomerization of cyclohexane *J. Am. Chem. Soc.* **104** 2290
- [101] Ashcroft J, Besnard M, Aquada V and Jonas J 1984 *Chem. Phys. Lett.* **110** 420
- [102] Ashcroft J and Xie C-L 1989 *J. Chem. Phys.* **90** 5386
- [103] Campbell D M, Mackowiak M and Jonas J 1992 *J. Chem. Phys.* **96** 2717
- [104] Wilson M and Chandler D 1990 *Chem. Phys.* **149** 11
- [105] Franck J and Rabinowitch E 1934 Some remarks about free radicals and photochemistry of solutions *Trans. Faraday Soc.* **30** 120
- [106] Noyes R M 1961 Effects of diffusion on reaction rates *Prog. React. Kinet.* **1** 129
- [107] Chuang T J, Hoffman G W and Eisinger L 1974 Picosecond studies of the cage effect and collision induced predissociation of iodine in liquids *Chem. Phys. Lett.* **25** 201
- [108] Langhoff C A, Moore B and DeMeuse M 1983 Diffusion theory and picosecond atom recombination *J. Chem. Phys.* **78** 1191
- [109] Kelley D F, Abul-Haj N A and Jang D J 1984 *J. Chem. Phys.* **80** 4105
- Harris A L, Brown J K and Harris C B 1988 *Ann. Rev. Phys. Chem.* **39** 341
- [110] Wang W, Nelson K A, Xiao L and Coker D F 1994 Molecular dynamics simulation studies of solvent cage effects on photodissociation in condensed phases *J. Chem. Phys.* **101** 9663–71
- Batista V S and Coker D F 1996 Nonadiabatic molecular dynamics simulation of photodissociation and geminate recombination of I₂ liquid xenon *J. Chem. Phys.* **105** 4033–54
- [111] Luther K and Troe J 1974 Photolytic cage effect of iodine in gases at high pressure *Chem. Phys. Lett.* **24** 85–90
- Dutoit J C, Zellweger J M and van den Bergh H 1990 *J. Chem. Phys.* **93** 242
-

- [112] Bunker D L and Davidson B S 1972 Photolytic cage effect. Monte Carlo experiments *J. Am. Chem. Soc.* **94** 1843
- [113] Murrell J N, Stace A J and Dammel R 1978 Computer simulation of the cage effect in the photodissociation of iodine *J. Chem. Soc. Faraday Trans. II* **74** 1532
- [114] Dardi P S and Dahler J S 1990 Microscopic models for iodine photodissociation quantum yields in dense fluids *J. Chem. Phys.* **93** 242–56
- [115] Dardi P S and Dahler J S 1993 A model for nonadiabatic coupling in the photodissociation of I₂-solvent complexes *J. Chem. Phys.* **98** 363–72
- [116] Northrup S H and Hynes J T 1979 Short range caging effects for reactions in solution. I. Reaction rate constants and short range caging picture *J. Chem. Phys.* **71** 871–83
- [117] Northrup S H and Hynes J T 1979 Short range caging effects for reactions in solution. II. Escape probability and time dependent reactivity *J. Chem. Phys.* **71** 884
- [118] Hynes J T 1985 The theory of reactions in solution *Theory of Chemical Reaction Dynamics* ed M Baer (Boca Raton, FL: CRC Press) pp 171–234
- [119] Tapia O and Bertran J (eds) 1996 Solvent effects and chemical reactivity *Understanding Chemical Reactivity* vol 17 (Dordrecht: Kluwer)
- [120] Zawadski A G and Hynes J T 1989 Radical recombination rate constants from gas to liquid phase *J. Phys. Chem.* **93** 7031–6

-1-

A3.7 Molecular reaction dynamics in the gas phase

Daniel M Neumark

A3.7.1 INTRODUCTION

The field of gas phase reaction dynamics is primarily concerned with understanding how the microscopic forces between atoms and molecules govern chemical reactivity. This goal is targeted by performing exacting experiments which yield measurements of detailed attributes of chemical reactions, and by developing state-of-the-art theoretical techniques in order to calculate accurate potential energy surfaces for reactions and determine the molecular dynamics that occur on these surfaces. It has recently become possible to compare experimental results with theoretical predictions on a series of benchmark reactions. This convergence of experiment and theory is leading to significant breakthroughs in our understanding of how the peaks and valleys on a potential energy surface can profoundly affect the measurable properties of a chemical reaction.

In most of gas phase reaction dynamics, the fundamental reactions of interest are bimolecular reactions,



and unimolecular photodissociation reactions,



There are significant differences between these two types of reactions as far as how they are treated experimentally and theoretically. Photodissociation typically involves excitation to an excited electronic state, whereas bimolecular reactions often occur on the ground-state potential energy surface for a reaction. In addition, the initial conditions are very different. In bimolecular collisions one has no control over the reactant orbital angular momentum (impact parameter), whereas in photodissociation one can start with cold molecules with total angular momentum $J \approx 0$. Nonetheless, many theoretical constructs and experimental methods can be applied to both types of reactions, and from the point of view of this chapter their similarities are more important than their differences.

The field of gas phase reaction dynamics has been extensively reviewed elsewhere [1, 2 and 3] in considerably greater detail than is appropriate for this chapter. Here, we begin by summarizing the key theoretical concepts and experimental techniques used in reaction dynamics, followed by a ‘case study’, the reaction $F + H_2 \rightarrow HF + H$, which serves as an illustrative example of these ideas.

A3.7.2 THEORETICAL BACKGROUND: THE POTENTIAL ENERGY SURFACE

Experimental and theoretical studies of chemical reactions are aimed at obtaining a detailed picture of the potential

-2-

energy surface on which these reactions occur. The potential energy surface represents the single most important theoretical construct in reaction dynamics. For N particles, this is a $3N - 6$ dimensional function $V(q_1 \dots q_{3N-6})$ that gives the potential energy as a function of nuclear internal coordinates. The potential energy surface for any reaction can, in principle, be found by solving the electronic Schrödinger equation at many different nuclear configurations and then fitting the results to various functional forms, in order to obtain a smoothly varying surface in multiple dimensions. In practice, this is extremely demanding from a computational perspective. Thus, much of theoretical reaction dynamics as recently as a few years ago was performed on highly approximate model surfaces for chemical reactions which were generated using simple empirical functions (the London–Eyring–Polanyi–Sato potential, for example [4]). The $H + H_2$ reaction was the first for which an accurate surface fitted to *ab initio* points was generated [5, 6]. However, recent conceptual and computational advances have made it possible to construct accurate surfaces for a small number of benchmark systems, including the $F + H_2$, $Cl + H_2$ and $OH + H_2$ reactions [7, 8 and 9]. Even in these systems, one must be concerned with the possibility that a single Born–Oppenheimer potential energy surface is insufficient to describe the full dynamics [10].

Let us consider the general properties of a potential energy surface for a bimolecular reaction involving three atoms, i.e. equation (A3.7.1) with A , B and C all atomic species. A three-atom reaction requires a three-dimensional function. It is more convenient to plot two-dimensional surfaces in which all coordinates but two are allowed to vary. Figure A3.7.1 shows a typical example of a potential energy surface contour plot for a collinear three-atom reaction. The dotted curve represents the minimum energy path, or reaction coordinate, that leads from reactants on the lower right to products on the upper left. The reactant and product valleys (often referred to as the entrance and exit valleys, respectively) are connected by the transition-state region, where the transformation from reactants to products occurs, and ends in the product valley at the upper left. The potential energy surface shown in Figure A3.7.1 is characteristic of a ‘direct’ reaction, in that there is a single barrier (marked by ‡ in Figure A3.7.1) along the minimum energy path in the transition-state region. In the other general class of bimolecular reaction, a ‘complex’ reaction, one finds a well rather than a barrier in the transition-state region.

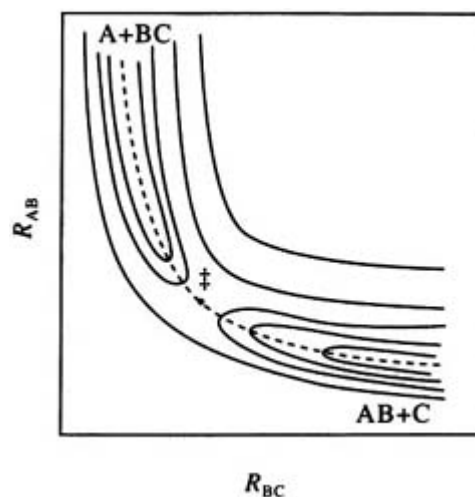


Figure A3.7.1. Two-dimensional contour plot for direct collinear reaction $A + BC \rightarrow AB + C$. Transition state is indicated by ‡.

-3-

The barrier on the surface in [figure A3.7.1](#) is actually a saddle point; the potential is a maximum along the reaction coordinate but a minimum along the direction perpendicular to the reaction coordinate. The classical transition state is defined by a slice through the top of the barrier perpendicular to the reaction coordinate. This definition holds for multiple dimensions as well; for N particles, the classical transition state is a saddle point that is unbound along the reaction coordinate but bound along the $3N - 7$ remaining coordinates. A cut through the surface at the transition state perpendicular to the reaction coordinate represents a $3N - 7$ dimensional dividing surface that acts as a ‘bottleneck’ between reactants and products. The nature of the transition state and, more generally, the region of the potential energy in the vicinity of the transition state (referred to above as the transition-state region) therefore plays a major role in determining many of the experimental observables of a reaction such as the rate constant and the product energy and angular distributions. For this reason, the transition-state region is the most important part of the potential energy surface from a computational (and experimental) perspective.

Once such an *ab initio* potential energy surface for a reaction is known, then all properties of the reaction can, in principle, be determined by carrying out multidimensional quantum scattering calculations. This is again computationally very demanding, and for many years it was more useful to perform classical and quasi-classical trajectory calculations to explore dynamics on potential energy surfaces [11]. The simpler calculations led to very valuable generalizations about reaction dynamics, showing, for example, that for an exothermic reaction with an entrance channel barrier, reactant translation was far more effective than vibration in surmounting the barrier and thus forming products, and are still very useful, since quantum effects in chemical reactions are often relatively small. However, recent conceptual and computational advances [12, 13 and 14] have now made it possible to carry out exact quantum scattering calculations on multidimensional potential energy surfaces, including the benchmark surfaces mentioned above. Comparison of such calculations with experimental observables provides a rigorous test of the potential energy surface.

A3.7.3 EXPERIMENTAL TECHNIQUES IN REACTION DYNAMICS

We now shift our focus to a general discussion of experimental chemical reaction dynamics. Given that the goal of these experiments is to gain an understanding of the reaction potential energy surface, it is important to perform experiments that can be interpreted as cleanly as possible in terms of the underlying surface. Hence, bimolecular and unimolecular reactions are often studied under ‘single-collision’ conditions, meaning

that the number density in the experiment is sufficiently low that each reactant atom or molecule undergoes at most one collision with another reactant or a photon during the course of the experiment, and the products are detected before they experience any collisions with bath gases, walls, etc. One can therefore examine the results of single-scattering events without concern for the secondary collisions and reactions that often complicate the interpretation of more standard chemical kinetics experiments. Moreover, the widespread use of supersonic beams in reaction dynamics experiments [15, 16] allows one to perform reactions under well defined initial conditions; typically the reactants are rotationally and vibrationally very cold, and the spread in collision energies (for bimolecular reactions) is narrow. The study of photodissociation reactions [2, 17] has been greatly facilitated by recent developments in laser technology, which now permit one to investigate photodissociation at virtually any wavelength over a spectral range extending from the infrared to vacuum ultraviolet (VUV).

What attributes of bimolecular and unimolecular reactions are of interest? Most important is the identity of the products, without which any further characterization is impossible. Once this is established, more detailed issues can be addressed. For example, in any exothermic reaction, one would like to determine how the excess energy is

-4-

partitioned among the translational, rotational, vibrational and electronic degrees of freedom of the products. Under the ideal of 'single-collision' conditions, one can measure the 'nascent' internal energy distribution of the products, i.e. the distribution resulting from the reaction before any relaxation (collisional or otherwise) has occurred. Measurements of the product angular distribution provide considerable insight into the topology and symmetry of the potential energy surface(s) on which the reaction occurs. More recently, the measurement of product alignment and orientation has become an area of intense interest; in photodissociation reactions, for example, one can determine if the rotational angular momentum of a molecular fragment is randomly oriented or if it tends to be parallel or perpendicular to the product velocity vector.

An incredible variety of experimental techniques have been developed over the years to address these issues. One of the most general is the crossed molecular beams method with mass spectrometric detection of the products, an experiment developed by Lee, Herschbach and co-workers [18, 19]. A schematic illustration of one version of the experiment is shown in figure A3.7.2. Two collimated beams of reactants cross in a vacuum chamber under single-collision conditions. The scattered products are detected by a rotatable mass spectrometer, in which the products are ionized by electron impact and mass selected by a quadrupole mass spectrometer. By measuring mass spectra as a function of scattering angle, one obtains angular distributions for all reaction products. In addition, by chopping either the products or one of the reactant beams with a rapidly spinning slotted wheel, one can determine the time of flight of each product from the interaction region, where the two beams cross, to the ionizer, and from this the product translational energy E_T can be determined at each scattering angle. The resulting product translational energy distributions $P(E_T)$ also contain information on the internal energy distribution of the products via conservation of energy, so long as the reactant collision energy is well defined.

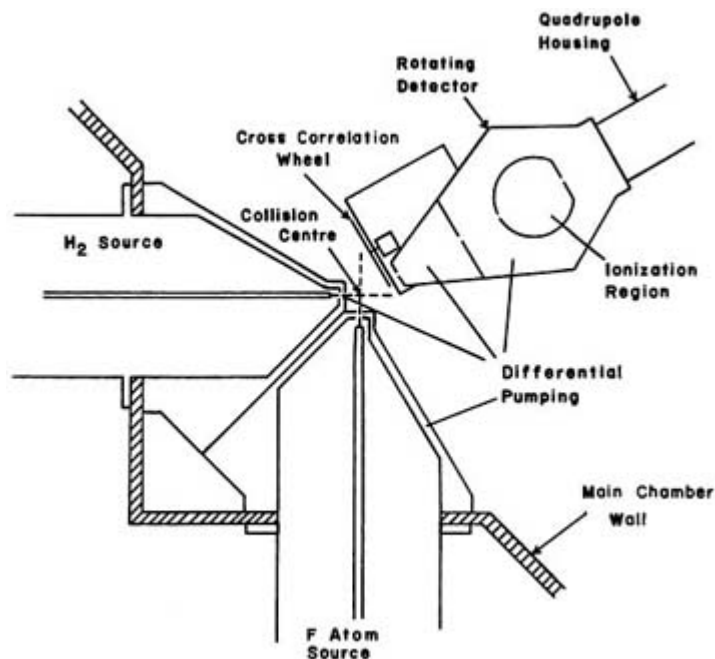


Figure A3.7.2. Schematic illustration of crossed molecular beams experiment for F + H₂ reaction.

-5-

In an important variation of this experiment, one of the reactant beams is replaced by a pulsed laser which photodissociates molecules in the remaining reactant beam. Use of a pulsed laser makes it straightforward to determine the product translational energy distribution by time of flight. This experiment, photofragment translational spectroscopy, was first demonstrated by Wilson [20, 21] and is now used in many laboratories [17].

Mass spectrometry, the primary detection method in the above crossed beams experiments, is a particularly general means of analysing reaction products, since no knowledge of the optical spectroscopy of the products is required. On the other hand, electron impact ionization often leads to extensive fragmentation, thereby complicating identification of the primary products. Very recently, tunable VUV radiation from synchrotrons has been used to ionize scattered products from both photodissociation [22] and bimolecular reactions [23]; other than the ionization mechanism, the instrument is similar in principle to that shown in figure A3.7.2. By choosing the VUV wavelength to lie above the ionization potential of the product of interest but below the lowest dissociative ionization threshold (i.e. the minimum energy for $AB + h\nu \rightarrow A^+ B + e^-$) one can eliminate fragmentation and thus simplify interpretation of the experiments.

A complementary approach to reaction dynamics centres on probing reaction products by optical spectroscopy. Optical spectroscopy often provides higher resolution on the product internal energy distribution than the measurement of translational energy distributions, but is less universally applicable than mass spectrometry as a detection scheme. If products are formed in electronically excited states, their emission spectra (electronic chemiluminescence) can be observed, but ground-state products are more problematic. Polanyi [24] made a seminal contribution in this field by showing that vibrationally excited products in their ground electronic state could be detected by spectrally resolving their spontaneous emission in the infrared; this method of 'infrared chemiluminescence' has proved of great utility in determining product vibrational and, less frequently, rotational distributions.

However, with the advent of lasers, the technique of 'laser-induced fluorescence' (LIF) has probably become the single most popular means of determining product-state distributions; an early example is the work by Zare and co-workers on Ba + HX ($X = F, Cl, Br, I$) reactions [25]. Here, a tunable laser excites an electronic transition of one of the products (the BaX product in this example), and the total fluorescence is detected as a

function of excitation frequency. This is an excellent means of characterizing molecular products with bound-bound electronic transitions and a high fluorescence quantum yield; in such cases the LIF spectra are often rotationally resolved, yielding rotational, vibrational and, for open shell species, fine-structure distributions. LIF has been used primarily for diatomic products since larger species often have efficient non-radiative decay pathways that deplete fluorescence, but there are several examples in which LIF has been used to detect polyatomic species as well.

LIF can provide more detail than the determination of the product internal energy distribution. By measuring the shape LIF profile for individual rotational lines, one can obtain Doppler profiles which yield information on the translational energy distribution of the product as well [26, 27]. In photodissociation experiments where the photolysis and probe laser are polarized, the Doppler profiles yield information on product alignment, i.e. the distribution of m_J levels for products in a particular rotational state J [28]. Experiments of this type have shown, for example, that the rotational angular momentum of the OH product from H_2O photodissociation tends to be perpendicular to \mathbf{v} [29], the vector describing the relative velocity of the products, whereas for H_2O_2 photodissociation [30] one finds J tends to be parallel to \mathbf{v} . These ‘vector correlation’ measurements [31, 32 and 33] are proving very useful in unravelling the detailed dynamics of photodissociation and, less frequently, bimolecular reactions.

The above measurements are ‘asymptotic’, in that they involve looking at the products of reaction long after the collision has taken place. These very valuable experiments are now complemented by ‘transition-state spectroscopy’

-6-

experiments, in which one uses frequency- or time-domain experiments to probe the very short-lived complex formed when two reactants collide [34]. For example, in our laboratory, we have implemented a transition-state spectroscopy experiment based on negative-ion photodetachment [35]. The principle of the experiment, in which a stable negative ion serves as a precursor for a neutral transition state, is illustrated in figure A3.7.3. If the anion geometry is similar to that of the transition state, then photodetachment of the anion will access the transition-state region on the neutral surface. The resulting photoelectron spectrum can give a vibrationally resolved picture of the transition-state dynamics, yielding the frequencies of the bound vibrational modes of the transition state (i.e. those perpendicular to the reaction coordinate) and thereby realizing the goal of transition-state spectroscopy. An example of the successful application of this technique is given below in the discussion of the $\text{F} + \text{H}_2$ reaction.

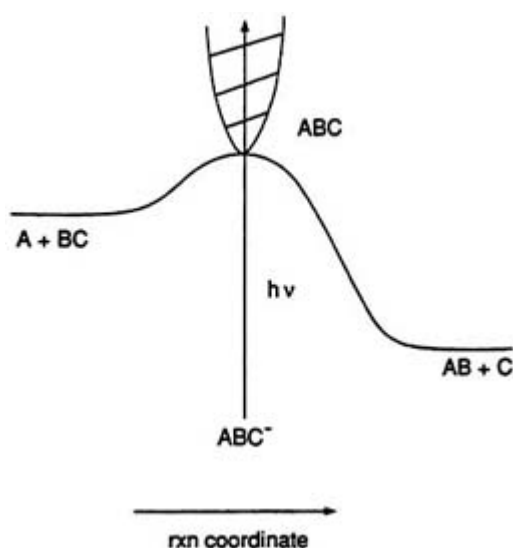


Figure A3.7.3. Principle of transition-state spectroscopy via negative-ion photodetachment.

Alternatively, one can take advantage of the developments in ultrafast laser technology and use femtosecond lasers to follow the course of a reaction in real time. In this approach, pioneered by Zewail [36], a unimolecular or bimolecular reaction is initiated by a femtosecond pump pulse, and a femtosecond probe pulse monitors some aspect of the reaction as a function of pump–probe delay time. The first example of such an experiment was the photodissociation of ICN [37]. Here the pump pulse excited ICN to a repulsive electronic state correlating to ground state $I + CN(X^2 \Sigma^+)$ products. The probe pulse excited the dissociating ICN to a second repulsive state correlating to excited $I + CN(B^2 \Sigma^+)$ products, and progress of the dissociation was monitored via LIF. If the probe pulse is tuned to be resonant with the $CN B \leftarrow X$ transition, the LIF signal rises monotonically on a 200 fs time scale, attributed to the time delay for the formation of CN product. On the other hand, at slightly redder probe wavelengths, the LIF signal rises then falls, indicative of the transient ICN^* species formed by the pump pulse. This experiment thus represented the first observation of a molecule in the act of falling apart.

In an elegant application of this method to bimolecular reactions, the reaction $H + CO_2 \rightarrow OH + CO$ was studied by forming the $CO_2 \cdot HI$ van der Waals complex, dissociating the HI moiety with the pump pulse, allowing the resulting H atom to react with the CO_2 , and then using LIF to probe the OH signal as a function of time [38]. This experiment represents the ‘real-time clocking’ of a chemical reaction, as it monitors the time interval between initiation of a bimolecular reaction and its completion.

-7-

The above discussion represents a necessarily brief summary of the aspects of chemical reaction dynamics. The theoretical focus of this field is concerned with the development of accurate potential energy surfaces and the calculation of scattering dynamics on these surfaces. Experimentally, much effort has been devoted to developing complementary asymptotic techniques for product characterization and frequency- and time-resolved techniques to study transition-state spectroscopy and dynamics. It is instructive to see what can be accomplished with all of these capabilities. Of all the benchmark reactions mentioned in [section A3.7.2](#), the reaction $F + H_2 \rightarrow HF + H$ represents the best example of how theory and experiment can converge to yield a fairly complete picture of the dynamics of a chemical reaction. Thus, the remainder of this chapter focuses on this reaction as a case study in reaction dynamics.

A3.7.4 CASE STUDY: THE F + H₂ REACTION

The energetics for the $F + H_2$ reaction is shown in figure A3.7.4. The reaction is exothermic by 32.07 kcal mol⁻¹, so that at collision energies above 0.5 kcal mol⁻¹, enough energy is available to populate HF vibrational levels up to and including $v = 3$. Hence the determination of the HF vibration–rotation distribution from this reaction has been of considerable interest. How might one go about this? Since HF does not have an easily accessible bound excited state, LIF is not an appropriate probe technique. On the other hand, the HF vibrational transitions in the infrared are exceedingly strong, and this is the spectral region where characterization of the HF internal energy distribution has been carried out.

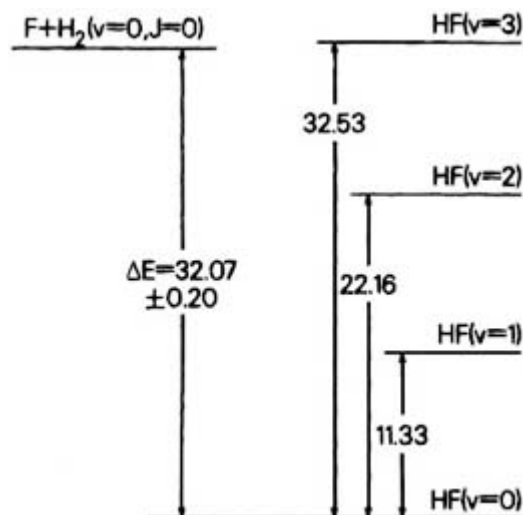


Figure A3.7.4. Energetics of the $F + H_2$ reaction. All energies in kcal mol^{-1} .

The first information on the HF vibrational distribution was obtained in two landmark studies by Pimentel [39] and Polanyi [24] in 1969; both studies showed extensive vibrational excitation of the HF product. Pimentel found that the $F + H_2$ reaction could pump an infrared chemical laser, i.e. the vibrational distribution was inverted, with the $HF(v = 2)$ population higher than that for the $HF(v = 1)$ level. A more complete picture was obtained by Polanyi by measuring and spectrally analysing the spontaneous emission from vibrationally excited HF produced by the reaction. This ‘infrared chemiluminescence’ experiment yielded relative populations of 0.29, 1 and 0.47 for the $HF(v = 1, 2$ and $3)$

-8-

vibrational levels, respectively. While improvements in these measurements were made in subsequent years, the numbers describing the vibrational populations have stayed approximately constant. The highly inverted vibrational distributions are characteristic of a potential energy surface for an exothermic reaction with a barrier in the entrance channel.

Spectroscopic determination of the HF rotational distribution is another story. In both the chemical laser and infrared chemiluminescence experiments, rotational relaxation due to collisions is faster or at least comparable to the time scale of the measurements, so that accurate determination of the nascent rotational distribution was not feasible. However, Nesbitt [40, 41] has recently carried out direct infrared absorption experiments on the HF product under single-collision conditions, thereby obtaining a full vibration–rotation distribution for the nascent products.

These spectroscopic probes have been complemented by studies using the crossed molecular beams technique. In these experiments, two well collimated and nearly monoenergetic beams of H_2 and F atoms cross in a large vacuum chamber. The scattered products are detected by a rotatable mass spectrometer, yielding the angular distribution of the reaction products. The experiment measures the transitional energy of the products via time of flight. Thus, one obtains the full transitional energy and angular distribution, $P(E_T, \theta)$, for the HF products. The first experiments of this type on the $F + D_2$ reaction were carried out by Lee [42] in 1970. Subsequent work by the Lee [43, 44] and Toennies [45, 46] groups on the $F + H_2$, D_2 and HD reactions has yielded a very complete characterization of the $P(E, \theta)$ distribution.

As an example, [figure A3.7.5](#) shows a polar contour plot of the HF product velocity distribution at a reactant collision energy of $E_{\text{coll}} = 1.84 \text{ kcal mol}^{-1}$ [43]. *p*- H_2 refers to *para*-hydrogen, for which most of the rotational population is in the $J = 0$ level under the experimental conditions used here. This plot is in the centre-of-mass (CM) frame of reference. F atoms are coming from the right, and H_2 from the left, and the

scattering angle θ is reference to the H_2 beam. The dashed circles ('Newton circles') represent the maximum speed of the HF product in a particular vibrational state, given by

$$v_{\max} = \frac{m_H}{M} \sqrt{\frac{2(\Delta E + E_{\text{coll}} - E_v)}{\mu}} \quad (\text{A3.7.3})$$

where ΔE is the exothermicity, E_v the vibrational energy, M is the total mass ($M = m_H + m_{\text{HF}}$) and $\mu = m_F m_H / M$ the reduced mass of the products. Thus, all the signal inside the $\nu = 3$ circle is from HF($\nu = 3$), all the signal inside the $\nu = 2$ circle is from HF($\nu = 2$) or HF($\nu = 3$), etc.

-9-

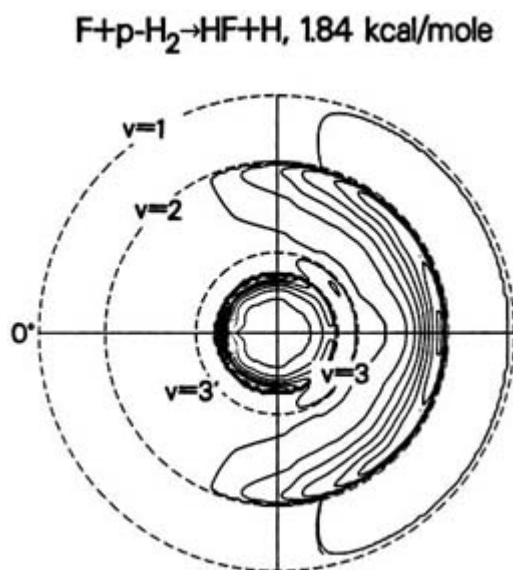


Figure A3.7.5. Velocity–flux contour plot for HF product from the reaction $F + \textit{para}\text{-}H_2 \rightarrow HF + H$ at a reactant collision energy of $1.84 \text{ kcal mol}^{-1}$.

An important feature of figure A3.7.5 is that the contributions from different HF vibrational levels, particularly the $\nu = 2$ and 3 levels, are very distinct, a result of relatively little rotational excitation of the HF ($\nu = 2$) products (i.e. if these products had sufficient rotational excitation, they would have the same translational energy as HF($\nu = 3$) product in low J levels). As a consequence, from figure A3.7.5 one can infer the angular distribution for each HF vibrational state, in other words, vibrationally state-resolved differential cross sections. These are quite different depending on the vibrational level. The HF($\nu = 2$) and ($\nu = 1$) products are primarily back-scattered with their angular distributions peaking at $\theta = \pi$, while the HF($\nu = 3$) products are predominantly forward-scattered, peaking sharply at $\theta = 0^\circ$. In general, backward-scattered products result from low impact parameter, head-on collisions, while forward-scattered products are a signature of higher impact parameter, glancing collisions. To understand the significance of these results, it is useful to move away from experimental results and consider the development of potential energy surfaces for this reaction.

Many potential energy surfaces have been proposed for the $F + H_2$ reaction. It is one of the first reactions for which a surface was generated by a high-level *ab initio* calculation including electron correlation [47]. The resulting surface (restricted to collinear geometries) was imperfect, but it had a low barrier ($1.66 \text{ kcal mol}^{-1}$) lying in the entrance channel, as expected for an exothermic reaction with a low activation energy ($\sim 1.0 \text{ kcal mol}^{-1}$). In the 1970s, several empirical surfaces were developed which were optimized so that classical trajectory calculations performed on these surfaces reproduced experimental results, primarily the rate

constant and HF vibrational energy distribution. One of these, the Muckerman V surface [48], was used in many classical and quantum mechanical scattering calculations up until the mid-1980s and provided a generally accepted theoretical foundation for the $F + H_2$ reaction. However, one notable feature of this surface was its rather stiff bend potential near the transition state. With such a potential, only near-collinear collisions were likely to lead to reaction. As a consequence, the HF product angular distribution found by scattering calculations on this surface was strongly back-scattered for all vibrational states. This is in marked disagreement with the experimental results in figure A3.7.5 which show the HF($\nu = 3$) distribution to be strongly forward-scattered.

-10-

At the time the experiments were performed (1984), this discrepancy between theory and experiment was attributed to quantum mechanical resonances that led to enhanced reaction probability in the HF($\nu = 3$) channel for high impact parameter collisions. However, since 1984, several new potential energy surfaces using a combination of *ab initio* calculations and empirical corrections were developed in which the bend potential near the barrier was found to be very flat or even non-collinear [49, 51], in contrast to the Muckerman V surface. In 1988, Sato [52] showed that classical trajectory calculations on a surface with a bent transition-state geometry produced angular distributions in which the HF($\nu = 3$) product was peaked at $\theta = 0^\circ$, while the HF($\nu = 2$) product was predominantly scattered into the backward hemisphere ($\theta \geq 90^\circ$), thereby qualitatively reproducing the most important features in figure A3.7.5.

At this point it is reasonable to ask whether comparing classical or quantum mechanical scattering calculations on model surfaces to asymptotic experimental observables such as the product energy and angular distributions is the best way to find the 'true' potential energy surface for the $F + H_2$ (or any other) reaction. From an experimental perspective, it would be desirable to probe the transition-state region of the $F + H_2$ reaction in order to obtain a more direct characterization of the bending potential, since this appears to be the key feature of the surface. From a theoretical perspective, it would seem that, with the vastly increased computational power at one's disposal compared to 10 years ago, it should be possible to construct a chemically accurate potential energy surface based entirely on *ab initio* calculations, with no reliance upon empirical corrections. Quite recently, both developments have come to pass and have been applied to the $F + H_2$ reaction.

The transition-state spectroscopy experiment based on negative-ion photodetachment described above is well suited to the study of the $F + H_2$ reaction. The experiment is carried out through measurement of the photoelectron spectrum of the anion FH_2^- . This species is calculated to be stable with a binding energy of about 0.20 eV with respect to $F^- + H_2$ [53]. Its calculated equilibrium geometry is linear and the internuclear distances are such that good overlap with the entrance barrier transition state is expected.

The photoelectron spectrum of FH_2^- is shown in figure A3.7.6 [54]. The spectrum is highly structured, showing a group of closely spaced peaks centred around 1 eV, and a smaller peak at 0.5 eV. We expect to see vibrational structure corresponding to the bound modes of the transition state perpendicular to the reaction coordinate. For this reaction with its entrance channel barrier, the reaction coordinate at the transition state is the $F \cdots H_2$ distance, and the perpendicular modes are the F–H–H bend and H–H stretch. The bend frequency should be considerably lower than the stretch. We therefore assign the closely spaced peaks to a progression in the F–H–H bend and the small peak at 0.5 eV to a transition-state level with one quantum of vibrational excitation in the H_2 stretch.

-11-

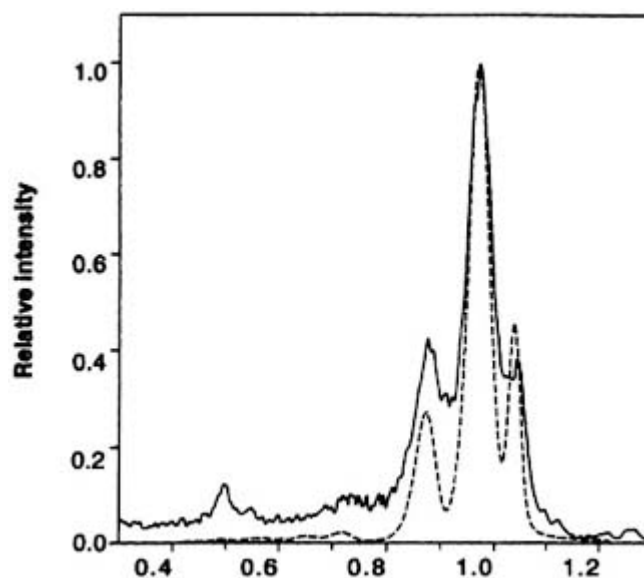


Figure A3.7.6. Photoelectron spectrum of FH_2^- . Here the F^- is complexed to *para*- H_2 . Solid curve: experimental results. Dashed curve: simulated spectrum from scattering calculation on *ab initio* surface.

The observation of a bend progression is particularly significant. In photoelectron spectroscopy, just as in electronic absorption or emission spectroscopy, the extent of vibrational progressions is governed by Franck–Condon factors between the initial and final states, i.e. the transition between the anion vibrational level ν'' and neutral level ν' is given by

$$I_{\nu''-\nu'} \propto |\langle \psi_{\nu''} | \psi_{\nu'} \rangle|^2 \quad (\text{A3.7.4})$$

where $\psi_{\nu'}$ and $\psi_{\nu''}$ are the neutral and anion vibrational wavefunctions, respectively. Since the anion is linear, a progression in a bending mode of the neutral species can only occur if the latter is bent. Hence the FH_2^- photoelectron spectrum implies that the FH_2^\ddagger transition state is bent.

While this experimental work was being carried out, an intensive theoretical effort was being undertaken by Werner and co-workers to calculate an accurate $\text{F} + \text{H}_2$ potential energy surface using purely *ab initio* methods. The many previous unsuccessful attempts indicated that an accurate calculation of the barrier height and transition-state properties requires both very large basis sets and a high degree of electron correlation; Werner incorporated both elements in his calculation. The resulting Stark–Werner (SW) surface [7] has a bent geometry at the transition state and a barrier of $1.45 \pm 0.25 \text{ kcal mol}^{-1}$. A two-dimensional contour plot of this potential near the transition state is shown in figure A3.7.7. The reason for the bent transition state is illuminating. The F atom has one half-filled p orbital and one might expect this to react most readily with H_2 by collinear approach of the reactants with the half-filled p orbital lined up with the internuclear axis of the H_2 molecule. On the other hand, at longer $\text{F} \cdots \text{H}_2$ distances, where electrostatic forces dominate, there is a minimum in the potential energy surface at a T-shaped geometry with the half-filled orbital perpendicular to the H–H bond. (This arises from the quadrupole–quadrupole interaction between the F and H_2 .)

The interplay between favourable reactivity at a collinear geometry and electrostatic forces favouring a T-shaped geometry leads to a bent geometry at the transition state.

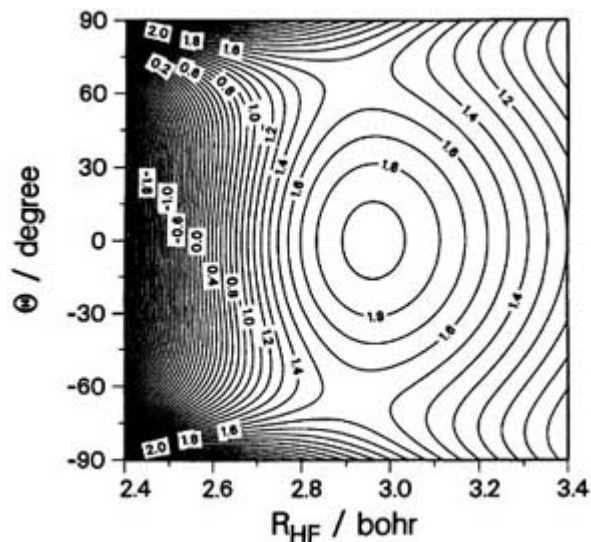


Figure A3.7.7. Two-dimensional contour plot of the Stark–Werner potential energy surface for the $F + H_2$ reaction near the transition state. Θ is the F–H–H bend angle.

How good is this surface? The first test was to simulate the FH_2 photoelectron spectrum. This calculation was carried out by Manolopoulos [54] and the result is shown as a dashed curve in figure A3.7.6. The agreement with experiment is excellent considering that no adjustable parameters are used in the calculation. In addition, Castillo *et al* [55, 56] and Aoiz *et al* [57] have performed quasi-classical and quantum scattering calculations on the SW surface to generate angular distributions for each HF product vibrational state for direct comparison to the molecular beam scattering results in figure A3.7.5. The state-specific forward-scattering of the HF($v = 3$) product is indeed reproduced in the quantum calculations and, to a somewhat lesser extent, in the quasi-classical calculations. The experimental product vibrational populations are also reproduced by the calculations. It therefore appears that scattering calculations on the SW surface agree with the key experimental results for the $F + H_2$ reaction.

What is left to understand about this reaction? One key remaining issue is the possible role of other electronic surfaces. The discussion so far has assumed that the entire reaction takes place on a single Born–Oppenheimer potential energy surface. However, three potential energy surfaces result from the interaction between an F atom and H_2 . The spin–orbit splitting between the $^2P_{3/2}$ and $^2P_{1/2}$ states of a free F atom is 404 cm^{-1} . When an F atom interacts with H_2 , the $^2P_{3/2}$ state splits into two states with A' and A'' symmetry ($^2\Sigma^+$ and $^2\Pi_{3/2}$, respectively, for collinear geometry) while the higher-lying $^2P_{1/2}$ state becomes an A' state ($^2\Pi_{1/2}$ for collinear geometry). Only the lower A' state correlates adiabatically to ground-state HF + H products; the other two states correlate to highly excited products and are therefore non-reactive in the adiabatic limit. In this limit, the excited $F(^2P_{1/2})$ state is completely unreactive.

Since this state is so low in energy, it is likely to be populated in the F atom beams typically used in scattering experiments (where pyrolysis or microwave/electrical discharges are used to generate F atoms), so the issue of its reactivity is important. The molecular beam experiments of Lee [43] and Toennies [45] showed no evidence for

reaction from the $F(^2P_{1/2})$ state. However, the recent work of Nesbitt [40, 41], in which the vibrational and rotational HF distribution was obtained by very high-resolution IR spectroscopy, shows more rotational excitation of the HF($v = 3$) product than should be energetically possible from reaction with the $F(^2P_{3/2})$ state. They therefore suggested that this rotationally excited product comes from the $F(^2P_{1/2})$ state. This work prompted an intensive theoretical study of spin–orbit effects on the potential energy surface and reaction

dynamics. A recent study by Alexander and co-workers [58] does predict a small amount of reaction from the $F(^2P_{1/2})$ state but concludes that the adiabatic picture is largely correct. The issue of whether a reaction can be described by a single Born–Oppenheimer surface is of considerable interest in chemical dynamics [10], and it appears that the effect of multiple surfaces must be considered to gain a complete picture of a reaction even for as simple a model system as the $F + H_2$ reaction.

A3.7.5 CONCLUSIONS AND PERSPECTIVES

This chapter has summarized some of the important concepts and results from what has become an exceedingly rich area of chemical physics. On the other hand, the very size of the field means that the vast majority of experimental and theoretical advances have been left out; the books referenced in the introduction provide a much more complete picture of the field.

Looking toward the future, two trends are apparent. First, the continued study of benchmark bimolecular and photodissociation reactions with increasing levels of detail is likely to continue and be extremely productive. Although many would claim that the ‘three-body problem’ is essentially solved from the perspective of chemical reaction dynamics, the possibility of multiple potential surfaces playing a role in the dynamics adds a new level of complexity even for well studied model systems such as the $F + H_2$ reaction considered here. Slightly more complicated benchmark systems such as the $OH + H_2$ and $OH + CO$ reactions present even more of a challenge to both experiment and theory, although considerable progress has been achieved in both cases.

However, in order to deliver on its promise and maximize its impact on the broader field of chemistry, the methodology of reaction dynamics must be extended toward more complex reactions involving polyatomic molecules and radicals for which even the primary products may not be known. There certainly have been examples of this: notably the crossed molecular beams work by Lee [59] on the reactions of O atoms with a series of hydrocarbons. In such cases the spectroscopy of the products is often too complicated to investigate using laser-based techniques, but the recent marriage of intense synchrotron radiation light sources with state-of-the-art scattering instruments holds considerable promise for the elucidation of the bimolecular and photodissociation dynamics of these more complex species.

REFERENCES

- [1] Levine R D and Bernstein R B 1987 *Molecular Reaction Dynamics and Chemical Reactivity* (New York: Oxford University Press)
 - [2] Schinke R 1993 *Photodissociation Dynamics* (Cambridge: Cambridge University Press)
 - [3] Scoles G (ed) 1988 *Atomic and Molecular Beam Methods* vols 1 and 2 (New York: Oxford University Press)
 - [4] Sato S 1955 *J. Chem. Phys.* **23** 592
 - [5] Siegbahn P and Liu B 1978 *J. Chem. Phys.* **68** 2457
-

- [6] Truhlar D G and Horowitz C J 1978 *J. Chem. Phys.* **68** 2466
- [7] Stark K and Werner H J 1996 *J. Chem. Phys.* **104** 6515
- [8] Alagia M *et al* 1996 *Science* **273** 1519
- [9] Alagia M, Balucani N, Casavecchia P, Stranges D, Volpi G G, Clary D C, Kliesch A and Werner H J 1996 *Chem. Phys.* **207** 389
- [10] Butler L J 1998 *Annu. Rev. Phys. Chem.* **49** 125
- [11] Polanyi J C 1972 *Acc. Chem. Res.* **5** 161

- [12] Miller W H 1990 *Annu. Rev. Phys. Chem.* **41** 245
- [13] Bowman J M and Schatz G C 1995 *Annu. Rev. Phys. Chem.* **46** 169
- [14] Schatz G C 1996 *J. Phys. Chem.* **100** 12 839
- [15] Anderson J B, Andres R P and Fenn J B 1966 *Adv. Chem. Phys.* **10** 275
- [16] Miller D R 1988 *Atomic and Molecular Beam Methods* vol 1, ed G Scoles (New York: Oxford University Press) p 14
- [17] Butler L J and Neumark D M 1996 *J. Phys. Chem.* **100** 12 801
- [18] Lee Y T, McDonald J D, LeBreton P R and Herschbach D R 1969 *Rev. Sci. Instrum.* **40** 1402
- [19] McDonald J D, LeBreton P R, Lee Y T and Herschbach D R 1972 *J. Chem. Phys.* **56** 769
- [20] Busch G E and Wilson K R 1972 *J. Chem. Phys.* **56** 3626
- [21] Busch G E and Wilson K R 1972 *J. Chem. Phys.* **56** 3638
- [22] Sun W Z, Yokoyama K, Robinson J C, Suits A G and Neumark D M 1999 *J. Chem. Phys.* **110** 4363
- [23] Blank D A, Hemmi N, Suits A G and Lee Y T 1998 *Chem. Phys.* **231** 261
- [24] Polanyi J C and Tardy D C 1969 *J. Chem. Phys.* **51** 5717
- [25] Cruse H W, Dagdigian P J and Zare R N 1973 *Discuss. Faraday* **55** 277
- [26] Ondrey G, van Veen N and Bersohn R 1983 *J. Chem. Phys.* **78** 3732
- [27] Vasudev R, Zare R N and Dixon R N 1984 *J. Chem. Phys.* **80** 4863
- [28] Greene C H and Zare R N 1983 *J. Chem. Phys.* **78** 6741
- [29] David D, Bar I and Rosenwaks S 1993 *J. Phys. Chem.* **97** 11 571
- [30] Gericke K-H, Klee S, Comes F J and Dixon R N 1986 *J. Chem. Phys.* **85** 4463
- [31] Dixon R N 1986 *J. Chem. Phys.* **85** 1866
- [32] Simons J P 1987 *J. Phys. Chem.* **91** 5378
- [33] Hall G E and Houston P L 1989 *Annu. Rev. Phys. Chem.* **40** 375
- [34] Polanyi J C and Zewail A H 1995 *Acc. Chem. Res.* **28** 119
- [35] Neumark D M 1993 *Acc. Chem. Res.* **26** 33
- [36] Khundkar L R and Zewail A H 1990 *Annu. Rev. Phys. Chem.* **41** 15
- [37] Dantus M, Rosker M J and Zewail A H 1987 *J. Chem. Phys.* **87** 2395
- [38] Scherer N F, Khundkar L R, Bernstein R B and Zewail A H 1987 *J. Chem. Phys.* **87** 1451
- [39] Parker J H and Pimentel G C 1969 *J. Chem. Phys.* **51** 91
- [40] Chapman W B, Blackmon B W and Nesbitt D J 1997 *J. Chem. Phys.* **107** 8193
- [41] Chapman W B, Blackmon B W, Nizkorodov S and Nesbitt D J 1998 *J. Chem. Phys.* **109** 9306
- [42] Schafer T P, Siska P E, Parson J M, Tully F P, Wong Y C and Lee Y T 1970 *J. Chem. Phys.* **53** 3385
- [43] Neumark D M, Wodtke A M, Robinson G N, Hayden C C and Lee Y T 1985 *J. Chem. Phys.* **92** 3045
- [44] Neumark D M, Wodtke A M, Robinson G N, Hayden C C, Shobotake K, Sparks R K, Schafer T P and Lee Y T 1985 *J. Chem. Phys.* **82** 3067
- [45] Faubel M, Martinezhaya B, Rusin L Y, Tappe U, Toennies J P, Aoiz F J and Banares L 1996 *Chem. Phys.* **207** 227
- [46] Baer M, Faubel M, Martinez-Haya B, Rusin L, Tappe U and Toennies J P 1999 *J. Chem. Phys.* **110** 10 231
- [47] Bender C F, Pearson P K, O'Neill S V and Schaefer H F 1972 *Science* **176** 1412

- [48] Muckerman J T 1971 *Theoretical Chemistry—Advances and Perspectives* vol 6A, ed H Eyring and D Henderson (New York: Academic) p 1
- [49] Brown F B, Steckler R, Schwenke D W, Truhlar D G and Garrett B C 1985 *J. Chem. Phys.* **82** 188
- [50] Lynch G C, Steckler R, Schwenke D W, Varandas A J C, Truhlar D G and Garrett B C 1991 *J. Chem. Phys.* **94** 7136
- [51] Mielke S L, Lynch G C and Truhlar D G and Schwenke D W 1993 *Chem. Phys. Lett.* **213** 10
- [52] Takayanagi T and Sato S 1988 *Chem. Phys. Lett.* **144** 191
- [53] Nichols J A, Kendall R A and Cole S J and Simons J 1991 *J. Phys. Chem.* **95** 1074
- [54] Manolopoulos D E, Stark K, Werner H J, Arnold D W, Bradforth S E and Neumark D M 1993 *Science* **262** 1852
- [55] Castillo J F, Manolopoulos D E, Stark K and Werner H J 1996 *J. Chem. Phys.* **104** 6531
- [56] Castillo J F, Hartke B, Werner H J, Aoiz F J, Banares L and MartinezHaya B 1998 *J. Chem. Phys.* **109** 7224
- [57] Aoiz F J, Banares L, MartinezHaya B, Castillo J F, Manolopoulos D E, Stark K and Werner H J 1997 *J. Phys. Chem. A* **101** 6403
- [58] Alexander M H, Werner H J and Manolopoulos D E 1998 *J. Chem. Phys.* **109** 5710
- [59] Lee Y T 1987 *Science* **236** 793

A3.8 Molecular reaction dynamics in condensed phases

Gregory A Voth

A3.8.0 INTRODUCTION

The effect of the condensed phase environment on chemical reaction rates has been extensively studied over the past few decades. The central framework for understanding these effects is provided by the transition state theory (TST) [1, 2] developed in the 1930s, the Kramers theory [3] of 1940, the Grote–Hynes [4] and related theories [5] of the 1980s and 1990s and the Yamamoto reactive flux correlation function formalism [6] as extended and further developed by a number of workers [7, 8]. Each of these seminal theoretical breakthroughs has, in turn, generated an enormous amount of research in its own right. There are many good reviews of this body of literature, some of which are cited in [5, 9, 10, 11 and 12]. It therefore serves no useful purpose to review the field again in the present chapter. Instead, the key issues involving condensed phase effects on chemical reactions will be organized around the primary theoretical concepts as they stand at the present time. Even more importantly, the gaps in our understanding and prediction of these effects will be highlighted. From this discussion it will become evident that, despite the large body of theoretical work in this field, there are significant questions that remain unanswered, as well as a need for greater contact between theory and experiment. The discussion here is by no means intended to be exhaustive, nor is the reference list comprehensive.

A3.8.1 THE REACTIVE FLUX

To begin, consider a system which is at equilibrium and undergoing a forward and reverse chemical reaction. For simplicity, one can focus on an isomerization reaction, but the discussion also applies to other forms of unimolecular reactions as well as to bimolecular reactions that are not diffusion limited. The equilibrium of the reaction is characterized by the mole fractions x_R and x_P of reactants and products, respectively, and an equilibrium constant K_{eq} . For gas phase reactions, it is commonplace to introduce the concept of the *minimum energy path* along some reaction coordinate, particularly if one is interested in microcanonical reaction rates. In condensed phase chemical dynamics, however, this concept is not useful. In fact, a search for the minimum energy path in a liquid phase reaction would lead one to the solid state! Instead, one considers a *free energy path* along the reaction coordinate q , and the dominant effect of a condensed phase environment is to change the nature of this path (i.e. its barriers and reactant and product wells, or minima). To illustrate this point, the free energy function along the reaction coordinate of an isomerizing molecule in the gas phase is shown by the full curve in [figure A3.8.1](#). In the condensed phase, the free energy function will almost always be modified by the interaction with the solvent, as shown by the broken curve in [figure A3.8.1](#). (It should be noted that, in the spirit of TST, the definition of the optimal reaction coordinate should probably be redefined for the condensed phase reaction, but for simplicity it can be taken to be the same coordinate as in the gas phase.) As can be seen from [figure A3.8.1](#), the solvent can modify the barrier height for the reaction, the location of the barrier along q , and the reaction free energy (i.e. the difference between the reactant and product minima). It may also introduce dynamical effects that are not apparent from the curve, and it is noted here that a classical framework has been implicitly used—the generalization to the quantum regime will be addressed in a later section.

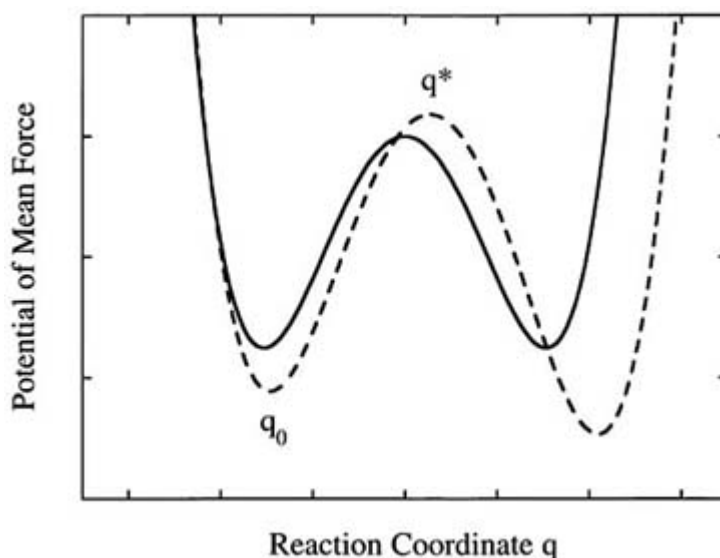


Figure A3.8.1 A schematic diagram of the PMF along the reaction coordinate for an isomerizing solute in the gas phase (full curve) and in solution (broken curve). Note the modification of the barrier height, the well positions, and the reaction free energy due to the interaction with the solvent.

It is worth discussing the fact that a free energy can be directly relevant to the rate of a *dynamical* process such as a chemical reaction. After all, a free energy function generally arises from an ensemble average over configurations. On the other hand, most condensed phase chemical rate constants are indeed thermally averaged quantities, so this fact may not be so surprising after all, although it should be quantified in a rigorous fashion. Interestingly, the free energy curve for a condensed phase chemical reaction (cf figure A3.8.1) can be viewed, in effect, as a natural consequence of Onsager's linear regression hypothesis as it is applied to condensed phase chemical reactions, along with some additional analysis and simplifications [7].

In the spirit of Onsager, if one imagines a relatively *small* perturbation of the populations of reactants and products away from their equilibrium values, then the regression hypothesis states that the decay of these populations back to their equilibrium values will follow the same time-dependent behaviour as the decay of correlations of *spontaneous* fluctuations of the reactant and product populations in the equilibrium system. In the condensed phase, it is this powerful principle that connects a macroscopic dynamical quantity such as a kinetic rate constant with equilibrium quantities such as a free energy function along a reaction pathway and, in turn, the underlying microscopic interactions which determine this free energy function. The effect of the condensed phase environment can therefore be largely understood in the equilibrium, or quasi-equilibrium, context in terms of the modifications of the free energy curve as shown in figure A3.8.1. As will be shown later, the remaining condensed phase effects which are not included in the equilibrium picture may be defined as being 'dynamical'.

The Onsager regression hypothesis, stated mathematically for the chemically reacting system just described, is given in the classical limit by

$$\frac{\Delta N_{\text{R}}(t)}{\Delta N_{\text{R}}(0)} = \frac{\langle \delta N_{\text{R}}(0) \delta N_{\text{R}}(t) \rangle}{\langle \delta N_{\text{R}}(0)^2 \rangle} \quad (\text{A 3.8.1})$$

where $\Delta N_{\text{R}}(t) = \bar{N}_{\text{R}}(t) - \langle N_{\text{R}} \rangle$ is the time-dependent difference between the number of reactant molecules $\bar{N}_{\text{R}}(t)$ arising from an initial non-equilibrium (perturbed) distribution and the final equilibrium number of the reactants $\langle N_{\text{R}} \rangle$. On the right-hand side of the equation, $\langle \delta N_{\text{R}}(t) = N_{\text{R}}(t) - \langle N_{\text{R}} \rangle$ is the instantaneous fluctuation

in the number of reactant molecules away from its equilibrium value in the canonical ensemble, and the notation $\langle \dots \rangle$ denotes the ensemble average over initial conditions.

The solution to the usual macroscopic kinetic rate equations for the reactant and product concentrations yields an expression for the left-hand side of (A3.8.1) that is equal to $\Delta N_R(t) = \Delta N_R(0) \exp(-t/\tau_{rxn})$, where τ_{rxn}^{-1} is the sum of the forward and reverse rate constants, k_f and k_r , respectively. The connection with the microscopic dynamics of the reactant molecule comes about from the right-hand side of (A3.8.1). In particular, in the dilute solute limit, the reactant and product states of the reacting molecule can be identified by the reactant and product population of functions $h_R[q(t)] = 1 - h_P[q(t)]$ and $h_P[q(t)]$, respectively, where $h_P[q(t)] \equiv h[q^* - q(t)]$ and $h(x)$ is the Heaviside step function. The product population function abruptly switches from a value of zero to one as the reaction coordinate trajectory $q(t)$ passes through the barrier maximum at q^* (cf. [Figure A3.8.1](#)). The important connection between the macroscopic (exponential) rate law and the decay of spontaneous fluctuations in the reactant populations, as specified by the function $h_R[q(t)] = 1 - h_P[q(t)]$ and in terms of the microscopic reaction coordinate q , is valid in a ‘coarse-grained’ sense in time, i.e. after a period of molecular-scale transients usually of the order of a few tens of femtoseconds. From the theoretical point of view, the importance of the connection outlined above cannot be overstated because it provides a link between the macroscopic (experimentally observed) kinetic phenomena and the molecular scale dynamics of the reaction coordinate in the equilibrium ensemble.

However, further analysis of the linear regression expression in A3.8.1 is required to achieve a useful expression for the rate constant both from a computational and a conceptual points of view. Such an expression was first provided by Yamamoto [6], but others have extended, validated, and expounded upon his analysis in considerable detail [7, 8]. The work of Chandler [7] in this regard is followed most closely here in order to demonstrate the places in which condensed phase effects can appear in the theory, and hence in the value of the thermal rate constant. The key mathematical step is to differentiate both sides of the linear regression formula in (A3.8.1) and then carefully analyse its expected behaviour for systems having a barrier height of at least several times $k_B T$. The resulting expression for the classical forward rate constant in terms of the so-called ‘reactive flux’ time correlation function is given by [6, 7 and 8]

$$\begin{aligned} k_f &= x_R^{-1} \langle \dot{h}_P[q(0)] h_P[q(t_{pl})] \rangle \\ &= x_R^{-1} \langle \dot{q}(0) \delta[q^* - q(0)] h_P[q(t_{pl})] \rangle \end{aligned} \quad (\text{A 3.8.2})$$

where x_R is the equilibrium mole fraction of the reactant. The classical rate constant is obtained from (A3.8.2) when the correlation function reaches a ‘plateau’ value at the time $t = t_{pl}$ after the molecular-scale transients have ended [7]. Upon inspection of the above expression, it becomes apparent that the classical rate constant can be calculated by averaging over trajectories initiated at the barrier top with a velocity Boltzmann distribution for the reaction coordinate and an equilibrium distribution in all other degrees of freedom of the system. Those trajectories are then weighted by their initial velocity and the initial flux over the barrier is correlated with the product state population function $h_P[q(t)]$. The time dependence of the correlation function is computed until the plateau value is reached, at which point it

becomes essentially constant and the numerical value of the thermal rate constant can be evaluated. An example of such a correlation function obtained through molecular dynamics simulations is shown in figure A3.8.2.

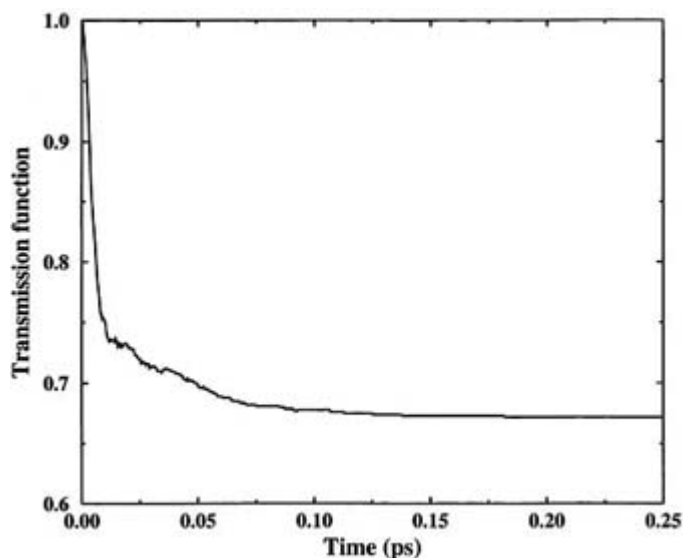


Figure A3.8.2 The correlation function $\kappa(t)$ for particular case of the reaction of methyl vinyl ketone with cyclopentadiene in water. The leveling-off of this function to reach a constant value at the plateau time t_{pl} is clearly seen.

It is important to recognize that the time-dependent behaviour of the correlation function during the molecular transient time seen in figure A3.8.2 has an important origin [7, 8]. This behaviour is due to trajectories that recross the transition state and, hence, it can be proven [7] that the classical TST approximation to the rate constant is obtained from A3.8.2 in the $t \rightarrow 0^+$ limit:

$$\begin{aligned} k_f^{\text{TST}} &= x_R^{-1} \lim_{t \rightarrow 0^+} \langle \dot{q}(0) \delta[q^* - q(0)] h_P[q(t)] \rangle \\ &= x_R^{-1} \langle h[\dot{q}(0)] \dot{q}(0) \delta[q^* - q(0)] \rangle. \end{aligned} \quad (\text{A } 3.8.3)$$

It is, of course, widely considered that the classical TST provides the central framework for the understanding of thermal rate constants (see the review article by Truhlar *et al* [13]) and also for quantifying the dominant effects of the considered phase in chemical reactions (see below).

In order to segregate the theoretical issues of condensed phase effects in chemical reaction dynamics, it is useful to rewrite the exact classical rate constant in (A3.8.2) as [5, 6, 7, 8, 9, 10 and 11]

$$k_f = \kappa k_f^{\text{TST}} \quad (\text{A } 3.8.4)$$

where κ is the dynamical correction factor (or ‘transmission coefficient’) which is given by

$$\begin{aligned} \kappa &= \frac{\langle \dot{q}(0) \delta[q^* - q(0)] h_P[q(t_{pl})] \rangle}{\langle h[\dot{q}(0)] \dot{q}(0) \delta[q^* - q(0)] \rangle} \\ &= \langle h_P[q(t_{pl})] \rangle_+ - \langle h_P[q(t_{pl})] \rangle_-. \end{aligned} \quad (\text{A } 3.8.5)$$

Here, the symbol $\langle \dots \rangle_{\pm}$ denotes an averaging over the flux-weighted distribution [7, 8] for positive or negative initial velocities of the reaction coordinate. In figure A3.8.2 is shown the correlation function $\kappa(t)$ for the particular case of the reaction of methyl vinyl ketone with cyclopentadiene in water. The leveling-off of this

function to reach a constant value at the plateau time t_{pl} is clearly seen. The effect of the condensed phase environment in a thermal rate constant thus appears *both* in the value of the TST rate constant and in the value of the dynamical correction factor in (A3.8.4). These effects will be described separately in the following sections, but it should be noted that the two quantities are not independent of each other in that they both depend on the choice of the reaction coordinate q . The ‘variational’ choice of q amounts to finding a definition of that coordinate that causes the value of κ to be as close to unity as possible, i.e. to minimize the number of recrossing trajectories. It seems clear that an important area of research for the future will be to define theoretically the ‘best’ reaction coordinate in a condensed phase chemical reaction—one in which the solvent is explicitly taken into account. In charge transfer reactions, for example, a collective solvent polarization coordinate can be treated as being coupled to a solute coordinate (see, for example, 14), but a more detailed and rigorous microscopic treatment of the full solution—phase reaction coordinate is clearly desirable for the future (see, for example, 15 for progress in this regard). Before describing the effects of a solvent on the thermal rate constants, it is worthwhile to first reconsider the above analysis in light of current experimental work on condensed phase dynamics and chemical reactions. The formalism outlined above, while exceptionally powerful in that it provides a link between microscopic dynamics and macroscopic chemical kinetics, is intended to help us calculate and analyse *only* thermal rate constants in *equilibrium* systems. The linear regression hypothesis provides the key line of analysis for this problem. To the extent that the thermal rate constant is the quantity of interest—and many times it *is* the primary quantity of interest—this theoretical approach would appear to be the best. However, in many experiments, for example nonlinear optical experiments involving intense laser pulses and/or photoinitiated chemical reactions, the system may initially be far from equilibrium and the above theoretical analysis may not be completely applicable. Furthermore, experimentally measured quantities such as vibrational or phase relaxation rates are often only indirectly related to the thermal rate constant. It would therefore appear that more theoretical effort will be required in the future to relate experimental measurements to the particular microscopic dynamics in the liquid phase that influence the outcome of such measurements and, in turn, to the more standard quantities such as the thermal rate constant.

A3.8.2 THE ACTIVATION FREE ENERGY AND CONDENSED PHASE EFFECTS

Having separated the dynamical from equilibrium (or, more accurately, quasi-equilibrium) effects, one can readily discover the origin of the activation free energy and define the concept of the potential of mean force by analysis of the expression for the TST rate constant, k_f^{TST} in (A3.8.3). The latter can be written as [7]

$$k_f^{\text{TST}} = \frac{(2\pi m\beta)^{-1/2}}{\int_{-\infty}^{q^*} dq \exp[-\beta V_{\text{eq}}(q)]} \exp[-\beta V_{\text{eq}}(q^*)] \quad (\text{A 3.8.6})$$

-6-

where $\beta = 1/k_B T$ and $V_{\text{eq}}(q)$ is the *potential of mean force* (PMF) along the reaction coordinate q . The latter quantity is all important for quantifying and understanding the effect of the condensed phase on the value of the thermal rate constant. It is defined as

$$V_{\text{eq}}(q) = -k_B T \ln \left[\int dq' d\mathbf{x} \delta(q - q') \exp[-\beta V'(q', \mathbf{x})] \right] + \text{constant} \quad (\text{A 3.8.7})$$

where \mathbf{x} are all coordinates of the condensed phase system other than the reaction coordinate, and $V(q, \mathbf{x})$ is the total potential energy function. The additive constant in (A3.8.7) is irrelevant to the value of the thermal rate constant in (A3.8.6). If the PMF around its minimum in the reactant state (cf. Figure A3.8.1) is expanded

quadratically, i.e. $V_{\text{eq}}(q) \approx V_{\text{eq}}(q_0) + (1/2)m\omega_0^2(q - q_0)^2$, then (A3.8.6) simplifies to [5, 7]

$$k_f^{\text{TST}} = \frac{\omega_0}{2\pi} \exp(-\beta \Delta F_{\text{cl}}^*) \quad (\text{A 3.8.8})$$

where the *activation free energy* of the system is defined as $\Delta F_{\text{cl}}^* = V_{\text{eq}}(q^*) - V_{\text{eq}}(q_0)$. The PMF is often decomposed as $V_{\text{eq}}(q) = v(q) + W_{\text{eq}}(q)$, where $v(q)$ is the intrinsic contribution to the PMF from the solute potential energy function and, therefore, by definition, $W_{\text{eq}}(q)$ is the contribution arising from the solute–solvent coupling. Figure A3.8.1 illustrates how the latter coupling is responsible for the condensed phase-induced change in the activation free energy, the reaction free energy, and the position of the reactant and product wells. Thus, within the context of the TST, one can conclude that the condensed phase enters into the picture in a ‘simple’ way through the aforementioned modifications of the reaction coordinate free energy profile in figure A3.8.1.

In principle, nothing more is necessary to understand the influence of the solvent on the TST rate constant than the modification of the PMF, and the resulting changes in the free energy barrier height should be viewed as the dominant effect on the rate since these changes appear in an exponential form. As an example, an error in calculating the solvent contribution to the barrier of 1 kcal mol^{-1} will translate into an error of a factor of four in the rate constant—a factor which is often larger than any dynamical and/or quantum effects such as those described in later sections. This is a compelling fact for the theorist, so it is therefore no accident that the accurate calculation of the solvent contribution to the activation free energy has become the primary focus of many theoretical and computational chemists. The successful completion of such an effort requires four things: (1) an accurate representation of the solute potential, usually from highly demanding *ab initio* electronic structure calculations; (2) an accurate representation of both the solvent potential and the solute–solvent coupling; (3) an accurate computational method to compute, with good statistics, the activation free energy in condensed phase systems; and (4) improved theoretical techniques, both analytical and computational, to identify the *microscopic origin* of the dominant contributions to the activation free energy and the relationship of these effects to experimental parameters such as pressure, temperature, solvent viscosity and polarity, etc. Each of these areas has in turn generated a significant number of theoretical papers over the past few decades—too many in fact to fairly cite them here—and many of these efforts have been major steps forward. There seems to be little dispute, however, that much work remains to be done in all of these areas. Indeed, one of the computational ‘grand challenges’ facing theoretical chemistry over the coming decades will surely be the *quantitative prediction* (better than a factor of two) of chemical reaction rates in highly complex systems. Some of this effort may, in fact, be driven by the needs of industry and government in, for example, the environmental fate prediction of pollutants.

A3.8.3 THE DYNAMICAL CORRECTION AND SOLVENT EFFECTS

While the TST estimate of the thermal rate constant is usually a good approximation to the true rate constant and contains most of the dominant solvent effects, the dynamical corrections to the rate can be important as well. In the classical limit, these corrections are responsible for a value of the dynamical correction factor κ in A3.8.4 that drops below unity. A considerable theoretical effort has been underway over the past 50 years to develop a general theory for the dynamical correction factor (see, for example, [5, 6, 7, 8, 9, 10, 11 and 12]). One approach to the problem is a direct calculation of κ using molecular dynamics simulation and the reactive flux correlation function formalism [7, 8, 16]. This approach obviously requires the numerically exact integration of Newton’s equations for the many-body potential energy surface and a good microscopic model of the condensed phase interactions. However, another approach [5, 9, 10, 11 and 12] has been to employ a *model* for the reaction coordinate dynamics around the barrier top, for example, the generalized Langevin

equation (GLE) given by

$$m\ddot{q}(t) = -\frac{dV_{\text{eq}}(q)}{dq} - \int_0^t dt' \eta(t-t'; q^*)\dot{q}(t') + \delta F(t). \quad (\text{A 3.8.9})$$

In this equation, m is the effective mass of the reaction coordinate, $\eta(t-t'; q^*)$ is the friction kernel calculated with the reaction coordinate ‘clamped’ at the barrier top, and $\delta F(t)$ is the fluctuating force from all other degrees of freedom with the reaction coordinate so configured. The friction kernel and force fluctuations are related by the fluctuation–dissipation relation

$$\eta(t; q^*) = \beta \langle \delta F(0)\delta F(t) \rangle_{q^*}. \quad (\text{A 3.8.10})$$

In the limit of a very rapidly fluctuating force, the above equation can sometimes be approximated by the simpler Langevin equation

$$m\ddot{q}(t) = -\frac{dV_{\text{eq}}(q)}{dq} - \hat{\eta}(0)\dot{q}(t) + \delta F(t) \quad (\text{A 3.8.11})$$

where $\hat{\eta}(0)$ is the so-called ‘static’ friction, $\hat{\eta}(0) = \int_0^\infty dt \eta(t; q^*)$

The GLE can be derived by invoking the linear response approximation for the response of the solvent modes coupled to the motion of the reaction coordinate.

It should be noted that the friction kernel is not in general independent of the reaction coordinate motion [17], i.e. a nonlinear response, so the GLE may have a limited range of validity [18, 19 and 20]. Furthermore, even if the equation is valid, the strength of the friction might be so great that the second and third terms on the right-hand side of (A3.8.9) could dominate the dynamics much more so than the force generated by the PMF. It should also be noted that, even though the friction in (A3.8.9) may be adequately approximated to be dynamically independent of the value of the reaction coordinate, the equation is still in general nonlinear, depending on the nature of the PMF. For non-quadratic

forms of the PMF, $V_{\text{eq}}(q)$, even the solution of the reactive dynamics from the model perspective of the GLE becomes a non-trivial problem.

Two central results have arisen from the GLE-based perspective on the dynamical correction factor. The first is the Kramers theory of 1940 [3], based on the simpler Langevin equation, while the second is the Grote–Hynes theory of 1980 [4]. Both have been extensively discussed and reviewed in the literature [5, 9, 10, 11 and 12]. The important insight of the Kramers theory is that the transmission coefficient for an isomerization or metastable escape reaction undergoes a ‘turnover’ as one increases the static friction from zero to large values. For weak damping (friction), the transmission coefficient is proportional to the friction, i.e. $\kappa \propto \hat{\eta}(0)$. This dependence arises because the barrier recrossings are caused by the slow energy diffusion (equilibration) in the reaction coordinate motion as it leaves the barrier region. For strong damping, on the other hand, the transmission coefficient is inversely proportional to the friction, i.e. $\kappa \propto 1/\hat{\eta}(0)$, because the barrier crossings are caused by the diffusive spatial motion of the reaction coordinate in the barrier region. For systems such as atom exchange reactions that do not involve a bound reactant state, only the spatial diffusion regime is predicted. The basic phenomenology of condensed phase activated rate processes, as mapped out by Kramers, captures the essential physics of the problem and remains the seminal work to this day.

The second key insight into the dynamical corrections to the TST was provided by the Grote–Hynes theory [4]. This theory highlights the importance of the time dependence of the friction and demonstrates how it may be taken into account at the leading order. In the overdamped regime this is done so through the insightful and compact Grote–Hynes (subscript GH) formula for the transmission coefficient [4].

$$\kappa_{\text{GH}} = \frac{\lambda_0^\ddagger}{\omega_{b,\text{eq}}} \quad \lambda_0^\ddagger = \frac{\omega_{b,\text{eq}}^2}{\lambda_0^\ddagger + \hat{\eta}(\lambda_0^\ddagger)/m} \quad (\text{A 3.8.12})$$

where $\hat{\eta}(z)$ is the Laplace transform of the friction kernel, i.e. $\hat{\eta}(z) = \int_0^\infty dt e^{-zt} \eta(t)$, and $\omega_{b,\text{eq}}$ is the magnitude of the unstable PMF barrier frequency. Importantly, the derivation of this formula assumes a quadratic approximation to the barrier $V_{\text{eq}}(q) \approx V_{\text{eq}}(q^*) - (1/2)m\omega_{b,\text{eq}}^2 (q - q^*)^2$ that may not always be a good one.

Research over the past decade has demonstrated that a multidimensional TST approach can also be used to calculate an even more accurate transmission coefficient than κ_{GH} for systems that can be described by the full GLE with a non-quadratic PMF. This approach has allowed for variational TST improvements [21] of the Grote–Hynes theory in cases where the nonlinearity of the PMF is important and/or for systems which have general nonlinear couplings between the reaction coordinate and the bath force fluctuations. The Kramers turnover problem has also been successfully treated within the context of the GLE and the multidimensional TST picture [22]. A multidimensional TST approach has even been applied [15] to a realistic model of an $S_{\text{N}}2$ reaction and may prove to be a promising way to elaborate the explicit microscopic origins of solvent friction. While there has been great progress toward an understanding and quantification of the dynamical corrections to the TST rate constant in the condensed phase, there are several quite significant issues that remain largely open at the present time. For example, even if the GLE were a valid model for calculating the dynamical corrections, it remains unclear how an accurate and predictive microscopic theory can be developed for the friction kernel $\eta(t)$ so that one does not have to resort to a molecular dynamics simulation [17] to calculate this quantity. Indeed, if one could compute the solvent friction along the reaction coordinate in such a manner, one could instead just calculate the exact rate

-9-

constant using the reactive-flux formalism. A microscopic theory for the friction is therefore needed to relate the friction along the reaction coordinate to the parameters varied by experimentalists such as pressure or solvent viscosity. No complete test of Kramers theory will ever be possible until such a theoretical effort is completed. Two possible candidates in the latter vein are the instantaneous normal mode theory of liquids [23] and the damped normal mode theory [24] for liquid state dynamics.

Another key issue remaining to be resolved is whether a one-dimensional GLE as in [A3.8.11](#) is the optimal choice of a dynamical model in the case of strong damping, or whether a two- or multi-dimensional GLE that explicitly includes coupling to solvation and/or intramolecular modes is more accurate and/or more insightful. Such an approach might, for example, allow better contact with nonlinear optical experiments that could measure the dynamics of such additional modes. It is also entirely possible that the GLE may not even be a good approximation to the true dynamics in many cases because, for example, the friction strongly depends on the position of the reaction coordinate. In fact, a strong solvent modification of the PMF usually ensures that the friction will be spatially dependent [25]. Several analytical studies have dealt with this issue (see, for example, [26, 27 and 28] and literature cited therein). Spatially-dependent friction is found to have an important effect on the dynamical correction in some instances, but in others the Grote–Hynes estimate is predicted to be robust [29]. Nevertheless, the question of the nonlinearity and the accurate modelling of real activated rate processes by the GLE remains an open one.

Another important issue has been identified by several authors [30, 31] which involves the participation of

intramolecular solute modes in defining the range of the energy diffusion-limited regime of condensed phase activated dynamics. In particular, if the coupling between the reaction coordinate and such modes is strong, then the Kramers turnover behaviour as a function of the solvent friction occurs at a significantly lower value of the friction than for the simple case of the reaction coordinate coupled to the solvent bath alone. In fact, the issue of whether the turnover can be experimentally observed at all in the condensed phase hinges on this issue. To date, it has remained a challenge to calculate the effective number of intramolecular modes that are strongly coupled to the reaction coordinate; no general theory yet exists to accomplish this important goal.

As a final point, it should again be emphasized that many of the quantities that are measured experimentally, such as relaxation rates, coherences and time-dependent spectral features, are complementary to the thermal rate constant. Their information content in terms of the underlying microscopic interactions may only be indirectly related to the value of the rate constant. A better theoretical link is clearly needed between experimentally measured properties and the common set of microscopic interactions, if any, that also affect the more traditional solution phase chemical kinetics.

A3.8.4 QUANTUM ACTIVATED RATE PROCESSES AND SOLVENT EFFECTS

The discussion thus far in this chapter has been centred on classical mechanics. However, in many systems, an explicit quantum treatment is required (not to mention the fact that it is the correct law of physics). This statement is particularly true for proton and electron transfer reactions in chemistry, as well as for reactions involving high-frequency vibrations.

The exact quantum expression for the activated rate constant was first derived by Yamamoto [6]. The resulting quantum reactive flux correlation function expression is given by

-10-

$$k_f = \frac{1}{x_R \hbar \beta} \int_0^{\hbar \beta} d\tau \langle \dot{h}_P(-i\tau) h_P(t_{pl}) \rangle \quad (\text{A 3.8.13})$$

where $h_P(t)$ is the Heisenberg product state population operator. As opposed to the classical case, however, the $t \rightarrow 0^+$ limit of this expression is always equal to zero [32] which ensures that an *entirely different* approach from the classical analysis must be adopted in order to formulate a quantum TST (QTST), as well as a theory for its dynamical corrections. An article by Truhlar *et al* [13] describes many of the efforts over the past 60 years to develop quantum versions of the TST, and many, if not most, of these efforts have been applicable to primarily low-dimensional gas phase systems. A QTST that is useful for condensed phase reactions is an extremely important theoretical goal since a direct numerical attack on the time-dependent Schrödinger equation for many-body systems is computationally prohibitive, if not impossible. (The latter fact seems to be true in the fundamental sense, i.e. there is an exponential scaling of the numerical effort with system size for the exact solution.) In this section, some of the leading candidates for a viable condensed phase QTST will now be briefly described. The discussion should by no means be considered complete.

As a result of several complementary theoretical efforts, primarily the path integral centroid perspective [33, 34 and 35], the periodic orbit [36] or instanton [37] approach and the ‘above crossover’ quantum activated rate theory [38], one possible candidate for a unifying perspective on QTST has emerged [39] from the ideas from [39, 40, 41 and 42]. In this theory, the QTST expression for the forward rate constant is expressed as [39]

$$k_f \approx \nu \frac{\text{Im } Q_b}{Q_R} \quad (\text{A 3.8.14})$$

where ν is a simple frequency factor, Q_R is the reactant partition function, and Q_b is the barrier ‘partition function’ which is to be interpreted in the appropriate asymptotic limit [39, 40, 41 and 42]. The frequency factor has the piecewise continuous form [39]

$$\nu = \begin{cases} \lambda_0^\ddagger/2\pi & \hbar\beta\lambda_0^\ddagger < 2\pi \\ (\hbar\beta)^{-1} & \hbar\beta\lambda_0^\ddagger \geq 2\pi \end{cases} \quad (\text{A 3.8.15})$$

while the barrier partition function is defined under most conditions as 39

$$Q_b = \int_{q_c \rightarrow i q_c} dq_c \rho_c(q_c). \quad (\text{A 3.8.16})$$

The quantity $\rho_c(q_c)$ is the Feynman path integral centroid density [43] that is understood to be expressed asymptotically as

$$\rho_c(q_c) \approx \rho_c(q^*) \exp[-\beta V_c''(q^*)(q_c - q^*)^2/2] \quad (\text{A 3.8.17})$$

where the quantum centroid potential of mean force is given by $V_c(q_c) = -k_B T \ln[\rho_c(q_c)] + \text{constant}$ and q^* is *defined*

-11-

to be the value of the reaction coordinate that gives the maximum value of $V_c(q)$ in the barrier region (i.e. it may differ [33, 35] from the maximum of the classical PMF along q). The path integral centroid density along the reaction coordinate is given by the Feynman path integral expression

$$\rho_c(q_c) = \int \cdots \int Dq(\tau) D\mathbf{x}(\tau) \delta(q_c - \tilde{q}_0) \exp\{-S[q(\tau), \mathbf{x}(\tau)]/\hbar\} \quad (\text{A 3.8.18})$$

which is a functional integral over all possible cyclic paths of the system coordinates weighted by the imaginary time action function [43]:

$$S[q(\tau), \mathbf{x}(\tau)] = \int_0^{\hbar\beta} d\tau \left\{ \frac{m}{2} \dot{q}(\tau)^2 + \sum_{i=1}^N \frac{m_i}{2} \dot{x}_i(\tau)^2 + V[q(\tau), \mathbf{x}(\tau)] \right\}. \quad (\text{A 3.8.19})$$

The key feature of A3.8.18 is that the centroids of the reaction coordinate Feynman paths are constrained to be at the position q_c . The centroid \tilde{q}_0 of a particular reaction coordinate path $q(\tau)$ is given by the zero-frequency Fourier mode, i.e.

$$\tilde{q}_0 = \frac{1}{\hbar\beta} \int_0^{\hbar\beta} d\tau q(\tau) \quad (\text{A 3.8.20})$$

Under most conditions, the sign of $V_c''(q^*)$ in (A3.8.17) is negative. In such cases, the centroid variable *naturally* appears in the theory 39, and the equation for the quantum thermal rate constant from (A3.8.14) – (A3.8.17) is then given by [39]

$$k_{\text{f}} \approx v \frac{\sqrt{2\pi/\beta|V_{\text{c}}''(q^*)|}}{Q_{\text{R}}} \exp[-\beta V_{\text{c}}(q^*)]. \quad (\text{A } 3.8.21)$$

It should be noted that in the cases where $V_{\text{c}}''(q^*) > 0$, the centroid variable becomes irrelevant to the quantum activated dynamics as defined by (A3.8.14) and the instanton approach [37] to evaluate Q_{b} based on the steepest descent approximation to the path integral becomes the approach one may take. Alternatively, one may seek a more generalized saddle point coordinate about which to evaluate A3.8.14. This approach has also been used to provide a unified solution for the thermal rate constant in systems influenced by non-adiabatic effects, i.e. to bridge the adiabatic and non-adiabatic (Golden Rule) limits of such reactions.

In the limit of reasonably high temperatures (above the so-called ‘crossover’ temperature), i.e. $\hbar\beta\lambda_{\text{0}}^{\ddagger} < 2\pi$, the above formula in A3.8.21 is best simplified further and approximately written as

$$k_{\text{f}} \approx k_{\text{GH}} \frac{(2\pi m\beta)^{-1/2}}{\int_{-\infty}^{q^*} dq_{\text{c}} \exp[-\beta V_{\text{c}}(q_{\text{c}})]} \exp[-\beta V_{\text{c}}(q^*)]. \quad (\text{A } 3.8.22)$$

-12-

This formula, aside from the prefactor κ_{GH} , is often referred to as the path integral quantum transition state theory (PI–QTST) formula 33. One clear strength of this formula is its clear analogy with and generalization of the classical TST formula in A3.8.6. In turn, this allows for an interpretation of solvent effects on quantum activated rate constants in terms of the quantum centroid potential of mean force in a fashion analogous to the classical case. The quantum activation free energy for highly-non-trivial systems can also be directly calculated with imaginary time path integral Monte Carlo techniques 44. Many such studies have now been carried out, but a single example will be described in the following section.

The preceding discussion has focused on the path integral centroid picture of condensed phase quantum activated dynamics, primarily because of its strong analogy with the classical case, the PMF, etc, as well as its computational utility for realistic problems. However, several recent complementary developments must be mentioned. The first is due to Pollak, Liao and Shao 45 who have significantly extended an earlier idea 30 in which the exact Heisenberg population operator in $h_{\text{p}}(t)$ in A3.8.13 is replaced by one for a parabolic barrier (plus some other important manipulations, such as symmetrization of the flux operator, that were not done in 30). The dynamical population operator then has an analytic form which in turn leads one to a purely analytic ‘quantum transition state theory’ approximation to A3.8.13. This approach, which in principle can be systematically improved upon through perturbation theory, has been demonstrated to be as accurate as the path integral centroid-based formulae in A3.8.21 and A3.8.22 above the crossover temperature.

A second recent development has been the application 46 of the initial value representation 47 to semiclassically calculate A3.8.13 (and/or the equivalent time integral of the ‘flux–flux’ correlation function). While this approach has to date only been applied to problems with simplified harmonic baths, it shows considerable promise for applications to realistic systems, particularly those in which the real solvent ‘bath’ may be adequately treated by a further classical or quasiclassical approximation.

A3.8.5 SOLVENT EFFECTS IN QUANTUM CHARGE TRANSFER PROCESSES

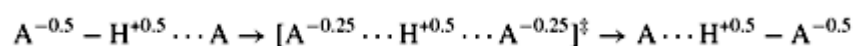
In this section, the results of a computational study 48 will be used to illustrate the effects of the solvent—and the significant *complexity* of these effects—in quantum charge transfer processes. The particular example

described here is for a ‘simple’ modelistic proton transfer reaction in a polar solvent. This study, while useful in its own right, also illustrates the level of detail and theoretical formalism that is likely to be necessary in the future to accurately study solvent effects in condensed phase charge transfer reactions, even at the equilibrium (quantum PMF) level.

Some obvious targets for quantum activated rate studies are proton, hydride, and hydrogen transfer reactions because they are of central importance in the solution phase and acid–base chemistry, as well as in biochemistry. These reactions are particularly interesting because they can involve large quantum mechanical effects and, since there is usually a redistribution of solute electronic charge density during the reaction, a substantial contribution to the activation free energy may have its origin from the solvent reorganization process. It is thought that intramolecular vibrations may also play a crucial role in modulating the reactive process by lowering the intrinsic barrier for the reaction.

Many of the condensed phase effects mentioned above have been studied computationally using the PI–QTST approach outlined in the first part of the last section. One such study [48](#) has focused on the model symmetric

-13-



three-body proton transfer reaction in a polar fluid with its dipole moment chosen to model methanol. The molecular group ‘A’ represents a generic proton donor/acceptor group.

After some straightforward manipulations of [A3.8.22](#), the PI–QTST estimate of the proton transfer rate constant can be shown to be given by [48](#)

$$k_f^{\text{PI-QTST}} = \frac{\omega_{c,0}}{2\pi} \exp(-\beta \Delta F_c^*) \quad (\text{A 3.8.23})$$

where $\omega_{c,0} = [V''_c(q_0)/m]^{1/2}$ and the quantum activation free energy is given by [48](#)

$$\begin{aligned} \Delta F_c^* &= -k_B T \ln[\rho_c(q^*)/\rho_c(q_0)] \\ &= -k_B T \ln[P_c(q_0 \rightarrow q^*)]. \end{aligned} \quad (\text{A 3.8.24})$$

The probability $P_c(q_0 \rightarrow q^*)$ to move the reaction coordinate centroid variable from the reactant configuration to the transition state is calculated [48](#) by path integral Monte Carlo techniques [44](#) combined with umbrella sampling [[48](#), [49](#)]. From the calculations on the model proton transfer system above, the quantum activation free energy curves are shown in [figure A3.8.3](#) for both a rigid and non-rigid (vibrating) intra-complex A–A (donor/acceptor) distance. Shown are both the activation curves for the complex in isolation and in the solvent. The effect of the solvent in the total activation free energy is immediately obvious, contributing 2–4 kcal mol⁻¹ to its overall value. One effect of the A–A distance fluctuations is a lowering of the quantum activation free energy (i.e. increased tunnelling) both when the solvent is present and when it is not. A second interesting effect becomes evident from a comparison of the curves for the systems with the rigid versus flexible A–A distance. The contribution to the quantum activation free energy from the solvent is reduced when the A–A distance can fluctuate, resulting in a rate that is 20 times higher than in the rigid case. This novel behaviour was found to arise from a *nonlinear* coupling between the intra-complex fluctuations and the solvent activation, resulting in a reduced dipole moment of the solute when there is an inward fluctuation of the A–A distance.

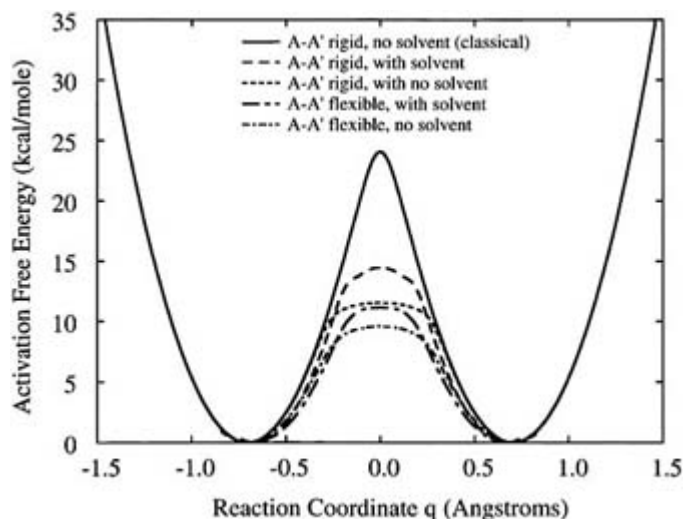


Figure A3.8.3 Quantum activation free energy curves calculated for the model A–H–A proton transfer reaction described 45. The full line is for the classical limit of the proton transfer solute in isolation, while the other curves are for different fully quantized cases. The rigid curves were calculated by keeping the A–A distance fixed. An important feature here is the direct effect of the solvent activation process on both the solvated rigid and flexible solute curves. Another feature is the effect of a fluctuating A–A distance which both lowers the activation free energy and reduces the influence of the solvent. The latter feature enhances the rate by a factor of 20 over the rigid case.

From the above PI–QTST studies, it was found that, in order to fully quantify the solvent effects for even a ‘simple’ model proton transfer reaction, one must deal with a number of complex, nonlinear interactions. Examples of other such interactions include the nonlinear dependence of the solute dipole on the position of the proton and the intrinsically nonlinear interactions arising from both solute and solvent polarizability effects 48. In the latter context, it was found that the solvent electronic polarizability modes *must* be treated quantum mechanically when studying their influence on the proton transfer activation free energy 48. (In general, the adequate treatment of electronic polarizability in a variety of condensed phase contexts is emerging as an extremely important problem in many contexts; condensed phase reactions may never be properly described until this problem is addressed.) The detailed calculations described above, while only for a model proton transfer system, clearly illustrate the significant challenge that lies ahead for those who hope to *quantitatively* predict the rates of computation phase chemical reactions through computer simulation.

A3.8.6 CONCLUDING REMARKS

In this chapter many of the basic elements of condensed phase chemical reactions have been outlined. Clearly, the material presented here represents just an overview of the most important features of the problem. There is an extensive literature on all of the issues described herein and, more importantly, there is still much work to be done before a complete understanding of the effects of condensed phase environments on chemical reactions can be achieved. The theorist and experimentalist alike can therefore look forward to many more years of exciting and challenging research in this important area of physical chemistry.

REFERENCES

- [1] Eyring H 1934 The activated complex in chemical reactions *J. Chem. Phys.* **3** 107
- [2] Wigner E 1937 Calculation of the rate of elementary associated reactions *J. Chem. Phys.* **5** 720
- [3] Kramers H A 1940 Brownian motion in field of force and diffusion model of chemical reactions *Physica* **7** 284
- [4] Grote R F and Hynes J T 1980 The stable states picture of chemical reactions. II. Rate constants for condensed and gas phase reaction models *J. Chem. Phys.* **73** 2715
Grote R F and Hynes J T 1981 Reactive modes in condensed phase reactions *J. Chem. Phys.* **74** 4465
- [5] Hänggi P, Talkner P and Borkovec M 1990 Reaction-rate theory: fifty years after Kramers *Rev. Mod. Phys.* **62** 251
- [6] Yamamoto T 1960 Quantum statistical mechanical theory of the rate of exchange chemical reactions in the gas phase *J. Chem. Phys.* **33** 281
- [7] Chandler D 1978 Statistical mechanics of isomerization dynamics in liquids and the transition state approximation *J. Chem. Phys.* **68** 2959
Montgomery J A Jr, Chandler D and Berne B J 1979 Trajectory analysis of a kinetic theory for isomerization dynamics in condensed phases *J. Chem. Phys.* **70** 4056
Rosenberg R O, Berne B J and Chandler D 1980 Isomerization dynamics in liquids by molecular dynamics *Chem. Phys. Lett.* **75** 162
- [8] Keck J 1960 Variational theory of chemical reaction rates applied to three-body recombinations *J. Chem. Phys.* **32** 1035
Anderson J B 1973 Statistical theories of chemical reactions. Distributions in the transition region *J. Chem. Phys.* **58** 4684
Bennett C H 1977 Molecular dynamics and transition state theory: the simulation of infrequent events *Algorithms for Chemical Computation (ACS Symposium Series No 46)* ed R E Christofferson (Washington, DC: American Chemical Society)
Hynes J T 1985 The theory of reactions in solution *The Theory of Chemical Reaction Dynamics* vol IV, ed M Baer (Boca Raton, FL: CRC Press)
Berne B J 1985 Molecular dynamics and Monte Carlo simulations of rare events *Multiple Timescales* ed J V Brackbill and B I Cohen (New York: Academic Press)
- [9] For reviews of theoretical work on the corrections to classical TST, see
Hynes J T 1985 *Ann. Rev. Phys. Chem.* **36** 573
Berne B J, Borkovec M and Straub J E 1988 Classical and modern methods in reaction rate theory *J. Phys. Chem.* **92** 3711
Nitzan A 1988 Activated rate processes in condensed phases: the Kramers theory revisited *Adv. Chem. Phys.* **70** 489
Onuchic J N and Wolynes P G 1988 Classical and quantum pictures of reaction dynamics in condensed matter: resonances, dephasing and all that *J. Phys. Chem.* **92** 6495
- [10] Fleming G and Hänggi P (eds) 1993 *Activated Barrier Crossing* (New Jersey: World Scientific)
- [11] Talkner P and Hänggi P (eds) 1995 *New Trends in Kramers' Reaction Rate Theory* (Dordrecht: Kluwer)
- [12] Warshel A 1991 *Computer Modeling of Chemical Reactions in Enzymes and Solutions* (New York: Wiley)
- [13] Truhlar D G, Garrett B C and Klippenstein S J 1996 Current status of transition-state theory *J. Phys. Chem.* **100** 12 771
- [14] van der Zwan G and Hynes J T 1982 Dynamical polar solvent effects on solution reactions: a simple continuum model *J. Chem. Phys.* **76** 2993
van der Zwan G and Hynes J T 1983 Nonequilibrium solvation dynamics in solution reactions *J. Chem. Phys.* **78** 4174
van der Zwan G and Hynes J T 1984 A simple dipole isomerization model for non-equilibrium solvation dynamics in reaction in polar solvents *Chem. Phys.* **90** 21
- [15] Pollak E 1993 Variational transition state theory for dissipative systems *Activated Barrier Crossing* ed G Fleming and P Hänggi (New Jersey: World Scientific) p 5
Gershinsky G and Pollak E 1995 Variational transition state theory: application to a symmetric exchange in water *J. Chem. Phys.* **103** 8501

- [16] Whitnell R M and Wilson K R 1993 *Reviews of Computational Chemistry* ed K B Lipkowitz and D B Boyd (New York: VCH)
- [17] Straub J E, Borkovec M and Berne B J 1987 On the calculation of dynamical friction on intramolecular degrees of freedom *J. Phys. Chem.* **91** 4995
Straub J E, Borkovec M and Berne B J 1988 Molecular dynamics study of an isomerizing diatomic Lennard-Jones fluid *J. Chem. Phys.* **89** 4833
Straub J E, Berne B J and Roux B 1990 Spatial dependence of time-dependent friction for pair diffusion in a simple fluid *J. Chem. Phys.* **93** 6804
- [18] Singh S, Krishnan R and Robinson G W 1990 Theory of activated rate processes with space-dependent friction *Chem.*

- [19] Straus J B and Voth G A 1992 Studies on the influence of nonlinearity in classical activated rate processes *J. Chem. Phys.* **96** 5460
- [20] Straus J B, Gomez-Llorente J M and Voth G A 1993 Manifestations of spatially-dependent friction in classical activated rate processes *J. Chem. Phys.* **98** 4082
- [21] See, for example, Pollak E 1986 Theory of activated rate processes: a new derivation of Kramers' expression *J. Chem. Phys.* **85** 865
Pollak E 1987 Transition state theory for photoisomerization rates of *trans*-stilbene in the gas and liquid phases *J. Chem. Phys.* **86** 3944
Pollak E, Tucker S C and Berne B J 1990 Variational transition state theory for reaction rates in dissipative systems *Phys. Rev. Lett.* **65** 1399
Pollak E 1990 Variational transition state theory for activated rate processes *J. Chem. Phys.* **93** 1116
Pollak E 1991 Variational transition state theory for reactions in condensed phases *J. Phys. Chem.* **95** 533
Frishman A and Pollak E 1992 Canonical variational transition state theory for dissipative systems: application to generalized Langevin equations *J. Chem. Phys.* **96** 8877
Berezhkovskii A M, Pollak E and Zitserman V Y 1992 Activated rate processes: generalization of the Kramers–Grote–Hynes and Langer theories *J. Chem. Phys.* **97** 2422
- [22] Pollak E, Grabert H and Hänggi P 1989 Theory of activated rate processes for arbitrary frequency dependent friction: solution of the turnover problem *J. Chem. Phys.* **91** 4073
- [23] Stratt R M and Maroncelli M 1996 Nonreactive dynamics in solution: the emerging molecular view of solvation dynamics and vibrational relaxation *J. Phys. Chem.* **100** 12 981
- [24] Cao J and Voth G A 1995 A theory for time correlation functions in liquids *J. Chem. Phys.* **103** 4211
- [25] Haynes G R and Voth G A 1993 The dependence of the potential of mean force on the solvent friction: consequences for condensed phase activated rate theories *J. Chem. Phys.* **99** 8005
- [26] Voth G A 1992 A theory for treating spatially-dependent friction in classical activated rate processes *J. Chem. Phys.* **97** 5908
- [27] Haynes G R, Voth G A and Pollak E 1993 A theory for the thermally activated rate constant in systems with spatially dependent friction *Chem. Phys. Lett.* **207** 309
- [28] Haynes G R, Voth G A and Pollak E 1994 A theory for the activated barrier crossing rate constant in systems influenced by space and time dependent friction *J. Chem. Phys.* **101** 7811
- [29] Haynes G R and Voth G A 1995 Reaction coordinate dependent friction in classical activated barrier crossing dynamics: when it matters and when it doesn't *J. Chem. Phys.* **103** 10 176
- [30] Borkovec M, Straub J E and Berne B J The influence of intramolecular vibrational relaxation on the pressure dependence of unimolecular rate constants *J. Chem. Phys.* **85** 146
Straub J E and Berne B J 1986 Energy diffusion in many dimensional Markovian systems: the consequences of the competition between inter- and intra-molecular vibrational energy transfer *J. Chem. Phys.* **85** 2999
Straub J E, Borkovec M and Berne B J 1987 Numerical simulation of rate constants for a two degree of freedom system in the weak collision limit *J. Chem. Phys.* **86** 4296
Borkovec M and Berne B J 1987 Activated barrier crossing for many degrees of freedom: corrections to the low friction result *J. Chem. Phys.* **86** 2444

Gershinsky G and Berne B J 1999 The rate constant for activated barrier crossing: the competition between IVR and energy transfer to the bath *J. Chem. Phys.* **110** 1053

- [31] Hershkovitz E and Pollak E 1997 Multidimensional generalization of the PGH turnover theory for activated rate processes *J. Chem. Phys.* **106** 7678
- [32] Voth G A, Chandler D and Miller W H 1989 Time correlation function and path integral analysis of quantum rate constants *J. Phys. Chem.* **93** 7009
- [33] Voth G A, Chandler D and Miller W H 1989 Rigorous formulation of quantum transition state theory and its dynamical corrections *J. Chem. Phys.* **91** 7749
Voth G A 1990 Analytic expression for the transmission coefficient in quantum mechanical transition state theory *Chem. Phys. Lett.* **170** 289
- [34] Gillan M J 1987 Quantum simulation of hydrogen in metals *Phys. Rev. Lett.* **58** 563
Gillan M J 1987 Quantum-classical crossover of the transition rate in the damped double well *J. Phys. C: Solid State Phys.* **20** 3621
- [35] Voth G A 1993 Feynman path integral formulation of quantum mechanical transition state theory *J. Phys. Chem.* **97** 8365

- [36] Miller W H 1975 Semiclassical limit of quantum mechanical transition state theory for nonseparable systems *J. Chem. Phys.* **62** 1899
- [37] See, for example, Coleman S 1979 *The Whys of Subnuclear Physics* ed A Zichichi (New York: Plenum)
- [38] Wolynes P G 1981 Quantum theory of activated events in condensed phases *Phys. Rev. Lett.* **47** 968
- [39] Cao J and Voth G A 1996 A unified framework for quantum activated rate processes: I. General theory *J. Chem. Phys.* **105** 6856
Cao J and Voth G A 1997 A unified framework for quantum activated rate processes: II. The nonadiabatic limit *J. Chem. Phys.* **106** 1769
- [40] Makarov D E and Topaler M 1995 Quantum transition-state theory below the crossover temperature *Phys. Rev. E* **52** 178
- [41] Stuchebrukhov A A 1991 Green's functions in quantum transition state theory *J. Chem. Phys.* **95** 4258
- [42] Zhu J J and Cukier R I 1995 An imaginary energy method-based formulation of a quantum rate theory *J. Chem. Phys.* **102** 4123
- [43] Feynman R P and Hibbs A R 1965 *Quantum Mechanics and Path Integrals* (New York: McGraw-Hill)
Feynman R P 1972 *Statistical Mechanics* (Reading, MA: Addison-Wesley)
- [44] For reviews of numerical path integral techniques, see
Berne B J and Thirumalai D 1987 *Ann. Rev. Phys. Chem.* **37** 401
Doll J D, Freeman D L and Beck T L 1990 *Adv. Chem. Phys.* **78** 61
Doll J D and Gubernatis J E (eds) 1990 *Quantum Simulations of Condensed Matter Phenomena* (Singapore: World Scientific)
- [45] Pollak E and Liao J-L 1998 A new quantum transition state theory *J. Chem. Phys.* **108** 2733
Shao J, Liao J-L and Pollak E 1998 Quantum transition state theory—perturbation expansion *J. Chem. Phys.* **108** 9711
Liao J-L and Pollak E 1999 A test of quantum transition state theory for a system with two degrees of freedom *J. Chem. Phys.* **110** 80
- [46] Wang H, Sun X and Miller W H 1998 Semiclassical approximations for the calculation of thermal rate constants for chemical reactions in complex molecular systems *J. Chem. Phys.* **108** 9726
Sun X, Wang H and Miller W H 1998 On the semiclassical description of quantum coherence in thermal rate constants *J. Chem. Phys.* **109** 4190
- [47] Miller W H 1998 Quantum and semiclassical theory of chemical reaction rates *Faraday Disc. Chem. Soc.* **110** 1
- [48] Lobaugh J and Voth G A 1994 A path integral study of electronic polarization and nonlinear coupling effects in condensed phase proton transfer reactions *J. Chem. Phys.* **100** 3039
- [49] Valleau J P and Torrie G M 1977 *Statistical Mechanics, Part A* ed B J Berne (New York: Plenum)
-

A3.9 Molecular reaction dynamics: surfaces

George R Darling, Stephen Holloway and Charles Rettner

A3.9.1 INTRODUCTION

Molecular reaction dynamics is concerned with understanding elementary chemical reactions in terms of the individual atomic and molecular forces and the motions that occur during the process of chemical change. In gas phase and condensed phase reactions (discussed in [section A3.7](#) and [section A3.8](#)) the reactants, products and all intermediates are in the same phase. This ‘reduces’ the complexity of such systems such that we need ‘only’ develop experimental and theoretical tools to treat one medium. In a surface reaction, the reactants derive from the gas phase, to which the products may or may not return, but the surface is a condensed phase exchanging energy with reactants and products and any intermediates in a nontrivial fashion. The electronic states of the surface may also play a role by changing the bonding within and between the various species, affecting the reaction as a heterogeneous catalyst (see [section A3.10](#)). Of course, the surface itself may be one of the reactants, as in the etching of silicon surfaces by halide molecules. Indeed, it might be argued that if the

reactants achieve thermal equilibrium with the surface, they have become part of a new surface, with properties differing from those of the clean surface.

An individual surface reaction may be the result of several steps occurring on very different timescales. For example, a simple bimolecular reaction $A(\text{gas}) + B(\text{gas}) \rightarrow AB(\text{gas})$ might proceed as follows: A strikes the surface, losing enough energy to stick (i.e. adsorb), B also adsorbs, A and B diffuse across the surface and meet to form AB, after some time AB acquires enough energy to escape (i.e. desorb) from the surface. Each part of this schematic process is itself complicated. In the initial collisions with the surface, the molecules can lose or gain energy, and this can be translational energy (i.e. from the centre-of-mass motion) or internal (rotational, vibrational etc) energy, or both. Internal energy can be exchanged for translational energy, or *vice versa*, or the molecule can simply fragment on impact. Thermalization (i.e. the attainment of thermal equilibrium with the surface) is a slower process, requiring possibly tens of bounces of the molecule on the surface. The subsequent diffusion of A and B towards each other is even slower, while the desorption of the product AB might occur as soon as it is formed, leaving the molecule with some of the energy released in the association step.

Why should we be interested in the dynamics of such complex systems? Apart from the intellectual rewards offered by this field, understanding reactions at surfaces can have great practical and economic value. Gas–surface chemical reactions are employed in numerous processes throughout the chemical and electronic industries. Heterogeneous catalysis lies at the heart of many synthetic cycles, and etching and deposition are key steps in the fabrication of microelectronic components. Gas–surface reactions also play an important role in the environment, from acid rain to the chemistry of the ozone hole. Energy transfer at the gas–surface interface influences flight, controls spacecraft drag, and determines the altitude of a slider above a computer hard disk. Any detailed understanding of such processes needs to be built on fundamental knowledge of the dynamics and kinetics at the molecular level.

For any given gas–surface reaction, the various elementary steps of energy transfer, adsorption, diffusion, reaction and desorption are inextricably linked. Rather than trying to study all together in a single system where they cannot easily be untangled, most progress has been made by probing the individual steps in carefully chosen systems [1, 2 and 3].

-2-

For example, energy transfer in molecule–surface collisions is best studied in nonreactive systems, such as the scattering and trapping of rare-gas atoms or simple molecules at metal surfaces. We follow a similar approach below, discussing the dynamics of the different elementary processes separately. The surface must also be ‘simplified’ compared to technologically relevant systems. To develop a detailed understanding, we must know exactly what the surface looks like and of what it is composed. This requires the use of surface science tools (section B1.19-26) to prepare very well-characterized, atomically clean and ordered substrates on which reactions can be studied under ultrahigh vacuum conditions. The most accurate and specific experiments also employ molecular beam techniques, discussed in [section B2.3](#).

A3.9.2 REACTION MECHANISMS

The basic paradigms of surface reaction dynamics originate in the pioneering studies of heterogeneous catalysis by Langmuir [4, 5 and 6]. Returning to our model bimolecular reaction $A(\text{gas}) + B(\text{gas}) \rightarrow AB(\text{gas})$, let us assume first that A adsorbs on, and comes into thermal equilibrium with, the surface. We categorize the reaction according to the behaviour of molecule B. For most surface reactions, B adsorbs and thermalizes on the surface before meeting and reacting with A, by way of a Langmuir–Hinshelwood mechanism. However, in some systems, AB can only be formed as a result of a direct collision of the incoming B with the adsorbed A. Such reactions, which are discussed in further detail in [section A3.9.6](#), are

said to occur by an Eley–Rideal mechanism. A schematic illustration of these processes is shown in [figure A3.9.1](#).

For a Langmuir–Hinshelwood reaction, we can expect the surface temperature to be an important variable determining overall reactivity because it determines how fast A and B diffuse across the surface. If the product AB molecules thermalize before desorption, the distribution of internal and translational energies in the gas phase will also reflect the surface temperature (yielding Boltzmann distributions that are modified by a dynamical factor related by the principle of detailed balance to the energetics of adsorption [7]). The main factor discriminating between the reaction schemes is that for a Langmuir–Hinshelwood reaction, the AB molecule can have no memory of the initial state and motion of the B molecule, but these should be evident in the AB products if the mechanism is of the Eley–Rideal type. These simple divisions are of course too black and white. In all probability, the two paradigms are actually extremes, with real systems reflecting aspects of both mechanisms [8].

-3-

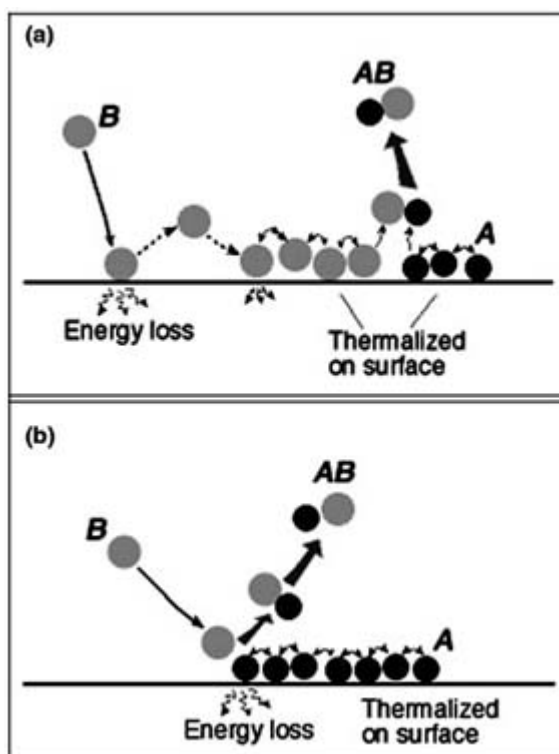


Figure A3.9.1. Schematic illustrations of (a) the Langmuir–Hinshelwood and (b) Eley–Rideal mechanisms in gas–surface dynamics.

A3.9.3 COLLISION DYNAMICS AND TRAPPING IN NONREACTIVE SYSTEMS

As with any collision process, to understand the dynamics of collisions we need an appreciation of the relevant forces and masses. Far from the surface, the incoming atom or molecule will experience the van der Waals attraction of the form

$$V(z) = -C/z^3 \tag{A3.9.1}$$

where z is the distance from the surface, and C is a constant dependent on the polarizability of the particle and the dielectric properties of the solid [9]. Close to the surface, where z is 0.1 nm for a nonreactive system, this attractive interaction is overwhelmed by repulsive forces (Pauli repulsion) due to the energy cost of orthogonalizing the overlapping electronic orbitals of the incoming molecule and the surface. The net result of van der Waals attraction and Pauli repulsion is a potential with a shallow well, the physisorption well, illustrated in figure A3.9.2. The depth of this well ranges from a few meV for He adsorption to ~ 30 meV for H_2 molecules on noble metal surfaces, and to ~ 100 meV for Ar or Xe on metal surfaces.

-4-

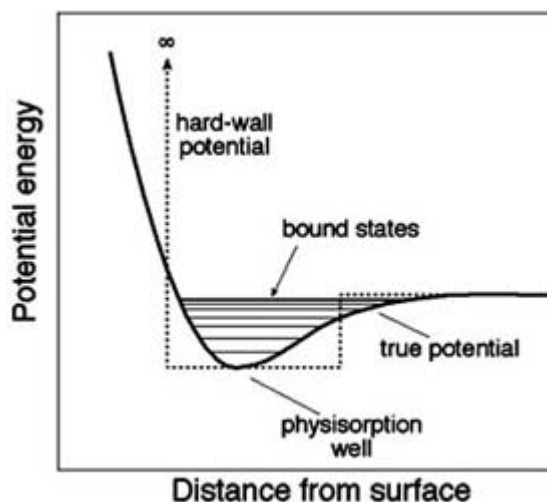


Figure A3.9.2. Interaction potential for an atom or molecule physisorbed on a surface. A convenient model is obtained by ‘squaring off’ the potential, which facilitates solution of the Schrödinger equation for the scattering of a quantum particle.

The van der Waals attraction arises from the interaction between instantaneous charge fluctuations in the molecule and surface. The molecule interacts with the surface as a whole. In contrast the repulsive forces are more short-range, localized to just a few surface atoms. The repulsion is, therefore, not homogeneous but depends on the point of impact in the surface plane, that is, the surface is corrugated.

A3.9.3.1 BINARY COLLISION (HARD-CUBE) MODEL

We can obtain an approximate description of the molecule–surface encounter using a binary collision model, with the projectile of mass m as one collision partner and a ‘cube’ having an effective mass, M , related to that of a surface atom, as the other partner. M depends on how close the projectile approaches the atoms of the surface, on the stiffness of the surface, and on the degree of corrugation of the repulsive potential. For rare-gas atoms interacting with metal surfaces, the surface electronic orbitals are delocalized and the repulsive interaction is effectively with a large cluster. The effective mass is correspondingly large, some 3–9 times the mass of a surface atom [10]. In other cases, such as O_2 colliding with Ag(111), the degree of corrugation and the effective mass may be closer to that expected for one atom [11].

Approximating the real potential by a square well and infinitely hard repulsive wall, as shown in figure A3.9.2 we obtain the hard cube model. For a well depth of W , conservation of energy and momentum lead [11, 12] to the very useful Baule formula for the translational energy loss, δE , to the substrate

$$\delta E = \frac{4\mu}{(1 + \mu)^2} (E + W) \quad (\text{A3.9.2})$$

where E is the initial translational energy of the projectile and μ is the ratio m/M . This formula shows us that the energy transfer increases with the mass of the projectile, reaching a maximum when projectile and cube masses are equal.

-5-

Of course the real projectile–surface interaction potential is not infinitely hard (cf figure A3.9.2). As E increases, the projectile can penetrate deeper into the surface, so that at its turning point (where it momentarily stops before reversing direction to return to the gas phase), an energetic projectile interacts with fewer surface atoms, thus making the effective cube mass smaller. Thus, we expect $\delta E/E$ to increase with E (and also with W since the well accelerates the projectile towards the surface).

The effect of surface temperature, T_S , can be included in this model by allowing the cube to move [12]. E becomes the translational energy in the frame of the centre-of-mass of projectile and cube; then we average the results over E , weighting with a Boltzmann distribution at T_S . This causes δE to decrease with increasing T_S , and when the thermal energy of the cube, kT_S , substantially exceeds E , the projectile actually gains energy in the collision! This is qualitatively consistent with experimental observations of the scattering of beams of rare-gas atoms from metal surfaces [14, 15].

A3.9.3.2 SCATTERING AND TRAPPING–DESORPTION DISTRIBUTIONS

Projectiles leaving the surface promptly after an inelastic collision have exchanged energy with the surface, yet their direction of motion and translational and internal energies are clearly related to their initial values. This is called direct-inelastic (DI) scattering. At low E , the projectile sees a surface in thermal motion. This motion dominates the final energy and angular distributions of the scattering, and so this is referred to as thermal scattering. As E becomes large, the projectile penetrates the surface more deeply, seeing more of the detailed atomic structure, and the interaction comes to be dominated by scattering from individual atoms. Eventually E becomes so large that the surface thermal motion becomes negligible, and the energy and angular distributions depend only on the atomic structure of the surface. This is known as the structure-scattering regime. Comparing experimental results with those of detailed classical molecular dynamics modelling of these phenomena can allow one to construct good empirical potentials to describe the projectile-surface interaction, as has been demonstrated for the Xe/Pt(111) [16] and Ar/Ag(111) [17] systems.

From equation (A3.9.2), we can see that at low E , the acceleration into the well dominates the energy loss, that is, δE does not reduce to zero with decreasing E . Below a critical translational energy, given by

$$E_c = \frac{4\mu W}{(1 - \mu)^2} \quad (\text{A3.9.3})$$

the projectile has insufficient energy remaining to escape from the well and it traps at the surface. Inclusion of surface temperature (cube motion) leads to a blurring of this cut-off energy so that trapping versus energy curves are predicted to be smoothed step functions. In fact, true trapping versus energy curves are closer to exponential in form, due to the combined effects of additional averaging over variations of W with surface site and with the orientation of the incident molecule. Additionally, transfer of motion normal to the surface to motion parallel to the surface, or into internal motions (rotations) can also lead to trapping, as we shall discuss below.

Trapped molecules can return to the gas phase once the thermal energy, kT_S , becomes comparable to the well depth. Having equilibrated with the surface, they have velocity, angular distribution and internal energies determined by T_S . This is visible in experiment as a scattering component (the trapping–desorption (TD) scattering component) with a very different appearance to the DI component, being peaked at and symmetrical about the surface normal, independently of the incidence conditions of the beam of projectiles.

Such behaviour has been seen in many systems, for example in the scattering of Ar from Pt(111) [10] as illustrated in [figure A3.9.3](#).

-6-

Figure A3.9.3. Time-of-flight spectra for Ar scattered from Pt(111) at a surface temperature of 100 K [10]. Points in the upper plot are actual experimental data. Curve through points is a fit to a model in which the bimodal distribution is composed of a sharp, fast moving (hence short flight time), direct-inelastic (DI) component and a broad, slower moving, trapping–desorption (TD) component. These components are shown separately in the lower curves. Parameters: $E = 12.5 \text{ kJ mol}^{-1}$; $\theta_i = 60^\circ$; $\theta_f = 40^\circ$; $T_s = 100 \text{ K}$.

A3.9.3.3 SELECTIVE ADSORPTION

Light projectiles impinging on a cold surface exhibit strong quantum behaviour in the scattering and trapping dynamics. Motion in the physisorption well is quantized normal to the surface, as indicated in [figure A3.9.2](#). Although in the gas phase the projectile can have any parallel momentum, when interacting with a perfect surface, the parallel momentum can only change by whole numbers of reciprocal lattice vectors (the wavevectors corresponding to wavelengths fitting within the surface lattice) [9]. The scattering is thus into special directions, forming a diffraction pattern, which is evident even for quite massive particles such as Ar [18]. These quantizations couple to yield maxima in the trapping probability when, to accommodate the gain in parallel momentum, the projectile must drop into one of the bound states in the z -direction. In other words, the quantized gain in parallel motion leaves the projectile with more translational energy than it had initially, but the excess is cancelled by the negative energy of the bound state [19]. This is an entirely elastic phenomenon, no energy loss to the substrate is required, simply a conversion of normal for parallel motion. The trapping is undone if the parallel momentum gain is reversed.

The energies of the selective adsorption resonances are very sensitive to the details of the physisorption potential. Accurate measurement allied to computation of bound state energies can be used to obtain a very accurate quantitative form for the physisorption potential, as has been demonstrated for helium atom scattering. For molecules, we have

-7-

the additional possibility of exchanging normal translations for rotational motion (the vibrational energies of light molecules are much larger than typical physisorption energies). Parallel momentum changes are effected by the surface corrugation, giving rise to corrugation mediated selective adsorption (CMSA). By analogy, rotational excitations produce rotation mediated selective adsorption (RMSA). Together these yield the acronym CRMSA. All such processes have been identified in the scattering of H_2 and its isotopomers from noble and simple metal surfaces [20]. Typical results are shown in [figure A3.9.4](#). The selective adsorption resonances show up as peaks in the trapping (minima in the reflectivity) because the long residence time at the surface increases the amount of energy lost to the substrate, resulting in sticking [21].

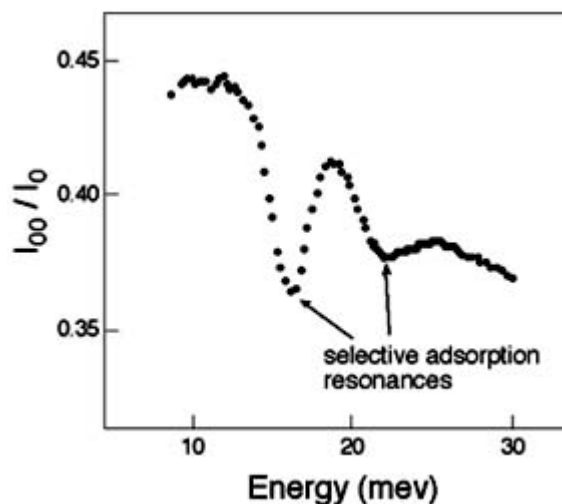


Figure A3.9.4. The ratio of specular reflectivity to incident beam intensity ratio for D_2 molecules scattering from a Cu(100) surface at 30 K [21].

A3.9.4 MOLECULAR CHEMISORPTION AND SCATTERING

Unlike physisorption, chemisorption results from strong attractive forces mediated by chemical bonding between projectile and surface [9]. There is often significant charge transfer between surface and molecule, as in the adsorption of O_2 on metal surfaces [22]. The characteristics, well depth, distance of minimum above the surface etc. can vary greatly with surface site [23]. The degree of charge transfer can also differ, such that in many systems we can speak of there being more than one chemisorbed species [24].

The chemisorption interaction is also very strongly dependent on the molecular orientation, especially for heteronuclear molecules. This behaviour is exemplified by NO adsorption on metal surfaces, where the N end is the more strongly bound. These anisotropic interactions lead to strong steric effects and consequent rotational excitation in the scattering dynamics. Rainbows are evident in the rotational distributions [25], as can be seen in [figure A3.9.5](#). These steric effects show up particularly strongly when the incident molecules are aligned prior to scattering (by magnetic fields) [26, 27].

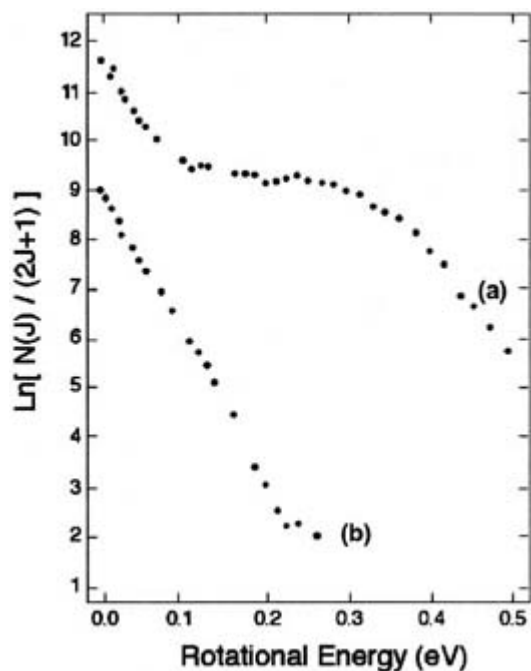


Figure A3.9.5. Population of rotational states versus rotational energy for NO molecules scattered from an Ag (111) surface at two different incidence energies and at $T_S = 520$ K [25]: (a) $E = 0.85$ eV, $\theta_i = 15^\circ$ and (b) $E = 0.09$ eV, $\theta_i = 15^\circ$. Results at $E = 0.85$ eV show a pronounced rotational rainbow.

The change of charge state of an adsorbed molecule leads to a change in the intramolecular bonding, usually a lengthening of the bond, which can result in vibrational excitation of the scattered molecule. Once again, this shows up in the scattering of NO from the Ag(111) surface [28], as shown in [figure A3.9.6](#). In this case, the vibrational excitation probability is dependent on both the translational energy and the surface temperature. The translational energy dependence is probably due to the fact that the closer the molecule is to a surface, the more extended the molecular bond becomes, that is, in the language of [section A3.9.5](#), the NO is trying to get round the elbow (see [figure A3.9.8](#)) to dissociate. It fails, and returns to the gas phase with increased vibrational energy. Surface temperature can enhance this process by supplying energy from the thermal motion of a surface atom towards the molecule [29, 30], but interaction with electronic excitations in the metal has also been demonstrated to be an efficient and likely source of energy transfer to the molecular vibrations [29, 30].

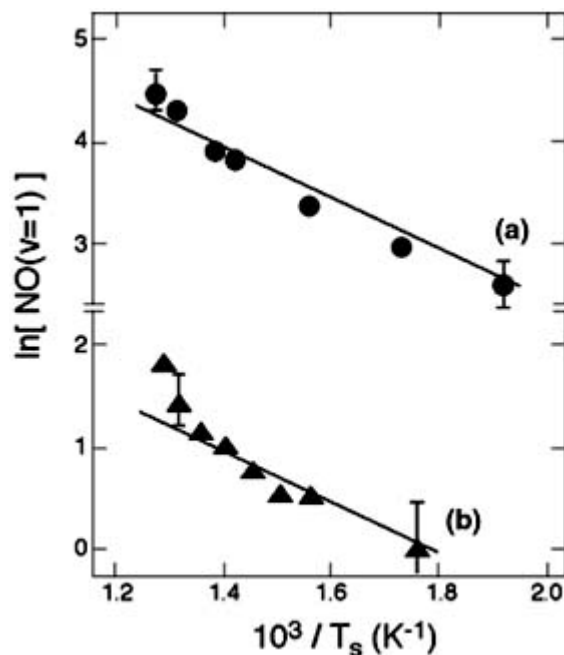


Figure A3.9.6. Population of the first excited vibrational state ($v = 1$) versus inverse of surface temperature for NO scattering from an Ag(111) surface [28]. Curves: (a) $E = 102 \text{ kJ mol}^{-1}$ and (b) $E = 9 \text{ kJ mol}^{-1}$.

A3.9.4.1 CHEMISORPTION AND PRECURSOR STATES

The chemisorption of a molecule is often a precursor [31] to further reactions such as dissociation (see section A3.9.5.2), that is, the molecule must reside in the precursor state exploring many configurations until finding that leading to a reaction. Where there is more than one distinct chemisorption state, one can act as a precursor to the other [32]. The physisorption state can also act as a precursor to chemisorption, as is observed for the $\text{O}_2/\text{Ag}(110)$ system [33].

The presence of a precursor breaks the dynamical motion into three parts [34]. First, there is the dynamics of trapping into the precursor state; secondly, there is (at least partial) thermalization in the precursor state; and, thirdly, the reaction to produce the desired species (possibly a more tightly bound chemisorbed molecule). The first two of these we can readily approach with the knowledge gained from the studies of trapping and sticking of rare-gas atoms, but the long timescales involved in the third process may perhaps more usefully be addressed by kinetics and transition state theory [35].

A3.9.5 DYNAMICS OF DISSOCIATION REACTIONS

A3.9.5.1 DIRECT DISSOCIATION OF DIATOMICS

The direct dissociation of diatomic molecules is the most well studied process in gas–surface dynamics, the one for which the combination of surface science and molecular beam techniques allied to the computation of total energies and detailed and painstaking solution of the molecular dynamics has been most successful. The result is a substantial body of knowledge concerning the importance of the various degrees of freedom (e.g. molecular rotation) to the reaction dynamics, the details of which are contained in a number of review articles [2, 36, 37, 38, 39, 40 and 41].

(A) LENNARD-JONES MODEL OF HYDROGEN DISSOCIATION

In the 1930s Lennard-Jones [42] introduced a model that is still in use today in discussions of the dissociation of molecules at surfaces. He proposed a description based on two potential energy curves. The first, describing the interaction of the intact molecule with the surface as a physisorption potential, is shown as curve (a) in figure A3.9.7. Coupled with this, there is a second potential describing the interaction of the two separately chemisorbed atoms with the surface (curve (b) in figure A3.9.7). In equilibrium the adsorbed atoms are located at the minimum, L, of curve (b). The difference between (a) and (b) far from the surface is the gas-phase molecular dissociation energy, D . A dissociation event occurs if a molecule approaches the surface until K, where it makes a radiationless transition from (a) to (b) becoming adsorbed as atoms.

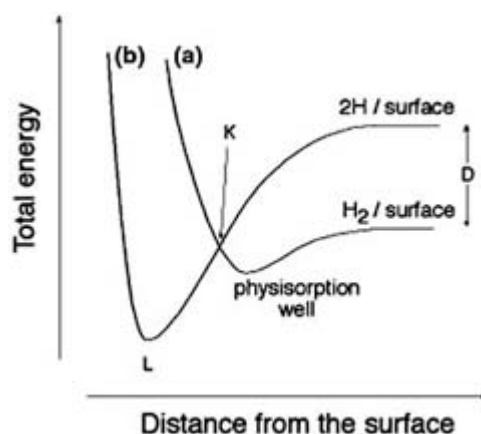


Figure A3.9.7. A representation of the Lennard-Jones model for dissociative adsorption of H_2 . Curves: (a) interaction of intact molecule with surface; (b) interaction of two separately chemisorbed atoms with surface.

There is an inconsistency in the model in that when changing from (a) to (b) the molecular bond is instantaneously elongated. Lennard-Jones noted that although one-dimensional potential energy curves (such as shown in figure A3.9.7 can prove of great value in discussions, ‘they do not lend themselves to generalization when more than one coordinate is necessary to specify a configuration’. In a quantitative theory there should be a number of additional curves between (a) and (b) corresponding to rotational and vibrational states of the molecule. In modern terms, we try to describe each

-11-

degree-of-freedom relevant to the problem with a separate dimension. The potential energy curves of [figure A3.9.7](#) then become a multidimensional surface, the potential energy surface (PES).

The model illustrated in [figure A3.9.7](#) is primarily diabatic, the molecule jumps suddenly from one type of bonding, represented by a potential energy curve, to another. However, much of the understanding in gas-surface dynamics derives from descriptions based on motion on a single adiabatic PES, usually the ground-state PES. In the Lennard-Jones model, this would approximately correspond to whichever of (a) and (b) has the lower energy. Although this approach is successful in describing H_2 dissociation, it will not be adequate for reactions involving very sudden changes of electronic state [43]. These may occur, for example, in the O_2 reaction with simple metal surfaces [44]; they are so energetic that they can lead to light or electron emission during reaction [45].

(B) INFLUENCE OF MOLECULAR VIBRATION ON REACTION

Dissociation involves extension of a molecular bond until it breaks and so it might seem obvious that the more energy we can put into molecular vibration, the greater the reactivity. However, this is not always so: the

existence of a vibrational enhancement of dissociation reveals something about the shape, or topography of the PES itself. This is illustrated in figure A3.9.8 which shows a generic elbow PES [37]. This two-dimensional PES describes the dynamics in the molecule–surface and intramolecular bond length coordinates only. Far from the surface, it describes the intramolecular bonding of the projectile by, for example, a Morse potential. Close to the surface at large bond length, the PES describes the chemisorption of the two atoms to the surface in similar fashion to curve (b) in figure A3.9.7. The curved region linking these two extremes is the interaction region (shaded in figure A3.9.8), where the bonding is changing from one type to another. It corresponds roughly to the curve crossing point, K, in the Lennard-Jones model.

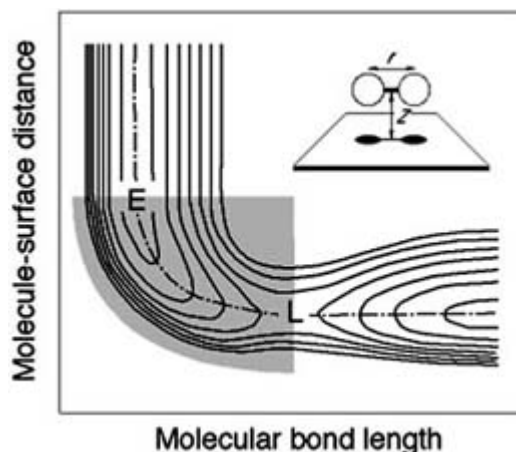


Figure A3.9.8. An elbow potential energy surface representing the dissociation of a diatomic in two dimensions—the molecular bond length and the distance from the molecule to the surface.

For vibrational effects in the dynamics, the location of the dissociation barrier within the curved interaction region of figure A3.9.8 is crucial. If the barrier occurs largely before the curved region, it is an ‘early’ barrier at point E, then vibration will not promote reaction as it occurs largely at right angles to the barrier. In contrast, if the barrier occurs when the bond is already extended, say at L (a ‘late’ barrier) in the figure, the vibration is now clearly helping the molecule to attack this barrier, and can substantially enhance reaction.

The consequences of these effects have been fully worked out, and agree with the Polanyi rules used in gas-phase scattering [46, 47]. Experimental observations of both the presence and absence of vibrational enhancement have been made, most clearly in hydrogen dissociation on metal surfaces. For instance, H₂ dissociation on Ni surfaces shows no vibrational enhancement [48, 49]. On Cu surfaces, however, vibrational enhancement of dissociation has been clearly demonstrated by using molecular beam techniques (section B2.6) to vary the internal and translational energies independently [50] and by examining the energy and state distributions of molecules undergoing the reverse of dissociation, the associative desorption reaction [51]. Figure A3.9.9 shows typical results presented in the form of dissociation versus translational energy curves backed out from the desorption data [52]. The curves corresponding to the vibrationally excited states clearly lie at lower energy than those for the vibrational ground-state, implying that some of the energy for the reaction comes from the H₂ vibration.

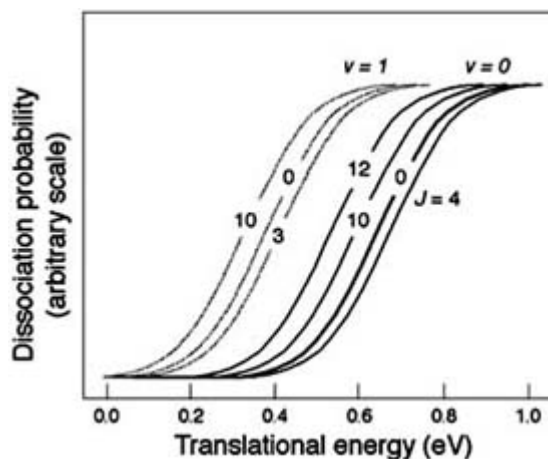


Figure A3.9.9. Dissociation probability versus incident energy for D_2 molecules incident on a Cu(111) surface for the initial quantum states indicated (v indicates the initial vibrational state and J the initial rotational state) [100]. For clarity, the saturation values have been scaled to the same value irrespective of the initial state, although in reality the saturation value is higher for the $v = 1$ state.

An important further consequence of curvature of the interaction region and a late barrier is that molecules that fail to dissociate can return to the gas-phase in vibrational states different from the initial, as has been observed experimentally in the H_2/Cu system [53, 55]. To undergo vibrational (de-)excitation, the molecules must round the elbow part way, but fail to go over the barrier, either because it is too high, or because the combination of vibrational and translational motions is such that the molecule moves across rather than over the barrier. Such vibrational excitation and de-excitation constrains the PES in that we require the elbow to have high curvature. Dissociation is not necessary, however, for as we have pointed out, vibrational excitation is observed in the scattering of NO from Ag(111) [55].

(C) ROTATIONAL EFFECTS: STERIC HINDRANCE AND CENTRIFUGAL ENHANCEMENT

Molecular rotation has two competing influences on the dissociation of diatomics [56, 57 and 58]. A molecule will only be able to dissociate if its bond is oriented correctly with respect to the plane of the surface. If the bond is parallel to the plane, then dissociation will take place, whereas if the molecule is end-on to the surface, dissociation requires one atom to be ejected into the gas phase. In most cases, this ‘reverse Eley–Rideal’ process is energetically very

unfavourable (although it does occur for very energetic reactions such as halide adsorption on Si surfaces, and possibly for O_2 adsorbing on reactive metals [59]). In general, molecules cannot dissociate when oriented end-on. The PES is, thus, highly corrugated in the molecular orientation coordinate. In consequence, increasing the rapidity of motion in this coordinate (i.e. increasing the rotational state) will make it more likely for the molecule to race past the small dissociation window at the parallel orientation, strike-off a more repulsive region of the PES and return to the gas phase. Therefore, dissociation is inhibited by increasing the rotational energy of the molecule. In opposition to this effect, the rotational motion can enhance reactivity when the dissociation barrier is late (i.e. occurs at extended bond length). As the molecule progresses through the interaction region, its bond begins to extend. This increases the moment of inertia and thus reduces the rotational energy. The rotational energy thus ‘lost’ feeds into the reaction coordinate, further stretching the molecular bond and enhancing the reaction. The combination of these two competing effects has been demonstrated in the $H_2/Cu(111)$ system. For the first few rotational states, increases in rotation reduce the dissociation (i.e. shift the dissociation curve to higher energy) as can be seen in [figure A3.9.9](#). Eventually, however, centrifugal enhancement wins out, and for the higher rotational states the dissociation curves are pushed to lower translational energies.

The strong dependence of the PES on molecular orientation also leads to strong coupling between rotational states, and hence rotational excitation/de-excitation in the scattering. This has been observed experimentally for H₂ scattering from Cu surfaces. Recent work has shown that for H₂ the changes in rotational state occur almost exclusively when the molecular bond is extended, that is, longer than the gas-phase equilibrium value [60].

(D) SURFACE CORRUGATION AND SITE SPECIFICITY OF REACTION

The idea that certain sites on a surface are especially active is common in the field of heterogeneous catalysis [61]. Often these sites are defects such as dislocations or steps. But surface site specificity for dissociation reactions also occurs on perfect surfaces, arising from slight differences in the molecule–surface bonding at different locations. This is so not only of insulator and semiconductor surfaces where there is strongly directional bonding, but also of metal surfaces where the electronic orbitals are delocalized. The site dependence of the reactivity manifests itself as a strong corrugation in the PES, which has been shown to exist by *ab initio* computation of the interaction PES for H₂ dissociation on some simple and noble metal surfaces [62, 63 and 64].

The dynamical implications of this corrugation appear straightforward: surface sites where the dissociation barrier is high (unfavourable reaction sites) should shadow those sites where the barrier is low (the favoured reaction sites) if the reacting molecule is incident at an angle to the surface plane. If we assume that the motion normal to the surface is important in traversing the dissociation barrier, then those molecules approaching at an angle should have lower dissociation probability than those approaching at normal incidence. This has indeed been observed in a number of dissociation systems [37], but a far more common observation is that the dissociation scales with the ‘normal energy’, $E_p = E \cos^2 \theta$, where E is the translational energy, and θ the angle of incidence of the beam with respect to the surface normal. Normal energy scaling, shown in [figure A3.9.10](#) implies that the motion parallel to the surface does not affect dissociation, and the surface appears flat.

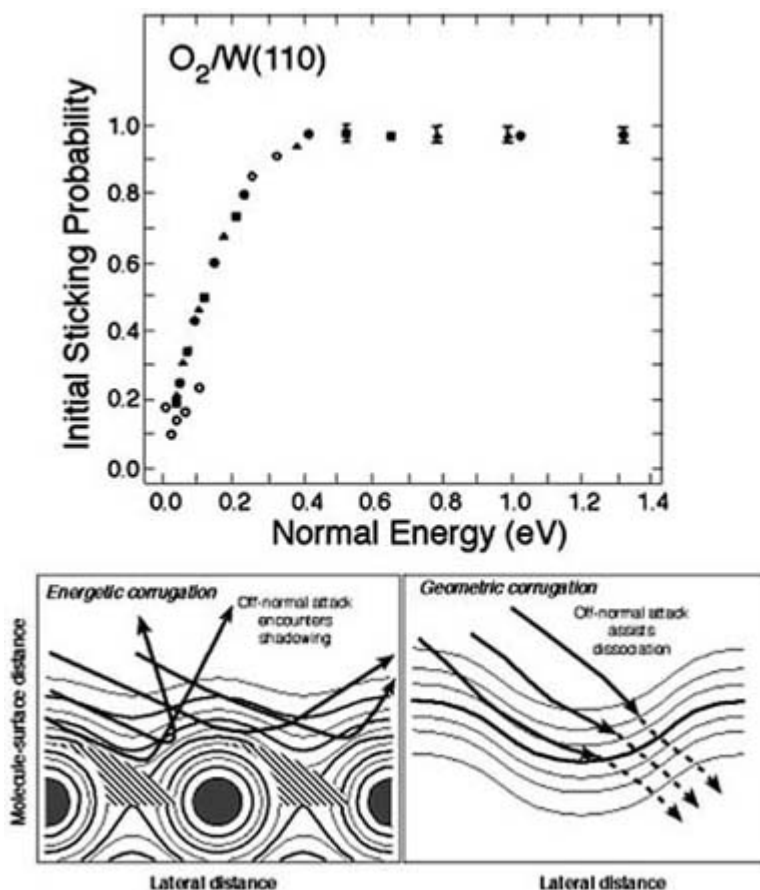


Figure A3.9.10. The dissociation probability for O₂ on W(110) [101] as a function of the normal energy, (upper). $T_g = 800$ K; θ : (●) 0°, (▲) 30° (■) 45° and (○) 60°. The normal energy scaling observed can be explained by combining the two surface corrugations indicated schematically (lower diagrams).

This difficulty has been resolved with the realization that the surface corrugation is not merely of the barrier energy, but of the distance of the barrier above the surface [65]. We then distinguish between *energetic corrugation* (the variation of the energetic height of the barrier) and *geometric corrugation* (a simple variation of the barrier location or shape). The two cases are indicated in figure A3.9.10. For energetic corrugation, the shadowing does lead to lower dissociation at off-normal incidence, but this can be counterbalanced by geometric corrugation, for which the parallel motion helps the molecule to attack the facing edge of the PES [65].

The site specificity of reaction can also be a state-dependent site specificity, that is, molecules incident in different quantum states react more readily at different sites. This has recently been demonstrated by Kroes and co-workers for the H₂/Cu(100) system [66]. Additionally, we can find reactivity dominated by certain sites, while inelastic collisions leading to changes in the rotational or vibrational states of the scattering molecules occur primarily at other sites. This spatial separation of the active site according to the change of state occurring (dissociation, vibrational excitation etc) is a very surface specific phenomenon.

(E) STEERING DOMINATED REACTION

A very extreme version of surface corrugation has been found in the nonactivated dissociation reactions of H₂ on W [67, 68], Pd and Rh systems. In these cases, the very strong chemisorption bond of the H atoms gives rise to a very large energy release when the molecule dissociates. In consequence, at certain sites on the surface, the molecule accelerates rapidly downhill into the dissociation state. At the unfavourable sites, there

are usually small dissociation barriers and, of course, molecules oriented end-on to the surface cannot dissociate. When we examine the dynamics of motion on such PESs, we find that the molecules are steered into the attractive downhill regions [69], away from the end-on orientation and away from the unfavourable reaction sites.

Steering is a very general phenomenon, caused by gradients in the PES, occurring in every gas–surface system [36]. However, for these nonactivated systems showing extreme variations in the PES, the steering dominates the dissociation dynamics. At the very lowest energies, most molecules have enough time to steer into the most favourable geometry for dissociation hence the dissociation probability is high. At higher E , there is less time for steering to be effective and the dissociation decreases. The general signature of a steering dominated reaction is, therefore, a dissociation probability that falls with increasing E [49, 70, 71], as shown in figure A3.9.11. This can be contrasted with the curve usually expected for direct dissociation, figure A3.9.10, one which increases with E because, as E increases, it is easier to overcome the barriers in unfavourable geometries.

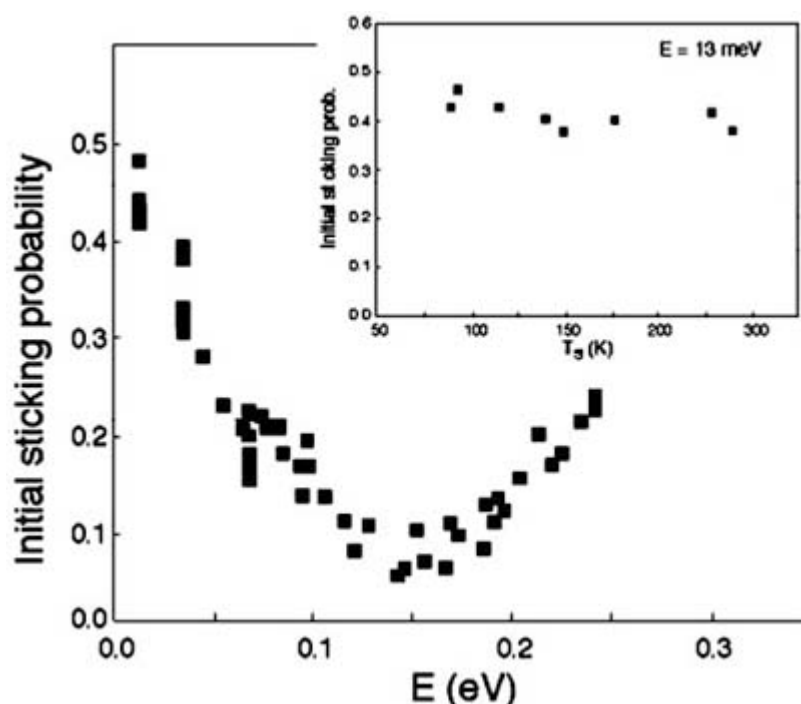


Figure A3.9.11. Dissociation of H_2 on the $W(100)-c(2 \times 2)-Cu$ surface as a function of incident energy [71]. The steering dominated reaction [102] is evident at low energy, confirmed by the absence of a significant surface temperature.

(F) SURFACE TEMPERATURE DEPENDENCE

Direct dissociation reactions are affected by surface temperature largely through the motion of the substrate atoms [72]. Motion of the surface atom towards the incoming molecule increases the likelihood of (activated) dissociation, while motion away decreases the dissociation probability. For low dissociation probabilities, the net effect is an enhancement of the dissociation by increasing surface temperature, as observed in the system $O_2/Pt\{100\}-hex-R0.7^\circ$ [73].

This interpretation is largely based on the results of cube models for the surface motion. It may also be that

the thermal disorder of the surface leads to slightly different bonding and hence different barrier heights. Increasing temperature also changes the populations of the excited electronic states in the surface, which may affect bonding. The contribution of these effects to the overall surface temperature dependence of reaction is presently not clear.

A3.9.5.2 DIRECT DISSOCIATION OF POLYATOMIC MOLECULES

Although understanding the dissociation dynamics of diatomic molecules has come a long way, that of polyatomic molecules is much less well-developed. Quite simply, this is due to the difficulty of computing adequate PESs on which to perform suitable dynamics, when there are many atoms. Quantum dynamics also becomes prohibitively expensive as the dimensionality of the problem increases. The dissociation of CH_4 (to $\text{H} + \text{CH}_3$) on metal surfaces is the most studied to date [74]. This shows dependences on molecular translational energy and internal state, as well as a strong surface temperature dependence, which has been interpreted in terms of thermally assisted quantum tunnelling through the dissociation barrier. More recent experimental work has shown complicated behaviour at low E , with the possible involvement of steering or trapping [75].

A3.9.5.3 PRECURSOR-MEDIATED DISSOCIATION

Precursor-mediated dissociation involves trapping in a molecularly chemisorbed state (or possibly several states) prior to dissociation. If the molecule thermalizes before dissociation, we can expect to observe the signature of trapping in the dissociation dynamics, that is, we expect increasing E and increasing surface temperature to decrease the likelihood of trapping, and hence of dissociation. This is exemplified by the dissociation of N_2 on $\text{W}(100)$ in the low energy regime [76], shown in figure A3.9.12.

The thermalization stage of this dissociation reaction is not amenable to modelling at the molecular dynamics level because of the long timescales required. For some systems, such as $\text{O}_2/\text{Pt}(111)$, a kinetic treatment is very successful [77]. However, in others, thermalization is not complete, and the internal energy of the molecule can still enhance reaction, as observed for $\text{N}_2/\text{Fe}(111)$ [78, 79] and in the dissociation of some small hydrocarbons on metal surfaces [80]. A detailed explanation of these systems is presently not available.

-17-

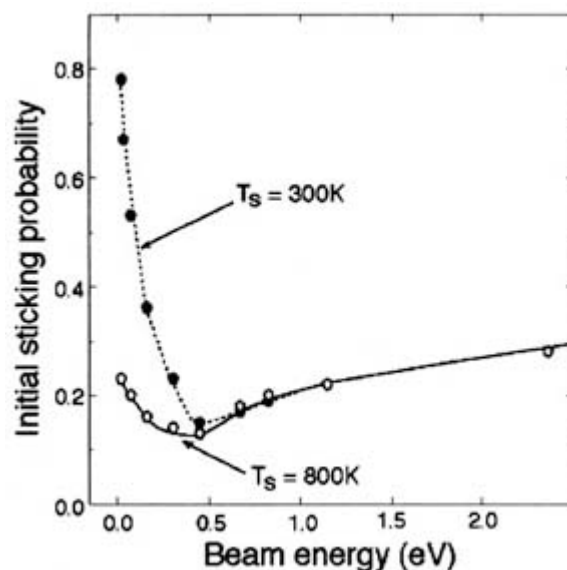


Figure A3.9.12. The dissociation probability of N_2 on the $\text{W}(100)$ surface as a function of the energy of the

molecular beam. The falling trend and pronounced surface temperature dependence are indicative of a precursor mediated reaction at low energies.

A3.9.6 ELEY–RIDEAL DYNAMICS

The idea that reactions can occur directly when an incident reagent strikes an adsorbate is strongly supported by detailed theoretical calculations. From early classical simulations on model PESs (e.g. for H on H/tungsten [81] and O on C/platinum [82]), to more recent theoretical studies of ER reactions employing quantum-mechanical models, it has been clearly established that product molecules should show a high degree of internal excitation. As noted in [section A3.9.2](#), certain highly facile gas–surface reactions can occur directly at the point of impact between an incident gas-phase reagent and an adsorbate; however, it is more likely that the incident reagent will ‘bounce’ a few times before reaction, this being to some degree accommodated in the process. A more useful working definition of an ER reaction is that it should occur before the reagents have become equilibrated at the surface. With this definition, we encompass hot-atom dynamics and what Harris and Kasemo have termed *precursor dynamics* [8]. Since the heat of adsorption of the incident reagent is not fully accommodated, ER reactions are far more exothermic than their LH counterparts.

Until relatively recently, the experimental evidence for the ER mechanism came largely from kinetic measurements, relating the rate of reaction to the incident flux and to the surface coverage and temperature [83]. For example, it has been found that the abstraction of halogens from Si(100) by incident H proceeds with a very small activation barrier, consistent with an ER mechanism. To *prove* that a reaction can occur on essentially a single gas–surface collision, however, dynamical measurements are required. The first definitive evidence for an ER mechanism was obtained in 1991, in a study showing that hyperthermal $\text{N}(\text{C}_2\text{H}_4)_3\text{N}$ can pick up a proton from a H/Pt(111) surface to give an ion with translational energy dependent on that of the incident molecule [84]. A year later, a study was reported of the formation of HD from H atoms incident on D/Cu(111), and from D incident on H/Cu(111) [85]. The angular

-18-

distribution of the HD was found to be asymmetrical about the surface normal and peaked on the opposite side of the normal to that of the incident atom. This behaviour proved that the reaction must occur before the incident atom reaches equilibrium with the surface. Moreover, the angular distribution was found to depend on the translational energy of the incident atom and on which isotope was incident, firmly establishing the operation of an ER mechanism for this elementary reaction.

Conceptually similar studies have since been carried out for the reaction of H atoms with Cl/Au(111). More recently, quantum-state distributions have been obtained for both the H + Cl/Au(111) [86, 87 and 88] and H(D) + D (H)/Cu(111) systems. The results of these studies are in good qualitative agreement with calculations. Even for the H(D) + D (H)/Cu(111) system [89], where we know that the incident atom cannot be significantly accommodated prior to reaction, reaction may not be direct. Detailed calculations yield much smaller cross sections for direct reaction than the overall experimental cross section, indicating that reaction may occur only after trapping of the incident atom [90].

Finally, it should also be clear that ER reactions do not necessarily yield a gas-phase product. The new molecule may be trapped on the surface. There is evidence for an ER mechanism in the addition of incident H atoms to ethylene and benzene on Cu(111) [91], and in the abstraction of H atoms from cyclohexane by incident D atoms [92], and the direct addition of H atoms to CO on Ru(001) [93].

A3.9.7 PHOTOCHEMISTRY

The interaction of light with both clean surfaces and those having adsorbed species has been a popular research topic over the past 10 years [94]. Our understanding of processes such as photodesorption, photodissociation and photoreaction is still at a very early stage and modelling has been largely performed on a system-by-system basis rather than any general theories being applicable. One of the most important aspects of performing photochemical reactions on surfaces, which has been well documented by Polanyi and co-workers is that it is possible to align species before triggering reactions that cannot be done in the gas phase. This is frequently referred to as surface aligned photochemistry [95]. One of the key issues when light, such as that from a picosecond laser, impinges a surface covered with an adsorbate is where the actual absorption takes place. Broadly speaking there are two possible choices either in the adsorbate molecule or the surface itself. Unfortunately, although it may seem that unravelling microscopic reaction mechanisms might be quite distinct depending on what was absorbing, this is not the case and considerable effort has been spent on deciding what the dynamical consequences are for absorption into either localized or extended electronic states [96].

Of lesser interest here for a laser beam incident upon a surface are the processes that occur due to surface heating. Of greater interest are those occasions when an electronic transition is initiated and a process occurs, for in these circumstances it becomes possible to ‘tune’ reactivity by an external agent. A good example of this is the UV photodissociation of a range of carbonyls on Si surfaces [97]. Here it was shown explicitly that 257 nm light can selectively excite the adsorbate and then dissociation ensues. An alternative story unfolds when NO is photodesorbed from Pt surfaces. Detailed experiment and modelling shows that, in this case, the initial excitation (absorption) event occurs in the metal substrate. Following this, the excess energy is transferred to the adsorbate by a hot electron which resides for about 10–12 fs before returning to the substrate. During this time, it is possible for the NO to gain sufficient energy to overcome the adsorption bond [98].

-19-

Finally, and most recently, femtosecond lasers have been employed to investigate reactions on surfaces, one good example is the oxidation of CO on a Ru surface [99]. One of the long outstanding problems in surface dynamics is to determine the energy pathways that are responsible for irreversible processes at the surface. Both phonons and electrons are capable of taking energy from a prethermalized adsorbate and because of the time required for converting electronic motion to nuclear motion, there is the possibility that measurements employing ultrashort-pulsed lasers might be able to distinguish the dominant pathway.

A3.9.8 OUTLOOK

Despite the considerable progress over the 1990s, the field of gas–surface reaction dynamics is still very much in its infancy. We have a relatively good understanding of hydrogen dissociation on noble metals but our knowledge of other gas–surface systems is far from complete. Even for other diatomic reagents such as N₂ or O₂ a great deal yet remains to be learned. Nevertheless, we believe that progress will take place even if in a slightly different fashion to that which is described here.

In parallel with the remarkable increase in computing power, particularly in desktop workstations, there have been significant advances also in the algorithmic development of codes that can calculate the potential energy (hyper-)surfaces that have been mentioned in this article. Most of the theoretical work discussed here has relied to a greater or lesser extent on potential energy surfaces being available from some secondary agency

and this, we believe, will not be the case in the future. Software is now available which will allow the dynamicist to calculate new potentials and then deploy them to evaluate state-to-state cross sections and reaction probabilities. Although new, detailed experimental data will provide guidance, a more general understanding of gas–surface chemistry will develop further as computational power continues to increase.

REFERENCES

- [1] Barker J A and Auerbach D J 1985 Gas–surface interactions and dynamics; thermal energy atomic and molecular beam studies *Surf. Sci.Rep.* **4** 1
- [2] Rettner C T and Ashfold M N R 1991 *Dynamics of Gas–Surface Interactions* (London: Royal Society of Chemistry)
- [3] Rettner C T, Auerbach D J, Tully J C and Kleyn A W 1996 Chemical dynamics at the gas–surface interface *J. Phys. Chem.* **100** 13 201
- [4] Langmuir I 1922 Chemical reactions on surfaces *Trans. Faraday Soc.* **17** 607
- [5] Holloway S 1993 Dynamics of gas–surface interactions *Surf. Sci.* **299/300** 656
- [6] Ertl G 1993 Reactions at well-defined surfaces *Surf.Sci.* **299/300** 742 3:08 PM 2/7/2002
- [7] Rettner C T, Michelsen H A and Auerbach D J 1993 From quantum-state-specific dynamics to reaction-rates—the dominant role of translational energy in promoting the dissociation of D₂ on Cu(111) under equilibrium conditions *Faraday Discuss.* **96** 17
- [8] Harris J and Kasemo B 1981 On precursor mechanisms for surface reactions *Surf. Sci.* **105** L281
- [9] Zangwill A 1988 *Physics at Surfaces* (Cambridge: Cambridge University Press)

-20-

- [10] Head-Gordon M, Tully J C, Rettner C T, Mullins C B and Auerbach D J 1991 On the nature of trapping and desorption at high surface temperatures: theory and experiments for the Ar–Pt(111) system *J. Chem. Phys.* **94** 1516
- [11] Spruit M E M, van den Hoek P J, Kuipers E W, Geuzebroek F and Kleyn A W 1989 Direct inelastic scattering of superthermal Ar, CO, NO and O₂ from Ag(111) *Surf. Sci.* **214** 591
- [12] Harris J 1987 Notes on the theory of atom–surface scattering *Phys.scr.* **36** 156
- [13] Harris J 1991 Mechanical energy transfer in particle–surface collisions *Dynamics of Gas–Surface Interactions* ed C T Rettner and M N R Ashfold (London: Royal Society of Chemistry) p 1
- [14] Hurst J E, Becker C A, Cowin J P, Janda K C, Auerbach D J and Wharton L 1979 Observation of direct inelastic scattering in the presence of trapping-desorption scattering: Xe on Pt(111) *Phys. Rev. Lett.* **43** 1175
- [15] Janda K C, Hurst J E, Cowin J P, Warton L and Auerbach D J 1983 Direct-inelastic and trapping-desorption scattering of N₂ and CH₄ from Pt(111) *Surf. Sci.* **130** 395
- [16] Barker J A and Rettner C T 1992 Accurate potential energy surface for Xe/Pt(111): a benchmark gas–surface interaction potential *J. Chem. Phys.* **97** 5844
- [17] Kirchner E J J, Kleyn A W and Baerends E J 1994 A comparative study of Ar/Ag(111) potentials *J. Chem. Phys.* **101** 9155
- [18] Schweizer E K and Rettner C T 1989 Quantum effects in the scattering of argon from 2H-W(100) *Phys. Rev. Lett.* **62** 3085
- [19] Lennard-Jones J E and Devonshire A F 1936 Diffraction and selective adsorption of atoms at crystal surfaces *Nature* **137** 1069
- [20] Andersson S, Wilzén L, Persson M and Harris J 1989 Sticking in the quantum regime: H₂ and D₂ on Cu(100) *Phys. Rev. B* **40** 8146
- [21] Persson M, Wilzén L and Andersson S 1990 Mean free path of a trapped physisorbed hydrogen molecule *Phys. Rev. B* **42** 5331
- [22] Backx C, de Groot C P M and Biloen P 1981 Adsorption of oxygen on Ag(110) studied by high resolution ELS and TPD

- [23] Gravil P A and Bird D M 1996 Chemisorption of O₂ on Ag(110) *Surf. Sci.* **352** 248
- [24] Campbell C T 1985 Atomic and molecular oxygen adsorption on Ag(111) *Surf. Sci.* **157** 43
- [25] Rettner C T 1991 Inelastic scattering of NO from Ag(111): Internal state, angle, and velocity resolved measurements *J. Chem. Phys.* **94** 734
- [26] Lahaye R J W E, Stolte S, Holloway S and Kleyn A W 1996 NO/Pt(111) orientation and energy dependence of scattering *J. Chem. Phys.* **104** 8301
- [27] Heinzmann U, Holloway S, Kleyn A W, Palmer R E and Snowdon K J 1996 Orientation in molecule–surface interactions *J. Phys. Condens. Matter* **8** 3245
- [28] Rettner C T, Kimman J, Fabre F, Auerbach D J and Morawitz H 1987 Direct vibrational excitation in gas–surface collisions of NO with Ag(111) *Surf. Sci.* **192** 107
- [29] Gates G A, Darling G R and Holloway S 1994 A theoretical study of the vibrational excitation of NO/Ag(111) *J. Chem. Phys.* **101** 6281
- [30] Groß A and Brenig W 1993 Vibrational excitation of NO in NO Ag scattering revisited *Surf. Sci.* **289** 335
- [31] Auerbach D J and Rettner C T 1987 Precursor states, myth or reality: a perspective from molecular beam studies *Kinetics of Interface Reactions* ed M Grunze and H J Kreuzer (Berlin: Springer) p 125
- [32] Luntz A C, Grimblot J and Fowler D 1989 Sequential precursors in dissociative chemisorption—O₂ on Pt(111) *Phys. Rev. B* **39** 12 903
-

- [33] Vattuone L, Boragno C, Pupo M, Restelli P, Rocca M and Valbusa U 1994 Azimuthal dependence of sticking probability of O₂ on Ag(110) *Phys. Rev. Lett.* **72** 510
- [34] Doren D J and Tully J C 1991 Dynamics of precursor-mediated chemisorption *J. Chem. Phys.* **94** 8428
- [35] Kang H C and Weinberg W H 1994 Kinetic modelling of surface rate processes *Surf. Sci.* **299/300** 755
- [36] DePristo A E and Kara A 1990 Molecule–surface scattering and reaction dynamics *Adv. Chem. Phys.* **77** 163
- [37] Darling G R and Holloway S 1995 The dissociation of diatomic molecules *Rep. Prog. Phys.* **58** 1595
- [38] Jacobs D C 1995 The role of internal energy and approach geometry in molecule–surface reactive scattering *J. Phys.: Condens. Matter* **7** 1023
- [39] Groß A 1996 Dynamical quantum processes of molecular beams at surfaces—hydrogen on metals *Surf. Sci.* **363** 1
- [40] Groß A 1998 Reactions at surfaces studied by *ab initio* dynamics calculations *Surf. Sci. Rep.* **32** 291
- [41] Kroes G J 1999 Six-dimensional quantum dynamics of dissociative chemisorption of H₂ on metal surfaces *Prog. Surf. Sci.* **60** 1
- [42] Lennard-Jones J E 1932 Processes of adsorption and diffusion on solid surfaces *Trans. Faraday Soc.* **28** 333
- [43] Kasemo B 1996 Charge transfer, electronic quantum processes, and dissociation dynamics in molecule–surface collisions *Surf. Sci.* **363** 22
- [44] Katz G, Zeiri Y and Kosloff R 1999 Non-adiabatic charge transfer process of oxygen on metal surfaces *Surf. Sci.* **425** 1
- [45] Böttcher A, Imbeck R, Morgante A and Ertl G 1990 Nonadiabatic surface reaction: Mechanism of electron emission in the Cs + O₂ system *Phys. Rev. Lett.* **65** 2035
- [46] Polanyi J C and Wong W H 1969 Location of energy barriers. I. Effect on the dynamics of reactions A + BC *J. Chem. Phys.* **51** 1439
- [47] Polanyi J C 1987 Some concepts in reaction dynamics *Science* **236** 680
- [48] Robota H J, Vielhaber W, Lin M C, Segner J and Ertl G 1985 Dynamics of the interaction of H₂ and D₂ with Ni(110) and Ni(111) surfaces *Surf. Sci.* **155** 101
- [49]

Rendulic K D, Anger G and Winkler A 1989 Wide-range nozzle beam adsorption data for the systems H₂/Ni and H₂/Pd (100) *Surf. Sci.* **208** 404

- [50] Hayden B E and Lamont C L A 1989 Coupled translational–vibrational activation in dissociative hydrogen adsorption on Cu(110) *Phys. Rev. Lett.* **63** 1823
- [51] Michelsen H A, Rettner C T and Auerbach D J 1993 The adsorption of hydrogen at copper surfaces: A model system for the study of activated adsorption *Surface Reactions* ed R J Madix (Berlin: Springer) p 123
- [52] Rettner C T, Michelsen H A and Auerbach D J 1995 Quantum-state-specific dynamics of the dissociative adsorption and associative desorption of H₂ at a Cu(111) surface *J. Chem. Phys.* **102** 4625
- [53] Rettner C T, Auerbach D J and Michelsen H A 1992 Observation of direct vibrational-excitation in collisions of H₂ and D₂ with a Cu(111) surface *Phys. Rev. Lett.* **68** 2547
- [54] Hodgson A, Moryl J, Traversaro P and Zhao H 1992 Energy transfer and vibrational effects in the dissociation and scattering of D₂ from Cu(111) *Nature* **356** 501
- [55] Rettner C T, Fabre F, Kimman J and Auerbach D J 1985 Observation of direct vibrational-excitation in gas–surface collisions—NO on Ag(111) *Phys. Rev. Lett.* **55** 1904
-

-22-

- [56] Beauregard J N and Mayne H R 1993 The role of reactant rotation and rotational alignment in the dissociative chemisorption of hydrogen on Ni(100) *Chem. Phys. Lett.* **205** 515
- [57] Darling G R and Holloway S 1993 Rotational effects in the dissociative adsorption of H₂ on Cu(111) *Faraday Discuss. Chem. Soc.* **96** 43
- [58] Darling G R and Holloway S 1994 Rotational motion and the dissociation of H₂ on Cu(111) *J. Chem. Phys.* **101** 3268
- [59] Wahnström G, Lee A B and Strömquist J 1996 Motion of ‘hot’ oxygen adatoms on corrugated metal surfaces *J. Chem. Phys.* **105** 326
- [60] Wang Z S, Darling G R and Holloway S 2000 Translation-to-rotational energy transfer in scattering of H₂ molecules from Cu(111) surfaces *Surf. Sci.* **458** 63
- [61] Somorjai G A 1994 The surface science of heterogeneous catalysis *Surf. Sci.* **299/300** 849
- [62] Bird D M, Clarke L J, Payne M C and Stich I 1993 Dissociation of H₂ on Mg(0001) *Chem. Phys. Lett.* **212** 518
- [63] White J A, Bird D M, Payne M and Stich I 1994 Surface corrugation in the dissociative adsorption of H₂ on Cu(100) *Phys. Rev. Lett.* **73** 1404
- [64] Hammer B, Scheffler M, Jacobsen K W and Nørskov J K 1994 Multidimensional potential energy surface for H₂ dissociation over Cu(111) *Phys. Rev. Lett.* **73** 1400
- [65] Darling G R and Holloway S 1994 The role of parallel momentum in the dissociative adsorption of H₂ at highly corrugated surfaces *Surf. Sci.* **304** L461
- [66] McCormack D A and Kroes G J 1999 A classical study of rotational effects in dissociation of H₂ on Cu(100) *Phys. Chem. Chem. Phys.* **1** 1359
- [67] White J A, Bird D M and Payne M C 1995 Dissociation of H₂ on W(100) *Phys. Rev. B* **53** 1667
- [68] Groß A, Wilke S and Scheffler M 1995 6-dimensional quantum dynamics of adsorption and desorption of H₂ at Pd(100)—steering and steric effects *Phys. Rev. Lett.* **75** 2718
- [69] Kay M, Darling G R, Holloway S, White J A and Bird D M 1995 Steering effects in non-activated adsorption *Chem. Phys. Lett.* **245** 311
- [70] Butler D A, Hayden B E and Jones J D 1994 Precursor dynamics in dissociative hydrogen adsorption on W(100) *Chem. Phys. Lett.* **217** 423
- [71] Butler D A and Hayden B E 1995 The indirect channel to hydrogen dissociation on W(100)c(2 × 2)Cu—evidence for a dynamical precursor *Chem. Phys. Lett.* **232** 542

- [72] Hand M R and Harris J 1990 Recoil effects in surface dissociation *J. Chem. Phys.* **92** 7610
- [73] Guo X-C, Bradley J M, Hopkinson A and King D A 1994 O₂ interaction with Pt{100}-hexR0.7°—scattering, sticking and saturating *Surf. Sci.* **310** 163
- [74] Luntz A C and Harris J 1991 CH₄ dissociation on metals—a quantum dynamics model *Surf. Sci.* **258** 397
- [75] Walker A V and King D A 1999 Dynamics of the dissociative adsorption of methans on Pt(110)-(1 × 2) *Phys.Rev. Lett.* **82** 5156
- [76] Rettner C T, Schweizer E K and Stein H 1990 Dynamics of chemisorption of N₂ on W(100): Precursor-mediated and activated dissociation *J. Chem. Phys.* **93** 1442
-

-23-

- [77] Rettner C T and Mullins C B 1991 Dynamics of the chemisorption of O₂ on Pt(111): Dissociation via direct population of a molecularly chemisorbed precursor at high incidence kinetic energy *J. Chem. Phys.* **94** 1626
- [78] Rettner C T and Stein H 1987 Effect of the translational energy on the chemisorption of N₂ on Fe(111): activated dissociation via a precursor state *Phys. Rev. Lett.* **59** 2768
- [79] Rettner C T and Stein H 1987 Effect of the vibrational energy on the dissociative chemisorption of N₂ on Fe(111) *J. Chem. Phys.* **87** 770
- [80] Luntz A C and Harris J 1992 The role of tunneling in precursor mediated dissociation: Alkanes on metal surfaces *J. Chem. Phys.* **96** 7054
- [81] Elkowitz A B, McCreery J H and Wolken G 1976 Dynamics of atom-adsorbed atom collisions: Hydrogen on tungsten *Chem. Phys.* **17** 423
- [82] Tully J C 1980 Dynamics of gas-surface interactions: reactions of atomic oxygen with adsorbed carbon on platinum *J. Chem. Phys.* **73** 6333
- [83] Weinberg W H 1991 Kinetics of surface reactions *Dynamics of Gas-Surface Interactions* ed C T Rettner and M N R Ashfold (London: Royal Society of Chemistry)
- [84] Kuipers E W, Vardi A, Danon A and Amirav A 1991 Surface-molecule proton transfer—a demonstration of the Eley-Rideal mechanism *Phys.Rev. Lett.* **66** 116
- [85] Rettner C T 1992 Dynamics of the direct reaction of hydrogen atoms adsorbed on Cu(111) with hydrogen atoms incident from the gas phase *Phys.Rev. Lett.* **69** 383
- [86] Lykke K R and Kay B D 1990 State-to-state inelastic and reactive molecular beam scattering from surfaces *Laser Photoionization and Desorption Surface Analysis Techniques* vol 1208, ed N S Nogar (Bellingham, WA: SPIE) p 1218
- [87] Rettner C T and Auerbach D J 1994 Distinguishing the direct and indirect products of a gas-surface reaction *Science* **263** 365
- [88] Rettner C T 1994 Reaction of an H-atom beam with Cl/Au(111)—dynamics of concurrent Eley-Rideal and Langmuir-Hinshelwood mechanisms *J. Chem. Phys.* **101** 1529
- [89] Rettner C T and Auerbach D J 1996 Quantum-state distributions for the HD product of the direct reaction of H(D)/Cu(111) with D(H) incident from the gas phase *J. Chem. Phys.* **104** 2732
- [90] Shalashilin D V, Jackson B and Persson M 1999 Eley-Rideal and hot atom reactions of H(D) atoms with D(H)-covered Cu(111) surfaces; quasiclassical studies *J. Chem. Phys.* **110** 11 038
- [91] Xi M and Bent B E 1992 Evidence for an Eley-Rideal mechanism in the addition of hydrogen atoms to unsaturated hydrocarbons on Cu(111) *J. Vac. Sci. Technol. B* **10** 2440
- [92] Xi M and Bent B E 1993 Reaction of deuterium atoms with cyclohexane on Cu(111)—hydrogen abstraction reactions by Eley-Rideal mechanisms *J. Phys. Chem.* **97** 4167
- [93] Xie J, Mitchell W J, Lyons K J and Weinberg W H 1994 Atomic hydrogen induced decomposition of chemisorbed formate at 100 K on the Ru(001) surface *J. Chem. Phys.* **101** 9195
- [94] Dai E H L and Ho W 1995 *Laser Spectroscopy and Photochemistry at Metal Surfaces* (Singapore: World Scientific)
- [95] Polanyi J C and Rieley H 1991 Photochemistry in the adsorbed state *Dynamics of Gas-Surface Interactions* ed C T Rettner and M N R Ashfold (London: Royal Society of Chemistry) p 329

- [96] Hasselbrink E 1994 State-resolved probes of molecular desorption dynamics induced by short-lived electronic excitations *Laser Spectroscopy and Photochemistry at Metal Surfaces* ed E H L Dai and W Ho (Hong Kong: World Scientific) p 685
- [97] Ho W 1994 Surface photochemistry *Surf. Sci.* **299/300** 996
-

-24-

- [98] Cavanagh R R, King D S, Stephenson J C and Heinz T F 1993 Dynamics of nonthermal reactions—femtosecond surface chemistry *J. Phys. Chem.* **97** 786
- [99] Bonn M, Funk S, Hess C, Denzler D N, Stampfl C, Scheffler M, Wolf M and Ertl G 1999 Phonon versus electron-mediated desorption and oxidation of CO on Ru(001) *Science* **285** 1042
- [100] Michelsen H A, Rettner C T, Auerbach D J and Zare R N 1993 Effect of rotation on the translational and vibrational energy dependence of the dissociative adsorption of D₂ on Cu(111) *J. Chem. Phys.* **98** 8294
- [101] Rettner C T, DeLouise L A and Auerbach D J 1986 Effect of incidence kinetic energy and surface coverage on the dissociative chemisorption of oxygen on W(110) *J. Chem. Phys.* **85** 1131
- [102] Darling G R, Kay M and Holloway S 1998 The steering of molecules in simple dissociation reactions *Surf. Sci.* **400** 314
-

FURTHER READING

Rettner C T and Ashfold M N R 1991 *Dynamics of Gas–Surface Interactions*. (London: Royal Society of Chemistry)

Darling G R and Holloway S 1995 The dissociation of diatomic molecules *Rep. Prog. Phys.* **58** 1595

Rettner C T, Auerbach D J, Tully J C and Kleyn A W 1996 Chemical dynamics at the gas–surface interface *J. Phys. Chem.* **100** 13 201

-1-

A3.10 Reactions on surfaces: corrosion, growth, etching and catalysis

Todd P St Clair and D Wayne Goodman

A3.10.1 INTRODUCTION

The impact of surface reactions on society is often overlooked. How many of us pause to appreciate integrated circuitry before checking email? Yet, without growth and etching reactions, the manufacturing of integrated circuits would be quite impractical. Or consider that in 1996, the United States alone consumed 123 billion gallons of gasoline [1]. The production of this gasoline from crude petroleum is accomplished by the petroleum industry using heterogeneous catalytic reactions. Even the control of automobile exhaust emissions, an obvious environmental concern, is achieved *via* catalytic reactions using ‘three-way catalysts’ that eliminate hydrocarbons, CO and NO_x. The study of these types of surface reactions and others is an exciting and rapidly changing field. Nevertheless, much remains to be understood at the atomic level regarding the interaction of gases and liquids with solid surfaces.

Surface science has thrived in recent years primarily because of its success at providing answers to fundamental questions. One objective of such studies is to elucidate the basic mechanisms that control surface reactions. For example, a goal could be to determine if CO dissociation occurs prior to oxidation over Pt catalysts. A second objective is then to extrapolate this microscopic view of surface reactions to the

corresponding macroscopic phenomena.

How are fundamental aspects of surface reactions studied? The surface science approach uses a simplified system to model the more complicated 'real-world' systems. At the heart of this simplified system is the use of well defined surfaces, typically in the form of oriented single crystals. A thorough description of these surfaces should include composition, electronic structure and geometric structure measurements, as well as an evaluation of reactivity towards different adsorbates. Furthermore, the system should be constructed such that it can be made increasingly more complex to more closely mimic macroscopic systems. However, relating surface science results to the corresponding real-world problems often proves to be a stumbling block because of the sheer complexity of these real-world systems.

Essential to modern surface science techniques is the attainment and maintenance of ultrahigh vacuum (UHV), which corresponds to pressures of the order of 10^{-10} Torr ($\sim 10^{-13}$ atm). At these pressures, the number of collisions between gas phase molecules and a surface are such that a surface can remain relatively contaminant-free for a period of hours. For example, in air at 760 Torr and 298 K the collision frequency is 3×10^{23} collisions $\text{cm}^{-2} \text{s}^{-1}$. Assuming a typical surface has 10^{15} atoms cm^{-2} , then each surface atom undergoes $\sim 10^8$ collisions per second. Clearly, a surface at 760 Torr has little chance of remaining clean. However, by lowering the pressure to 10^{-10} Torr, the collision frequency decreases to approximately 10^{10} collisions $\text{cm}^{-2} \text{s}^{-1}$, corresponding to a collision with a surface atom about every 10^5 s. Decreasing the pressure is obviously a solution to maintaining a clean sample, which itself is crucial to sustaining well characterized surfaces during the course of an experiment.

Modern UHV chambers are constructed from stainless steel. The principal seals are metal-on-metal, thus the use of greases is avoided. A combination of pumps is normally used, including ion pumps, turbomolecular pumps, cryopumps and mechanical (roughing) pumps. The entire system is generally heatable to ~ 500 K. This 'bakeout' for a period of

-2-

10–20 h increases gas desorption rates from the internal surfaces, ultimately resulting in lower pressure. For further reading on vacuum technology, including vacuum and pump theory, see [2, 3].

The importance of low pressures has already been stressed as a criterion for surface science studies. However, it is also a limitation because real-world phenomena do not occur in a controlled vacuum. Instead, they occur at atmospheric pressures or higher, often at elevated temperatures, and in conditions of humidity or even contamination. Hence, a major thrust in surface science has been to modify existing techniques and equipment to permit detailed surface analysis under conditions that are less than ideal. The scanning tunnelling microscope (STM) is a recent addition to the surface science arsenal and has the capability of providing atomic-scale information at ambient pressures and elevated temperatures. Incredible insight into the nature of surface reactions has been achieved by means of the STM and other *in situ* techniques.

This chapter will explore surface reactions at the atomic level. A brief discussion of corrosion reactions is followed by a more detailed look at growth and etching reactions. Finally, catalytic reactions will be considered, with a strong emphasis on the surface science approach to catalysis.

A3.10.2 CORROSION

A3.10.2.1 INTRODUCTION

Corrosion is a frequently encountered phenomenon in which a surface undergoes changes associated with

exposure to a reactive environment. While materials such as plastics and cement can undergo corrosion, the term corrosion more commonly applies to metal surfaces. Rust is perhaps the most widely recognized form of corrosion, resulting from the surface oxidation of an iron-containing material such as steel. Economically, corrosion is extremely important. It has been estimated that annual costs associated with combating and preventing corrosion are 2–3% of the gross national product for industrialized countries. Equipment damage is a major component of the costs associated with corrosion. There are also costs related to corrosion prevention, such as implementation of anti-corrosive paints or other protective measures. Finally, there are indirect losses, such as plant shutdowns, when equipment or facilities need repair or replacement.

Most metals tend to corrode in an environment of air and/or water, forming metal oxides or hydrated oxides. Whether or not such a reaction is possible is dictated by the thermodynamics of the corrosion reaction. If the reaction has a negative Gibbs free energy of formation, then the reaction is thermodynamically favoured. While thermodynamics determines whether a particular reaction can occur or not, the *rate* of the corrosion reaction is determined by kinetic factors. A number of variables can affect the corrosion rate, including temperature, pH and passivation, which is the formation of a thin protective film on a metal surface. Passivation can have a tremendous influence on the corrosion rate, often reducing it to a negligible amount.

Since metals have very high conductivities, metal corrosion is usually electrochemical in nature. The term electrochemical is meant to imply the presence of an electrode process, i.e. a reaction in which free electrons participate. For metals, electrochemical corrosion can occur by loss of metal atoms through anodic dissolution, one of the fundamental corrosion reactions. As an example, consider a piece of zinc, hereafter referred to as an electrode, immersed in water. Zinc tends to dissolve in water, setting up a concentration of Zn^{2+} ions very near the electrode

-3-

surface. The term *anodic dissolution* arises because the area of the surface where zinc is *dissolving* to form Zn^{2+} is called the *anode*, as it is the source of positive current in the system. Because zinc is oxidized, a concentration of electrons builds up on the electrode surface, giving it a negative charge. This combination of negatively charged surface region with positively charged near-surface region is called an electrochemical double layer. The potential across the layer, called the electrode potential, can be as much as ± 1 V.

In moist environments, water is present either at the metal interface in the form of a thin film (perhaps due to condensation) or as a bulk phase. Figure A3.10.1 schematically illustrates another example of anodic dissolution where a droplet of slightly acidic water (for instance, due to H_2SO_4) is in contact with an Fe surface in air [4]. Because Fe is a conductor, electrons are available to reduce O_2 at the edges of the droplets. The electrons are then replaced by the oxidation reaction of Fe to Fe^{2+} (forming $FeSO_4$ if H_2SO_4 is the acid), and the rate of corrosion is simply the current induced by metal ions leaving the surface.

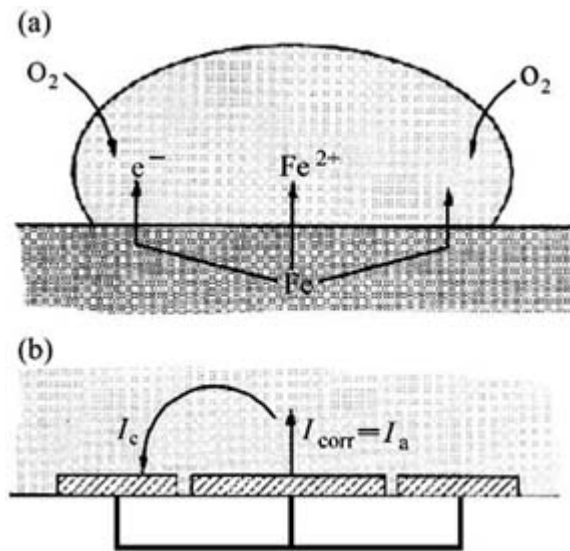


Figure A3.10.1 (a) A schematic illustration of the corrosion process for an oxygen-rich water droplet on an iron surface. (b) The process can be viewed as a short-circuited electrochemical cell [4].

Corrosion protection of metals can take many forms, one of which is passivation. As mentioned above, passivation is the formation of a thin protective film (most commonly oxide or hydrated oxide) on a metallic surface. Certain metals that are prone to passivation will form a thin oxide film that displaces the electrode potential of the metal by +0.5–2.0 V. The film severely hinders the diffusion rate of metal ions from the electrode to the solid–gas or solid–liquid interface, thus providing corrosion resistance. This decreased corrosion rate is best illustrated by anodic polarization curves, which are constructed by measuring the net current from an electrode into solution (the corrosion current) under an applied voltage. For passivable metals, the current will increase steadily with increasing voltage in the so-called active region until the passivating film forms, at which point the current will rapidly decrease. This behaviour is characteristic of metals that are susceptible to passivation.

Another method by which metals can be protected from corrosion is called alloying. An alloy is a multi-component solid solution whose physical and chemical properties can be tailored by varying the alloy composition.

-4-

For example, copper has relatively good corrosion resistance under non-oxidizing conditions. It can be alloyed with zinc to yield a stronger material (brass), but with lowered corrosion resistance. However, by alloying copper with a passivating metal such as nickel, both mechanical and corrosion properties are improved. Another important alloy is steel, which is an alloy between iron (>50%) and other alloying elements such as carbon.

Although alloying can improve corrosion resistance, brass and steel are not completely resistant to attack and often undergo a form of corrosion known as selective corrosion (also called de-alloying or leaching). De-alloying consists of the segregation of one alloy component to the surface, followed by the removal of this surface component through a corrosion reaction. De-zincification is the selective leaching of zinc from brasses in an aqueous solution. The consequences of leaching are that mechanical and chemical properties change with compositional changes in the alloy.

As an example of the effect that corrosion can have on commercial industries, consider the corrosive effects of salt water on a seagoing vessel. Corrosion can drastically affect a ship's performance and fuel consumption over a period of time. As the hull of a steel boat becomes corroded and fouled by marine growths, the

performance of the ship declines because of increased frictional drag. Therefore, ships are drydocked periodically to restore the smoothness of the hull. Figure A3.10.2 shows the loss of speed due to corrosion and marine fouling between annual drydockings for a ship with a steel hull [5]. As corrosion effects progressively deteriorated the hull and as marine growth accumulated, the ship experienced an overall loss of speed even after drydocking and an increased fuel consumption over time. It is clear that there is strong economic motivation to implement corrosion protection.

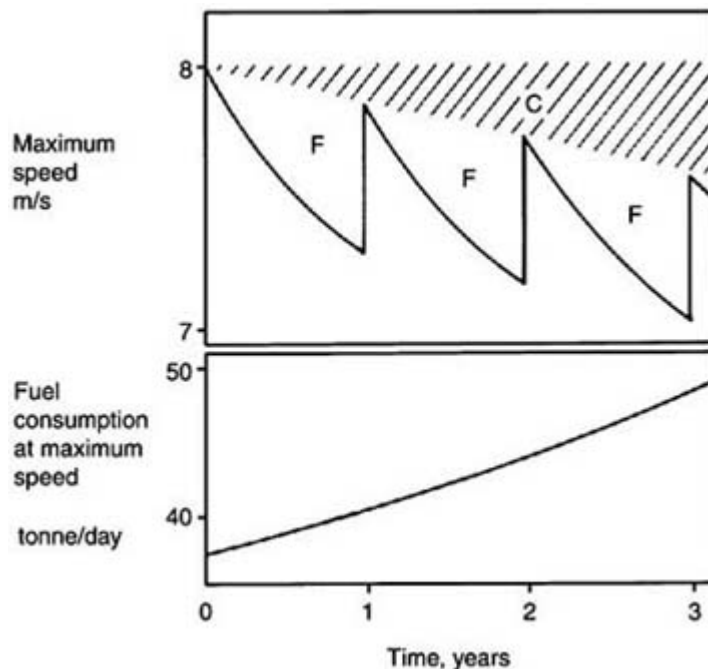


Figure A3.10.2 The influence of corrosion (C) and marine fouling (F) on the performance of a steel ship drydocked annually for cleaning and painting [5].

Surface science studies of corrosion phenomena are excellent examples of *in situ* characterization of surface reactions. In particular, the investigation of corrosion reactions with STM is promising because not only can it be used to study solid–gas interfaces, but also solid–liquid interfaces.

A3.10.2.2 SURFACE SCIENCE OF CORROSION

(A) THE ROLE OF SULFUR IN CORROSION

STM has been used to study adsorption on surfaces as it relates to corrosion phenomena [6, 7]. Sulfur is a well known corrosion agent and is often found in air (SO_2 , H_2S) and in aqueous solution as dissolved anions (HSO_3^-) or dissolved gas (H_2S). By studying the interaction of sulfur with surfaces, insights can be gained into the fundamental processes governing corrosion phenomena. A Ni(111) sample with 10 ppm sulfur bulk impurity was used to study sulfur adsorption by annealing the crystal to segregate the sulfur to the surface [8]. Figure A3.10.3 shows a STM image of a S-covered Ni(111) surface. It was found that sulfur formed islands preferentially near step edges, and that the Ni surface reconstructed under the influence of sulfur adsorption. This reconstruction results in surface sites that have fourfold symmetry rather than threefold symmetry as on the unreconstructed (111) surface. Furthermore, the fourfold symmetry sites are similar to those found on unreconstructed Ni(100), demonstrating the strong influence that sulfur adsorption has on this surface. The mechanism by which sulfur leads to corrosion of nickel surfaces is clearly linked to the ability of sulfur to weaken Ni–Ni bonds.

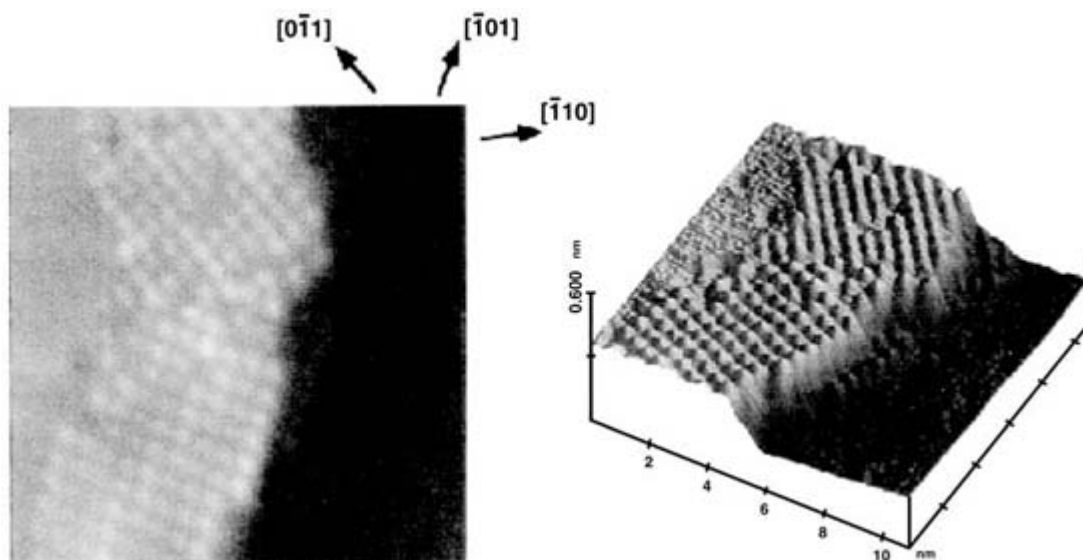


Figure A3.10.3 STM images of the early stages of sulfur segregation on Ni(111). Sulfur atoms are seen to preferentially nucleate at step edges [8].

-6-

(B) ANODIC DISSOLUTION IN ALLOYS

This weakening of Ni–Ni surface bonds by adsorbed sulfur might lead one to expect that the corrosion rate should increase in this case. In fact, an increased anodic dissolution rate was observed for Ni₃Fe (100) in 0.05 M H₂SO₄ [9]. Figure A3.10.4 shows the anodic polarization curves for clean and S-covered single-crystal alloy surfaces. While both surfaces show the expected current increase with potential increase, the sulfur-covered surface clearly has an increased rate of dissolution. In addition, the sulfur coverage (measured using radioactive sulfur, ³⁵S) does not decrease even at the maximum dissolution rate, indicating that adsorbed sulfur is not consumed by the dissolution reaction. Instead, surface sulfur simply enhances the rate of dissolution, as expected based on the observation above that Ni–Ni bonds are significantly weakened by surface sulfur.

Figure A3.10.4 The effect of sulfur on the anodic polarization curves from a Ni_{0.25}Fe(100) alloy in 0.05 M H₂SO₄. θ is the sulfur (³⁵S) coverage [6].

The nature of copper dissolution from CuAu alloys has also been studied. CuAu alloys have been shown to have a surface Au enrichment that actually forms a protective Au layer on the surface. The anodic polarization curve for CuAu alloys is characterized by a critical potential, E_c , above which extensive Cu dissolution is observed [10]. Below E_c , a smaller dissolution current arises that is approximately potential-independent. This critical potential depends not only on the alloy composition, but also on the solution composition. STM was used to investigate the mechanism by which copper is selectively dissolved from a CuAu₃ electrode in solution [11], both above and below the critical potential. At potentials below E_c , it was found that, as copper dissolves, vacancies agglomerate on the surface to form voids one atom deep. These voids grow two-dimensionally with increasing Cu dissolution while the second atomic layer remains undisturbed. The fact that the second atomic layer is unchanged suggests that Au atoms from the first layer are filling

-7-

in holes left by Cu dissolution. In sharp contrast, for potentials above E_c , massive Cu dissolution results in a rough surface with voids that grow both parallel and perpendicular to the surface, suggesting a very fast dissolution process. These *in situ* STM observations lend insight into the mechanism by which Cu dissolution occurs in CuAu₃ alloys.

The characterization of surfaces undergoing corrosion phenomena at liquid–solid and gas–solid interfaces remains a challenging task. The use of STM for *in situ* studies of corrosion reactions will continue to shape the atomic-level understanding of such surface reactions.

A3.10.3 GROWTH

A3.10.3.1 INTRODUCTION

Thin crystalline films, or overlayers, deposited onto crystalline substrates can grow in such a way that the substrate lattice influences the overlayer lattice. This phenomenon is known as *epitaxy*; if the deposited material is different from (the same as) the substrate, the process is referred to as heteroepitaxy (homoepitaxy). Epitaxial growth is of interest for several reasons. First, it is used prevalently in the semiconductor industry for the manufacture of III/V and II/VI semiconductor devices. Second, novel phases have been grown epitaxially by exploiting such phenomena as lattice mismatch and strain. These new phases have physical and chemical properties of interest to science and engineering. Finally, fundamental catalytic studies often focus on modelling oxide-supported metal particles by depositing metal films on oxide single crystals and thin films and, in many cases, these oxide and metal films grow epitaxially.

When considering whether growth will occur epitaxially or not, arguments can be made based on geometrical considerations, or row matching. This concept is based on the idea that the overlayer must sit on minima of the substrate corrugation potential to minimize the interaction energy. For example, consider the illustration of epitaxial growth in [figure A3.10.5](#) where an fcc(111) monolayer has been overlaid on a bcc(110) surface [12]. [Figure A3.10.5\(a\)](#) shows that the overlayer must be expanded or contracted in two directions to obtain row matching. [Figure A3.10.5\(b\)](#) shows, however, that rotation of the overlayer by 5.26° results in row matching along the most close-packed row of the lattices. Epitaxial growth clearly provides a pathway to energetically favourable atomic arrangements.

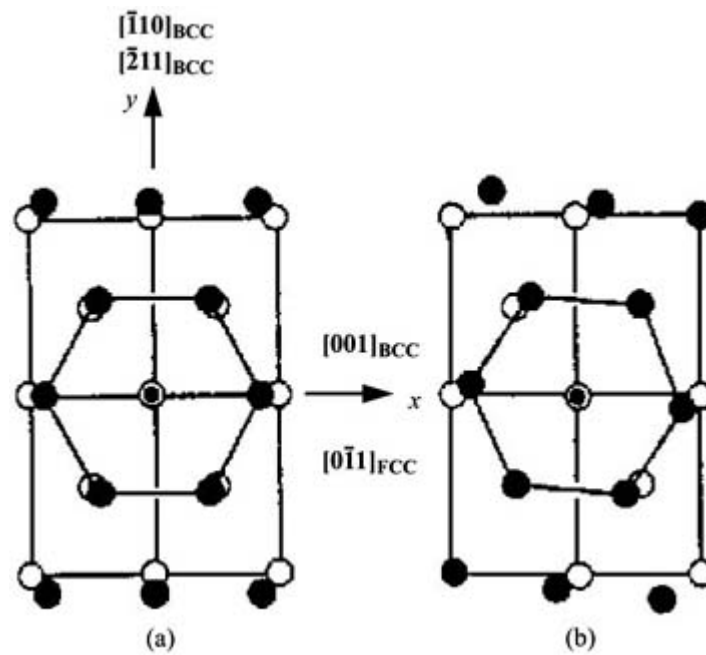


Figure A3.10.5 An fcc(111) monolayer (full circles) overlaid onto a bcc(110) substrate (open circles). (a) fcc [011] parallel to bcc[001]. (b) 5.26° rotation relative to (a). The lattice constants were chosen to produce row-matching in (b) [12].

The influence of the substrate lattice makes it energetically favourable for two materials to align lattices. On the other hand, if two lattices are misaligned or mismatched in some other way, then lattice strain may result. This lattice strain can lead to a metastable atomic arrangement of the deposited material. In other words, an overlayer can respond to lattice strain by adopting a crystal structure that differs from its normal bulk structure in order to row-match the substrate lattice. This phenomenon is known as pseudomorphy. For example, Cu (fcc) deposited on a Pd(100) surface will grow epitaxially to yield a pseudomorphic fcc overlayer [13]. However, upon increasing the copper film thickness, a body-centred tetragonal (bct) metastable phase, one not normally encountered for bulk copper, was observed. This phase transformation is due to a high degree of strain in the fcc overlayer.

Another example of epitaxy is tin growth on the (100) surfaces of InSb or CdTe ($a = 6.49 \text{ \AA}$) [14]. At room temperature, elemental tin is metallic and adopts a bct crystal structure ('white tin') with a lattice constant of 5.83 \AA . However, upon deposition on either of the two above-mentioned surfaces, tin is transformed into the diamond structure ('grey tin') with $a = 6.49 \text{ \AA}$ and essentially no misfit at the interface. Furthermore, since grey tin is a semiconductor, then a novel heterojunction material can be fabricated. It is evident that epitaxial growth can be exploited to synthesize materials with novel physical and chemical properties.

A3.10.3.2 FILM GROWTH TECHNIQUES

There are several design parameters which distinguish film growth techniques from one another, namely generation of the source atom/molecule, delivery to the surface and the surface condition. The source molecule can be generated in a number of ways including vapour produced thermally from solid and liquid sources, decomposition of organometallic

compounds and precipitation from the liquid phase. Depending on the pressures used, gas phase atoms and

molecules impinging on the surface may be in viscous flow or molecular flow. This parameter is important to determining whether atom–atom (molecule–molecule) collisions, which occur in large numbers at pressures higher than UHV, can affect the integrity of the atom (molecule) to be deposited. The condition of the substrate surface may also be a concern: elevating the surface temperature may alter the growth kinetics, or the surface may have to be nearly free of defects and/or contamination to promote the proper growth mode. Two film growth techniques, molecular beam epitaxy (MBE) and vapour phase epitaxy (VPE) will be briefly summarized below. These particular techniques were chosen because of their relevance to UHV studies. The reader is referred elsewhere for more detailed discussions of the various growth techniques [15, 16 and 17].

MBE is accomplished under UHV conditions with pressures of the order of $\sim 10^{-10}$ Torr. By using such low pressures, the substrate surface and deposited thin films can be kept nearly free of contamination. In MBE, the material being deposited is usually generated in UHV by heating the source material to the point of evaporation or sublimation. The gas phase species is then focused in a molecular beam onto the substrate surface, which itself may be at an elevated temperature. The species flux emanating from the source can be controlled by varying the source temperature and the species flux arriving at the surface can be controlled by the use of mechanical shutters. Precise control of the arrival of species at the surface is a very important characteristic of MBE because it allows the growth of epitaxial films with very abrupt interfaces. Several sources can be incorporated into a single vacuum chamber, allowing doped semiconductors, compounds or alloys to be grown. For instance, MBE is used prevalently in the semiconductor industry to grow GaAs/Al_xGa_{1-x}As layers and, in such a situation, a growth chamber would be outfitted with Ga, As and Al deposition sources. Because of the compatibility of MBE with UHV surface science techniques, it is often the choice of researchers studying fundamentals of thin-film growth.

A second technique, VPE, is also used for surface science studies of overlayer growth. In VPE, the species being deposited can be generated in several ways, including vaporization of a liquid precursor into a flowing gas stream or sublimation of a solid precursor. VPE generates an unfocused vapour or cloud of the deposited material, rather than a collimated beam as in MBE. Historically, VPE played a major role in the development of III/V semiconductors. Currently, VPE is used as a tool for studying metal growth on oxides, an issue of importance to the catalysis community.

The following two sections will focus on epitaxial growth from a surface science perspective with the aim of revealing the fundamentals of thin-film growth. As will be discussed below, surface science studies of thin-film deposition have contributed greatly to an atomic-level understanding of nucleation and growth.

A3.10.3.3 THERMODYNAMICS

The number of factors affecting thin-film growth is largely dependent upon the choice of growth technique. The overall growth mechanism may be strongly influenced by three factors: mass transport, thermodynamics and kinetics. For instance, for an exothermic (endothermic) process, increasing (decreasing) the surface temperature will decrease (increase) the growth rate for a thermodynamically limited process. On the other hand, if temperature has no effect on the growth rate, then the process may be limited by mass transport, which has very little dependence on the substrate temperature. Another test of mass transport limitations is to increase the total flow rate to the surface while keeping the partial pressures constant—if the growth rate is influenced, then mass transport limitations should be considered. Alternatively, if the substrate orientation is found to influence the growth rates, then the process is very likely kinetically limited. Thus, through a relatively straightforward analysis of the parameters affecting macroscopic

quantities, such as growth rate, a qualitative description of the growth mechanism can be obtained. The

growth of epitaxial thin films by vapour deposition in UHV is a non-equilibrium kinetic phenomenon. At thermodynamic equilibrium, atomic processes are required to proceed in opposite directions at equal rates. Hence, a system at equilibrium must have equal adsorption and desorption rates, as well as equal cluster growth and cluster decay rates. If growth were occurring under equilibrium conditions, then there would be no net change in the amount of deposited material on the surface. Typical growth conditions result in systems far from equilibrium, so film growth is usually limited by kinetics considerations. Thermodynamics does play an important role, however, as will be discussed next.

Thermodynamics can lend insight into the expected growth mode by examination of energetics considerations. The energies of importance are the surface free energy of the overlayer, the interfacial energy between the substrate and the overlayer, and the surface free energy of the substrate. Generally, if the free energy of the overlayer plus the interface energy is greater than the free energy of the substrate, then Frank–van der Merwe (FM) growth will occur [18]. FM growth, also known as layer-by-layer growth, is characterized by the completion of a surface overlayer before the second layer begins forming. However, if the free energy of the overlayer plus the interface energy is less than the free energy of the substrate then the growth mode is Volmer–Weber (VW) [18]. VW, or three-dimensional (3D), growth yields 3D islands or clusters that coexist with bare patches of substrate. There is also a third growth mode, called Stranski–Krastanov (SK), which can be described as one or two monolayers of growth across the entire surface subsequently followed by the growth of 3D islands [18]. In SK growth, the sum of the surface free energy of the overlayer plus interface energy is initially greater than that of the substrate, resulting in the completion of the first monolayer, after which the surface free energy of the overlayer plus interface energy becomes greater than that of the substrate, resulting in 3D growth. It should be stressed that the energetic arguments for these growth modes are only valid for equilibrium processes. However, these descriptions provide good models for the growth modes experimentally observed even under non-equilibrium conditions.

A3.10.3.4 NUCLEATION AND GROWTH

The process of thin-film growth from an atomic point of view consists of the following stages: adsorption, diffusion, nucleation, growth and coarsening. Adsorption is initiated by exposing the substrate surface to the deposition source. As described above, this is a non-equilibrium process, and the system attempts to restore equilibrium by forming aggregates. The adatoms randomly walk during the diffusion process until two or more collide and subsequently nucleate to form a small cluster. A rate-limiting step is the formation of some critical cluster size, at which point cluster growth becomes more probable than cluster decay. The clusters increase in size during the growth stage, with the further addition of adatoms leading to island formation. Growth proceeds at this stage according to whichever growth mode is favoured. Once deposition has ceased, further island morphological changes occur during the coarsening stage, whereby atoms in small islands evaporate and add to other islands or adsorb onto available high-energy adsorption sites such as step edge sites. For an excellent review on the atomic view of epitaxial metal growth, see [19].

Experimentally, the variable-temperature STM has enabled great strides to be made towards understanding nucleation and growth kinetics on surfaces. The evolution of overlayer growth can be followed using STM from the first stages of adatom nucleation through the final stages of island formation. The variable-temperature STM has also been crucial to obtaining surface diffusion rates. In such cases, however, the importance of tip–sample interactions must be considered. Typically, low tunnelling currents are best because under these conditions the tip is further from the surface, thereby reducing the risk of tip–sample interactions.

Much effort in recent years has been aimed at modelling nucleation at surfaces and several excellent reviews exist [20, 21 and 22]. Mean-field nucleation theory is one of these models and has a simple picture at its core.

In the nucleation stage, an atom arriving at the surface from the gas phase adsorbs and then diffuses at a particular rate until it collides with another surface adatom to form a dimer. If the dimers are assumed to be stable (so that no decay occurs) and immobile (so that no diffusion occurs) then, as deposition proceeds, the concentration of dimers will increase approximately linearly until it is roughly equal to the concentration of monomers. At this point, the probability of an atom colliding with a dimer is comparable to the probability of an adatom colliding with another adatom, hence growth and nucleation compete. Once the island density has saturated, i.e. no more clusters are being formed, then the adatom mean free path is equal to the mean island separation and further deposition results in island growth. At coverages near 0.5 monolayers (ML), islands begin to coalesce and the island density decreases.

This simple and idealistic picture of nucleation and growth from mean field nucleation theory was found to be highly descriptive of the Ag/Pt(111) system at 75 K (figure A3.10.6) [23]. Figure A3.10.6 shows a series of STM images of increasing Ag coverage on Pt(111) and demonstrates the transition from nucleation to growth. At very low coverages ((a) and (b)), the average cluster size is 2.4 and 2.6 atoms, respectively, indicating that dimers and trimers are the predominant surface species. However, when the coverage was more than doubled from (a) to (b), the mean island size remained relatively constant. This result clearly indicates that deposition at these low coverages is occurring in the nucleation regime. By increasing the coverage to 0.03 ML, the Ag mean island size doubled to 6.4 atoms and the island density increased, indicating that nucleation and growth were competing. Finally, after increasing the coverage even further (d), the mean island size doubled again, while the island density saturated, suggesting that a pure growth regime dominated, with little or no nucleation occurring.

Growth reactions at surfaces will certainly continue to be the focus of much research. In particular, the synthesis of novel materials is an exciting field that holds much promise for the nanoscale engineering of materials. Undoubtedly, the advent of STM as a means of investigating growth reactions on the atomic scale will influence the future of nanoscale technology.

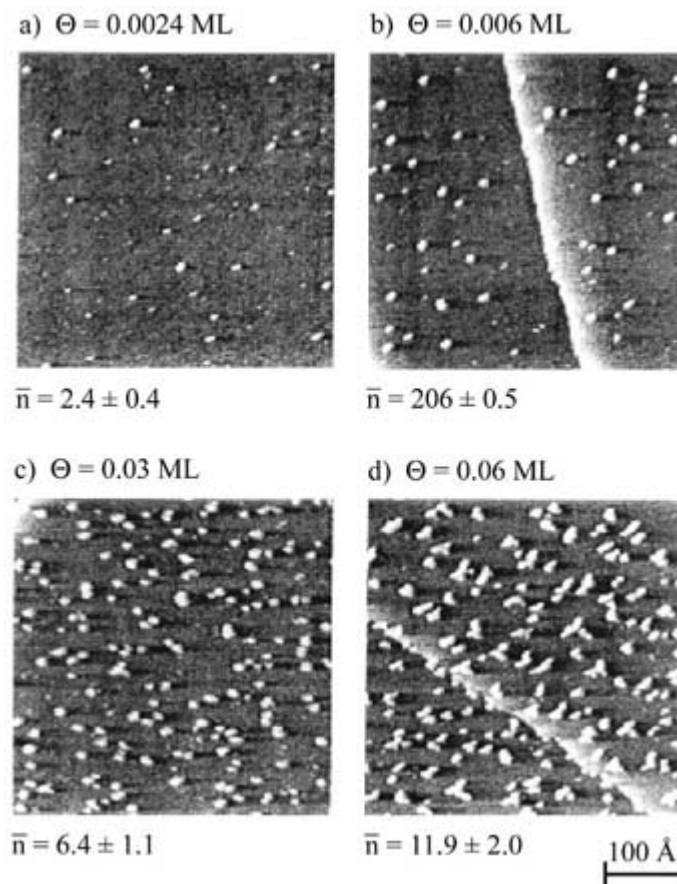


Figure A3.10.6 A series of STM images for Ag/Pt(111) at 75 K showing the transition from nucleation to growth [23]. Coverages (Θ) and mean island sizes (\bar{n}) are indicated.

A3.10.4 ETCHING

A3.10.4.1 INTRODUCTION

Etching is a process by which material is removed from a surface. The general idea behind etching is that by interaction of an etch atom or molecule with a surface, a surface species can be formed that is easily removed. The use of a liquid to etch a surface is known as *wet etching*, while the use of a gas to etch a surface is known as *dry etching*. Wet etching has been employed since the late Middle Ages. The process then was rather simple and could be typified as follows. The metal to be etched was first coated with a wax, or in modern vernacular, a mask. Next, a pattern was cut into the wax to reveal the metal surface beneath. Then, an acid was used to etch the exposed metal, resulting in a patterned surface. Finally, the mask was removed to reveal the finished product. Modern methods are considerably more technologically advanced, although the general principles behind etching remain unchanged.

Both wet and dry etching are used extensively in the semiconductor processing industry. However, wet etching has limitations that prevent it being used to generate micron or submicron pattern sizes for GaAs etching. The most serious of these limitations is called substrate undercutting, which is a phenomenon where etch rates parallel and perpendicular to the surface are approximately equal (isotropic etching). Substrate

undercutting is much less prevalent for silicon surfaces than GaAs surfaces, thus wet etching is more commonly used to etch silicon surfaces. Generally, when patterning surfaces, anisotropic etching is preferred, where etch rates perpendicular to the surface exceed etch rates parallel to the surface. Hence, in cases of undercutting, an ill defined pattern typically results. In the early 1970s, dry etching (with CF_4/O_2 , for example) became widely used for patterning. Dry methods have a distinct advantage over wet methods, namely anisotropic etching.

A form of anisotropic etching that is of some importance is that of orientation-dependent etching, where one particular crystal face is etched at a faster rate than another crystal face. A commonly used orientation-dependent wet etch for silicon surfaces is a mixture of KOH in water and isopropanol. At approximately 350 K, this etchant has an etch rate of $0.6 \mu\text{m min}^{-1}$ for the Si(100) plane, $0.1 \mu\text{m min}^{-1}$ for the Si(110) plane and $0.006 \mu\text{m min}^{-1}$ for the Si(111) plane [24]. These different etch rates can be exploited to yield anisotropically etched surfaces.

Semiconductor processing consists of a number of complex steps, of which etching is an integral step. [Figure A3.10.7](#) shows an example of the use of etching [25] in which the goal of this particular process is to remove certain parts of a film, while leaving the rest in a surface pattern to serve as, for example, interconnection paths. This figure illustrates schematically how etching paired with a technique called photolithography can be used to manufacture a semiconductor device. In this example, the substrate enters the manufacturing stream covered with a film (for example, a SiO_2 film on a Si wafer). A liquid thin-film called a photoresist (denoted 'positive resist' or 'negative resist', as explained below) is first placed on the wafer, which is then spun at several thousand rotations per minute to spread out the film and achieve a uniform coating. Next, the wafer is exposed through a mask plate to an ultraviolet (UV) light source. The UV photons soften certain resists (positive resists) and harden others (negative resists). Next, a developer solution is used to remove the susceptible area, leaving behind the remainder according to the mask pattern. Then, the wafer is etched to remove all of the surface film not protected by the photoresist. Finally, the remaining photoresist is removed, revealing a surface with a patterned film. Thus the role of etching in semiconductor processing is vital and it is evident that motivation exists to explore etching reactions on a fundamental level.

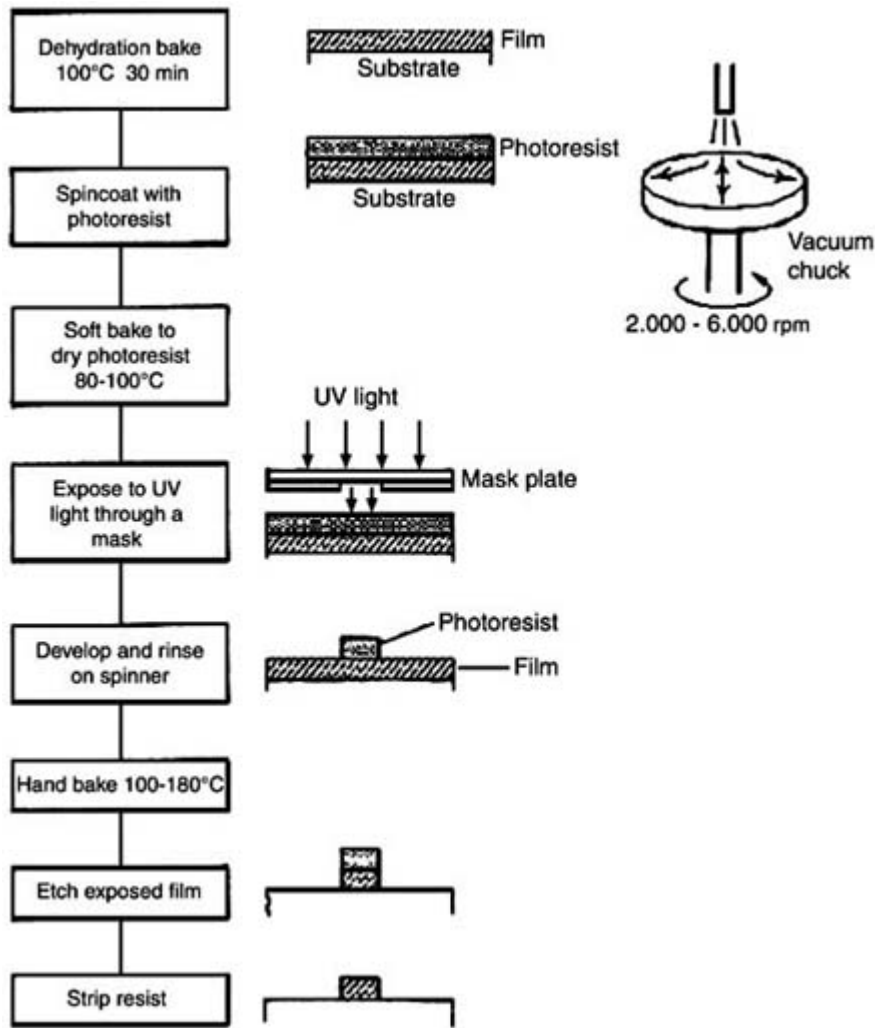


Figure A3.10.7 The role of etching in photolithography [25].

A3.10.4.2 DRY ETCHING TECHNIQUES

It has already been mentioned that dry etching involves the interaction of gas phase molecules/atoms with a surface. More specifically, dry etching utilizes either plasmas that generate reactive species, or energetic ion beams to etch surfaces. Dry etching is particularly important to GaAs processing because, unlike silicon, there are no wet etching methods that result in negligible undercutting. Dry etching techniques can be characterized by either chemical or physical etching mechanisms. The chemical mechanisms tend to be more selective, i.e. more anisotropic, and tend to depend strongly on the specific material being etched. Several dry etch techniques will be briefly discussed below. For a more comprehensive description of these and other techniques, the reader is referred to the texts by Williams [26] or Sugawara [27].

Ion milling is a dry etch technique that uses a physical etching mechanism. In ion milling, ions of an inert gas are generated and then accelerated to impinge on a surface. The etching mechanism is simply the bombardment of these energetic ions on the surface, resulting in erosion. The energy of the ions can be controlled by varying the accelerating voltage, and it may be possible to change the selectivity by varying the angle of incidence.

Plasma etching is a term used to describe any dry etching process that utilizes reactive species generated from a gas plasma. For semiconductor processing, a low-pressure plasma, also called a glow discharge, is used. The glow discharge is characterized by pressures in the range 0.1–5 Torr and electron energies of 1–10 eV. The simplest type of plasma reactor consists of two parallel plates in a vacuum chamber filled with a gas at low pressure. A radio frequency (RF) voltage is applied between the two plates, generating plasma that emits a characteristic glow. Reactive radicals are produced by the plasma, resulting in a collection of gas phase species that are the products of collisions between photons, electrons, ions and atoms or molecules. These chemically reactive species can then collide with a nearby surface and react to form a volatile surface species, thereby etching the surface.

Reactive ion etching (RIE) is distinguished from plasma etching by the fact that the surface reactions are enhanced by the kinetic energy of the incoming reactive species. This type of chemical mechanism is referred to as a kinetically assisted chemical reaction, and very often results in highly anisotropic etching. RIE is typically performed at low pressures (0.01–0.1 Torr) and is used industrially to etch holes in GaAs.

Dry etching is a commonly used technique for creating highly anisotropic, patterned surfaces. The interaction of gas phase etchants with surfaces is of fundamental interest to understanding such phenomena as undercutting and the dependence of etch rate on surface structure. Many surface science studies aim to understand these interactions at an atomic level, and the next section will explore what is known about the etching of silicon surfaces.

A3.10.4.3 ATOMIC VIEW OF ETCHING

On the atomic level, etching is composed of several steps: diffusion of the etch molecules to the surface, adsorption to the surface, subsequent reaction with the surface and, finally, removal of the reaction products. The third step, that of reaction between the etchant and the surface, is of considerable interest to the understanding of surface reactions on an atomic scale. In recent years, STM has given considerable insight into the nature of etching reactions at surfaces. The following discussion will focus on the etching of silicon surfaces [28].

Figure A3.10.8 schematically depicts a Si(100) surface (a) being etched to yield a rough surface (b) and a more regular surface (c). The surfaces shown here are seen to consist of steps, terraces and kinks, and clearly have a three-dimensional character, rather than the two-dimensional character of an ideally flat, smooth surface. The general etching mechanism is based on the use of halogen molecules, the principal etchants used in dry etching. Upon adsorption on silicon at room temperature, Br₂ dissociates to form bromine atoms, which react with surface silicon atoms. Then, if an external source of energy is provided, for example by heating Si (100) to 900 K, SiBr₂ forms and desorbs, revealing the silicon atom(s) beneath and completing the etching process. Depending upon the relative desorption energies from various surface sites, the surface could be etched quite differently, as seen in figure A3.10.8(b) and figure A3.10.8(c).

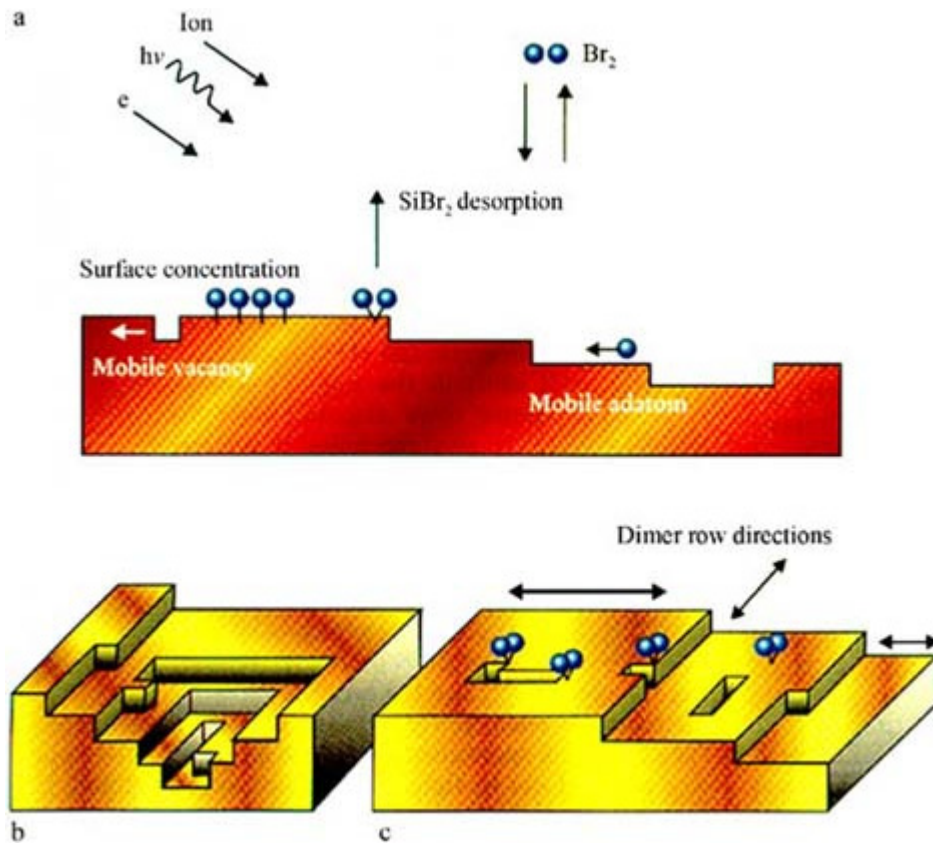


Figure A3.10.8 Depiction of etching on a Si(100) surface. (a) A surface exposed to Br_2 as well as electrons, ions and photons. Following etching, the surface either becomes highly anisotropic with deep etch pits (b), or more regular (c), depending on the relative desorption energies for different surface sites [28].

Semiconductors such as silicon often undergo rearrangements, or reconstructions, at surface boundaries to lower their surface free energy. One way of lowering the surface free energy is the reduction of dangling bonds, which are non-bonding orbitals that extend (dangle) into the vacuum. Si(111) undergoes a complex (7×7) reconstruction that was ultimately solved using STM. [Figure A3.10.9\(a\)](#) shows an STM image of the reconstructed Si(111) surface [29]. This reconstruction reduces the number of dangling bonds from 49 to 19 per unit cell.

The (7×7) reconstruction also affects the second atomic layer, called the rest layer. The rest layer is composed of silicon atoms arranged in triangular arrays that are separated from one another by rows of silicon dimers. [Figure A3.10.9\(b\)](#) shows the exposed rest layer following bromine etching at 675 K [29]. It is noteworthy that the rest layer does not reconstruct to form a new (7×7) surface. The stability of the rest layer following etching of (7×7)-Si(111) is due to the unique role of the halogen. The silicon adlayer is removed by insertion of bromine atoms into Si–Si dimer bonds. Once this silicon adlayer is gone, the halogen stabilizes the silicon rest layer by reacting with the dangling bonds, effectively inhibiting surface reconstruction to a (7×7) phase. Unfortunately, the exposure of the rest layer makes etching more difficult because to form SiBr_2 , bromine atoms must insert into stronger Si–Si bonds.

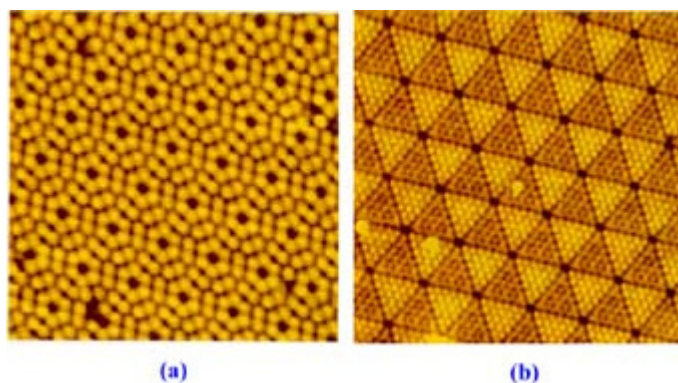


Figure A3.10.9 STM images of Si(111) surfaces before (a) and after (b) etching by bromine at 675 K. In (a) the (7×7) reconstructed surface is seen. In (b), the rest layer consisting of triangular arrays of Si atoms has been exposed by etching [28]. Both images show a $17 \times 17 \text{ nm}^2$ area.

Si(100) reconstructs as well, yielding a (1×2) surface phase that is formed when adjacent silicon atoms bond through their respective dangling bonds to form a more stable silicon dimer. This reconstructed bonding results in a buckling of the surface atoms. Furthermore, because Si–Si dimer bonds are weaker than bulk silicon bonds, the reconstruction actually facilitates etching. For a comprehensive discussion on STM studies of reconstructed silicon surfaces, see [30].

Si(100) is also etched by Br_2 , although in a more dramatic fashion. [Figure A3.10.10](#) shows a STM image of a Si(100) surface after etching at 800 K [28]. In this figure, the dark areas are etch pits one atomic layer deep. The bright rows running perpendicular to these pits are silicon dimer chains, which are composed of silicon atoms that were released from terraces and step edges during etching. The mechanism by which Si(100) is etched has been deduced from STM studies. After Br_2 dissociatively adsorbs to the surface, a bromine atom bonds to each silicon atom in the dimer pairs. SiBr_2 is the known desorption product and so the logical next step is the formation of a surface SiBr_2 species. This step can occur by the breaking of the Si–Si dimer bond and the transfer of a bromine atom from one of the dimer atoms to the other. Then, if enough energy is available to overcome the desorption barrier, SiBr_2 will desorb, leaving behind a highly uncoordinated silicon atom that will migrate to a terrace and eventually re-dimerize. On the other hand, if there is not enough energy to desorb SiBr_2 , then the Br atom would transfer back to the original silicon atom, and a silicon dimer bond would again be formed. In this scenario, SiBr_2 desorption is essential to the etching process.

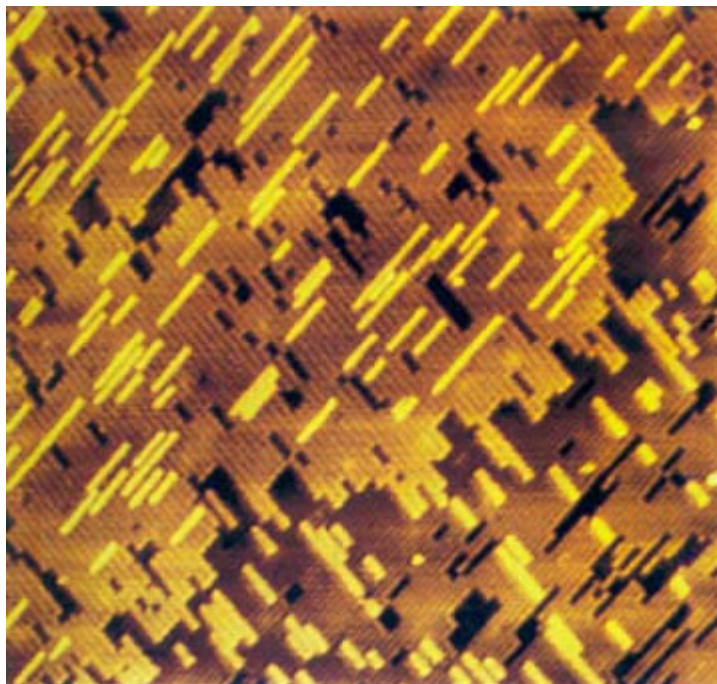


Figure A3.10.10 STM image ($55 \times 55 \text{ nm}^2$) of a Si(100) surface exposed to molecular bromine at 800 K. The dark areas are etch pits on the terraces, while the bright rows that run perpendicular to the terraces are Si dimer chains. The dimer chains consist of Si atoms released from terraces and step edges during etching [28].

Another view of the Si(100) etching mechanism has been proposed recently [28]. Calculations have revealed that the most important step may actually be the escape of the bystander silicon atom, rather than SiBr_2 desorption. In this way, the SiBr_2 becomes trapped in a state that otherwise has a very short lifetime, permitting many more desorption attempts. Preliminary results suggest that indeed this vacancy-assisted desorption is the key step to etching Si(100) with Br_2 .

The implementation of tools such as the STM will undoubtedly continue to provide unprecedented views of etching reactions and will deepen our understanding of the phenomena that govern these processes.

A3.10.5 CATALYTIC REACTIONS

A3.10.5.1 INTRODUCTION

A catalyst is a material that accelerates a reaction rate towards thermodynamic equilibrium conversion without itself being consumed in the reaction. Reactions occur on catalysts at particular sites, called ‘active sites’, which may have different electronic and geometric structures than neighbouring sites. Catalytic reactions are at the heart of many chemical industries, and account for a large fraction of worldwide chemical production. Research into fundamental aspects of catalytic reactions has a strong economic motivating factor: a better understanding of the catalytic process

may lead to the development of a more efficient catalyst. While the implementation of a new catalyst based on surface science studies has not yet been realized, the investigation of catalysis using surface science methods has certainly shaped the current understanding of catalytic reactions. Several recommended texts on catalysis

can be found in [31, 32 and 33].

Fundamental studies in catalysis often incorporate surface science techniques to study catalytic reactions at the atomic level. The goal of such experiments is to characterize a catalytic surface before, during and after a chemical reaction; this is no small task. The characterization of these surfaces is accomplished using a number of modern analytical techniques. For example, surface compositions can be determined using x-ray photoelectron spectroscopy (XPS) or Auger electron spectroscopy (AES). Surface structures can be probed using low-energy electron diffraction (LEED) or STM. In addition, a number of techniques are available for detecting and identifying adsorbed species on surfaces, such as infrared reflection absorption spectroscopy, high-resolution electron energy-loss spectroscopy (HREELS) and sum frequency generation (SFG).

As with the other surface reactions discussed above, the steps in a catalytic reaction (neglecting diffusion) are as follows: the adsorption of reactant molecules or atoms to form bound surface species, the reaction of these surface species with gas phase species or other surface species and subsequent product desorption. The global reaction rate is governed by the slowest of these elementary steps, called the rate-determining or rate-limiting step. In many cases, it has been found that either the adsorption or desorption steps are rate determining. It is not surprising, then, that the surface structure of the catalyst, which is a variable that can influence adsorption and desorption rates, can sometimes affect the overall conversion and selectivity.

Industrial catalysts usually consist of one or more metals supported on a metal oxide. The supported metal can be viewed as discrete single crystals on the support surface. Changes in the catalyst structure can be achieved by varying the amount, or 'loading', of the metal. An increased loading should result in a particle size increase, and so the relative population of a particular crystal face with respect to other crystal faces may change. If a reaction rate on a per active site basis changes as the metal loading changes, then the reaction is deemed to be structure sensitive. The surface science approach to studying structure-sensitive reactions has been to examine the chemistry that occurs over different crystal orientations. In general, these studies have shown that close-packed, atomically smooth metal surfaces such as (111) and (100) fcc and (110) bcc surfaces are less reactive than more open, rough surfaces such as fcc(110) and bcc(111). The remaining task is then to relate the structure sensitivity results from single-crystal studies to the activity results over real-world catalysts.

Surface science studies of catalytic reactions certainly have shed light on the atomic-level view of catalysis. Despite this success, however, two past criticisms of the surface science approach to catalysis are that the pressure regimes (usually 10^{-10} Torr) and the materials (usually low-surface-area single crystals) are far removed from the high pressures and high-surface-area supported catalysts used industrially. These criticisms have been termed the 'pressure gap' and the 'materials gap'. To combat this criticism, much research in the last 30 years has focused on bridging these gaps, and many advances have been made that now suggest these criticisms are no longer warranted.

A3.10.5.2 EXPERIMENTAL

(A) BRIDGING THE PRESSURE GAP

The implementation of high-pressure reaction cells in conjunction with UHV surface science techniques allowed the first true *in situ* postmortem studies of a heterogeneous catalytic reaction. These cells permit exposure of a sample to ambient pressures without any significant contamination of the UHV environment. The first such cell was internal to the main vacuum chamber and consisted of a metal bellows attached to a reactor cup [34]. The cup could be translated using a hydraulic piston to envelop the sample, sealing it from

the surrounding UHV by means of a copper gasket. Once isolated from the vacuum, the activity of the enclosed sample for a given reaction could be measured at elevated pressures. Following the reaction, the high-pressure cell was evacuated and then retracted, exposing the sample again to the UHV environment, at which point any number of surface science techniques could be used to study the 'spent' catalyst surface.

Shortly thereafter, another high-pressure cell design appeared [35]. This design consisted of a sample mounted on a retractable bellows, permitting the translation of the sample to various positions. The sample could be retracted to a high-pressure cell attached to the primary chamber and isolated by a valve, thereby maintaining UHV in the primary chamber when the cell was pressurized for catalytic studies. The reactor could be evacuated following high-pressure exposures before transferring the sample back to the main chamber for analysis.

A modification to this design appeared several years later (figure A3.10.11) [36, 37]. In this arrangement, the sample rod can be moved easily between the UHV chamber and the high-pressure cell without any significant increase in chamber pressure. Isolation of the reaction cell from UHV is achieved by a differentially pumped sliding seal mechanism (figure A3.10.12) whereby the sample rod is pushed through the seals until it is located in the high-pressure cell. Three spring-loaded, differentially pumped Teflon seals are used to isolate the reaction chamber from the main chamber by forming a seal around the sample rod. Differential pumping is accomplished by evacuating the space between the first and second seals (on the low-pressure side) by a turbomolecular pump and the space between the second and third seals (on the high-pressure side) by a mechanical (roughing) pump. Pressures up to several atmospheres can be maintained in the high-pressure cell while not significantly raising the pressure in the attached main chamber.

The common thread to these designs is that a sample can be exposed to reaction conditions and then studied using surface science methods without exposure to the ambient. The drawback to both of these designs is that the samples are still being analysed under UHV conditions *before* and *after* the reaction under study. The need for *in situ* techniques is clear.

-21-

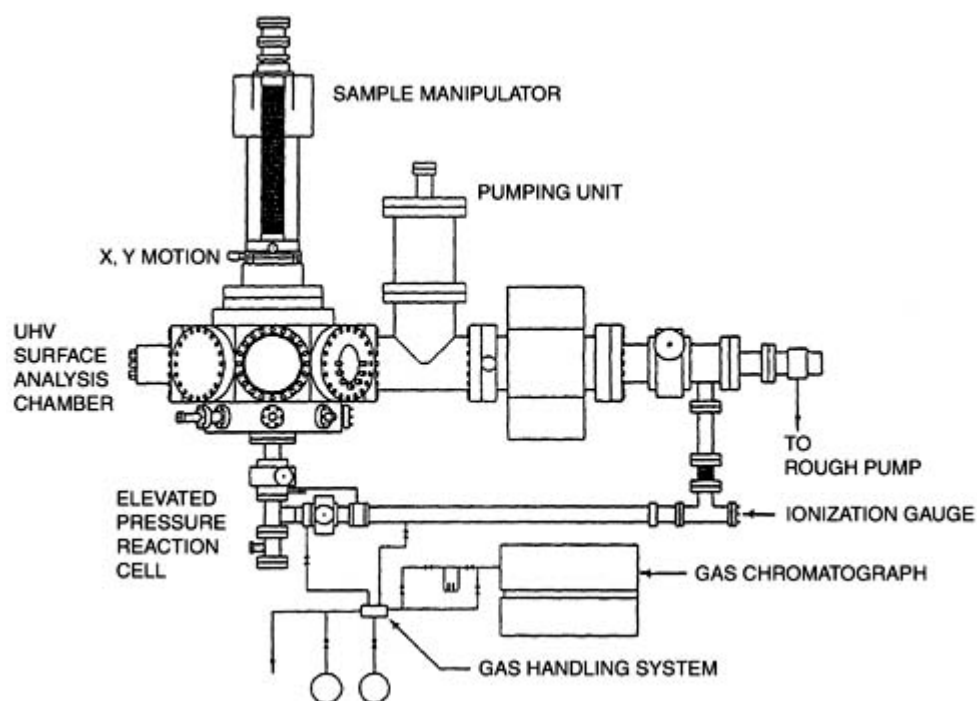


Figure A3.10.11 Side view of a combined high-pressure cell and UHV surface analysis system [37].

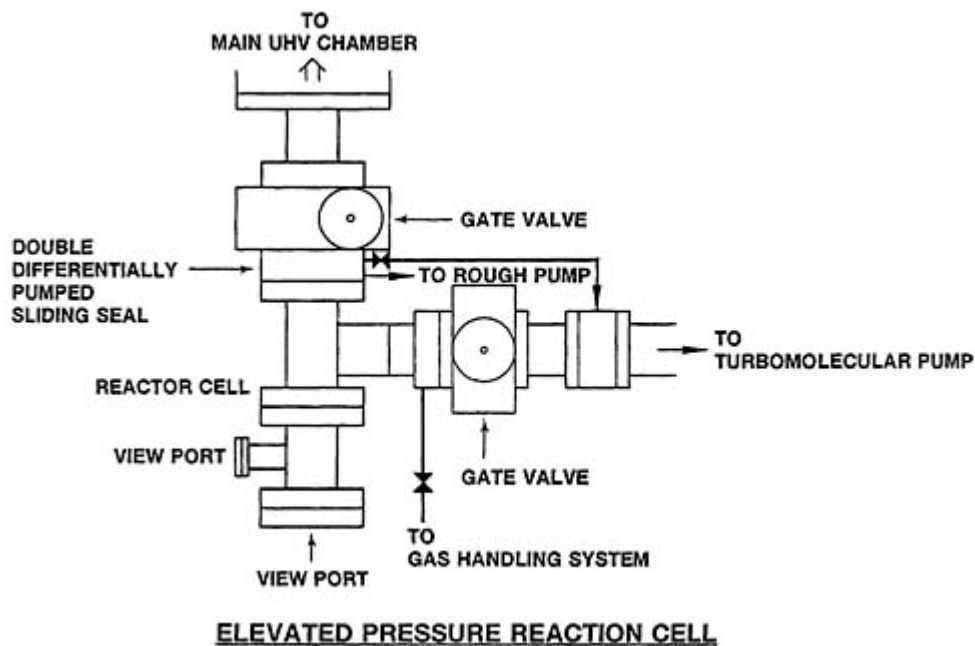


Figure A3.10.12 Side view of the high-pressure cell showing the connections to the UHV chamber, the turbomolecular pump and the gas handling system. The differentially pumped sliding seal is located between the high-pressure cell and the UHV chamber [37].

Two notable *in situ* techniques are at the forefront of the surface science of catalysis: STM and SFG. STM is used to investigate surface structures while SFG is used to investigate surface reaction intermediates. The significance of both techniques is that they can operate over a pressure range of 13 orders of magnitude, from 10^{-10} to 10^3 Torr, i.e. they are truly *in situ* techniques. STM has allowed the visualization of surface structures under ambient conditions and has shed light on adsorbate-induced morphological changes that occur at surfaces, for both single-crystal metals and metal clusters supported on oxide single crystals. Studies of surface reactions with SFG have given insight into reaction mechanisms previously investigated under non-ideal pressure or temperature constraints. Both SFG and STM hold promise as techniques that will contribute greatly to the understanding of catalytic reactions under *in situ* conditions.

(B) BRIDGING THE MATERIALS GAP

Single crystals are traditionally used in UHV studies because they provide an opportunity to well characterize a surface. However, as discussed above, single crystals are quite different from industrial catalysts. Typically, such catalysts consist of supported particles that can have multiple crystal orientations exposed at the surface. Therefore, an obstacle in attempting surface science studies of catalysis is the preparation of a surface in such a way that it mimics a real-world catalyst.

One criterion necessary for using charged-particle spectroscopies such as AES and EELS is that the material being investigated should be conductive. This requisite prevents problems such as charging when using electron spectroscopies and ensures homogeneous heating during thermal desorption studies. A problem then with investigating oxide surfaces for use as metal supports is that many are insulators or semiconductors. For example, alumina and silica are often used as oxide supports for industrial catalysts, yet both are insulators at room temperature, severely hindering surface science studies of these materials. However, thin-films of these and other oxides can be deposited onto metal substrates, thus providing a conductive substrate (*via* tunnelling)

for use with electron spectroscopies and other surface science techniques.

Thin oxide films may be prepared by substrate oxidation or by vapour deposition onto a suitable substrate. An example of the former method is the preparation of silicon oxide thin-films by oxidation of a silicon wafer. In general, however, the thickness and stoichiometry of a film prepared by this method are difficult to control. On the other hand, vapour deposition, which consists of evaporating the parent metal in an oxidizing environment, allows precise control of the film thickness. The extent of oxidation can be controlled by varying the O_2 pressure (lower O_2 pressures can lead to lower oxides) and the film thickness can be controlled by monitoring the deposition rate. A number of these thin metal oxide films have been prepared by vapour deposition, including SiO_2 , Al_2O_3 , MgO , TiO_2 and NiO [38].

MgO films have been grown on a $Mo(100)$ substrate by depositing Mg onto a clean $Mo(100)$ sample in O_2 ambient at 300 K [39, 40]. LEED results indicated that MgO grows epitaxially at an optimum O_2 pressure of 10^{-7} Torr, with the (100) face of MgO parallel to the $Mo(100)$ surface. Figure A3.10.13 shows a ball model illustration of the $MgO(100)$ overlayer on $Mo(100)$. The chemical states of Mg and O were also probed as a function of the O_2 pressure during deposition by AES and XPS. It was found that as the O_2 pressure was increased, the metallic Mg^0 ($L_{2,3}VV$) Auger transition at 44.0 eV decreased while a new transition at 32.0 eV increased. The transition at 32.0 eV was assigned to a Mg^{2+} ($L_{2,3}VV$) transition due to the formation of MgO . When the O_2 pressure reached 10^{-7} Torr, the Mg^{2+} feature dominated the AES spectrum while the Mg^0 feature completely diminished. XPS studies confirmed the LEED and AES results, verifying that MgO was formed at the optimal O_2 pressure. Furthermore, the Mg 2p and O 1s XPS peaks from the MgO film had the same binding energy (BE) and peak shape as the Mg 2p and O 1s peaks from an MgO single crystal. Both AES and XPS indicated that the stoichiometry of the film was MgO . Further annealing in O_2 did

-23-

not increase the oxygen content of the film, which supports the fact that no evidence of Mg suboxides was found. This MgO film was successfully used to study the nature of surface defects in Li -doped MgO as they relate to the catalytic oxidative coupling of methane.

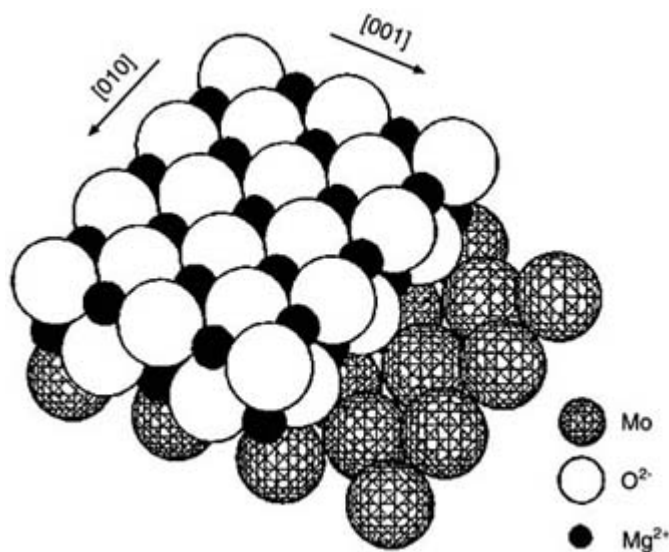


Figure A3.10.13 Ball model illustration of an epitaxial MgO overlayer on $Mo(100)$ [38].

The deposition of titanium oxide thin-films on $Mo(110)$ represents a case where the stoichiometry of the film is sensitive to the deposition conditions [41]. It was found that both TiO_2 and Ti_2O_3 thin-films could be made,

depending on the Ti deposition rate and the O₂ background pressure. Lower deposition rates and higher O₂ pressures favoured the formation of TiO₂. The two compositionally different films could be distinguished in several ways. Different LEED patterns were observed for the different films: TiO₂ exhibited a (1 × 1) rectangular periodicity, while Ti₂O₃ exhibited a (1 × 1) hexagonal pattern. XPS Ti 2p data clearly differentiated the two films as well, showing narrow peaks with a Ti 2p_{3/2} BE of 459.1 eV for TiO₂ and broad peaks with a Ti 2p_{3/2} BE of 458.1 eV for Ti₂O₃. From LEED and HREELS results, it was deduced that the surfaces grown on Mo(110) were TiO₂(100) and Ti₂O₃(0001). Therefore, it is clear that vapour deposition allows control over thickness and extent of oxidation and is certainly a viable method for producing thin oxide films for use as model catalyst supports.

Metal vapour deposition is a method that can be used to conveniently prepare metal clusters for investigation under UHV conditions. The deposition is accomplished using a doser constructed by wrapping a high-purity wire of the metal to be deposited around a tungsten or tantalum filament that can be resistively heated. After sufficient outgassing, which is the process of heating the doser to remove surface and bulk impurities, then a surface such as an oxide can be exposed to the metal emanating from the doser to yield a model oxide-supported metal catalyst.

Model catalysts such as Au/TiO₂(110) have been prepared by metal vapour deposition [42]. [Figure A3.10.14](#) shows a STM image of 0.25 ML (1 ML = 1.387 × 10¹⁵ atoms cm⁻²) Au/TiO₂(110). These catalysts were tested for CO oxidation to compare to conventional Au catalysts. It is well known that for conventional Au catalysts there is an optimal Au cluster size (~3 nm) that yields a maximum CO oxidation rate. This result was duplicated by measuring the CO oxidation rate over model Au/TiO₂(110), where the cluster sizes were varied by manipulating the deposition

amounts. There is a definite maximum in the CO oxidation activity at a cluster size of approximately 3.5 nm. Furthermore, investigation of the cluster electronic properties using scanning tunnelling spectroscopy (STS) revealed a correlation between the cluster electronic structure and the maximum in CO oxidation activity. Pd/SiO₂/Mo(100) model catalysts were also prepared and were found to have remarkably similar kinetics for CO oxidation when compared to Pd single crystals and conventional silica-supported Pd catalysts [43]. These results confirm that metal vapour deposition on a suitable substrate is a viable method for producing model surfaces for UHV studies.

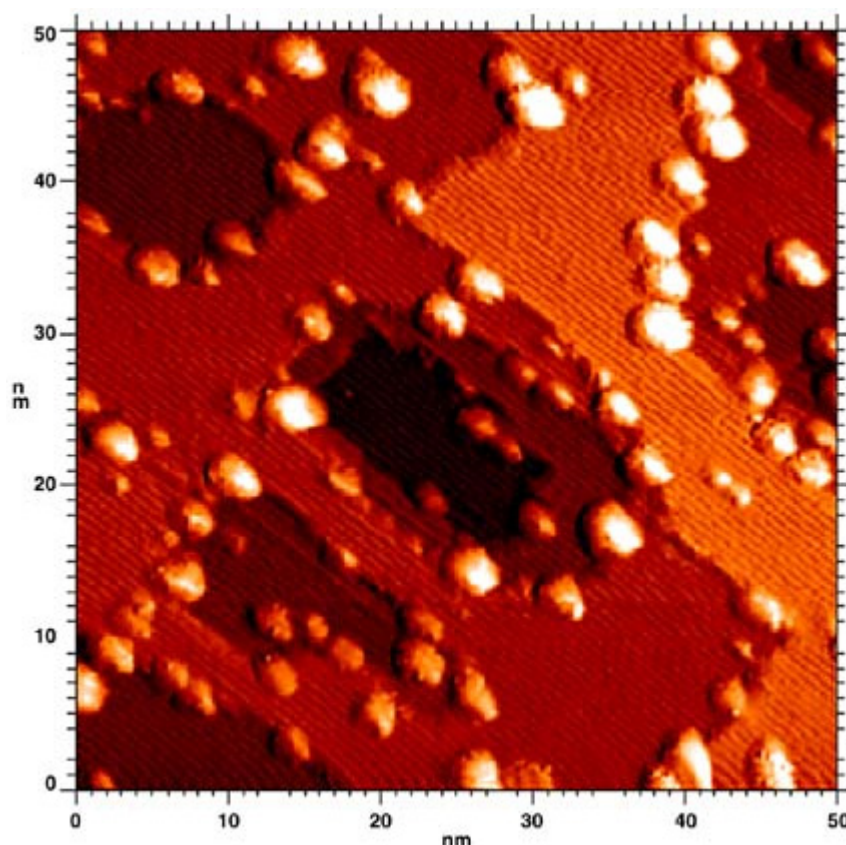


Figure A3.10.14 STM image of 0.25 ML Au vapour-deposited onto $\text{TiO}_2(110)$. Atomic resolution of the substrate is visible as parallel rows. The Au clusters are seen to nucleate preferentially at step edges.

Another method by which model-supported catalysts can be made is electron beam lithography [44]. This method entails spin-coating a polymer solution onto a substrate and then using a collimated electron beam to damage the polymer surface according to a given pattern. Next, the damaged polymer is removed, exposing the substrate according to the electron beam pattern, and the sample is coated with a thin metal film. Finally, the polymer is removed from the substrate, taking with it the metal film except where the metal was bound to the substrate, leaving behind metal particles of variable size. This technique has been used to prepare Pt particles with 50 nm diameters and 15 nm heights on an oxidized silicon support [44]. It was found that ethylene hydrogenation reaction rates on the model catalysts agreed well with turnover rates on Pt single crystals and conventional Pt-supported catalysts.

A3.10.5.3 ATOMIC-LEVEL VIEWS OF CATALYSIS

(A) NH_3 SYNTHESIS: $\text{N}_2 + 3\text{H}_2 \leftrightarrow 2\text{NH}_3$

Ammonia has been produced commercially from its component elements since 1909, when Fritz Haber first demonstrated the viability of this process. Bosch, Mittasch and co-workers discovered an excellent promoted Fe catalyst in 1909 that was composed of iron with aluminium oxide, calcium oxide and potassium oxide as promoters. Surprisingly, modern ammonia synthesis catalysts are nearly identical to that first promoted iron catalyst. The reaction is somewhat exothermic and is favoured at high pressures and low temperatures, although, to keep reaction rates high, moderate temperatures are generally used. Typical industrial reaction conditions for ammonia synthesis are 650–750 K and 150–300 atm. Given the technological importance of the

ammonia synthesis reaction, it is not surprising that surface science techniques have been used to thoroughly study this reaction on a molecular level [45, 46].

As mentioned above, a structure-sensitive reaction is one with a reaction rate that depends on the catalyst structure. The synthesis of ammonia from its elemental components over iron surfaces is an example of a structure-sensitive reaction. Figure A3.10.15 demonstrates this structure sensitivity by showing that the rate of NH_3 formation at 20 atm and 600–700 K has a clear dependence on the surface structure [47]. The (111) and (211) Fe faces are much more active than the (100), (210) and (110) faces. Figure A3.10.16 depicts the different Fe surfaces for which ammonia synthesis was studied in figure A3.10.15. The coordination of the different surface atoms is denoted in each drawing. Surface roughness is often associated with higher catalytic activity, however in this case the (111) and (210) surfaces, both of which can be seen to be atomically rough, have distinctly different catalytic activities. Closer inspection of these surfaces reveals that the (111) and (211) faces have a C_7 site in common, i.e. a surface Fe atom with seven nearest neighbours. The high catalytic activity of the (111) and (211) Fe faces has been proposed to be due to the presence of these C_7 sites.

-26-

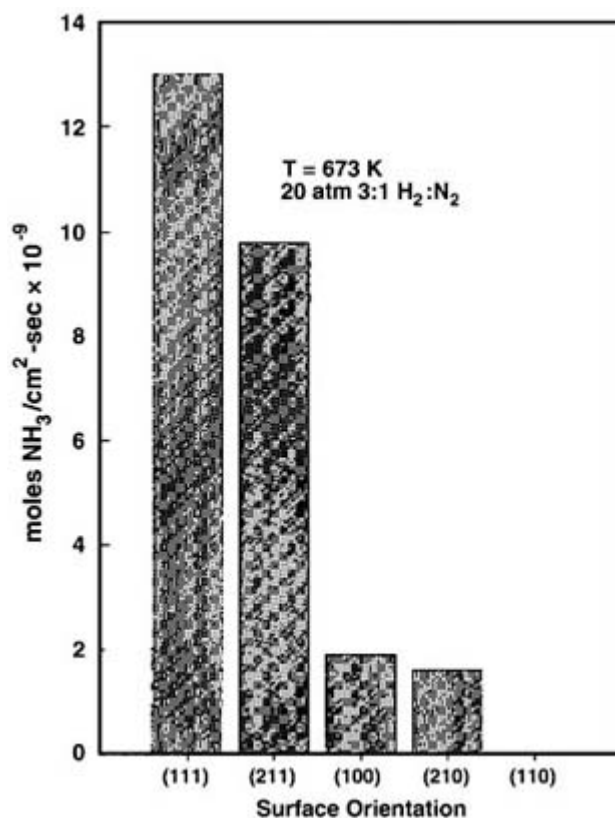


Figure A3.10.15 NH_3 synthesis activity of different Fe single-crystal orientations [32]. Reaction conditions were 20 atm and 600–700 K.

-27-

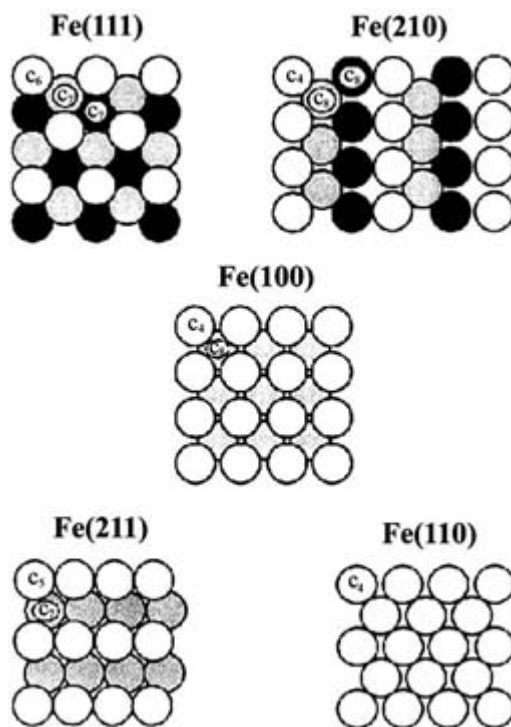


Figure A3.10.16 Illustrations of the surfaces in [figure A3.10.15](#) for which ammonia synthesis activity was tested. The coordination of the surface atoms is noted in the figure [32].

It is widely accepted that the rate-determining step in NH_3 synthesis is the dissociative adsorption of N_2 , depicted in a Lennard-Jones potential energy diagram in [figure A3.10.17](#) [46]. This result is clearly illustrated by examining the sticking coefficient (the adsorption rate divided by the collision rate) of N_2 on different Fe crystal faces ([figure A3.10.18](#)) [48]. The concentration of surface nitrogen on the Fe single crystals at elevated temperatures in UHV was monitored with AES as a function of N_2 exposure. The sticking coefficient is proportional to the slope of the curves in [figure A3.10.18](#). The initial sticking coefficients increase in the order $(110) < (100) < (111)$, which is the same trend observed for the ammonia synthesis catalytic activity at high-pressure (20 atm). This result indicates that the pressure gap for ammonia synthesis can be overcome: the kinetics results obtained in UHV conditions can be readily extended to the kinetics results obtained under high-pressure reaction conditions.

-28-

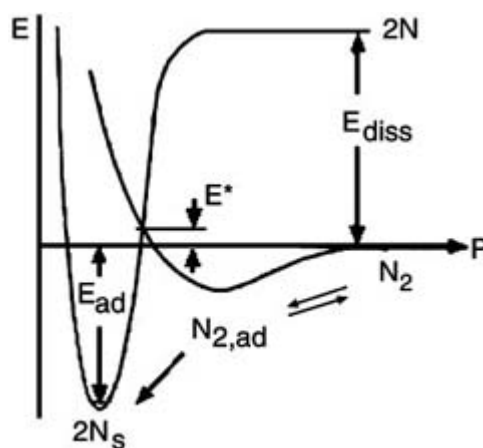


Figure A3.10.17 Potential energy diagram for the dissociative adsorption of N_2 [46].

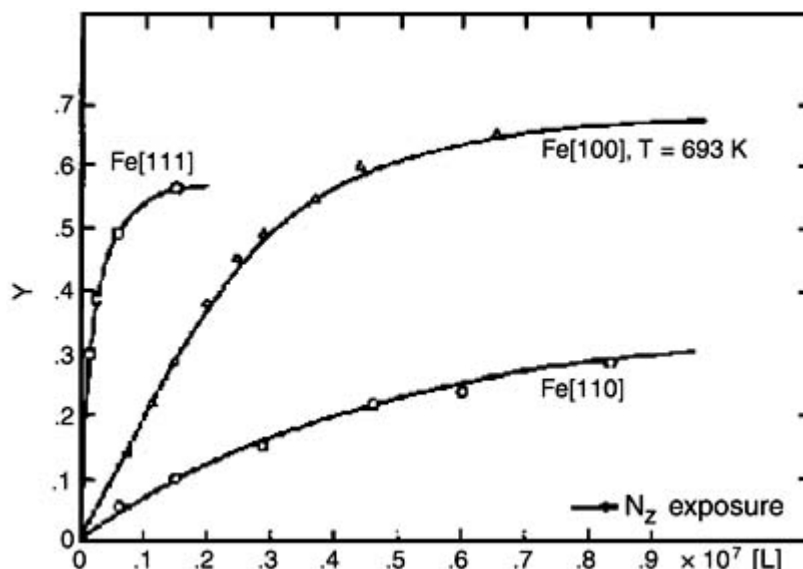


Figure A3.10.18 Surface concentration of nitrogen on different Fe single crystals following N_2 exposure at elevated temperatures in UHV [48].

Further work on modified Fe single crystals explored the role of promoters such as aluminium oxide and potassium [49, 50 and 51]. It was found that the simple addition of aluminium oxide to Fe single crystal surfaces decreased the ammonia synthesis rate proportionally to the amount of Fe surface covered, indicating no favourable interaction between Fe and aluminium oxide under those conditions. However, by exposing an aluminium-oxide-modified Fe surface to water vapour, the surface was oxidized, inducing a favourable interaction between Fe and the Al_xO_y . This interaction resulted in a 400-fold increase in ammonia synthesis activity for $Al_xO_y/Fe(110)$ as compared to Fe(110) and an activity for $Al_xO_y/Fe(110)$ comparable to that of Fe(111). Interestingly, aluminium-oxide-modified Fe(111) showed no change in activity. The increase in activity for $Al_xO_y/Fe(110)$ to that of Fe(111) suggests a possible reconstruction

of the catalyst surface, in particular that Fe(111) and Fe(211) surfaces may be formed. These surfaces have C_7 sites and so the formation of crystals with these orientations could certainly lead to an enhancement in catalytic activity. Thus, the promotion of Fe ammonia synthesis catalysts by Al_xO_y appears to be primarily a geometric effect.

The addition of potassium to Fe single crystals also enhances the activity for ammonia synthesis. Figure A3.10.19 shows the effect of surface potassium concentration on the N_2 sticking coefficient. There is nearly a 300-fold increase in the sticking coefficient as the potassium concentration reaches $\sim 1.5 \times 10^{14}$ K atoms cm^{-2} . Not only does the sticking coefficient increase, but with the addition of potassium as a promoter, N_2 molecules are bound more tightly to the surface, with the adsorption energy increasing from 30 to 45 $kJ\ mol^{-1}$. A consequence of the lowering of the N_2 potential well is that the activation energy for dissociation (E^* in Figure A3.10.17) also decreases. Thus, the promotion of Fe ammonia synthesis catalysts by potassium appears to be primarily an electronic effect.

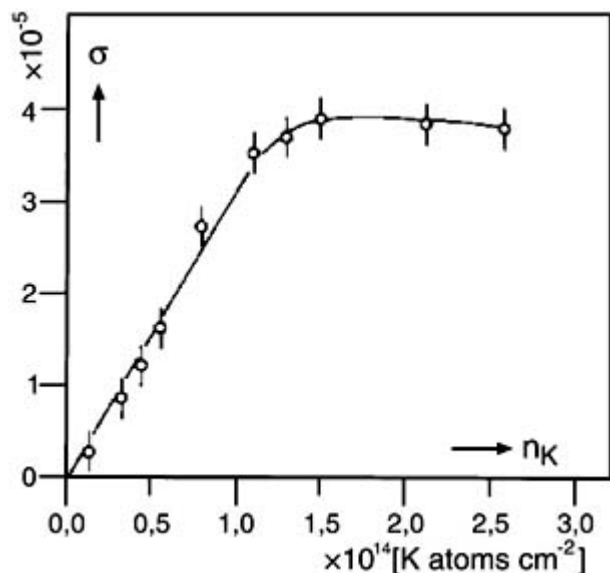
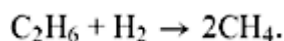


Figure A3.10.19 Variation of the initial sticking coefficient of N_2 with increasing potassium surface concentration on Fe(100) at 430 K [50].

(B) ALKANE HYDROGENOLYSIS

Alkane hydrogenolysis, or cracking, involves the dissociation of a larger alkane molecule to a smaller alkane molecule. For example, ethane hydrogenolysis in the presence of H_2 yields methane:



Cracking (or hydrocracking, as it is referred to when carried out in the presence of H_2) reactions are an integral part of petroleum refining. Hydrocracking is used to lower the average molecular weight (MW) of a higher MW hydrocarbon mixture so that it can then be blended and sold as gasoline. The interest in the fundamentals of catalytic cracking reactions is strong and it has been thoroughly researched.

Ethane hydrogenolysis has been shown to be structure sensitive over nickel catalysts [43], as seen in figure A3.10.20 where methane formation rates are plotted for both nickel single crystals and a conventional, supported nickel catalyst. There is an obvious difference in the rates over Ni(111) and Ni(100), and it is evident that the rate also changes as a function of particle size for the supported Ni catalysts. In addition, differences in activation energy were observed: for Ni(111) the activation energy is 192 kJ mol^{-1} , while for Ni(100) the activation energy is 100 kJ mol^{-1} . It is noteworthy that there is overlap between the hydrogenolysis rates over supported Ni catalysts with the Ni single crystals. The data suggest that small Ni particles are composed primarily of Ni(100) facets while large Ni particles are composed primarily of Ni(111) facets. In fact, this has been observed for fcc materials where surfaces with a (111) orientation are more commonly observed after thermally induced sintering. The structure sensitivity of this reaction over Ni surfaces has been clearly demonstrated.

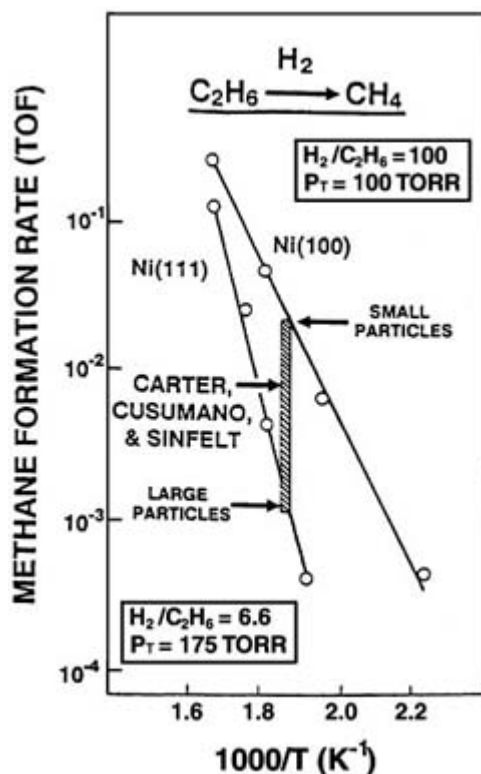


Figure A3.10.20 Arrhenius plot of ethane hydrogenolysis activity for Ni(100) and Ni(111) at 100 Torr and $\text{H}_2/\text{C}_2\text{H}_6 = 100$. Also included is the hydrogenolysis activity on supported Ni catalysts at 175 Torr and $\text{H}_2/\text{C}_2\text{H}_6 = 6.6$ [43].

The initial step in alkane hydrogenolysis is the dissociative adsorption, or ‘reactive sticking’ of the alkane. One might suspect that this first step may be the key to the structure sensitivity of this reaction over Ni surfaces. Indeed, the reactive sticking of alkanes has been shown to depend markedly on surface structure [52]. Figure A3.10.21 shows the buildup of surface carbon due to methane decomposition ($P_{\text{methane}} = 1.00$ Torr) over three single-crystal Ni surfaces at 450 K. The rate of methane decomposition is obviously dependent upon the surface structure with the decomposition rate increasing in the order (111) < (100) < (110). It can be seen that, initially, the rates of methane decomposition are

similar for Ni(100) and (110), while Ni(111) has a much lower reaction rate. With increasing reaction time, i.e. increasing carbon coverage, the rate over Ni(110) continues to increase linearly while both Ni(111) and (100) exhibit a nonlinear dependence. This linear dependence over Ni(111) may be due to either the formation of carbon islands or a reduced carbon coverage dependence as compared to Ni(111) and (100).

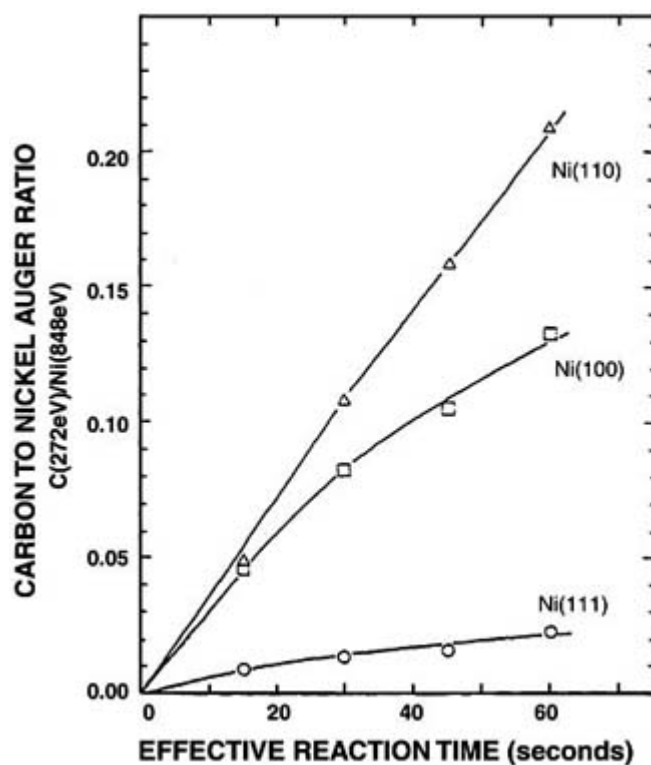


Figure A3.10.21 Methane decomposition kinetics on low-index Ni single crystals at 450 K and 1.00 Torr methane [43].

Hydrogenolysis reactions over Ir single crystals and supported catalysts have also been shown to be structure sensitive [53, 54 and 55]. In particular, it was found that the reactivity tracked the concentration of low-coordination surface sites. Figure A3.10.22 shows ethane selectivity (selectivity is reported here because both ethane and methane are products of butane cracking) for *n*-butane hydrogenolysis over Ir(111) and the reconstructed surface Ir(110)-(1 × 2), as well as two supported Ir catalysts. There are clear selectivity differences between the two Ir surfaces, with Ir(110)-(1 × 2) having approximately three times the ethane selectivity of Ir(111). There is also a similarity seen between the ethane selectivity on small Ir particles and Ir(110)-(1 × 2), and between the ethane selectivity on large Ir particles and Ir(111).

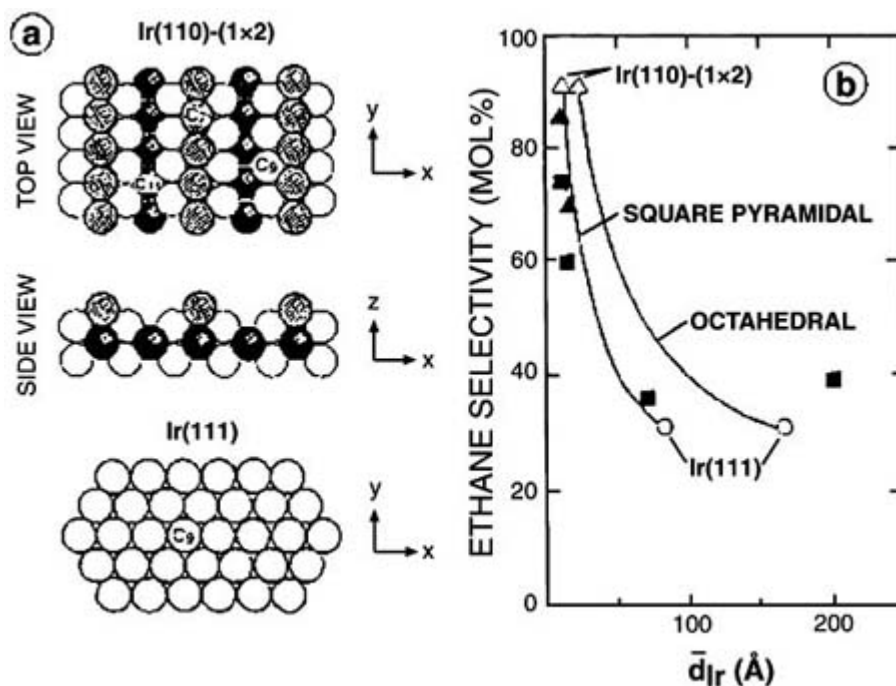


Figure A3.10.22 Relationship between selectivity and surface structure for *n*-butane hydrogenolysis on iridium. (a) Illustrations of the Ir(110)-(1 × 2) and Ir(111) surfaces. The *z*-axis is perpendicular to the plane of the surface. (b) Selectivity for C₂H₆ production (mol% total products) for *n*-butane hydrogenolysis on both Ni single crystals and supported catalysts at 475 K. The effective particle size for the single crystal surfaces is based on the specified geometric shapes [43]. Δ Ir/Al₂O₃; \square Ir/SiO₂.

The mechanisms by which *n*-butane hydrogenolysis occurs over Ir(110)-(1 × 2) and Ir(111) are different. The high ethane selectivity of Ir(110)-(1 × 2) has been attributed to the ‘missing row’ reconstruction that the (110) surface undergoes (figure A3.10.22). This reconstruction results in the exposure of a highly uncoordinated C₇ site that is sterically unhindered. These C₇ sites are capable of forming a metallocyclopentane (a five-membered ring consisting of four carbons and an Ir atom) which, based on kinetics and surface carbon coverages, has been suggested as the intermediate for this reaction [56, 57]. It has been proposed that the crucial step in this reaction mechanism over the reconstructed (110) surface is the reversible cleavage of the central C–C bond. On the other hand, the hydrogenolysis of *n*-butane over Ir(111) is thought to proceed by a different mechanism, where dissociative chemisorption of *n*-butane and hydrogen are the first steps. Then, the adsorbed hydrocarbon undergoes the irreversible cleavage of the terminal C–C bond. It is evident that surface structure plays an important role in hydrogenolysis reactions over both nickel and iridium surfaces.

(C) CO OXIDATION: $2\text{CO} + \text{O}_2 \rightarrow 2\text{CO}_2$

The oxidation of CO to CO₂, which is essential to controlling automobile emissions, has been extensively studied because of the relative simplicity of this reaction. CO oxidation was the first reaction to be studied using the surface science approach and is perhaps the most well understood heterogeneous catalytic reaction [58]. The simplicity of CO oxidation by O₂ endears itself to surface science studies. Both reactants are diatomic molecules whose adsorption

on single-crystal surfaces has been widely studied, and presumably few steps are necessary to convert CO to CO₂. Surface science studies of CO and O₂ adsorption on metal surfaces have provided tremendous insight

into the mechanism of the CO–O₂ reaction. The mechanism over platinum surfaces has been unequivocally established and the reaction has shown structure insensitivity over platinum [59], palladium [55, 59, 60] and rhodium surfaces [61, 62].

Although dissociative adsorption is sometimes observed, CO adsorption on platinum group metals typically occurs molecularly and this will be the focus of the following discussion. Figure A3.10.23 illustrates schematically the donor–acceptor model (first proposed by Blyholder [63]) for molecular CO chemisorption on a metal such as platinum. The bonding of CO to a metal surface is widely accepted to be similar to bond formation in a metal carbonyl. Experimental evidence indicates that the 5σ highest occupied molecular orbital (HOMO), which is regarded as a lone pair on the carbon atom, bonds to the surface by donating charge to unoccupied density of states (DOS) at the surface. Furthermore, this surface bond can be strengthened by back-donation, which is the transfer of charge from the surface to the 2π* lowest unoccupied molecular orbital (LUMO). An effect of this backbonding is that the C–O bond weakens, as seen by a lower C–O stretch frequency for adsorbed CO (typically <2100 cm⁻¹) than for gas phase CO (2143 cm⁻¹).

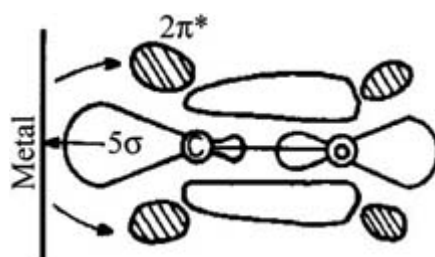


Figure A3.10.23 Schematic diagram of molecular CO chemisorption on a metal surface. The model is based on a donor–acceptor scheme where the CO 5σ HOMO donates charge to surface unoccupied states and the surface back-donates charge to the CO 2π LUMO [58].

Ultraviolet photoelectron spectroscopy (UPS) results have provided detailed information about CO adsorption on many surfaces. Figure A3.10.24 shows UPS results for CO adsorption on Pd(110) [58] that are representative of molecular CO adsorption on platinum surfaces. The difference result in (c) between the clean surface and the CO-covered surface shows a strong negative feature just below the Fermi level (E_F), and two positive features at ~8 and 11 eV below E_F . The negative feature is due to suppression of emission from the metal d states as a result of an anti-resonance phenomenon. The positive features can be attributed to the 4σ molecular orbital of CO and the overlap of the 5σ and 1π molecular orbitals. The observation of features due to CO molecular orbitals clearly indicates that CO molecularly adsorbs. The overlap of the 5σ and 1π levels is caused by a stabilization of the 5σ molecular orbital as a consequence of forming the surface–CO chemisorption bond.

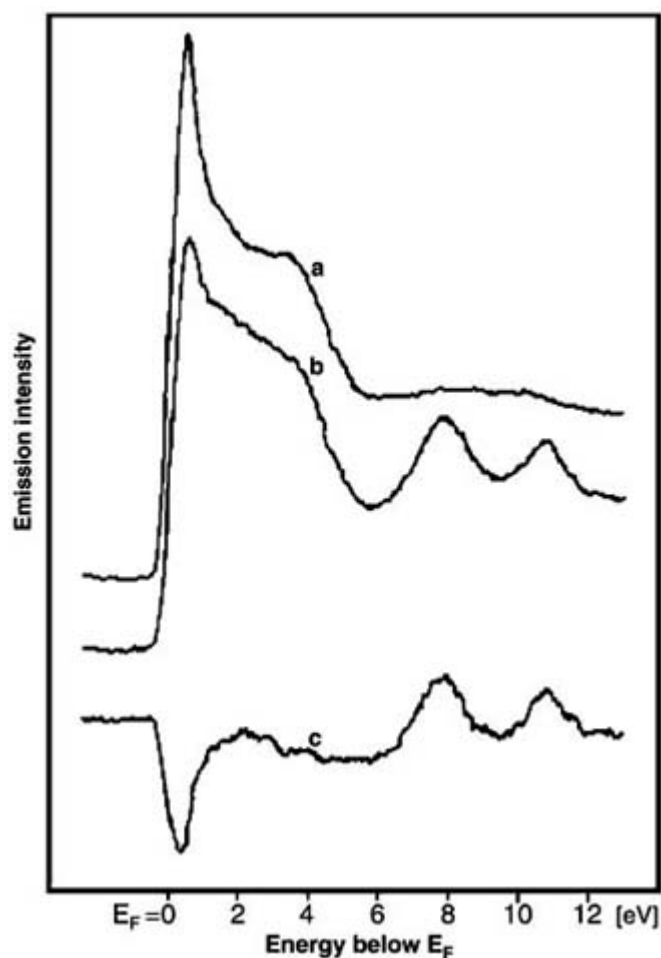
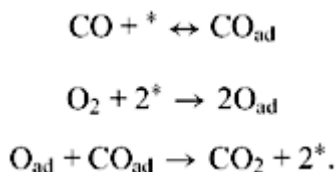


Figure A3.10.24 UPS data for CO adsorption on Pd(110). (a) Clean surface. (b) CO-dosed surface. (c) Difference spectrum (b-a). This spectrum is representative of molecular CO adsorption on platinum metals [58].

The adsorption of O₂ on platinum surfaces is not as straightforward as CO adsorption because molecular and dissociative adsorption can occur, as well as oxide formation [58]. However, molecular adsorption has been observed only at very low temperatures, where CO oxidation rates are negligible, hence this form of adsorbed oxygen will not be discussed here. UPS data indicate dissociative adsorption of O₂ on platinum surfaces at temperatures >100 K, and isotopic exchange measurements support this finding as well. The oxygen atoms resulting from O₂ dissociation can be either chemisorbed oxygen or oxygen in the form of an oxide. The two types of oxygen are distinguished by noting that oxide oxygen is located beneath the surface ('subsurface') while chemisorbed oxygen is located on the surface. Experimentally, the two types of oxygen are discernible by AES, XPS and UPS. In general, it has been found that as long as pressure and temperature are kept fairly low, the most likely surface oxygen species will be chemisorbed. Therefore, when formulating a mechanism for reaction under these general conditions, only chemisorbed oxygen needs to be considered.

The mechanism for CO oxidation over platinum group metals has been established from a wealth of data, the analysis of which is beyond the scope of this chapter. It is quite evident that surface science provided the foundation for this mechanism by directly showing that CO adsorbs molecularly and O₂ adsorbs

dissociatively. The mechanism is represented below (* denotes an empty surface site):



The first step consists of the molecular adsorption of CO. The second step is the dissociation of O₂ to yield two adsorbed oxygen atoms. The third step is the reaction of an adsorbed CO molecule with an adsorbed oxygen atom to form a CO₂ molecule that, at room temperature and higher, desorbs upon formation. To simplify matters, this desorption step is not included. This sequence of steps depicts a Langmuir–Hinshelwood mechanism, whereby reaction occurs between two adsorbed species (as opposed to an Eley–Rideal mechanism, whereby reaction occurs between one adsorbed species and one gas phase species). The role of surface science studies in formulating the CO oxidation mechanism was prominent.

CO oxidation by O₂ is a structure-insensitive reaction over rhodium catalysts [61, 62]. Figure A3.10.25 illustrates this structure insensitivity by demonstrating that the activation energies over supported Rh catalysts and a Rh(111) single crystal (given by the slope of the line) were nearly identical. Furthermore, the reaction rates over both supported Rh/Al₂O₃ and single crystal Rh (111) surfaces were also remarkably similar. Thus, the reaction kinetics were quite comparable over both the supported metal particles and the single crystal surfaces, and no particle size effect (structure sensitivity) was observed.

-36-

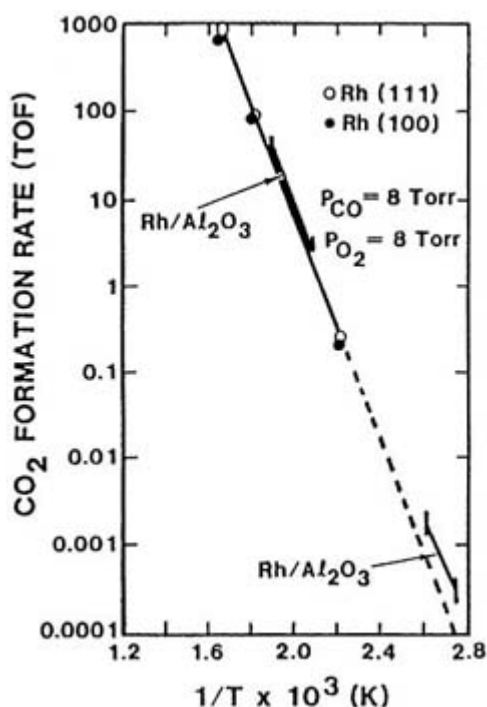


Figure A3.10.25 Arrhenius plots of CO oxidation by O₂ over Rh single crystals and supported Rh/Al₂O₃ at P_{CO} = P_{O₂} = 0.01 atm [43]. The dashed line in the figure is the predicted behaviour based on the rate constants for CO and O₂ adsorption and desorption on Rh under UHV conditions.

The study of catalytic reactions using surface science techniques has been fruitful over the last 30 years. Great strides have been made towards understanding the fundamentals of catalytic reactions, particularly by

bridging the material and pressure gaps. The implementation of *in situ* techniques and innovative model catalyst preparation will undoubtedly shape the future of catalysis.

REFERENCES

- [1] National Transportation Statistics 1998 (US Department of Transportation) table 4-3
- [2] Roth A 1990 *Vacuum Technology* 3rd edn (New York: Elsevier)
- [3] O'Hanlon J F 1989 *A User's Guide to Vacuum Technology* 2nd edn (New York: Wiley)
- [4] Atkins P W 1990 *Physical Chemistry* 4th edn (New York: Freeman)
- [5] West J M 1986 *Basic Corrosion and Oxidation* 2nd edn (Chichester: Ellis Horwood) p 220
- [6] Marcus P 1998 Surface science approach of corrosion phenomena *Electrochim. Acta* **43** 109
- [7] Itaya K 1998 *In situ* scanning tunneling microscopy in electrolyte solutions *Prog. Surf. Sci.* **58** 121
- [8] Maurice V, Kitakatsu N, Siegers M and Marcus P 1997 Low coverage sulfur induced reconstruction of Ni(111) *Surf. Sci.* **373** 307
-
- [9] Marcus P, Teissier A and Oudar J 1984 The influence of sulfur on the dissolution and the passivation of a nickel-iron alloy. 1. Electrochemical and radiotracer measurements *Corrosion Sci.* **24** 259
- [10] Moffat T P, Fan F R F and Bard A 1991 Electrochemical and scanning tunneling microscopic study of dealloying of Cu₃Au *J. Electrochem. Soc.* **138** 3224
- [11] Chen S J, Sanz F, Ogletree D F, Hallmark V M, Devine T M and Salmeron M 1993 Selective dissolution of copper from Au-rich Au-Cu alloys: an electrochemical STS study *Surf. Sci.* **292** 289
- [12] Dahmen U 1982 Orientation relationships in precipitation systems *Acta Metall.* **30** 63
- [13] Hahn E, Kampshoff E, Wälchli N and Kern K 1995 Strain driven fcc-bct phase transition of pseudomorphic Cu films on Pd (100) *Phys. Rev. Lett.* **74** 1803
- [14] Zangwill A 1988 *Physics at Surfaces* (Cambridge: Cambridge University Press) pp 427–8
- [15] Foord J S, Davies G J and Tsang W S 1997 *Chemical Beam Epitaxy and Related Techniques* (New York: Wiley)
- [16] Panish M B and Temkin H 1993 *Gas Source Molecular Beam Epitaxy* (New York: Springer)
- [17] Stringfellow G B 1989 *Organometallic Vapor-Phase Epitaxy* (San Diego, CA: Academic)
- [18] Zangwill A 1988 *Physics at Surfaces* (Cambridge: Cambridge University Press) pp 428–32
- [19] Brune H 1998 Microscopic view of epitaxial metal growth: nucleation and aggregation *Surf. Sci. Rep.* **31** 121
- [20] Lewis B and Anderson J C 1978 *Nucleation and Growth of Thin Films* (New York: Academic)
- [21] Venables J A, Spiller G D T and Hanbucken M 1984 Nucleation and growth of thin-films *Rep. Prog. Phys.* **47** 399
- [22] Stoyanov S and Kashchiev D 1981 Thin film nucleation and growth theories: a confrontation with experiment *Current Topics in Materials Science* vol 7, ed E Kaldis (Amsterdam: North-Holland) p 69
- [23] Brune H, Röder H, Boragno C and Kern K 1994 Microscopic view of nucleation on surfaces *Phys. Rev. Lett.* **73** 1955
- [24] Sze S M 1985 *Semiconductor Devices* (New York: Wiley) p 456
- [25] Cooke M J 1990 *Semiconductor Devices* (New York: Prentice-Hall) p 181
- [26] Williams R 1990 *Modern GaAs Processing Methods* (Norwood: Artech House)
- [27] Sugawara M 1998 *Plasma Etching: Fundamentals and Applications* (New York: Oxford University Press)
- [28] Boland J J and Weaver J H 1998 A surface view of etching *Phys. Today* **51** 34
- [29] Boland J J and Villarrubia J S 1990 Formation of Si(111)-(1 × 1)Cl *Phys. Rev. B* **41** 9865
- [30] Wiesendanger R 1994 *Scanning Probe Microscopy and Spectroscopy* (Cambridge: Cambridge University Press)

- [31] Thomas J M and Thomas W J 1996 *Principles and Practice of Heterogeneous Catalysis* (Weinheim: VCH)
- [32] Somorjai G A 1993 *Introduction to Surface Chemistry and Catalysis* (New York: Wiley)
- [33] Masel R I 1996 *Principles of Adsorption and Reaction on Solid Surfaces* (New York: Wiley)
- [34] Blakely D W, Kozak E I, Sexton B A and Somorjai G A 1976 New instrumentation and techniques to monitor chemical surface reactions over a wide pressure range (10^{-8} to 10^5 Torr) in the same apparatus *J. Vac. Sci. Technol.* **13** 1091
- [35] Goodman D W, Kelley R D, Madey T E and Yates J T Jr 1980 Kinetics of the hydrogenation of CO over a single crystal nickel catalyst *J. Catal.* **63** 226
- [36] Campbell R A and Goodman D W 1992 A new design for a multitechnique ultrahigh vacuum surface analysis chamber with high-pressure capabilities *Rev. Sci. Instrum.* **63** 172
-

-38-

- [37] Szanyi J and Goodman D W 1993 Combined elevated pressure reactor and ultrahigh vacuum surface analysis system *Rev. Sci. Instrum.* **64** 2350
- [38] Goodman D W 1996 Chemical and spectroscopic studies of metal oxide surfaces *J. Vac. Sci. Technol. A* **14** 1526
- [39] Wu M-C, Estrada C A, Corneille J S, He J-W and Goodman D W 1991 Synthesis and characterization of ultrathin MgO films on Mo(100) *Chem. Phys. Lett.* **472** 182
- [40] He J-W, Corneille J S, Estrada C A, Wu M-C and Goodman D W 1992 CO interaction with ultrathin MgO films on a Mo (100) surface studied by IRAS, TPD, and XPS *J. Vac. Sci. Technol. A* **10** 2248
- [41] Guo Q and Goodman D W Vanadium oxide thin-films grown on rutile $\text{TiO}_2(110)-(1 \times 1)$ and (1×2) surfaces *Surf. Sci.* **437** 38
- [42] Valden M, Lai X and Goodman D W 1998 Onset of catalytic activity of gold clusters on titania with the appearance of nonmetallic properties *Science* **281** 1647
- [43] Goodman D W 1995 Model studies in catalysis using surface science probes *Chem. Rev.* **95** 523
- [44] Jacobs P W and Somorjai G A 1997 Conversion of heterogeneous catalysis from art to science: the surface science of heterogeneous catalysis *J. Mol. Catal. A* **115** 389
- [45] Ertl G 1983 *Catalysis: Science and Technology* vol 4, ed J R Anderson and M Boudart (Heidelberg: Springer)
- [46] Ertl G 1990 Elementary steps in heterogeneous catalysis *Angew. Chem., Int. Ed. Engl.* **29** 1219
- [47] Strongin D R, Carrazza J, Bare S R and Somorjai G A 1987 The importance of C_7 sites and surface roughness in the ammonia synthesis reaction over iron *J. Catal.* **103** 213
- [48] Ertl G 1991 *Catalytic Ammonia Synthesis: Fundamentals and Practice, Fundamentals and Applied Catalysis* ed J R Jennings (New York: Plenum)
- [49] Bare S R, Strongin D R and Somorjai G A 1986 Ammonia synthesis over iron single crystal catalysts—the effects of alumina and potassium *J. Phys. Chem.* **90** 4726
- [50] Ertl G, Lee S B and Weiss M 1982 Adsorption of nitrogen on potassium promoted Fe(111) and (100) surfaces *Surf. Sci.* **114** 527
- [51] Paal Z, Ertl G and Lee S B 1981 Interactions of potassium, nitrogen, and oxygen with polycrystalline iron surfaces *Appl. Surf. Sci.* **8** 231
- [52] Beebe T P, Goodman D W, Kay B D and Yates J T Jr 1987 Kinetics of the activated dissociation adsorption of methane on low index planes of nickel single crystal surfaces *J. Chem. Phys.* **87** 2305
- [53] Wu M-C, Estrada C A, Corneille J S and Goodman D W 1996 Model surface studies of metal oxides: adsorption of water and methanol on ultrathin MgO films on Mo(100) *J. Chem. Phys.* **96** 3892
- [54] Xu X and Goodman D W 1992 New approach to the preparation of ultrathin silicon dioxide films at low temperature *Appl. Phys. Lett.* **61** 774
- [55] Szanyi J, Kuhn W K and Goodman D W 1994 CO oxidation on palladium: 2. A combined kinetic-infrared reflection absorption spectroscopic study of Pd(100) *J. Phys. Chem.* **98** 2978
- [56] Engstrom J R, Goodman D W and Weinberg W H 1986 Hydrogenolysis of n-butane over the (111) and (110)- (1×2) surfaces of iridium: a direct correlation between catalytic selectivity and surface structure *J. Am. Chem. Soc.* **108** 4653
- [57] Engstrom J R, Goodman D W and Weinberg W H 1988 Hydrogenolysis of ethane, propane, n-butane and neopentane

over the (111) and (110)-(1 × 2) surfaces of iridium *J. Am. Chem. Soc.* **110** 8305

- [58] Engel T and Ertl G 1978 Elementary steps in the catalytic oxidation of carbon monoxide on platinum metals *Adv. Catal.* **28** 1
- [59] Berlowitz P J and Goodman D W 1988 Kinetics of CO oxidation on single crystal Pd, Pt, and Ir *J. Phys. Chem.* **92** 5213
-

-39-

- [60] Szanyi J and Goodman D W 1994 CO oxidation on palladium: 1. A combined kinetic-infrared reflection absorption spectroscopic study of Pd(111) *J. Phys. Chem.* **98** 2972
- [61] Oh S H, Fisher G B, Carpenter J E and Goodman D W 1986 Comparative kinetic studies of CO-O₂ and CO-NO reactions over single crystal and supported rhodium catalysts *J. Catal.* **100** 360
- [62] Berlowitz P J, Goodman D W, Peden C H F and Blair D S 1988 Kinetics of CO oxidation by O₂ or NO on Rh(111) and Rh(100) single crystals *J. Phys. Chem.* **92** 1563
- [63] Blyholder G 1964 Molecular orbital view of chemisorbed carbon monoxide *J. Phys. Chem.* **68** 2772
-

-1-

A3.11 Quantum mechanics of interacting systems: scattering theory

George C Schatz

A3.11.1 INTRODUCTION

Quantum scattering theory is concerned with transitions between states which have a continuous energy spectrum, i.e., which are unbound. The most common application of scattering theory in chemical physics is to collisions involving atoms, molecules and/or electrons. Such collisions can produce many possible results, ranging from elastic scattering to reaction and fragmentation. Scattering theory can also be used to describe collisions of atoms, molecules and/or electrons with solid surfaces and it also has application to many kinds of dynamical process in solids. These latter include collisions of conduction electrons in a metal with impurities or with particle surfaces, or collisions of collective wave motions such as phonons with impurities, or adsorbates. Scattering theory is also involved in describing the interaction of light with matter, including applications to elastic and inelastic light scattering, photoabsorption and emission. Additionally, there are many processes where continuum states of particles are coupled to continuum states of electromagnetic radiation, including photodissociation of molecules and photoemission from surfaces.

While the basic formalism of quantum scattering theory can be found in a variety of general physics textbooks [1, 2, 3, 4, 5, 6 and 7] and textbooks that are concerned with scattering theory in a broad sense [8, 9, 10 and 11], many problems in chemical physics require special adaptation of the theory. For example, in collisions of particles with surfaces, angular momentum conservation is not important, but linear momentum conservation can be crucial. Also, in many collision problems involving atoms and molecules, the de Broglie wavelength is short compared to the distances over which the particles interact strongly, making classical or semiclassical theory useful. One especially important feature associated with scattering theory applications in chemical physics is that the forces between the interacting particles can usually be determined with reasonable accuracy (in principle to arbitrary accuracy), so explicit forms for the Hamiltonian governing particle motions are

available. Often these forces are quite complicated, so it is not possible to develop analytical solutions to the scattering theory problem. However, numerical solutions are possible, so a significant activity among researchers in this field is the development of numerical methods for solving scattering problems. There are a number of textbooks which consider scattering theory applications of more direct relevance to problems in chemical physics [[12](#), [13](#), [14](#), [15](#), [16](#), [17](#) and [18](#)], as well as numerous monographs that have a narrower focus within the field [[19](#), [20](#), [21](#), [22](#), [23](#), [24](#), [25](#), [26](#), [27](#), [28](#) and [29](#)].

Much of what one needs to know about scattering theory can be understood by considering a particle moving in one dimension governed by a potential that allows it to move freely except for a range of coordinates where there is a feature such as a barrier or well that perturbs the free particle motion. Our discussion will therefore begin with this simple problem ([section A3.11.2](#)). Subsequently ([section A3.11.3](#)) more complete versions of scattering theory will be developed that apply to collisions involving particles having internal degrees of freedom. There are both time dependent and time independent versions of scattering theory, and both of these theories will be considered. In [section A3.11.4](#), the numerical methods that are used to calculate scattering theory properties are considered, including time dependent and independent approaches. Also presented ([section A3.11.5](#)) are scattering theory methods for determining information that has been summed and averaged over many degrees of freedom (such as in a Boltzmann distribution).

-2-

Finally, in [section A3.11.6](#), scattering theory methods based on classical and semiclassical mechanics are described.

There are a variety of topics that will not be considered, but it is appropriate to provide references for further reading. The development in this paper assumes that the Born–Oppenheimer approximation applies in collisions between atoms and molecules and thus the nuclear motion is governed by a single potential energy surface. However there are many important problems where this approximation breaks down and multiple coupled potential energy surfaces are involved, with nonadiabatic transitions taking place during the scattering process. The theory of such processes is described in many places, such as [[14](#), [15](#), [23](#), [25](#)].

Other topics that have been omitted include the description of scattering processes using Feynman path integrals [[18](#), [19](#)] and the description of scattering processes with more than two coupled continua (i.e., where three or more independent particles are produced, as in electron impact ionization [[30](#)] or collision induced dissociation) [[31](#)]. Our treatment of resonance effects in scattering processes (i.e., the formation of metastable intermediate states) is very brief as this topic is commonly found in textbooks and one monograph is available [[26](#)]. Finally, it should be mentioned that the theory of light scattering is not considered; interested readers should consult textbooks such as that by Newton [[8](#)].

A3.11.2 QUANTUM SCATTERING THEORY FOR A ONE-DIMENSIONAL POTENTIAL FUNCTION

A3.11.2.1 HAMILTONIAN; BOUNDARY CONDITIONS

The problem of interest in this section is defined by the simple one-dimensional Hamiltonian

$$\hat{H} = \frac{\hat{p}^2}{2m} + V(x) \tag{A3.11.1}$$

where $V(x)$ is the potential energy function, examples of which are pictured in [figure A3.11.1](#). The potentials

shown are of two general types: those which are constant in the limit of $x \rightarrow \pm\infty$ [figure A3.11.1\(a\)](#) and [figure A3.11.1\(b\)](#), and those which are constant in the limit of $x \rightarrow -\infty$ and are infinite in the limit of $x \rightarrow +\infty$ [figure A3.11.1\(c\)](#) (of course this potential could be flipped around if one wants). In the former case, one can have particles moving at constant velocity in both asymptotic limits ($x \rightarrow \pm\infty$), so there are two physically distinct processes that can be described, namely, scattering in which the particle is initially moving to the right in the limit $x \rightarrow -\infty$, and scattering in which the particle is initially moving to the left in the limit $x \rightarrow +\infty$. In the latter case [figure A3.11.1\(c\)](#), the only physically interesting situation involves the particle initially moving to the left in the limit $x \rightarrow +\infty$. The former case is appropriate for describing a chemical reaction where there is either a barrier [figure A3.11.1\(a\)](#) or a well [figure A3.11.1\(b\)](#). It is also relevant to the scattering of an electron from the surface of a metal, where either transmission or reflection can occur. In [figure A3.11.1\(c\)](#), only reflection can occur, such as happens in elastic collisions of atoms, or low energy collisions of molecules with surfaces.

-3-

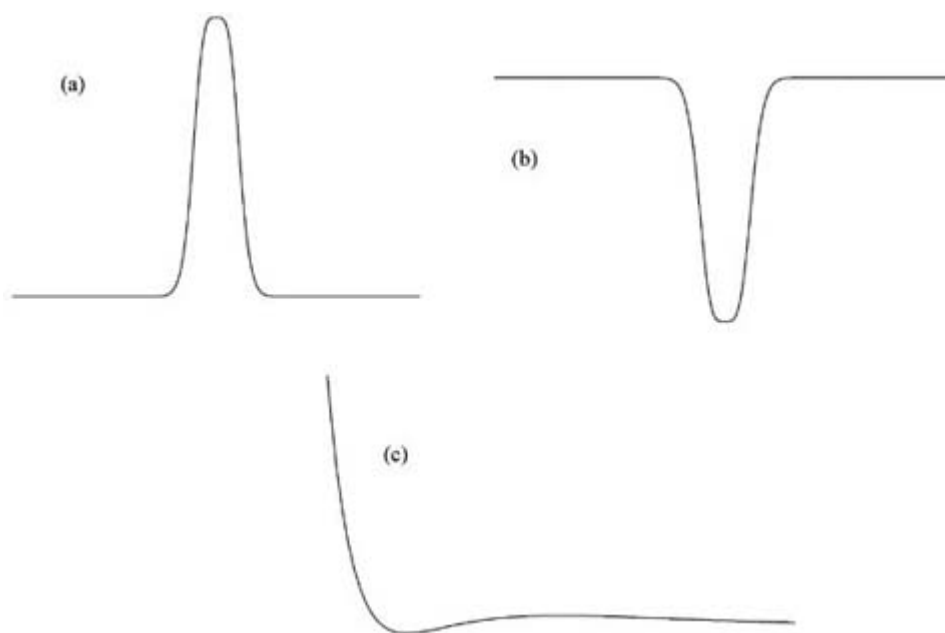


Figure A3.11.1. Potential associated with the scattering of a particle in one dimension. The three cases shown are (a) barrier potential, (b) well potential and (c) scattering off a hard wall that contains an intermediate well.

The physical question to be answered for [figure A3.11.1\(a\)](#) and [figure A3.11.1\(b\)](#) is: what is the probability P that a particle incident with an energy E from the left at $x \rightarrow -\infty$ will end up moving to the right at $x \rightarrow +\infty$? In the case of [figure A3.11.1\(c\)](#) only reflection can occur. However the change in phase of the wavefunction that occurs in this reflection is often of interest. In the following treatment the detailed theory associated with [figure A3.11.1\(a\)](#) and [figure A3.11.1\(b\)](#) will be considered. Eventually we will see that [figure A3.11.1\(c\)](#) is a subset of this theory.

The classical expression for the transmission probability associated with [figure A3.11.1\(a\)](#) or [figure A3.11.1\(b\)](#) is straightforward, namely

(1) $P(E) = 0$ if $V(x) \geq E$ for any x

(2) $P(E) = 1$ if $V(x) < E$ for all x .

The quantum solution to this problem is much more difficult for a number of reasons. First, it is important to know how to define what we mean by a particle moving in a given direction when $V(x)$ is constant. Secondly, one must determine the probability that the particle is moving in any specified direction at any desired

location and, third, we need to be able to solve the Schrödinger equation for the potential $V(x)$.

A3.11.2.2 WAVEPACKETS IN ONE DIMENSION

To understand how to describe a particle moving in a constant potential, consider the case of a free particle for which $V(x) = 0$. In this case the time-dependent Schrödinger equation is

-4-

$$i\hbar \frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial x^2} \quad (\text{A3.11.2})$$

and if one invokes the usual procedure for separating the time and spatial parts of this equation, it can readily be shown that one possible solution is

$$\psi_k(x, t) = e^{-iEt/\hbar} e^{ikx} \quad (\text{A3.11.3})$$

where

$$E = \frac{\hbar^2 k^2}{2m} \quad (\text{A3.11.4})$$

is the particle's energy and $\hbar k$ is its linear momentum. Note that both energy and momentum of the particle are exactly specified in this solution. As might be expected from the uncertainty principle, the location of the particle is therefore completely undetermined. As a result, this solution to the Schrödinger equation, even though correct, is not useful for describing the scattering processes of interest.

In order to localize the particle, it is necessary to superimpose wavefunctions ψ^k with different momenta k . A very general way to do this is to construct a wavepacket, defined through the integral

$$\begin{aligned} \psi_{\text{wp}}(x, t) &= \int_{-\infty}^{\infty} dk C(k) \psi_k(x, t) \\ &= \int_{-\infty}^{\infty} dk C(k) e^{ikx} e^{-i\hbar k^2 t/2m} \end{aligned} \quad (\text{A3.11.5})$$

where $C(k)$ is a function which tells us how much of each momentum $\hbar k$ is contained in the wavepacket. If the particle is to move with roughly a constant velocity, $C(k)$ must be peaked at some k which is taken to be k_0 .

One function which accomplishes this is the Gaussian

$$C(k) = \sqrt{\frac{a}{2\pi^{3/2}}} \exp[-a^2(k - k_0)^2/2] \quad (\text{A3.11.6})$$

where a measures the width of the packet. Substituting this into equation (A3.11.5), the result is:

$$(\text{A3.11.7})$$

$$\psi_{\text{wp}}(x, t) = \pi^{-1/4} [a(1 + i\hbar t/ma^2)]^{-1/2} \exp \left[-\frac{(x - \hbar k_0 t/m)^2}{2a^2(1 + i\hbar t/ma^2)} + ik_0 x - \frac{i\hbar t}{2ma^2} \right].$$

-5-

The absolute square of this wavefunction is $|\psi_{\text{wp}}(x, t)|^2$

$$|\psi_{\text{wp}}|^2 = \pi^{-1/2} a^{-1} (1 + \hbar^2 t^2/m^2 a^4)^{-1/2} \exp \left[-\frac{(x - \hbar k_0 t/m)^2}{a^2(1 + \hbar^2 t^2/m^2 a^4)} \right]. \quad (\text{A3.11.8})$$

This is a Gaussian function which peaks at $x = \hbar k_0 t/m$, moving to the right with a momentum $\hbar k_0$. The width of this peak is

$$\Delta = a[(\ln 2)(1 + \hbar^2 t^2/m^2 a^4)]^{1/2} \quad (\text{A3.11.9})$$

which starts out at $\Delta = a(\ln 2)^{1/2}$ at $t = 0$ and increases linearly with time for large t . This increase in width means that wavepacket spreads as it moves. This is an inevitable consequence of the fact that the wavepacket was constructed with a distribution of momentum components, and is a natural consequence of the uncertainty principle. Note that the wavefunction in [equation \(A3.11.7\)](#) still satisfies the Schrödinger [equation \(A3.11.2\)](#).

One can show that the expectation value of the Hamiltonian operator for the wavepacket in [equation \(A3.11.7\)](#) is:

$$\langle \hat{H} \rangle = \frac{\hbar^2 k_0^2}{2m} + \frac{\hbar^2}{4ma^2}. \quad (\text{A3.11.10})$$

The first term is what one would expect to obtain classically for a particle of momentum $\hbar k_0$, and it is much bigger than the second term provided $k_0 a \gg 1$. Since the de Broglie wavelength λ is $2\pi/k_0$, this condition is equivalent to the statement that the size of the wavepacket be much larger than the de Broglie wavelength.

It is also notable that the spreading of the wavepacket can be neglected for times t such that $t \ll ma^2/\hbar$. In this time interval the centre of the wavepacket will have moved a distance $(k_0 a)a$. Under the conditions noted above for which $k_0 a \gg 1$, this distance will be many times larger than the width of the packet.

A3.11.2.3 WAVEPACKETS FOR THE COMPLETE SCATTERING PROBLEM

The generalization of the treatment of the previous section to the determination of a wavepacket for the Hamiltonian in [equation \(A3.11.1\)](#) is accomplished by writing the solution as follows:

$$\psi_{\text{wp}}(x, t) = \int_{-\infty}^{\infty} dk C(k) \psi_k(x) e^{-iE_k t/\hbar} \quad (\text{A3.11.11})$$

where ψ_k is the solution of the time-independent Schrödinger equation

-6-

$$\hat{H}\psi_k = E_k\psi_k \quad (\text{A3.11.12})$$

for an energy E_k . By substituting [equation \(A3.11.11\)](#) into the time-dependent Schrödinger equation one can readily show that ψ_{wp} is a solution.

However, it is important to make sure that ψ_{wp} satisfies the desired boundary conditions initially and finally. Part of this is familiar already, since we have already demonstrated in [equation \(A3.11.3\)](#), [equation \(A3.11.5\)](#) and [equation \(A3.11.7\)](#) that use of $\psi_k = e^{ikx}$ and a Gaussian $C(k)$ gives a Gaussian wavepacket which moves with momentum $\hbar k_0$. This is the behaviour that is of interest initially ($t \rightarrow -\infty$) in the limit of $x \rightarrow -\infty$.

At the end of the collision ($t \rightarrow +\infty$) one expects to see part of the wavepacket moving to the right for $x \rightarrow \infty$ (the transmitted part) and part of it moving to the left for $x \rightarrow -\infty$ (the reflected part). Both this and the $t \rightarrow -\infty$ boundary condition can be satisfied by requiring that

$$\psi_k(x) \underset{x \rightarrow -\infty}{=} e^{ikx} + R e^{-ikx} \quad (\text{A3.11.13a})$$

$$\underset{x \rightarrow +\infty}{=} T e^{-i\bar{k}x} \quad (\text{A3.11.13b})$$

where R and T are as yet undetermined coefficients that will be discussed later and $\bar{k} = (2m(E - V_0)/\hbar^2)^{1/2}$ where V_0 is the value of the potential in the limit $x \rightarrow \infty$. Note that V_0 specifies the energy difference between the potential in the right and left asymptotic limits, and it has been assumed that $E > V_0$, as otherwise there could not be travelling waves in the $x \rightarrow \infty$ limit.

To prove that [equation \(A3.11.13\)](#) gives a wavepacket which satisfies the desired boundary conditions, we note that substitution of [equation \(3.11.13\)](#) into [equation \(A3.11.11\)](#) gives us two wavepackets which roughly speaking are given by

$$\psi_{\text{wp}} \underset{x \rightarrow -\infty}{\approx} e^{-(x - \hbar k_0 t/m)^2/2a^2} + R e^{-(x + \hbar k_0 t/m)^2/2a^2}. \quad (\text{A3.11.14a})$$

In the $t \rightarrow -\infty$ limit, only the first term, representing a packet moving to the right, has a peak in the $x \rightarrow -\infty$ region (the left asymptotic region). The second term peaks in the right asymptotic region but this is irrelevant as [equation A3.11.14](#) does not apply there. Thus, in the left asymptotic region the second term is negligible and all we have is a packet moving to the right. For $t \rightarrow +\infty$, [equation A3.11.14](#) still applies in the left asymptotic region, but now it is the second term which peaks and this packet moves to the left.

Now substitute [equation A3.11.13](#) into [equation \(A3.11.11\)](#). Ignoring various unimportant terms, we obtain

$$\psi_{\text{wp}} \underset{x \rightarrow +\infty}{\approx} T e^{-(x - \hbar \bar{k}_0 t/m)^2/2a^2}. \quad (\text{A3.11.14b})$$

This formula represents a packet moving to the right centred at $x = \hbar \bar{k}_0 t/m$. For $t \rightarrow -\infty$, this is negligible in the right asymptotic region, so the wavefunction is zero there, while for $t \rightarrow +\infty$ this packet is large for $x \rightarrow +\infty$ just as we wanted.

A3.11.2.4 FLUXES AND PROBABILITIES

Now let us use the wavepackets just discussed to extract the physically measurable information about our problem, namely, the probabilities of reflection and transmission. As long as the wavepackets do not spread much during the collision, these probabilities are given by the general definition:

$$\text{probability} = \frac{|\text{total flux outgoing for process of interest}|}{|\text{total flux incident}|} \quad (\text{A3.11.15})$$

where the flux is the number of particles per unit time that cross a given point (that cross a given surface in three dimensions), and the total flux is the spatial integral of the instantaneous flux. Classically the flux is just ρv where ρ is the density of particles (particles per unit length in one dimension) and v is the velocity of the particles. In quantum mechanics, the flux I is defined as

$$I = \text{Re}[\psi^* \hat{v} \psi] \quad (\text{A3.11.16})$$

where \hat{v} is the velocity operator ($\hat{v} = (-i\hbar/m)\partial/\partial x$ in one dimension) and Re implies that only the real part of $\psi^* \hat{v} \psi$ is to be used.

To see how equation (A3.11.16) works, substitute [equation \(A3.11.7\)](#) into (A3.11.16). Under the condition that wavepacket spreading is small (i.e., $\hbar t/ma^2 \ll 1$) we obtain

$$I = \left(\frac{\hbar k_0}{m} \right) \pi^{-1/2} a^{-1} \exp[-(x - \hbar k_0 t/m)^2/a^2] \quad (\text{A3.11.17})$$

which is just $v_0 |\psi_{\text{wp}}|^2$ where v_0 is the initial most probable velocity ($v_0 = \hbar k_0/m$). In view of [equation A3.11.14\(a\)](#), this is just the incident flux. The integral of this quantity over all space (the total flux) is $I_{\text{tot}}^{\text{inc}} = v_0$.

For the reflected wave associated with [equation \(A3.11.13a\)](#), the total outgoing flux is $I_{\text{tot}}^{\text{out}} = |R|^2 v_0$ so the reflection probability P_R is

$$P_R = |R|^2. \quad (\text{A3.11.18a})$$

-8-

A similar calculation of the transmission probability gives

$$P_T = \frac{\bar{v}_0}{v_0} |T|^2 \quad (\text{A3.11.18b})$$

where

$$\bar{v}_0 \equiv \frac{\hbar \bar{k}}{m}. \quad (\text{A3.11.19})$$

A3.11.2.5 TIME-INDEPENDENT APPROACH TO SCATTERING

Note from [equation \(A3.11.18a\)](#), [equation \(A3.11.18\)](#) that all of the physically interesting information about the scattering process involves the coefficients R and T which are properties of the time *independent* wavefunction ψ_k obtained from [equations \(A3.11.12\)](#) with the boundary conditions in [equations \(A3.11.13\)](#). As a result, we can use scattering theory completely in a time independent picture. This picture can be thought of as related to the time dependent picture by the superposition of many Gaussian incident wavepackets to form a plane wave. The important point to remember in using time independent solutions is that the asymptotic solution given by [equations \(A3.11.13\)](#) involves waves moving to the left and right that should be treated *separately* in calculating fluxes since these solutions do not contribute at the same time to the evolution of $\psi_{\text{wp}}(x, t)$ in the $t \rightarrow \pm\infty$ limits. As a result, fluxes are evaluated by substituting either the left or right moving wavepacket parts of [equations \(A3.11.13\)](#) into [\(A3.11.16\)](#).

A3.11.2.6 SCATTERING MATRIX

It is useful to rewrite the asymptotic part of the wavefunction as

$$\psi_k(x) \underset{x \rightarrow -\infty}{=} e^{ikx} + S_{11} e^{-ikx} \quad (\text{A3.11.20a})$$

$$\underset{x \rightarrow +\infty}{=} S_{12} (k/\bar{k})^{1/2} e^{-i\bar{k}x} \quad (\text{A3.11.20b})$$

where the coefficients S_{11} and S_{12} are two elements of a 2×2 matrix known as the scattering (S) matrix. The other two elements are associated with a different scattering solution in which the incident wave at $t \rightarrow -\infty$ moves to the left in the $x \rightarrow +\infty$ region. The boundary conditions on this solution are

$$\begin{aligned} \psi_{\bar{k}} \underset{x \rightarrow +\infty}{=} e^{-i\bar{k}x} + S_{22} e^{+i\bar{k}x} \\ \underset{x \rightarrow -\infty}{=} S_{21} (\bar{k}/k)^{1/2} e^{-i\bar{k}x}. \end{aligned} \quad (\text{A3.11.21})$$

-9-

The S matrix has a number of important properties, one of which is that it is *unitary*. Mathematically this means that $\mathbf{S}^+ \mathbf{S} = 1$ where \mathbf{S}^+ is the Hermitian conjugate (transpose of complex conjugate) of \mathbf{S} . This property comes from the equation of continuity, which says that for any solution ψ to the time dependent Schrödinger equation,

$$\frac{\partial |\psi|^2}{\partial t} + \frac{\partial I}{\partial x} = 0 \quad (\text{A3.11.22})$$

where I is the flux from [equation \(A3.11.16\)](#). [equation \(A3.11.22\)](#) can be proved by substitution of [equation \(A3.11.16\)](#) and the time dependent Schrödinger equation into [\(A3.11.22\)](#).

If $\psi = \psi_k(x)e^{iEt/\hbar}$, $|\psi|^2$ is time independent, so equation (A3.11.22) reduces to $\partial I/\partial x = 0$, which implies I is a constant (i.e., flux is conserved), independent of x . If so then the evaluation of I at $x \rightarrow +\infty$ and at $x \rightarrow -\infty$ should give the same result. By directly substituting equations (A3.11.13) into (A3.11.16) one finds

$$\begin{aligned} I &= \frac{\hbar k}{m} (1 - |S_{11}|^2) \\ &= \frac{\hbar k}{m} |S_{12}|^2 \end{aligned} \quad (\text{A3.11.23})$$

and since these two have to be equal, we find that

$$|S_{11}|^2 + |S_{12}|^2 = 1 \quad (\text{A3.11.24})$$

which indicates that the sum of the reflected and transmitted probabilities has to be unity. This is one of the equations that is implied by unitarity of the S matrix. The other equations can be obtained by using the solution ψ_k (equation (A3.11.21)) and by using a generalized flux that is defined by

$$I_{k\bar{k}} = \text{Re}(\psi_k^* \hat{v} \psi_{\bar{k}}). \quad (\text{A3.11.25})$$

Another useful property of the S matrix is that it is *symmetric*. This property follows from conservation of the fluxlike expression

$$\tilde{I}_{k\bar{k}} = \text{Re}(\psi_k v \psi_{\bar{k}}) \quad (\text{A3.11.26})$$

which differs from equation (A3.11.25) in the absence of a complex conjugate in the wavefunction ψ_k . The symmetry property of \mathbf{S} implies that S_{12} in equation (A3.11.20b) equals S_{21} in equation (A3.11.21). Defining the probability matrix \mathbf{P} by the relation

$$P_{ij} = |S_{ij}|^2 \quad (\text{A3.11.27})$$

we see that symmetry of \mathbf{S} implies equal probabilities for the $i \rightarrow j$ and $j \rightarrow i$ transitions. This is a statement of the principle of *microscopic reversibility* and it arises from the time reversal symmetry associated with the Schrödinger equation.

The probability matrix plays an important role in many processes in chemical physics. For chemical reactions, the probability of reaction is often limited by tunnelling through a barrier, or by the formation of metastable states (resonances) in an intermediate well. Equivalently, the conductivity of a molecular wire is related to the probability of transmission of conduction electrons through the junction region between the wire and the electrodes to which the wire is attached.

A3.11.2.7 GREEN'S FUNCTIONS FOR SCATTERING

Now let us write down the Schrödinger equation (A3.11.12) using equation (A3.11.1) for H and assuming that V_0 in figure A3.11.1 is zero. The result can be written

$$\left(\frac{d^2}{dx^2} + k^2\right) \psi_k(x) = \frac{2m}{\hbar^2} V(x) \psi_k(x). \quad (\text{A3.11.28})$$

One way to solve this is to invert the operator on the left hand side, thereby converting this differential equation into an integral equation. The general result is

$$\psi_k(x) = \varphi_k(x) + \frac{2m}{\hbar^2} \int_{-\infty}^{\infty} G_0(x, x') V(x') \psi_k(x') dx' \quad (\text{A3.11.29})$$

where G_0 is called the Green function associated with the operator $d^2/dx^2 + k^2$ and φ_k is a solution of the homogeneous equation that is associated with equation (A3.11.28), namely

$$\left(\frac{d^2}{dx^2} + k^2\right) \varphi_k(x) = 0. \quad (\text{A3.11.30})$$

To determine $G_0(x, x')$, it is customary to reexpress equation (A3.11.28) in a Fourier representation. Let $F_k(k')$ be the Fourier transform of $\psi_k(x)$. Taking the Fourier transform of equation (A3.11.28), we find

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{ik'x} (k^2 - k'^2) F_k(k') dk' = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{ik'x} B(k') dk' \quad (\text{A3.11.31})$$

-11-

where

$$B(k') = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-ik'x} \frac{2m}{\hbar^2} V(x) \psi_k(x) dx. \quad (\text{A3.11.32})$$

Equation (A3.11.31) implies

$$F_k(k') = \frac{B(k')}{k^2 - k'^2} \quad (\text{A3.11.33})$$

and upon inverting the Fourier transform we find

$$\psi_k(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{ik'x} \frac{B(k')}{k^2 - k'^2} dk' = \frac{2m}{\hbar^2} \int_{-\infty}^{\infty} G_0(x, x') V(x') \psi_k(x') dx' \quad (\text{A3.11.34})$$

where the Green function is given by

$$G_0(x, x') = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ik'(x-x')} (k^2 - k'^2)^{-1} dk'. \quad (\text{A3.11.35})$$

The evaluation of the integral in equation (A3.11.35) needs to be done carefully as there is a pole at $k' = \pm k$. A standard trick to do it involves replacing k by $k \pm i\varepsilon$ where ε is a small positive constant that will be set to zero in the end. This reduces equation (A3.11.35) to

$$G_0(x, x') = \frac{1}{4\pi k} \lim_{\varepsilon \rightarrow 0} \int_{-\infty}^{\infty} \exp[ik'(x - x')] \left(\frac{1}{k - k' \pm i\varepsilon} + \frac{1}{k + k' \pm i\varepsilon} \right) dk'. \quad (\text{A3.11.36})$$

This integral can be done by contour integration using the contours in figure A3.11.2. For the $+i\varepsilon$ choice, the contour in figure A3.11.2(a) is appropriate for $x < x'$ as the circular part has a negative imaginary k' which makes $e^{ik'(\xi - \xi')}$ vanish for $|k'| \rightarrow \infty$. Likewise for $x > x'$, we want to use the contour in figure A3.11.2(b) as this makes the imaginary part of k' positive along the circular part. In either case, the integral along the real axis equals the full contour integral, and the latter is determined by the residue theorem to be $2\pi i$ times the residue at the pole which is encircled by the contour.

-12-

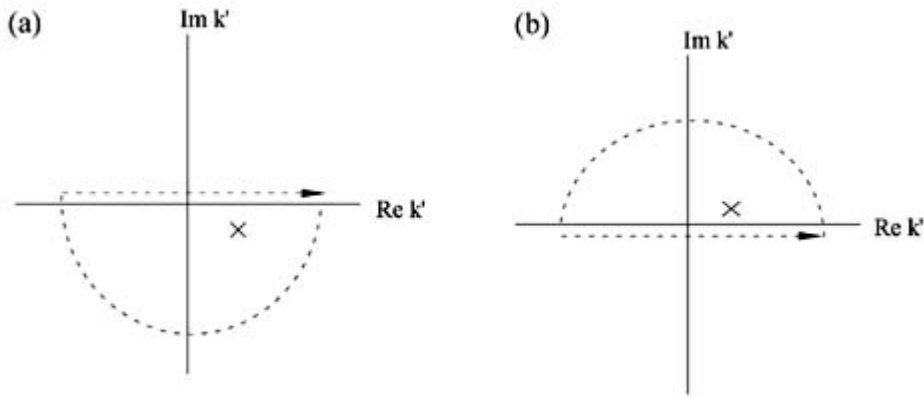


Figure A3.11.2. Integration contours used to evaluate equation (A3.11.36) (a) for $x < x'$, (b) for $x > x'$.

The pole is at $k' = -k - i\varepsilon$ for the contour in figure A3.11.2(a) and at $k' = k + i\varepsilon$ for figure A3.11.2(b). This gives us

$$G_0^+(x, x') = \begin{cases} \left(\frac{-i}{2k} \right) e^{-ik(x-x')} & \text{for } x < x' \\ \left(\frac{-i}{2k} \right) e^{ik(x-x')} & \text{for } x > x' \end{cases} \quad (\text{A3.11.37})$$

which we will call the ‘plus’ wave free particle Green function G_0^+ . A different Green function (‘minus’ wave) is obtained by using $-i\varepsilon$ in the above formulas. It is

$$G_0^-(x, x') = \begin{cases} \left(\frac{i}{2k} \right) e^{ik(x-x')} & \text{for } x < x' \\ \left(\frac{i}{2k} \right) e^{-ik(x-x')} & \text{for } x > x'. \end{cases} \quad (\text{A3.11.38})$$

Upon substitution of G_0 into equation (A3.11.29) we generate the following integral equation for the solution ψ_k^* that is associated with G_0^+ :

$$(\text{A3.11.39})$$

$$\begin{aligned}\psi_k^+(x) = \varphi_k(x) - \int_{-\infty}^x \left(\frac{i}{2k}\right) e^{ik(x-x')} \frac{2m}{\hbar^2} V(x') \psi_k^+(x') dx' \\ - \int_x^{\infty} \left(\frac{i}{2k}\right) e^{-ik(x-x')} \frac{2m}{\hbar^2} V(x') \psi_k^+(x') dx' .\end{aligned}$$

For $x \rightarrow \pm\infty$, it is possible to make ψ_k^+ look like [equation \(A3.11.20\)](#) by setting $\varphi_k(x) = e^{ikx}$. This shows that the plus Green function is associated with scattering solutions in which outgoing waves move to the right in the $x \rightarrow \infty$ limit. For $x \rightarrow -\infty$, [equation \(A3.11.39\)](#) becomes

$$\psi_k^+ \underset{x \rightarrow -\infty}{=} e^{ikx} - e^{-ikx} \int_{-\infty}^{\infty} \left(\frac{i}{2k}\right) e^{ikx'} \frac{2m}{\hbar^2} V(x') \psi_k^+(x') dx' . \quad (\text{A3.11.40})$$

-13-

By comparison with [equation A3.11.20\(a\)](#), we see that

$$S_{11} = -\frac{i}{2k} \int_{-\infty}^{\infty} e^{ikx'} \frac{2m}{\hbar^2} V(x') \psi_k^+(x') dx' \quad (\text{A3.11.41})$$

which is an integral that can be used to calculate S_{11} provided that ψ_k^+ is known. One can similarly show that

$$S_{12} = 1 - \frac{i}{2k} \int_{-\infty}^{\infty} e^{ikx'} \frac{2m}{\hbar^2} V(x') \psi_k^+(x') dx' . \quad (\text{A3.11.42})$$

The other S matrix components S_{21} and S_{22} can be obtained from the G_0^- Green function.

A3.11.2.8 BORN APPROXIMATION

If $V(x)$ is ‘small’, ψ_k^+ will not be perturbed much from what it would be if $V(x) = 0$. If so, then we can approximate ψ_k^+ and obtain

$$S_{11} = -\frac{i}{2k} \int_{-\infty}^{\infty} e^{2ikx'} \frac{2m}{\hbar^2} V(x') dx' \quad (\text{A3.11.43a})$$

$$S_{12} = 1 - \frac{i}{2k} \int_{-\infty}^{\infty} \frac{2m}{\hbar^2} V(x') dx' . \quad (\text{A3.11.43b})$$

This is the one dimensional version of what is usually called the *Born approximation* in scattering theory. The transition probability obtained from [equation A3.11.43\(b\)](#) is

$$P_{11} = \frac{m^2}{\hbar^2 p^2} \left| \int_{-\infty}^{\infty} e^{2ikx'} V(x') dx' \right|^2 \quad (\text{A3.11.44})$$

where $p = \hbar k$ is the momentum. Note that this approximation simplifies the evaluation of transition probabilities to performing an integral.

A number of improvements to the Born approximation are possible, including *higher order* Born approximations (obtained by inserting lower order approximations to ψ_k^* into equation (A3.11.40), then the result into (A3.11.41) and (A3.11.42)), and the *distorted wave* Born approximation (obtained by replacing the free particle approximation for ψ_k^* by the solution to a Schrödinger equation that includes part of the interaction potential). For chemical physics

-14-

applications, the distorted wave Born approximation is the most often used approach, as the approximation of ψ_k^* by a plane wave is rarely of sufficient accuracy to be even qualitatively useful. However, even the distorted wave Born approximation is poorly convergent for many applications, so other exact and approximate methods need to be considered.

A3.11.2.9 VARIATIONAL METHODS

A completely different approach to scattering involves writing down an expression that can be used to obtain S directly from the wavefunction, and which is stationary with respect to small errors in the wavefunction. In this case one can obtain the scattering matrix element by variational theory. A recent review of this topic has been given by Miller [32]. There are many different expressions that give S as a functional of the wavefunction and, therefore, there are many different variational theories. This section describes the Kohn variational theory, which has proven particularly useful in many applications in chemical reaction dynamics. To keep the derivation as simple as possible, we restrict our consideration to potentials of the type plotted in figure A3.11.1(c) where the wavefunction vanishes in the limit of $x \rightarrow -\infty$, and where the S matrix is a scalar property so we can drop the matrix notation.

The Kohn variational approximation states that for a trial wavefunction $\tilde{\Psi}$ which has the asymptotic form

$$\tilde{\Psi} \underset{x \rightarrow \infty}{\sim} v^{-1/2} (e^{-ikx} - e^{-ikx} \tilde{S}) \quad (\text{A3.11.45})$$

that the quantity

$$S = \tilde{S} + \frac{i}{\hbar} \langle \tilde{\Psi} | \hat{H} - E | \tilde{\Psi} \rangle \quad (\text{A3.11.46})$$

is stationary with respect to variations in $\tilde{\Psi}$, and $S = S_{\text{ex}}$ where S_{ex} is the exact scattering matrix when $\tilde{\Psi} = \psi_{\text{exact}}$. Note that $\tilde{\Psi}$ is not complex conjugated in calculating $\langle \tilde{\Psi} |$.

To prove this we expand $\tilde{\Psi}$ about the exact wavefunction ψ_{ex} , that is, we let

$$\tilde{\Psi} = \psi_{\text{ex}} + \delta\Psi. \quad (\text{A3.11.47})$$

ψ_{ex} here is assumed to have the asymptotic form

$$\Psi_{\text{ex}} \underset{x \rightarrow \infty}{\sim} -v^{-1/2}(e^{-ikx} - e^{-ikx} S_{\text{ex}}). \quad (\text{A3.11.48})$$

-15-

This means that

$$S\Psi \underset{x \rightarrow \infty}{\sim} v^{-1/2} e^{-ikx} \delta S \quad (\text{A3.11.49})$$

where $\delta S = S - S_{\text{ex}}$. Then we see that

$$\begin{aligned} S &= S_{\text{ex}} + \delta S + \frac{i}{\hbar} \langle \Psi_{\text{ex}} + \delta\Psi | \hat{H} - E | \Psi_{\text{ex}} + \delta\Psi \rangle \\ &= S_{\text{ex}} + \delta S + \frac{i}{\hbar} \langle \Psi_{\text{ex}} | \hat{H} - E | \delta\Psi \rangle + \mathcal{O}(\delta\Psi^2) \end{aligned} \quad (\text{A3.11.50})$$

since $\langle \hat{H} - E | \Psi_{\text{ex}} \rangle = 0$.

Now use integration by parts twice to show that

$$\left\langle \Psi_{\text{ex}} \left| \frac{d^2}{dx^2} \right| \delta\Psi \right\rangle = +(\Psi_{\text{ex}} \delta\Psi' - \Psi_{\text{ex}}' \delta\Psi) \Big|_{-\infty}^{\infty} + \left\langle \delta\Psi \left| \frac{d^2}{dx^2} \right| \Psi_{\text{ex}} \right\rangle \quad (\text{A3.11.51})$$

which means that

$$\langle \Psi_{\text{ex}} | \hat{H} - E | \delta\Psi \rangle = \frac{-\hbar^2}{2\mu} (\Psi_{\text{ex}} \delta\Psi' - \Psi_{\text{ex}}' \delta\Psi) \Big|_{-\infty}^{\infty} + \langle \delta\Psi | \hat{H} - E | \Psi_{\text{ex}} \rangle. \quad (\text{A3.11.52})$$

The last term vanishes, and so does the first at $x = -\infty$. The nonzero part is then

$$\begin{aligned} \frac{-\hbar^2}{2\mu} (-v^{-1/2}(e^{-ikx} - e^{ikx} S_{\text{ex}})v^{-1/2}(ik) e^{-ikx} \delta S + v^{-1/2}(-ik) \\ (e^{-ikx} + e^{ikx} S_{\text{ex}})v^{-1/2} e^{ikx} \delta S) \\ = \frac{-\hbar^2}{2\mu} \left(2 \frac{-ik}{v} \delta S \right) = i\hbar \delta S. \end{aligned} \quad (\text{A3.11.53})$$

So overall

$$S = S_{\text{ex}} + \delta S + \frac{i}{\hbar} (i\hbar \delta S) + \mathcal{O}(\delta\Psi^2) = S_{\text{ex}} + \mathcal{O}(\delta\Psi^2) \quad (\text{A3.11.54})$$

-16-

which means that the deviations from the exact result are of second order. This means that S is stationary with respect to variations in the trial function. Later ([section A3.11.4](#)) we will show how the variational approach can be used in practical applications where the scattering wavefunction is expanded in terms of basis functions.

A3.11.3 MULTICHANNEL QUANTUM SCATTERING THEORY; SCATTERING IN THREE DIMENSIONS

In this section we consider the generalization of quantum scattering theory to problems with many degrees of freedom, and to problems where the translational motion takes place in three dimensions rather than one. The simplest multidimensional generalization is to consider two degrees of freedom, and we will spend much of our development considering this, as it contains the essence of the complexity that can arise in what is called ‘multichannel’ scattering theory. Moreover, models containing two degrees of freedom are of use throughout the field of chemical physics. For example, this model can be used to describe the collision of an atom with a diatomic molecule with the three atoms constrained to be collinear so that only vibrational motion in the diatomic molecule needs to be considered in addition to translational motion of the atom relative to the molecule. This model is commonly used in studies of *vibrational energy transfer* [29] where the collision causes changes in the vibrational state of the molecule. In addition, this model can be used to describe *reactive* collisions wherein an atom is transferred to form a new diatomic molecule [23, 23 and 24]. We will discuss both of these processes in the following two sections ([A3.11.3.1](#) and [A3.11.3.2](#)).

The treatment of translational motion in three dimensions involves representation of particle motions in terms of plane waves $e^{i\mathbf{k}\cdot\mathbf{r}}$ where the wavevector \mathbf{k} specifies the direction of motion in addition to the magnitude of the velocity. For problems involving the motion of isolated particles, i.e., gas phase collisions, all problems can be represented in terms of eigenfunctions of the total angular momentum, which is a conserved quantity. The relationship between these eigenfunctions and the plane wave description of particle motions leads to the concept of a *partial wave expansion*, something that is used throughout the field of chemical physics. This is described in the third part of this [section \(A3.11.3.3\)](#).

Problems in chemical physics which involve the collision of a particle with a surface do not have rotational symmetry that leads to partial wave expansions. Instead they have two dimensional translational symmetry for motions parallel to the surface. This leads to expansion of solutions in terms of diffraction eigenfunctions. This theory is described in the literature [33].

A3.11.3.1 MULTICHANNEL SCATTERING—COUPLED CHANNEL EQUATIONS

Consider the collision of an atom (denoted A) with a diatomic molecule (denoted BC), with motion of the atoms constrained to occur along a line. In this case there are two important degrees of freedom, the distance R between the atom and the centre of mass of the diatomic, and the diatomic internuclear distance r . The Hamiltonian in terms of these coordinates is given by:

$$\hat{H} = \frac{\hat{p}_R^2}{2\mu_{A,BC}} + \frac{\hat{p}_r^2}{2\mu_{BC}} + V(R, r) \quad (\text{A3.11.55})$$

where $\mu_{A,BC}$ is the reduced mass associated with motion in the R coordinate, and μ_{BC} is the corresponding diatom reduced mass. Note that this Hamiltonian can be derived by starting with the Hamiltonian of the

independent atoms and separating out the motion of the centre of mass. The second form (A3.11.56) arises by replacing the momentum operators by their usual quantum mechanical expressions.

$$= -\frac{\hbar^2}{2\mu_{A,BC}} \frac{\partial^2}{\partial R^2} - \frac{\hbar^2}{2\mu_{BC}} \frac{\partial^2}{\partial r^2} + V(R, r) \quad (\text{A3.11.56})$$

We concentrate in this section on solving the time-independent Schrödinger equation, which, as we learned from [section A3.11.2.5](#), is all we need to do to generate the physically meaningful scattering information. If BC does not dissociate then it is reasonable to use the BC eigenfunctions as a basis for expanding the scattering wavefunction. Assume that as $R \rightarrow \infty$, $V(R, r) \rightarrow V_{BC}(r)$. Then the BC eigenfunctions are solutions to

$$\left(\frac{-\hbar^2}{2\mu_{BC}} \frac{d^2}{dr^2} + V_{BC}(r) \right) \varphi_v(r) = \epsilon_v \varphi_v(r) \quad (\text{A3.11.57})$$

where ϵ_v is the vibrational eigenvalue. The expansion of Ψ in terms of the BC eigenfunctions is thus given by

$$\Psi(R, r) = \sum_v \varphi_v(r) g_v(R) \quad (\text{A3.11.58})$$

where the g_v are unknown functions to be determined. This equation is called a *coupled channel* expansion. Substituting this into the Schrödinger equation, we find

$$\begin{aligned} \frac{-\hbar^2}{2\mu_{A,BC}} \sum_v \varphi_v(r) \frac{d^2 g_v}{dR^2} + \sum_v g_v(R) \left[\frac{-\hbar^2}{2\mu_{BC}} \frac{d^2}{dr^2} + V_{BC}(r) \right] \varphi_v(r) \\ + \sum_v (V(R, r) - V_{BC}(r)) g_v(R) \varphi_v(r) \\ = E \sum_v g_v(R) \varphi_v(r). \end{aligned} \quad (\text{A3.11.59})$$

Now rearrange, multiply by $\varphi_{v'}$, and integrate to obtain

$$\frac{-\hbar^2}{2\mu_{A,BC}} \frac{d^2 g_v}{dR^2} = (E - \epsilon_v) g_v - \sum_{v'} \langle \varphi_v | V - V_{BC} | \varphi_{v'} \rangle g_{v'} \quad (\text{A3.11.60})$$

or

$$\frac{d^2 g_v(R)}{dR^2} = \sum_{v'} U_{vv'}(R) g_{v'}(R) \quad (\text{A3.11.61})$$

where

$$U_{vv'} = \frac{2\mu_{A,BC}}{\hbar^2}(E - \epsilon_v) \delta_{vv'} + \frac{2\mu_{BC}}{\hbar^2} \langle \phi_v | V - V_{BC} | \phi_{v'} \rangle. \quad (\text{A3.11.62})$$

In matrix–vector form these *coupled-channel* equations are

$$\frac{d^2 \mathbf{g}}{dR^2} = \mathbf{U} \mathbf{g} \quad (\text{A3.11.63})$$

where \mathbf{g} is the vector formed using the g_v as elements and \mathbf{U} is a matrix whose elements are $U_{vv'}$. Note that the internal states may be either *open* or *closed*, depending on whether the energy E is above or below the internal energy ϵ_v . Only the open states (often termed *open channels*) have measurable scattering properties, but the closed channels can be populated as intermediates during the collision, sometimes with important physical consequences. In the following discussion we confine our discussion to the open channels. The boundary conditions on the open channel solutions are:

$$\mathbf{g}(R) \rightarrow \mathbf{0} \text{ as } R \rightarrow 0 \quad (\text{A3.11.64})$$

provided that the potential is repulsive at short range, and

$$\mathbf{g}(R) \rightarrow \mathbf{v}^{-1/2} (\mathbf{e}^{-ikR} - \mathbf{e}^{ikR} \mathbf{S}) \text{ as } R \rightarrow \infty. \quad (\text{A3.11.65})$$

Here we have collected the N independent g that correspond to different incoming states for N open channels into a matrix \mathbf{g} (where the *sans serif* bold notation is again used to denote a square matrix). Also we have the matrices

$$(\mathbf{v})_{vv'} = v_v \delta_{vv'} \quad (\text{A3.11.66})$$

$$(\mathbf{k})_{vv'} = k_v \delta_{vv'} \quad (\text{A3.11.67})$$

where

$$k_v = \frac{\sqrt{2\mu_{A,BC}}}{\hbar^2} (E - \epsilon_v) \quad (\text{A3.11.68})$$

-19-

$$v_v = \frac{\hbar k_v}{\mu_{A,BC}} \quad (\text{A3.11.69})$$

$$\mathbf{S} = S_{vv'}. \quad (\text{A3.11.70})$$

\mathbf{S} is the *scattering matrix*, analogous to that defined earlier. As before, the probabilities for transitions between states v and v' are

$$P_{vv'} = |S_{vv'}|^2. \quad (\text{A3.11.71})$$

Often in numerical calculations we determine solutions $\mathbf{g}(R)$ that solve the Schrödinger equations but do not satisfy the asymptotic boundary condition in (A3.11.65). To solve for \mathbf{S} , we rewrite equation (A3.11.65) and its derivative with respect to R in the more general form:

$$\mathbf{g} = (\mathbf{I} - \mathbf{O}\mathbf{S})\mathbf{A} \quad (\text{A3.11.72})$$

$$\mathbf{g}' = (\mathbf{I}' - \mathbf{O}'\mathbf{S})\mathbf{A} \quad (\text{A3.11.73})$$

where the incoming and outgoing asymptotic solutions are:

$$\mathbf{I} = \kappa^{-1/2} e^{-i\kappa R} \quad (\text{A3.11.74})$$

$$\mathbf{O} = \kappa^{-1/2} e^{i\kappa R}. \quad (\text{A3.11.75})$$

\mathbf{A} is a coefficient matrix that is designed to transform between solutions that obey arbitrary boundary conditions and those which obey the desired boundary conditions. \mathbf{A} and \mathbf{S} can be regarded as unknowns in equation (A3.11.72) and equation (A3.11.73). This leads to the following expression for \mathbf{S} :

$$\mathbf{S} = \mathbf{W}^{-1}(\mathbf{I}'\mathbf{g} - \mathbf{I}\mathbf{g}')(\mathbf{O}'\mathbf{g} - \mathbf{O}\mathbf{g}')^{-1}\mathbf{W} \quad (\text{A3.11.76})$$

where

$$\mathbf{W} = \mathbf{O}\mathbf{I}' - \mathbf{O}'\mathbf{I}. \quad (\text{A3.11.77})$$

The present derivation can easily be generalized to systems with an arbitrary number of internal degrees of freedom, and it leads to coupled channel equations identical with equation (A3.11.63), where the coupling terms (A3.11.62) are expressed as matrix elements of the interaction potential using states which depend on these internal degrees of

-20-

freedom. These internal states could, in principle, have a continuous spectrum but, in practice, if there are multiple continuous degrees of freedom then it is most useful to reformulate the problem to take this into account. One particularly important case of this sort arises in the treatment of reactive collisions, where the atom B is transferred from C to A, leading to the formation of a new arrangement of the atoms with its own scattering boundary conditions. We turn our attention to this situation in the next section.

A3.11.3.2 REACTIVE COLLISIONS

Let us continue with the atom–diatom collinear collision model, this time allowing for the possibility of the reaction $A + BC \rightarrow AB + C$. We first introduce *mass-scaled* coordinates, as these are especially convenient to describe rearrangements, using

$$(\text{A3.11.78})$$

$$R' = \left(\frac{\mu_{A,BC}}{m} \right)^{1/2} R$$

$$r' = \left(\frac{\mu_{BC}}{m} \right)^{1/2} r. \quad (\text{A3.11.79})$$

The choice of m in these formulas is arbitrary, but it is customary to take either $m = 1$ or

$$m = \sqrt{\frac{m_A m_B m_C}{m_A + m_B + m_C}} = \sqrt{\mu_{A,BC} \mu_{BC}}. \quad (\text{A3.11.80})$$

Either choice is invariant to permutation of the atom masses.

In terms of these coordinates, the Hamiltonian of [equation \(A3.11.55\)](#) becomes

$$H = \frac{\hat{P}_R^2}{2\mu_{A,BC}} + \frac{\hat{P}_r^2}{2\mu_{BC}} + V = \frac{\hat{P}_R'^2 + \hat{P}_r'^2}{2m} + V. \quad (\text{A3.11.81})$$

One nice thing about H in mass-scaled coordinates is that it is identical to the Hamiltonian of a mass point moving in two dimensions. This is convenient for visualizing trajectory motions or wavepackets, so the mass-scaled coordinates are commonly used for plotting data from scattering calculations.

Another reason why mass-scaled coordinates are useful is that they simplify the transformation to the Jacobi coordinates that are associated with the products $AB + C$. If we define S as the distance from C to the centre of mass of AB , and s as the AB distance, mass scaling is accomplished via

$$S' = \sqrt{\frac{\mu_{C,AB}}{m}} S \quad (\text{A3.11.82})$$

-21-

$$s' = \sqrt{\frac{\mu_{AB}}{m}} s. \quad (\text{A3.11.83})$$

The Hamiltonian in terms of product coordinates is

$$\hat{H} = \frac{\hat{P}_{S'}^2 + \hat{P}_{s'}^2}{2m} + V \quad (\text{A3.11.84})$$

and the transformation between reagent and product coordinates is given by:

$$\begin{aligned} S' &= r' \sin \beta + R' \cos \beta \\ s' &= -r' \cos \beta + R' \sin \beta \end{aligned} \quad (\text{A3.11.85})$$

where the angle β is defined by:

$$\tan \beta = \sqrt{\frac{m_B(m_A + m_B + m_C)}{m_A m_C}}. \quad (\text{A3.11.86})$$

Equation (A3.11.85) implies that the $R', r' \rightarrow S', s'$ transformation is orthogonal, a point which is responsible for the similarities between the Hamiltonian expressed in terms of reagent and product mass-scaled coordinates ((A3.11.81) and (A3.11.84)). In fact, the reagent to product transformation can be thought of as a rotation by an angle β followed by a flip in the sign of s' . The angle β is sometimes called the 'skew' angle, and it can vary between 0 and 90°, as determined by equation (A3.11.86). If $m_A = m_B = m_C$ (i.e., all three masses are identical, as in the reaction $\text{H} + \text{H}_2$), then $\beta = 60^\circ$, while for $m_B \gg m_A, m_C$, $\beta \rightarrow 90^\circ$ and $m_B \ll m_A m_C$ gives $\beta \rightarrow 0$.

Although the Schrödinger equation associated with the $\text{A} + \text{BC}$ reactive collision has the same form as for the nonreactive scattering problem that we considered previously, it *cannot* be solved by the coupled-channel expansion used then, as the reagent vibrational basis functions cannot directly describe the product region (for an expansion in a finite number of terms). So instead we need to use alternative schemes of which there are many.

One possibility is to use *hyperspherical coordinates*, as these enable the use of basis functions which describe reagent and product internal states in the same expansion. Hyperspherical coordinates have been extensively discussed in the literature [34, 35 and 36] and in the present application they reduce to polar coordinates (ρ, η) defined as follows:

$$\rho = \sqrt{R'^2 + r'^2} = \sqrt{S'^2 + s'^2} \quad 0 \leq \rho \leq \infty \quad (\text{A3.11.87})$$

$$\eta = \tan^{-1} \frac{r'}{R'} \quad 0 \leq \eta \leq \beta. \quad (\text{A3.11.88})$$

-22-

Hyperspherical coordinates have the properties that η motion is always bound since $\eta = 0$ and $\eta = \beta$ correspond to cases where two of the three atoms are on top of one another, yielding a very repulsive potential. Also, $\rho \rightarrow 0$ is a repulsive part of the potential, while large ρ takes us to the reagent and product valleys.

To develop coupled-channel methods to solve the Schrödinger equation, we first transform the Hamiltonian (A3.11.81) to hyperspherical coordinates, yielding:

$$\hat{H} = \frac{-\hbar^2}{2m} \left(\frac{\partial^2}{\partial \rho^2} + \frac{1}{\rho} \frac{\partial}{\partial \rho} + \frac{1}{\rho^2} \frac{\partial^2}{\partial \eta^2} \right) + V. \quad (\text{A3.11.89})$$

Now define a new wavefunction $\chi = \rho^{+1/2} \psi$. Then

$$\hat{H} \psi = \frac{-\hbar^2}{2m} \left(\frac{\partial^2}{\partial \rho^2} + \frac{1}{\rho} \frac{\partial}{\partial \rho} + \frac{1}{\rho^2} \frac{\partial^2}{\partial \eta^2} \right) \rho^{-1/2} \chi + V \rho^{-1/2} \chi. \quad (\text{A3.11.90})$$

After cancelling out a factor $\rho^{-1/2}$ and regrouping, we obtain a new version of the Schrödinger equation in which the first derivative term has been eliminated.

$$\left\{ \frac{-\hbar^2}{2m} \left[\frac{\partial^2}{\partial \rho^2} + \frac{1}{4\rho^2} + \frac{1}{\rho^2} \frac{\partial^2}{\partial \eta^2} \right] + V \right\} \chi = E \chi. \quad (\text{A3.11.91})$$

Now select out the η -dependent part to define vibrational functions at some specific ρ which we call $\bar{\rho}$.

$$-\frac{\hbar^2}{2m\bar{\rho}^2} \frac{d^2}{d\eta^2} \varphi_n + V(\bar{\rho}, \eta) \varphi_n = \varepsilon_n \varphi_n \quad (\text{A3.11.92})$$

with the boundary condition that $\varphi_n \rightarrow 0$ as $\eta \rightarrow 0$ and $\eta \rightarrow \beta$. For large $\bar{\rho}$ the φ_n will become eigenfunctions of the reagent and product diatomics.

To set up coupled-channel equations, use the expansion

$$\psi = \rho^{-1/2} \sum_n \varphi_n(\eta) g_n(\rho). \quad (\text{A3.11.93})$$

This leads to

$$\frac{-\hbar^2}{2m} \frac{d^2}{d\rho^2} \sum_n \varphi_n g_n + \frac{-\hbar^2}{2m\rho^2} \sum_n g_n \frac{d^2 \varphi_n}{d\eta^2} + \left(V - \frac{\hbar^2}{8m\rho^2} \right) \sum_n \varphi_n g_n = E \sum_n \varphi_n g_n. \quad (\text{A3.11.94})$$

-23-

Now substitute the $\rho = \bar{\rho}$ solution for $d^2 \varphi_n / d\eta^2$ from above, multiply by φ_n and integrate over η . This gives

$$\frac{d^2 \mathbf{g}}{d\rho^2} = \mathbf{U} \mathbf{g} \quad (\text{A3.11.95})$$

where

$$U_{nn'} = \frac{2m}{\hbar^2} \left\langle \varphi_n \left| V(\rho, \eta) - \frac{\bar{\rho}^2}{\rho^2} V(\bar{\rho}, \eta) \right| \varphi_{n'} \right\rangle + \frac{2m}{\hbar^2} \delta_{nn'} \left(\varepsilon_n \frac{\bar{\rho}^2}{\rho^2} - \frac{\hbar^2}{8m\rho^2} - E \right). \quad (\text{A3.11.96})$$

This equation may be solved by the same methods as used with the nonreactive coupled-channel equations (discussed later in [section A3.11.4.2](#)). However, because $V(\rho, \eta)$ changes rapidly with ρ , it is desirable to periodically change the expansion basis set φ_n . To do this we divide the range of ρ to be integrated into 'sectors' and within each sector choose a $\bar{\rho}$ (usually the midpoint) to define local eigenfunctions. The coupled-channel equations just given then apply within each sector, but at sector boundaries we change basis sets. Let $\bar{\rho}_1$ and $\bar{\rho}_2$ be the $\bar{\rho}$ associated with adjacent sectors. Then, at the sector boundary ρ_b we require

$$\Psi_1(\eta, \rho_b) = \Psi_2(\eta, \rho_b) \quad (\text{A3.11.97})$$

or

$$\sum_n \varphi_n(\eta, \bar{\rho}_1) g_{nn_i}^{(1)}(\rho_b) = \sum_n \varphi_n(\eta, \bar{\rho}_2) g_{nn_i}^{(2)}(\rho_b). \quad (\text{A3.11.98})$$

Multiply by $\varphi_n(\varepsilon, \bar{\rho}_2)$ and integrate to obtain

$$\mathbf{g}^{(2)} = \mathbf{S}^{21} \mathbf{g}^{(1)} \quad (\text{A3.11.99})$$

where

$$S_{nn'}^{21} = \langle \varphi_n(\eta, \bar{\rho}_2) | \varphi_{n'}(\eta, \bar{\rho}_1) \rangle. \quad (\text{A3.11.100})$$

The corresponding derivative transformation is:

$$\frac{d\mathbf{g}^{(2)}}{d\rho} = \mathbf{S}^{21} \frac{d\mathbf{g}^{(1)}}{d\rho}. \quad (\text{A3.11.101})$$

-24-

This scheme makes it possible to propagate \mathbf{g} from small ρ where \mathbf{g} should vanish to large ρ where an asymptotic analysis can be performed.

To perform the asymptotic analysis we need to first write down the proper asymptotic solution. Clearly we want some solutions with incoming waves in the reagents, then outgoing waves in both reagents and products and other solutions with the reagent and product labels interchanged. One way to do this is to define a matrix of incoming waves \mathbf{I} and outgoing waves \mathbf{O} such that

$$I_{\alpha v' \alpha' v} = \delta_{\alpha \alpha'} \delta_{v v'} \begin{cases} e^{-ik_v R'} & \alpha = 1 \\ e^{-ik_v S'} & \alpha = 2 \end{cases} \quad (\text{A3.11.102a})$$

$$O_{\alpha v \alpha' v'} = \delta_{\alpha \alpha'} \delta_{v v'} \begin{cases} e^{ik_v R'} & \alpha = 1 \\ e^{ik_v S'} & \alpha = 2 \end{cases} \quad (\text{A3.11.102b})$$

where α is an arrangement channel label such that $\alpha = 1$ and 2 correspond to the ‘reagents’ and ‘products’. Also let $\varphi_{\alpha v}$ be the reagent or product vibrational function. Then the asymptotic solution is

$$\Psi_{R' S' \rightarrow \infty} \sim \sum_{\alpha v} \varphi_{\alpha v}(v_\alpha)^{-1/2} \left(I_{\alpha v \alpha' v'} - \sum_{\alpha'' v''} S_{\alpha v \alpha'' v''} O_{\alpha'' v'' \alpha v} \right). \quad (\text{A3.11.103})$$

We have expressed Ψ in terms of Jacobi coordinates as this is the coordinate system in which the vibrations and translations are separable. The separation does not occur in hyperspherical coordinates except at $\rho = \infty$, so it is necessary to interrelate coordinate systems to complete the calculations. There are several approaches for doing this. One way is to project the hyperspherical solution onto Jacobi’s before performing the asymptotic analysis, i.e.

$$(\text{A3.11.104})$$

$$\rho^{-1/2} \sum_v \varphi_v(\eta) g_{vv'}(\rho) = \sum_v \varphi_v(r) G_{vv'}(R).$$

The \mathbf{G} matrix is then obtained by performing the quadrature

$$G_{vv'} = \int dr \varphi_v(r) \sum_{v''} \rho^{-1/2} \varphi_{v''}(\eta) g_{v''v'}(\rho) \quad (\text{A3.11.105})$$

where $\rho(R, r)$, $\eta(R, r)$ are to be substituted as needed into the right hand side.

A3.11.3.3 SCATTERING IN THREE DIMENSIONS

All the theory developed up to this point has been limited in the sense that translational motion (the continuum degree of freedom) has been restricted to one dimension. In this section we discuss the generalization of this to three dimensions for collision processes where space is isotropic (i.e., collisions in homogeneous phases, such as in a

-25-

vacuum, but not collisions with surfaces). We begin by considering collisions involving a single particle in three dimensions; the multichannel case is considered subsequently.

The biggest change associated with going from one to three dimensional translational motion refers to asymptotic boundary conditions. In three dimensions, the initial scattering wavefunction for a single particle is represented by a plane wave $e^{i\mathbf{k}\cdot\mathbf{r}}$ moving in a direction which we denote with the wavevector \mathbf{k} . Scattering then produces outgoing spherical waves as $t \rightarrow \infty$ weighted by an amplitude $f_k(\theta)$ which specifies the scattered intensity as a function of the angle θ between \mathbf{k} and the observation direction. Mathematically the time independent boundary condition analogous to [equation \(A3.11.13a\)](#), [equation \(A3.11.13b\)](#) is:

$$\psi_{\mathbf{k}}(\mathbf{r}) \underset{r \rightarrow \infty}{=} e^{i\mathbf{k}\cdot\mathbf{r}} + f_k(\theta) \frac{e^{ikr}}{r}. \quad (\text{A3.11.106})$$

Note that for potentials that depend only on the scalar distance r between the colliding particles, the amplitude $f_k(\theta)$ does not depend on the azimuthal angle associated with the direction of observation.

The measurable quantity in a three dimensional scattering experiment is the differential cross section $d\sigma_k(\theta)/d\Omega$. This is defined as

$$\frac{d\sigma_k(\theta)}{d\Omega} = \frac{|\text{outgoing radial flux}|}{|\text{total incident flux}|} \quad (\text{A3.11.107})$$

where outgoing flux refers to the radial velocity operator $\hat{v}_r = -i\hbar\partial/\partial r$. Substitution of [equation \(A3.11.106\)](#) into [\(A3.11.107\)](#) using [\(A3.11.16\)](#) yields

$$d\sigma_k(\theta)/d\Omega = |f_k(\theta)|^2. \quad (\text{A3.11.108})$$

It is convenient to expand $f_k(\theta)$ in a basis of Legendre polynomials $P_\ell(\cos \theta)$ (as these define the natural

angular eigenfunctions associated with motion in three dimensions). Here we write:

$$I(\theta) = \sum_n \frac{b_n}{\sin \theta} \left(\frac{db}{d\theta} \right)_n. \quad (\text{A3.11.109})$$

We call this a *partial wave expansion*. To determine the coefficients a_ℓ^k , one matches asymptotic solutions to the radial Schrödinger equation with the corresponding partial wave expansion of equation (A3.11.106). It is customary to write the asymptotic radial Schrödinger equation solution as

$$\psi_{\ell m}(r, \theta, \varphi) \underset{r \rightarrow \infty}{=} \frac{1}{r} Y_{\ell m}(\theta, \varphi) (e^{-i(kr - \ell\pi/2)} - S_\ell e^{i(kr - \ell\pi/2)}) \quad (\text{A3.11.110})$$

-26-

where S_ℓ is the scattering matrix for the ℓ th partial wave and m is the projection quantum number associated with ℓ . Unitarity of the scattering matrix implies that S_ℓ can be written as $\exp(2i\delta_\ell)$ where δ_ℓ is a real quantity known as the *phase shift*.

The asymptotic partial wave expansion of [equation \(A3.11.106\)](#) can be developed using the identity

$$e^{ik \cdot r} = e^{ikr \cos \theta} = \sum_{\ell=0}^{\infty} i^\ell (2\ell + 1) j_\ell(kr) P_\ell(\cos \theta) \quad (\text{A3.11.111})$$

where $j_\ell(kr)$ is a spherical Bessel function. At large r , the spherical Bessel function reduces to

$$j_\ell(kr) \underset{r \rightarrow \infty}{=} \frac{\sin(kr - \ell\pi/2)}{kr}. \quad (\text{A3.11.112})$$

If equation (A3.11.112) is then used to evaluate (A3.11.111) after substitution of the latter into [\(A3.11.106\)](#) and if [equation \(A3.11.109\)](#) is also substituted into [\(A3.11.106\)](#) and the result for each ℓ and m is equated to [\(A3.11.110\)](#), one finds that only $m = 0$ contributes, and that

$$a_\ell^k = \frac{(2\ell + 1)}{2ik} (S_\ell - 1). \quad (\text{A3.11.113})$$

From [equation \(A3.11.108\)](#), [equation \(A3.11.109\)](#) and [equation \(A3.11.113\)](#) one then finds

$$\frac{d\sigma_k(\theta)}{d\Omega} = \frac{1}{4k^2} \left| \sum_{\ell} (2\ell + 1) P_\ell(\cos \theta) (S_\ell - 1) \right|^2 \quad (\text{A3.11.114a})$$

$$= \frac{1}{k^2} \left| \sum_{\ell} (2\ell + 1) P_\ell(\cos \theta) e^{i\delta_\ell} \sin \delta_\ell \right|^2. \quad (\text{A3.11.114b})$$

This differential cross section may be integrated over scattering angles to define an integral cross section σ as follows:

$$\sigma = 2\pi \int_0^\pi \frac{d\sigma_\ell(\theta)}{d\Omega} \sin\theta d\theta = \frac{\pi}{k^2} \sum_\ell (2\ell + 1) |S_\ell - 1|^2 \quad (\text{A3.11.115a})$$

$$= \frac{4\pi}{k^2} \sum_\ell (2\ell + 1) \sin^2 \delta_\ell. \quad (\text{A3.11.115b})$$

-27-

Equations [A3.11.114\(b\)](#) and [A3.11.115\(b\)](#) are in a form that is convenient to use for potential scattering problems. One needs only to determine the phase shift δ_ℓ for each ℓ , then substitute into these equations to determine the cross sections. Note that in the limit of large ℓ , δ_ℓ must vanish so that the infinite sum over partial waves ℓ will converge. For most potentials of interest to chemical physics, the calculation of δ_ℓ must be done numerically.

Equation [A3.11.115\(a\)](#) is also useful as a form that enables easy generalization of the potential scattering theory that we have just derived to multistate problems. In particular, if we imagine that we are interested in the collision of two molecules A and B starting out in states n_A and n_B and ending up in states n'_A and n'_B , then the asymptotic wavefunction analogous to [equation \(A3.11.106\)](#) is

$$\psi_{n_A n_B \rightarrow n'_A n'_B} = \exp(i\mathbf{k}_{n_A n_B} \cdot \mathbf{r}) |n_A n_B\rangle + r^{-1} \sum_{n'_A n'_B} f_{n_A n_B \rightarrow n'_A n'_B}(\theta) \exp(i\mathbf{k}_{n'_A n'_B} \cdot \mathbf{r}) |n'_A n'_B\rangle \quad (\text{A3.11.116})$$

where the scattering amplitude f is now labelled by the initial and final state indices. Integral cross sections are then obtained using the following generalization of [equation A3.11.115\(a\)](#):

$$\sigma_{n_A n_B \rightarrow n'_A n'_B} = \frac{\pi}{k_{n_A n_B}^2} \sum_J (2J + 1) |S_{n_A n_B \rightarrow n'_A n'_B}^J - \delta_{n_A n_B, n'_A n'_B}|^2 \quad (\text{A3.11.117})$$

where S is the multichannel scattering matrix, δ is the Kronecker delta function and J is the total angular momentum (i.e., the vector sum of the orbital angular momentum ℓ plus the angular momenta of the molecules A and B). Here the sum is over J rather than ℓ , because ℓ is not a conserved quantity due to coupling with angular momenta in the molecules A and B.

A3.11.4 COMPUTATIONAL METHODS AND STRATEGIES FOR SCATTERING PROBLEMS

In this section we present several numerical techniques that are commonly used to solve the Schrödinger equation for scattering processes. Because the potential energy functions used in many chemical physics problems are complicated (but known to reasonable precision), new numerical methods have played an important role in extending the domain of application of scattering theory. Indeed, although much of the formal development of the previous sections was known 30 years ago, the numerical methods (and computers) needed to put this formalism to work have only been developed since then.

This section is divided into two sections: the first concerned with time-dependent methods for describing the evolution of wavepackets and the second concerned with time-independent methods for solving the time independent Schrödinger equation. The methods described are designed to be representative of what is in use,

but not exhaustive. More detailed discussions of time-dependent and time-independent methods are given in the literature [37, 38].

-28-

A3.11.4.1 TIME-DEPENDENT WAVEPACKET METHODS

(A) OVERALL STRATEGY

The methods described here are all designed to determine the time evolution of wavepackets that have been previously defined. This is only one of several steps for using wavepackets to solve scattering problems. The overall procedure involves the following steps:

- (a) First, choose an initial wavepacket $\psi(x,t)$ that describes the range of energies and initial conditions that we want to simulate and which is numerically as well behaved as possible. Typically, this is chosen to be a Gaussian function of the translational coordinate x , with mean velocity and width chosen to describe the range of interest. In making this choice, one needs to consider how the spatial part of the Schrödinger equation is to be handled, i.e., whether the dependence of the wavepacket on spatial coordinates is to be represented on a grid, or in terms of basis functions.
- (b) Second, one propagates this wavepacket in time using one of the methods described below, for a sufficient length of time to describe the scattering process of interest.
- (c) Third, one calculates the scattering information of interest, such as the outgoing flux.

Typically, the ratio of this to the incident flux determines the transition probability. This information will be averaged over the energy range of the initial wavepacket, unless one wants to project out specific energies from the solution. This projection procedure is accomplished using the following expression for the energy resolved (time-independent) wavefunction in terms of its time-dependent counterpart:

$$\psi^+(E) = \frac{1}{a(E)} \int_{-\infty}^{\infty} e^{iEt} \psi(t) dt \quad (\text{A3.11.118})$$

where

$$a(E) = \frac{1}{v^{1/2}} \langle e^{-ikR} | \psi(0) \rangle. \quad (\text{A3.11.119})$$

(B) SECOND ORDER DIFFERENCING

A very simple procedure for time evolving the wavepacket is the second order differencing method. Here we illustrate how this method is used in conjunction with a fast Fourier transform method for evaluating the spatial coordinate derivatives in the Hamiltonian.

If we write the time-dependent Schrödinger equation as $\partial\psi/\partial t = -(i/\hbar)\hat{H}\psi$, then, after replacing the time derivative by a central difference, we obtain

$$\frac{\partial\psi}{\partial t} = \frac{\psi(t + \Delta t) - \psi(t - \Delta t)}{2\Delta t}. \quad (\text{A3.11.120})$$

After rearranging this becomes

$$\psi(t + \Delta t) = \psi(t - \Delta t) - \frac{2i\Delta t}{\hbar} \hat{H} \psi(t). \quad (\text{A3.11.121})$$

To invoke this algorithm, we need to evaluate $\hat{H}\psi = (\hat{T} + \hat{V})\psi$. If ψ is represented on a uniform grid in coordinate x , an effective scheme is to use fast Fourier transforms (FFTs) to evaluate $\hat{T}\psi$. Thus, in one dimension we have, with n points on the grid,

$$\tilde{\psi}(k_j, t) = \sum_{m=1}^n e^{ik_j x_m} \psi(x_m, t) \quad (\text{A3.11.122})$$

where the corresponding momentum grid is

$$k_j = \frac{2\pi j}{n\Delta x}. \quad (\text{A3.11.123})$$

Differentiation of (A3.11.122) then gives

$$\hat{T} \tilde{\psi}(k_j, t) = \frac{p_j^2}{2m} \tilde{\psi}(k_j, t) \quad (\text{A3.11.124a})$$

where $p_j = \hbar k_j$. This expression can be inverted to give

$$\hat{T} \psi(x_\ell) = \frac{1}{n} \sum_{j=-n/2}^{n/2-1} e^{-ik_j x_\ell} \frac{p_j^2}{2m} \tilde{\psi}(k_j, t). \quad (\text{A3.11.124b})$$

This expression, in combination with (A3.11.122), determines the action of the kinetic energy operator on the wavefunction at each grid point. The action of \hat{V} is just $V(x_\ell)\psi(x_\ell)$ at each grid point.

(C) SPLIT-OPERATOR OR FEIT-FLECK METHOD

A more powerful method for evaluating the time derivative of the wavefunction is the split-operator method [39]. Here we start by formally solving $i\hbar\partial\psi/\partial t = \hat{H}\psi$ with the solution $\psi(t) = e^{-i\hat{H}t/\hbar}\psi(0)$. Note that H is assumed to be time-independent. Now imagine evaluating the propagator $e^{-i\hat{H}t/\hbar}$ over a short time interval.

$$e^{-i\hat{H}\Delta t/\hbar} = e^{-i(\hat{T}+\hat{V})\Delta t/\hbar} \approx e^{-i\hat{V}\Delta t/2\hbar} e^{-i\hat{T}\Delta t/\hbar} e^{-i\hat{V}\Delta t/2\hbar}. \quad (\text{A3.11.125})$$

Evidently, this formula is not exact if \hat{T} and \hat{V} do not commute. However for short times it is a good approximation, as can be verified by comparing terms in Taylor series expansions of the middle and right-hand expressions in (A3.11.125). This approximation is intrinsically unitary, which means that scattering information obtained from this calculation automatically conserves flux.

The complete propagator is then constructed by piecing together N time steps, leading to

$$e^{-i\hat{H}\Delta t/\hbar} = e^{-i\hat{V}\Delta t/2\hbar} e^{-i\hat{T}\Delta t/\hbar} e^{-i\hat{V}\Delta t/\hbar} \dots e^{-i\hat{V}\Delta t/\hbar} e^{-i\hat{T}\Delta t/\hbar} e^{-i\hat{V}\Delta t/2\hbar}. \quad (\text{A3.11.126})$$

To evaluate each term we can again do it on a grid, using FFTs as described above to evaluate $e^{-i\hat{T}\Delta t/\hbar}$.

(D) CHEBYSHEV METHOD

Another approach [40] is to expand $e^{-i\hat{H}\Delta t/\hbar}$ in terms of Chebyshev polynomials and to evaluate each term in the polynomial at the end of the time interval. Here a Chebyshev expansion is chosen as it gives the most uniform convergence in representing the exponential over the chosen time interval. The time interval Δt is typically chosen to be several hundred of the time steps that would be used in the second order differencing or split-operator methods. Although the Chebyshev method is not intrinsically unitary, it is capable of much higher accuracy than the second order differencing or split-operator methods [41].

In order to apply this method it is necessary to scale \hat{H} to lie in a certain finite interval which is usually chosen to be $(-1, 1)$. Thus, if V_{\max} and V_{\min} are estimates of the maximum and minimum potentials and T_{\max} is the maximum kinetic energy, we use

$$\hat{H}_{\text{norm}} = [\hat{H} - (R + V_{\min})]/R = \frac{\hat{H} - V_{\min}}{R} - 1 \quad (\text{A3.11.127})$$

where

$$R = (T_{\max} + V_{\max} - V_{\min})/2. \quad (\text{A3.11.128})$$

This choice restricts the range of values of \hat{H}_{norm} to the interval $(0, 1)$. Then the propagator becomes

$$e^{-i\hat{H}\Delta t/\hbar} = e^{-i\hat{H}_{\text{norm}}R\Delta t/\hbar} e^{-i(R+V_{\min})\Delta t/\hbar}. \quad (\text{A3.11.129})$$

Now we replace the first exponential in the right-hand side of (A3.11.129) by a Chebyshev expansion as follows:

$$e^{-i\hat{H}\Delta t/\hbar} = e^{-i(R+V_{\min})\Delta t/\hbar} \sum_{k=0}^N C_k J_k(R\Delta t/\hbar) i^k T_k(-\hat{H}_{\text{norm}}) \quad (\text{A3.11.130})$$

where T_k is a Chebyshev polynomial, and J_k is a Bessel function. The coefficients C_k are fixed at $C_0 = 1$, $C_k = 2$, $k < 0$.

To apply this method, the J_k are calculated once and stored while the T_k are generated using the recursion formula:

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x) \quad (\text{A3.11.131})$$

with $T_0 = 1$, $T_1 = x$. Actually the T_k are never explicitly stored, as all we really want is T_k operating onto a wavefunction. However, the recursion formula is still used to generate this, so the primary computational step involves \hat{H}_{norm} operating onto wavefunctions. This can be done using FFTs as discussed previously.

(E) SHORT ITERATIVE LANCZOS METHOD

Another approach involves starting with an initial wavefunction ψ_0 , represented on a grid, then generating $\hat{H}\psi_0$, and consider that this, after orthogonalization to ψ_0 , defines a new state vector. Successive applications \hat{H} can now be used to define an orthogonal set of vectors which defines as a *Krylov space* via the iteration: ($n = 0, \dots, N$)

$$\beta_{n+1}|\psi_{n+1}\rangle = (\hat{H} - \alpha_n)|\psi_n\rangle - \beta_n|\psi_{n-1}\rangle \quad (\text{A3.11.132})$$

where

$$\alpha_n = \langle \psi_n | \hat{H} | \psi_n \rangle \quad \beta_{n+1} = \langle \psi_n | \hat{H} | \psi_{n+1} \rangle. \quad (\text{A3.11.133})$$

The Hamiltonian in this vector space is

$$\mathbf{H} = \begin{pmatrix} \alpha_0 & \beta_1 & 0 & 0 & 0 \\ \beta_1 & \alpha_1 & \beta_2 & 0 & 0 \\ 0 & \beta_2 & \alpha_2 & \beta_3 & 0 \\ 0 & 0 & \beta_2 & \alpha_3 & \beta_4 \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}. \quad (\text{A3.11.134})$$

Here \mathbf{H} forms an $N \times N$ matrix, where N is the dimensionality of the space and is generally much smaller than the number of grid points.

Now diagonalize \mathbf{H} , calling the eigenvalues λ_k and eigenvectors $T_{\ell k}$. Numerically, this is a very efficient process due to the tridiagonal form of (A3.11.134). The resulting eigenvalues and eigenvectors are then used to propagate for a short time Δt via

$$[\psi(t + \Delta t)]_j = [\psi_0^+(t)\psi_0(t)]_j^{1/2} \sum_{\ell=0}^N \sum_{k=0}^{\ell} \frac{(\psi(t)_\ell)_j}{[\psi(t)_\ell^+ \psi(t)_\ell]_j^{1/2}} T_{\ell k} e^{-i\lambda_k \Delta t/\hbar} T_{0k} \quad (\text{A3.11.135})$$

where $(\psi_\ell)_j$ are coefficients that transfer between the j th grid point and the ℓ th order Krylov space in (A3.11.132).

A3.11.4.2 TIME-INDEPENDENT METHODS

Here we discuss several methods that are commonly used to propagate coupled-channel equations, and we also present a linear algebra method for applying variational theory. The coupled-channel equations are coupled ordinary differential equations, so they can in principle be solved using any one of a number of standard methods for doing this (Runge–Kutta, predictor–corrector etc). However these methods are very inefficient for this application and a number of alternatives have been developed which take advantage of specific features of the problems being solved.

(A) GORDON-TYPE METHODS

In many kinds of atomic and molecular collision problem the wavefunction has many oscillations because the energy is high (i.e., $g(R) \approx e^{ikR}$ and k is large). In this case it is useful to expand $g(R)$ in terms of oscillatory solutions to some reference problem that is similar to the desired one and then regard the expansion coefficients as the quantity being integrated, thereby removing most or all of the oscillations from the time dependence of the coefficients.

For example, suppose that we divide coordinate space into steps ΔR , then evaluate \mathbf{U} (in equation (A3.11.61)) at the middle of each step and regard this as the reference for propagation within this step. Further, let us diagonalize \mathbf{U} , calling the eigenvalues u_k and eigenvectors \mathbf{T} . Then, as long as the variation in eigenvalues and eigenvectors can be neglected in each step, the Schrödinger equation solution within each step is easily expressed in terms of sin and cos $w_k \Delta R$, where $w_k = \sqrt{-u_k}$ (or exponential solutions if $u_k < 0$). In particular, if $\mathbf{g}(R_0)$ is the solution at the beginning of each step, then $\mathbf{T}\mathbf{g}$ transforms into the diagonalized representation and $\mathbf{T}\mathbf{g}'$ is the corresponding derivative. The complete solution at the end of each step would then be

$$\mathbf{g}(R_1) = \mathbf{T}^{-1}(\mathbf{w}^{-1} \sin(\mathbf{w}\Delta R)\mathbf{T}\mathbf{g}'(R_0) + \cos(\mathbf{w}\Delta R)\mathbf{T}\mathbf{g}(R_0)) \quad (\text{A3.11.136})$$

$$\mathbf{g}'(R_1) = \mathbf{T}^{-1}(\cos(\mathbf{w}\Delta R)\mathbf{T}\mathbf{g}'(R_0) - \mathbf{w} \sin(\mathbf{w}\Delta R)\mathbf{T}\mathbf{g}(R_0)). \quad (\text{A3.11.137})$$

In principle, one can do better by allowing for R -dependence to \mathbf{U} and \mathbf{T} . If we allow them to vary linearly with R , then we have Gordon's method [42]. However, the higher order evaluation in this case leads to a much more cumbersome theory that is often less efficient even though larger steps can be used.

One problem with using this method (or any method that propagates ψ) is that in regions where $u_k > 0$, the so-called 'closed' channels, the solutions increase exponentially. If such solutions exist for some channels while others are still open, the closed-channel solutions can become numerically dominant (i.e., so much bigger that they overwhelm the open-channel solutions to within machine precision and, after a while, all channels propagate as if they are closed).

To circumvent this, it is necessary to 'stabilize' the solutions periodically. Typically this is done by multiplying $\mathbf{g}(R)$ and $\mathbf{g}'(R)$ by some matrix \mathbf{h} that 'orthogonalizes' the solutions as best one can. For example, this can be done using $\mathbf{h} = \mathbf{g}^{-1}(R_s)$ where R_s is the value of R at the end of the 'current' step. Thus, after stabilization, the new \mathbf{g} and \mathbf{g}' are:

$$\mathbf{g}_{\text{new}}(R_s) = \mathbf{g}_{\text{old}}(R_s)\mathbf{g}_{\text{old}}^{-1}(R_4) = \mathbf{I} \quad (\text{A3.11.138})$$

$$\mathbf{g}'_{\text{new}}(R_s) = \mathbf{g}'_{\text{old}}(R_s)\mathbf{g}_{\text{old}}^{-1}(R_4). \quad (\text{A3.11.139})$$

One consequence of performing the stabilization procedure is that the initial conditions that correspond to the current $\mathbf{g}(R)$ are changed each time stabilization is performed. However this does not matter as long the initial $\mathbf{g}(R)$ value corresponds to the limit $R \rightarrow 0$ as then all one needs is for $\mathbf{g}(R)$ to be small (i.e., the actual value is not important).

(B) LOG DERIVATIVE PROPAGATION

One way to avoid the stabilization problem just mentioned is to propagate the log derivative matrix $\mathbf{Y}(R)$ [43]. This is defined by

$$\mathbf{Y}(R) = \mathbf{g}'(R)\mathbf{g}^{-1}(R) \quad (\text{A3.11.140})$$

and it remains well behaved numerically even when $\mathbf{g}(R)$ grows exponentially. The differential equation obeyed by \mathbf{Y} is

$$\mathbf{Y}'(R) = \mathbf{g}''(R)\mathbf{g}^{-1}(R) - \mathbf{g}'(R)\mathbf{g}^{-1}(R)\mathbf{g}'(R)\mathbf{g}^{-1}(R) = \mathbf{U}\mathbf{g}(R)\mathbf{g}^{-1}(R) - \mathbf{Y}^2(R) = \mathbf{U} - \mathbf{Y}^2(R). \quad (\text{A3.11.141})$$

It turns out that one cannot propagate \mathbf{Y} using standard numerical methods because $|\mathbf{Y}|$ blows up whenever $|\mathbf{g}|$ is zero. To circumvent this one must propagate \mathbf{Y} by ‘invariant imbedding’. The basic idea here is to construct a propagator \mathbf{Y} which satisfies

$$\begin{pmatrix} \mathbf{g}'(R') \\ \mathbf{g}'(R'') \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_1(R', R'') & \mathbf{Y}_2(R', R'') \\ \mathbf{Y}_3(R', R'') & \mathbf{Y}_4(R', R'') \end{pmatrix} \begin{pmatrix} -\mathbf{g}(R') \\ \mathbf{g}(R'') \end{pmatrix} \quad (\text{A3.11.142})$$

where R' and R'' might form the beginning and end of a propagation step. Assuming for the moment that we know what the \mathbf{Y}_i are, then the evolution of \mathbf{Y} is as follows

$$\mathbf{Y}(R'') = \mathbf{g}'(R'')\mathbf{g}^{-1}(R'') = -\mathbf{Y}_3\mathbf{g}(R')\mathbf{g}^{-1}(R'') + \mathbf{Y}_4\mathbf{g}(R'')\mathbf{g}^{-1}(R'') \quad (\text{A3.11.143})$$

$$\mathbf{Y}(R') = \mathbf{g}'(R')\mathbf{g}^{-1}(R') = -\mathbf{Y}_1\mathbf{g}(R')\mathbf{g}^{-1}(R') + \mathbf{Y}_2\mathbf{g}(R'')\mathbf{g}^{-1}(R'). \quad (\text{A3.11.144})$$

So the overall result is

$$\mathbf{Y}(R'') = \mathbf{Y}_4 - \mathbf{Y}_3[\mathbf{Y}(R') + \mathbf{Y}_1]^{-1}\mathbf{Y}_2. \quad (\text{A3.11.145})$$

To solve for the \mathbf{Y} , we begin by solving a reference problem wherein the coupling matrix is assumed diagonal with constant couplings within each step. (These could be accomplished by diagonalizing \mathbf{U} , but it would be better to avoid this work and use the diagonal \mathbf{U} matrix elements.) Then, in terms of the reference \mathbf{U} (which we call \mathbf{U}_d), we have

$$\mathbf{g}(R_1) = \mathbf{U}_d^{-1} \sin(\mathbf{U}_d \Delta R) \mathbf{g}'(R_0) + \cos(\mathbf{U}_d \Delta R) \mathbf{g}(R_0) \quad (\text{A3.11.146})$$

$$\mathbf{g}'(R_1) = \cos \mathbf{U}_d \Delta R \mathbf{g}'(R_0) - \mathbf{U}_d \sin \mathbf{U}_d \Delta R \mathbf{g}(R_0) \quad (\text{A3.11.147})$$

Now rearrange these to:

$$\mathbf{g}'(R_0) = \mathbf{U}_d (\sin^{-1} \mathbf{U}_d \Delta R)^{-1} \mathbf{g}(R_1) - \mathbf{U}_d \cot \mathbf{U}_d \Delta R \mathbf{g}(R_0) \quad (\text{A3.11.148})$$

$$\mathbf{g}'(R_1) = \mathbf{U}_d \cot \mathbf{U}_d \Delta R \mathbf{g}(R_1) - \mathbf{U}_d (\sin \mathbf{U}_d \Delta R)^{-1} \mathbf{g}(R_0) \quad (\text{A3.11.149})$$

which can be written:

$$\begin{pmatrix} \mathbf{g}'(R_0) \\ \mathbf{g}'(R_1) \end{pmatrix} = \begin{pmatrix} \mathbf{U}_d \cot \mathbf{U}_d \Delta R & \mathbf{U}_d (\sin \mathbf{U}_d \Delta R)^{-1} \\ \mathbf{U}_d (\sin \mathbf{U}_d \Delta R)^{-1} & \mathbf{U}_d \cot \mathbf{U}_d \Delta R \end{pmatrix} \begin{pmatrix} -\mathbf{g}(R_0) \\ \mathbf{g}(R_1) \end{pmatrix}. \quad (\text{A3.11.150})$$

Note that $|\mathbf{U}_d \Delta R| < 0$ is required for meaningful results and thus ΔR cannot be too large. By comparing equation (A3.11.142) and equation (A3.11.150), we find:

$$\mathbf{Y}_1 = \mathbf{Y}_4 = \mathbf{U}_d \cot \mathbf{U}_d \Delta R \quad (\text{A3.11.151})$$

$$\mathbf{Y}_2 = \mathbf{Y}_3 = \mathbf{U}_d (\sin \mathbf{U}_d \Delta R)^{-1}. \quad (\text{A3.11.152})$$

The standard log-derivative propagator now corrects for the difference between \mathbf{U} and \mathbf{U}_d using a Simpson-rule integration. The specific formulas are

$$\mathbf{y}_1 \rightarrow \mathbf{y}_1 + \mathbf{Q}(R_0) \quad (\text{A3.11.153})$$

$$\mathbf{y}_2 \rightarrow \mathbf{y}_2 \quad (\text{A3.11.154})$$

$$\mathbf{y}_3 \rightarrow \mathbf{y}_3 \quad (\text{A3.11.155})$$

$$\mathbf{y}_4 \rightarrow \mathbf{y}_4 + \mathbf{Q}(R_1). \quad (\text{A3.11.156})$$

Then for a step divided into two halfsteps, at $R = a, c, b$, write $c = 1/2(a+b)$, $\Delta R = (b-a)/2$, $\Delta \mathbf{u} = \mathbf{U} - \mathbf{U}_d$, $a = R_0$, $b = R_1$. This leads to the following expression for \mathbf{Q} :

$$\mathbf{Q}(a) = \frac{\Delta R}{e} \Delta \mathbf{u}(a) \quad (\text{A3.11.157})$$

$$\mathbf{Q}(c) = \frac{1}{2} \left(1 - \frac{\Delta R^2}{6} \Delta \mathbf{u}(c) \right)^{-1} \frac{4\Delta R}{3} \Delta \mathbf{u}(c) \quad (\text{A3.11.158})$$

$$\mathbf{Q}(b) = \frac{\Delta R}{3} \Delta \mathbf{u}(b). \quad (\text{A3.11.159})$$

Propagation then proceeds from $R \rightarrow 0$ to large R , then the scattering matrix is easily connected to \mathbf{Y} at large R .

(C) VARIATIONAL CALCULATIONS

Now let us return to the Kohn variational theory that was introduced in [section A3.11.2.8](#). Here we demonstrate how [equation \(A3.11.46\)](#) may be evaluated using basis set expansions and linear algebra. This discussion will be restricted to scattering in one dimension, but generalization to multidimensional problems is very similar.

To construct $\tilde{\Psi}$, we use the basis expansion

$$\tilde{\Psi} = -u_0 + \sum_{t=1}^N u_t(r) C_t \quad (\text{A3.11.160})$$

where $u_0(r)$ is a special basis function which asymptotically looks like $u_0(r) \sim v^{-1/2} e^{-ikr}$, and $\mathbf{u}_1 = \mathbf{u}_0^*$ is the outgoing wavepart, multiplied by a coefficient C_1 which is \tilde{S} . Typically the complete form of u_0 is chosen to be

$$u_0 = v^{-1/2} f(r) e^{-ikr} \quad (\text{A3.11.161})$$

where $f(r)$ is a function which is unity at larger r and vanishes at small r . The functions u_2, u_3, \dots are taken to be square integrable and the coefficients C_1, \dots, C_N are to be variationally modified. Now substitute $\tilde{\Psi}$ into the expression for S . This gives

$$\begin{aligned} S &= \tilde{S} + \frac{i}{\hbar} \left\langle -u_0 + \sum_t u_t C_t \left| \hat{H} - E \right| -u_0 + \sum_t u_t C_t \right\rangle \\ &= \tilde{S} + \frac{i}{\hbar} \langle u_0 | \hat{H} - E | u_0 \rangle - \frac{i}{\hbar} \sum_t \langle u_0 | \hat{H} - E | u_t \rangle C_t \\ &\quad - \frac{i}{\hbar} \sum_t \langle u_t | \hat{H} - E | u_0 \rangle C_t + \frac{i}{\hbar} \sum_t \sum_{t'} \langle u_t | \hat{H} - E | u_{t'} \rangle C_t C_{t'}. \end{aligned} \quad (\text{A3.11.162})$$

Let us define a matrix \mathbf{M} via

$$M_{0,0} = \langle u_0 | \hat{H} - E | u_0 \rangle \quad (\text{A3.11.163})$$

$$(M_0)_t = \langle u_t | \hat{H} - E | u_0 \rangle \quad (\text{A3.11.164})$$

$$(M)_{t,t'} = \langle u_t | \hat{H} - E | u_{t'} \rangle. \quad (\text{A3.11.165})$$

Also, employ integration by parts to convert (plus $\tilde{S} = C_1$), yielding

$$\frac{-i}{\hbar} \sum_t \langle u_0 | \hat{H} - E | u_t \rangle C_t = \frac{-i}{\hbar} \sum_t \langle u_t | \hat{H} - E | u_0 \rangle C_t - \frac{-i}{\hbar} (i\hbar \tilde{S}). \quad (\text{A3.11.166})$$

This replaces (A3.11.162) with

$$S = \frac{i}{\hbar} \left(M_{0,0} - 2 \sum_t C_t M_{t,0} + \sum_{t,t'} C_t C_{t'} M_{t,t'} \right). \quad (\text{A3.11.167})$$

Now apply the variational criterion as follows:

$$\frac{\partial}{\partial C_t} S = 0 = \frac{i}{\hbar} \left(-2M_{t,0} + 2 \sum_{t'} C_{t'} M_{t,t'} \right). \quad (\text{A3.11.168})$$

This leads to:

$$\mathbf{M}\mathbf{C} = M_0 \quad (\text{A3.11.169})$$

and thus:

$$\mathbf{C} = \mathbf{M}^{-1} M_0 \quad (\text{A3.11.170})$$

and the S matrix is given by:

$$S = \frac{i}{\hbar} (M_{0,0} - 2M_0^+ \mathbf{M}^{-1} M_0 + M_0^+ \mathbf{M}^{-1} \mathbf{M} \mathbf{M}^{-1} M_0) = \frac{i}{\hbar} (M_{0,0} - M_0^+ \mathbf{M}^{-1} M_0). \quad (\text{A3.11.171})$$

This converts the calculation of S to the evaluation of matrix elements together with linear algebra operations. Generalizations of this theory to multichannel calculations exist and lead to a result of more or less the same form.

A3.11.5 CUMULATIVE REACTION PROBABILITIES

A special feature of quantum scattering theory as it applies to chemical reactions is that in many applications

it is only the cumulative reaction probability (CRP) that is of interest in determining physically measurable properties such as the reactive rate constant. This probability P_{cum} (also denoted $N(E)$) is obtained from the S matrix through the formula:

$$P_{\text{cum}} = \sum_i \sum_f |S_{if}|^2. \quad (\text{A3.11.172})$$

Note that the sums are restricted to the portion of the full S matrix that describes reaction (or the specific reactive process that is of interest). It is clear from this definition that the CRP is a highly averaged property where there is no information about individual quantum states, so it is of interest to develop methods that determine this probability directly from the Schrödinger equation rather than indirectly from the scattering matrix. In this section we first show how the CRP is related to the physically measurable rate constant, and then we discuss some rigorous and approximate methods for directly determining the CRP. Much of this discussion is adapted from Miller and coworkers [44, 45].

A3.11.5.1 RATE CONSTANTS

Consider first a gas phase bimolecular reaction ($A + B \rightarrow C + D$). If we consider that the reagents are approaching each other with a relative velocity v , then the total flux of A moving toward B is just vC_A where C_A is the concentration of A (number of A per unit volume (or per unit length in one dimension)). If σ is the integral cross section for reaction between A and B for a given velocity v (σ is the reaction probability in one dimension), then for every B, the number of reactive collisions per unit time is $\sigma v C_A$. The total number of reactive collisions per unit time per unit volume (or per unit length in one dimension) is then $\sigma v C_A C_B$ where C_B is the concentration of B. Equating this to the rate constant k times $C_A C_B$ leads us to the conclusion that

$$k = \sigma v. \quad (\text{A3.11.173})$$

This rate constant refers to reactants which all move with a velocity v whereas the usual situation is such that we have a Boltzmann distribution of velocities. If so then the rate constant is just the average of (A3.11.173) over a Boltzmann distribution P_B :

$$k(T) = \int_0^\infty P_B(v) v \sigma(v) dv. \quad (\text{A3.11.174})$$

This expression is still oversimplified, as it ignores the fact that the molecules A and B have internal states and that the cross section σ depends on these states; σ depends also on the internal states of the products C and D. Letting the indices i and f denote the internal states of the reagents and products respectively, we find that σ in equation (A3.11.174) must be replaced by $\sum_f \sigma_{if}$ and the Boltzmann average must now include the internal states.

Thus, equation (A3.11.174) becomes:

$$k(T) = \sum_i p_B(i) \int_0^\infty P_B(v_i) v_i \sum_f \sigma_{if}(v_i) dv_i$$

where $p_B(i)$ is the internal state Boltzmann distribution.

Now let us write down explicit expressions for $p_B(i)$, $P_B(v_i)$ and σ_{if} . Denoting the internal energy for a given state i as ε_i and the relative translational energy as $E_i = 1/2\mu v_i^2$, we have (in three dimensions)

$$p_B(i) = e^{-\varepsilon_i/kT} / Q_{\text{int}} \quad (\text{A3.11.176})$$

and

$$P_B = 4\pi(\mu/2\pi kT)^{3/2} v_i^2 \exp(-\mu v_i^2/2kT) \quad (\text{A3.11.177})$$

where Q_{int} is the internal state partition function.

The cross section σ_{if} is related to the partial wave reactive scattering matrix $\times S_{if}^J$ through the partial wave sum (i.e., [equation \(A3.11.117\)](#) evaluated for $n_A n_B \neq n_A n_B$).

$$\sigma_{if} = \frac{\pi}{k_i^2} \sum_J (2J+1) |S_{if}^J|^2 \quad (\text{A3.11.178})$$

where $k_i = \mu v_i / \hbar$. Now substitute [equation \(A3.11.176\)](#), [equation \(A3.11.177\)](#) and [equation \(A3.11.178\)](#) into [\(A3.11.175\)](#). Replacing the integral over v_i by one over E_i leads us to the expression

$$k(T) = \frac{(2\pi\hbar)^2}{Q_{\text{int}}(2\pi\mu kT)^{3/2}} \sum_i e^{-\varepsilon_i/kT} \int_0^\infty e^{-E_i/kT} \sum_J (2J+1) \sum_f |S_{if}^J|^2 dE_i. \quad (\text{A3.11.179})$$

If we now change the integration variable from E_i to the total energy $E = E_i + \varepsilon_i$, we can rewrite [equation \(A3.11.179\)](#) as

$$k(T) = \frac{kT}{h} \frac{1}{Q_{\text{int}} Q_{\text{trans}}} \int_0^\infty e^{-E/kT} P_{\text{cum}}(E) dE/kT \quad (\text{A3.11.180})$$

where Q_{trans} is the translational partition function per unit volume:

$$Q_{\text{trans}} = \left(\frac{2\pi\mu kT}{h^2} \right)^{3/2} \quad (\text{A3.11.181})$$

and P_{cum} is the cumulative reaction probability that we wrote down in [equation \(A3.11.172\)](#), but generalized to include a sum over the conserved total angular momentum J weighted by the usual $2J + 1$ degeneracy:

$$P_{\text{cum}}(E) = \sum_J (2J + 1) \sum_i \sum_f |S_{if}^J|^2. \quad (\text{A3.11.182})$$

Note that in deriving [equation \(A3.11.180\)](#), we have altered the lower integration limit in [equation \(A3.11.182\)](#) from zero to $-\varepsilon_i$ by defining S_{if}^J to be zero for $E_i < 0$.

In one physical dimension, [equation \(A3.11.180\)](#) still holds, but Q_{trans} is given by its one dimensional counterpart and [\(A3.11.172\)](#) is used for the CRP.

A3.11.5.2 TRANSITION STATE THEORY

The form of [equation \(A3.11.182\)](#) is immediately suggestive of statistical approximations. If we assume that the total reaction probability $\sum_f |S_{if}^J|^2$ is zero for $E < E_i^\ddagger$ and unity for $E \geq E_i^\ddagger$ where E_i^\ddagger is the energy of a critical bottleneck (commonly known as the transition state) then

$$P_{\text{cum}}^\ddagger = \sum_J (2J + 1) \sum_i h(E - E_i^\ddagger) \quad (\text{A3.11.183})$$

where h is a Heaviside (step) function which is unity for positive arguments and zero for negative arguments, and we have added the subscript i to E_i^\ddagger since the bottleneck energies will in general be dependent on internal state.

[Equation \(A3.11.183\)](#) is simply a formula for the number of states energetically accessible at the transition state and [equation \(A3.11.180\)](#) leads to the thermal average of this number. If we imagine that the states of the system form a continuum, then $P_{\text{cum}}^\ddagger(E)$ can be expressed in terms of a density of states ρ as in

$$P_{\text{cum}}^\ddagger(E) = \int_0^E \rho^\ddagger(\varepsilon) d\varepsilon. \quad (\text{A3.11.184})$$

Substituting this into the integral in [equation \(A3.11.180\)](#) and inverting the order of integration, one obtains

-40-

$$\int_0^\infty e^{-E/kT} \left(\int_0^E \rho^\ddagger(\varepsilon) d\varepsilon \right) dE/kT = \int_0^\infty \rho^\ddagger(\varepsilon) \left(\int_\varepsilon^\infty e^{-E/kT} dE/kT \right) d\varepsilon. \quad (\text{A3.11.185})$$

The inner integral on the right-hand side is just $e^{-\varepsilon/kT}$, so [equation \(A3.11.185\)](#) reduces to the transition state partition function (leaving out relative translation):

$$Q^\ddagger = \int_0^\infty e^{-\varepsilon/kT} \rho^\ddagger(\varepsilon) d\varepsilon. \quad (\text{A3.11.186})$$

Using this in [equation \(A3.11.180\)](#) gives the following

$$k(T) = \frac{kT}{h} \frac{Q^\ddagger}{Q_{\text{int}} Q_{\text{trans}}}. \quad (\text{A3.11.187})$$

This is commonly known as the *transition state theory* approximation to the rate constant. Note that all one needs to do to evaluate (A3.11.187) is to determine the partition function of the reagents and transition state, which is a problem in statistical mechanics rather than dynamics. This makes transition state theory a very useful approach for many applications. However, what is left out are two potentially important effects, tunnelling and barrier recrossing, both of which lead to CRPs that differ from the sum of step functions assumed in [\(A3.11.183\)](#).

A3.11.5.3 EXACT QUANTUM EXPRESSIONS FOR THE CUMULATIVE REACTION PROBABILITY

An important development in the quantum theory of scattering in the last 20 years has been the development of exact expressions which directly determine either $P_{\text{cum}}(E)$ or the thermal rate constant $k(T)$ from the Hamiltonian H . Formally, at least, these expressions avoid the determination of scattering wavefunctions and any information related to the internal states of the reagents or products. The fundamental derivations in this area have been presented by Miller [44] and by Schwartz *et al* [45].

The basic expression of $P_{\text{cum}}(E)$ is

$$P_{\text{cum}}(E) = \frac{1}{2} (2\pi\hbar)^2 \text{Tr}[\hat{F} \delta(E - \hat{H}) \hat{F} \delta(E - \hat{H})] \quad (\text{A3.11.188})$$

where \hat{F} is the symmetrized flux operator:

$$\hat{F} = \frac{1}{2} \left\{ \frac{\hat{p}}{m} \delta(s) + \delta(s) \frac{\hat{p}}{m} \right\}. \quad (\text{A3.11.189})$$

Note that [equation \(A3.11.188\)](#) includes a quantum mechanical trace, which implies a sum over states. The states used for this evaluation are arbitrary as long as they form a complete set and many choices have been considered in recent work. Much of this work has been based on wavepackets [46] or grid point basis functions [47].

An exact expression for the thermal rate constant is given by:

$$k = Q^{-1} \int_0^\infty dt C_f(t) \quad (\text{A3.11.190})$$

where $C_f(t)$ is a flux–flux correlation function

$$C_f(t) = \text{Tr}(\hat{F} e^{i\hat{H}t_c/\hbar} \hat{F} e^{-i\hat{H}t_c/\hbar}). \quad (\text{A3.11.191})$$

Here t_c is a complex time which is given by $t_c = t - i\hbar/2kT$. Methods for evaluating this equation have included path integrals [45], wavepackets [48, 49] and direct evaluation of the trace in square integrable basis sets [50].

A3.11.6 CLASSICAL AND SEMICLASSICAL SCATTERING THEORY

Although the primary focus of this article is on quantum scattering theory, it is important to note that classical and semiclassical approximations play an important role in the application of scattering theory to problems in chemical physics. The primary reason for this is that the de Broglie wavelength associated with motions of atoms and molecules is typically short compared to the distances over which these atoms and molecules move during a scattering process. There are exceptions to this of course, in the limits of low temperature and energy, and for light atoms such as hydrogen atoms, but for a very broad sampling of problems the dynamics is close to the classical limit.

A3.11.6.1 CLASSICAL SCATTERING THEORY FOR A SINGLE PARTICLE

Consider collisions between two molecules A and B. For the moment, ignore the structure of the molecules, so that each is represented as a particle. After separating out the centre of mass motion, the classical Hamiltonian that describes this problem is

$$H = \frac{1}{2}\mu \dot{r}^2 + V(r) \quad (\text{A3.11.192})$$

where the reduced mass is $\mu = m_A m_B / (m_A + m_B)$ and the potential V only depends on the distance r between the particles. Because of the spherical symmetry of the potential, motion of the system is confined to a plane. It is convenient to use polar coordinates, r, θ, ϕ and to choose the plane of motion such that $\phi = 0$. In this case the orbital angular momentum is:

$$|L| = |r \times p_r| = \mu r^2 \dot{\theta}. \quad (\text{A3.11.193})$$

-42-

Since angular momentum is conserved, [equation \(A3.11.192\)](#) may be rearranged to give the following implicit equation for the time dependence of r :

$$\int_{r_1}^{r_2} \frac{dr}{\sqrt{(E - L^2/2\mu r^2 - V)2/\mu}} = t_2 - t_1. \quad (\text{A3.11.194})$$

The time dependence of θ can then be obtained by integrating [\(A3.11.193\)](#).

The physical situation of interest in a scattering problem is pictured in [figure A3.11.3](#). We assume that the initial particle velocity v is coincident with the z axis and that the particle starts at $z = -\infty$, with $x = b = \text{impact parameter}$, and $y = 0$. In this case, $L = \mu v b$. Subsequently, the particle moves in the x, z plane in a trajectory that might be as pictured in [figure A3.11.4](#) (here shown for a hard sphere potential). There is a point of closest approach, i.e., $r = r_2$ (inner turning point for r motions) where

$$E = \frac{L^2}{2\mu r_{\infty}^2} + V(r_{\infty}). \quad (\text{A3.11.195})$$

If we define $t_1 = 0$ at $r = r_{\infty}$, then the explicit trajectory motion is determined by

$$t = \int_{r_{\infty}}^{r(t)} \frac{dr}{\sqrt{(E - L^2/2\mu r^2 - V(r))2/\mu}} \quad (\text{A3.11.196})$$

$$\theta(t) = \pi + L \int_{-\infty}^t \frac{dr}{\mu r(t')^2}. \quad (\text{A3.11.197})$$

The final scattering angle θ is defined using $\theta = \theta(t = \infty)$. There will be a correspondence between b and θ that will tend to look like what is shown in [figure A3.11.5](#) for a repulsive potential (here given for the special case of a hard sphere potential).

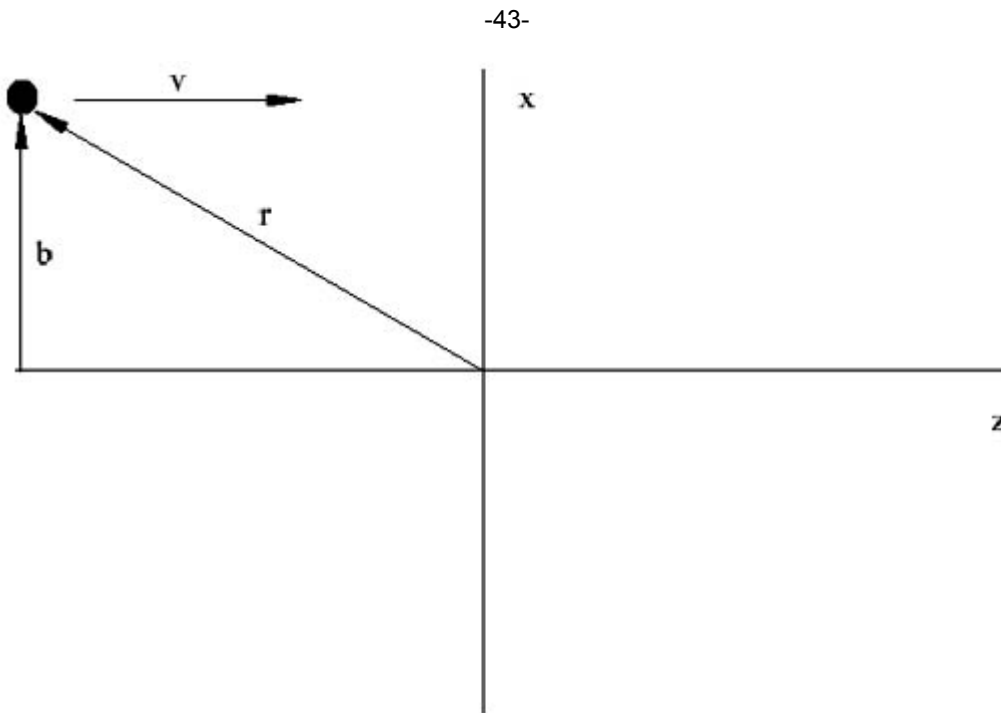


Figure A3.11.3. Coordinates for scattering of a particle from a central potential.

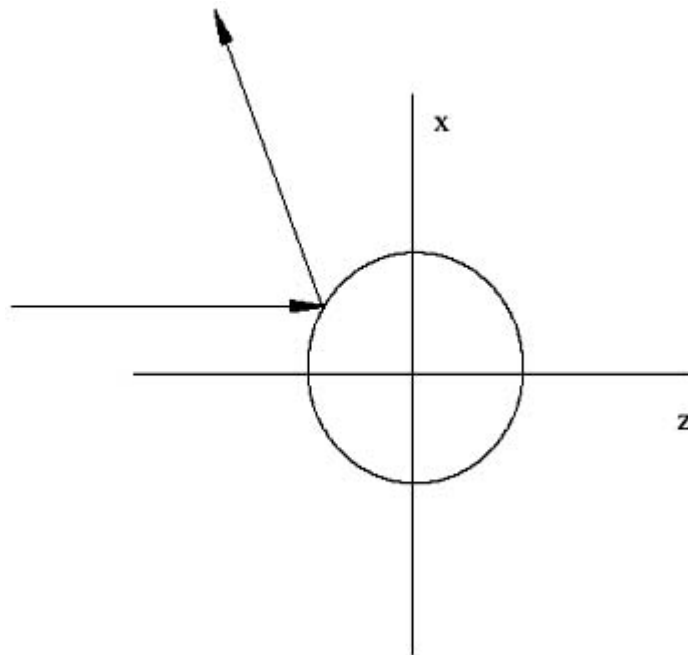


Figure A3.11.4. Trajectory associated with a particle scattering off a hard sphere potential.

-44-

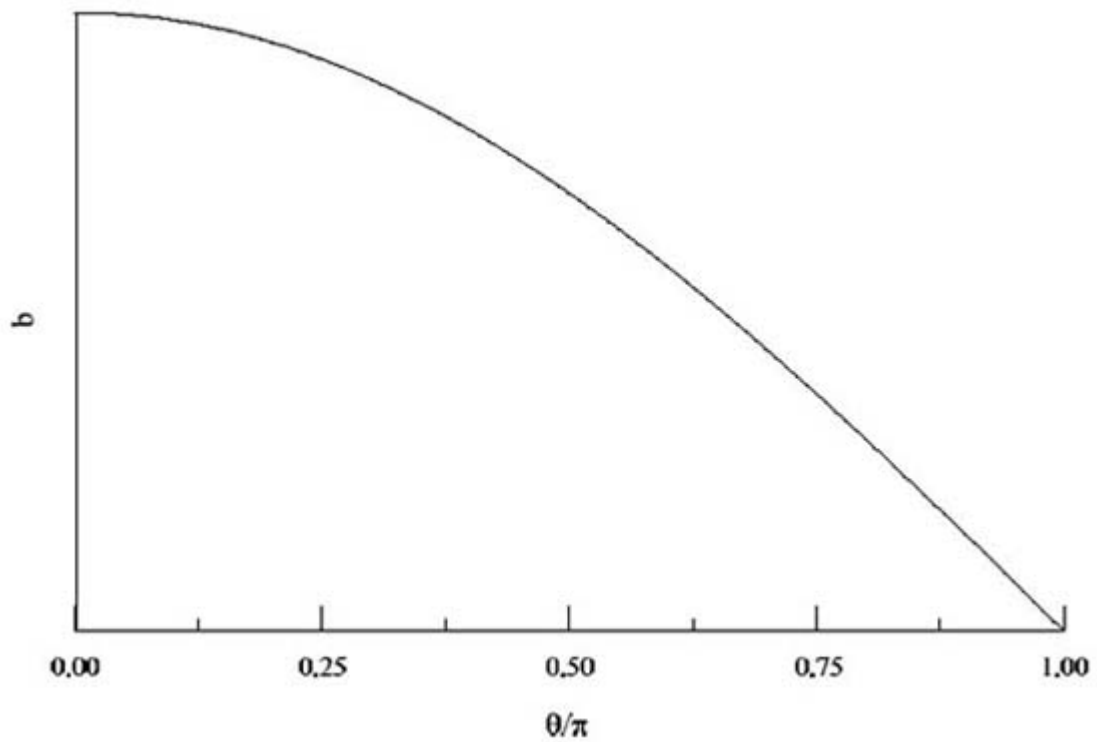


Figure A3.11.5. Typical dependence of b on τ (shown for a hard sphere potential).

In an ensemble of collisions, the impact parameters are distributed randomly on a disc with a probability distribution $P(b)$ that is defined by $P(b) db = 2\pi b db$. The *cross section* $d\sigma$ is then defined by

$$d\sigma = 2\pi b db. \quad (\text{A3.11.198})$$

Now $d\sigma = (d\sigma/d\Omega) d\Omega$ or $(d\sigma/2\pi d \cos\theta) 2\pi d \cos\theta = [d\sigma/2\pi d(\cos\theta)] 2\pi \sin\theta d\theta = I(\omega) 2\pi \sin\theta d\theta$ where $I(\omega)$ is the differential cross section. Therefore

$$I(\omega) = \frac{b db}{\sin\theta d\theta} = \frac{b}{\sin\theta} \left| \frac{db}{d\theta} \right| \quad (\text{A3.11.199})$$

where the absolute value takes care of the case when $db/d\theta < 0$. The integral cross section is

$$\sigma = 2\pi \int_0^\pi I \sin\theta d\theta = 2\pi \int_0^{b_{\max}} b db \quad (\text{A3.11.200})$$

-45-

where b_{\max} is the value of the impact parameter that is associated with a scattering angle of 0 (i.e., scattering in the forward direction). Note that for a potential with infinite range (one that does not go to zero until $r = \infty$), the cross section predicted by (A3.11.200) is infinite. This is not generally correct, except for a Coulomb ($1/r$) potential. This is a classical artifact; the corresponding quantum mechanical result is finite.

A simple example of a finite range potential is the *hard sphere*, for which $V(r) = 0$ for $r > a$, $V(r) = \infty$ for $r < a$. By geometry one can show that $2\phi + \theta = \pi$ and $\sin\phi = b/a$. Therefore

$$b = a \sin\left(\frac{\pi}{2} - \frac{\theta}{2}\right) = a \cos\theta/2 \quad (\text{A3.11.201})$$

$$\frac{db}{d\theta} = \frac{-a}{2} \sin\theta/2 \quad (\text{A3.11.202})$$

$$I(\omega) = \frac{a \cos\theta/2 (a/2) \sin\theta/2}{\sin\theta} = \frac{a^2}{4} \quad (\text{A3.11.203})$$

and

$$\sigma = 2\pi \frac{a^2}{4} \int \sin\theta d\theta = \pi a^2. \quad (\text{A3.11.204})$$

This shows that the differential cross section is independent of angle for this case, and the integral cross section is, just as expected, the area of the circle associated with the radius of the sphere. More generally it is important to note that there can be many trajectories which give the same θ for different b' . In this case the DCS is just the sum over trajectories.

$$f_k(\theta) = \sum_l a_l^k P_l(\cos\theta). \quad (\text{A3.11.205})$$

An explicit result for the differential cross section (DCS) can be obtained by substituting $L = pb = \mu vb$ into the

following expression:

$$\frac{\dot{\theta}}{\dot{r}} = \frac{d\theta}{dr} = \frac{L/\mu r^2}{\sqrt{(2/\mu)(E - V - L^2/2\mu r^2)}} = \frac{b/r^2}{\sqrt{(1 - V/E - (2\mu E)b^2/2\mu E r^2)}} \quad (\text{A3.11.206})$$

$$= \frac{b/r^2}{\sqrt{1 - V/E - b^2/r^2}}$$

To integrate this expression, we note that θ starts at π when $r = \infty$, then it decreases while r decreases to its turning point, then r retraces back to ∞ while θ continues to evolve back to π . The *total change* in θ is then *twice* the integral

-46-

$$\theta = \pi - 2b \int_{r_c}^{\infty} \frac{dr}{r^2} \left(1 - \frac{V}{E} - \frac{b^2}{r^2}\right)^{-1/2}. \quad (\text{A3.11.207})$$

Note that θ obtained this way can be negative. Because of cylindrical symmetry, only $|\theta|$ (or $\theta \bmod \pi$) means anything.

For a typical interatomic potential such as a 6-12 potential, $\theta(b)$ looks like figure A3.11.6 rather than A3.11.5. This shows that for some θ there are three b (one for positive θ and two for negative θ) that contribute to the DCS. The θ where the number of contributing trajectories changes value are sometimes called *rainbow angles*. At these angles, the classical differential cross sections have singularities.

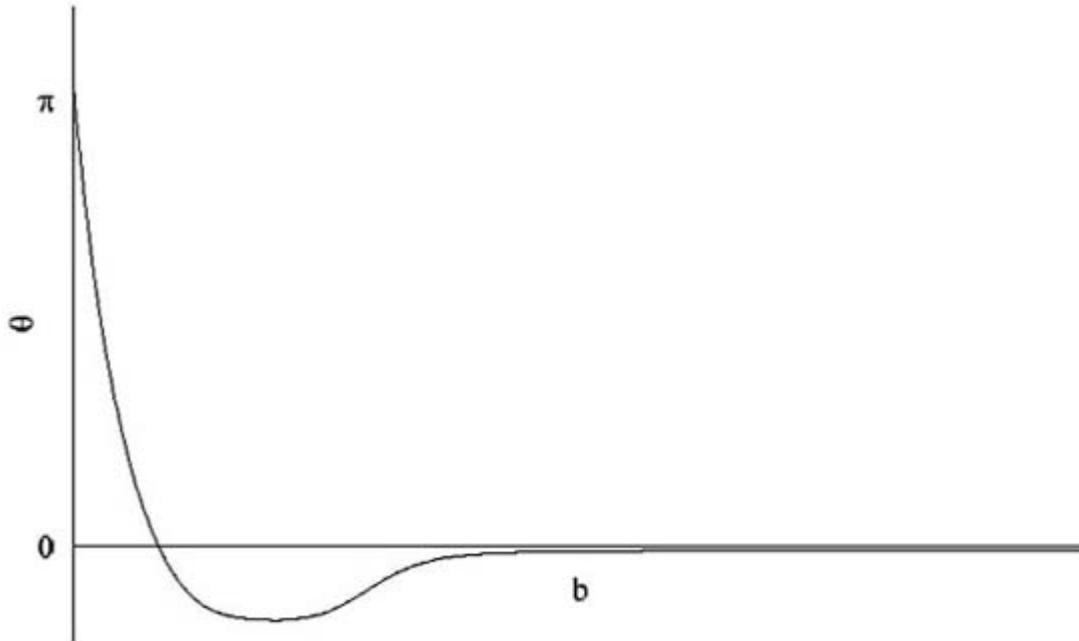


Figure A3.11.6. Dependence of scattering angle τ on impact parameter for a 6-12 potential.

To generalize what we have just done to reactive and inelastic scattering, one needs to calculate numerically integrated trajectories for motions in many degrees of freedom. This is most convenient to develop in space-fixed Cartesian coordinates. In this case, the classical equations of motion (Hamilton's equations) are given by:

$$\dot{x}_i = \frac{\partial H}{\partial p_i} = \frac{p_i}{m_i} \quad (\text{A3.11.208})$$

-47-

$$\dot{p}_i = \frac{\partial H}{\partial x_i} = -\frac{\partial V}{\partial x_i} \quad (\text{A3.11.209})$$

where m_i is the mass associated with the i th degree of freedom and the second equality applies to Cartesian coordinates. Methods for solving these equations of motion have been described in review articles, as have procedures for defining initial conditions [27]. Note that for most multidimensional problems it is necessary to average over initial conditions that represent the internal motions of the species undergoing collision. These averages are often determined by Monte Carlo integration (i.e., randomly sampling the coordinates that need to be averaged over). The initial conditions may be chosen from canonical or microcanonical ensembles, or they may be chosen to mimic an initially prepared quantum state. In the latter case, the trajectory calculation is called a 'quasiclassical' trajectory calculation.

A3.11.6.3 SEMICLASSICAL THEORY

The obvious defect of classical trajectories is that they do not describe quantum effects. The best known of these effects is tunnelling through barriers, but there are others, such as effects due to quantization of the reagents and products and there are a variety of interference effects as well. To circumvent this deficiency, one can sometimes use semiclassical approximations such as WKB theory. WKB theory is specifically for motion of a particle in one dimension, but the generalizations of this theory to motion in three dimensions are known and will be mentioned at the end of this section. More complete descriptions of WKB theory can be found in many standard texts [1, 2, 3, 4 and 5, 18].

(A) WKB THEORY

In WKB theory, one generates a wavefunction that is valid in the $\hbar \rightarrow 0$ limit using a linear combination of exponentials of the form

$$\psi(x) = A(x) e^{iS(x)/\hbar} \quad (\text{A3.11.210})$$

where $A(x)$ and $S(x)$ are real (or sometimes purely imaginary) functions that are derived from the Hamiltonian. This expression is, of course, very familiar from scattering theory applications described above (A3.11.2), where $A(x)$ is a constant, and $S(x)$ is kx . More generally, by substituting (A3.11.210) into the time independent Schrödinger equation in one dimension, and expanding $A(x)$ and $S(x)$ in powers of \hbar , one can show that the leading terms representing $S(x)$ have the form:

$$(\text{A3.11.211})$$

$$A(x) = [E - V(x)]^{-1/4}$$

$$S(x) = \pm \int \sqrt{\frac{2m}{\hbar^2} (E - V(x))} dx.$$

Note that the integrand in $S(x)$ is just the classical momentum $p(x)$, so $S(x)$ is the classical *action* function. In addition, $A(x)$ is proportional to $p^{-1/2}$, which means that $|\psi|_2$ is proportional to the inverse of the classical velocity of the particle.

-48-

This is just the usual classical expression for the probability density. Note that $A(x)$ and $S(x)$ are real as long as motion is classically allowed, meaning that $E > V(x)$. If $E < V(x)$, then $S(x)$ becomes imaginary and $\psi(x)$ involves real rather than complex exponentials. At the point of transition between allowed and forbidden regions, i.e., at the so-called *turning points* of the classical motion, $A(x)$ becomes infinite and the solutions above are not valid. However, it is possible to ‘connect’ the solutions on either side of the turning point using ‘connection formulas’ that are determined from exact solutions to the Schrödinger equation near the turning point. The reader should consult the standard textbooks [1, 2, 3, 4 and 5, 18] for a detailed discussion of this.

In applications to scattering theory, one takes linear combinations of functions of the form (A3.11.210) to satisfy the desired boundary conditions and one uses the connection formulas to determine wavefunctions that are valid for all values of x . By examining the asymptotic forms of the wavefunction, scattering information can be determined. For example, in applications to scattering from a central potential, one can solve the radial Schrödinger equation using WKB theory to determine the phase shift for elastic scattering. The explicit result depends on how many turning points there are in the radial motion.

(B) SCATTERING THEORY FOR MANY DEGREES OF FREEDOM

For multidimensional problems, the generalization of WKB theory to the description of scattering problems is often called Miller–Marcus or classical S -matrix theory [51]. The reader is referred to review articles for a more complete description of this theory [52].

Another theory which is used to describe scattering problems and which blends together classical and quantum mechanics is the semiclassical wavepacket approach [53]. The basic procedure comes from the fact that wavepackets which are initially Gaussian remain Gaussian as a function of time for potentials that are constant, linear or quadratic functions of the coordinates. In addition, the centres of such wavepackets evolve in time in accord with classical mechanics. We have already seen one example of this with the free particle wavepacket of equation (A3.11.7). Consider the general quadratic Hamiltonian (still in one dimension but the generalization to many dimensions is straightforward)

$$\hat{H} = -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + V_0 + V_x(x - x_t) + \frac{1}{2} V_{xx}(x - x_t)^2. \quad (\text{A3.11.212})$$

The Gaussian wavepacket is written as

$$\psi(x, t) = \exp[(i/\hbar)\alpha_t(x - x_t)^2 + (i/\hbar)p_t(x - x_t) + (i/\hbar)\gamma_t]. \quad (\text{A3.11.213})$$

Here x_t and p_t are real time dependent quantities that specify the average position and momentum of the wavepacket ($p_t = \langle p \rangle, x_t = \langle x \rangle$) and α_t and γ_t are complex functions which determine the width, phase and normalization of the wavepacket.

Inserting equation (A3.11.213) into $\hat{H}\psi = i\hbar\partial\psi/\partial t$, and using equation (A3.11.212), leads to the following relation:

-49-

$$-\dot{\alpha}_t(x-x_t)^2 + (2\alpha_t\dot{x}_t - \dot{p}_t)(x-x_t) - \dot{\gamma}_t + p_t\dot{x}_t] \psi = \left[\left(\frac{2}{m} \right) \alpha_t^2 + \frac{1}{2} V_{xx} \right] (x-x_t)^2 + (2\alpha_t p_t/m + V_x)(x-x_t) + V_0 - i\hbar\alpha_t/m + p_t^2/2m \} \psi. \quad (\text{A3.11.214})$$

Comparing coefficients of like powers of $(x-x_t)$ then gives us three equations involving the four unknowns:

$$\dot{\alpha}_t = -(2/m)\alpha_t^2 - V_{xx}/2 \quad (\text{A3.11.215a})$$

$$2\alpha_t\dot{x}_t - \dot{p}_t = 2\alpha_t p_t/m + V_x \quad (\text{A3.11.215b})$$

$$\dot{\gamma}_t = i\hbar\alpha_t/m + p_t\dot{x}_t - V_0 - p_t^2/2m. \quad (\text{A3.11.215c})$$

To develop an additional equation, we simply make the *ansatz* that the first term on the left-hand side of equation (3.11.215b) equals the first term on the right-hand side and similarly with the second term. This immediately gives us Hamilton's equations

$$\dot{x}_t = p_t/m \quad (\text{A3.11.216a})$$

$$\dot{p}_t = -V_x \quad (\text{A3.11.216b})$$

from which it follows that x_t and p_t are related through the classical Hamiltonian function

$$H = p_t^2/2m + V_0 = E. \quad (\text{A3.11.217})$$

Equations (A3.11.216) can then be cast in the general form

$$\dot{x}_t = \partial H / \partial p_t \quad (\text{A3.11.218a})$$

$$-\dot{p}_t = \partial H / \partial x_t \quad (\text{A3.11.218b})$$

and the remaining two equations in equation (A3.11.215) become

$$\dot{\alpha}_t = -(2/m)\alpha_t^2 - \frac{1}{2} V_{xx} \quad (\text{A3.11.219a})$$

$$\dot{\gamma}_t = i\hbar\alpha_t/m + p_t\dot{x}_t - E. \quad (\text{A3.11.219b})$$

It is not difficult to show that, for a constant potential, equation (A3.11.218) and equation (A3.11.219) can be solved to give the free particle wavepacket in [equation \(A3.11.7\)](#). More generally, one can solve equation (A3.11.218) and equation (A3.11.219) numerically for *any* potential, even potentials that are not quadratic, but the solution obtained will be exact only for potentials that are constant, linear or quadratic. The deviation between the exact and Gaussian wavepacket solutions for other potentials depends on how close they are to being *locally quadratic*, which means

-50-

how well the potential can be approximated by a quadratic potential over the width of the wavepacket. Note that although this theory has many classical features, the $\hbar \rightarrow 0$ limit has not been used. This circumvents problems with singularities in the wavefunction near classical turning points that cause trouble in WKB theory.

REFERENCES

- [1] Messiah A 1965 *Quantum Mechanics* (Amsterdam: North-Holland)
- [2] Schiff L I 1968 *Quantum Mechanics* (New York: McGraw-Hill)
- [3] Merzbacher E 1970 *Quantum Mechanics* (New York: Wiley)
- [4] Davydov A S 1976 *Quantum Mechanics* (Oxford: Pergamon)
- [5] Sakurai J J 1985 *Modern Quantum Mechanics* (Menlo Park: Benjamin-Cummings)
- [6] Adhi Kari S K and Kowolski K L 1991 *Dynamical Collision Theory and Its Applications* (New York: Academic)
- [7] Cohen-Tannouji C, Diu B and Laloë F 1977 *Quantum Mechanics* (New York: Wiley)
- [8] Newton R G 1982 *Scattering Theory of Waves and Particles* 2nd edn (New York: McGraw-Hill)
- [9] Rodberg L S and Thaler R M 1967 *Introduction to the Quantum Theory of Scattering* (New York: Academic)
- [10] Roman P 1965 *Advanced Quantum Theory* (Reading, MA: Addison-Wesley)
- [11] Simons J and Nichols J 1997 *Quantum Mechanics in Chemistry* (New York: Oxford University Press)
- [12] Levine R D 1969 *Quantum Mechanics of Molecular Rate Processes* (London: Oxford University Press)
- [13] Child M S 1974 *Molecular Collision Theory* (New York: Academic)
- [14] Nikitin E E 1974 *Theory of Elementary Atomic and Molecular Processes in Gases* (Oxford: Clarendon)
- [15] Massey H S W 1979 *Atomic and Molecular Collisions* (London: Taylor and Francis)
- [16] Levine R D and Bernstein R B 1987 *Molecular Reaction Dynamics and Chemical Reactivity* (New York: Oxford University Press)
- [17] Murrell J N and Bosanac S D 1989 *Introduction to the Theory of Atomic and Molecular Collisions* (New York: Wiley)
- [18] Schatz G C and Ratner M A 1993 *Quantum Mechanics in Chemistry* (Englewood Cliffs, NJ: Prentice-Hall)

- [19] Miller W H (ed) 1976 *Dynamics of Molecular Collisions* (New York: Plenum)
- [20] Bernstein R B (ed) 1979 *Atom–Molecule Collision Theory. A Guide for the Experimentalist* (New York: Plenum)
- [21] Truhlar D G (ed) 1981 *Potential Energy Surfaces and Dynamics Calculations* (New York: Plenum)
-

-51-

- [22] Bowman J M (ed) 1983 *Molecular Collision Dynamics* (Berlin: Springer)
- [23] Baer M (ed) 1985 *The Theory of Chemical Reaction Dynamics* (Boca Raton, FL: Chemical Rubber Company)
- [24] Clary D C 1986 *The Theory of Chemical Reaction Dynamics* (Boston: Reidel)
- [25] Baer M and Ng C-Y (eds) 1992 *State-Selected and State-to-State Ion–Molecule Reaction Dynamics Part 2. Theory (Adv. Chem. Phys. 72)* (New York: Wiley)
- [26] Truhlar D G (ed) 1984 *Resonances in Electron–Molecule Scattering, van der Waals Complexes, and Reactive Chemical Dynamics (ACS Symp. Ser. 263)* (Washington, DC: American Chemical Society)
- [27] Thompson D L 1998 *Modern Methods for Multidimensional Dynamics Computations in Chemistry* (Singapore: World Scientific)
- [28] Hase W L (ed) 1998 *Comparisons of Classical and Quantum Dynamics (Adv. in Classical Trajectory Methods III)* (Greenwich, CT: JAI Press)
- [29] Mullin A S and Schatz G C (eds) 1997 *Highly Excited Molecules: Relaxation, Reaction and Structure (ACS Symp. Ser. 678)* (Washington, DC: American Chemical Society)
- [30] Rost J M 1998 Semiclassical s-matrix theory for atomic fragmentation *Phys. Rep.* **297** 272–344
- [31] Kaye J A and Kuppermann A 1988 Mass effect in quantum-mechanical collision-induced dissociation in collinear reactive atom diatomic molecule collisions *Chem. Phys.* **125** 279–91
- [32] Miller W H 1994 S-matrix version of the Kohn variational principle for quantum scattering theory of chemical reactions *Adv. Mol. Vibrations and Collision Dynamics* vol 2A, ed J M Bowman (Greenwich, CT: JAI Press) pp 1–32
- [33] Mayne H R 1991 Classical trajectory calculations on gas-phase reactive collisions *Int. Rev. Phys. Chem.* **10** 107–21
- [34] Delves L M 1959 Tertiary and general-order collisions *Nucl. Phys.* **9** 391–9
Delves L M 1960 Tertiary and general-order collisions (II) *Nucl. Phys.* **20** 275–308
- [35] Smith F T 1962 A symmetric representation for three-body problems. I. Motion in a plane *J. Math. Phys.* **3** 735–48
Smith F T and Whitten R C 1968 Symmetric representation for three body problems. II. Motion in space *J. Math. Phys.* **9** 1103–13
- [36] Kuppermann A 1997 Reactive scattering with row-orthonormal hyperspherical coordinates. 2. Transformation properties and Hamiltonian for tetraatomic systems *J. Phys. Chem. A* **101** 6368–83
- [37] Kosloff R 1994 Propagation methods for quantum molecular-dynamics *Annu. Rev. Phys. Chem.* **45** 145–78
- [38] Balint-Kurti G G, Dixon R N and Marston C C 1992 Grid methods for solving the Schrödinger equation and time-dependent quantum dynamics of molecular photofragmentation and reactive scattering processes *Int. Rev. Phys. Chem.* **11** 317–44

- [39] Feit M D and Fleck J A Jr 1983 Solution of the Schrödinger equation by a spectral method. II. Vibrational energy levels of triatomic molecules *J. Chem. Phys.* **78** 301–8
-

-52-

- [40] Tal-Ezer H and Kosloff R 1984 An accurate and efficient scheme for propagating the time dependent Schrödinger equation *J. Chem. Phys.* **81** 3967–71
- [41] Leforestier C *et al* 1991 Time-dependent quantum mechanical methods for molecular dynamics *J. Comput. Phys.* **94** 59–80
- [42] Gordon R G 1969 Constructing wave functions for bound states and scattering *J. Chem. Phys.* **51** 14–25
- [43] Manolopoulos D E 1986 An improved log derivative method for inelastic scattering *J. Chem. Phys.* **85** 6425–9
- [44] Miller W H 1974 Quantum mechanical transition state theory and a new semiclassical model for reaction rate constants *J. Chem. Phys.* **61** 1823–34
- [45] Miller W H, Schwartz S D and Tromp J W 1983 Quantum mechanical rate constants for bimolecular reactions *J. Chem. Phys.* **79** 4889–98
- [46] Zhang D H and Light J C 1996 Cumulative reaction probability via transition state wave packets *J. Chem. Phys.* **104** 6184–91
- [47] Manthe U, Seideman T and Miller W H 1993 Full-dimensional quantum-mechanical calculation of the rate-constant for the $\text{H}_2 + \text{OH} \rightarrow \text{H}_2\text{O} + \text{H}$ reaction *J. Chem. Phys.* **99** 10 078–81
- [48] Wahnstrom G and Metiu H 1988 Numerical study of the correlation function expressions for the thermal rate coefficients in quantum systems *J. Phys. Chem.* **92** 3240–52
- [49] Thachuk M and Schatz G C 1992 Time dependent methods for calculating thermal rate coefficients using flux correlation functions *J. Chem. Phys.* **97** 7297–313
- [50] Day P N and Truhlar D G 1991 Benchmark calculations of thermal reaction rates. II. Direct calculation of the flux autocorrelation function for a canonical ensemble *J. Chem. Phys.* **94** 2045–56
- [51] Miller W H 1970 Semiclassical theory of atom–diatom collisions: path integrals and the classical S matrix *J. Chem. Phys.* **53** 1949–59
Marcus R A 1970 Extension of the WKB method to wave functions and transition probability amplitudes (S-matrix) for inelastic or reactive collisions *Chem. Phys. Lett.* **7** 525–32
- [52] Miller W H 1971 Semiclassical nature of atomic and molecular collisions *Accounts Chem. Res.* **4** 161–7
Miller W H 1974 Classical-limit quantum mechanics and the theory of molecular collisions *Adv. Chem. Phys.* **25** 69–177
Miller W H 1975 Classical S-matrix in molecular collisions *Adv. Chem. Phys.* **30** 77–136
- [53] Heller E 1975 Time dependent approach to semiclassical dynamics *J. Chem. Phys.* **62** 1544–55
-

FURTHER READING

Basic quantum mechanics textbooks that include one or more chapters on scattering theory as applied to physics:

Messiah A 1965 *Quantum Mechanics* (Amsterdam: North-Holland)

Schiff L I 1968 *Quantum Mechanics* (New York: McGraw-Hill)

Merzbacher E 1970 *Quantum Mechanics* (New York: Wiley)

Davydov A S 1976 *Quantum Mechanics* (Oxford: Pergamon)

Cohen-Tannouji C, Diu B and Laloë F 1997 *Quantum Mechanics* (New York: Wiley)

Sakurai J J 1985 *Modern Quantum Mechanics* (Menlo Park: Benjamin-Cummings)

Books that are entirely concerned with scattering theory (only the Levine text is concerned with applications in chemical physics):

Roman P 1965 *Advanced Quantum Theory* (Reading, MA: Addison-Wesley)

Rodberg L S and Thaler R M 1967 *Introduction to the Quantum Theory of Scattering* (New York: Academic)

Levine R D 1969 *Quantum Mechanics of Molecular Rate Processes* (London: Oxford)

Newton R G 1982 *Scattering Theory of Waves and Particles* 2nd edn (New York: McGraw-Hill)

Adhi Kari S K and Kowolski K L 1991 *Dynamical Collision Theory and Its Applications* (New York: Academic)

Murrell J N and Bosanac S D 1989 *Introduction to the Theory of Atomic and Molecular Collisions* (New York: Wiley)

Books with a more chemical bent that include chapters on scattering theory and related issues.

Child M S 1974 *Molecular Collision Theory* (New York: Academic)

Nikitin E E 1974 *Theory of Elementary Atomic and Molecular Processes in Gases* (Oxford: Clarendon)

Massey H S W 1979 *Atomic and Molecular Collisions* (London: Taylor and Francis)

Levine R D and Bernstein R B 1987 *Molecular Reaction Dynamics and Chemical Reactivity* (New York: Oxford University Press)

Schatz G C and Ratner M A 1993 *Quantum Mechanics in Chemistry* (Englewood Cliffs, NJ: Prentice-Hall)

Simons J and Nichols J 1997 *Quantum Mechanics in Chemistry* (New York: Oxford)

A3.12 Statistical mechanical description of chemical kinetics: RRKM

William L Hase

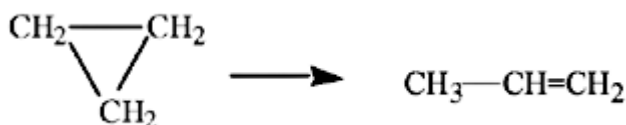
A3.12.1 INTRODUCTION

As reactants transform to products in a chemical reaction, reactant bonds are broken and reformed for the products. Different theoretical models are used to describe this process ranging from time-dependent classical or quantum dynamics [1,2], in which the motions of individual atoms are propagated, to models based on the postulates of statistical mechanics [3]. The validity of the latter models depends on whether statistical mechanical treatments represent the actual nature of the atomic motions during the chemical reaction. Such a statistical mechanical description has been widely used in unimolecular kinetics [4] and appears to be an accurate model for many reactions. It is particularly instructive to discuss statistical models for unimolecular reactions, since the model may be formulated at the elementary microcanonical level and then averaged to obtain the canonical model.

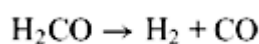
Unimolecular reactions are important in chemistry, physics, biochemistry, materials science, and many other areas of science and are denoted by



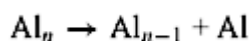
where the asterisk denotes that the unimolecular reactant A contains sufficient internal vibrational/rotational energy to decompose. (Electronic excitation may also promote decomposition of A, but this topic is outside the purview of this presentation.) The energy is denoted by E and must be greater than the unimolecular decomposition threshold energy E_0 . There are three general types of potential energy profiles for unimolecular reactions (see figure A3.12.1). One type is for an isomerization reaction, such as cyclopropane isomerization



for which there is a substantial potential energy barrier separating the two isomers. The other two examples are for unimolecular dissociation. In one case, as for formaldehyde dissociation



there is a potential energy barrier for the reverse association reaction. In the other, as for aluminium cluster dissociation



there is no barrier for association.

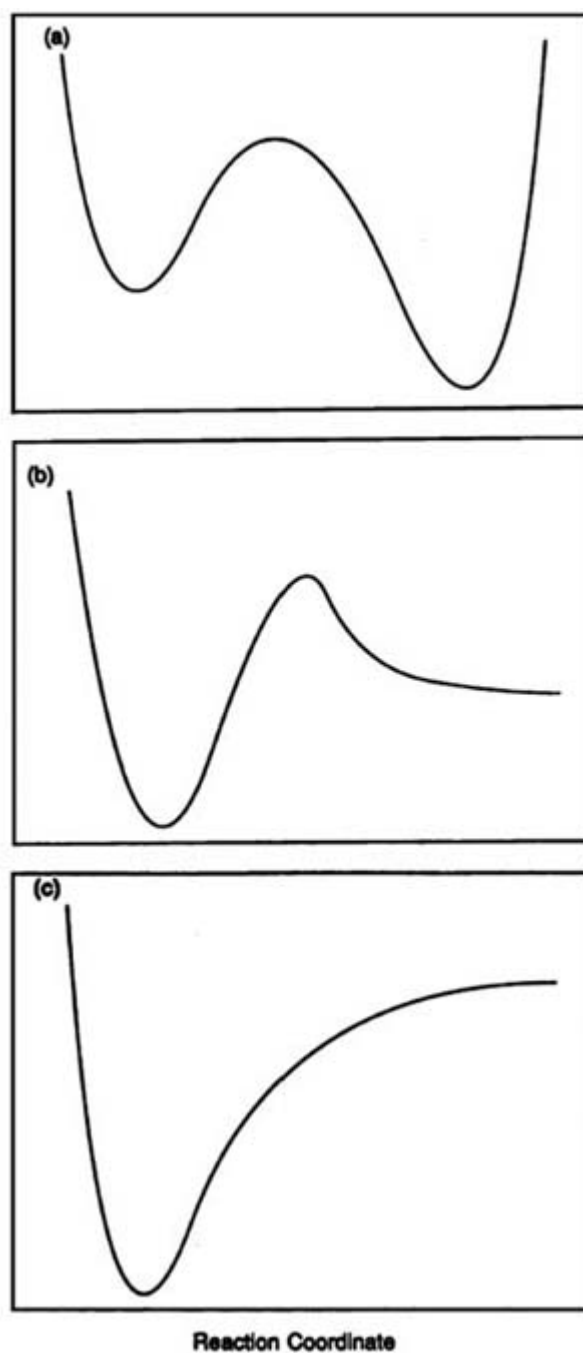
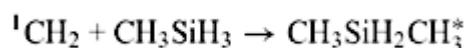
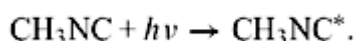


Figure A3.12.1. Schematic potential energy profiles for three types of unimolecular reactions. (a) Isomerization. (b) Dissociation where there is an energy barrier for reaction in both the forward and reverse directions. (c) Dissociation where the potential energy rises monotonically as for rotational ground-state species, so that there is no barrier to the reverse association reaction. (Adapted from [5].)

A number of different experimental methods may be used to energize the unimolecular reactant A. For example, energization can take place by the potential energy release in chemical reaction, i.e.



or by absorption of a single photon,



Extensive discussions of procedures for energizing molecules are given elsewhere [5].

Quantum mechanically, the time dependence of the initially prepared state of A^* is given by its wavefunction $\Psi(t)$, which may be determined from the equation of motion

$$i\hbar \frac{d\Psi(t)}{dt} = \hat{H}\Psi(t). \quad (\text{A3.12.2})$$

At the unimolecular threshold of moderate to large size molecules (e.g. C_2H_6 to peptides), there are many vibrational/rotational states within the experimental energy resolution dE and the initial state of A^* may decay by undergoing transitions to other states and/or decomposing to products. The former is called intramolecular vibrational energy redistribution (IVR) [6]. The probability amplitude versus time of remaining in the initially prepared state is given by

$$C(t) = \langle \Psi(0) | \Psi(t) \rangle \quad (\text{A3.12.3})$$

and is comprised of contributions from both IVR and unimolecular decomposition. The time dependence of the unimolecular decomposition may be constructed by evaluating $|\Psi(t)|^2$ inside the potential energy barrier, within the reactant region of the potential energy surface.

In the statistical description of unimolecular kinetics, known as Rice–Ramsperger–Kassel–Marcus (RRKM) theory [4,7,8], it is assumed that complete IVR occurs on a timescale much shorter than that for the unimolecular reaction [9]. Furthermore, to identify states of the system as those for the reactant, a dividing surface [10], called a transition state, is placed at the potential energy barrier region of the potential energy surface. The assumption implicit in RRKM theory is described in the next section.

A3.12.2 FUNDAMENTAL ASSUMPTION OF RRKM THEORY: MICROCANONICAL ENSEMBLE

RRKM theory assumes a microcanonical ensemble of A^* vibrational/rotational states within the energy interval $E \rightarrow E + dE$, so that each of these states is populated statistically with an equal probability [4]. This assumption of a microcanonical distribution means that the unimolecular rate constant for A^* only depends on energy, and not on the manner in which A^* is energized. If $N(0)$ is the number of A^* molecules excited at $t = 0$ in accord with a microcanonical ensemble, the microcanonical rate constant $k(E)$ is then defined by

$$\left. \frac{-dN(t)}{dt} \right|_{t=0} = k(E)N(t)|_{t=0}. \quad (\text{A3.12.4})$$

The rapid IVR assumption of RRKM theory means that a microcanonical ensemble is maintained as the A* molecules decompose so that, at any time t , $k(E)$ is given by

$$\frac{-dN(t)}{dt} = k(E)N(t). \quad (\text{A3.12.5})$$

As a result of the fixed time-independent rate constant $k(E)$, $N(t)$ decays exponentially, i.e.

$$N(t) = N(0) \exp[-k(E)t]. \quad (\text{A3.12.6})$$

A RRKM unimolecular system obeys the ergodic principle of statistical mechanics[11].

The quantity $-dN(t)/[N(t)dt]$ is called the lifetime distribution $P(t)$ [12] and according to RRKM theory is given by

$$P(t) = k(E) \exp[-k(E)t]. \quad (\text{A3.12.7})$$

[Figure A3.12.2\(a\)](#) illustrates the lifetime distribution of RRKM theory and shows random transitions among all states at some energy high enough for eventual reaction (toward the right). In reality, transitions between quantum states (though coupled) are not equally probable: some are more likely than others. Therefore, transitions between states must be sufficiently rapid and disorderly for the RRKM assumption to be mimicked, as qualitatively depicted in [figure A3.12.2\(b\)](#). The situation depicted in these figures, where a microcanonical ensemble exists at $t = 0$ and rapid IVR maintains its existence during the decomposition, is called *intrinsic* RRKM behaviour [9].

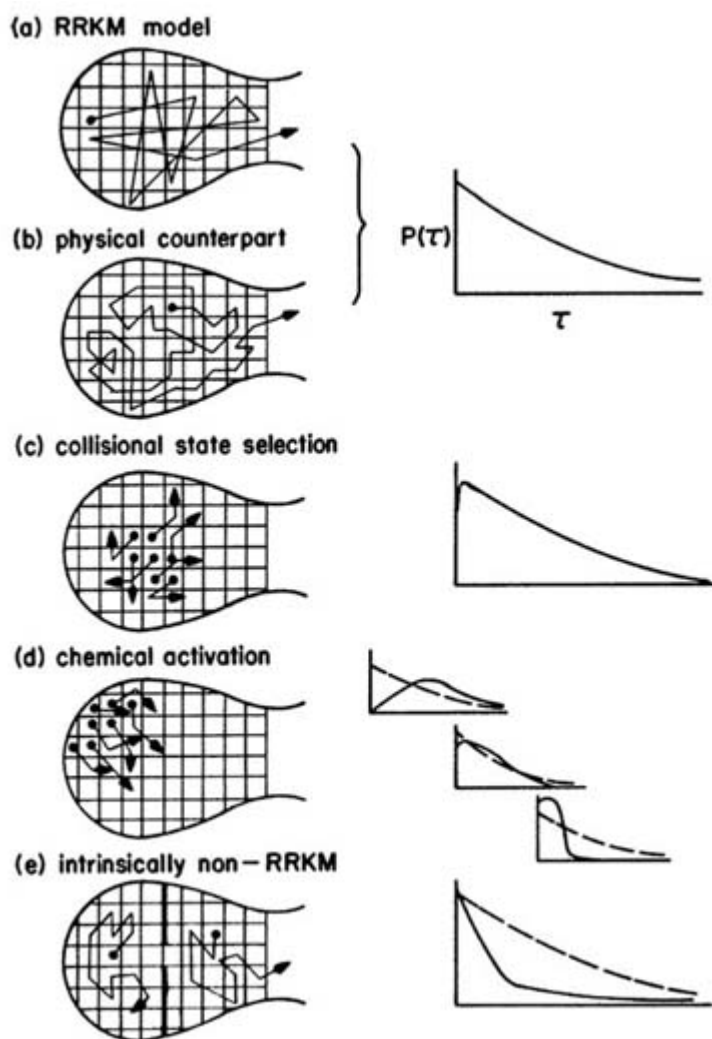


Figure A3.12.2. Relation of state occupation (schematically shown at constant energy) to lifetime distribution for the RRKM theory and for various actual situations. Dashed curves in lifetime distributions for (d) and (e) indicate RRKM behaviour. (a) RRKM model. (b) Physical counterpart of RRKM model. (c) Collisional state selection. (d) Chemical activation. (e) Intrinsically non-RRKM. (Adapted from [9].)

The lifetime distribution will depend in part on the manner in which the energy needed for reaction is supplied. In many experiments, such as photoactivation and chemical activation, the molecular vibrational/rotational states are excited non-randomly. Regardless of the pattern of the initial energizing, the RRKM model of rapid IVR requires the distribution of states to become microcanonical in a negligibly short time. Three different possible lifetime distributions are represented by figure A3.12.2(d). As shown in the middle, the lifetime distribution may be similar to that of RRKM theory. In other cases, the probability of a short lifetime with respect to reaction may be enhanced or reduced, depending on the location of the initial excitation within the molecule. These are examples of *apparent* non-RRKM behaviour [9] arising from the initial non-random excitation. If there are strong internal couplings, $P(t)$ will become

that of RRKM theory, (equation A3.12.5), after rapid IVR. A classic example of apparent non-RRKM behaviour is described below in section A3.12.8.1.

A situation that arises from the intramolecular dynamics of A^* and completely distinct from apparent non-RRKM behaviour is *intrinsic* non-RRKM behaviour [9]. By this, it is meant that A^* has a non-random $P(t)$ even if the internal vibrational states of A^* are prepared randomly. This situation arises when transitions between individual molecular vibrational/rotational states are slower than transitions leading to products. As a result, the vibrational states do not have equal dissociation probabilities. In terms of classical phase space dynamics, slow transitions between the states occur when the reactant phase space is *metrically decomposable* [13,14] on the timescale of the unimolecular reaction and there is at least one *bottleneck* [9] in the molecular phase space other than the one defining the transition state. An intrinsic non-RRKM molecule decays non-exponentially with a time-dependent unimolecular rate constant or exponentially with a rate constant different from that of RRKM theory.

The above describes the fundamental assumption of RRKM theory regarding the intramolecular dynamics of A^* . The RRKM expression for $k(E)$ is now derived.

A3.12.3 THE RRKM UNIMOLECULAR RATE CONSTANT

A3.12.3.1 DERIVATION OF THE RRKM $k(E)$

As discussed above, to identify states of the system as those for the reactant A^* , a dividing surface is placed at the potential energy barrier region of the potential energy surface. This is a classical mechanical construct and classical statistical mechanics is used to derive the RRKM $k(E)$ [4].

In the vicinity of the dividing surface, it is assumed that the Hamiltonian for the system may be separated into the two parts

$$H = H_1 + H' \quad (\text{A3.12.8})$$

where H_1 defines the energy for the conjugate coordinate and momentum pair q_1, p_1 and H' gives the energy for the remaining conjugate coordinates and momenta. This special coordinate q_1 is called the *reaction coordinate*. Reactive systems which have a total energy $H = E$ and a value for q_1 which lies between q_1^\ddagger and $q_1^\ddagger + dq_1^\ddagger$ called microcanonical transition states. The reaction coordinate potential at the transition state is E_0 . The RRKM $k(E)$ is determined from the rate at which these transition states form products.

The hypersurface formed from variations in the system's coordinates and momenta at $H(p, q) = E$ is the microcanonical system's phase space, which, for a Hamiltonian with $3n$ coordinates, has a dimension of $6n - 1$. The assumption that the system's states are populated statistically means that the population density over the whole surface of the phase space is uniform. Thus, the ratio of molecules at the dividing surface to the total molecules $[dN(q_1^\ddagger, p_1^\ddagger)/N]$

may be expressed as a ratio of the phase space at the dividing surface to the total phase space. Thus, at any instant in time, the ratio of molecules whose reaction coordinate and conjugate momentum have values that range from q_1^\ddagger to $q_1^\ddagger + dq_1^\ddagger$ and from p_1^\ddagger to $p_1^\ddagger + dp_1^\ddagger$ to the total number of molecules is given by

$$\frac{dN(q_1^\ddagger, p_1^\ddagger)}{N} = \frac{dq_1^\ddagger dp_1^\ddagger \int \dots \int_{H=E-E_1^\ddagger-E_0} dq_2^\ddagger \dots dq_{3n}^\ddagger dp_2^\ddagger \dots dp_{3n}^\ddagger}{\int \dots \int_{H=E} dq_1 \dots dq_{3n} dp_1 \dots dp_{3n}} \quad (\text{A3.12.9})$$

where E_1^\ddagger is the translational energy in the reaction coordinate. One can think of this expression as a reactant–transition state equilibrium constant for a microcanonical system. The term $dq_1^\ddagger dp_1^\ddagger$ divided by Planck's constant is the number of translational states in the reaction coordinate and the surface integral in the numerator divided by h^{3n-1} is the density of states for the $3n - 1$ degrees of freedom orthogonal to the reaction coordinate. Similarly, the surface integral in the denominator is the reactant density of states multiplied by h^{3n} .

To determine $k(E)$ from equation (A3.12.9) it is assumed that transition states with positive p_1^\ddagger form products. Noting that $p_1^\ddagger = \mu_1 dq_1^\ddagger/dt$, where μ_1 is the reduced mass of the separating fragments, all transition states that lie within q_1^\ddagger and $q_1^\ddagger + dq_1^\ddagger$ with positive p_1^\ddagger will cross the transition state toward products in the time interval $dt = \mu_1 dq_1^\ddagger/p_1^\ddagger$. Inserting this expression into equation (A3.12.9), one finds that the reactant-to-product rate (i.e. flux) through the transition state for momentum p_1^\ddagger is

$$\frac{dN(q_1^\ddagger, p_1^\ddagger)}{dt} = \frac{N \frac{p_1^\ddagger dq_1^\ddagger}{\mu_1} \int \dots \int_{H=E-E_1^\ddagger-E_0} dq_2^\ddagger \dots dq_{3n}^\ddagger dp_2^\ddagger \dots dp_{3n}^\ddagger}{\int \dots \int_{H=E} dq_1 \dots dq_{3n} dp_1 \dots dp_{3n}} \quad (\text{A3.12.10})$$

Since the energy in the reaction coordinate is $E_1^\ddagger = p_1^{\ddagger 2}/2\mu_1$, its derivative is $dE_1^\ddagger = p_1^\ddagger dp_1^\ddagger/\mu_1$ so that equation (A3.12.10) can be converted into

$$\frac{dN(q_1^\ddagger, p_1^\ddagger)}{dt} = \frac{N dE_1^\ddagger \int \dots \int_{H=E-E_1^\ddagger-E_0} dq_2^\ddagger \dots dq_{3n}^\ddagger dp_2^\ddagger \dots dp_{3n}^\ddagger}{\int \dots \int_{H=E} dq_1 \dots dq_{3n} dp_1 \dots dp_{3n}} \quad (\text{A.12.11})$$

This equation represents the reaction rate at total energy E with a fixed energy in the reaction coordinate E_1^\ddagger and may be written as

$$dN(E, E_1^\ddagger)/dt = k(E, E_1^\ddagger) N dE_1^\ddagger \quad (\text{A3.12.12})$$

-8-

where $k(E, E_1^\ddagger)$ is a unimolecular rate constant. As discussed above, the integrals in [equation \(A3.12.11\)](#) are densities of states ρ , so $k(E, E_1^\ddagger)$ becomes

$$k(E, E_1^\ddagger) = \frac{\rho^\ddagger(E - E_0 - E_1^\ddagger)}{h\rho(E)} \quad (\text{A3.12.13})$$

To find the total reaction flux, [equation \(A3.12.12\)](#) must be integrated between the limits E_1^\ddagger equal to 0 and $E - E_0$, so that

$$\frac{dN}{dt} = \frac{\int_0^{E-E_0} dN(E, E_1^\ddagger)}{dt} = N \int_0^{E-E_0} k(E, E_1^\ddagger) dE_1^\ddagger = k(E)N \quad (\text{A3.12.14})$$

where, using equation (A3.12.13), $k(E)$ is given by

$$k(E) = \frac{\int_0^{E-E_0} \rho^\ddagger(E - E_0 - E_1^\ddagger) dE_1^\ddagger}{h\rho(E)} = \frac{N^\ddagger(E - E_0)}{h\rho(E)}. \quad (\text{A3.12.15})$$

The term $N^\ddagger(E - E_0)$ is the sum of states at the transition state for energies from 0 to $E - E_0$. Equation (A3.12.15) is the RRKM expression for the unimolecular rate constant.

Only in the high-energy limit does classical statistical mechanics give accurate sums and densities of state [15]. Thus, in general, quantum statistical mechanics must be used to calculate a RRKM $k(E)$ which may be compared with experiment [16]. A comparison of classical and quantum harmonic (see below) RRKM rate constants for $\text{C}_2\text{H}_5 \rightarrow \text{H} + \text{C}_2\text{H}_4$ is given in figure A3.12.3 [17]. The energies used for the classical calculation are with respect to the reactant's and transition state's potential minima. For the quantum calculation the energies are with respect to the zero-point levels. If energies with respect to the zero-point levels were used in the classical calculation, the classical $k(E)$ would be appreciably smaller than the quantum value [16].

-9-

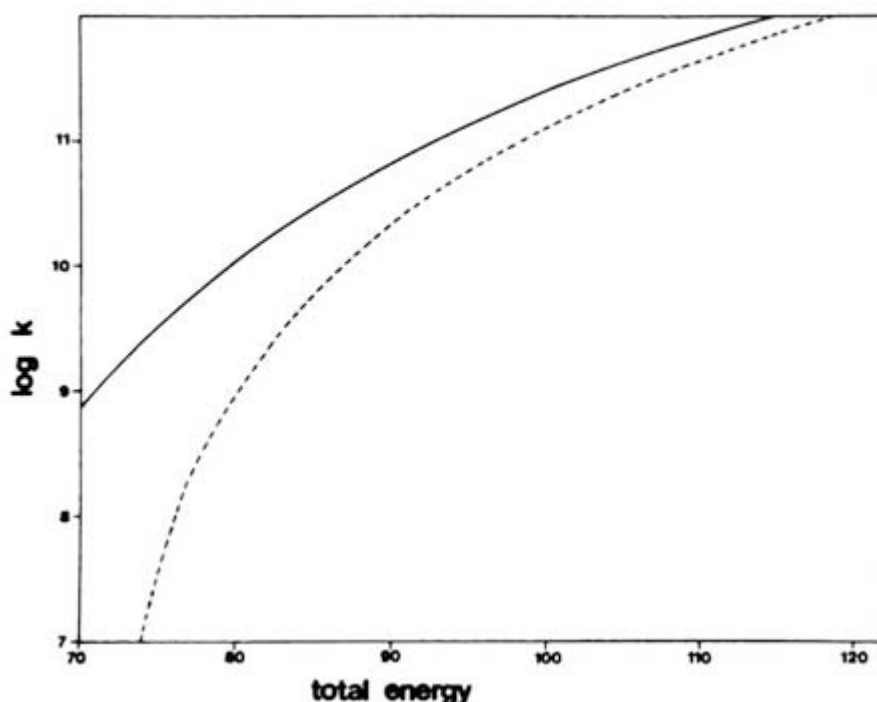


Figure A3.12.3. Harmonic RRKM unimolecular rate constants for $\text{C}_2\text{H}_5 \rightarrow \text{H} + \text{C}_2\text{H}_4$ dissociation: classical state counting (solid curve), quantum state counting (dashed curve). Rate constant is in units of s^{-1} and energy in kcal mol^{-1} . (Adapted from [17].)

RRKM theory allows some modes to be uncoupled and not exchange energy with the remaining modes [16]. In quantum RRKM theory, these uncoupled modes are not *active*, but are *adiabatic* and stay in fixed quantum states n during the reaction. For this situation, equation (A3.12.15) becomes

$$k(E, \mathbf{n}) = \frac{N^\ddagger[E - E_0(\mathbf{n}), \mathbf{n}]}{h\rho(E, \mathbf{n})}. \quad (\text{A3.12.16})$$

In addition to affecting the number of active degrees of freedom, the fixed n also affects the unimolecular threshold $E_0(n)$. Since the total angular momentum j is a constant of motion and quantized according to

$$j = \sqrt{J(J+1)}\hbar \quad (\text{A3.12.17})$$

the quantum number J is fixed during the unimolecular reaction. This may be denoted by explicitly including J in equation (A3.12.16), i.e.

$$k(E, J, \mathbf{n}) = \frac{N^\ddagger(E, J, \mathbf{n})}{\rho(E, J, \mathbf{n})}. \quad (\text{A3.12.18})$$

-10-

The treatment of angular momentum is discussed in detail below in [section A3.12.4.3](#).

A3.12.3.2 K(E) AS AN AVERAGE FLUX

The RRKM rate constant is often expressed as an average classical flux through the transition state [18, 19 and 20]. To show that this is the case, first recall that the density of states $\rho(E)$ for the reactant may be expressed as

$$\rho(E) = \frac{d}{dE} \int \dots \int dq_1 \dots dq_{3n} dp_1 \dots dp_{3n} \theta(E - H) / h^{3n} \quad (\text{A3.12.19})$$

where θ is the Heaviside function, i.e. $\theta(x) = 1$ for $x > 0$ and $\theta(x) = 0$ for $x < 0$. Since the delta and Heaviside functions are related by $\delta(x) = d\theta(x)/dx$, equation (A3.12.19) becomes

$$\rho(E) = \int \dots \int dq_1 \dots dq_{3n} dp_1 \dots dp_{3n} \delta(E - H) / h^{3n}. \quad (\text{A3.12.20})$$

From [equation \(A3.12.11\)](#), [equation \(A3.12.12\)](#), [equation \(A3.12.13\)](#), [equation \(A3.12.14\)](#) and [equation \(A3.12.15\)](#) and the discussion above, the RRKM rate constant may be written as

$$k(E) = \frac{\int_0^{E-E_0} \left[\int \dots \int dq_2^\ddagger \dots dq_{3n}^\ddagger dp_2^\ddagger \dots dp_{3n}^\ddagger \delta(E' - H) \right] dH}{\int \dots \int dq_1 \dots dq_{3n} dp_1 \dots dp_{3n} \delta(E - H)}. \quad (\text{A3.12.21})$$

The inner multiple integral is the transition state's density of states at energy E' , and also the numerator in [equation \(A3.12.13\)](#), which gives the transition states sum of states $N^\ddagger(E - E_0)$ when integrated from $E' = 0$ to $E' = E - E_0$. Using Hamilton's equation $dH/dp_1 = q_1$, dH in the above equation may be replaced by $q_1 dp_1$. Also, from the definition of the delta function

$$\int \delta(q_1 - q_1^\ddagger) dq_1 = 1. \quad (\text{A3.12.22})$$

This expression may be inserted into the numerator of the above equation, without altering the equation. Making the above two changes and noting that $\delta(q_1 - q_1^\ddagger)$ specifies the transition state, so that the \ddagger superscript to the transition state's coordinates and momenta may be dropped, equation (A3.12.21) becomes

$$k(E) = \frac{\int \dots \int \dot{q}_1 dq_1 \dots dq_{3n} dp_1 \dots dp_{3n} \delta(q_1 - q_1^\ddagger) \delta(E - H)}{\int \dots \int dq_1 \dots dq_{3n} dp_1 \dots dp_{3n} \delta(E - H)}. \quad (\text{A3.12.23})$$

-11-

The rate constant is an average of $\dot{q}_1 \delta(q_1 - q_1^\ddagger)$, with positive \dot{q}_1 , for a microcanonical ensemble $H = E$ and may be expressed as

$$k(E) = \langle \dot{q}_1 \delta(q_1 - q_1^\ddagger) \rangle. \quad (\text{A3.12.24})$$

The RRKM rate constant written this way is seen to be an average flux through the transition state.

A3.12.3.3 VARIATIONAL RRKM THEORY

In deriving the RRKM rate constant in section A3.12.3.1, it is assumed that the rate at which reactant molecules cross the transition state, in the direction of products, is the same rate at which the reactants form products. Thus, if any of the trajectories which cross the transition state in the product direction return to the reactant phase space, i.e. recross the transition state, the actual unimolecular rate constant will be smaller than that predicted by RRKM theory. This one-way crossing of the transition state, with no recrossing, is a fundamental assumption of transition state theory [21]. Because it is incorporated in RRKM theory, this theory is also known as microcanonical transition state theory.

As a result of possible recrossings of the transition state, the classical RRKM $k(E)$ is an upper bound to the correct classical microcanonical rate constant. The transition state should serve as a bottleneck between reactants and products, and in *variational RRKM theory* [22] the position of the transition state along q_1 is varied to minimize $k(E)$. This minimum $k(E)$ is expected to be the closest to the truth. The quantity actually minimized is $N^\ddagger(E - E_0)$ in equation (A3.12.15), so the operational equation in variational RRKM theory is

$$\frac{dN^\ddagger[E - E_0(q_1)]}{dq_1} = 0 \quad (\text{A3.12.25})$$

where $E_0(q_1)$ is the potential energy as a function of q_1 . The minimum in $N^\ddagger[E - E_0(q_1)]$ is identified by $q_1 = q_1^\ddagger$ and this value for q_1 , with the smallest sum of states, is expected to be the best bottleneck for the reaction.

For reactions with well defined potential energy barriers, as in figure A3.12.1(a) and figure A3.12.1(b) the variational criterion places the transition state at or very near this barrier. The variational criterion is particularly important for a reaction where there is no barrier for the reverse association reaction: see figure A3.12.1(c). There are two properties which gave rise to the minimum in $N^\ddagger[E - E_0(q_1)]$ for such a reaction.

As q_1 is decreased the potential energy $E_0(q_1)$ decreases and the energy available to the transition state $E - E_0(q_1)$ increases. This has the effect of increasing the sum of states. However, as q_1 is decreased, the intermolecular anisotropic forces between the dissociating fragments increase, which has the effect of decreasing the available phase space and, thus, the sum of states. The combination of these two effects gives rise to a minimum in $N^\ddagger[E - E_0(q_1)]$. Plots of the sum of states versus q_1 are shown in [figure A3.12.4](#) for three model potentials of the $C_2H_6 \rightarrow 2CH_3$ dissociation reaction [23].

-12-

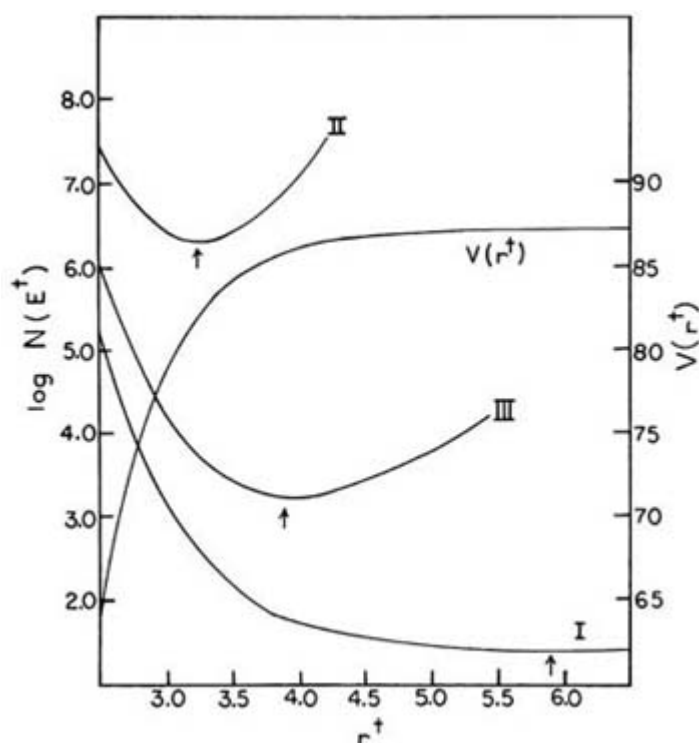


Figure A3.12.4. Plots of $N^\ddagger[E - E_0(q_1)]$ versus q_1 for three models of the $C_2H_6 \rightarrow 2CH_3$ potential energy function. r^\ddagger represents q_1 and the term on the abscissa represents N^\ddagger . (Adapted from [23].)

Variational RRKM theory is particularly important for unimolecular dissociation reactions, in which vibrational modes of the reactant molecule become translations and rotations in the products [22]. For $CH_4 \rightarrow CH_3 + H$ dissociation there are three vibrational modes of this type, i.e. the C---H stretch which is the reaction coordinate and the two degenerate H---CH₃ bends, which first transform from high-frequency to low-frequency vibrations and then hindered rotors as the H---C bond ruptures. These latter two degrees of freedom are called transitional modes [24,25]. $C_2H_6 \rightarrow 2CH_3$ dissociation has five transitional modes, i.e. two pairs of degenerate CH₃ rocking/rotational motions and the CH₃ torsion.

To calculate $N^\ddagger(E - E_0)$, the non-torsional transitional modes have been treated as vibrations as well as rotations [26]. The former approach is invalid when the transitional mode's barrier for rotation is low, while the latter is inappropriate when the transitional mode is a vibration. Harmonic frequencies for the transitional modes may be obtained from a semi-empirical model [23] or by performing an appropriate normal mode analysis as a function of the reaction path for the reaction's potential energy surface [26]. Semiclassical quantization may be used to determine anharmonic energy levels for the transitional modes [27].

The intermolecular Hamiltonian of the product fragments is used to calculate the sum of states of the transitional modes, when they are treated as rotations. The resulting model [28] is nearly identical to phase space theory [29].

if the distance between the product fragments' centres-of-mass is assumed to be the reaction coordinate [30]. A more complete model is obtained by using a generalized reaction coordinate, which may contain contributions from different motions, such as bond stretching and bending, as well as the above relative centre-of-mass motion [31].

Variational RRKM calculations, as described above, show that a unimolecular dissociation reaction may have two variational transition states [32, 33, 34, 35 and 36], i.e. one that is a tight vibrator type and another that is a loose rotator type. Whether a particular reaction has both of these variational transition states, at a particular energy, depends on the properties of the reaction's potential energy surface [33, 34 and 35]. For many dissociation reactions there is only one variational transition state, which smoothly changes from a loose rotator type to a tight vibrator type as the energy is increased [26].

A3.12.4 APPROXIMATE MODELS FOR THE RRKM RATE CONSTANT

A3.12.4.1 CLASSICAL HARMONIC OSCILLATORS: RRK THEORY

The classical mechanical RRKM $k(E)$ takes a very simple form, if the internal degrees of freedom for the reactant and transition state are assumed to be harmonic oscillators. The classical sum of states for s harmonic oscillators is [16]

$$N(E) = \frac{E^s}{s! \prod_{i=1}^s h \nu_i}. \quad (\text{A3.12.26})$$

The density $\rho(E) = dN(E)/dE$ is then

$$\rho(E) = \frac{E^{s-1}}{(s-1)! \prod_{i=1}^s h \nu_i}. \quad (\text{A3.12.27})$$

The reactant density of states in [equation \(A3.12.15\)](#) is given by the above expression for $\rho(E)$. The transition state's sum of states is

$$N^\ddagger(E - E_0) = \frac{(E - E_0)^{s-1}}{(s-1)! \prod_{i=1}^{s-1} h \nu_i^\ddagger}. \quad (\text{A3.12.28})$$

Inserting [equation \(A3.12.27\)](#) and [equation \(A3.12.28\)](#) into [equation \(A3.12.15\)](#) gives

$$k(E) = \frac{\prod_{i=1}^s \nu_i}{\prod_{i=1}^{s-1} \nu_i^\ddagger} \left(\frac{E - E_0}{E} \right)^{s-1}. \quad (\text{A3.12.29})$$

If the ratio of the products of vibrational frequencies is replaced by v , [equation \(A3.12.29\)](#) becomes

$$k(E) = v \left(\frac{E - E_0}{E} \right)^{s-1} . \quad (\text{A3.12.30})$$

which is the Rice–Ramsperger–Kassel (RRK) unimolecular rate constant [16,37]. Thus, the $k(E)$ of RRK theory is the classical harmonic limit of RRKM theory.

A3.12.4.2 QUANTUM HARMONIC OSCILLATORS

Only in the high-energy limit does classical statistical mechanics give accurate values for the sum and density of states terms in [equation \(A3.12.15\)](#) [3,14]. Thus, to determine an accurate RRKM $k(E)$ for the general case, quantum statistical mechanics must be used. Since it is difficult to make anharmonic corrections, both the molecule and transition state are often assumed to be a collection of harmonic oscillators for calculating the sum $N^\ddagger(E - E_0)$ and density $\rho(E)$. This is somewhat incongruous since a molecule consisting of harmonic oscillators would exhibit intrinsic non-RRKM dynamics.

With the assumption of harmonic oscillators, the molecule's quantum energy levels are

$$E(\mathbf{n}) = \sum_{i=1}^s n_i h\nu_i . \quad (\text{A3.12.31})$$

The same expression holds for the transition state, except that the sum is over $s - 1$ oscillators and the frequencies are the ν_i^\ddagger . The Beyer–Swinehart algorithm [38] makes a very efficient direct count of the number of quantum states between an energy of zero and E . The molecule's density of states is then found by finite difference, i.e.

$$\rho(E) = \frac{N(E + \Delta E/2) - N(E - \Delta E/2)}{\Delta E} \quad (\text{A3.12.32})$$

where $N(E + \Delta E/2)$ is the sum of states at energy $E + \Delta E/2$. The transition state's harmonic $N^\ddagger(E - E_0)$ is counted directly by the Beyer–Swinehart algorithm. This harmonic model is used so extensively to calculate a value for the RRKM $k(E)$ that it is easy to forget that RRKM theory is not a harmonic theory.

A3.12.4.3 OVERALL ROTATION

Regardless of the nature of the intramolecular dynamics of the reactant A^* , there are two constants of the motion in a unimolecular reaction, i.e. the energy E and the total angular momentum j . The latter ensures the rotational quantum number J is fixed during the unimolecular reaction and the quantum RRKM rate constant is specified as $k(E, J)$.

(A) SEPARABLE VIBRATION/ROTATION

For a RRKM calculation without any approximations, the complete vibrational/rotational Hamiltonian for the unimolecular system is used to calculate the reactant density and transition state's sum of states. No approximations are made regarding the coupling between vibration and rotation. However, for many molecules the exact nature of the coupling between vibration and rotation is uncertain, particularly at high energies, and a model in which rotation and vibration are assumed separable is widely used to calculate the quantum RRKM $k(E, J)$ [4,16]. To illustrate this model, first consider a linear polyatomic molecule which decomposes via a linear transition state. The rotational energy for the reactant is assumed to be that for a rigid rotor, i.e.

$$E_r = J(J + 1)\hbar^2/2I. \quad (\text{A3.12.33})$$

The same expression applies to the transition state's rotational energy $E_r^\ddagger(J)$ except that the moment of inertia I is replaced by I^\ddagger . Since the quantum number j is fixed, the active energies for the reactant and transition state are $[E - E_r(J)]$ and $[E - E_0 - E_r^\ddagger(J)]$, respectively. The RRKM rate constant is denoted by

$$k(E, J) = \frac{N[E - E_0 - E_r^\ddagger(J)]}{h\rho[E - E_r(J)]} \quad (\text{A3.12.34})$$

where N^\ddagger and ρ are the sum and density of states for the vibrational degrees of freedom. Each j level is $(2J + 1)^2$ degenerate, which cancels for the sum and density.

(B) THE K QUANTUM NUMBER: ADIABATIC OR ACTIVE

The degree of freedom in [equation \(A3.12.18\)](#), which has received considerable interest regarding its activity or adiabaticity, is the one associated with the K rotational quantum number for a symmetric or near-symmetric top molecule [39,40]. The quantum number K represents the projection of J onto the molecular symmetry axis. Coriolis coupling can mix the $2J + 1$ K levels for a particular J and destroy K as a good quantum number. For this situation K is considered an active degree of freedom. On the other hand, if the Coriolis coupling is weak, the K quantum number may retain its integrity and it may be possible to measure the unimolecular rate constant as a function of K as well as of E and J . For this case, K is an adiabatic degree of freedom.

It is straightforward to introduce active and adiabatic treatments of K into the widely used RRKM model which represents vibration and rotation as separable and the rotations as rigid rotors [41,42]. For a symmetric top, the rotational energy is given by

$$E_r(J, K) = \frac{J(J + 1)\hbar^2}{2I_a} + \left(\frac{1}{I_c} - \frac{1}{I_a} \right) K^2\hbar^2. \quad (\text{A3.12.35})$$

If K is adiabatic, a molecule containing total vibrational-rotational energy E and, in a particular J, K level, has a vibrational density of states $\rho[E - E_r(J, K)]$. Similarly, the transition state's sum of states for the same E, J , and K is $N^\ddagger[E - E_0 - E_r^\ddagger(J, K)]$. The RRKM rate constant for the K adiabatic model is

$$k(E, J, K) = \frac{N^\ddagger[E - E_0 - E_t^\ddagger(J, K)]}{h\rho[E - E_r(J, K)]}. \quad (\text{A3.12.36})$$

Mixing the $2J + 1$ K levels, for the K active model, results in the following sums and densities of states:

$$N^\ddagger(E, J) = \sum_{K=-J}^J N^\ddagger[E - E_0 - E_t^\ddagger(J, K)] \quad (\text{A3.12.37})$$

$$\rho(E, J) = \sum_{K=-J}^J \rho[E - E_r(J, K)]. \quad (\text{A3.12.38})$$

The RRKM rate constant for the K active model is

$$k(E, J) = \frac{\sum_{K=-J}^J N^\ddagger[E - E_0 - E_t^\ddagger(J, K)]}{h \sum_{K=-J}^J \rho[E - E_r(J, K)]}. \quad (\text{A3.12.39})$$

In these models the treatment of K is the same for the molecule and transition state. It is worthwhile noting that *mixed mode RRKM models* are possible in which K is treated differently in the molecule and transition state [39].

A3.12.5 ANHARMONIC EFFECTS

In the above section a harmonic model is described for calculating RRKM rate constants with harmonic sums and densities of states. This rate constant, denoted by $k_h(E, J)$, is related to the actual anharmonic RRKM rate constant by

$$k(E, J) = f_{\text{anh}}(E, J)k_h(E, J) = f_{\text{anh}}(E, J) \frac{N_h^\ddagger(E, J)}{h\rho_h(E, J)} \quad (\text{A3.12.40})$$

where $NN_h^\ddagger(E, J)$ and $\rho_h(E, J)$ are the harmonic approximations to the sum and density of states. The anharmonic correction, $f_{\text{anh}}(E, J)$, is obviously

$$f_{\text{anh}}(E, J) = \frac{N_{\text{anh}}^\ddagger(E, J)/N_h^\ddagger(E)}{\rho_{\text{anh}}(E, J)/\rho_h(E, J)} = \frac{f_{\text{anh}, N^\ddagger}(E, J)}{f_{\text{anh}, \rho}(E, J)} \quad (\text{A3.12.41})$$

the ratio of anharmonic corrections for the sum and density. For energies near the unimolecular threshold, where the transition state energy $E - E_0$ is small, anharmonicity in the transition state may be negligible, so

that $f_{\text{anh}}(E)$ may be well approximated by $1/f_{\text{anh},\rho(E)}$ [43]. However, for higher energies, anharmonicity is expected to become important also for the transition state.

There is limited information available concerning anharmonic corrections to RRKM rate constants. Only a few experimental studies have investigated the effect of anharmonicity on the reactant's density of states (see below). To do so requires spectroscopic information up to very high energies. It is even more difficult to measure anharmonicities for transition state energy levels [44,45]. If the potential energy surface is known for a unimolecular reactant, anharmonic energy levels for both the reactant and transition state may be determined in principle from large-scale quantum mechanical variational calculations. Such calculations are more feasible for the transition state with energy $E - E_0$ than for the reactant with the much larger energy E . Such calculations have been limited to relatively small molecules [4].

The bulk of the information about anharmonicity has come from classical mechanical calculations. As described above, the anharmonic RRKM rate constant for an analytic potential energy function may be determined from either equation (A3.12.4) [13] or equation (A3.12.24) [46] by sampling a microcanonical ensemble. This rate constant and the one calculated from the harmonic frequencies for the analytic potential give the anharmonic correction $f_{\text{anh}}(E, J)$ in equation (A3.12.41). The transition state's anharmonic classical sum of states is found from the phase space integral

$$N_{\text{anh}}^{\ddagger}(E, J) = \int \dots \int dq_2 \dots dq_{3n} dp_2 \dots dp_{3n} \theta(E - H) / h^{3n-1} \quad (\text{A3.12.42})$$

which may be combined with the harmonic sum $NN_{\text{h}}^{\ddagger}(E, J)$ to give $f_{\text{anh},N_{\text{h}}^{\ddagger}}(E, J)$. The classical anharmonic correction to the reactant's density of states, $f_{\text{anh},\rho}(E, J)$, may be obtained in a similar manner.

A3.12.5.1 MOLECULES WITH A SINGLE MINIMUM

Extensive applications of RRKM theory have been made to unimolecular reactions, for which there is a single potential energy minimum for the reactant molecule [4,47]. For such reactions, the anharmonic correction $f_{\text{anh}}(E, J)$ is usually assumed to be unity and the harmonic model is used to calculate the RRKM $k(E, J)$. Though this is a widely used approach, uncertainties still remain concerning its accuracy. Anharmonic densities of states for formaldehyde [48] and acetylene [49] obtained from high-resolution spectroscopic experiments at energies near their unimolecular thresholds, are 11 and 6 times larger, respectively, than their harmonic densities. From calculations with analytic potential energy functions at energies near the unimolecular thresholds, the HCN quantum anharmonic density of states is 8 times larger than the harmonic value [50] and the classical anharmonic density of states for the model alkyl radical HCC is 3–5 times larger than the harmonic value [51]. There is a sense that the anharmonic correction may become less important for larger molecules, since the average energy per mode becomes smaller [4]. However, as shown below, this assumption is clearly not valid for large fluxional molecules with multiple minima.

Analytic expressions have been proposed for making anharmonic corrections for molecules with a single potential minimum [52,53,54,55 and 56]. Haarhoff [52] derived an expression for this correction factor by describing the molecules' degrees of freedom as a collection of Morse oscillators. One of the limitations of this model is that it is difficult to assign Morse parameters to non-stretching degrees of freedom. Following the spirit of Haarhoff's work, Troe [54] formulated the correction factor

$$f_{\text{anh},\rho}(E) = \prod_{i=1}^m \left(1 + \frac{E/D_i}{2s-3} \right) \quad (\text{A3.12.43})$$

for a molecule with s degrees of freedom, m of which are Morse stretches. The remaining $s - m$ degrees of freedom are assumed to be harmonic oscillators. The D_i are the individual Morse dissociation energies. To account for bend–stretch coupling, i.e. the attenuation of bending forces as bonds are stretched [51], Troe amended equation (A3.12.43) to give [55]

$$f_{\text{anh},\rho}(E) = \sum_{i=1}^m \left(1 + \frac{E/D_i}{2s-3} \right)^2. \quad (\text{A3.12.44})$$

The above expressions are empirical approaches, with m and D_i as parameters, for including an anharmonic correction in the RRKM rate constant. The utility of these equations is that they provide an analytic form for the anharmonic correction. Clearly, other analytic forms are possible and may be more appropriate. For example, classical sums of states for H–C–C, H–C=C, and H–C≡C hydrocarbon fragments with Morse stretching and bend–stretch coupling anharmonicity [51] are fit accurately by the exponential

$$f_{\text{anh},N}(E) = \exp(bE). \quad (\text{A3.12.45})$$

The classical anharmonic density of states is then [56]

$$f_{\text{anh},\rho}(E) = \exp(bE)[1 + bE/s]. \quad (\text{A3.12.46})$$

Modifying equation (A3.12.45) to represent the transition state's sum of states, the anharmonic correction to the RRKM rate constant becomes

$$f_{\text{anh}}(E) = \frac{\exp[b^*(E - E_0)]}{\exp(bE)[1 + bE/s]}. \quad (\text{A3.12.47})$$

This expression, and variations of it, have been used to fit classical anharmonic microcanonical $k(E, J)$ for unimolecular decomposition [56].

A3.12.5.2 FLUXIONAL MOLECULES WITH MULTIPLE MINIMA

Anharmonic corrections are expected to be very important for highly fluxional molecules such as clusters and macromolecules [30]. Figure A3.12.5 illustrates a possible potential energy curve for a fluxional molecule. There are multiple minima (i.e. conformations) separated by barriers much lower than that for dissociation. Thus, a moderately excited fluxional molecule may undergo rapid transitions between its many conformations, and all will contribute to the molecule's unimolecular rate constant. Many different conformations are expected for the products, but near the dissociation threshold E_0 , only one set of product conformations is accessible. As the energy is increased, thresholds for other product conformations are reached. For energies near E_0 , there is very little excess energy in the transition state and the harmonic approximation for the lowest energy product conformation should be very good for the transition state's sum

of states. Thus, for $E \approx E_0$ the anharmonic correction in [equation \(A3.12.40\)](#) is expected to primarily result from anharmonicity in the reactant density of states.

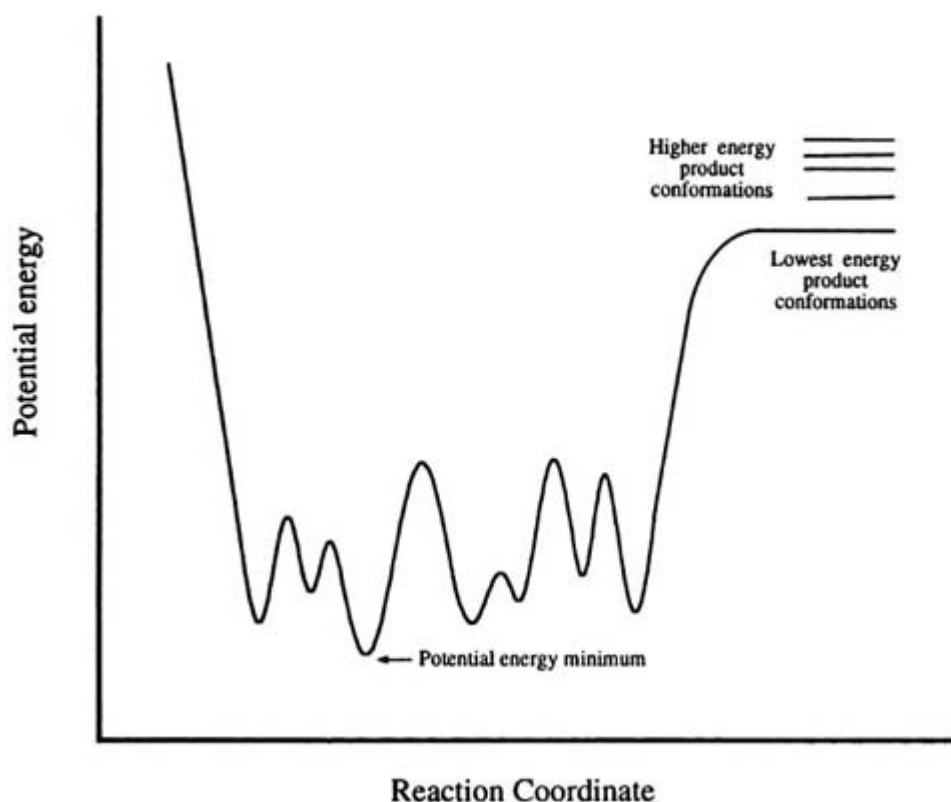


Figure A3.12.5. A model reaction coordinate potential energy curve for a fluxional molecule. (Adapted from [30].)

The classical anharmonic RRKM rate constant for a fluxional molecule may be calculated from classical trajectories by following the initial decay of a microcanonical ensemble of states for the unimolecular reactant, as given by [equation \(A3.12.4\)](#). Such a calculation has been performed for dissociation of the Al_6 and Al_{13} clusters using a model analytic potential energy function written as a sum of Lennard–Jones and Axelrod–Teller potentials [30]. Structures of some of the Al_6 minima, for the potential function, are shown in [figure A3.12.6](#). The deepest potential minimum has

C_{2h} symmetry and a classical $\text{Al}_6 \rightarrow \text{Al}_5 + \text{Al}$ dissociation energy E_0 of $43.8 \text{ kcal mol}^{-1}$. For energies 30–80 kcal mol^{-1} in excess of this E_0 , the value of f_{anh} determined from the trajectories varies from 200 to 130. The harmonic RRKM rate constants are based on the deepest potential energy minima for the reactant and transition state, and calculated for a reaction path degeneracy of 6. As discussed above, even larger corrections are expected at lower energies, particularly for $E \approx E_0$, where anharmonicity in the transition state does not contribute $B_{\text{anh}}(E)$. However, because of the size of Al_6 and its long unimolecular lifetime, it becomes impractical to simulate the classical dissociation of Al_6 for energies in excess of E_0 much smaller than 30 kcal mol^{-1} . For the bigger cluster Al_{13} , the anharmonic correction varies from 5500 to 1200 for excess energies in the range of 85–185 kcal mol^{-1} [30]. These calculations illustrate the critical importance of including anharmonic corrections when calculating accurate RRKM rate constants for fluxional molecules.

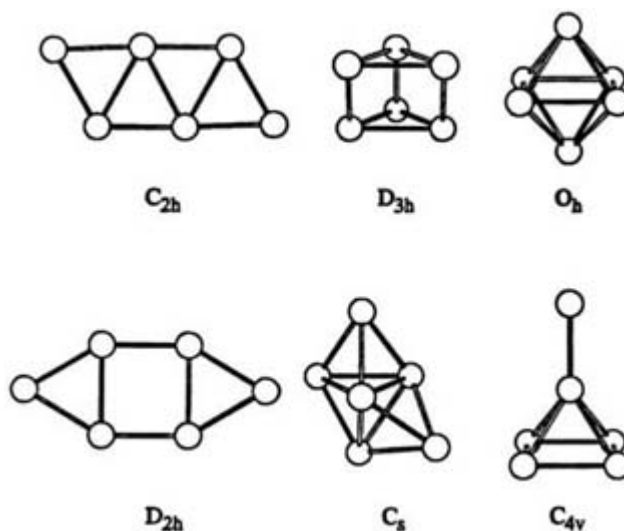


Figure A3.12.6. Structures for some of the potential energy minima for Al_6 . The unimolecular thresholds for the C_{2h} , D_{3h} , C_s , O_h , C_{4v} , and D_h minima are 43.8, 40.0, 39.6, 38.8, 31.4 and 20.9 kcal mol $^{-1}$, respectively. (Adapted from [40].)

In the above discussion it was assumed that the barriers are low for transitions between the different conformations of the fluxional molecule, as depicted in [figure A3.12.5](#) and therefore the transitions occur on a timescale much shorter than the RRKM lifetime. This is the rapid IVR assumption of RRKM theory discussed in [section A3.12.2](#). Accordingly, an initial microcanonical ensemble over all the conformations decays exponentially. However, for some fluxional molecules, transitions between the different conformations may be slower than the RRKM rate, giving rise to bottlenecks in the unimolecular dissociation [4,57]. The ensuing lifetime distribution, [equation \(A3.12.7\)](#), will be non-exponential, as is the case for *intrinsic* non-RRKM dynamics, for an initial microcanonical ensemble of molecular states.

A3.12.6 CLASSICAL DYNAMICS OF INTRAMOLECULAR MOTION AND UNIMOLECULAR DECOMPOSITION

A3.12.6.1 NORMAL-MODE HAMILTONIAN: SLATER THEORY

The classical mechanical model of unimolecular decomposition, developed by Slater [18], is based on the normal-mode harmonic oscillator Hamiltonian. Although this Hamiltonian is rigorously exact only for small displacements from the molecular equilibrium geometry, Slater extended it to the situation where molecules are highly vibrationally energized, undergo large amplitude motions and decompose. Since there are no couplings in the normal-mode Hamiltonian, the energies in the individual normal modes do not vary with time. This is the essential difference from the RRKM theory which treats a molecule as a collection of coupled modes which freely exchange energy.

The normal-mode harmonic oscillator classical Hamiltonian is

$$H = \sum_{i=1}^s \frac{(P_i^2 + \lambda_i Q_i^2)}{2} \quad (\text{A3.12.48})$$

where $\lambda_i = 4\pi^2\nu_i^2$. Solving the classical equations of motion for this Hamiltonian gives rise to quasiperiodic motion [58] in which each normal-mode coordinate Q_i varies with time according to

$$Q_i = Q_i^0 \cos(2\pi\nu_i t + \delta_i) \quad (\text{A3.12.49})$$

where Q_i^0 is the amplitude and δ_i the phase of the motion. Thus, if an energy $E_i = (p_i^2 + \lambda_i Q_i^2)/2$ and phase δ_i are chosen for each normal mode, the complete intramolecular motion of the energized molecule may be determined for this particular initial condition.

Reaction is assumed to have occurred if a particular internal coordinate q , such as a bond length, attains a critical extension q^\ddagger . In the normal-mode approximation, the displacement d of internal coordinates and normal-mode coordinates Q are related through the linear transformation

$$d = LQ. \quad (\text{A3.12.50})$$

The transformation matrix L is obtained from a normal-mode analysis performed in internal coordinates [59,60]. Thus, as the evolution of the normal-mode coordinates versus time is evaluated from equation (A3.12.49), displacements in the internal coordinates and a value for q are found from equation (A3.12.50). The variation in q with time results from a superposition of the normal modes. At a particular time, the normal-mode coordinates may phase together so that q exceeds the critical extension q^\ddagger , at which point decomposition is assumed to occur.

-22-

The preceding discussion gives the essential details of the Slater theory. Energy does not flow freely within the molecule and attaining the critical reaction coordinate extension is not a statistically random process as in RRKM theory, but depends on the energies and phases of the specific normal modes excited. If a microcanonical ensemble is chosen at $t = 0$, Slater theory gives an initial decay rate which agrees with the RRKM value. However, Slater theory gives rise to intrinsic non-RRKM behaviour [12,13]. The trajectories are quasiperiodic and each trajectory is restricted to a particular type of motion and region of phase space. Thus, as specific trajectories react, other trajectories cannot fill up unoccupied regions of phase space. As a result, a microcanonical ensemble is not maintained during the unimolecular decomposition. In addition, some of the trajectories may be unreactive and trapped in the reactant region of phase space.

Overall, the Slater theory is unsuccessful in interpreting experiments. Many unimolecular rate constants and reaction paths are consistent with energy flowing randomly within the molecule [4,36]. If one considers the nature of classical Hamiltonians for actual molecules, it is not surprising that the Slater theory performs so poorly. For example, in Slater theory, the intramolecular and unimolecular dynamics of the molecule conform to the symmetry of the molecular vibrations. Thus, if normal modes of a particular symmetry type are excited (e.g. in-plane vibrations) a decomposition path of another symmetry type (e.g. out-of-plane dissociation) cannot occur. This path requires excitation of out-of-plane vibrations. Normal modes of different symmetry types for actual molecules are coupled by *Coriolis* vibrational-rotational interactions [61]. Similarly, nonlinear resonance interactions couple normal modes of vibration, allowing transfer of energy [62,63]. Not including these effects is a severe shortcoming of Slater theory. Clearly, understanding the classical intramolecular motion of vibrationally excited molecules requires one to go beyond the normal-mode model.

A3.12.6.2 COUPLED ANHARMONIC HAMILTONIANS

The first classical trajectory study of unimolecular decomposition and intramolecular motion for realistic anharmonic molecular Hamiltonians was performed by Bunker [12,13]. Both intrinsic RRKM and non-RRKM dynamics was observed in these studies. Since this pioneering work, there have been numerous additional studies [9,17,30,64,65,66 and 67] from which two distinct types of intramolecular motion, chaotic and quasiperiodic [14], have been identified. Both are depicted in figure A3.12.7. Chaotic vibrational motion is not regular as predicted by the normal-mode model and, instead, there is energy transfer between the modes. If all the modes of the molecule participate in the chaotic motion and energy flow is sufficiently rapid, an initial microcanonical ensemble is maintained as the molecule dissociates and RRKM behaviour is observed [9]. For non-random excitation initial apparent non-RRKM behaviour is observed, but at longer times a microcanonical ensemble of states is formed and the probability of decomposition becomes that of RRKM theory.

-23-

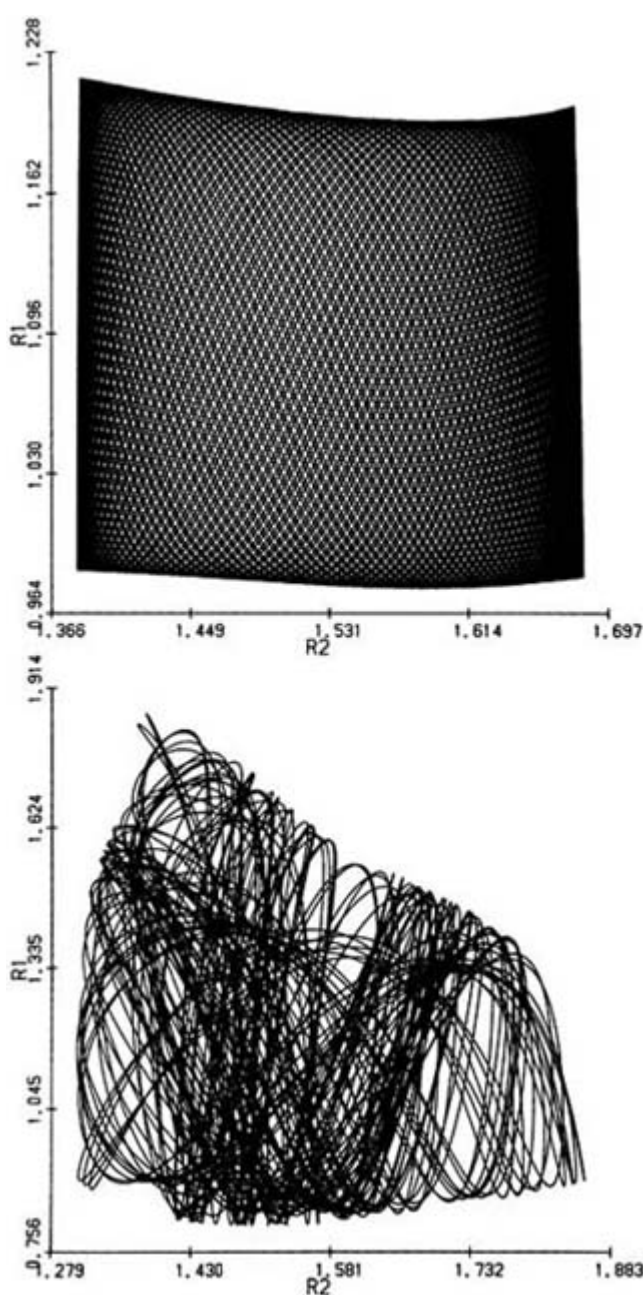


Figure A3.12.7. Two trajectories for a model HCC Hamiltonian. Top trajectory is for $n_{\text{HC}}=0$ and $n_{\text{CC}}=2$, and

is quasiperiodic. Bottom trajectory is for $n_{\text{HC}}=5$ and $n_{\text{CC}}=0$, and is chaotic. R_1 is the HC bond length and R_2 the CC bond length. Distance is in Å ngstroms (Å). (Adapted from [121] and [4].)

-24-

Quasiperiodic motion is regular as assumed by the Slater theory. The molecule's vibrational motion may be represented by a superposition of individual modes, each containing a fixed amount of energy. For some cases these modes resemble normal modes, but they may be identifiably different. Thus, although actual molecular Hamiltonians contain potential and kinetic energy coupling terms, they may still exhibit regular vibrational motion with no energy flow between modes as predicted by the Slater theory. The existence of regular motion for coupled systems is explained by the Kolmogorov–Arnold–Moser (KAM) theorem [4,68].

Extensive work has been done to understand quasiperiodic and chaotic vibrational motion of molecular Hamiltonians [14,58,63,68,69]. At low levels of excitation, quasiperiodic normal mode motion is observed. However, as the energy is increased, nonlinear resonances [14,62,63,68] result in the flow of energy between the normal modes, giving rise to chaotic trajectories. In general, the fraction of the trajectories, which are chaotic at a fixed energy, increases as the energy is increased. With increase in energy the nature of the quasiperiodic trajectories may undergo a transition from normal mode to another type of motion, e.g. local mode [70,71]. In many cases the motion becomes totally chaotic before the unimolecular threshold energy is reached, so that the intramolecular dynamics is ergodic. Though this implies intrinsic RRKM behaviour for energies above the threshold, the ergodicity must occur on a timescale much shorter than the RRKM lifetime for a system to be intrinsically RRKM.

For some systems quasiperiodic (or nearly quasiperiodic) motion exists above the unimolecular threshold, and intrinsic non-RRKM lifetime distributions result. This type of behaviour has been found for Hamiltonians with low unimolecular thresholds, widely separated frequencies and/or disparate masses [12,62,65]. Thus, classical trajectory simulations performed for realistic Hamiltonians predict that, for some molecules, the unimolecular rate constant may be strongly sensitive to the modes excited in the molecule, in agreement with the Slater theory. This property is called *mode specificity* and is discussed in the next section.

It is of interest to consider the classical/quantal correspondence for the above different types of classical motion. If the motion within the classical phase space is ergodic so that the decomposing molecules can be described by a microcanonical ensemble, classical RRKM theory will be valid. However, the classical and quantal RRKM rate constants may be in considerable disagreement. This results from an incorrect treatment of zero-point energy in the classical calculations [17,72] and is the reason quantum statistical mechanics is needed to calculate an accurate RRKM rate constant: see the discussion following equation (A3.12.15). With the energy referenced at the bottom of the well, the total internal energy of the dissociating molecule is $E = E^* + E_z^*$ where E^* is the internal energy of the molecule and E_z^* is its zero-point energy. The classical dissociation energy is D_e , and the energy available to the dissociating molecule at the classical barrier is $E - D_e$. Because the quantal threshold is $D_e + E_z^\ddagger$ where E_z^\ddagger is the zero-point energy at the barrier, the classical threshold is lower than the quantal one by E_z^\ddagger . For large molecules with a large E_z^\ddagger and/or for low levels of excitation the classical RRKM rate constant is significantly larger than the quantal one. Only in the high-energy limit are they equal; see figure A3.12.3.

Quasiperiodic trajectories, with an energy greater than the unimolecular threshold, are trapped in the reactant region of phase space and will not dissociate. These trajectories correspond to quantum mechanical compound-state resonances $|n\rangle$ (discussed in the next section), which have complex eigenvalues. Applying semiclassical mechanics to the trajectories [73, 74, 75 and 76] gives energies E_n , wavefunctions ψ_n , and unimolecular rate constants k_n for these resonances. A classical microcanonical ensemble for an energized

molecule may consist of quasiperiodic, chaotic, and ‘vague tori’ trajectories [77]. The lifetimes of trajectories for the latter may yield correct quantum k_n for resonance states[4].

A3.12.7 STATE-SPECIFIC UNIMOLECULAR DECOMPOSITION

The quantum dynamics of unimolecular decomposition may be studied by solving the time-dependent Schrödinger equation, i.e. equation (A3.12.2). For some cases the dissociation probability of the molecule is sufficiently small that one can introduce the concept of quasi-stationary states. Such states are commonly referred to as resonances, since the energy of the unimolecular product(s) in the continuum is in resonance with (i.e. matches) the energy of a vibrational/rotational level of the unimolecular reactant. For unimolecular reactions there are two types of resonance states. A shape resonance occurs when a molecule is temporarily trapped by a fairly high and wide potential energy barrier and decomposes by tunnelling. The second type of resonance, called a Feshbach or compound-state resonance, arises when energy is initially distributed between molecular vibrational/rotational degrees of freedom which are not strongly coupled to the decomposition reaction coordinate motion, so that there is a time lag for unimolecular dissociation.

In a time-dependent picture, resonances can be viewed as localized wavepackets composed of a superposition of continuum wavefunctions, which qualitatively resemble bound states for a period of time. The unimolecular reactant in a resonance state moves within the potential energy well for a considerable period of time, leaving it only when a fairly long time interval τ has elapsed; τ may be called the lifetime of the almost stationary resonance state.

Solving the time-dependent Schrödinger equation for resonance states [78] one obtains a set of complex eigenvalues, which may be written in the form

$$E_n^0 = E_n - i\Gamma_n/2 \quad (\text{A3.12.51})$$

where E_n and Γ_n are positive constants. The constant E_n , the real component to the eigenvalue, gives the position of the resonance in the spectrum. It is easy to see the physical significance of complex energy values. The time factor in the wavefunction of a quasi-stationary state is of the form

$$\exp[-(i/\hbar)E_n^0 t] = \exp[-(i/\hbar)E_n t] \exp[-(\Gamma_n/2\hbar)t]. \quad (\text{A3.12.52})$$

Hence, all probabilities given by the squared modulus of the wavefunction decrease as $\exp[-(\Gamma_n/\hbar)t]$ with time, that is

$$|\psi_n(t)|^2 = |\psi_n(0)|^2 \exp[-(\Gamma_n/\hbar)t]. \quad (\text{A3.12.53})$$

In particular, the probability of finding the unimolecular reactant within its potential energy well decreases according to this law. Thus Γ_n determines the lifetime of the state and the state specific unimolecular rate constant is

$$k_n = \Gamma_n/\hbar = 1/\tau_n \quad (\text{A3.12.54})$$

where τ_n is the state's lifetime.

-26-

The energy spectrum of the resonance states will be quasi-discrete; it consists of a series of broadened levels with Lorentzian lineshapes whose full-width at half-maximum Γ is related to the lifetime by $\Gamma = \hbar\tau$. The resonances are said to be isolated if the widths of their levels are small compared with the distances (spacings) between them, that is

$$\Gamma_n \ll 1/\rho(E) \quad (\text{A3.12.55})$$

where $\rho(E)$ is the density of states for the energized molecule. A possible absorption spectrum for a molecule with isolated resonances is shown in figure A3.12.8. Below the unimolecular threshold E_0 , the absorption lines for the molecular eigenstates are very narrow and are only broadened by interaction of the excited molecule with the radiation field. However, above E_0 the excited states leak toward product space, which gives rise to widths for the resonances in the spectrum. Each resonance has its own characteristic width (i.e. lifetime). As the linewidths broaden and/or the number of resonance states in an energy interval increases, the spectrum may no longer be quasi-discrete since the resonance lines may overlap, that is

$$\Gamma_n \gg 1/\rho(E). \quad (\text{A3.12.56})$$

It is of interest to determine when the linewidth $\Gamma(E)$ associated with the RRKM rate constant $k(E)$ equals the average distance $\rho(E)^{-1}$ between the reactant energy levels. From equation (A3.12.54) $\Gamma(E) = \hbar k(E)$ and from the RRKM rate constant expression equation (A3.12.15) $\rho(E)^{-1} = \hbar 2\pi K(E)/N^\ddagger(E - E_0)$. Equating these two terms gives $N^\ddagger(E - E_0) = 2\pi$, which means that the linewidths, associated with RRKM decomposition, begin to overlap when the transition state's sum of states exceeds six.

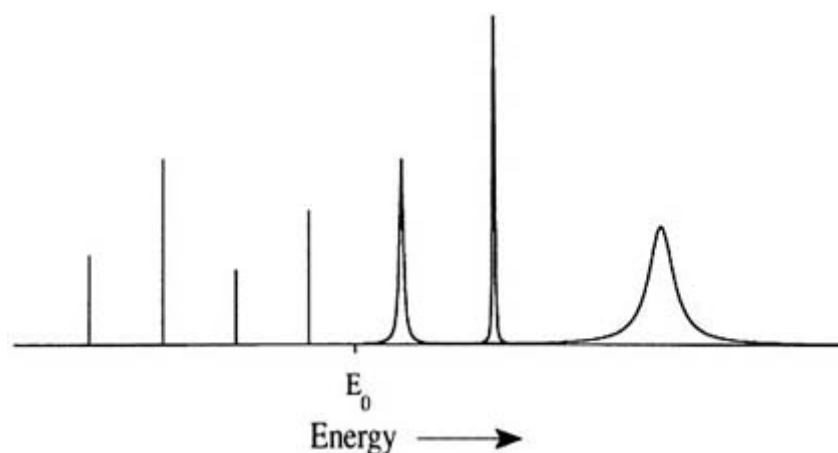


Figure A3.12.8. Possible absorption spectrum for a molecule which dissociates via isolated compound-state resonances. E_0 is the unimolecular threshold. (Adapted from [4].)

The theory of isolated resonances is well understood and is discussed below. Mies and Krauss [79,80] and Rice [81] were pioneers in treating unimolecular rate theory in terms of the decomposition of isolated Feshbach resonances.

A3.12.7.1 ISOLATED REACTANT RESONANCE STATES MODE SPECIFICITY

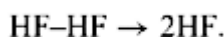
The observation of decomposition from isolated compound-state resonances does not necessarily imply mode-specific unimolecular decomposition. Nor is mode specificity established by the presence of fluctuations in state-specific rate constants for resonances within a narrow energy interval. What is required for mode-specific unimolecular decomposition is a distinguishable and, thus, assignable pattern (or patterns) in the positions of resonance states in the spectrum. Identifying such patterns in a spectrum allows one to determine which modes in the molecule are excited when forming the resonance state. It is, thus, possible to interpret particularly large or small state-specific rate constants in terms of mode-specific excitations. Therefore, mode specificity means there are exceptionally large or small state-specific rate constants depending on which modes are excited.

The ability to assign a group of resonance states, as required for mode-specific decomposition, implies that the complete Hamiltonian for these states is well approximated by a zero-order Hamiltonian with eigenfunctions $\phi_i(m)$ [58]. The ϕ_i are product functions of a zero-order orthogonal basis for the reactant molecule and the quantity m represents the quantum numbers defining ϕ_i . The wavefunctions ψ_n for the compound state resonances are given by

$$\psi_n = \sum_i c_{in} \phi_i(m). \quad (\text{A3.12.57})$$

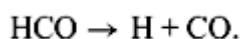
Resonance states in the spectra, which are assignable in terms of zero-order basis $\phi_i(m)$, will have a predominant expansion coefficient c_{in} . Hose and Taylor [58] have argued that for an assignable level $c_{in}^2 > 0.5$ for one of the expansion coefficients. The quasiperiodic and ‘vague tori’ trajectories for energies above the unimolecular threshold, discussed in the previous section, are the classical analogue of these quantum mode specific resonance states.

Mode specificity has been widely observed in the unimolecular decomposition of van der Waals molecules [82], e.g.



A covalent bond (or particular normal mode) in the van der Waals molecule (e.g. the I_2 bond in $\text{I}_2\text{-He}$) can be selectively excited, and what is usually observed experimentally is that the unimolecular dissociation rate constant is orders of magnitude smaller than the RRKM prediction. This is thought to result from weak coupling between the excited high-frequency intramolecular mode and the low-frequency van der Waals intermolecular modes [83]. This coupling may be highly mode specific. Exciting the two different HF stretch modes in the $(\text{HF})_2$ dimer with one quantum results in lifetimes which differ by a factor of 24 [84]. Other van der Waals molecules studied include $(\text{NO})_2$ [85], NO-HF [86], and $(\text{C}_2\text{H}_4)_2$ [87].

There are fewer experimental examples of mode specificity for the unimolecular decomposition of covalently bound molecules. One example is the decomposition of the formyl radical HCO, namely



Well defined progressions are seen in the stimulated emission pumping spectrum (SEP) [88,89] so that quantum numbers may be assigned to the HCO resonance states, and lifetimes for these states may be associated with the degree of excitation in the HC stretch, CO stretch and HCO bend vibrational modes, denoted by quantum numbers ν_1 , ν_2 , and ν_3 , respectively. States with large ν_1 and large excitations in the HC stretch have particularly short lifetimes, while states with large ν_2 and large excitations in the CO stretch have particularly long lifetimes. Short lifetimes for states with a large ν_1 might be expected, since the reaction coordinate for dissociation is primarily HC stretching motion. The mode specific effects are illustrated by the nearly isoenergetic (ν_1, ν_2, ν_3) resonance states (0, 4, 5), (1, 5, 1) and (0, 7, 0) whose respective energies (i.e. position in spectrum) are 12373, 12487 and 12544 cm^{-1} and whose respective linewidths Γ are 42, 55 and 0.72 cm^{-1} .

Time-dependent quantum mechanical calculations have also been performed to study the HCO resonance states [90,91]. The resonance energies, linewidths and quantum number assignments determined from these calculations are in excellent agreement with the experimental results.

Mode specificity has also been observed for $\text{HOCl} \rightarrow \text{Cl} + \text{OH}$ dissociation [92, 93 and 94]. For this system, many of the states are highly mixed and unassignable (see below). However, resonance states with most of the energy in the OH bond, e.g. $\nu_{\text{OH}} = 6$, are assignable and have unimolecular rate constants orders of magnitude smaller than the RRKM prediction [92, 93 and 94]. The lifetimes of these resonances have a very strong dependence on the J and K quantum numbers of HOCl.

(A) STATISTICAL STATE SPECIFICITY

In contrast to resonance states which may be assigned quantum numbers and which may exhibit mode-specific decomposition, there are states which are intrinsically unassignable. Because of extensive couplings, a zero-order Hamiltonian and its basis set cannot be found to represent the wavefunctions ψ_n for these states. The spectrum for these states is irregular without patterns, and fluctuations in the k_n are related to the manner in which the ψ_n are randomly distributed in coordinate space. Thus, the states are intrinsically unassignable and have no good quantum numbers apart from the total energy and angular momentum. Energies for these resonance states do not fit into a pattern, and states with particularly large or small rate constants are simply random occurrences in the spectrum. For the most statistical (i.e. non-separable) situation, the expansion coefficients in [equation \(A3.12.56\)](#) are random variables, subject only to the normalization and orthogonality conditions

$$\sum_n c_{in}^2 = 1 \quad \text{and} \quad \sum_i c_{in} c_{im} = 0. \quad (\text{A3.12.58})$$

If all the resonance states which form a microcanonical ensemble have random ψ_n , and are thus intrinsically unassignable, a situation arises which is called *statistical state-specific behaviour* [95]. Since the wavefunction coefficients of the ψ_n are Gaussian random variables when projected onto ϕ_i basis functions for any zero-order representation [96], the distribution of the state-specific rate constants k_n will be as statistical as possible. If these k_n within the energy interval $E \rightarrow E + dE$ form a continuous distribution, Levine [97] has argued that the probability of a particular k is given by the Porter–Thomas [98] distribution

$$P(k) = \frac{\nu}{2\bar{k}} \left(\frac{\nu k}{2\bar{k}} \right)^{((\nu-2)/2)} \frac{\exp(-\nu k/2\bar{k})}{\Gamma(1/2\nu)} \quad \text{A3.12.59}$$

where \bar{k} is the average state-specific unimolecular rate constant within the energy interval $E \rightarrow E + dE$,

$$\bar{k} = \int_0^\infty k P(k) dk \quad \text{(A3.12.60)}$$

and ν is the ‘effective number of decay channels’. Equation (A3.12.59) is derived in statistics as the probability distribution

$$X_\nu^2 = x_1^2 + x_2^2 + \dots + x_\nu^2 \quad \text{(A3.12.61)}$$

where the νx_i are each independent Gaussian distributions [96]. Increasing ν reduces the variance of the distribution $P(k)$.

The connection between the Porter–Thomas $P(k)$ distribution and RRKM theory is made through the parameters \bar{k} and ν . Waite and Miller [99] have studied the relationship between the average of the statistical state-specific rate constants k and the RRKM rate constant $k(E)$ by considering a separable (uncoupled) two-dimensional Hamilton, $H = H_x + H_y$, whose decomposition path is tunnelling through a potential energy barrier along the x -coordinate. They found that the average of the state-specific rate constants for a microcanonical ensemble \bar{k} is the same as the RRKM rate constant $k(E)$. Though insightful, this is not a general result since the tunnelling barrier defines the dividing surface, with no recrossings, which is needed to derive RRKM from classical (not quantum) mechanical principles (see section A3.12.3). For state-specific decomposition which does not occur by tunnelling, a dividing surface cannot be constructed for a quantum calculation. However, the above analysis is highly suggestive that \bar{k} may be a good approximation to the RRKM $k(E)$.

The parameter ν in equation (A3.12.59) has also been related to RRKM theory. Polik *et al* [80] have shown that for decomposition by quantum mechanical tunnelling

$$\nu = \left[\sum_n \kappa(E - E_n^\ddagger) \right]^2 / \sum_n \kappa(E - E_n^\ddagger)^2 \quad \text{(A3.12.62)}$$

where $\kappa(E - E_n^\ddagger)$ is a one-dimensional tunnelling probability through a potential barrier and E_n^\ddagger is the vibrational energy in the $3N - 7$ modes orthogonal to the tunnelling coordinate. If the energy is sufficiently low that all the tunnelling probabilities are much less than 1 and one makes a parabolic approximation to the tunnelling probabilities [96,100], equation (A3.12.62) becomes

$$\nu = \prod_{k=1}^{3N-7} \coth(\pi \omega_k^\ddagger / \omega_b) \quad \text{(A3.12.63)}$$

where the ω_k^\ddagger are the $3N - 7$ frequencies for the modes orthogonal to the tunnelling coordinate and ω_b is the

barrier frequency. The interesting aspect of [equation \(A3.12.63\)](#) is that it shows ν to be energy independent in the tunnelling region. On the other hand, for energies significantly above the barrier so that $\kappa(E - E_{\ddagger}^{\ddagger}) = 1$, it is easy to show [\[96,100\]](#) that

$$\nu = N^{\ddagger}(E) \quad (\text{A3.12.64})$$

where $N^{\ddagger}(E)$ is the sum of states for the transition state. In this energy region, ν rapidly increases with increase in energy and the $P(k)$ distribution becomes more narrowly peaked. Statistical state-specific behaviour has been observed in experimental SEP studies of $\text{H}_2\text{CO} \rightarrow \text{H}_2 + \text{CO}$ dissociation [\[44,48\]](#) and quantum mechanical scattering calculations of $\text{HO}_2 \rightarrow \text{H} + \text{O}_2$ dissociation [\[101,102\]](#). The state-specific rate constants for the latter calculation are plotted in [figure A3.12.9](#). For both of these dissociations the RRKM rate constant and the average of the state-specific quantum rate constants for a small energy interval ΔE are in good agreement. Similarly, the fluctuations in the resonance rate constants are well represented by the Porter–Thomas distribution. That HO_2 dissociation is statistical state-specific, while HCO dissociation is mode specific, is thought to arise from the deeper potential energy well and associated greater couplings and density of states for HO_2 .

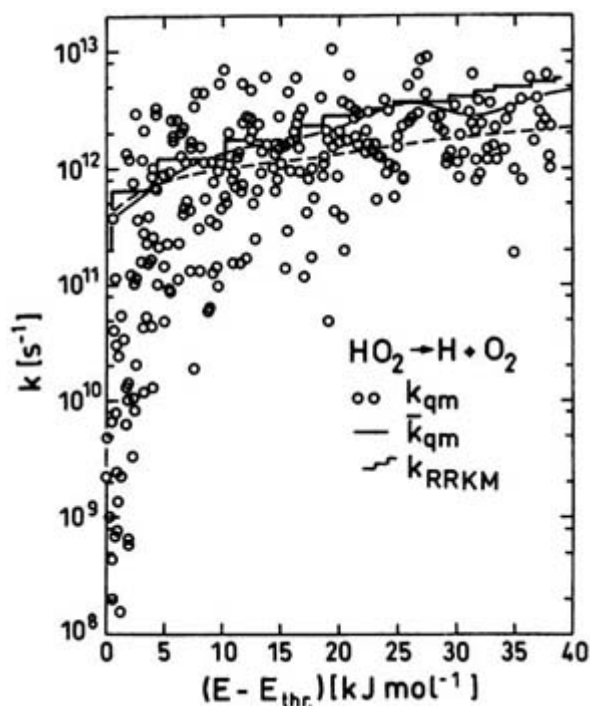


Figure A3.12.9. Comparison of the unimolecular dissociation rates for $\text{HO}_2 \rightarrow \text{H} + \text{O}_2$ as obtained from the quantum mechanical resonances (k_{qm} , open circles) and from variational transition state RRKM (k_{RRKM} , step function). E_{thr} is the threshold energy for dissociation. Also shown is the quantum mechanical average (solid line) as well as the experimental prediction for $J=0$ derived from a simplified SACM analysis of high pressure unimolecular rate constants. (Adapted from [\[101\]](#).)

A microcanonical ensemble of isolated resonances decays according to

$$N(t, E) = \sum_n \exp(-k_n t).$$

If the state-specific rate constants are assumed continuous, equation (A3.12.65) can be written as [103]

$$N(t, E) = N_0 \int_0^\infty \exp(-kt) P(k) \mathbf{d}(k) \quad (\text{A3.12.66})$$

where N_0 is the total number of molecules in the microcanonical ensemble. For the Porter–Thomas $P(k)$ distribution, $N(t, E)$ becomes [103,104]

$$N(t, E)/N_0 = (1 + 2\bar{k}t/v)^{-v/2}. \quad (\text{A3.12.67})$$

The expression for $N(t, E)$ in equation (A3.12.67) has been used to study [103,104] how the Porter–Thomas $P(k)$ affects the collision-averaged monoenergetic unimolecular rate constant $k(\omega, E)$ [105] and the Lindemann–Hinshelwood unimolecular rate constant $k_{\text{uni}}(\omega, T)$ [47]. The Porter–Thomas $P(k)$ makes $k(\omega, E)$ pressure dependent [103]. It equals \bar{k} in the high-pressure $\omega \rightarrow \infty$ limit and $[(v-2)/v]\bar{k}$ in the $\omega \rightarrow 0$ low-pressure limit. $P(k)$ only affects $k_{\text{uni}}(\omega, T)$ in the intermediate pressure regime [40,104], and has no effect on the high- and low-pressure limits. This type of analysis has been applied to $\text{HO}_2 \rightarrow \text{H} + \text{O}_2$ resonance states [106], which decay in accord with the Porter–Thomas $P(k)$. Deviations between the $k_{\text{uni}}(\omega, T)$ predicted by the Porter–Thomas and exponential $P(k)$ are more pronounced for the model in which the rotational quantum number K is treated as adiabatic than the one with K active.

A3.12.7.2 INDIVIDUAL TRANSITION STATE LEVELS

The prediction of RRKM theory is that at low energies, where $N^\ddagger(E)$ is small, there are incremental increases in $N^\ddagger(E)$ and resulting in steps in $k(E)$. The minimum rate constant is at the threshold where $N^\ddagger(E) = 1$, i.e. $k(E_0) = 1/\rho(E_0)$. Steps are then expected in $k(E)$ as $N^\ddagger(E)$ increases by unit amounts. This type of behaviour has been observed in experiments for $\text{NO}_2 \rightarrow \text{NO} + \text{O}$ [107,108], $\text{CH}_2\text{CO} \rightarrow \text{CH}_2 + \text{CO}$ [44] and $\text{CH}_3\text{CHO} \rightarrow \text{CH}_3 + \text{HCO}$ [109] dissociation. These experiments do not directly test the rapid IVR assumption of RRKM theory, since steps are expected in $N^\ddagger(E)$ even if all the states of the reactant do not participate in $\rho(E)$. However, if the measured threshold rate constant $k(E_0)$ equals the inverse of the accurate anharmonic density of states for the reactant (difficult to determine), RRKM theory is verified.

If properly interpreted [110], the above experiments provide information about the energy levels of the transition state, i.e. figure A3.12.10. For $\text{NO}_2 \rightarrow \text{NO} + \text{O}$ dissociation, there is no barrier for the reverse association reaction, and it has been suggested that the steps in its $k(E)$ may arise from quantization of the transition state's O---NO bending mode [107,108]. Ketene (CH_2CO) dissociates on both singlet and triplet surfaces. The triplet surface has a saddlepoint, at which the transition state is located, and the steps in $k(E)$ for this surface are thought to result from excitation in the transition state's CH_2 wag and C---CO bending vibrations [44]. The singlet ketene surface does not have a barrier for the reverse $^1\text{CH}_2 + \text{CO}$ association and the small steps in $k(E)$ for dissociation on this surface are attributed to CO free

rotor energy levels for a loose transition state [44]. The steps for acetaldehyde dissociation [109] have been associated with the torsional and C---CO bending motions at the transition state.

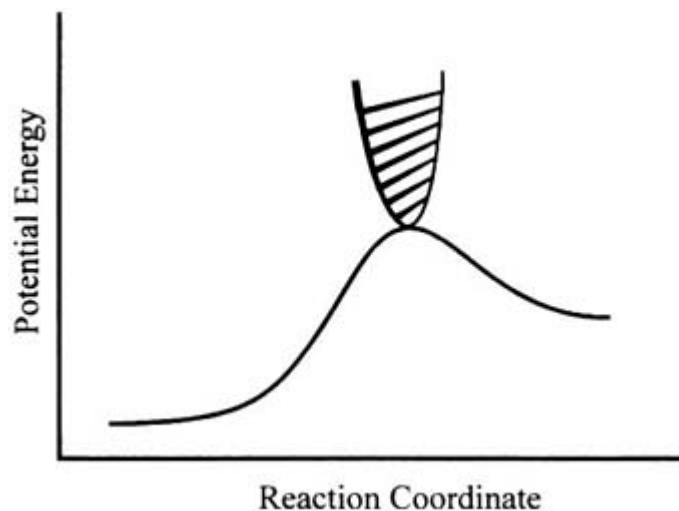


Figure A3.12.10. Schematic diagram of the one-dimensional reaction coordinate and the energy levels perpendicular to it in the region of the transition state. As the molecule's energy is increased, the number of states perpendicular to the reaction coordinate increases, thereby increasing the rate of reaction. (Adapted from [4].)

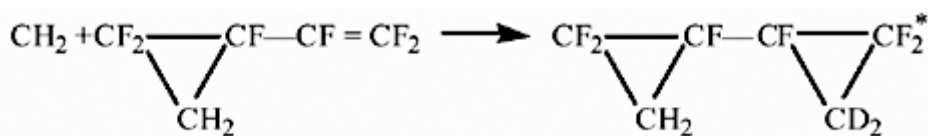
Detailed analyses of the above experiments suggest that the apparent steps in $k(E)$ may not arise from quantized transition state energy levels [110,111]. Transition state models used to interpret the ketene and acetaldehyde dissociation experiments are not consistent with the results of high-level *ab initio* calculations [110,111]. The steps observed for NO_2 dissociation may originate from the opening of electronically excited dissociation channels [107,108]. It is also of interest that RRKM-like steps in $k(E)$ are not found from detailed quantum dynamical calculations of unimolecular dissociation [91,101,102,112]. More studies are needed of unimolecular reactions near threshold to determine whether there are actual quantized transition states and steps in $k(E)$ and, if not, what is the origin of the apparent steps in the above measurements of $k(E)$.

A3.12.8 EXAMPLES OF NON-RRKM DECOMPOSITION

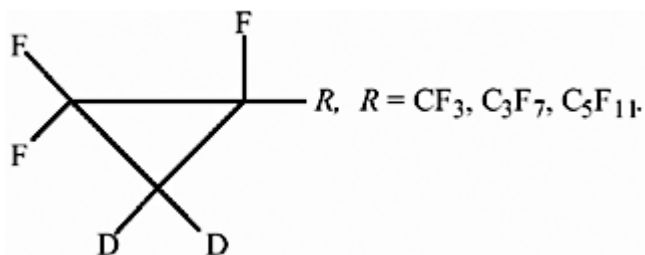
A3.12.8.1 APPARENT NON-RRKM

Apparent non-RRKM behaviour occurs when the molecule is excited non-randomly and there is an initial non-RRKM decomposition before IVR forms a microcanonical ensemble (see section A3.12.2). Reaction pathways, which have non-competitive RRKM rates, may be promoted in this way. Classical trajectory simulations were used in early studies of apparent non-RRKM dynamics [113,114].

To detect the initial apparent non-RRKM decay, one has to monitor the reaction at short times. This can be performed by studying the unimolecular decomposition at high pressures, where collisional stabilization competes with the rate of IVR. The first successful detection of apparent non-RRKM behaviour was accomplished by Rabinovitch and co-workers [115], who used chemical activation to prepare vibrationally excited hexafluorobicyclopropyl- d_2 :



The molecule decomposes by elimination of CF_2 , which should occur with equal probabilities from each ring when energy is randomized. However, at pressures in excess of 100 Torr there is a measurable increase in the fraction of decomposition in the ring that was initially excited. From an analysis of the relative product yield *versus* pressure, it was deduced that energy flows between the two cyclopropyl rings with a rate of only $3 \times 10^9 \text{ s}^{-1}$. In a related set of experiments Rabinovitch *et al* [116] studied the series of chemically activated fluoroalkyl cyclopropanes:



The chemically activated molecules are formed by reaction of $^1\text{CH}_2$ with the appropriate fluorinated alkene. In all these cases apparent non-RRKM behaviour was observed. As displayed in [figure A3.12.11](#) the measured unimolecular rate constants are strongly dependent on pressure. The large rate constant at high pressure reflects an initial excitation of only a fraction of the total number of vibrational modes, i.e. initially the molecule behaves smaller than its total size. However, as the pressure is decreased, there is time for IVR to compete with dissociation and energy is distributed between a larger fraction of the vibrational modes and the rate constant decreases. At low pressures each rate constant approaches the RRKM value.

-34-

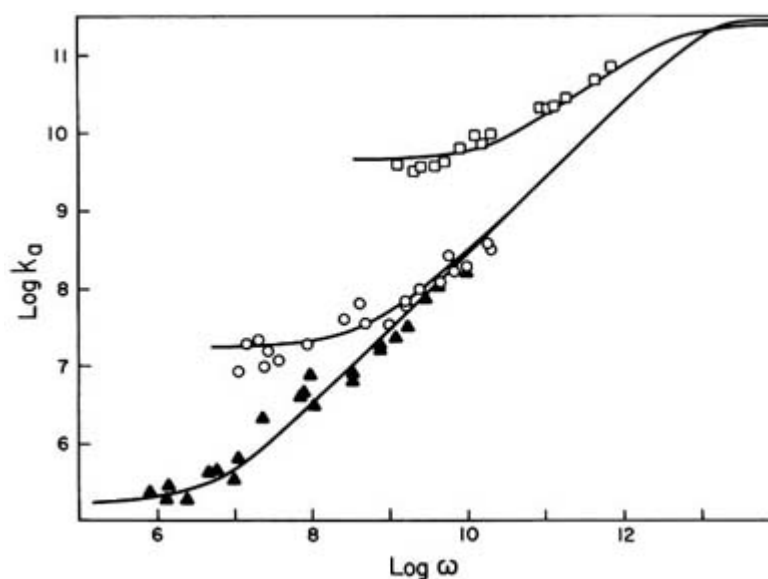
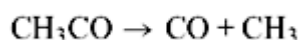


Figure A3.12.11. Chemical activation unimolecular rate constants *versus* ω for fluoroalkyl cyclopropanes. The \square , \circ and \blacktriangle points are for $R=\text{CF}_3$, C_3F_7 , and C_5F_{11} , respectively. (Adapted from [116].)

Apparent non-RRKM dynamics has also been observed in time-resolved femtosecond (fs) experiments in a collision-free environment [117]. An experimental study of acetone illustrates this work. Acetone is dissociated to the CH_3 and CH_3CO (acetyl) radicals by a fs laser pulse. The latter which dissociates by the channel



is followed in real time by fs mass spectrometry to measure its unimolecular rate constant. It is found to be $2 \times 10^{12} \text{ s}^{-1}$ and ~ 10 times smaller than the RRKM value, which indicates the experimental excitation process does not put energy in the C–C reaction coordinate and the rate constant value and short timescale reflects restricted IVR and non-RRKM kinetics.

A3.12.8.2 INTRINSIC NON-RRKM

As discussed in section A3.12.2, intrinsic non-RRKM behaviour occurs when there is at least one bottleneck for transitions between the reactant molecule's vibrational states, so that IVR is slow and a microcanonical ensemble over the reactant's phase space is not maintained during the unimolecular reaction. The above discussion of mode-specific decomposition illustrates that there are unimolecular reactions which are intrinsically non-RRKM. Many van der Waals molecules behave in this manner [4,82]. For example, in an initial microcanonical ensemble for the $(\text{C}_2\text{H}_4)_2$ van der Waals molecule both the $\text{C}_2\text{H}_4 \cdots \text{C}_2\text{H}_4$ intermolecular modes and C_2H_4 intramolecular modes are excited with equal probabilities. However, this microcanonical ensemble is not maintained as the dimer dissociates. States with energy in the intermolecular modes react more rapidly than do those with the C_2H_4 intramolecular modes excited [85].

-35-

Furthermore, IVR is not rapid between the C_2H_4 intramolecular modes and different excitation patterns of these modes result in different dissociation rates. As a result of these different timescales for dissociation, the relative populations of the vibrational modes of the C_2H_4 dimer change with time.

Similar behaviour is observed in both experiments and calculations for $\text{HCO} \rightarrow \text{H} + \text{CO}$ dissociation [88, 89, 90 AND 91] and in calculations for the $\text{X}^- \cdots \text{CH}_3\text{Y}$ ion–dipole complexes, which participate in $\text{S}_{\text{N}}2$ nucleophilic substitution reactions [118]. HCO states with HC excitation dissociate more rapidly than do those with CO excitation and, thus, the relative population of HC to CO excitation decreases with time. The unimolecular dynamics of the $\text{X}^- \cdots \text{CH}_3\text{Y}$ complex is similar to that for van der Waals complexes. There is weak coupling between the $\text{X}^- \cdots \text{CH}_3\text{Y}$ intermolecular modes and the CH_3Y intramolecular modes, and the two sets of modes react on different timescales.

Definitive examples of intrinsic non-RRKM dynamics for molecules excited near their unimolecular thresholds are rather limited. Calculations have shown that intrinsic non-RRKM dynamics becomes more pronounced at very high energies, where the RRKM lifetime becomes very short and dissociation begins to compete with IVR [119]. There is a need for establishing quantitative theories (i.e. not calculations) for identifying which molecules and energies lead to intrinsic non-RRKM dynamics. For example, at thermal energies the unimolecular dynamics of the $\text{Cl}^- \cdots \text{CH}_3\text{Cl}$ complex is predicted to be intrinsically non-RRKM [118], while experiments have shown that simply replacing one of the H-atoms of CH_3Cl with a CN group leads to intrinsic RRKM dynamics for the $\text{Cl}^- \cdots \text{ClCH}_2\text{CN}$ complex [120]. This difference is thought to arise from a deeper potential energy well and less of a separation between vibrational frequencies for the $\text{Cl}^- \cdots \text{ClCH}_2\text{CN}$ complex. For the $\text{Cl}^- \cdots \text{CH}_3\text{Cl}$ complex the three intermolecular vibrational frequencies are 64(2) and 95 cm^{-1} , while the lowest intramolecular frequency is the C–Cl stretch of 622 cm^{-1} [118]. Thus, very

high-order resonances are required for energy transfer from the intermolecular to intramolecular modes. In contrast, for the $\text{Cl}^- \cdots \text{ClCH}_2\text{CN}$ complex there is less of a hierarchy of frequencies, with 44, 66 and 118 cm^{-1} for the intermolecular modes and 207, 367, 499, 717, ... for the intramolecular ones [120]. Here there are low-order resonances which couple the intermolecular and intramolecular modes. It would be very useful if one could incorporate such molecular properties into a theoretical model to predict intrinsic RRKM and non-RRKM behaviour.

ACKNOWLEDGMENTS

The author wishes to thank the many graduate students, post-doctorals and collaborators who have worked with him on this topic. Of particular importance are the many discussions the author has had with Tom Baer, Reinhard Schinke and Jürgen Troe concerning unimolecular dynamics. Special thanks are given to Fleming Crim, Martin Quack and Kihyung Song for their valuable comments in preparing this chapter.

REFERENCES

- [1] Levine R D and Bernstein R B 1987 *Molecular Reaction Dynamics and Chemical Reactivity* (New York: Oxford University Press)
 - [2] Schinke R 1993 *Photodissociation Dynamics* (New York: Cambridge University Press)
 - [3] McQuarrie D A 1973 *Statistical Thermodynamics* (New York: Harper and Row)
-
- 36-
- [4] Baer T and Hase W L 1996 *Unimolecular Reaction Dynamics. Theory and Experiments* (New York: Oxford University Press)
 - [5] Steinfeld J I, Francisco J S and Hase W L 1999 *Chemical Kinetics and Dynamics* 2nd edn (Upper Saddle River, NJ: Prentice-Hall)
 - [6] Uzer T 1991 Theories of intramolecular vibrational energy transfer *Phys. Rep.* **199** 73–146
 - [7] Marcus R A 1952 Unimolecular dissociations and free radical recombination reactions *J. Chem. Phys.* **20** 359–64
 - [8] Rosenstock H M, Wallenstein M B, Wahrhaftig A L and Eyring H 1952 Absolute rate theory for isolated systems and the mass spectra of polyatomic molecules *Proc. Natl Acad. Sci. USA* **38** 667–78
 - [9] Bunker D L and Hase W L 1973 On non-RRKM unimolecular kinetics: molecules in general and CH_3NC in particular *J. Chem. Phys.* **59** 4621–32
 - [10] Miller W H 1976 Importance of nonseparability in quantum mechanical transition-state theory *Acc. Chem. Res.* **9** 306–12
 - [11] Chandler D 1987 *Introduction to Modern Statistical Mechanics* (New York: Oxford University Press)
 - [12] Bunker D L 1964 Monte Carlo calculations. IV. Further studies of unimolecular dissociation *J. Chem. Phys.* **40** 1946–57
 - [13] Bunker D L 1962 Monte Carlo calculation of triatomic dissociation rates. I. N_2O and O_3 *J. Chem. Phys.* **37** 393–403

- [14] Gutzwiller M C 1990 *Chaos in Classical and Quantum Mechanics* (New York: Springer)
- [15] Slater J C 1951 *Quantum Theory of Matter* (New York: McGraw-Hill)
- [16] Robinson P J and Holbrook K A 1972 *Unimolecular Reactions* (New York: Wiley)
- [17] Hase W L and Buckowski D G 1982 Dynamics of ethyl radical decomposition. II. Applicability of classical mechanics to large-molecule unimolecular reaction dynamics *J. Comp. Chem.* **3** 335–43
- [18] Slater N B 1959 *Theory of Unimolecular Reactions* (Ithaca, NY: Cornell University Press)
- [19] Miller W H 1976 Unified statistical model for 'complex' and 'direct' reaction mechanisms *J. Chem. Phys.* **65** 2216–23
- [20] Doll J D 1980 A unified theory of dissociation *J. Chem. Phys.* **73** 2760–2
- [21] Truhlar D G and Garrett B C 1980 Variational transition-state theory *Acc. Chem. Res.* **13** 440–8
- [22] Hase W L 1983 Variational unimolecular rate theory *Acc. Chem. Res.* **16** 258–64
- [23] Hase W L 1972 Theoretical critical configuration for ethane decomposition and methyl radical recombination *J. Chem. Phys.* **57** 730–3
-

-37-

- [24] Lin Y N and Rabinovitch B S 1970 A simple quasi-accommodation model of vibrational energy transfer *J. Phys. Chem.* **74** 3151–9
- [25] Wardlaw D M and Marcus R A 1984 RRKM reaction rate theory for transition states of any looseness *Chem. Phys. Lett.* **110** 230–4
- [26] Hase W L and Wardlaw D M 1989 *Bimolecular Collisions* ed M N R Ashfold and J E Baggott (London: Royal Society of Chemistry) p 171
- [27] Song K, Peslherbe G H, Hase W L, Dobbyn A J, Stumpf M and Schinke R 1995 Comparison of quantum and semiclassical variational transition state models for $\text{HO}_2 \rightarrow \text{H} + \text{O}_2$ microcanonical rate constants *J. Chem. Phys.* **103** 8891–900
- [28] Wardlaw D M and Marcus R A 1987 On the statistical theory of unimolecular processes *Adv. Chem. Phys.* **70** 231–63
- [29] Chesnavich W J and Bowers M T 1979 *Gas Phase Ion Chemistry* vol 1, ed M T Bowers (New York: Academic) p 119
- [30] Peslherbe G H and Hase W L 1996 Statistical anharmonic unimolecular rate constants for the dissociation of fluxional molecules. Application to aluminum clusters *J. Chem. Phys.* **105** 7432–47
- [31] Klippenstein S J 1992 Variational optimizations in the Rice–Ramsperger–Kassel–Marcus theory calculations for unimolecular dissociations with no reverse barrier *J. Chem. Phys.* **96** 367–71
- [32] Chesnavich W J, Bass L, Su T and Bowers M T 1981 Multiple transition states in unimolecular reactions: a transition state switching model. Application to C_4H_8^+ *J. Chem. Phys.* **74** 2228–46
- [33] Hu X and Hase W L 1989 Properties of canonical variational transition state theory for association reactions without potential energy barriers *J. Phys. Chem.* **93** 6029–38
- [34] Song K and Chesnavich W J 1989 Multiple transition states in chemical reactions: variational transition state theory studies of the HO_2 and HeH_2^+ systems *J. Chem. Phys.* **91** 4664–78

- [35] Song K and Chesnavich W J 1990 Multiple transition state in chemical reactions. II. The effect of angular momentum in variational studies of HO₂ and HeH₂⁺ systems *J. Chem. Phys.* **93** 5751–9
- [36] Klippenstein S J, East A L L and Allen W D A 1994 First principles theoretical determination of the rate constant for the dissociation of singlet ketene *J. Chem. Phys.* **101** 9198–201
- [37] Kassel L S 1928 Studies in homogeneous gas reactions. II. Introduction of quantum theory *J. Phys. Chem.* **32** 1065–79
- [38] Beyer T and Swinehart D R 1973 Number of multiply-restricted partitions *ACM Commun.* **16** 379
- [39] Zhu L, Chen W, Hase W L and Kaiser E W 1993 Comparison of models for treating angular momentum in RRKM calculations with vibrator transition states. Pressure and temperature dependence of Cl+C₂H₂ association *J. Phys. Chem.* **97** 311–22
-

-38-

- [40] Hase W L 1998 Some recent advances and remaining questions regarding unimolecular rate theory *Acc. Chem. Res.* **31** 659–65
- [41] Quack M and Troe J 1974 Specific rate constants of unimolecular processes. II. Adiabatic channel model *Ber. Bunsenges. Phys. Chem.* **78** 240–52
- [42] Miller W H 1979 Tunneling corrections to unimolecular rate constants, with applications to formaldehyde *J. Am. Chem. Soc.* **101** 6810–14
- [43] Bunker D L and Pattengill M 1968 Monte Carlo calculations. VI. A re-evaluation of the RRKM theory of unimolecular reaction rates *J. Chem. Phys.* **48** 772–6
- [44] Green W H, Moore C B and Polik W F 1992 Transition states and rate constants for unimolecular reactions *Ann. Rev. Phys. Chem.* **43** 591–626
- [45] Ionov S I, Brucker G A, Jaques C, Chen Y and Wittig C 1993 Probing the NO₂ → NO+O transition state via time resolved unimolecular decomposition *J. Chem. Phys.* **99** 3420–35
- [46] Viswanathan R, Raff L M and Thompson D L 1984 Monte Carlo random walk calculations of unimolecular dissociation of methane *J. Chem. Phys.* **81** 3118–21
- [47] Gilbert R G and Smith S C 1990 *Theory of Unimolecular and Recombination Reactions* (London: Blackwell)
- [48] Polik W F, Guyer D R and Moore C B 1990 Stark level-crossing spectroscopy of S₀ formaldehyde eigenstates at the dissociation threshold *J. Chem. Phys.* **92** 3453–70
- [49] Abramson E, Field R W, Imre D, Innes K K and Kinsey J L 1985 Fluorescence and stimulated emission S₁ → S₀ spectra of acetylene: regular and ergodic regions *J. Chem. Phys.* **85** 453–65
- [50] Wagner A F, Kiefer J H and Kumaran S S 1992 The importance of hindered rotation and other anharmonic effects in the thermal dissociation of small unsaturated molecules: application to HCN *Twenty-Fourth Symposium on Combustion* (Combustion Institute) 613–19
- [51] Bhuiyan L B and Hase W L 1983 Sum and density of states for anharmonic polyatomic molecules. Effect of bend-stretch coupling *J. Chem. Phys.* **78** 5052–8
- [52] Haarhoff P C 1963 The density of vibrational energy levels of polyatomic molecules *Mol. Phys.* **7** 101–17
- [53] Forst W 1971 Methods for calculating energy-level densities *Chem. Rev.* **71** 339–56

- [54] Troe J 1983 Specific rate constants $k(E, J)$ for unimolecular bond fissions *J. Chem. Phys.* **79** 6017–29
- [55] Troe J 1995 Simplified models for anharmonic numbers and densities of vibrational states. I. Application to NO_2 and H_3^+ *Chem. Phys.* **190** 381–92
- [56] Song K and Hase W L 1999 Fitting classical microcanonical unimolecular rate constants to a modified RRK expression: anharmonic and variational effects *J. Chem. Phys.* **110** 6198–207
- [57] Duffy L M, Keister J W and Baer T 1995 Isomerization and dissociation in competition. The pentene ion story *J. Phys. Chem.* **99** 17 862–71
-

-39-

- [58] Hose G and Taylor H A 1982 A quantum analog to the classical quasiperiodic motion *J. Chem. Phys.* **76** 5356–64
- [59] Wilson E B Jr, Decius J C and Cross P C 1955 *Molecular Vibrations* (New York: McGraw-Hill)
- [60] Califano S 1976 *Vibrational States* (New York: Wiley)
- [61] Herzberg G 1945 *Molecular Spectra and Molecular Structure. II. Infrared and Raman Spectra of Polyatomic Molecules* (New York: Van Nostrand-Reinhold)
- [62] Oxtoby D W and Rice S A 1976 Nonlinear resonance and stochasticity in intramolecular energy exchange *J. Chem. Phys.* **65** 1676–83
- [63] W L Hase (ed) 1992 *Advances in Classical Trajectory Methods. 1. Intramolecular and Nonlinear Dynamics* (London: JAI)
- [64] Sloane C S and Hase W L 1977 On the dynamics of state selected unimolecular reactions: chloroacetylene dissociation and predissociation *J. Chem. Phys.* **66** 1523–33
- [65] Wolf R J and Hase W L 1980 Quasiperiodic trajectories for a multidimensional anharmonic classical Hamiltonian excited above the unimolecular threshold *J. Chem. Phys.* **73** 3779–90
- [66] Viswanathan R, Thompson D L and Raff L M 1984 Theoretical investigations of elementary processes in the chemical vapor deposition of silicon from silane. Unimolecular decomposition of SiH_4 *J. Chem. Phys.* **80** 4230–40
- [67] Sorescu D, Thompson D L and Raff L M 1994 Statistical effects in the skeletal inversion of bicyclo (2.1.0)pentane *J. Chem. Phys.* **101** 3729–41
- [68] Lichtenberg A J and Leiberman M A 1992 *Regular and Chaotic Dynamics* 2nd edn (New York: Springer)
- [69] Brickmann J, Pfeiffer R and Schmidt P C 1984 The transition between regular and chaotic dynamics and its influence on the vibrational energy transfer in molecules after local preparation *Ber. Bunsenges. Phys. Chem.* **88** 382–97
- [70] Jaffé C and Brumer P 1980 Local and normal modes: a classical perspective *J. Chem. Phys.* **73** 5646–58
- [71] Sibert E L III, Reinhardt W P and Hynes J T 1982 Classical dynamics of energy transfer between bonds in ABA triatomics *JCP* **77** 3583–94
- [72] Marcus R A 1977 Energy distributions in unimolecular reactions *Ber. Bunsenges. Phys. Chem.* **81** 190–7
- [73] Miller W H 1974 Classical-limit quantum mechanics and the theory of molecular collisions *Adv. Chem. Phys.* **25** 69–177

- [74] Child M S 1991 *Semiclassical Mechanics with Molecular Applications* (New York: Oxford University Press)
- [75] Marcus R A 1973 Semiclassical theory for collisions involving complexes (compound state resonances) and for bound state systems *Faraday Discuss. Chem. Soc.* **55** 34–44
- [76] Heller E J 1983 The correspondence principle and intramolecular dynamics *Faraday Discuss. Chem. Soc.* **75** 141–53
-

-40-

- [77] Shirts R B and Reinhardt W P 1982 Approximate constants of motion for classically chaotic vibrational dynamics: vague tori, semiclassical quantization, and classical intramolecular energy flow *J. Chem. Phys.* **77** 5204–17
- [78] Landau L D and Lifshitz E M 1965 *Quantum Mechanics* (London: Addison-Wesley)
- [79] Mies F H and Krauss M 1966 Time-dependent behavior of activated molecules. High-pressure unimolecular rate constant and mass spectra *J. Chem. Phys.* **45** 4455–68
- [80] Mies F H 1969 Resonant scattering theory of association reactions and unimolecular decomposition. II. Comparison of the collision theory and the absolute rate theory *J. Chem. Phys.* **51** 798–807
- [81] Rice O K 1971 On the relation between unimolecular reaction and predissociation *J. Chem. Phys.* **55** 439–46
- [82] Miller R E 1988 The vibrational spectroscopy and dynamics of weakly bound neutral complexes *Science* **240** 447–53
- [83] Ewing G E 1980 Vibrational predissociation in hydrogen-bonded complexes *J. Chem. Phys.* **72** 2096–107
- [84] Huang Z S, Jucks K W and Miller R E 1986 The vibrational predissociation lifetime of the HF dimer upon exciting the 'free-H' stretching vibration *J. Chem. Phys.* **85** 3338–41
- [85] Casassa M P, Woodward A M, Stephenson J C and Kind D S 1986 Picosecond measurements of the dissociation rates of the nitric oxide dimer $v_1 = 1$ and $v_4 = 1$ levels *J. Chem. Phys.* **85** 6235–7
- [86] Lovejoy C M and Nesbitt D J 1989 The infrared spectra of NO–HF isomers *J. Chem. Phys.* **90** 4671–80
- [87] Fischer G, Miller R E, Vohralik P F and Watts R O 1985 Molecular beam infrared spectra of dimers formed from acetylene, methyl acetylene and ethene as a function of source pressure and concentration *J. Chem. Phys.* **83** 1471–7
- [88] Tobiason J D, Dunlap J R and Rohlfing E A 1995 The unimolecular dissociation of HCO: a spectroscopic study of resonance energies and widths *J. Chem. Phys.* **103** 1448–69
- [89] Stock C, Li X, Keller H-M, Schinke R and Temps F 1997 Unimolecular dissociation dynamics of highly vibrationally excited DCO (\bar{X}^2A'). I. Investigation of dissociative resonance states by stimulated emission pumping spectroscopy *J. Chem. Phys.* **106** 5333–58
- [90] Wang D and Bowman J M 1995 Complex L^2 calculations of bound states and resonances of HCO and DCO *Chem. Phys. Lett.* **235** 277–85
- [91] Stumpf M, Dobbyn A J, Mordaunt D H, Keller H-M, Fluethmann H and Schinke R 1995 Unimolecular dissociations of HCO, HNO, and HO₂: from regular to irregular dynamics *Faraday Discuss.* **102** 193–213
- [92] Barnes R J, Dutton G and Sinha A 1997 Unimolecular dissociation of HOCl near threshold: quantum state and time-resolved studies *J. Phys. Chem. A* **101** 8374–7

- [93] Callegari A, Rebstein J, Muentner J S, Jost R and Rizzo T R 1999 The spectroscopy and intramolecular vibrational energy redistribution dynamics of HOCl in the $\nu(\text{OH}) = 6$ region, probed by infrared-visible double resonance overtone excitation *J. Chem. Phys.* **111** 123–33
-

-41-

- [94] Skokov S and Bowman J M 1999 Variation of the resonance width of HOCl ($6\nu_{\text{OH}}$) with total angular momentum: comparison between *ab initio* theory and experiment *J. Chem. Phys.* **110** 9789–92
- [95] Hase W L, Cho S-W, Lu D-H and Swamy K N 1989 The role of state specificity in unimolecular rate theory *Chem. Phys.* **139** 1–13
- [96] Polik W F, Guyer D R, Miller W H and Moore C B 1990 Eigenstate-resolved unimolecular reaction dynamics: ergodic character of S_0 formaldehyde at the dissociation threshold *J. Chem. Phys.* **92** 3471–84
- [97] Levine R D 1987 Fluctuations in spectral intensities and transition rates *Adv. Chem. Phys.* **70** 53–95
- [98] Porter C E and Thomas R G 1956 Fluctuations of nuclear reaction widths *Phys. Rev.* **104** 483–91
- [99] Waite B A and Miller W H 1980 Model studies of mode specificity in unimolecular reaction dynamics *J. Chem. Phys.* **73** 3713–21
- [100] Miller W H, Hernandez R, Moore C B and Polik W F A 1990 Transition state theory-based statistical distribution of unimolecular decay rates with application to unimolecular decomposition of formaldehyde *J. Chem. Phys.* **93** 5657–66
- [101] Stumpf M, Dobbyn A J, Keller H-M, Hase W L and Schinke R 1995 Quantum mechanical study of the unimolecular dissociation of HO_2 : a rigorous test of RRKM theory *J. Chem. Phys.* **102** 5867–70
- [102] Dobbyn A J, Stumpf M, Keller H-M and Schinke R 1996 Theoretical study of the unimolecular dissociation $\text{HO}_2 \rightarrow \text{H} + \text{O}_2$. II. Calculation of resonant states, dissociation rates, and O_2 product state distributions *J. Chem. Phys.* **104** 8357–81
- [103] Lu D-H and Hase W L 1989 Monoenergetic unimolecular rate constants and their dependence on pressure and fluctuations in state-specific unimolecular rate constants *J. Phys. Chem.* **93** 1681–3
- [104] Miller W H 1988 Effect of fluctuations in state-specific unimolecular rate constants on the pressure dependence of the average unimolecular reaction rate *J. Phys. Chem.* **92** 4261–3
- [105] Rabinovitch B S and Setser D W 1964 Unimolecular decomposition and some isotope effects of simple alkanes and alkyl radicals *Adv. Photochem.* **3** 1–82
- [106] Song K and Hase W L 1998 Role of state specificity in the temperature- and pressure-dependent unimolecular rate constants for $\text{HO}_2 \rightarrow \text{H} + \text{O}_2$ dissociation *J. Phys. Chem. A* **102** 1292–6
- [107] Ionov S I, Brucker G A, Jaques C, Chen Y and Wittig C 1993 Probing the $\text{NO}_2 \rightarrow \text{NO} + \text{O}$ transition state via time resolved unimolecular decomposition *J. Chem. Phys.* **99** 3420–35
- [108] Miyawaki J, Yamanouchi K and Tsuchiya S 1993 State-specific unimolecular reaction of NO_2 just above the dissociation threshold *J. Chem. Phys.* **99** 254–64
- [109] Leu G-H, Huang C-L, Lee S-H, Lee Y-C and Chen I-C 1998 Vibrational levels of the transition state and rate of dissociation of triplet acetaldehyde *J. Chem. Phys.* **109** 9340–50
-

- [110] King R A, Allen W D and Schaefer H F III 2000 On apparent quantized transition-state thresholds in the photofragmentation of acetaldehyde *J. Chem. Phys.* **112** 5585–92
- [111] Gezelter J D and Miller W H 1996 Dynamics of the photodissociation of triplet ketene *J. Chem. Phys.* **104** 3546–54
- [112] Schinke R, Beck C, Grebenshchikov S Y and Keller H-M 1998 Unimolecular dissociation: a state-specific quantum mechanical perspective *Ber. Bunsenges. Phys. Chem.* **102** 593–611
- [113] Hase W L 1976 *Modern Theoretical Chemistry. 2. Dynamics of Molecular Collisions* part B, ed W H Miller (New York: Plenum) p 121
- [114] Hase W L 1981 *Potential Energy Surfaces and Dynamics Calculations* ed D G Truhlar (New York: Plenum) p 1
- [115] Rynbrandt J D and Rabinovitch B S 1971 Direct demonstration of nonrandomization of internal energy in reacting molecules. Rate of intramolecular energy relaxation *J. Chem. Phys.* **54** 2275–6
- [116] Meagher J F, Chao K J, Barker J R and Rabinovitch B S 1974 Intramolecular vibrational energy relaxation. Decomposition of a series of chemically activated fluoroalkyl cyclopropanes *J. Phys. Chem.* **78** 2535–43
- [117] Kim S K, Guo J, Baskin J S and Zewail A H 1996 Femtosecond chemically activated reactions: concept of nonstatistical activation at high thermal energies *J. Phys. Chem.* **100** 9202–5
- [118] Hase W L 1994 Simulations of gas-phase chemical reactions: applications to S_N2 nucleophilic substitution *Science* **266** 998–1002
- [119] Shalashilin D V and Thompson D L 1996 Intrinsic non-RRK behavior: classical trajectory, statistical theory, and diffusional theory studies of a unimolecular reaction *J. Chem. Phys.* **105** 1833–45
- [120] Wladkowski B D, Lim K F, Allen W D and Brauman J I 1992 The S_N2 exchange reaction $ClCH_2CN + Cl^- \rightarrow Cl^- + ClCH_2CN$: experiment and theory *J. Am. Chem. Soc.* **114** 9136–53
- [121] Hase W L 1982 *J. Phys. Chem.* **86** 2873–9
-

FURTHER READING

Forst W 1973 *Theory of Unimolecular Reactions* (New York: Academic)

Hase W L 1976 *Modern Theoretical Chemistry, Dynamics of Molecular Collisions* part B, ed W H Miller (New York: Plenum) p 121

Gilbert R G and Smith S C 1990 *Theory of Unimolecular and Recombination Reactions* (London: Blackwell Scientific)

Uzer T 1991 *Phys. Rep.* **199** 73–146

Baer T and Hase W L 1996 *Unimolecular Reaction Dynamics. Theory and Experiments* (New York: Oxford University Press)

Thompson D L 1999 *Int. Rev. Phys. Chem.* **17** 547–69

Quack M and Troe J 1981 *Theoretical Chemistry: Advances and Perspectives* vol 6B, ed E Henderson (New York: Academic) p 199

A 3.13 Energy redistribution in reacting systems

Roberto Marquardt and Martin Quack

A 3.13.1 INTRODUCTION

Energy redistribution is the key primary process in chemical reaction systems, as well as in reaction systems quite generally (for instance, nuclear reactions). This is because many reactions can be separated into two steps:

(a) activation of the reacting species R, generating an energized species R*:



(b) reaction of the energized species to give products.



The first step (A3.13.1) is a general process of energy redistribution, whereas the second step (A3.13.2) is the genuine reaction step, occurring with a specific rate constant at energy E . This abstract reaction scheme can take a variety of forms in practice, because both steps may follow a variety of quite different mechanisms. For instance, the reaction step could be a barrier crossing of a particle, a tunnelling process or a nonadiabatic crossing between different potential hypersurfaces to name just a few important examples in chemical reactions.

The first step, which is the topic of the present chapter, can again follow a variety of different mechanisms. For instance, the energy transfer could happen within a molecule, say from one initially excited chemical bond to another, or it could involve radiative transfer. Finally, the energy transfer could involve a collisional transfer of energy between different atoms or molecules. All these processes have been recognized to be important for a very long time. The basic idea of collisional energization as a necessary primary step in chemical reactions can be found in the early work of van't Hoff [1] and Arrhenius [2, 3], leading to the famous Arrhenius equation for thermal chemical reactions (see also [chapter A3.4](#))

$$k(T) = A(T) \exp\left(-\frac{E_A(T)}{RT}\right). \quad (\text{A 3.13.3})$$

This equation results from the assumption that the actual reaction step in thermal reaction systems can happen only in molecules (or collision pairs) with an energy exceeding some threshold energy E_0 which is close, in general, to the Arrhenius activation energy defined by equation (A3.13.3). Radiative energization is at the basis of classical photochemistry (see e.g. [4, 3 and 7] and [chapter B2.5](#)) and historically has had an interesting sideline in the radiation

theory of unimolecular reactions [8], which was later superseded by the collisional Lindemann mechanism [9]. Recently, radiative energy redistribution has received new impetus through coherent and incoherent multiphoton excitation [10].

In this chapter we shall first outline the basic concepts of the various mechanisms for energy redistribution, followed by a very brief overview of collisional intermolecular energy transfer in chemical reaction systems. The main part of this chapter deals with true intramolecular energy transfer in polyatomic molecules, which is a topic of particular current importance. Stress is placed on basic ideas and concepts. It is not the aim of this chapter to review in detail the vast literature on this topic; we refer to some of the key reviews and books [11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, and 32] and the literature cited therein. These cover a variety of aspects of the topic and further, more detailed references will be given throughout this review. We should mention here the energy transfer processes, which are of fundamental importance but are beyond the scope of this review, such as electronic energy transfer by mechanisms of the Förster type [33, 34] and related processes.

A 3.13.2 BASIC CONCEPTS FOR INTER- AND INTRAMOLECULAR ENERGY TRANSFER

The processes summarized by equation (A3.13.1) can follow quite different mechanisms and it is useful to classify them and introduce the appropriate nomenclature as well as the basic equations.

A3.13.2.1 PROCESSES INVOLVING INTERACTION WITH THE ENVIRONMENT (BIMOLECULAR AND RELATED)

(a) The first mechanism concerns bimolecular, collisional energy transfer between two molecules or atoms and molecules. We may describe such a mechanism by



or more precisely by defining quantum energy levels for both colliding species, e.g.



This is clearly a process of intermolecular energy transfer, as energy is transferred between two molecular species. Generally one may, following chapter A.3.4.5, combine the quantum labels of M and R into one level index (I for initial and F for final) and define a cross section σ_{FI} for this energy transfer. The specific rate constant $k_{FI}(E_{tI})$ for the energy transfer with the collision energy E_{tI} is given by:

$$k_{FI}(E_{tI}) = \sigma_{FI}(E_{tI}) \sqrt{\frac{2E_{tI}}{\mu}} \quad (\text{A 3.13.6})$$

with the reduced mass:

$$E_{Mi} + E_{Ri} + E_{t,I} = E_{Mf} + E_{Rf} + E_{t,F}. \quad (\text{A } 3.13.7)$$

We note that, by energy conservation, the following equation must hold:

$$E_{Mi} + E_{Ri} + E_{t,I} = E_{Mf} + E_{Rf} + E_{t,F}. \quad (\text{A } 3.13.8)$$

Some of the internal (rovibronic) energy of the atomic and molecular collision partners is transformed into extra translational energy $\Delta E_t = E_{t,F} - E_{t,I}$ (or consumed, if ΔE_t is negative). If one averages over a thermal distribution of translational collision energies, one obtains the thermal rate constant for collisional energy transfer:

$$k_{FI}(T) = \left(\frac{8k_B T}{\pi \mu} \right)^{1/2} \int_0^\infty x \exp(-x) \sigma_{FI}(k_B T x) dx. \quad (\text{A } 3.13.9)$$

We note here that the quantum levels denoted by the capital indices I and F may contain numerous energy eigenstates, i.e. are highly degenerate, and refer to [chapter A3.4](#) for a more detailed discussion of these equations. The integration variable in equation (A3.13.9) is $x = E_{t,I} / k_B T$.

(b) The second mechanism, which is sometimes distinguished from the first although it is similar in kind, is obtained when we assume that the colliding species M does not change its internal quantum state. This special case is frequently realized if M is an inert gas atom in its electronic ground state, as the energies needed to generate excited states of M would then greatly exceed the energies available in ordinary reaction systems at modest temperatures. This type of mechanism is frequently called *collision induced intramolecular energy transfer*, as internal energy changes occur only *within the molecule* R. One must note that in general there is transfer of energy between *intermolecular translation* and intramolecular rotation and vibration in such a process, and thus the nomenclature ‘intramolecular’ is somewhat unfortunate. It is, however, widely used, which is the reason for mentioning it here. In the following, we shall not make use of this nomenclature and shall summarize mechanisms (a) and (b) as one class of bimolecular, intermolecular process. We may also note that, for mechanism (b) one can define a cross section σ_{FI} and rate constant k_{FI} between individual, nondegenerate quantum states *i* and *f* and obtain special equations analogous to [equation \(A3.13.5\)](#), [equation \(A3.13.4\)](#) and [equation \(A3.13.3\)](#), which we shall not repeat in detail. Indeed, one may then have cross sections and rates between different individual quantum states *i* and *f* of the same energy and thus no transfer of energy to translation. In this very special case, the redistribution of energy would indeed be entirely ‘intramolecular’ within R.

-4-

(c) The third mechanism would be transfer of energy between molecules and the radiation field. These processes involve absorption, emission or Raman scattering of radiation and are summarized, in the simplest case with one or two photons, in [equation \(A3.13.10\)](#), [equation \(A3.13.11\)](#) and [equation \(A3.13.12\)](#):



$$R_{i^{\nu}} + h\nu_i \rightarrow R_{f^{\nu}} + h\nu_f \quad (\text{Raman scattering}). \quad (\text{A 3.13.12})$$

In the case of polarized, but otherwise incoherent statistical radiation, one finds a rate constant for radiative energy transfer between initial molecular quantum states i and final states f :

$$k_{\bar{f}i} = \frac{8\pi^3}{h^2} \frac{I_{\nu}^z}{(4\pi\epsilon_0)c} |M_{\bar{f}i}^z|^2 \quad (\text{A 3.13.13})$$

where $I_{\nu}^z = dI^z / d\nu$ is the intensity per frequency bandwidth of radiation and $M_{\bar{f}i}^z$ is the electric dipole transition moment in the direction of polarization. For unpolarized random spatial radiation of density $\rho(\nu)$ per volume and frequency, I_{ν}^z / c must be replaced by $\rho(\nu) / 3$, because of random orientation, and the rate of induced transitions (absorption or emission) becomes:

$$\begin{aligned} k_{\bar{f}i}^{\text{induced}} &= B_{\bar{f}i} \rho(\nu) \\ &= \frac{8\pi^3}{3h^2(4\pi\epsilon_0)} \rho(\nu) |M_{\bar{f}i}|^2. \end{aligned} \quad (\text{A 3.13.14})$$

$B_{\bar{f}i}$ is the Einstein coefficient for induced emission or absorption, which is approximately related to the absolute value of the dipole transition moment $|M_{\bar{f}i}|$, to the integrated cross section $G_{\bar{f}i}$ for the transition and to the Einstein coefficient $A_{\bar{f}i}$ for spontaneous emission [10]:

$$B_{\bar{f}i} = \frac{c}{h} G_{\bar{f}i} = \frac{c^3}{8\pi h\nu_{\bar{f}i}^3} A_{\bar{f}i} \quad (\text{A 3.13.15})$$

with

$$G_{\bar{f}i} = \int_{\text{line}} \sigma_{\bar{f}i}(\nu) \nu^{-1} d\nu \quad (\text{A 3.13.16})$$

-5-

and $\sigma_{\bar{f}i}(\nu)$ the frequency dependent absorption cross section. In [equation \(A3.13.15\)](#), $\nu_{\bar{f}i} = |E_f - E_i| / h$. Equation (A3.13.17) is a simple, useful formula relating the integrated cross section and the electric dipole transition moment as dimensionless quantities, in the electric dipole approximation [10, 100]:

$$\frac{G_{\bar{f}i}}{\text{pm}^2} \approx 41.624 \left| \frac{M_{\bar{f}i}}{\text{Debye}} \right|^2. \quad (\text{A 3.13.17})$$

From these equations one also finds the rate coefficient matrix for thermal radiative transitions including absorption, induced and spontaneous emission in a thermal radiation field following Planck's law [35]:

$$k_{\bar{f}i} = A_{\bar{f}i} \frac{\text{sign}(E_f - E_i)}{\exp((E_f - E_i)/k_B T) - 1}. \quad (\text{A 3.13.18})$$

Finally, if one has a condition with incoherent radiation of a small band width $\Delta\nu$ exciting a broad absorption band with $\sigma(\nu \pm \Delta\nu) \approx \sigma(\nu)$, one finds:

$$k_{fi}^{\text{induced}} = \frac{\sigma(\nu)}{h\nu} I \quad (\text{A 3.13.19})$$

where I is the radiation intensity. For a detailed discussion refer to [10]. The problem of coherent radiative excitation is considered in [section A3.13.4](#) and [section A3.13.5](#) in relation to intramolecular vibrational energy redistribution.

(d) The fourth mechanism is purely intramolecular energy redistribution. It is addressed in the next section.

A3.13.2.2 STRICTLY MONOMOLECULAR PROCESSES IN ISOLATED MOLECULES

Purely intramolecular energy transfer occurs when energy migrates within an isolated molecule from one part to another or from one type of motion to the other. Processes of this type include the vast field of molecular electronic radiationless transitions which emerged in the late 1960s [36], but more generally any type of intramolecular motion such as intramolecular vibrational energy redistribution (IVR) or intramolecular vibrational–rotational energy redistribution (IVRR) and related processes [37, 38 and 39]. These processes will be discussed in [section A3.13.5](#) in some detail in terms of their full quantum dynamics. However, in certain situations a statistical description with rate equations for such processes can be appropriate [38].

[Figure A3.13.1](#) illustrates our general understanding of intramolecular energy redistribution in isolated molecules and shows how these processes are related to ‘intermolecular’ processes, which may follow any of the mechanisms discussed in the previous section. The horizontal bars represent levels of nearly degenerate states of an isolated molecule.

-6-

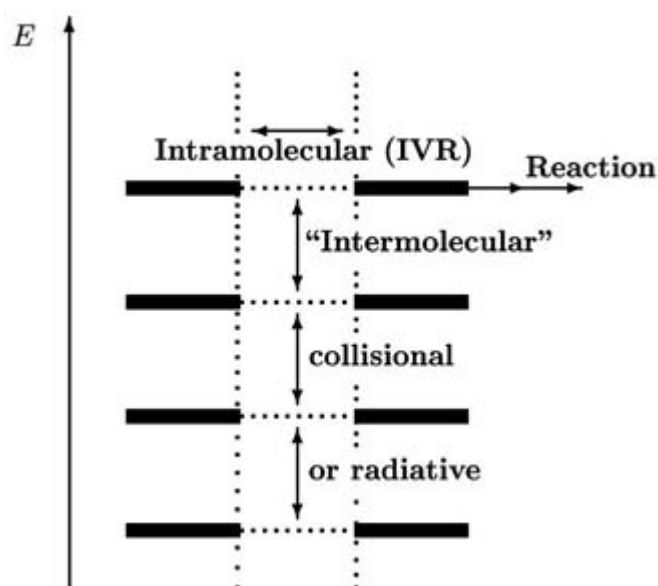


Figure A3.13.1. Schematic energy level diagram and relationship between ‘intermolecular’ (collisional or radiative) and intramolecular energy transfer between states of isolated molecules. The fat horizontal bars indicate thin energy shells of nearly degenerate states.

Having introduced the basic concepts and equations for various energy redistribution processes, we will now

discuss some of them in more detail.

A 3.13.3 COLLISIONAL ENERGY REDISTRIBUTION PROCESSES

A3.13.3.1 THE MASTER EQUATION FOR COLLISIONAL RELAXATION REACTION PROCESSES

The fundamental kinetic master equations for collisional energy redistribution follow the rules of the kinetic equations for all elementary reactions. Indeed an energy transfer process by inelastic collision, [equation \(A3.13.5\)](#), can be considered as a somewhat special ‘reaction’. The kinetic differential equations for these processes have been discussed in the general context of [chapter A3.4](#) on gas kinetics. We discuss here some special aspects related to collisional energy transfer in reactive systems. The general master equation for relaxation and reaction is of the type [[11](#), [12](#) and [13](#), [15](#), [25](#), [40](#), [41](#)]:

$$\frac{dc_j(t)}{dt} = F(\{c_k(t)\}) \quad (\text{A 3.13.20})$$

$$c_j(t = 0) = c_{j0}. \quad (\text{A 3.13.21})$$

-7-

The index j can label quantum states of the same or different chemical species. [Equation \(A3.13.20\)](#) corresponds to a generally stiff initial value problem [[42](#), [43](#)]. In matrix notation one may write:

$$\frac{dc(t)}{dt} = F[c(t)] \quad (\text{A 3.13.22})$$

$$c(t = 0) = c_0. \quad (\text{A 3.13.23})$$

There is no general, simple solution to this set of coupled differential equations, and thus one will usually have to resort to numerical techniques [[42](#), [43](#)] (see also [chapter A3.4](#)).

A3.13.3.2 THE MASTER EQUATION FOR COLLISIONAL AND RADIATIVE ENERGY REDISTRIBUTION UNDER CONDITIONS OF GENERALIZED FIRST-ORDER KINETICS

There is one special class of reaction systems in which a simplification occurs. If collisional energy redistribution of some reactant occurs by collisions with an excess of ‘heat bath’ atoms or molecules that are considered kinetically structureless, and if furthermore the reaction is either unimolecular or occurs again with a reaction partner M having an excess concentration, then one will have generalized first-order kinetics for populations p_j of the energy levels of the reactant, i.e. with

$$\frac{dp_j}{dt} = \sum_{k \neq j} (K'_{jk} p_k - K'_{kj} p_j) - k_j p_j \quad (\text{A 3.13.24})$$

$$\frac{dp}{dt} = \mathbf{K}p \quad (\text{A 3.13.25})$$

In equation (A3.13.24), k_j is the specific rate constant for reaction from level j , and K'_{jk} are energy transfer rate coefficients. With appropriate definition of a rate coefficient matrix \mathbf{K} one has, in matrix notation,

$$\frac{dp}{dt} = \mathbf{K}p \quad (\text{A 3.13.26})$$

where for $j \neq i$

$$K_{ji}(T) = \left(\frac{8k_B T}{\pi \mu} \right)^{1/2} [M] \int_0^\infty x \exp(-x) \sigma_{ji}(k_B T x) dx. \quad (\text{A 3.13.27})$$

-8-

(see [equation \(A3.13.9\)](#)) and

$$-K_{jj}(T) = k_j + \sum_{k \neq j} K_{kj}(T). \quad (\text{A3.13.28})$$

The master [equation \(A3.13.26\)](#) applies also, under certain conditions, to radiative excitation with rate coefficients for radiative energy transfer being given by [equation \(A3.13.13\)](#), [equation \(A3.13.14\)](#), [equation \(A3.13.15\)](#), [equation \(A3.13.16\)](#), [equation \(A3.13.17\)](#), [equation \(A3.13.18\)](#) and [equation \(A3.13.19\)](#), depending on the case, or else by more general equations [10]. Finally, the radiative and collisional rate coefficients may be considered together to be important at the same time in a given reaction system, if time scales for these processes are of the appropriate order of magnitude. The solution of [equation \(A3.13.26\)](#) is given by:

$$p(t) = \exp(\mathbf{K}t)p(0). \quad (\text{A 3.13.29})$$

This solution can be obtained explicitly either by matrix diagonalization or by other techniques (see [chapter A3.4](#) and [42, 43]). In many cases the discrete quantum level labels in [equation \(A3.13.24\)](#) can be replaced by a continuous energy variable and the populations by a population density $p(E)$, with replacement of the sum by appropriate integrals [11]. This approach can be made the starting point of useful analytical solutions for certain simple model systems [11, 19, 44, 45 and 46].

While the time dependent populations $p_j(t)$ may generally show a complicated behaviour, certain simple limiting cases can be distinguished and characterized by appropriate parameters:

(a) The long time steady state limit (formally $t \rightarrow \infty$) is described by the largest eigenvalue λ_1 of \mathbf{K} . Since all λ_j are negative, λ_1 has the smallest absolute value [35, 47]. In this limit one finds [47] (with the reactant fraction $F_R = \sum_j p_j$):

$$-\frac{d \ln(F_R(t))}{dt} = -\frac{d \ln(\sum_j p_j(t))}{dt} = k_{\text{uni}} = -\lambda_1. \quad (\text{A3.13.30})$$

Thus, this eigenvalue λ_1 determines the unimolecular steady-state reaction rate constant.

(b) The second largest eigenvalue λ_2 determines ideally the relaxation time towards this steady state, thus:

$$\tau_{\text{relax}}^{-1} = -\lambda_2. \quad (\text{A 3.13.31})$$

More generally, further eigenvalues must be taken into account in the relaxation process.

-9-

(c) It is sometimes useful to define an incubation time τ_{inc} by the limiting equation for steady state:

$$-\ln(F_{\text{R}}^{\text{st}}(t)) = -\lambda_1(t - \tau_{\text{inc}}). \quad (\text{A 3.13.32})$$

Figure A3.13.2 illustrates the origin of these quantities. Refer to [47] for a detailed mathematical discussion as well as the treatment of radiative laser excitation, in which incubation phenomena are important. Also refer to [11] for some classical examples in thermal systems.

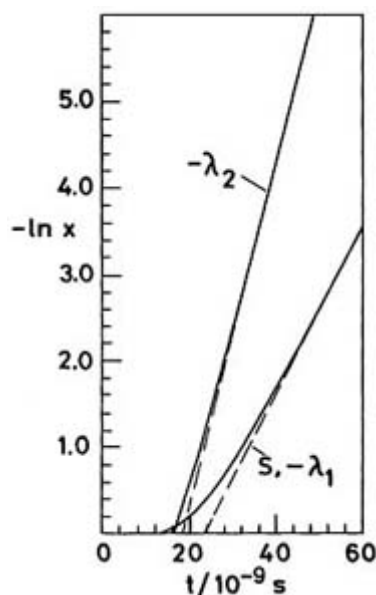


Figure A3.13.2. Illustration of the analysis of the master equation in terms of its eigenvalues λ_1 and λ_2 for the example of IR-multiphoton excitation. The dashed lines give the long time straight line limiting behaviour. The full line to the right-hand side is for $x = F_{\text{R}}(t)$ with a straight line of slope $-\lambda_1 = k_{\text{uni}}^{-1}$. The intercept of the corresponding dashed line (F_{R}^{st}) indicates τ_{inc} (see equation (A3.13.32)). The left-hand line is for $x = |F_{\text{R}} - F_{\text{R}}^{\text{st}}|$ with limiting slope $-\lambda_2 = \tau_{\text{relax}}^{-1}$ (see text and [47]).

As a rule, in thermal unimolecular reaction systems at modest temperatures, λ_1 is well separated from the other eigenvalues, and thus the time scales for incubation and ‘relaxation’ are well separated from the steady-state reaction time scale $\tau_{\text{reaction}} = k_{\text{uni}}^{-1}$. On the other hand, at high temperatures, k_{uni} , τ_{relax}^{-1} and τ_{inc}^{-1} may merge. This is illustrated in figure A3.13.3 for the classic example of thermal unimolecular dissociation [48, 49, 50 and 51]:



Note that in the ‘low pressure limit’ of unimolecular reactions ([chapter A3.4](#)), the unimolecular rate constant k_{uni} is entirely dominated by energy transfer processes, even though the relaxation and incubation rates (τ_{relax}^{-1} and τ_{inc}^{-1}) may be much faster than k_{uni} .

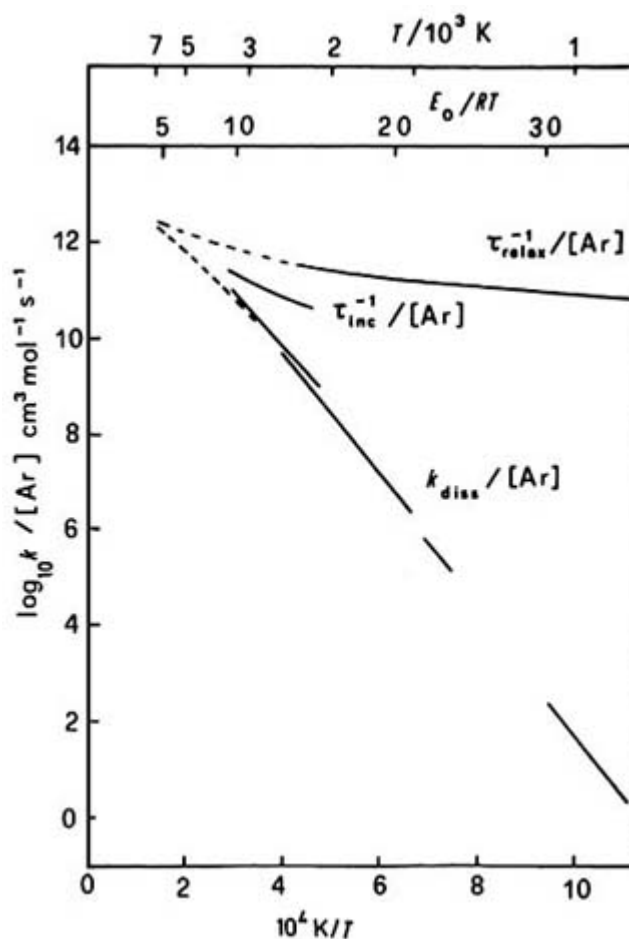


Figure A3.13.3. Dissociation ($k_{\text{uni}} = k_{\text{diss}}$), incubation (τ_{inc}^{-1}) and relaxation (τ_{relax}^{-1}) rate constants for the reaction $N_2O \rightarrow N_2 + O$ at low pressure in argon (from [11], see discussion in the text for details and references to the experiments).

The master equation treatment of energy transfer in even fairly complex reaction systems is now well established and fairly standard [52]. However, the rate coefficients k_{ij} for the individual energy transfer processes must be established and we shall discuss some aspects of this matter in the following section.

A3.13.3.3 MECHANISMS OF COLLISIONAL ENERGY TRANSFER

Collisional energy transfer in molecules is a field in itself and is of relevance for kinetic theory ([chapter A3.1](#)), gas phase kinetics ([chapter A3.4](#)), RRKM theory ([chapter A3.12](#)), the theory of unimolecular reactions in general,

as well as the kinetics of laser systems [53]. [Chapter C3.3](#), [Chapter C3.4](#) and [Chapter C3.5](#) treat these subjects in detail. We summarize those aspects that are of importance for mechanistic considerations in chemically reactive systems.

We start from a model in which collision cross sections or rate constants for energy transfer are compared with a reference quantity such as average Lennard-Jones collision cross sections or the usually cited Lennard-Jones collision frequencies [54]

$$Z_{LJ} = \pi \sigma_{AB}^2 \left(\frac{8k_B T}{\pi \mu_{AB}} \right)^{1/2} \Omega_{AB}^{(2,2)*} \quad (\text{A3.13.34})$$

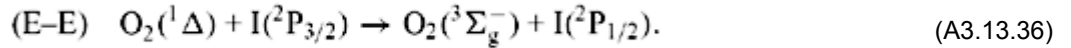
where σ_{AB} is the Lennard-Jones parameter and $\Omega_{AB}^{(2,2)*}$ is the reduced collision integral [54], calculated from the binding energy ϵ and the reduced mass μ_{AB} for the collision in the Lennard-Jones potential

$$V(r) = 4\epsilon \left[\left(\frac{\sigma_{AB}}{r} \right)^{12} - \left(\frac{\sigma_{AB}}{r} \right)^6 \right]. \quad (\text{A3.13.35})$$

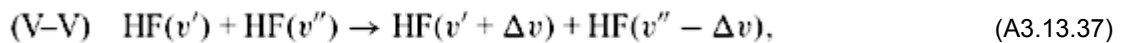
Given such a reference, we can classify various mechanisms of energy transfer either by the probability that a certain energy transfer process will occur in a ‘Lennard-Jones reference collision’, or by the average energy transferred by one ‘Lennard-Jones collision’.

With this convention, we can now classify energy transfer processes either as resonant, if $|\Delta E_t|$ defined in [equation \(A3.13.8\)](#) is small, or non-resonant, if it is large. Quite generally the rate of resonant processes can approach or even exceed the Lennard-Jones collision frequency (the latter is possible if other long-range potentials are actually applicable, such as by permanent dipole–dipole interaction).

Resonant processes of some importance include resonant electronic to electronic energy transfer (E–E), such as the pumping process of the iodine atom laser



Another near resonant process is important in the hydrogen fluoride laser, [equation \(A3.13.37\)](#), where vibrational to vibrational energy transfer is of interest:



where Δv is the number of vibrational quanta exchanged. If HF were a harmonic oscillator, ΔE_t would be zero (perfect resonance). In practice, because of anharmonicity, the most important process is exothermic, leading to increasing excitation v' of some of the HF molecules with successive collisions [55, 56], because the exothermicity drives this process to high v' as long as plenty of HF(v'') with low v'' are available.

Resonant rotational to rotational (R–R) energy transfer may have rates exceeding the Lennard-Jones collision frequency because of long-range dipole–dipole interactions in some cases. Quasiresonant vibration to rotation transfer (V–R) has recently been discussed in the framework of a simple model [57].

‘Non-resonant’ processes include vibration–translation (V–T) processes with transfer probabilities decreasing to very small values for diatomic molecules with very high vibrational frequencies, of the order of 10^{-4} and less for the probability of transferring a quantum in a collision. Also, vibration to rotation (V–R) processes frequently have low probabilities, of the order of 10^{-2} , if ΔE_t is relatively large. Rotation to translation (R–T)

processes are generally fast, with probabilities near 1. Also, the R–V–T processes in collisions of large polyatomic molecules have high probabilities, with average energies transferred in one Lennard-Jones collision being of the order of a few kJ mol^{-1} [11, 25], or less in collisions with rare gas atoms. As a general rule one may assume collision cross sections to be small, if ΔE_t is large [11, 58, 59].

In the experimental and theoretical study of energy transfer processes which involve some of the above mechanisms, one should distinguish processes in atoms and small molecules and in large polyatomic molecules. For small molecules a full theoretical quantum treatment is possible and even computer program packages are available [60, 62 and 63], with full state to state characterization. A good example are rotational energy transfer theory and experiments on $\text{He} + \text{CO}$ [64]:



On the experimental side, small molecule energy transfer experiments may use molecular beam techniques [65, 66 and 67] (see also [chapter C3.3](#) for laser studies).

In the case of large molecules, instead of the detailed quantum state characterization implied in the cross sections σ_{fi} and rate coefficients K_{fi} of the master [equation \(A3.13.24\)](#), one derives more coarse grained information on ‘levels’ covering a small energy bandwidth around E and E' (with an optional notation $K_{fi}(E', E)$) or finally energy transfer probabilities $P(E', E)$ for a transition from energy E to energy E' in a highly excited large polyatomic molecule where the density of states $\rho(E')$ is very large, for example in a collision with a heat bath inert gas atom [11]. Such processes can currently be modelled by classical trajectories [68, 69 and 70].

Experimental access to the probabilities $P(E', E)$ for energy transfer in large molecules usually involves techniques providing just the first moment of this distribution, i.e. the average energy $\langle \Delta E \rangle$ transferred in a collision. Such methods include UV absorption, infrared fluorescence and related spectroscopic techniques [11, 28, 71, 72, 73 and 74]. More advanced techniques, such as kinetically controlled selective ionization (KCSI [74]) have also provided information on higher moments of $P(E', E)$, such as $\langle (\Delta E)^2 \rangle$.

The standard mechanisms of collisional energy transfer for both small and large molecules have been treated extensively and a variety of scaling laws have been proposed to simplify the complicated body of data [58, 59, 75]. To conclude, one of the most efficient special mechanisms for energy transfer is the quasi-reactive process involving chemically bound intermediates, as in the example of the reaction:



-13-

Such processes transfer very large amounts of energy in one collision and have been treated efficiently by the statistical adiabatic channel model [11, 19, 30, 76, 77, 78 and 79]. They are quite similar mechanistically to chemical activation systems. One might say that in such a mechanism one may distinguish three phases:

(a) Formation of a bound collision complex AB:



(b) IVRR in this complex:



(c) Finally, dissociation of the internally, statistically equilibrated complex:



That is, rapid IVR in the long lived intermediate is an essential step. We shall treat this important process in the next section, but mention here in passing the observation of so-called ‘supercollisions’ transferring large average amounts of energy $\langle \Delta E \rangle$ in one collision [80], even if intermediate complex formation may not be important.

A 3.13.4 INTRAMOLECULAR ENERGY TRANSFER STUDIES IN POLYATOMIC MOLECULES

In this section we review our understanding of IVR as a special case of intramolecular energy transfer. The studies are based on calculations of the time evolution of vibrational wave packets corresponding to middle size and large amplitude vibrational motion in polyatomic molecules. An early example for the investigation of wave packet motion as a key to understanding IVR and its implication on reaction kinetics using experimental data is given in [81]. Since then, many other contributions have helped to increase our knowledge using realistic potential energy surfaces, mainly for two- and three-dimensional systems, and we give a brief summary of these results below.

A3.13.4.1 IVR AND CLASSICAL MECHANICS

Before undergoing a substantial and, in many cases, practically irreversible, change of geometrical structure within a chemical reaction, a molecule may often perform a series of vibrations in the multidimensional space around its equilibrium structure. This applies in general to reactions that take place entirely in the bound electronic ground state and in many cases to reactions that start in the electronic ground state near the equilibrium structure, but evolve into highly excited states above the reaction threshold energy. In the latter case, within the general scheme of [equation \(A3.13.1\)](#) a reaction is thought to be induced by a sufficiently energetic pulse of electromagnetic radiation or by collisions with adequate high-energy collision partners. In the first case, a reaction is thought to be the last step after

-14-

a chain of excitation steps has transferred enough energy into the molecule to react either thermally, by collisions, or coherently, for instance by irradiation with infrared laser pulses. These pulses can be tuned to adequately excite vibrations along the reaction coordinate, the amplitudes of which become gradually larger until the molecule undergoes a sufficiently large structural change leading to the chemical reaction.

Vibrational motion is thus an important primary step in a general reaction mechanism and detailed investigation of this motion is of utmost relevance for our understanding of the dynamics of chemical reactions. In classical mechanics, vibrational motion is described by the time evolution $\vec{q}(t)$ and $\vec{p}(t)$ of general internal position and momentum coordinates. These time dependent functions are solutions of the classical equations of motion, e.g. Newton’s equations for given initial conditions $\vec{q}(t_0) = q_0$ and $\vec{p}(t_0) = p_0$. The definition of initial conditions is generally limited in precision to within experimental uncertainties Δq_0 and Δp_0 , more fundamentally related by the Heisenberg principle $\Delta q_0 \Delta p_0 \gtrsim h/4\pi$. Therefore, we need to

consider an initial distribution $F_0(q-q_0, p-p_0)$, with widths Δq_0 and Δp_0 and the time evolution $F_t(q-\tilde{q}(t), p-\tilde{p}(t))$, which may be quite different from the initial distribution F_0 , depending on the integrability of the dynamical system. Ideally, for classical, integrable systems, vibrational motion may be understood as the motion of narrow, well localized distributions $F(q-\tilde{q}(t), p-\tilde{p}(t))$ (ideally δ -functions in a strict mathematical sense), centred around the solutions of the classical equations of motion. In this picture we wish to consider initial conditions that correspond to localized vibrational motion along specific manifolds, for instance a vibration that is induced by elongation of a single chemical bond (local mode vibrations) as a result of the interaction with some external force, but it is also conceivable that a large displacement from equilibrium might be induced along a single normal coordinate. Independent of the detailed mechanism for the generation of localized vibrations, harmonic transfer of excitation may occur when such a vibration starts to extend into other manifolds of the multidimensional space, resulting in trajectories that draw Lissajous figures in phase space, and also in configuration space [82] (see also [83]). Furthermore, if there is anharmonic interaction, IVR may occur. In [84, 85] this type of IVR was called classical intramolecular vibrational redistribution (CIVR).

A3.13.4.2 IVR AND QUANTUM MECHANICS

In time-dependent quantum mechanics, vibrational motion may be described as the motion of the wave packet $|\psi(q, t)\rangle^2$ in configuration space, e.g. as defined by the possible values of the position coordinates q . This motion is given by the time evolution of the wave function $\psi(q, t)$, defined as the projection $\langle q | \psi(t) \rangle$ of the time-dependent quantum state $|\psi(t)\rangle$ on configuration space. Since the quantum state is a complete description of the system, the wave packet defining the probability density can be viewed as the quantum mechanical counterpart of the classical distribution $F(q-\tilde{q}(t), p-\tilde{p}(t))$. The time dependence is obtained by solution of the time-dependent Schrödinger equation

$$i \frac{\hbar}{2\pi} \frac{d|\psi(t)\rangle}{dt} = \hat{H}|\psi(t)\rangle \quad (\text{A 3.13.43})$$

where \hbar is the Planck constant and \hat{H} is the Hamiltonian of the system under consideration. Solutions depend on initial conditions $|\psi(t_0)\rangle$ and may be formulated using the time evolution operator $\hat{U}(t, t_0)$:

$$|\psi(t)\rangle = \hat{U}(t, t_0)|\psi(t_0)\rangle. \quad (\text{A 3.13.44})$$

-15-

Alternatively, in the case of incoherent (e.g. statistical) initial conditions, the density matrix operator $\hat{P}(t) = |\psi(t)\rangle \langle \psi(t)|$ at time t can be obtained as the solution of the Liouville–von Neumann equation:

$$\hat{P}(t) = \hat{U}(t, t_0) \hat{P}(t_0) \hat{U}^\dagger(t, t_0) \quad (\text{A 3.13.45})$$

where $\hat{U}^\dagger(t, t_0)$ is the adjoint of the time evolution operator (in strictly conservative systems, the time evolution operator is unitary and $\hat{U}^\dagger(t, t_0) = \hat{U}^{-1}(t, t_0) = \hat{U}(t_0, t)$).

The calculation of the time evolution operator in multidimensional systems is a formidable task and some results will be discussed in this section. An alternative approach is the calculation of semi-classical dynamics as demonstrated, among others, by Heller [86, 87 and 88], Marcus [89, 90], Taylor [91, 92], Metiu [93, 94] and coworkers (see also [83] as well as the review by Miller [95] for more general aspects of semiclassical dynamics). This method basically consists of replacing the δ -function distribution in the true classical calculation by a Gaussian distribution in coordinate space. It allows for a simulation of the vibrational

quantum dynamics to the extent that interference effects in the evolving wave packet can be neglected. While the application of semi-classical methods might still be of some interest for the simulation of quantum dynamics in large polyatomic molecules in the near future, as a natural extension of classical molecular dynamics calculations [68, 96], full quantum mechanical calculations of the wave packet evolution in smaller polyatomic molecules are possible with the currently available computational resources. Following earlier spectroscopic work and three-dimensional quantum dynamics results [81, 97, 98, 99 and 100], Wyatt and coworkers have recently demonstrated applications of full quantum calculations to the study of IVR in fluoroform, with nine degrees of freedom [101, 102] and in benzene [103], considering all 30 degrees of freedom [104]. Such calculations show clearly the possibilities in the computational treatment of quantum dynamics and IVR. However, remaining computational limitations restrict the study to the lower energy regime of molecular vibrations, when all degrees of freedom of systems with more than three dimensions are treated. Large amplitude motion, which shows the inherently quantum mechanical nature of wave packet motion and is highly sensitive to IVR, cannot yet be discussed for such molecules, but new results are expected in the near future, as indicated in recent work on ammonia [105, 106], formaldehyde and hydrogen peroxide [106, 107 and 108], and hydrogen fluoride dimer [109, 110 and 111] including all six internal degrees of freedom.

A key feature in quantum mechanics is the dispersion of the wave packet, i.e. the loss of its Gaussian shape. This feature corresponds to a delocalization of probability density and is largely a consequence of anharmonicities of the potential energy surface, both the ‘diagonal’ anharmonicity, along the manifold in which the motion started, and ‘off diagonal’, induced by anharmonic coupling terms between different manifolds in the Hamiltonian. Spreading of the wave packet into different manifolds is thus a further important feature of IVR. In [84, 85] this type of IVR was called delocalization quantum intramolecular vibrational redistribution (DIVR). DIVR plays a central role for the understanding of statistical theories for unimolecular reactions in polyatomic molecules [84, 97], as will be discussed below.

A3.13.4.3 IVR WITHIN THE GENERAL SCHEME OF ENERGY REDISTRIBUTION IN REACTIVE SYSTEMS

As in classical mechanics, the outcome of time-dependent quantum dynamics and, in particular, the occurrence of IVR in polyatomic molecules, depends both on the Hamiltonian and the initial conditions, i.e. the initial quantum mechanical state $|\psi(t_0)\rangle$. We focus here on the time-dependent aspects of IVR, and in this case such initial conditions always correspond to the preparation, at a time t_0 , of superposition states of molecular (spectroscopic) eigenstates involving at least two distinct vibrational energy levels. Strictly, IVR occurs if these levels involve at least two distinct

-16-

vibrational manifolds in terms of which the total (vibrational) Hamiltonian is not separable [84]. In a time-independent view, this requirement states that the wave functions belonging to the two spectroscopic states are spread in a non-separable way over the configuration space spanned by at least two different vibrational modes. The conceptual framework for the investigation of IVR may be sketched within the following scheme, which also mirrors the way we might investigate IVR in the time-dependent approach, both theoretically and experimentally:

$$|\psi(t_{-1})\rangle \xrightarrow{\hat{U}_{\text{prep}}(t_0, t_{-1})} |\psi(t_0)\rangle \xrightarrow{\hat{U}_{\text{free}}(t, t_0)} |\psi(t)\rangle. \quad (\text{A 3.13.46})$$

In a first time interval $[t_{-1}, t_0]$ of the scheme (A3.13.46), a superposition state is prepared. This step corresponds to the step in [equation \(A3.13.1\)](#). One might think of a time evolution $|\psi(t_{-1})\rangle \rightarrow |\psi(t_0)\rangle = \hat{U}_{\text{prep}}(t_0, t_{-1})|\psi(t_{-1})\rangle$, where $|\psi(t_{-1})\rangle$ may be a molecular eigenstate and \hat{U}_{prep} is the time evolution operator obtained from the interaction with an external system, to be specified below. The probability distribution $|\psi(q, t_n)|^2$ is expected to be approximatively localized in configuration space, such that $|\psi(q^*, t_n)|^2 > 0$ for

position coordinates $q^* \in \mathcal{M}^*$ belonging to some specific manifold \mathcal{M}^* and $|\psi(q, t_0)|^2 \approx 0$ for coordinates $q \in \mathcal{M}$ belonging to the complementary manifold $\mathcal{M} = \overline{\mathcal{M}^*}$. In a second time interval $[t_0, t_1]$, the superposition state $|\psi(t_0)\rangle$ has a free evolution into states $|\psi(t)\rangle = \hat{U}_{\text{free}}(t, t_0)|\psi(t_0)\rangle$. This step corresponds to the intermediate step equation (A3.13.47), occurring between the steps described before by equation (A3.13.1) and equation (A3.13.2) (see also equation (A3.13.41)):

$$\mathbf{R}^* \rightarrow \mathbf{R}^{**}. \quad (\text{A } 3.13.47)$$

IVR is present if $|\psi(q, t)|^2 > 0$ is observed for $t > t_0$ also for $q \in \mathcal{M}$. IVR may of course also occur during the excitation process, if its time scale is comparable to that of the excitation.

In the present section, we concentrate on coherent preparation by irradiation with a properly chosen laser pulse during a given time interval. The quantum state at time t_{-1} may be chosen to be the vibrational ground state $|\phi_0^{(g)}\rangle$ in the electronic ground state. In principle, other possibilities may also be conceived for the preparation step, as discussed in section A3.13.1, section A3.13.2 and section A3.13.3. In order to determine superposition coefficients within a realistic experimental set-up using irradiation, the following questions need to be answered: (1) What are the eigenstates? (2) What are the electric dipole transition matrix elements? (3) What is the orientation of the molecule with respect to the laboratory fixed (linearly or circularly) polarized electric field vector of the radiation? The first question requires knowledge of the potential energy surface, or the Hamiltonian $\hat{H}_0(p, q)$ of the isolated molecule, the second that of the vector valued surface $\vec{\mu}(q)$ of the electric dipole moment. This surface yields the operator, which couples spectroscopic states by the impact of an external irradiation field and thus directly affects the superposition procedure. The third question is indeed of great importance for comparison with experiments aiming at the measurement of internal wave packet motion in polyatomic molecules and has recently received much attention in the treatment of molecular alignment and orientation [112, 113], including non-polar molecules [114, 115]. To the best of our knowledge, up to now explicit calculations of multidimensional wave packet evolution in polyatomic molecules have been performed upon neglect of rotational degrees of freedom, i.e. only internal coordinates have been considered, although calculations on coherent excitation in ozone level structures with rotation exist [116, 117], which could be interpreted in terms of wave packet evolution. A more detailed discussion of this point will be given below for a specific example.

-17-

A3.13.4.4 CONCEPTS OF COMPUTATIONAL METHODS

There are numerous methods for solving the time dependent Schrödinger equation (A3.13.43), and some of them were reviewed by Kosloff [118] (see also [119, 120]). Whenever projections of the evolving wave function on the spectroscopic states are useful for the detailed analysis of the quantum dynamics (and this is certainly the case for the detailed analysis of IVR), it is convenient to express the Hamiltonian based on spectroscopic states $|\phi_n\rangle$:

$$\hat{H}_0 = \sum_n \frac{\hbar}{2\pi} \omega_n |\phi_n\rangle \langle \phi_n| \quad (\text{A3.13.48})$$

where ω_n are the eigenfrequencies. For an isolated molecule $\hat{H} = \hat{H}_0$ in equation (A3.13.43) and the time evolution operator is of the form

$$(\text{A3.13.49})$$

$$\hat{U}(t, t_0) = \sum_n \exp(-i\omega_n(t - t_0)) |\phi_n\rangle \langle \phi_n|.$$

The time-dependent wave function is then given by the expression:

$$\psi(q, t) = \sum_n c_n^0 \exp(-i\omega_n t) \phi_n(q). \quad (\text{A3.13.50})$$

Here, $\phi_n(q) = \langle q | \phi_n \rangle$ are the wave functions of the spectroscopic states and the coefficients c_n^0 are determined from the initial conditions

$$\psi(q, t_0) = \sum_n c_n^0 \phi_n(q), \quad c_n^0 = \langle \phi_n | \psi(t_0) \rangle. \quad (\text{A3.13.51})$$

Equation (A3.13.49) describes the spectroscopic access to quantum dynamics. Clearly, when the spectral structure becomes too congested, i.e. when there are many close lying frequencies ω_n , calculation of all spectroscopic states becomes difficult. However, often it is not necessary to calculate all states when certain model assumptions can be made. One assumption concerns the separation of time scales. When there is evidence for a clear separation of time scales for IVR, only part of the spectroscopic states need to be considered for fast evolution. Typically, these states have large frequency separations, and considering only such states means neglecting the fine-grained spectral structure as a first approximation. An example for separation of time scales is given by the dynamics of the alkyl CH chromophore in CHXYZ compounds, which will be discussed below. This group span a three-dimensional linear space of stretching and bending vibrations. These vibrations are generally quite strongly coupled, which is manifested by the occurrence of a Fermi resonance in the spectral structure throughout the entire vibrational energy space. As we will see, the corresponding time evolution and IVR between these modes takes place in less than 1 ps, while other modes become involved in the dynamics on much longer time scales (10 ps to ns, typically). The assumption for time scale separation and IVR on the subpicosecond time scale for the alkyl CH chromophore was founded on the basis of

-18-

spectroscopic data nearly 20 years ago [98, 121]. The first results on the nature of IVR in the CH chromophore system and its role in IR photochemistry were also reported by that time [122, 123], including results for the acetylenic CH chromophore [124] and results obtained from first calculations of the wave packet motion [81]. The validity of this assumption has recently been confirmed in the case of CHF_3 both experimentally, from the highly resolved spectral structure of highly excited vibrational overtones [125, 126], and theoretically, including all nine degrees of freedom for modestly excited vibrational overtones up to 6000 cm^{-1} [102].

A3.13.4.5 IVR DURING AND AFTER COHERENT EXCITATION: GENERAL ASPECTS

Modern photochemistry (IR, UV or VIS) is induced by coherent or incoherent radiative excitation processes [4, 5, 6 and 7]. The first step within a photochemical process is of course a preparation step within our conceptual framework, in which time-dependent states are generated that possibly show IVR. In an ideal scenario, energy from a laser would be deposited in a spatially localized, large amplitude vibrational motion of the reacting molecular system, which would then possibly lead to the cleavage of selected chemical bonds. This is basically the central idea behind the concepts for a ‘mode selective chemistry’, introduced in the late 1970s [127], and has continuously received much attention [10, 117, 122, 128, 129, 130, 131, 132, 133, 134

and 135]. In a recent review [136], IVR was interpreted as a ‘molecular enemy’ of possible schemes for mode selective chemistry. This interpretation is somewhat limited, since IVR represents more complex features of molecular dynamics [37, 84, 134], and even the opposite situation is possible. IVR can indeed be selective with respect to certain structural features [85, 97] that may help mode selective reactive processes after tailored laser excitation [137].

To be more specific, we assume that for a possible preparation step the Hamiltonian might be given during the preparation time interval $[t_{-1}, t_0]$ by the expression:

$$\hat{H} = \hat{H}_0 + \hat{H}_1(t) \quad (\text{A 3.13.52})$$

where \hat{H}_0 is the Hamiltonian of the isolated molecule and \hat{H}_1 is the interaction Hamiltonian between the molecule and an external system. In this section, we limit the discussion to the case where the external system is the electromagnetic radiation field. For the interaction with a classical electromagnetic field with electric field vector $\vec{E}(t)$, the interaction Hamiltonian is given by the expression:

$$\hat{H}_1(t) = -\hat{\mu}\vec{E}(t). \quad (\text{A 3.13.53})$$

where $\hat{\mu}$ is the operator of the electric dipole moment. When we treat the interaction with a classical field in this way, we implicitly assume that the field will remain unaffected by the changes in the molecular system under consideration. More specifically, its energy content is assumed to be constant. The energy of the radiation field is thus not explicitly considered in the expression for the total Hamiltonian and all operators acting on states of the field are replaced by their time-dependent expectation values. These assumptions are widely accepted, whenever the number of photons in each field mode is sufficiently large. For a coherent, monochromatic, polarized field with intensity $I = \sqrt{\epsilon_0/\mu_0}|\vec{E}|^2 \approx 1 \text{ MW cm}^{-2}$ *in vacuo*, which is a typical value used in laser chemical experiments in the gas phase at low pressures, the number N_v of mid infrared photons existing in a cavity of volume $V = 1 \text{ m}^3$ is [138, p498]:

-19-

$$N_v = \frac{IV}{c_0 h \nu} \approx 10^{21}. \quad (\text{A 3.13.54})$$

Equation (A3.13.54) legitimates the use of this semi-classical approximation of the molecule–field interaction in the low-pressure regime. Since $\hat{H}_1(t)$ is explicitly time dependent, the time evolution operator is more complicated than in equation A3.13.49. However, the time-dependent wave function can still be written in the form

$$\psi(q, t) = \sum_n c_n(t) \phi_n(q) \quad (\text{A 3.13.55})$$

with time-dependent coefficients that are obtained by solving the set of coupled differential equations

$$i \frac{dc_n(t)}{dt} = \sum_{n'} \{W_{nn'} + V_{nn'}(t)\} c_{n'}(t) \quad (\text{A 3.13.56})$$

where $W_{nn'} = \delta_{nn'} \omega_n$ ($\delta_{nn'}$ is the Kronecker symbol, ω_n were defined in equation (A3.13.48)) and

$$\begin{aligned}
V_{nn'}(t) &= \frac{2\pi}{h} \langle \phi_n | \hat{H}_i(t) | \phi_{n'} \rangle \\
&= -\frac{2\pi}{h} \langle \phi_n | \hat{\mu} | \phi_{n'} \rangle \vec{E}(t).
\end{aligned}
\tag{A 3.13.57}$$

The matrix elements $\langle \phi_n | \hat{\mu} | \phi_{n'} \rangle$ are multidimensional integrals $\int \phi_n^*(q) \vec{\mu}(q) \phi_{n'}(q) d\mathbf{r}$ of the vector valued dipole moment surface. The time-independent part of the coupling matrix elements in equation (A3.13.57) can also be cast into the practical formula

$$V_{nn'}^0 / (2\pi c_0 \text{ cm}^{-1}) = -0.46093 \langle \phi_n | \hat{\mu}_\alpha | \phi_{n'} \rangle \sqrt{I_0 / \text{MW cm}^{-2}},
\tag{A 3.13.58}$$

where α is the direction of the electric field vector of the linearly polarized radiation field with maximal intensity I_0 . The solution of equation (A3.13.56) may still be quite demanding, depending on the size of the system under consideration. However, it has become a practical routine procedure to use suitable approximations such as the QRA (quasiresonant approximation) or Floquet treatment [35, 122, 129] and programmes for the numerical solution are available [139, 140].

A3.13.4.6 ELECTRONIC EXCITATION IN THE FRANCK–CONDON LIMIT AND IVR

At this stage we may distinguish between excitation involving different electronic states and excitation occurring within the same electronic (ground) state. When the spectroscopic states are located in different electronic states, say the ground (g) and excited (e) states, one frequently assumes the Franck–Condon approximation to be applicable:

$$\langle \phi_n^{(g)} | \hat{\mu} | \phi_{n'}^{(e)} \rangle \approx \vec{\mu}_{ge} \langle \phi_n^{(g)} | \phi_{n'}^{(e)} \rangle.
\tag{A 3.13.59}$$

Such electronic excitation processes can be made very fast with sufficiently intense laser fields. For example, if one considers monochromatic excitation with a wavenumber in the UV region ($60\,000 \text{ cm}^{-1}$) and a coupling strength $(\vec{\mu}_{ge} \vec{E}) / hc \approx 4000 \text{ cm}^{-1}$ (e.g. $\mu_{ge} \approx 1 \text{ Debye}$ in equation (A3.13.59), $I \approx 50 \text{ TW cm}^{-2}$), excitation occurs within 1 fs [141]. During such a short excitation time interval the relative positions of the nuclei remain unchanged (Franck approximation). Within these approximations, if one starts the preparation step in the vibrational ground state $|\phi_0^g\rangle$, the resulting state $|\psi(t_0)\rangle$ at time t_0 has the same probability distribution as the vibrational ground state. However, it is now transferred into the excited electronic state where it is no longer stationary, since it is a superposition state of vibrational eigenstates in the excited electronic state:

$$|\psi(t_0)\rangle = \sum_n \langle \phi_n^{(e)} | \phi_0^g \rangle |\phi_n^{(e)}\rangle.
\tag{A 3.13.60}$$

Often the potential energy surfaces for the ground and excited states are fairly different, i.e. with significantly different equilibrium positions. The state $|\psi(t_0)\rangle$ will then correspond to a wave packet, which has nearly a Gaussian shape with a centre position that is largely displaced from the minimal energy configuration on the excited surface and, since the Franck approximation can be applied, the expectation value of the nuclear linear momentum vanishes. In a complementary view, the superposition state of equation (A3.13.60) defines the manifold \mathcal{M}^* in configuration space. It is often referred to as the ‘bright’ state, since its probability density defines a region in configuration space, the Franck–Condon region, which has been reached by the irradiation

field through mediation by the electric dipole operator. After the preparation step, the wave packet most likely starts to move along the steepest descent path from the Franck–Condon region. One possibility is that it proceeds to occupy other manifolds, which were not directly excited. The occupation of the remaining, ‘dark’ manifolds (e.g. $\overline{\mathcal{M}^*}$) by the time-dependent wave packet is a characteristic feature of IVR.

Studies of wave packet motion in excited electronic states of molecules with three and four atoms were conducted by Schinke, Engel and collaborators, among others, mainly in the context of photodissociation dynamics from the excited state [142, 143 and 144] (for an introduction to photodissociation dynamics, see [7], and also more recent work [145, 146, 147, 148 and 149] with references cited therein). In these studies, the dissociation dynamics is often described by a time-dependent displacement of the Gaussian wave packet in the multidimensional configuration space. As time goes on, this wave packet will occupy different manifolds (from where the molecule possibly dissociates) and this is identified with IVR. The dynamics may be described within the Gaussian wave packet method [150], and the vibrational dynamics is then of the classical IVR type (CIVR [84]). The validity of this approach depends on the dissociation rate on the one hand, and the rate of delocalization of the wave packet on the other hand. The occurrence of DIVR often receives less attention in the discussions of photodissociation dynamics mentioned above. In [148], for instance, details of the wave packet motion by means of snapshots of the probability density are missing, but a delocalization of the wave packet probably takes place, as may be concluded from inspection of [figure 5](#) therein.

A 3.13.5 IVR IN THE ELECTRONIC GROUND STATE: THE EXAMPLE OF THE CH CHROMOPHORE

A3.13.5.1 REDISTRIBUTION DURING AND AFTER COHERENT EXCITATION

A system that shows IVR with very fast spreading of the wave packet, i.e. DIVR in the subpicosecond time range, is that of the infrared alkyl CH chromophore, which will be used in the remaining part of this chapter to discuss IVR as a result of a mode specific excitation within the electronic ground state. The CH stretching and bending modes of the alkyl CH chromophore in CHXYZ compounds are coupled by a generally strong Fermi resonance [100, 151]. [Figure A3.13.4](#) shows the shape of the potential energy surface for the symmetrical compound CHD_3 as contour line representations of selected one- and two-dimensional sections (see figure caption for a detailed description). The important feature is the curved shape of the $V(Q_s, Q_b)$ potential section ($V(Q_s, Q_b)$ being similarly curved), which indicates a rather strong anharmonic coupling. This feature is characteristic for compounds of the type CHXYZ [84, 100, 151, 152 and 153]. Q_s , Q_b and Q_{b_2} are (mass weighted) normal coordinates of the CH stretching and bending motion, with symmetry A_1 and E, respectively, in the C_{3v} point group of symmetrical CHD_3 . A change of Q_s is a concerted motion of all atoms along the z -axis, defined in [figure A3.13.5](#). However, displacements along Q_s are small for the carbon and deuterium atoms, and large for the hydrogen atom. Thus, this coordinate essentially describes a stretching motion of the CH bond (along the z -axis). In the same way, Q_b and Q_{b_2} describe bending motions of the CH bond along the x - and y -axis, respectively (see [figure A3.13.5](#)). In the one-dimensional sections the positions of the corresponding spectroscopic states are drawn as horizontal lines. On the left-hand side, in the potential section $V(Q_{b_2})$, a total of 800 states up to an energy equivalent wave number of $25\,000\text{ cm}^{-1}$ has been considered. These energy levels may be grouped into semi-isoenergetic shells defined by multiplets of states with a constant chromophore quantum number $N = v_s + \frac{1}{2}v_b = 0, \frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}, \dots$, where v_s and v_b are quantum numbers of effective basis states (‘Fermi modes’ [97, 152, 154]) that are strongly coupled by a 2:1 Fermi resonance. These multiplets give rise to spectroscopic polyads and can be well distinguished in the lower energy region, where the density of states is low.

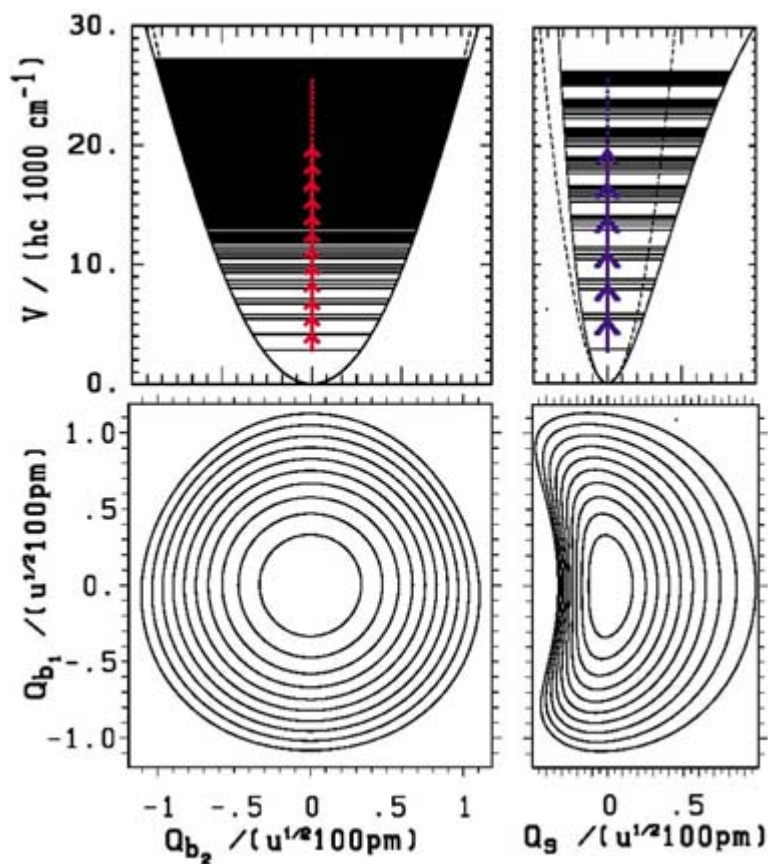


Figure A3.13.4. Potential energy cuts along the normal coordinate subspace pertaining to the CH chromophore in CHD_3 . Q_{b1} is the A' coordinate in C_s symmetry, essentially changing structure along the x -axis see also [Figure A3.13.5](#), and Q_{b2} is the A'' coordinate, essentially changing structure along the y -axis. Contour lines show equidistant energies at wave number differences of 3000 cm^{-1} up to $30\,000 \text{ cm}^{-1}$. The upper curves are one-dimensional cuts along Q_{b2} (left) and Q_s (right). The dashed curves in the two upper figures show harmonic potential curves (from [154]).

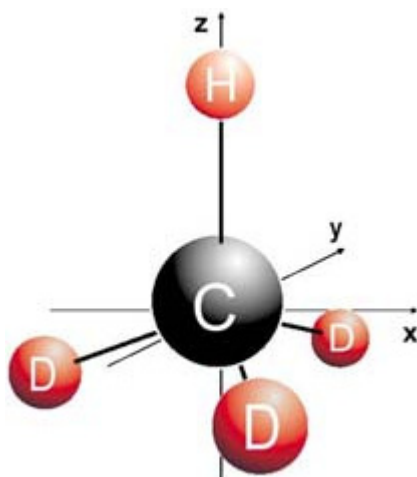


Figure A3.13.5. Coordinates and axes used to describe the wave packet dynamics of the CH chromophore in

CHX₃ or CHXYZ compounds.

In the potential section $V(Q_s)$, shown on the right hand side of [figure A3.13.4](#) the subset of A_1 energy states is drawn. This subset contains only multiplets with integer values of the chromophore quantum number $N = 0, 1, 2, \dots$. This reduction allows for an easier visualization of the multiplet structure and also represents the subset of states that are strongly coupled by the parallel component of the electric dipole moment (see discussion in the following paragraph). The excitation dynamics of the CH chromophore along the stretching manifold can indeed be well described by restriction to this subset of states [[97](#), [154](#)].

Excitation specificity is a consequence of the shape of the electric dipole moment surface. For the alkyl CH chromophore in CHX₃ compounds, the parallel component of the dipole moment, i.e. the component parallel to the symmetry axis, is a strongly varying function of the CH stretching coordinate, whereas it changes little along the bending manifolds [[155](#), [156](#)]. Excitation along this component will thus induce preparation of superposition states lying along the stretching manifold, preferentially. These states thus constitute the ‘bright’ manifold in this example. The remaining states define the ‘dark’ manifolds and any substantial population of these states during or after such an excitation process can thus be directly linked to the existence of IVR. On the other hand, the perpendicular components of the dipole moment vector are strongly varying functions of the bending coordinates. For direct excitation along one of these components, states belonging to the bending manifolds become the ‘bright’ states and any appearance of a subsequent stretching motion can be interpreted as arising from IVR.

The following discussion shall illustrate our understanding of structural changes along ‘dark’ manifolds in terms of wave packet motion as a consequence of IVR. [Figure A3.13.6](#) shows the evolution of the wave packet for the CH chromophore in CHF₃ during the excitation step along the parallel (stretching) coordinate [[97](#)]. The potential surface in the CH chromophore subspace is similar to that for CHD₃ ([figure A3.13.4](#) above), with a slightly more curved form in the stretching–bending representation (figures are shown in [[97](#), [151](#)]). The laser is switched on at a given time t_{-1} , running thereafter as a continuous, monochromatic irradiation up to time t_0 , when it is switched off. Thus, the electric field vector is given as

-24-

$$\vec{E}(t) = h(t - t_{-1})h(t_0 - t)\vec{E}_0 \cos(\omega_L t), \quad (\text{A3.13.61})$$

where $h(t)$ is the Heaviside unit step function, \vec{E}_0 is the amplitude of the electric field vector and $\omega_L = 2\pi c\tilde{\nu}_L$ its angular frequency. Excitation parameters are the irradiation intensity $I_0 = 30 \text{ TW cm}^{-2}$, which corresponds to a maximal electric field strength $E_0 \approx 3.4 \times 10^{10} \text{ V m}^{-1}$, and wave number $\tilde{\nu}_L = 2832.42 \text{ cm}^{-1}$, which lies in the region of the fundamental for the CH stretching vibration (see arrows in the potential cut $V(Q_s)$ of [figure \(A3.13.4\)](#) . The figure shows snapshots of the time evolution of the wave packet between 50 and 70 fs after the beginning of the irradiation ($t_{-1} = 0$ here). On the left-hand side, contour maps of the time-dependent, integrated probability density

$$|\psi(Q_s, Q_b, t)|^2 = \int_{\varphi_b} |\psi(Q_s, Q_b, \varphi_b, t)|^2 d\varphi_b \quad (\text{A3.13.62})$$

are shown, where Q_s is the coordinate for the stretching motion and $Q_b = \sqrt{Q_{b_1}^2 + Q_{b_2}^2}$, $\varphi_b = \arctan(Q_{b_2} / Q_{b_1})$ are polar representations of the bending coordinates Q_{b_1} and Q_{b_2} . Additionally, contour curves of the potential energy surface are drawn at the momentary energy of the wave packet. This energy is defined as:

$$(\text{A3.13.63})$$

$$E(t) = \sum_n E_n p_n(t)$$

where

$$p_n(t) = c_n^*(t)c_n(t) \quad (\text{A3.13.64})$$

are the time-dependent populations of the spectroscopic states during the preparation step (the complex coefficients $c_n(t)$ in equation (A3.13.64) are calculated according to equation (A3.13.55), the spectroscopic energies $E_n = \frac{h}{2\pi}\omega_n$ are defined in equation (A3.13.48); the dashed curves indicate the quantum mechanical uncertainty which arises from the superposition of molecular eigenstates). The same evolution is repeated on the right-hand side of the figure as a three-dimensional representation.

-25-

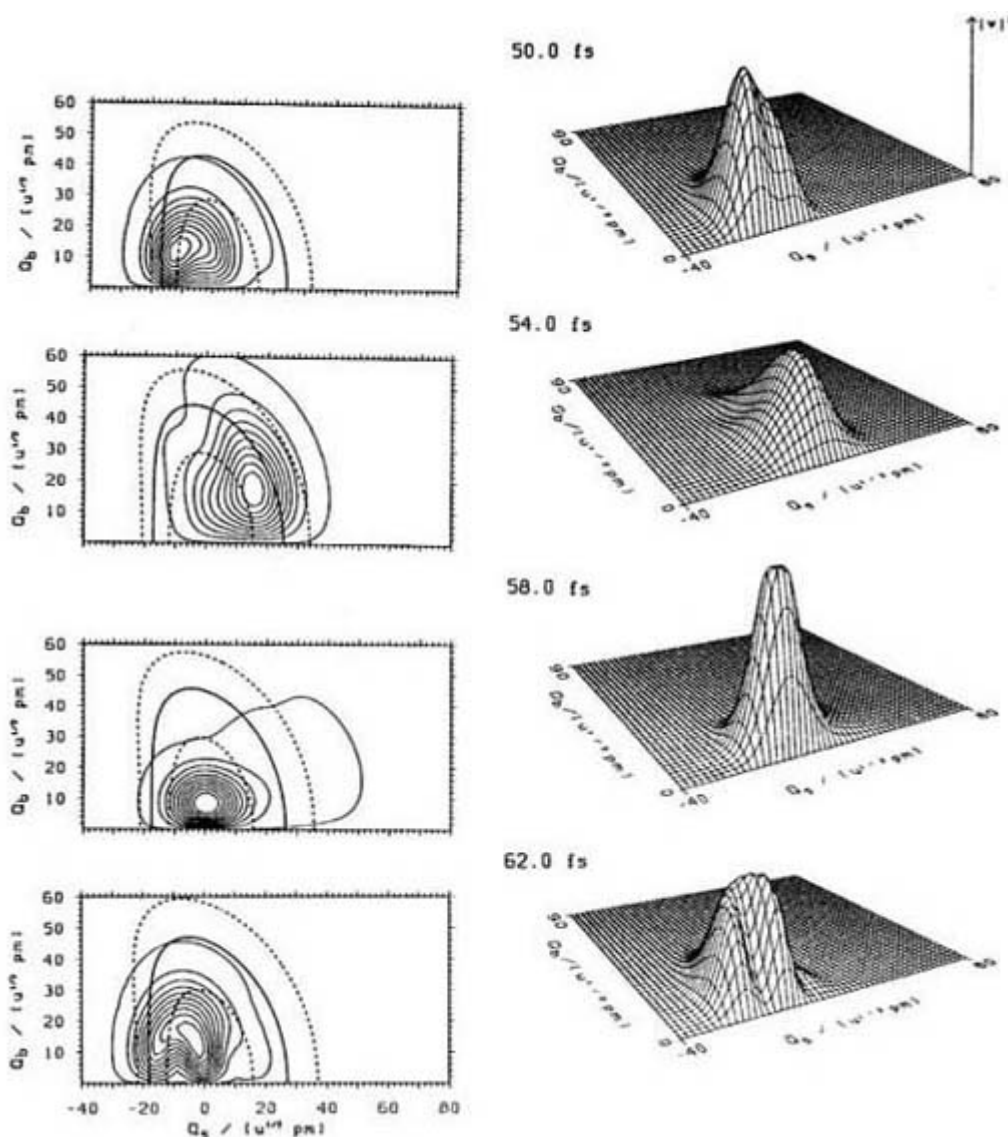


Figure A3.13.6. Time evolution of the probability density of the CH chromophore in CHF_3 after 50 fs of irradiation with an excitation wave number $\tilde{\nu}_L = 2832.42 \text{ cm}^{-1}$ at an intensity $I_0 = 30 \text{ TW cm}^{-2}$. The contour lines of equiprobability density in configuration space have values $2 \times 10^{-5} \text{ u}^{-1} \text{ pm}^{-2}$ for the lowest line shown and distances between the lines of 24, 15, 29 and $20 \times 10^{-5} \text{ u}^{-1} \text{ pm}^{-2}$ in the order of the four images

shown. The averaged energy of the wave packet corresponds to 6000 cm^{-1} (roughly 3100 cm^{-1} absorbed) with a quantum mechanical uncertainty of $\pm 3000 \text{ cm}^{-1}$ (from [97]).

-26-

In the treatment adopted in [97], the motion of the CF_3 frame is implicitly considered in the dynamics of the normal modes. Indeed, the integrand $|\psi(Q_s, Q_b, \phi_b, t)|^2$ in equation (A3.13.62) is to be interpreted as probability density for the change of the CHF_3 structure in the subspace of the CH chromophore, as defined by the normal coordinates Q_s , Q_b and ϕ_b , irrespective of the molecular structure and its change in the remaining space. This interpretation is also valid beyond the harmonic approximation, as long as the structural change in the CH chromophore space can be dynamically separated from that of the rest of the molecule. The assumption of dynamical separation is well confirmed, both from experiment and theory, at least during the first 1000 fs of motion of the CH chromophore.

When looking at the snapshots in figure A3.13.6 we see that the position of maximal probability oscillates back and forth along the stretching coordinate between the walls at $Q_s = -20$ and $+25 \sqrt{u}$ pm, with an approximate period of 12 fs, which corresponds to the classical oscillation period $\tau = 1/\nu$ of a pendulum with a frequency $\nu = c_0 \tilde{\nu} \approx 8.5 \times 10^{13} \text{ s}^{-1}$ and wave number $\tilde{\nu} = 2850 \text{ cm}^{-1}$. Indeed, the motion of the whole wave packet approximately follows this oscillation and, when it does so, the wave packet motion is semiclassical. In harmonic potential wells the motion of the wave packet is always semiclassical [157, 158 and 159]. However, since the potential surface of the CH chromophore is anharmonic, some gathering and spreading out of the wave packet is observable on top of the semiclassical motion. It is interesting to note that, at this ‘initial’ stage of the excitation step, the motion of the wave packet is nearly semiclassical, though with modest amplitudes of the oscillations, despite the anharmonicity of the stretching potential.

The later time evolution is shown in figure A3.13.7 between 90 and 100 fs, and in figure A3.13.8, between 390 and 400 fs, after the beginning of the excitation (time step t_{-1}). Three observations are readily made: first, the amount of energy absorbed by the chromophore has increased, from 3000 cm^{-1} in figure A3.13.6, to 6000 cm^{-1} in figure A3.13.7 and $12\,000 \text{ cm}^{-1}$ in figure A3.13.8. Second, the initially semiclassical motion has been replaced by a more irregular motion of probability density, in which the original periodicity is hardly visible. Third, the wave packet starts to occupy nearly all of the energetically available region in configuration space, thus escaping from the initial, ‘bright’ manifolds into the ‘dark’ manifolds. From these observations, the following conclusions may be directly drawn: IVR of the CH chromophore in fluoroform is fast (in the subpicosecond time scale); IVR sets in already during the excitation process, i.e. when an external force field is driving the molecular system along a well prescribed path in configuration space (the ‘bright’ manifold); IVR is of the delocalization type (DIVR). Understanding these observations is central for the understanding of IVR and they are discussed as follows:

(a) A more detailed analysis of quantum dynamics shows that the molecular system, represented by the group of vibrations pertaining to the CH chromophore in this example, absorbs continuously more energy as time goes on. Let the absorbed energy be $E_{\text{abs}} = N_{\text{v,abs}}(h/2\pi)\omega_L$, where $N_{\text{v,abs}}$ is the mean number of absorbed photons. Since the carrier frequency of the radiation field is kept constant at a value close to the fundamental of the stretching oscillation, $\omega_L \approx \omega_N = 1 - \omega_N = 0$ (N being the chromophore quantum number here), this means that the increase in absorbed energy is a consequence of the stepwise multiphoton excitation process, in which each vibrational level serves as a new starting level for further absorption of light after it has itself been significantly populated. This process is schematically represented, within the example for CHD_3 , by the sequence of upright arrows shown on the right-hand side of figure A3.13.4 $N_{\text{v}, \alpha\beta\sigma}$ is thus a smoothly increasing function of time.

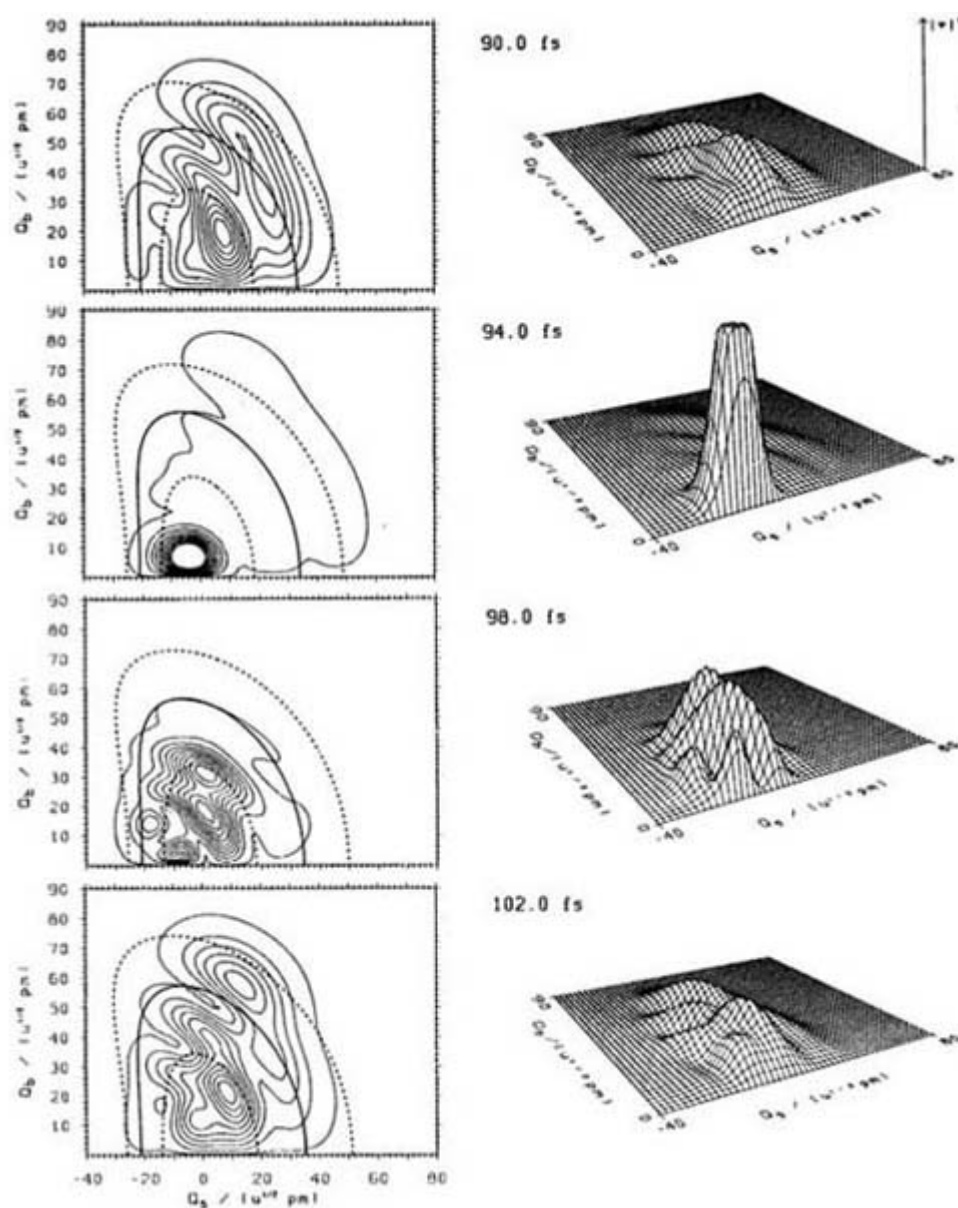


Figure A3.13.7. Continuation of the time evolution for the CH chromophore in CHF_3 after 90 fs of irradiation (see also figure A3.13.6). Distances between the contour lines are $10, 29, 16$ and $9 \times 10^{-5} \text{ u}^{-1} \text{ pm}^{-2}$ in the order of the four images shown. The averaged energy of the wave packet corresponds to 9200 cm^{-1} (roughly 6300 cm^{-1} absorbed) with a quantum mechanical uncertainty of $\pm 5700 \text{ cm}^{-1}$ (from [97]).

(b) The disappearance of the semiclassical type of motion and, thus, the delocalization of the wave packet, is understood to follow the onset of dephasing. With increasing energy, both the effective anharmonic couplings between the ‘bright’ stretching mode and the ‘dark’ bending modes, as well as the diagonal anharmonicity of

the ‘bright’ mode increase. The larger the anharmonicity, the larger the deviation from a purely harmonic behaviour, in which the wave packet keeps on moving in a semiclassical way. In quantum mechanics, the increase in anharmonicity of an oscillator leads to an effective broadening $\Delta\nu_{\text{eff}} > 0$ in the distribution of frequencies of high-probability transitions—for transitions induced by the electric dipole operator usually those with a difference of ± 1 in the oscillator quantum number (for the harmonic oscillator $\Delta\nu_{\text{eff}} = 0$). On the other hand, these are the transitions which play a major role in the stepwise multiphoton excitation of molecular vibrations. A broadening of the frequency distribution invariably leads to a broadening of the distribution of relative phases of the time-dependent coefficients $c_n(t)$ in [equation \(A3.13.55\)](#). Although the sum in [equation \(A3.13.55\)](#) is entirely coherent, one might introduce an effective coherence time defined by:

$$\tau_{c,\text{eff}} = 1/\Delta\nu_{\text{eff}}. \quad (\text{A 3.13.65})$$

For the stretching oscillations of the CH chromophore in CHF_3 $\tau_{c,\text{eff}} \approx 100$ fs. Clearly, typical coherence time ranges depend on both the molecular parameters and the effectively absorbed amount of energy during the excitation step, which in turn depends on the coupling strength of the molecule–radiation interaction. A more detailed study of the dispersion of the wave packet and its relationship with decoherence effects was carried out in [106]. In [97] an excitation process has been studied for the model of two anharmonically coupled, resonant harmonic oscillators (i.e. with at least one cubic coupling term) but under similar conditions as for the CH chromophore in fluoroform discussed here. When the cubic coupling parameter is chosen to be very small compared with the diagonal parameters of the Hamilton matrix, the motion of the wave packet is indeed semiclassical for very long times (up to 600 ps) and, moreover, the wave packet does probe the bending manifold without significantly changing its initial shape. This means that, under appropriate conditions, IVR can also be of the classical type within a quantum mechanical treatment of the dynamics. Such conditions require, for instance, that the band width $h\Delta\nu_{\text{eff}}$ be smaller than the resonance width (power broadening) of the excitation process.

(c) The third observation, that the wave packet occupies nearly all of the energetically accessible region in configuration space, has a direct impact on the understanding of IVR as a rapid promotor of microcanonical equilibrium conditions. Energy equipartition preceding a possible chemical reaction is the main assumption in quasiequilibrium statistical theories of chemical reaction dynamics (‘RRKM’ theory [161, 162 and 163], ‘transition state’ theory [164, 165] but also within the ‘statistical adiabatic channel model’ [76, 77]; see also [chapter A3.12](#) and further recent reviews on varied and extended forms of statistical theories in [25, 166, 167, 168, 169, 170, 171 and 172]). In the case of CHF_3 one might conclude from inspection of the snapshots at the later stage of the excitation dynamics (see [figure A3.13.8](#)) that after 400 fs the wave packet delocalization is nearly complete. Moreover, this delocalization arises here from a fully coherent, isolated evolution of a system consisting of one molecule and a coherent radiation field (laser). Of course, within the common interpretation of the wave packet as a probability distribution in configuration space, this result means that, for an ensemble of identically prepared molecules, vibrational motion is essentially delocalized at this stage and vibrational energy is nearly equipartitioned.

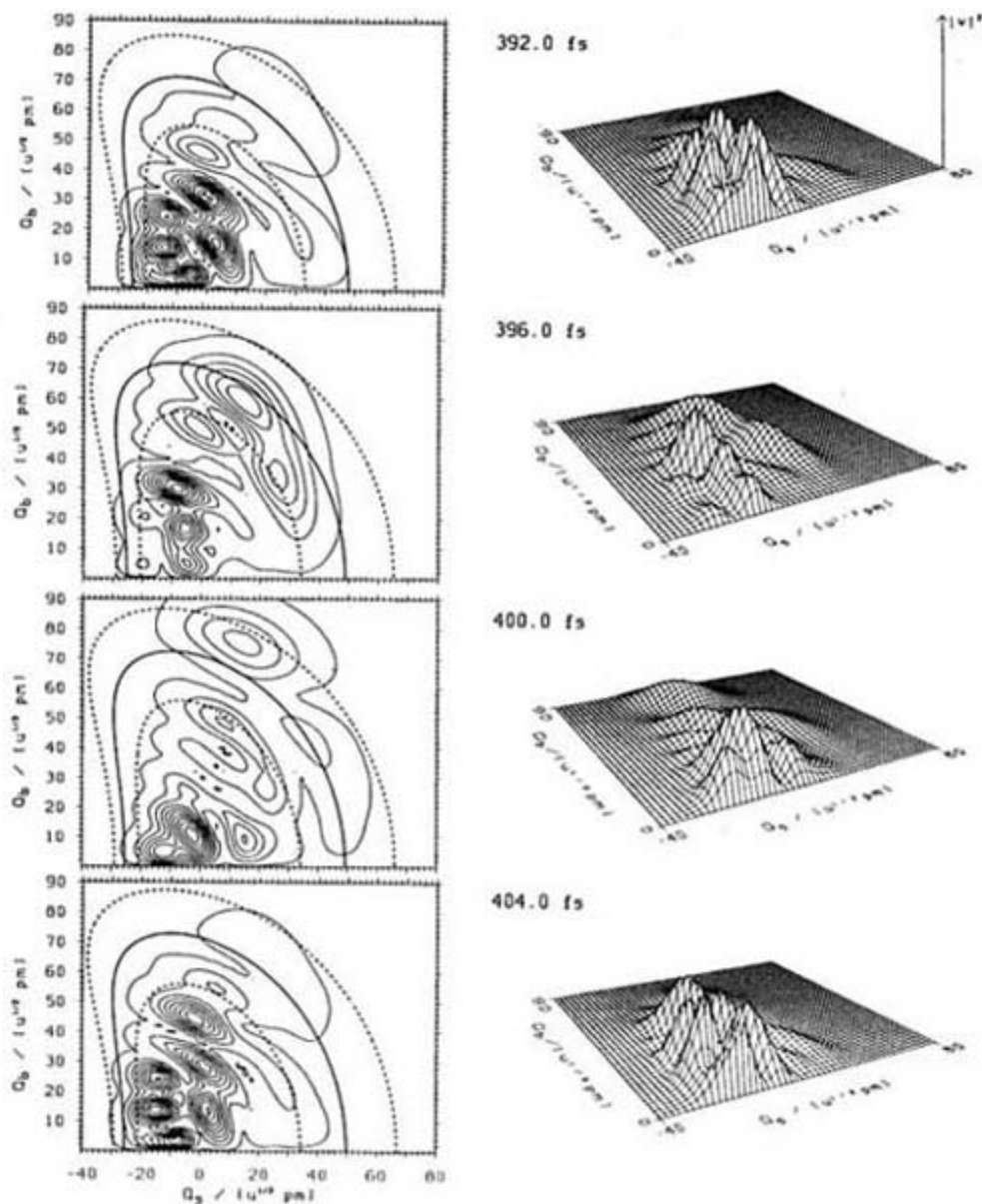


Figure A3.13.8. Continuation of the time evolution for the CH chromophore in CHF_3 after 392 fs of irradiation (see also [figure A3.13.6](#) and [figure A3.13.7](#)). Distances between the contour lines are 14 , 12 , 13 and $14 \times 10^{-5} u^{-1} \text{ pm}^{-2}$ in the order of the four images shown. The averaged energy of the wave packet corresponds to $15\,000 \text{ cm}^{-1}$ (roughly $12\,100 \text{ cm}^{-1}$ absorbed) with a quantum mechanical uncertainty of $\pm 5800 \text{ cm}^{-1}$ (from [97]).

-30-

However, the wave packet does not occupy all of the energetically accessible region. A more detailed analysis of populations [97, table IV] reveals that, during the excitation process, the absorbed energy is inhomogeneously distributed among the set of molecular eigenstates of a given energy shell (such a shell is represented by all nearly iso-energetic states belonging to one of the multiplets shown on the right-hand side of [figure A3.13.4](#)). Clearly, equipartition of energy is attained, if all states of an energy shell are equally populated. The microcanonical probability distribution in configuration space may then be represented by a typical member of the microcanonical ensemble, defined e.g. by the wave function

$$\psi_{\text{micro}} \approx \frac{1}{\sqrt{N_{\text{shell}}}} \sum_{n \in \text{shell}} \exp(-i\varphi_n^{\text{random}}) \phi_n \quad (\text{A3.13.66})$$

where N_{shell} denotes the number of nearly iso-energetic states ϕ_n of a shell and ϕ_n^{random} is a random phase. Such a state is shown in [figure A3.13.9](#). When comparing this state with the state generated by multiphoton excitation, the two different kinds of superposition that lead to these wave packets must, of course, be distinguished. In the stepwise multiphoton excitation, the time evolved wave packet arises from a superposition of many states in several multiplets (with roughly constant averaged energy after some excitation time and a large energy uncertainty). The microcanonical distribution is given by the superposition of states in a single multiplet (of the same averaged energy but much smaller energy uncertainty). In the case of the CH chromophore in CHF_3 studied in this example, the distribution of populations within a molecular energy shell is not homogeneous during the excitation process because the multiplets are not ideally centred at the multiphoton resonance levels and their energy range is effectively too large when compared to the resonance width of the excitation process (power broadening). If molecular energy shells fall entirely within the resonance width of the excitation, such as in the model systems of two harmonic oscillators studied in [97], population distribution within a shell becomes more homogeneous [97, table V]. However, as discussed in that work, equidistribution of populations does not imply that the wave packet is delocalized. Indeed, the contrary was shown to occur. If the probability distribution in configuration space is to delocalize, the relative phases between the superposition states must follow an irregular evolution, such as in a random phase ensemble, in addition to equidistribution of population. Thus, one statement would therefore be that IVR is not complete, although very fast, during the multiphoton excitation of CHF_3 . Excitation and redistribution are indeed two concurring processes. In the limit of weak field excitation, in the spectroscopic regime, the result is a superposition of essentially two eigenstates (the ground and an excited state, for instance). Within the ‘bright’ state concept, strong IVR will be revealed by an instantaneous delocalization of probability density, both in the ‘bright’ and the ‘dark’ manifolds, as soon as the excited state is populated, because the excited state is, of course, a superposition state of states from both manifolds. On the other hand, strong field stepwise IR multiphoton excitation promotes, in a first step, the deposition of energy in a spatially localized, time-dependent molecular structure. Simultaneously, IVR starts to induce redistribution of this energy among other modes. The redistribution becomes apparent after some time has passed and is expected to be of the DIVR type, at least on longer time scales. DIVR may lead to a complete redistribution in configuration space, if the separation between nearly iso-energetic states is small compared to the power broadening of the excitation field. However, under such conditions, at least during an initial stage of the dynamics, CIVR will dominate.

-31-

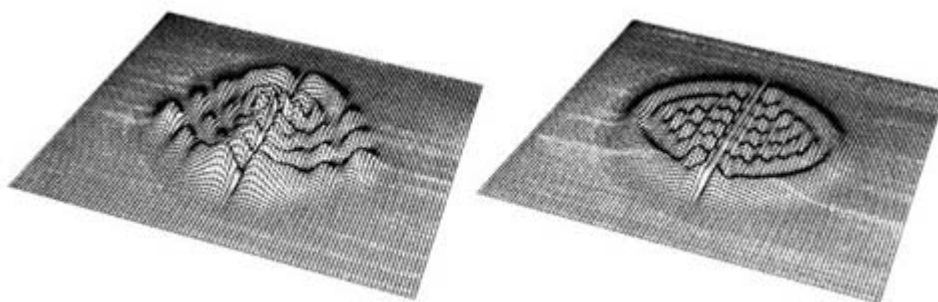


Figure A3.13.9. Probability density of a microcanonical distribution of the CH chromophore in CHF_3 within the multiplet with chromophore quantum number $N = 6$ ($N_{\text{shell}} = N + 1 = 7$). Representations in configuration space of stretching (Q_s) and bending (Q_b) coordinates (see text following (equation (A3.13.62)) and [figure A3.13.10](#)). Left-hand side: typical member of the microcanonical ensemble of the multiplet with $N = 6$ (random phases, (equation (A3.13.66))). Right-hand side: microcanonical density $P_{\text{micro}} = \frac{1}{N_{\text{shell}}} \sum_{n \in \text{shell}} |\phi_n|^2$ for the multiplet with $N = 6$ ($N_{\text{shell}} = 7$). Adapted from [81].

In view of the foregoing discussion, one might ask what is a typical time evolution of the wave packet for the *isolated* molecule, what are typical time scales and, if initial conditions are such that an entire energy shell participates, does the wave packet resulting from the coherent dynamics look like a microcanonical

distribution? Such studies were performed for the case of an initially pure stretching ‘Fermi mode’ ($v_s, v_b = 0$), with a high stretching quantum number, e.g. $v_s = 6$. It was assumed that such a state might be prepared by irradiation with some hypothetical laser pulse, without specifying details of the pulse. The energy of that state is located at the upper end of the energy range of the corresponding multiplet [81, 152, 154], which has a total of $N_{\text{shell}} = 7$ states. Such a state couples essentially to all remaining states of that multiplet. The corresponding evolution of the isolated system is shown as snapshots after the preparation step ($t_0 = 0$) in [figure A3.13.10](#). The wave packet starts to spread out from the initially occupied stretching manifold (along the coordinate axis denoted by Q_s) into the bending manifold (Q_b) within the first 30–45 fs of evolution (left-hand side). Later on, it remains delocalized most of the time (as shown at the time steps 80, 220 and 380 fs, on the right-hand side) with exceptional partial recovery of the initial conditions at some isolated times (such as at 125 fs). The shape of the distribution at 220 fs is very similar to that of a typical member of the microcanonical ensemble in [figure A3.13.9](#) above. However, in [figure A3.13.9](#), the relative phases between the seven superposition states were drawn from a random number generator, whereas in [figure A3.13.10](#) they result from a fully coherent and deterministic propagation of a wave function.

-32-

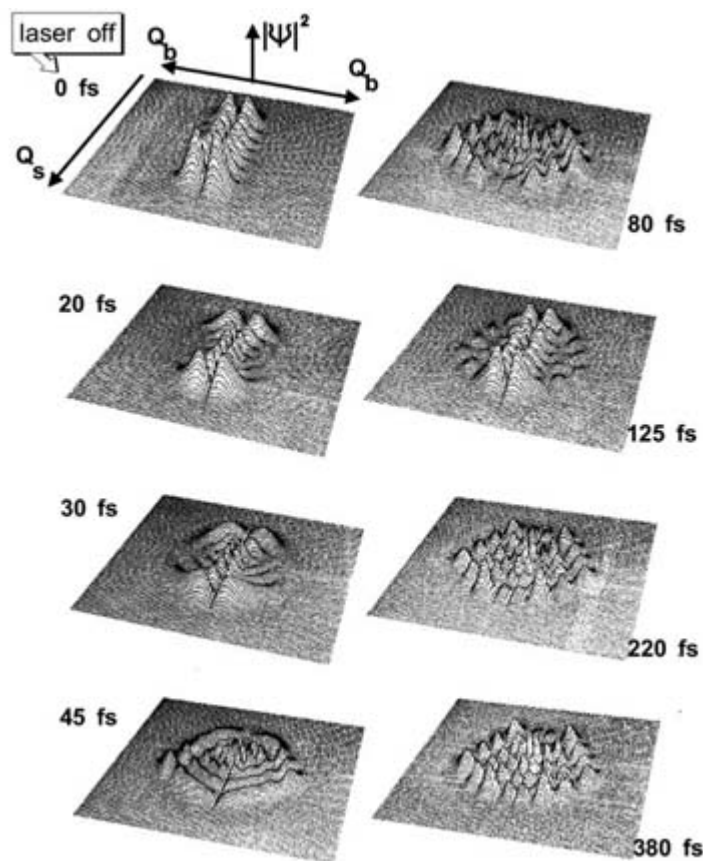


Figure A3.13.10. Time-dependent probability density of the isolated CH chromophore in CHF_3 . Initially, the system is in a ‘Fermi mode’ with six quanta of stretching and zero of bending motion. The evolution occurs within the multiplet with chromophore quantum number $N = 6$ ($N_{\text{shell}} = N + 1 = 7$). Representations are given in the configuration space of stretching (Q_s) and bending (Q_b) coordinates (see text following [equation \(A3.13.62\)](#)): Q_b is strictly a positive quantity, and there is always a node at $Q_b = 0$; the mirrored representation at $Q_b = 0$ is artificial and serves to improve visualization). Adapted from [81].

IVR in the example of the CH chromophore in CHF_3 is thus at the origin of a redistribution process which is, despite its coherent nature, of a statistical character. In CHD_3 , the dynamics after excitation of the stretching manifold reveals a less complete redistribution process in the same time interval [97]. The reason for this is a smaller effective coupling constant k'_{sbb} between the ‘Fermi modes’ of CHD_3 (by a factor of four) when

compared to that of CHF_3 . In [97] it was shown that redistribution in CHD_3 becomes significant in the picosecond time scale. However, on that time scale, the dynamical separation of time scales is probably no longer valid and couplings to modes pertaining to the space of CD_3 vibrations may become important and have additional influence on the redistribution process.

-33-

A3.13.5.2 IVR AND TIME-DEPENDENT CHIRALITY

IVR in the CH chromophore system may also arise from excitation along the bending manifolds. Bending motions in polyatomic molecules are of great importance as primary steps for reactive processes involving isomerization and similar, large amplitude changes of internal molecular structure. At first sight, the one-dimensional section of the potential surface along the out-of-plane CH bending normal coordinate in CHD_3 , shown in [figure A3.13.4](#) is clearly less anharmonic than its one-dimensional stretching counterpart, also shown in that figure, even up to energies in the wave number region of $30\,000\text{ cm}^{-1}$. This suggests that coherent sequential multiphoton excitation of a CH bending motion, for instance along the x -axis in [figure A3.13.5](#) may induce a quasiclassical motion of the wave packet along that manifold [159, 160], which is significantly longer lived than the motion induced along the stretching manifold under similar conditions (see discussion above). Furthermore, the two-dimensional section in the CH bending subspace, spanned by the normal coordinates in the lower part of [figure A3.13.4](#) is approximately isotropic. This corresponds to an almost perfect $C_{\infty\omega}$ symmetry with respect to the azimuthal angle φ (in the xy plane of [figure A3.13.5](#), and is related to the approximate conservation of the bending vibration angular momentum ℓ_b [152, 173]. This implies that the direct anharmonic coupling between the degenerate bending manifolds is weak. However, IVR between these modes might be mediated by the couplings to the stretching mode. An interesting question is then to what extent such a coupling scheme might lead to a motion of the wave packet with quasiclassical exchange of vibrational energy between the two bending manifolds, following paths which could be described by classical vibrational mechanics, corresponding to CIVR. Understanding quasiclassical exchange mechanisms of large amplitude vibrational motion opens one desirable route of exerting control over molecular vibrational motion and reaction dynamics. In [154] these questions were investigated by considering the CH bending motion in CHD_3 and the asymmetric isotopomers CHD_2T and CHDT_2 . The isotopic substitution was investigated with the special goal of a theoretical study of the coherent generation of dynamically chiral, bent molecular structures [174] and of the following time evolution. It was shown that IVR is at the origin of a coherent racemization dynamics which is superposed to a very fast, periodic exchange of left- and right-handed chiral structures ('stereomutation' reaction, period of roughly 20 fs, comparable to the period of the bending motion) and sets in after typically 300–500 fs. The main results are reviewed in the discussion of [figure A3.13.11](#) [figure A3.13.12](#) and [figure A3.13.13](#).

-34-

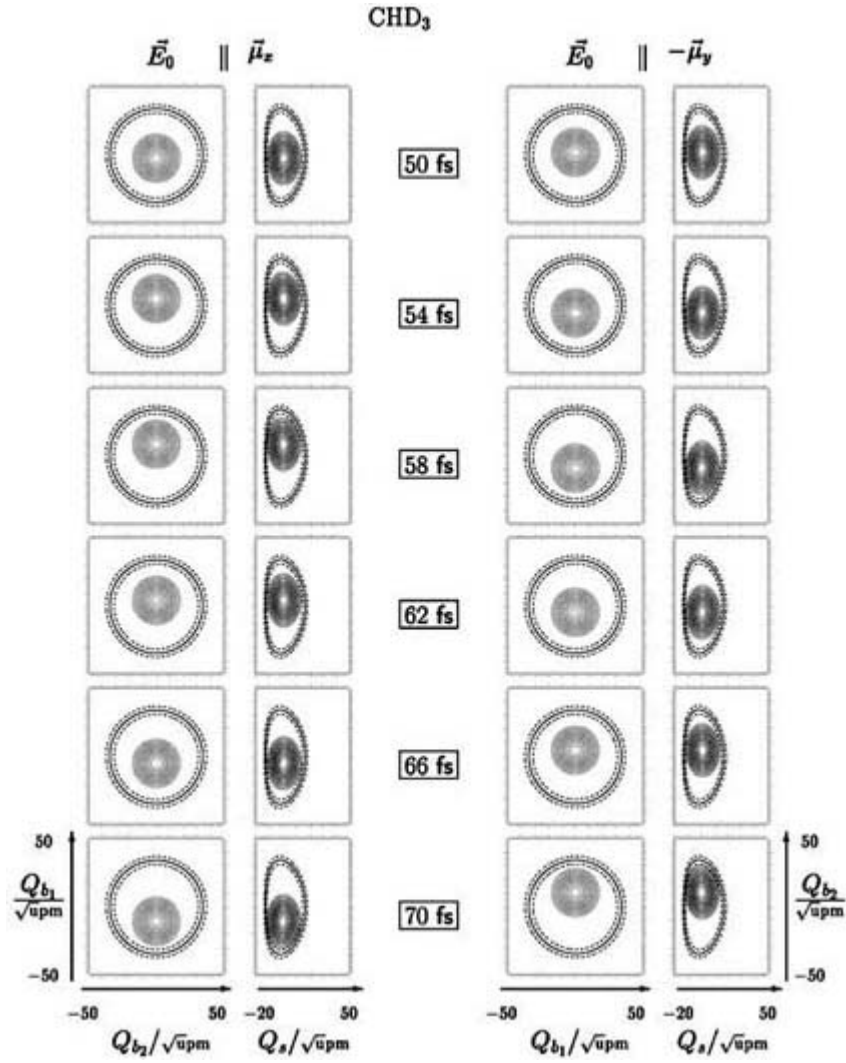


Figure A3.13.11. Illustration of the time evolution of reduced two-dimensional probability densities $|\psi_{bb}|^2$ and $|\psi_{sb}|^2$, for the excitation of CHD₃ between 50 and 70 fs (see [154] for further details). The full curve is a cut of the potential energy surface at the momentary absorbed energy corresponding to 3000 cm⁻¹ during the entire time interval shown here (≈ 6000 cm⁻¹, if zero point energy is included). The dashed curves show the energy uncertainty of the time-dependent wave packet, approximately 500 cm⁻¹. Left-hand side: excitation along the x -axis (see figure A3.13.5). The vertical axis in the two-dimensional contour line representations is the Q_{b1} -axis, the horizontal axes are Q_{b2} and Q_s , for $|\psi_{bb}|^2$ and $|\psi_{sb}|^2$, respectively. Right-hand side: excitation along the y -axis, but with the field vector pointing into the negative y -axis. In the two-dimensional contour line representations, the vertical axis is the Q_{b2} -axis, the horizontal axes are Q_{b1} and Q_s , for $|\psi_{bb}|^2$ and $|\psi_{sb}|^2$, respectively. The lowest contour line has the value $44 \times 10^{-5} \text{ u}^{-1} \text{ pm}^{-2}$, the distance between them is $7 \times 10^{-5} \text{ u}^{-1} \text{ pm}^{-2}$. Maximal values are nearly constant for all the images in this figure and correspond to $140 \times 10^{-5} \text{ u}^{-1} \text{ pm}^{-2}$ for $|\psi_{bb}|^2$ and $180 \times 10^{-5} \text{ u}^{-1} \text{ pm}^{-2}$ for $|\psi_{sb}|^2$.

The wave packet motion of the CH chromophore is represented by simultaneous snapshots of two-dimensional representations of the time-dependent probability density distribution

$$|\psi_{sb}(t, Q_s, Q_{b_i})|^2 = \int_{Q_{b_j}} dQ_{b_j} |\psi(t, Q_s, Q_{b_i}, Q_{b_j})|^2 \quad (i \neq j) \quad (\text{A3.13.67})$$

and

$$|\psi_{bb}(t, Q_{b_1}, Q_{b_2})|^2 = \int_{Q_s} dQ_s |\psi(t, Q_s, Q_{b_1}, Q_{b_2})|^2. \quad (\text{A3.13.68})$$

Such a sequence of snapshots, calculated in intervals of 4 fs, is shown as a series of double contour line plots on the left-hand side of [figure A3.13.11](#) (the outermost row shows the evolution of $|\psi_{bb}|^2$, equation (A3.13.68), the innermost row is $|\psi_{sb}|^2$, equation (A3.13.67), at the same time steps). This is the wave packet motion in CHD₃ for excitation with a linearly polarized field along the x -axis at 1300 cm⁻¹ and 10 TW cm⁻² after 50 fs of excitation. At this point a more detailed discussion regarding the orientational dynamics of the molecule is necessary. Clearly, the polarization axis is defined in a laboratory fixed coordinate system, while the bending axes are fixed to the molecular frame. Thus, exciting internal degrees of freedom along specific axes in the internal coordinate system requires two assumptions: the molecule must be oriented or aligned with respect to the external polarization axis, and this state should be stationary, at least during the relevant time scale for the excitation process. It is possible to prepare oriented states [[112](#), [114](#), [115](#)] in the gas phase, and such a state can generally be represented as a superposition of a large number of rotational eigenstates. Two questions become important then: How fast does such a rotational superposition state evolve? How well does a purely vibrational wave packet calculation simulate a more realistic calculation which includes rotational degrees of freedom, i.e. with an initially oriented rotational wave packet? The second question was studied recently by full dimensional quantum dynamical calculations of the wave packet motion of a diatomic molecule during excitation in an intense infrared field [[175](#)], and it was verified that rotational degrees of freedom may be neglected whenever vibrational–rotational couplings are not important for intramolecular rotational–vibrational redistribution (IVRR) [[84](#)]. Regarding the first question, because of the large rotational constant of methane, the time scales on which an initially oriented state of the free molecule is maintained are likely to be comparatively short and it would also be desirable to carry out calculations that include rotational states explicitly. Such calculations were done, for instance, for ozone at modest excitations [[116](#), [117](#)], but they would be quite difficult for the methane isotopomers at the high excitations considered in the present example.

The multiphoton excitation scheme corresponding to excitation along the x -axis is shown by the upright arrows on the left-hand side of [figure A3.13.4](#). In the convention adopted in [[154](#)], nuclear displacements along Q_{b_1} occur along the x -axis, displacements along Q_{b_2} are directed along the y -axis. One observes a semiclassical, nearly periodic motion of the wave packet along the excited manifold with a period of approximately 24 fs, corresponding to the frequency of the bending vibrations in the wave number region around 1500 cm⁻¹. At this stage of the excitation process, the motion of the wave packet is essentially one-dimensional, as seen from the trajectory followed by the maximum of the probability distribution and its practically unchanged shape during the oscillations back and forth between the turning points. The latter lie on the potential energy section defined by the momentary energy $E(t)$ of the wave packet, as described above,

and describe the classically accessible region in configuration space. These potential energy sections are shown by the continuous curves in the figures, which are surrounded by dotted curves describing the energy uncertainty.

The sequence on the right-hand side of [figure A3.13.11](#) shows wave packets during the excitation along the y -axis. Here, excitation was chosen to be antiparallel to the y -axis ($E_0 \parallel -\mu_y$). This choice induces a phase shift of π between the two wave packets shown in the figure, in addition to forcing oscillations along different directions. Excitation along the y -axis can be used to generate dynamically chiral structures. If the excitation laser field is switched off, e.g. at time step 70 fs after beginning the excitation, the displacement of the wave packet clearly corresponds to a bent molecular structure with angle $\vartheta \approx 10^\circ$ (e.g. in the xy plane of figure

A3.13.5 . This structure will, of course, also change with time for the isolated molecule, and one expects this change to be oscillatory, like a pendulum, at least initially. Clearly, IVR will play some role, if not at this early stage, then at some later time. One question is, will it be CIVR or DIVR? When studying this question with the isotopically substituted compounds CHD_2T and CHDT_2 , the y -axis being perpendicular to the C_s mirror plane, a bent CH chromophore corresponds to a chiral molecular structure with a well defined chirality quantum number, say R . As time evolves, the wave packet moves to the other side of the symmetry plane, $Q_{b_2} = 0$, implying a change of chirality. In this context, the enantiomeric excess can be defined by the probability

$$P_R(t) = \int_{-\infty}^0 dQ_{b_2} |\psi_{b_2}(t, Q_{b_2})|^2 \quad (\text{A3.13.69})$$

for right-handed ('R') chiral structures ($P_L(t) = 1 - P_R(t)$ is the probability for left-handed ('L') structures), where

$$|\psi_{b_2}(t, Q_{b_2})|^2 = \int_{Q_s} \int_{Q_{b_1}} dQ_s dQ_{b_1} |\psi(t, Q_s, Q_{b_1}, Q_{b_2})|^2. \quad (\text{A3.13.70})$$

The time evolution of $P_R(t)$ is shown in [figure A3.13.12](#) for the field free motion of wave packets for CHD_2T and CHDT_2 prepared by a preceding excitation along the y -axis.

In the main part of each figure, the evolution of P_R calculated within the stretching and bending manifold of states for the CH chromophore is shown (full curve). The dashed curve shows the evolution of P_R within a one-dimensional model, in which only the Q_{b_2} bending manifold is considered during the dynamics. Within this model there is obviously no IVR, and comparison of the full with the dashed curves helps to visualize the effect of IVR. The insert on the left-hand side shows a survey of the evolution of P_R for the one-dimensional model during a longer time interval of 2 ps, while the insert on the right-hand side shows the evolution of P_R for the calculation within the full three-dimensional stretching and bending manifold of states during the same time interval of 2 ps. The three-dimensional calculations yield a fast, initially nearly periodic, evolution, with an approximate period of 20 fs, which is superimposed by a slower decay of probability corresponding to an overall decay of enantiomeric excess $|D_{\text{abs}}(t)| = |1 - 2P_R(t)|$ on a time scale of 300–400 fs for both CHD_2T and CHDT_2 . The decay is clearly more pronounced for CHD_2T ([figure A3.13.12](#)a)). The first type of evolution corresponds to a stereomutation reaction, while the second can be interpreted as racemization. A further question is then related to the origin of this racemization.

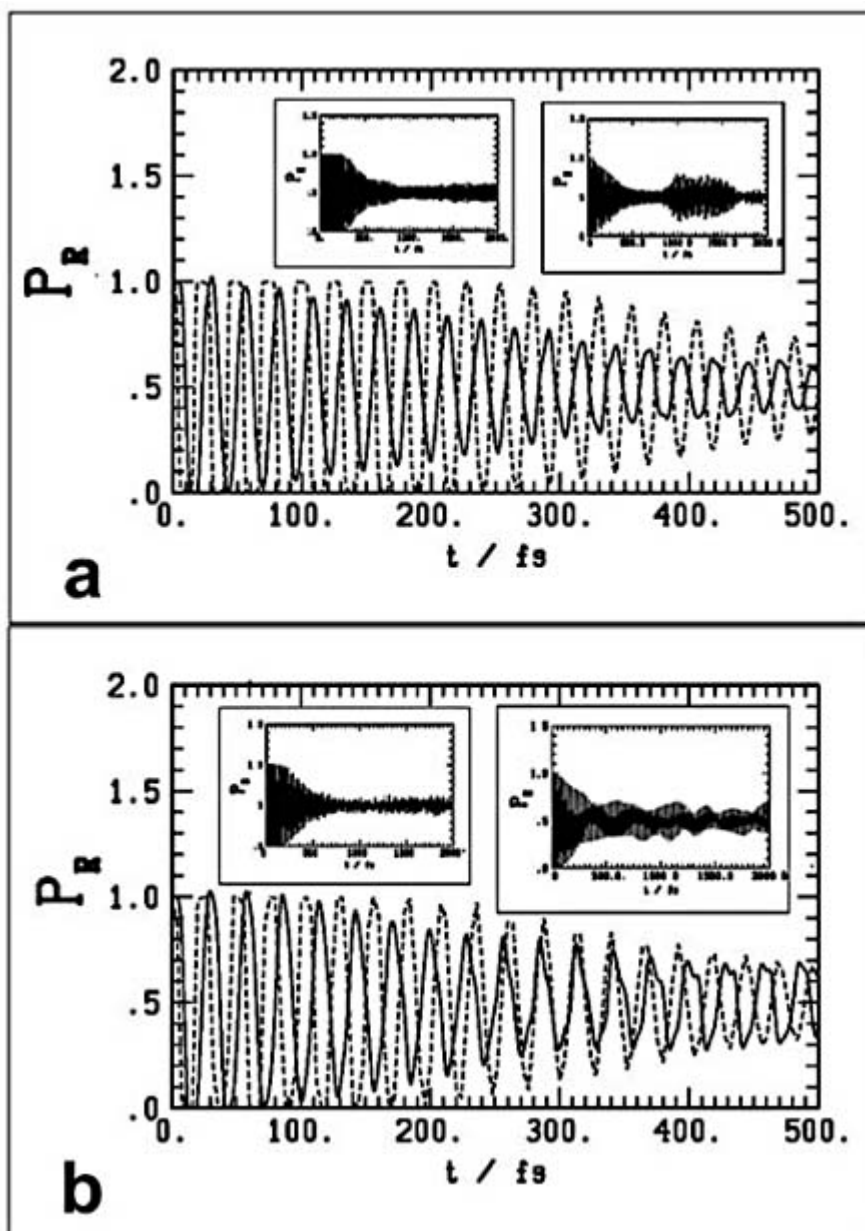


Figure A3.13.12. Evolution of the probability for a right-handed chiral structure $P_R(t)$ (full curve, see (equation (A3.13.69))) of the CH chromophore in CHD_2T (a) and CHDT_2 (b) after preparation of chiral structures with multiphoton laser excitation, as discussed in the text (see also [154]). For comparison, the time evolution of P_R according to a one-dimensional model including only the Q_{b2} bending mode (dashed curve) is also shown. The left-hand side insert shows the time evolution of P_R within the one-dimensional calculations for a longer time interval; the right-hand insert shows the P_R time evolution within the three-dimensional calculation for the same time interval (see text).

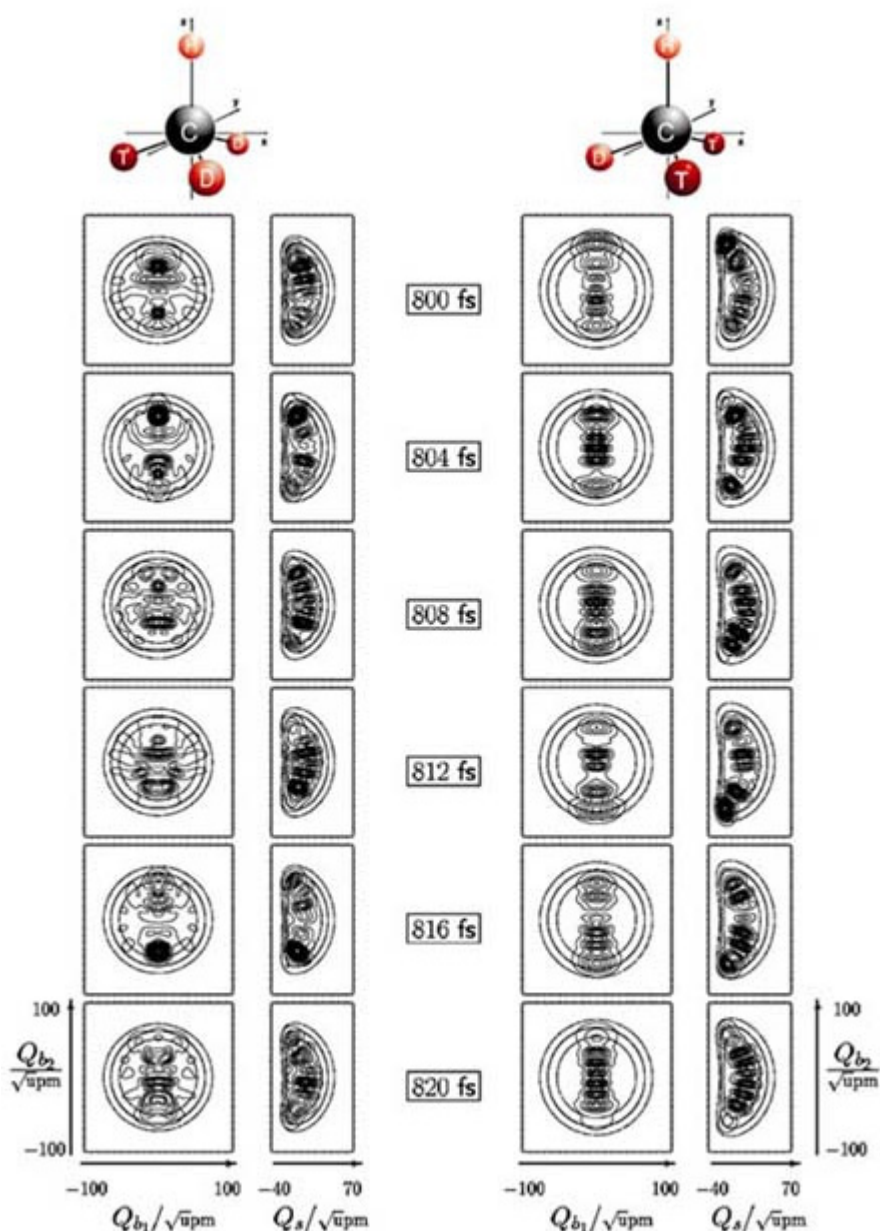


Figure A3.13.13. Illustration of the time evolution of reduced two-dimensional probability densities $|\psi_{bb}|^2$ and $|\psi_{sb}|^2$, for the isolated CHD_2T (left-hand side) and CHDT_2 (right-hand side) after 800 fs of free evolution. At time 0 fs the wave packets corresponded to a localized, chiral molecular structure (from [154]). See also text and figure A3.13.11.

Figure A3.13.14 shows the wave packet motion for CHD_2T and CHDT_2 , roughly 800 fs after the initially localized, chiral structure has been generated. Comparison with the wave packet motion allows for the conclusion that racemization is induced by the presence of DIVR between all vibrational modes of the CH chromophore. However, while DIVR is quite complete for CHD_2T , after excitation along the y axis, it is only two-dimensional for CHDT_2 . A localized exchange of vibrational energy in terms of CIVR has not been observed at any intermediate time step. Racemization is stronger for CHD_2T , for which DIVR occurs in the full three-dimensional subspace of the CH chromophore, under the present conditions. It is less pronounced for CHDT_2 , which has a higher degree of localization of the wave packet motion. In comparison with the one-dimensional calculations in figure A3.13.12, it becomes evident that there is a decay of the overall

enantiomeric excess for CHD_2T , as well as for CHDT_2 , also in the absence of IVR. The decay takes place on a time scale of 500–1000 fs and is a consequence of the dephasing of the wave packet due to the diagonal anharmonicity of the bending motion. This decay may, of course, also be interpreted as racemization. However, it is much less complete than racemization in the three-dimensional case and clearly of secondary importance for the enantiomeric decay in the first 200 fs.

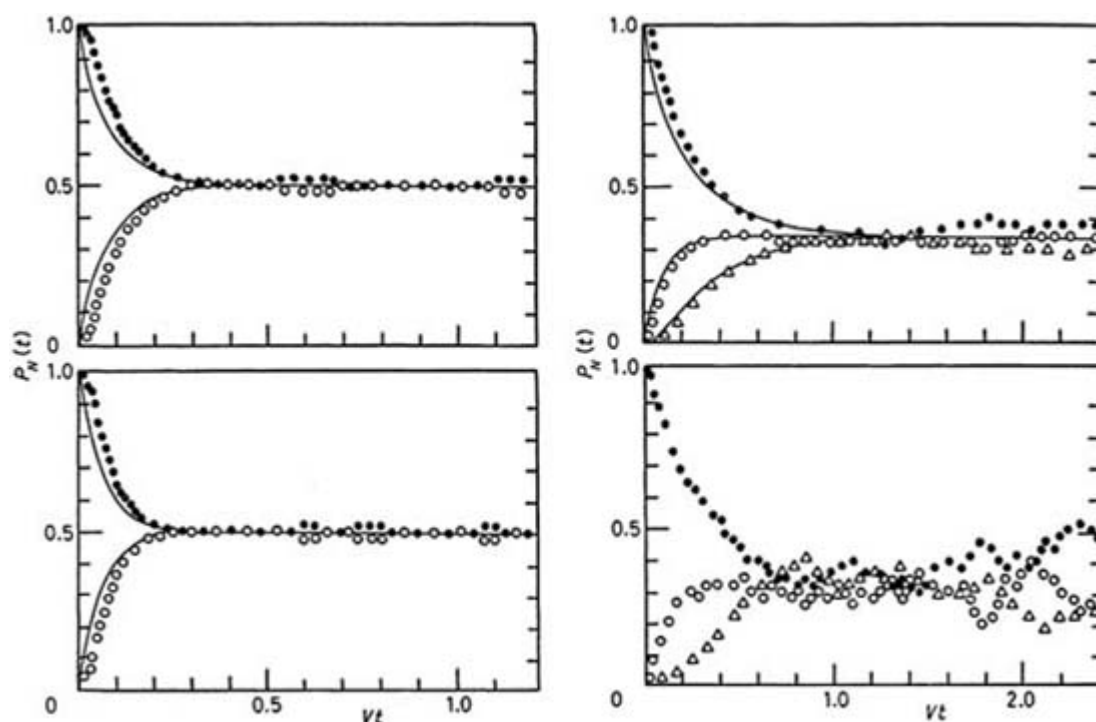


Figure A3.13.14. Illustration of the quantum evolution (points) and Pauli master equation evolution (lines) in quantum level structures with two levels (and 59 states each, left-hand side) and three levels (and 39 states each, right-hand side) corresponding to a model of the energy shell IVR (horizontal transition in [figure A3.13.1](#)). From [38]. The two-level structure (left) has two models: $|V_{ij}|^2 = \text{const}$ and random signs (upper part), random V_{ij} , but $-V_m \leq V_{ij} \leq V_m$ (lower part). The right-hand side shows an evolution with initial diagonal density matrix (upper part) and a single trajectory (lower part).

A 3.13.6 STATISTICAL MECHANICAL MASTER EQUATION TREATMENT OF INTRAMOLECULAR ENERGY REDISTRIBUTION IN REACTIVE MOLECULES

The previous sections indicate that the full quantum dynamical treatment of IVR in an intermediate size molecule even under conditions of ‘coherent’ excitation shows phenomena reminiscent of relaxation and equilibration. This suggests that, in general, at very high excitations in large polyatomic molecules with densities of states easily exceeding the order of 10^{10} cm^{-1} (or about 10^9 molecular states in an energy interval corresponding to 1 J mol^{-1}), a statistical master equation treatment may be possible [38, 122]. Such an approach has been justified by quantum simulations in model systems as well as analytical considerations [38], following early ideas in the derivation of the statistical mechanical Pauli equation [176]. [Figure A3.13.14](#) shows the kinetic behaviour in such model systems. The ‘coarse grained’ populations of groups of quantum states (‘levels’ with less than 100 states, indexed by capital letters I and J) at the same total energy show very similar behaviour if calculated from the Schrödinger equation, e.g. [equation \(A3.13.43\)](#), or the Pauli equation [\(A3.13.71\)](#),

$$p(t) = \mathbf{Y}(t)p(0), \quad (\text{A 3.13.71})$$

with \mathbf{Y} being given by:

$$\mathbf{Y}(t) = \exp(\mathbf{K}t), \quad (\text{A 3.13.72})$$

and the rate coefficient matrix elements in the limit of perturbation theory

$$K_{IJ} = 2\pi |V_{IJ}|^2 / \delta_I. \quad (\text{A 3.13.73})$$

In equation (A3.13.73), δ_I is the average angular frequency distance between quantum states within level I and $|V_{ij}|^2$ is the average square coupling matrix element (as angular frequency) between the quantum states in levels I and J (of total number of states N_I and N_J , respectively) and is given by:

$$|V_{IJ}|^2 = \frac{1}{N_I} \frac{1}{N_J} \sum_{i=1}^{N_I} \sum_{j=1}^{N_J} |V_{ij}|^2. \quad (\text{A 3.13.74})$$

Figure A3.13.14 seems to indicate that the Pauli equation is a strikingly good approximation for treating IVR under these conditions, involving even relatively few quantum states. This is, however, only true in this simple manner because we assume in the model that all the couplings are randomly distributed around their ‘typical’ values. This excludes any symmetry selection rules for couplings within the set of quantum states considered [177]. More generally, one has to consider only sets of quantum states with the same set of good (conserved) quantum numbers in such a master equation treatment. It is now well established that even in complex forming collisions leading to maximum energy transfer by IVR (section A3.13.3.3), conserved quantum numbers such as nuclear spin symmetry

-41-

and parity lead to considerable restrictions [177, 178]. More generally, one has to identify approximate symmetries on short time scales, which lead to further restrictions on the density of strongly coupled states [179]. Thus, the validity of a statistical master equation treatment for IVR in large polyatomic molecules is not obvious *a priori* and has to be established individually for at least classes of molecular systems, if not on a case by case basis.

Figure A3.13.15 shows a scheme for such a Pauli equation treatment of energy transfer in highly excited ethane, e.g. equation (A3.13.75), formed at energies above both thresholds for dissociation in chemical activation:



The figure shows the migration of energy between excited levels of the ultimately reactive C–C oscillator, the total energy being constant at $E/hc = 41\,000\text{ cm}^{-1}$ with a CC dissociation threshold of $31\,000\text{ cm}^{-1}$. The energy balance is thus given by:

(A 3.13.76)

$$E_{\text{tot}} = E_{\text{C-C}} + \sum_{i=1}^{17} E_i.$$

The microcanonical equilibrium distributions are governed by the densities ρ_{s-1} in the $(s-1) = 17$ oscillators (figure A3.13.15):

$$p_{\text{micro}}(n) = \rho_{s-1}^{(n)} \left\{ \sum_k \rho_{s-1}^{(k)} \right\}^{-1} \quad (\text{A 3.13.77})$$

where the 17 remaining degrees of freedom of ethane form essentially a ‘heat bath’. The kinetic master equation treatment of this model leads to steady-state populations shown in figure A3.13.16. This illustrates that the steady-state populations under conditions where reaction equation (A3.13.75) competes with IVR differ from the microcanonical equilibrium populations at high energy, and both differ from thermal distributions shown as lines (quantum or classical). Whereas the deviation from a thermal distribution is well understood and handled by standard statistical theories such as RRKM (chapter A3.12) and the statistical adiabatic channel model [76], the deviation from the microcanonical distribution would lead to an intramolecular nonequilibrium effect on the rates of reaction which so far has not been well investigated experimentally [37, 38 and 39].

-42-

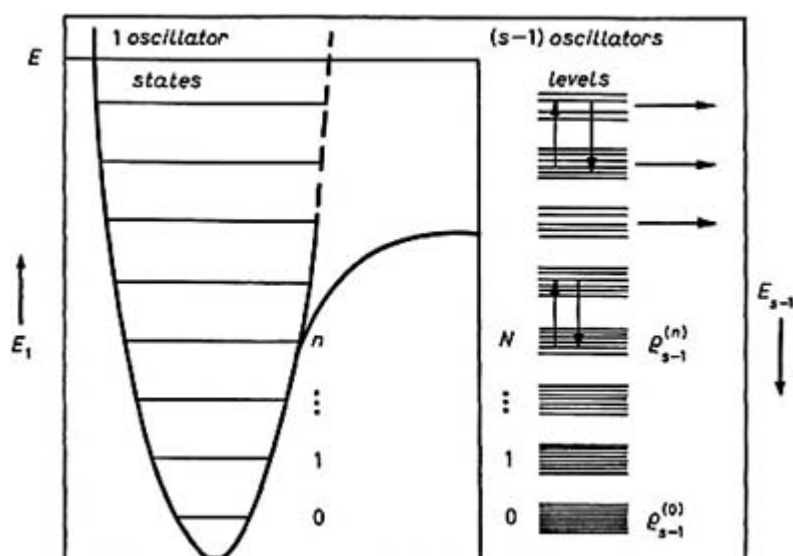


Figure A3.13.15. Master equation model for IVR in highly excited C_2H_6 . The left-hand side shows the quantum levels of the reactive CC oscillator. The right-hand side shows the levels with a high density of states from the remaining 17 vibrational (and torsional) degrees of freedom (from [38]).

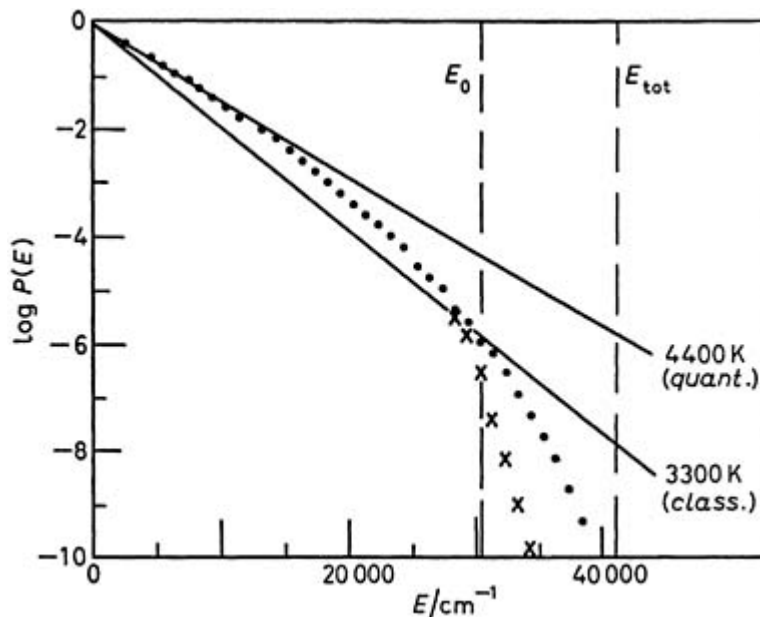


Figure A3.13.16. Illustration of the level populations (corresponding to the C–C oscillator states) from various treatments in the model of figure A3.13.15 for C_2H_6 at a total energy $E = (hc) 41\,000\text{ cm}^{-1}$ and a threshold energy $E_0 = (hc) 31\,000\text{ cm}^{-1}$. The points are microcanonical equilibrium distributions. The crosses result from the solution of the master equation for IVR at steady state and the lines are thermal populations at the temperatures indicated (from [38]: quant. is calculated with quantum densities of states, class. with classical mechanical densities.).

A 3.13.7 SUMMARIZING OVERVIEW ON ENERGY REDISTRIBUTION IN REACTING SYSTEMS

It has been understood for more than a century that energy redistribution is a key process in chemical reactions, including in particular the oldest process of chemical technology used by mankind: fire or combustion, where both radiative and collisional processes are relevant. Thus one might think that this field has reached a stage of maturity and saturation. Nothing could be further from the truth. While collisional energy transfer is now often treated in reaction systems in some detail, as is to some extent routine in unimolecular reactions, there remain plenty of experimental and theoretical challenges. In the master equation treatments, which certainly should be valid here, one considers a statistical, macroscopic reaction system consisting of reactive molecules in a mixture, perhaps an inert gas heat bath.

The understanding of the second process considered in this chapter, intramolecular energy redistribution within a single molecular reaction system, is still in its infancy. It is closely related to the challenge of finding possible schemes to control the dynamics of atoms in molecules and the related change of molecular structure during the course of a chemical reaction [10, 117, 154, 175], typically in the femtosecond time scale, which has received increasing attention in the last few decades [180, 181 and 182]. The border between fully quantum dynamical treatments, classical mechanical theories and, finally, statistical master equations for IVR type processes needs to be explored further experimentally and theoretically in the future. Unravelling details of the competition between energy redistribution and reaction in individual molecules remains an important task for the coming decades [37, 38, 39 and 40].

REFERENCES

- [1] van't Hoff J H 1884 *Études de Dynamique Chimique* (Amsterdam: F. Müller)
 - [2] Arrhenius S 1889 Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren *Z. Phys. Chem.* **4** 226–48
 - [3] Arrhenius S 1899 Zur Theorie der chemischen Reaktionsgeschwindigkeiten *Z. Phys. Chem.* **28** 317–35
 - [4] Calvert J G and Pitts J N 1966 *Photochemistry* (New York: Wiley)
 - [5] Simons J P 1977 The dynamics of photodissociation *Gas Kinetics and Energy Transfer* vol 2, ed P G Ashmore and R J Donovan (London: The Chemical Society) pp 58–95
 - [6] Turro N J 1978 *Modern Molecular Photochemistry* (Menlo Park, CA: Benjamin-Cummings)
 - [7] Schinke R 1993 *Photodissociation Dynamics* (Cambridge: Cambridge University Press)
 - [8] Perrin J 1922 On the radiation theory of chemical action *Trans. Faraday Soc.* **17** 546–72
 - [9] Lindemann F A 1922 Discussion contributions on the radiation theory of chemical action *Trans. Faraday Soc.* **17** 598–9
-

-44-

- [10] Quack M 1998 Multiphoton excitation *Encyclopedia of Computational Chemistry* vol 3, ed R P von Schleyer *et al* (New York: Wiley) pp 1775–91
- [11] Quack M and Troe J 1976 Unimolecular reactions and energy transfer of highly excited molecules *Gas Kinetics and Energy Transfer* vol 2, ch 5, ed P G Ashmore and R J Donovan (London: The Chemical Society) pp 175–238 (a review of the literature published up to early 1976)
- [12] Nikitin E E 1974 *Theory of Elementary Atomic and Molecular Processes in Gases* (Oxford: Clarendon)
- [13] Troe J 1975 Unimolecular reactions: experiment and theory *Physical Chemistry. An Advanced Treatise* vol VIB, ed H Eyring, D Henderson and W Jost (New York: Academic) pp 835–929
- [14] Rice S A 1975 Some comments on the dynamics of primary photochemical processes *Excited States* ed E C Lim (New York: Academic) pp 111–320
- [15] Light J C, Ross J and Shuler K E 1969 Rate coefficients, reaction cross sections and microscopic reversibility *Kinetic Processes in Gases and Plasmas* ed A R Hochstim (New York: Academic) pp 281–320
- [16] Stockburger M 1973 *Organic Molecular Photophysics* vol 1, ed J Birks (New York: Wiley) p 57
- [17] Montroll E W and Shuler K E 1958 The application of the theory of stochastic processes to chemical kinetics *Adv. Chem. Phys.* **1** 361–99
- [18] Pritchard H O 1975 *Reaction Kinetics* vol 1, ed P G Ashmore (London: The Chemical Society)
- [19] Quack M and Troe J 1981 Current aspects of unimolecular reactions *Int. Rev. Phys. Chem.* **1** 97–147
- [20] Golden D M and Benson S W 1975 *Physical Chemistry. An Advanced Treatise* vol VII (New York: Academic) p 57
- [21] Tardy D C and Rabinovitch B S 1977 Intermolecular vibrational energy transfer in thermal unimolecular systems *Chem. Rev.* **77** 369–408

- [22] Troe J 1978 Atom and radical recombination reactions *Ann. Rev. Phys. Chem.* **29** 223–50
- [23] Hippler H and Troe J 1989 *Advances in Gas Phase Photochemistry and Kinetics* ed M N R Ashfold and J E Baggott (London: Royal Society of Chemistry) pp 209–62
- [24] Kraijnovitch D J, Parmenter C S and Catlett D L Jr 1987 State-to-state vibrational transfer in atom-molecule collisions. Beams vs. bulbs *Chem. Rev.* **87** 237–88
- [25] Gilbert R G and Smith S C 1990 *Theory of Unimolecular and Recombination Reactions* (London: Blackwell)
- [26] Lambert J D 1977 *Vibrational and Rotational Relaxation in Gases* (Oxford: Oxford University Press)
- [27] Weitz E and Flynn G W 1981 Vibrational energy flow in the ground electronic states of polyatomic molecules *Adv. Chem. Phys.* **47** 185–235
- [28] Oref I and Tardy D C 1990 Energy transfer in highly excited large polyatomic molecules *Chem. Rev.* **90** 1407–45
- [29] Weston R E and Flynn G W 1992 Relaxation of molecules with chemically significant amounts of vibrational energy: the dawn of the quantum state resolved era *Ann. Rev. Phys. Chem.* **43** 559–89
-

-45-

- [30] Howard M J and Smith I W M 1983 The kinetics of radical-radical processes in the gas phase *Prog. Reaction Kin.* **12** 57–200
- [31] Orr B J and Smith I W M 1987 Collision-induced vibrational energy transfer in small polyatomic molecules *J. Phys. Chem.* **91** 6106–19
- [32] Flynn G W, Parmenter C S and Wodtke A M 1996 Vibrational energy transfer *J. Phys. Chem.* **100** 12 817–38
- [33] Förster Th 1948 Zwischenmolekulare Energiewanderung und Fluoreszenz *Ann. Phys.* **2** 55–75
- [34] Juzeliunas G and Andrews D L 2000 Quantum electrodynamics of resonance energy transfer *Adv. Chem. Phys.* **112** 357–410
- [35] Quack M 1982 Reaction dynamics and statistical mechanics of the preparation of highly excited states by intense infrared radiation *Adv. Chem. Phys.* **50** 395–473
- [36] Jortner J, Rice S A and Hochstrasser R M 1969 Radiationless transitions in photochemistry *Adv. Photochem.* **7** 149
- [37] Quack M and Kutzelnigg W 1995 Molecular spectroscopy and molecular dynamics: theory and experiment *Ber. Bunsenges. Phys. Chem.* **99** 231–45
- [38] Quack M 1981 Statistical mechanics and dynamics of molecular fragmentation *Nuovo Cimento B* **63** 358–77
- [39] Quack M 1995 Molecular femtosecond quantum dynamics between less than yoctoseconds and more than days: experiment and theory *Femtosecond Chemistry* ed J Manz and L Woeste (Weinheim: Verlag Chemie) pp 781–818
- [40] Oppenheim I, Shuler K E and Weiss G H 1977 *Stochastic Processes in Chemical Physics, The Master Equation* (Cambridge, MA: MIT Press)
- [41] van Kampen N G 1981 *Stochastic Processes in Physics and Chemistry* (Amsterdam: North-Holland)

- [42] Gear C W 1971 *Numerical Initial Value Problems in Ordinary Differential Equations* (Englewood Cliffs, NJ: Prentice-Hall)
- [43] Shampine S 1994 *Numerical Solutions of Ordinary Differential Equations* (New York: Chapman and Hall)
- [44] Troe J 1977 Theory of thermal unimolecular reactions at low pressures. I. Solutions of the master equation *J. Chem. Phys.* **66** 4745–57
- [45] Troe J 1977 Theory of thermal unimolecular reactions at low pressures. II. Strong collision rate constants. Applications *J. Chem. Phys.* **66** 4758
- [46] Troe J 1983 Theory of thermal unimolecular reactions in the fall-off range. I. Strong collision rate constants *Ber. Bunsenges. Phys. Chem.* **87** 161–9
- [47] Quack M 1979 Master equations for photochemistry with intense infrared light *Ber. Bunsenges. Phys. Chem.* **83** 757–75
- [48] Dove J E, Nip W and Teitelbaum H 1975 *Proc. XVth Int. Symp. on Combustion* (The Combustion Institute) p 903
-

-46-

- [49] Olschewski H A, Troe J and Wagner H G 1966 Niederdruckbereich und Hochdruckbereich des unimolekularen N_2O -Zerfalls *Ber. Bunsenges. Phys. Chem.* **70** 450
- [50] Martinengo A, Troe J and Wagner H G 1966 *Z. Phys. Chem. (Frankfurt)* **51** 104
- [51] Johnston H S 1951 Interpretation of the data on the thermal decomposition of nitrous oxide *J. Chem. Phys.* **19** 663–7
- [52] Venkatesh P K, Dean A M, Cohen M H and Carr R W 1999 Master equation analysis of intermolecular energy transfer in multiple-well, multiple-channel unimolecular reactions. II. Numerical methods and application to the mechanism of the $\text{C}_2\text{H}_5 + \text{O}_2$ reaction *J. Chem. Phys.* **111** 8313
- [53] Smith K and Thomson R M 1978 *Computer Modelling of Gas Lasers* (New York: Plenum)
- [54] Hirschfelder J D, Curtiss C F and Bird R B 1964 *Molecular Theory of Gases and Liquids* (New York: Wiley)
- [55] Tabor M, Levine R D, Ben-Shaul A and Steinfeld J I 1979 Microscopic and macroscopic analysis of non-linear master equations: vibrational relaxation of diatomic molecules *Mol. Phys.* **37** 141–58
- [56] Treanor C E, Rich J W and Rehm R G 1968 Vibrational relaxation of anharmonic oscillators with exchange-dominated collisions *J. Chem. Phys.* **48** 1798–807
- [57] McCaffery A J 1999 Quasiresonant vibration–rotation transfer: a kinematic interpretation *J. Chem. Phys.* **111** 7697
- [58] Levine R D and Bernstein R B 1987 *Molecular Reaction Dynamics and Chemical Reactivity* (New York: Oxford University Press)
- [59] Steinfeld J I and Klemperer W 1965 Energy-transfer processes in monochromatically excited iodine molecules. I. Experimental results *J. Chem. Phys.* **42** 3475–97
- [60] Gianturco F A 1979 *The Transfer of Molecular Energies by Collision* (Heidelberg: Springer)

- [61] Bowman J M (ed) 1983 *Molecular Collision Dynamics* (Berlin: Springer)
- [62] Hutson J M and Green S 1994 MOLSCAT computer code, version 14, distributed by Collaborative Computational Project No 6 of the Engineering and Physical Sciences Research Council (UK)
- [63] Alexander M H and Manolopoulos D E 1987 A stable linear reference potential algorithm for solution of the quantum close-coupled equations in molecular scattering theory *J. Chem. Phys.* **86** 2044–50
- [64] Bodo E, Gianturco F A and Paesani F 2000 Testing intermolecular potentials with scattering experiments: He–CO rotationally inelastic collisions *Z. Phys. Chem., NF* **214** 1013–34
- [65] Fluendy M A D and Lawley K P 1973 *Applications of Molecular Beam Scattering* (London: Chapman and Hall)
- [66] Faubel M and Toennies J P 1977 Scattering studies of rotational and vibrational excitation of molecules *Adv. Atom. Mol. Phys.* **13** 229
- [67] Faubel M 1983 Vibrational and rotational excitation in molecular collisions *Adv. Atom. Mol. Phys.* **19** 345
- [68] Bunker D L 1971 *Methods in Computational Physics* vol 10, ed B Alder (New York: Academic)
-

-47-

- [69] Lenzer T, Luther K, Troe J, Gilbert R G and Lim K F 1995 Trajectory simulations of collisional energy transfer in highly excited benzene and hexafluorobenzene *J. Chem. Phys.* **103** 626–41
- [70] Grigoleit U, Lenzer T and Luther K 2000 Temperature dependence of collisional energy transfer in highly excited aromatics studied by classical trajectory calculations *Z. Phys. Chem., NF* **214** 1065–85
- [71] Barker J R 1984 Direct measurement of energy transfer in rotating large molecules in the electronic ground state *J. Chem. Phys.* **88** 11
- [72] Miller L A and Barker J R 1996 Collisional deactivation of highly vibrationally excited pyrazine *J. Chem. Phys.* **105** 1383–91
- [73] Hippler H, Troe J and Wendelken H J 1983 Collisional deactivation of vibrationally highly excited polyatomic molecules. II. Direct observations for excited toluene *J. Chem. Phys.* **78** 6709
- [74] Hold U, Lenzer T, Luther K, Reihs K and Symonds A C 2000 Collisional energy transfer probabilities of highly excited molecules from kinetically controlled selective ionization (KCSI). I. The KCSI technique: experimental approach for the determination of $P(E',E)$ in the quasicontinuous energy range *J. Chem. Phys.* **112** 4076–89
- [75] Steinfeld J I, Ruttenberg P, Millot G, Fanjoux G and Lavorel B 1991 Scaling laws for inelastic collision processes in diatomic molecules *J. Phys. Chem.* **95** 9638–47
- [76] Quack M and Troe J 1974 Specific rate constants of unimolecular processes II. Adiabatic channel model *Ber. Bunsenges. Phys. Chem.* **78** 240–52
- [77] Quack M and Troe J 1998 Statistical adiabatic channel model *Encyclopedia of Computational Chemistry* vol 4, ed P von Ragué Schleyer *et al* (New York: Wiley) pp 2708–26
- [78] Quack M and Troe J 1975 Complex formation in reactive and inelastic scattering: statistical adiabatic channel model of unimolecular processes III *Ber. Bunsenges. Phys. Chem.* **79** 170–83
- [79] Quack M 1979 Quantitative comparison between detailed (state selected) *relative* rate data and averaged (thermal) *absolute* rate data for complex forming reactions *J. Phys. Chem.* **83** 150–8

- [80] Clary D C, Gilbert R G, Bernshtein V and Oref I 1995 Mechanisms for super collisions *Faraday Discuss. Chem. Soc.* **102** 423–33
- [81] Marquardt R, Quack M, Stohner J and Sutcliffe E 1986 Quantum-mechanical wavepacket dynamics of the CH group in the symmetric top X_3CH compounds using effective Hamiltonians from high-resolution spectroscopy *J. Chem. Soc., Faraday Trans. 2* **82** 1173–87
- [82] Herzberg G 1966 *Molecular Spectra and Molecular Structure III. Electronic Spectra and Electronic Structure of Polyatomic Molecules* (New York: Van Nostrand-Reinhold) (reprinted in 1991)
- [83] Rice S A 1981 An overview of the dynamics of intramolecular transfer of vibrational energy *Adv. Chem. Phys.* **47** 117–200
- [84] Beil A, Luckhaus D, Quack M and Stohner J 1997 Intramolecular vibrational redistribution and unimolecular reactions: concepts and new results on the femtosecond dynamics and statistics in CHBrCIF *Ber. Bunsenges. Phys. Chem.* **101** 311–28

-48-

- [85] Quack M 1993 Molecular quantum dynamics from high resolution spectroscopy and laser chemistry *J. Mol. Struct.* **292** 171–95
- [86] Heller E J 1975 Time-dependent approach to semiclassical dynamics *J. Chem. Phys.* **62** 1544–55
- [87] Heller E J 1981 The semiclassical way to molecular spectroscopy *Acc. Chem. Res.* **14** 368–78
- [88] Heller E J 1983 The correspondence principle and intramolecular dynamics *Faraday Discuss. Chem. Soc.* **75** 141–53
- [89] Noid D W, Koszykowski M L and Marcus R A 1981 Quasiperiodic and stochastic behaviour in molecules *Ann. Rev. Phys. Chem.* **32** 267–309
- [90] Marcus R A 1983 On the theory of intramolecular energy transfer *Faraday Discuss. Chem. Soc.* **75** 103–15
- [91] Hose G and Taylor H S 1982 A quantum analog to the classical quasiperiodic motion *J. Chem. Phys.* **76** 5356–64
- [92] Wyatt R E, Hose G and Taylor H S 1983 Mode-selective multiphoton excitation in a model system *Phys. Rev. A* **28** 815–28
- [93] Heather R and Metiu H 1985 Some remarks concerning the propagation of a Gaussian wave packet trapped in a Morse potential *Chem. Phys. Lett.* **118** 558–63
- [94] Sawada S and Metiu H 1986 A multiple trajectory theory for curve crossing problems obtained by using a Gaussian wave packet representation of the nuclear motion *J. Chem. Phys.* **84** 227–38
- [95] Miller W H 1974 Classical-limit quantum mechanics and the theory of molecular collisions *Adv. Chem. Phys.* **25** 69–177
- [96] van Gunsteren W F and Berendsen H J C 1990 Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry *Angew. Chem. Int. Ed. Engl.* **29** 992–1023
- [97] Marquardt R and Quack M 1991 The wave packet motion and intramolecular vibrational redistribution in CHX_3 molecules under infrared multiphoton excitation *J. Chem. Phys.* **95** 4854–67
- [98] Dübal H-R and Quack M 1980 Spectral bandshape and intensity of the C–H chromophore in the

infrared spectra of CF_3H and $\text{C}_4\text{F}_9\text{H}$ *Chem. Phys. Lett.* **72** 342–7

Dübal H-R and Quack M 1981 High resolution spectroscopy of fluoroform *Chem. Phys. Lett.* **80** 439–44

- [99] Dübal H-R and Quack M 1984 Tridiagonal Fermi resonance structure in the IR spectrum of the excited CH chromophore in CF_3H *J. Chem. Phys.* **81** 3779–91
- [100] Quack M 1990 Spectra and dynamics of coupled vibrations in polyatomic molecules *Ann. Rev. Phys. Chem.* **41** 839–74
- [101] Maynard A T, Wyatt R E and lung C 1995 A quantum dynamical study of CH overtones in fluoroform. I. A nine-dimensional *ab initio* surface, vibrational spectra and dynamics *J. Chem. Phys.* **103** 8372–90
- [102] Maynard A T, Wyatt R E and lung C 1997 A quantum dynamical study of CH overtones in fluoroform. II. Eigenstates of the $\nu_{\text{CH}} = 1$ and $\nu_{\text{CH}} = 2$ regions *J. Chem. Phys.* **106** 9483–96

-49-

- [103] Wyatt R E, lung C and Leforestier C 1992 Quantum dynamics of overtone relaxation in benzene. II. Sixteen-mode model for relaxation from $\text{CH}(\nu = 3)$ *J. Chem. Phys.* **97** 3477–86
- [104] Minehardt T A, Adcock J D and Wyatt R E 1999 Quantum dynamics of overtone relaxation in 30-mode benzene: a time-dependent local mode analysis for $\text{CH}(\nu = 2)$ *J. Chem. Phys.* **110** 3326–34
- [105] Gatti F, lung C, Leforestier C and Chapuisat X 1999 Fully coupled 6D calculations of the ammonia vibration–inversion–tunneling states with a split Hamiltonian pseudospectral approach *J. Chem. Phys.* **111** 7236–43
- [106] Luckhaus D 2000 6D vibrational quantum dynamics: generalized coordinate discrete variable representation and (a)diabatic contraction *J. Chem. Phys.* **113** 1329–47
- [107] Fehrensen B, Luckhaus D and Quack M 1999 Mode selective stereomutation tunnelling in hydrogen peroxide isotopomers *Chem. Phys. Lett.* **300** 312–20
- [108] Fehrensen B, Luckhaus D and Quack M 1999 Inversion tunneling in aniline from high resolution infrared spectroscopy and an adiabatic reaction path Hamiltonian approach *Z. Phys. Chem., NF* **209** 1–19
- [109] Zhang D H, Wu Q, Zhang J Z H, von Dirke M and Bai Z 1995 Exact full dimensional bound state calculations for $(\text{HF})_2$, $(\text{DF})_2$ and HFDF *J. Chem. Phys.* **102** 2315–25
- [110] Quack M and Suhm M A 1998 Spectroscopy and quantum dynamics of hydrogen fluoride clusters *Advances in Molecular Vibrations and Collision Dynamics, Vol. III Molecular Clusters* ed J Bowman and Z Bai (JAI Press) pp 205–48
- [111] Qiu Y and Bai Z 1998 Vibration–rotation–tunneling dynamics of $(\text{HF})_2$ and $(\text{HCl})_2$ from full-dimensional quantum bound state calculations *Advances in Molecular Vibrations and Collision Dynamics, Vol. I–II Molecular Clusters* ed J Bowman and Z Bai (JAI Press) pp 183–204
- [112] Loesch H J and Remscheid A 1990 Brute force in molecular reaction dynamics: a novel technique for measuring steric effects *J. Chem. Phys.* **93** 4779–90
- [113] Seideman T 1995 Rotational excitation and molecular alignment in intense laser fields *J. Chem. Phys.* **103** 7887–96

- [114] Friedrich B and Herschbach D 1995 Alignment and trapping of molecules in intense laser fields *Phys. Rev. Lett.* **74** 4623–6
- [115] Kim W and Felker P M 1996 Spectroscopy of pendular states in optical-field-aligned species *J. Chem. Phys.* **104** 1147–50
- [116] Quack M and Sutcliffe E 1983 Quantum interference in the IR-multiphoton excitation of small asymmetric-top molecules: ozone *Chem. Phys. Lett.* **99** 167–72
- [117] Quack M and Sutcliffe E 1984 The possibility of mode-selective IR-multiphoton excitation of ozone *Chem. Phys. Lett.* **105** 147–52
- [118] Kosloff R 1994 Propagation methods for quantum molecular dynamics *Ann. Rev. Phys. Chem.* **45** 145–78
- [119] Dey B D, Askar A and Rabitz H 1998 Multidimensional wave packet dynamics within the fluid dynamical formulation of the Schrödinger equation *J. Chem. Phys.* **109** 8770–82

-50-

- [120] Quack M 1992 Time dependent intramolecular quantum dynamics from high resolution spectroscopy and laser chemistry *Time Dependent Quantum Molecular Dynamics: Experiment and Theory. Proc. NATO ARW 019/92 (NATO ASI Ser. Vol 299)* ed J Broeckhove and L Lathouwers (New York: Plenum) pp 293–310
- [121] Quack M 1981 *Faraday Discuss. Chem. Soc.* **71** 309–11, 325–6, 359–64 (Discussion contributions on flexible transition states and vibrationally adiabatic models; statistical models in laser chemistry and spectroscopy; normal, local, and global vibrational states)
- [122] Quack M 1982 The role of intramolecular coupling and relaxation in IR-photochemistry *Intramolecular Dynamics, Proc. 15th Jerusalem Symp. on Quantum Chemistry and Biochemistry (Jerusalem, Israel, 29 March–1 April 1982)* ed J Jortner and B Pullman (Dordrecht: Reidel) pp 371–90
- [123] von Puttkamer K, Dübal H-R and Quack M 1983 Time-dependent processes in polyatomic molecules during and after intense infrared irradiation *Faraday Discuss. Chem. Soc.* **75** 197–210
- [124] von Puttkamer K, Dübal H R and Quack M 1983 Temperature-dependent infrared band structure and dynamics of the CH chromophore in $C_4F_9-C\equiv C-H$ *Chem. Phys. Lett.* **4–5** 358–62
- [125] Segall J, Zare R N, Dübal H R, Lewerenz M and Quack M 1987 Tridiagonal Fermi resonance structure in the vibrational spectrum of the CH chromophore in CHF_3 . II. Visible spectra *J. Chem. Phys.* **86** 634–46
- [126] Boyarkin O V and Rizzo T R 1996 Secondary time scales of intramolecular vibrational energy redistribution in CF_3H studied by vibrational overtone spectroscopy *J. Chem. Phys.* **105** 6285–92
- [127] Schulz P A, Sudbo A S, Kraijnovitch D R, Kwok H S, Shen Y R and Lee Y T 1979 Multiphoton dissociation of polyatomic molecules *Ann. Rev. Phys. Chem.* **30** 395–409
- [128] Zewail A H 1980 Laser selective chemistry—is it possible? *Phys. Today* Nov, 27–33
- [129] Quack M 1978 Theory of unimolecular reactions induced by monochromatic infrared radiation *J. Chem. Phys.* **69** 1282–307
- [130] Quack M, Stohner J and Sutcliffe E 1985 Time-dependent quantum dynamics of the picosecond

vibrational IR-excitation of polyatomic molecules *Time-Resolved Vibrational Spectroscopy, Proc. 2nd Int. Conf. Emil-Warburg Symp. (Bayreuth-Bischofsgrün, Germany, 3–7 June 1985)* ed A Laubereau and M Stockburger (Berlin: Springer) pp 284–8

- [131] Mukamel S and Shan K 1985 On the selective elimination of intramolecular vibrational redistribution using strong resonant laser fields *Chem. Phys. Lett.* **5** 489–94
- [132] Lupo D W and Quack M 1987 IR-laser photochemistry *Chem. Rev.* **87** 181–216
- [133] von Puttkamer K and Quack M 1989 Vibrational spectra of (HF)₂, (HF)_n and their D-isotopomers: mode selective rearrangements and nonstatistical unimolecular decay *Chem. Phys.* **139** 31–53
- [134] Quack M 1991 Mode selective vibrational redistribution and unimolecular reactions during and after IR-laser excitation *Mode Selective Chemistry* ed J Jortner, R D Levine and B Pullman (Dordrecht: Kluwer) pp 47–65
- [135] Crim F F 1993 Vibrationally mediated photodissociation: exploring excited state surfaces and controlling decomposition pathways *Ann. Rev. Phys. Chem.* **44** 397–428

-51-

- [136] Nesbitt D J and Field R W 1996 Vibrational energy flow in highly excited molecules: role of intramolecular vibrational redistribution *J. Phys. Chem.* **100** 12 735–56
- [137] He Y, Pochert J, Quack M, Ranz R and Seyfang G 1995 Discussion contributions on unimolecular reactions dynamics *J. Chem. Soc. Faraday Discuss.* **102** 358–62, 372–5
- [138] Siegman A E 1986 *Lasers* (Oxford: Oxford University Press)
- [139] Quack M and Sutcliffe E 1986 Program 515. URIMIR: unimolecular reactions induced by monochromatic infrared radiation *QCPE Bull.* **6** 98
- [140] Marquardt R, Quack M and Stohner J, at press
- [141] Marquardt R and Quack M 1996 Radiative excitation of the harmonic oscillator with applications to stereomutation in chiral molecules *Z. Phys. D* **36** 229–37
- [142] Cotting R, Huber J R and Engel V 1993 Interference effects in the photodissociation of FNO *J. Chem. Phys.* **100** 1040–8
- [143] Schinke R and Huber J R 1995 Molecular dynamics in excited electronic states—time-dependent wave packet studies *Femtosecond Chemistry: Proc. Berlin Conf. Femtosecond Chemistry (Berlin, March 1993)* (Weinheim: Verlag Chemie)
- [144] Schinke R and Huber J R 1993 Photodissociation dynamics of polyatomic molecules. The relationship between potential energy surfaces and the breaking of molecular bonds *J. Phys. Chem.* **97** 3463
- [145] Meier C and Engel V 1995 Pump–probe ionization spectroscopy of a diatomic molecule: sodium molecule as a prototype example *Femtosecond Chemistry: Proc. Berlin Conf. Femtosecond Chemistry (Berlin, March 1993)* (Weinheim: Verlag Chemie)
- [146] Meyer S and Engel V 1997 Vibrational revivals and the control of photochemical reactions *J. Phys. Chem.* **101** 7749–53
- [147] Flöthmann H, Beck C, Schinke R, Woywod C and Domcke W 1997 Photodissociation of ozone in the Chappuis band. II. Time-dependent wave packet calculations and interpretation of diffuse vibrational

structures *J. Chem. Phys.* **107** 7296–313

- [148] Loettgers A, Untch A, Keller H-M, Schinke R, Werner H-J, Bauer C and Rosmus P 1997 *Ab initio* study of the photodissociation of HCO in the first absorption band: three-dimensional wave packet calculations including the $\tilde{X}^2 A'$ $\tilde{A}^2 A''$ Renner–Teller coupling *J. Chem. Phys.* **106** 3186–204
- [149] Keller H-M and Schinke R 1999 The unimolecular dissociation of HCO. IV. Variational calculation of Siegert states *J. Chem. Phys.* **110** 9887–97
- [150] Braun M, Metiu H and Engel V 1998 Molecular femtosecond excitation described within the Gaussian wave packet approximation *J. Chem. Phys.* **108** 8983–8
- [151] Dübäl H-R, Ha T-K, Lewerenz M and Quack M 1989 Vibrational spectrum, dipole moment function, and potential energy surface of the CH chromophore in CHX₃ molecules *J. Chem. Phys.* **91** 6698–713

-52-

- [152] Lewerenz M and Quack M 1988 Vibrational spectrum and potential energy surface of the CH chromophore in CHD₃ *J. Chem. Phys.* **88** 5408–32
- [153] Marquardt R, Sanches Gonçalves N and Sala O 1995 Overtone spectrum of the CH chromophore in CHI₃ *J. Chem. Phys.* **103** 8391–403
- [154] Marquardt R, Quack M and Thanopoulos I 2000 Dynamical chirality and the quantum dynamics of bending vibrations of the CH chromophore in methane isotopomers *J. Phys. Chem. A* **104** 6129–49
- [155] Ha T-K, Lewerenz M, Marquardt R and Quack M 1990 Overtone intensities and dipole moment surfaces for the isolated CH chromophore in CHD₃ and CHF₃: experiment and *ab initio* theory *J. Chem. Phys.* **93** 7097–109
- [156] Hollenstein H, Marquardt R, Quack M and Suhm M A 1994 Dipole moment function and equilibrium structure of methane in an analytical, anharmonic nine-dimensional potential surface related to experimental rotational constants and transition moments by quantum Monte Carlo calculations *J. Chem. Phys.* **101** 3588–602
- [157] Schroedinger E 1926 Der stetige Übergang von der Mikro- zur Makromechanik *Naturwissenschaften* **14** 664–6
- [158] Kerner E H 1958 Note on the forced and damped oscillator in quantum mechanics *Can. J. Phys.* **36** 371–7
- [159] Marquardt R and Quack M 1989 Infrared-multiphoton excitation and wave packet motion of the harmonic and anharmonic oscillators: exact solutions and quasiresonant approximation *J. Chem. Phys.* **90** 6320–7
- [160] Marquardt R and Quack M 1989 Molecular motion under the influence of a coherent infrared-laser field *Infrared Phys.* **29** 485–501
- [161] Rice O K and Ramsperger H C 1927 Theories of unimolecular gas reactions at low pressures *J. Am. Chem. Soc.* **49** 1617–29
- [162] Kassel L S 1928 Studies in homogeneous gas reactions I *J. Phys. Chem.* **32** 225–42
- [163] Marcus R A and Rice O K 1951 The kinetics of the recombination of methyl radicals and iodine

atoms *J. Phys. Colloid. Chem.* **55** 894–908

- [164] Evans M G and Polanyi M 1935 Some applications of the transition state method to the calculation of reaction velocities, especially in solution *Trans. Faraday Soc.* **31** 875–94
- [165] Eyring H 1935 The activated complex in chemical reactions *J. Chem. Phys.* **3** 107–15
- [166] Hofacker L 1963 Quantentheorie chemischer Reaktionen *Z. Naturf. A* **18** 607–19
- [167] Robinson P J and Holbrook K A 1972 *Unimolecular Reactions* (New York: Wiley)
- [168] Quack M and Troe J 1981 Statistical methods in scattering *Theor. Chem. Adv. Perspect. B* **6** 199–276
- [169] Truhlar D G, Garrett B C and Klippenstein S J 1996 Current status of transition state theory *J. Phys. Chem.* **100** 12 771–800
- [170] Baer T and Hase W L 1996 *Unimolecular Reaction Dynamics. Theory and Experiment* (New York: Oxford University Press)

-53-

- [171] Holbrook K A, Pilling M J and Robertson S H 1996 *Unimolecular Reactions* 2nd edn (New York: Wiley)
- [172] Quack M 1990 The role of quantum intramolecular dynamics in unimolecular reactions *Phil. Trans. R. Soc. Lond. A* **332** 203–20
- [173] Luckhaus D and Quack M 1993 The role of potential anisotropy in the dynamics of the CH chromophore in CHX₃ (C_{3v}) symmetric tops *Chem. Phys. Lett.* **205** 277–84
- [174] Quack M 1989 Structure and dynamics of chiral molecules *Angew. Chem. Int. Ed. Engl.* **28** 571–86
- [175] Hervé S, Le Quéré F and Marquardt R 2001 Rotational and vibrational wave packet motion during the IR multiphoton excitation of HF *J. Chem. Phys.* **114** 826–35
- [176] Pauli W 1928 Über das H-Theorem vom Anwachsen der Entropie vom Standpunkt der neuen Quantenmechanik *Probleme der Modernen Physik (Festschrift zum 60. Geburtstag A. Sommerfelds)* ed P Debye (Leipzig: Hirzel) pp 30–45
- [177] Quack M 1977 Detailed symmetry selection rules for reactive collisions *Mol. Phys.* **34** 477–504
- [178] Cordonnier M, Uy D, Dickson R M, Kew K E, Zhang Y and Oka T 2000 Selection rules for nuclear spin modifications in ion-neutral reactions involving H₃⁺ *J. Chem. Phys.* **113** 3181–93
- [179] Quack M 1985 On the densities and numbers of rovibronic states of a given symmetry species: rigid and nonrigid molecules, transition states and scattering channels *J. Chem. Phys.* **82** 3277–83
- [180] Manz J and Wöste L (ed) 1995 *Femtosecond Chemistry: Proc. Berlin Conf. Femtosecond Chemistry (Berlin, March, 1993)* (Weinheim: Verlag Chemie)
- [181] Gaspard P and Burghardt I (ed) 1997 *XXth Solvay Conf. on Chemistry: Chemical Reactions and their Control on the Femtosecond Time Scale (Adv. Chem. Phys. 101)* (New York: Wiley)
- [182] *Femtochemistry 97* 1998 (The American Chemical Society) (*J. Phys. Chem.* **102** (23))
-

FURTHER READING

Ashmore P G and Donovan R J (ed) 1980 *Specialists Periodical Report: Gas Kinetics and Energy Transfer* vol 1 (1975), vol 2 (1977), vol 3 (1978), vol 4 (1980) (London: The Royal Society of Chemistry)

Bunsen. Discussion on 'Molecular spectroscopy and molecular dynamics. Theory and experiment' *Ber. Bunsengesellschaft Phys. Chem.* **99** 231–582

Bunsen. Discussion on 'Intramolecular processes' *Ber. Bunsenges. Phys. Chem.* **92** 209–450

Bunsen. Discussion on 'Unimolecular reactions' *Ber. Bunsenges. Phys. Chem.* **101** 304–635

Faraday Discuss. Chem. Soc. 1983 Intramolecular kinetics, No 75

-54-

Faraday Discuss. Chem. Soc. 1986 Dynamics of molecular photofragmentation, No 82

Faraday Discuss. Chem. Soc. 1995 Unimolecular dynamics, No 112

Gilbert R G and Smith S C 1990 *Theory of Unimolecular and Recombination Reactions* (Oxford: Blackwell)

Holbrook K, Pilling M J and Robertson S H 1996 *Unimolecular Reactions* (New York: Wiley)

Levine R D and Bernstein R B 1987 *Molecular Reaction Dynamics and Chemical Reactivity* (New York: Oxford University Press)

Quack M 1982 *Adv. Chem. Phys.* **50** 395–473

Quack M and Troe J 1981 Statistical methods in scattering *Theor. Chem.: Adv. Perspect.* B **6** 199–276

-1-

A3.14 Nonlinear reactions, feedback and self-organizing reactions

Stephen K Scott

A3.14.1 INTRODUCTION

A3.14.1.1 NONLINEARITY AND FEEDBACK

In the reaction kinetics context, the term 'nonlinearity' refers to the dependence of the (overall) reaction rate on the concentrations of the reacting species. Quite generally, the rate of a (simple or complex) reaction can be defined in terms of the rate of change of concentration of a reactant or product species. The variation of this rate with the extent of reaction then gives a 'rate–extent' plot. Examples are shown in [figure A3.14.1](#). In

the case of a first-order reaction, curve (i) in [figure A3.14.1\(a\)](#), the rate–extent plot gives a straight line: this is the only case of ‘linear kinetics’. For all other concentration dependences, the rate–extent plot is ‘nonlinear’: curves (ii) and (iii) in [figure A3.14.1\(a\)](#) correspond to second-order and half-order kinetics respectively. For all the cases in [figure A3.14.1\(a\)](#), the reaction rate is maximal at zero extent of reaction, i.e. at the beginning of the reaction. This is characteristic of ‘deceleratory’ processes. A different class of reaction types, [figure A3.14.1\(b\)](#), show rate–extent plots for which the reaction rate is typically low for the initial composition, but increases with increasing extent of reaction during an ‘acceleratory’ phase. The maximum rate is then achieved for some non-zero extent, with a final deceleratory stage as the system approaches complete reaction (chemical equilibrium). Curves (i) and (ii) in [figure A3.14.1\(b\)](#) are idealized representations of rate–extent curves observed in isothermal processes exhibiting ‘chemical feedback’. Feedback arises when a chemical species, typically an intermediate species produced from the initial reactants, influences the rate of (earlier) steps leading to its own formation. Positive feedback arises if the intermediate accelerates this process; negative feedback (or inhibition) arises if there is a retarding effect. Such feedback may arise chemically through ‘chain-branching’ or ‘autocatalysis’ in isothermal systems. Feedback may also arise through thermal effects: if the heat released through an exothermic process is not immediately lost from the system, the temperature of the reacting mixture will rise. A reaction showing a typical overall Arrhenius temperature dependence will thus show an increase in overall rate, potentially giving rise to further self-heating. Curve (iii) in [figure A3.14.1\(b\)](#) shows the rate–extent curve for an exothermic reaction under adiabatic conditions. Such feedback is the main driving force for the process known as combustion: endothermic reactions can similarly show self-cooling and inhibitory feedback. Specific examples of the origin of feedback in a range of chemical systems are presented below.

-2-

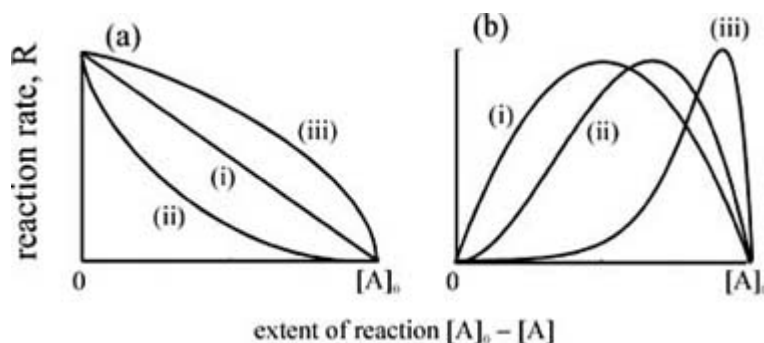
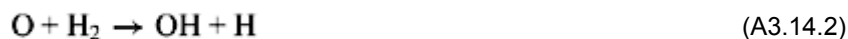


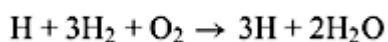
Figure A3.14.1. Rate-extent plots for (a) deceleratory and (b) acceleratory systems.

The branching cycle involving the radicals H, OH and O in the $\text{H}_2 + \text{O}_2$ reaction involves the three elementary steps



In step (1) and step (2) there is an increase from one to two ‘chain carriers’. (For brevity, step (x) is used to refer to equation (A3.14.x) throughout.) Under typical experimental conditions close to the first and second explosion limits (see [section A3.14.2.3](#)), step (2) and step (3) are fast relative to the rate determining step (1).

Combining (1) + (2) + 2 × (3) gives the overall stoichiometry



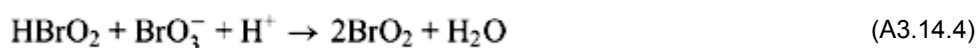
so there is a net increase of 2 H atoms per cycle. The rate of this overall step is governed by the rate of step (1), so we obtain

$$d[\text{H}]/dt = +2k_1[\text{O}_2][\text{H}]$$

where the + sign indicates that the rate of production of H atoms increases proportionately with that concentration.

-3-

In the bromate–iron clock reaction, there is an autocatalytic cycle involving the species intermediate species HBrO_2 . This cycle is comprised of the following non-elementary steps:



Step (5) is rapid due to the radical nature of BrO_2 , so the overall stoichiometric process given by (4) + 2 × (5), has the form



and an effective rate law

$$d[\text{HBrO}_2]/dt = +k_4[\text{BrO}_3^-][\text{H}^+][\text{HBrO}_2]$$

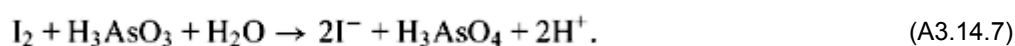
again showing increasing rate of production as the concentration of HBrO_2 increases.

In ‘Landolt’-type reactions, iodate ion is reduced to iodide through a sequence of steps involving a reductant species such as bisulfite ion (HSO_3^-) or arsenous acid (H_3AsO_3). The reaction proceeds through two overall stoichiometric processes. The Dushman reaction involves the reaction of iodate and iodide ions

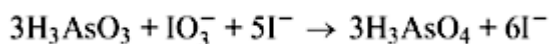


with an empirical rate law $R_\alpha = (k_1 + k_2[\text{I}^-])[\text{IO}_3^-][\text{I}^-][\text{H}^+]^2$.

The iodine produced in the Dushman process is rapidly reduced to iodide via the Roebuck reaction



If the initial concentrations are such that $[H_3AsO_3]_0/[IO_3^-]_0 > 3$, the system has excess reductant. In this case, the overall stoichiometry is given by (6) + 3 × (7) to give

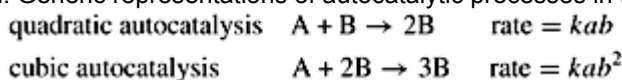


-4-

i.e. there is a net production of one iodide ion, with an overall rate given by R_α . At constant pH and for conditions such that $k_2[I^-] \gg k_1$, this can be approximated by

$$d[I^-]/dt = k[IO_3^-][I^-]^2$$

where $k = k_2[H^+]^2$. This again has the autocatalytic form, but now with the growth proportional to the square of the autocatalyst (I^-) concentration. Generic representations of autocatalytic processes in the form



where a and b are the concentrations of the reactant A and autocatalyst B respectively, are represented in [figure A3.14.1\(b\)](#) as curves (i) and (ii). The bromate–iron reaction corresponds to the quadratic type and the Landolt system to the cubic form.

A3.14.1.2 SELF-ORGANIZING SYSTEMS

‘Self-organization’ is a phrase referring to a range of behaviours exhibited by reacting chemical systems in which nonlinear kinetics and feedback mechanisms are operating. Examples of such behaviour include ignition and extinction, oscillations and chaos, spatial pattern formation and chemical wave propagation. There is a formal distinction between thermodynamically closed systems (no exchange of matter with the surroundings) and open systems [1]. In the former, the reaction will inevitably attain a unique state of chemical equilibrium in which the forward and reverse rates of every step in the overall mechanism become equal (detailed balance). This equilibrium state is temporally stable (the system cannot oscillate about equilibrium) and spatially uniform (under uniform boundary conditions). However, nonlinear responses such as oscillation in the concentrations of intermediate species can be exhibited as a transient phenomenon, provided the system is assembled with initial species concentrations sufficiently ‘far from’ the equilibrium composition (as is frequently the case). The ‘transient’ evolution may last for an arbitrary long (perhaps even a geological timescale), but strictly finite, period.

A3.14.2 CLOCK REACTIONS, CHEMICAL WAVES AND IGNITION

A3.14.2.1 CLOCK REACTIONS

The simplest manifestation of nonlinear kinetics is the clock reaction—a reaction exhibiting an identifiable ‘induction period’, during which the overall reaction rate (the rate of removal of reactants or production of final products) may be practically indistinguishable from zero, followed by a comparatively sharp ‘reaction event’ during which reactants are converted more or less directly to the final products. A schematic evolution of the reactant, product and intermediate species concentrations and of the reaction rate is represented in [figure A3.14.2](#). Two typical mechanisms may operate to produce clock behaviour.

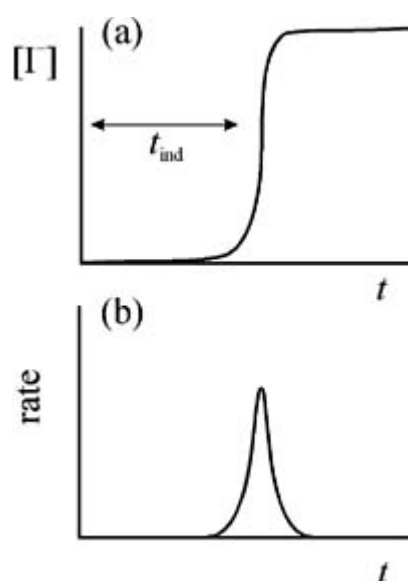
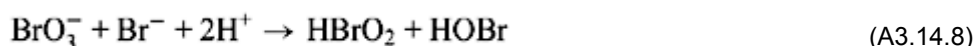


Figure A3.14.2. Characteristic features of a clock reaction, illustrated for the Landolt reaction, showing (a) variation of product concentration with induction period followed by sharp ‘reaction event’; (b) variation of overall reaction rate with course of reaction.

The Landolt reaction (iodate + reductant) is prototypical of an autocatalytic clock reaction. During the induction period, the absence of the feedback species (here iodide ion, assumed to have virtually zero initial concentration and formed from the reactant iodate only via very slow ‘initiation’ steps) causes the reaction mixture to become ‘kinetically frozen’. There is reaction, but the intermediate species evolve on concentration scales many orders of magnitude less than those of the reactant. The induction period depends on the initial concentrations of the major reactants in a manner predicted by integrating the overall rate cubic autocatalytic rate law, given in [section A3.14.1.1](#).

The bromate–ferroin reaction has a quadratic autocatalytic sequence, but in this case the induction period is determined primarily by the time required for the concentration of the ‘inhibitor’ bromide ion to fall to a critical low value through the reactions



Bromide ion acts as an inhibitor through step (9) which competes for HBrO₂ with the rate determining step for the autocatalytic process described previously, [step \(4\)](#) and [step \(5\)](#), Step (8) and Step (9) constitute a pseudo-first-order removal of Br[−] with HBrO₂ maintained in a low steady-state concentration. Only once [Br[−]] < [Br[−]]_{cr} = k₃[BrO₃[−]]/k₂ does [step \(3\)](#) become effective, initiating the autocatalytic growth and oxidation.

Clock-type induction periods occur in the spontaneous ignition of hydrocarbon–oxygen mixtures [2], in the setting of concrete and the curing of polymers [3]. A related phenomenon is the induction period exhibited

during the self-heating of stored material leading to thermal runaway [4]. A wide variety of materials stored in bulk are capable of undergoing a slow, exothermic oxidation at ambient temperatures. The consequent self-heating (in the absence of efficient heat transfer) leads to an increase in the reaction rate and, therefore, in the subsequent rate of heat release. The Semenov and Frank–Kamenetskii theories of thermal runaway address the relative rates of heat release and heat loss under conditions where the latter is controlled by Newtonian cooling and by thermal conductivity respectively. In the Frank–Kamenetskii form, the heat balance equation shows that the following condition applies for a steady-state balance between heat transfer and heat release:

$$\kappa \nabla^2 T - (-\Delta H)c^n A e^{-E/RT} = 0$$

where κ is the thermal conductivity and ∇^2 is the Laplacian operator appropriate to the particular geometry. The boundary condition specifies that the temperature must have some fixed value equal to the surrounding temperature T_a at the edge of the reacting mass: the temperature will then exceed this value inside the reacting mass, varying from point to point and having a maximum at the centre. Steady-state solutions are only possible if the group of quantities $(-\Delta H)a_0^2 c^n E A e^{-E/RT_a} / \kappa R T_a^2$, where a_0 is the half-width of the pile, is less than some critical value. If this group exceeds the critical value, thermal runaway occurs. For marginally supercritical situations where thermal balance is almost achieved, the runaway is preceded by an induction period as the temperature evolves on the Fourier time scale, $t_F = c_p a_0^2 / \kappa$, where c_p is the heat capacity. For large piles of low thermal conductivity, this may be of the order of months.

A3.14.2.2 REACTION–DIFFUSION FRONTS

A ‘front’ is a thin layer of reaction that propagates through a mixture, converting the initial reactants to final products. It is essentially a clock reaction happening in space. If the mixture is one of fuel and oxidant, the resulting front is known as a flame. In each case, the unreacted mixture is held in a kinetically frozen state due to the virtual absence of the feedback species (autocatalyst or temperature). The reaction is initiated locally to some point; for example, by seeding the mixture with the autocatalyst or providing a ‘spark’. This causes the reaction to occur locally, producing a high autocatalyst concentration/high temperature. Diffusion/conduction of the autocatalyst/heat then occurs into the surrounding mixture, initiating further reaction there. Front/flames propagate through this combination of diffusion and reaction, typically adopting a constant velocity which depends on the diffusion coefficient/thermal diffusivity and the rate coefficient for the reaction [5]. In each case, the speed c has the form $c \propto \sqrt{Dk}$. In gravitational fields, convective effects may arise due to density differences between the reactants ahead and the products behind the front. This difference may arise from temperature changes due to an exothermic/endothermic reaction or due to changes in molar volume between reactants and products—in some cases the two processes occur and may compete. In solid-phase combustion systems, such as those employed in self-propagating high-temperature synthesis (SHS) of materials, the steady flame may become unstable and a pulsing or oscillating flame develop—a feature also observed in propagating polymerization fronts [6].

A3.14.2.3 IGNITION, EXTINCTION AND BISTABILITY

In flow reactors there is a continuous exchange of matter due to the inflow and outflow. The species concentrations do not now attain the thermodynamic chemical equilibrium state—the system now has steady states which constitute a balance between the reaction rates and the flow rates. The steady-state concentrations (and temperature if the reaction is exo/endothermic) depend on the operating conditions through experimental parameters such as the flow rate. A plot of this dependence gives the steady-state locus, see [figure A3.14.3](#). With feedback reactions, this locus may fold back on itself, the fold points corresponding to critical conditions

for ignition or extinction—the plot is also then known as a ‘bifurcation diagram’. Between these points, the system exhibits bistability, as either the upper or the lower branch can be accessed; so the system may have different net reaction rates for identical operating conditions. Starting with a long residence time (low flow rates), the system lies on the ‘thermodynamic branch’, with a steady-state composition close to the equilibrium state. As the residence time is decreased (flow rate is increased), so the steady-state extent of conversion decreases, but at the turning point in the locus a further decrease in residence time causes the system to drop onto the lower, flow branch. This jump is known as ‘washout’ for solution-phase reactions and ‘extinction’ in combustion. The system now remains on the flow branch, even if the residence time is increased: there is hysteresis, with the system jumping back to the thermodynamic branch at the ‘ignition’ turning point in the locus. Many reactions exhibiting clock behaviour in batch reactors show ignition and extinction in flow systems. The determination of such bifurcation diagrams is a classic problem in chemical reactor engineering and of great relevance to the safe and efficient operation of flow reactors in the modern chemical industry. More complex steady-state loci, with isolated branches or multiple fold points leading to three accessible competing states have been observed in systems ranging from autocatalytic solution-phase reactions, smouldering combustion and in catalytic reactors [7]. Bistability has been predicted from certain models of atmospheric chemistry [8]. In the $\text{H}_2 + \text{O}_2$ and other branched-chain reactions, a balance equation for the radical species expresses the condition for a steady-state radical concentration. The condition for an ‘ignition limit’, i.e. for the marginal existence of a steady state, is that the branching and termination rates just balance. This can be expressed in terms of a ‘net branching factor’, $\phi = k_b - k_t$ where k_b and k_t are the pseudo-first-order rate constants for branching and termination respectively. For the hydrogen–oxygen system at low pressures, this has the form

$$2k_b[\text{O}_2] = k_{t1}[\text{O}_2][\text{M}] + k_{t2}$$

where k_{t1} corresponds to a three-body termination process (with [M] being the total gas concentration) and k_{t2} to a surface removal of H-atoms. This condition predicts a folded curve on the pressure–temperature plane—the first and second explosion limits, see [figure A3.14.4](#).

-8-

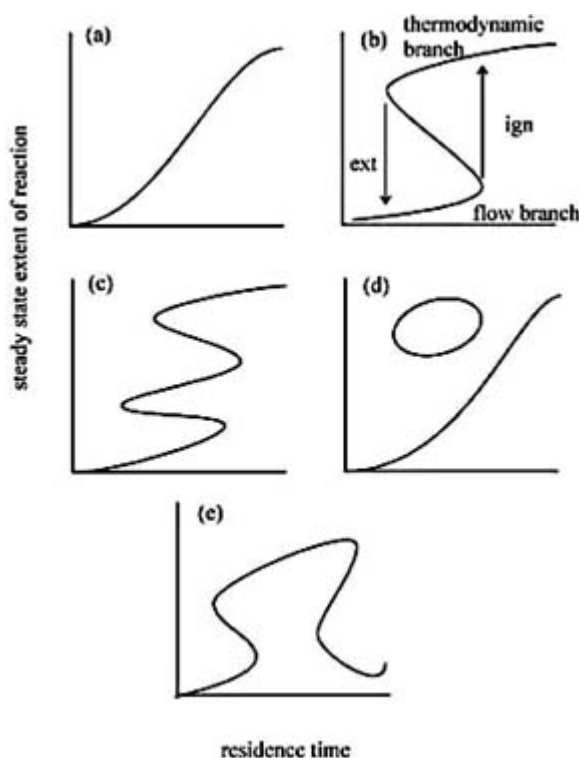


Figure A3.14.3. Example bifurcation diagrams, showing dependence of steady-state concentration in an open system on some experimental parameter such as residence time (inverse flow rate): (a) monotonic dependence; (b) bistability; (c) tristability; (d) isola and (e) mushroom.

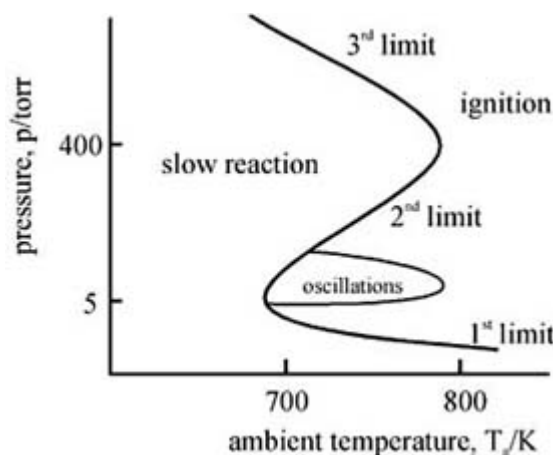


Figure A3.14.4. $P-T_a$ ignition limit diagram for $H_2 + O_2$ system showing first, second and third limits as appropriate to a closed reactor. The first and second limits have similar positions in a typical flow reactor, for which there is also a region of oscillatory ignition as indicated.

A3.14.3 OSCILLATIONS AND CHAOS

Despite previous worries about restrictions imposed by thermodynamics, the ability of homogeneous isothermal chemical systems to support long-lived (although strictly transient) oscillations in the concentrations of *intermediate* species even in closed reactors is now clearly established [9]. The reaction system studied in greatest detail is the Belousov–Zhabotinsky (BZ) reaction [10, 11 and 12], although the CIMA/CDIMA system involving chlorine dioxide, iodine and malonic acid is also of importance [13]. In flow reactors, oscillations amongst the concentrations of all species, including the reactants and products, are possible.

A3.14.3.1 THE BELOUSOV–ZHABOTINSKY REACTION

The BZ reaction involves the oxidation of an organic molecule (citric acid, malonic acid (MA)) by an acidified bromate solution in the presence of a redox catalyst such as the ferroin/ferrin or Ce^{3+}/Ce^{4+} couples. For a relatively wide range of initial reactant concentrations in a well-stirred beaker, the reaction may exhibit a short induction period, followed by a series of oscillations in the concentration of several intermediate species and also in the colour of the solution. The response of a bromide-ion-selective electrode and of a Pt electrode (responding to the redox couple) for such a system is shown in figure A3.14.5. Under optimal conditions, several hundred excursions are observed. In the redox catalyst concentrations, the oscillations are of apparently identical amplitude and only minutely varying period: the bromide ion concentration increases slowly with each complete oscillation and it is the slow build-up of this inhibitor, coupled with the consumption of the initial reactants, that eventually causes the oscillations to cease (well before the system approaches its equilibrium concentration).

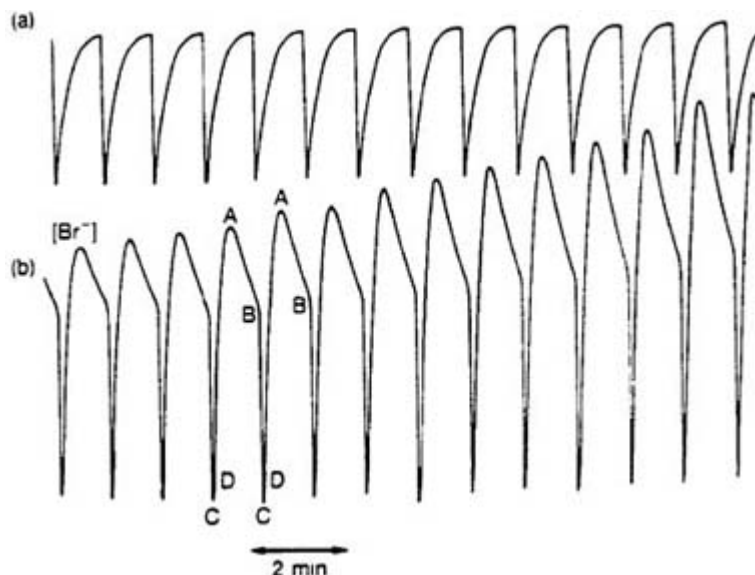


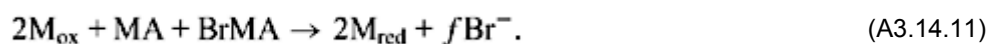
Figure A3.14.5. Experimental records from Pt and Br⁻-ion-sensitive electrode for the BZ reaction in batch showing regular oscillatory response.

-10-

The basic features of the oscillatory mechanism of the BZ reaction are given by the Field–Koros–Noyes (FKN) model [14]. This involves three ‘processes’—A, B and C. Process A involves [step \(8\)](#) and [step \(9\)](#) from [section A3.14.2.1](#), leading to removal of ‘inhibitor’ bromide ion. Process B involves [step \(3\)](#) and [step \(4\)](#) from [Section A3.14.1.1](#) and gives the autocatalytic oxidation of the catalyst. This growth is limited partly by the disproportionation reaction



The ‘clock’ is reset through process C. Bromomalonic acid, BrMA, is a by-product of processes A and B (possibly from HOBr) which reacts with the oxidized form of the redox catalyst. This can be represented as



Here, f is a stoichiometric factor and represents the number of bromide ions produced through this overall process for each two catalyst ions reduced. Because of the complex nature of this process, involving various radical species, f can lie in a range between 0 and ~ 3 depending on the $[\text{BrO}_3^-]/[\text{MA}]$ ratio and the $[\text{H}^+]$ concentration (note that these may change during the reaction).

The behaviour of the BZ system can be modelled semi-quantitatively by the ‘oregonator’ model [15]:

$$\frac{dx}{dt} = \frac{1}{\varepsilon} \left\{ x(1-x) - fz \frac{(x-q)}{(x+q)} \right\}$$

$$\frac{dz}{dt} = x - z$$

$$\frac{dx}{dt} = \frac{1}{\varepsilon} \left\{ x(1-x) - fz \frac{(x-q)}{(x+q)} \right\}$$

$$\frac{dz}{dt} = x - z$$

where x and z are (scaled) concentrations of HBrO_2 and M_{ox} respectively, and $\varepsilon = k_{11}[\text{Org}]/k_5[\text{BrO}_3^-][\text{H}^+]$ and $q = 2k_8k_{10}/k_9k_4$ are parameters depending on the rate coefficients and the initial concentrations [16], with $[\text{Org}]$ being the total concentration of organic species ($\text{MA} + \text{BrMA}$). Oscillations are observed in this model for $0.5 < f < 1 + \sqrt{2}$. More advanced models and detailed schemes account for the difference between systems with ferrion and cerium ion catalysts and for the effect of oxygen on the reaction [17].

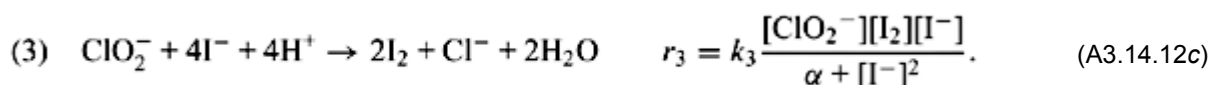
Under some conditions, it is observed that complex oscillatory sequences develop even in batch systems, typically towards the end of the oscillatory phase of the reaction. Transient ‘chaos’—see [section A3.14.3.3](#)—appears to be established [18].

-11-

In flow reactors, both simple period-1 oscillations (every oscillation has the same amplitude and period as its predecessor) and more complex periodic states can be established and sustained indefinitely. The first report of chemical chaos stemmed from the BZ system [19] (approximately contemporaneously with observations in the biochemical peroxidase reaction [20]). These observations were made in systems with relatively high flow rate and show complexity increasing through a sequence of ‘mixed-mode’ wave forms comprising one large excursion followed by n small peaks, with n increasing as the flow rate is varied. Subsequent period-doubling and other routes to chaos have been found in this system at low flow rates [21]. A relatively simple two-variable extension of the oregonator model can adequately describe these complex oscillations and chaos.

A3.14.3.2 THE CIMA/CDIMA SYSTEM

The reaction involving chlorite and iodide ions in the presence of malonic acid, the CIMA reaction, is another that supports oscillatory behaviour in a batch system (the chlorite–iodide reaction being a classic ‘clock’ system: the CIMA system also shows reaction–diffusion wave behaviour similar to the BZ reaction, see [section A3.14.4](#)). The initial reactants, chlorite and iodide are rapidly consumed, producing ClO_2 and I_2 which subsequently play the role of ‘reactants’. If the system is assembled from these species initially, we have the CDIMA reaction. The chemistry of this oscillator is driven by the following overall processes, with the empirical rate laws as given:



The concentrations of the major reactants ClO_2 and I_2 , along with H^+ , are treated as constants, so this is a two-variable scheme involving the concentrations of ClO_2^- and I^- . Step (12) constitutes the main feedback

process, which here is an inhibitory channel, with the rate decreasing as the concentration of iodide ion increases (for large $[I^-]$ the rate is inversely proportional to the concentration). Again, exploiting dimensionless terms, the governing rate equations for u (a scaled $[I^-]$) and v (scaled $[ClO_2^-]$) can be written as [22, 23]:

$$\frac{du}{dt} = a - u - \frac{4uv}{1+u^2} \frac{dv}{dt} = b \left(u - \frac{uv}{1+u^2} \right)$$

where a and b are constants depending on the rate coefficients and the initial concentrations of the reactants.

-12-

Another important reaction supporting nonlinear behaviour is the so-called FIS system, which involves a modification of the iodate–sulfite (Landolt) system by addition of ferrocyanide ion. The Landolt system alone supports bistability in a CSTR: the addition of an extra feedback channel leads to an oscillatory system in a flow reactor. (This is a general and powerful technique, exploiting a feature known as the ‘cross-shaped diagram’, that has led to the design of the majority of known solution-phase oscillatory systems in flow reactors [25].) The FIS system is one member of the important class of pH oscillators in which H^+ acts as an autocatalyst in the oxidation of a weak acid to produce a strong acid. Elsewhere, oscillations are observed in important chemical systems such as heterogeneously catalysed reactions or electrochemical and electrodisolution reactions [26].

A3.14.3.3 COMBUSTION SYSTEMS

Oscillatory behaviour occurs widely in the oxidation of simple fuels such as H_2 , CO and hydrocarbons. Even in closed reactors, the $CO + O_2$ reaction shows a ‘lighthouse effect’, with up to 100 periodic emissions of chemiluminescence accompanying the production of electronically excited CO_2 . Although also described as ‘oscillatory ignition’, each ‘explosion’ is associated with less than 1% fuel consumption and can be effectively isothermal, even for this strongly exothermic reaction. Many hydrocarbons exhibit ‘cool flame’ oscillations in closed reactors, with typically between two and seven ‘bursts’ of light emission (from excited HCHO) and reaction, accompanied by self-heating of the reacting mixture. In continuous flow reactors, these modes can be sustained indefinitely. Additionally, true periodic ignitions occur for both the $CO + O_2$ and $H_2 + O_2$ systems [27, 28]. The $p-T_a$ ‘ignition’ diagram for the $CO + O_2$ reaction under typical experimental conditions is shown in figure A3.14.6. Example oscillations observed at various locations on this diagram are displayed in figure A3.14.7. Within the region marked ‘complex oscillations’ the simple period-1 oscillation is replaced by waveforms that have different numbers of excursions in the repeating unit. The complexity develops through a ‘period-doubling’ sequence, with the waveform having 2^n oscillations per repeating unit, with n increasing with T_a . The range of experimental conditions over which the higher-order periodicities exist decreases in a geometric progression, with $n \rightarrow \infty$ leading to an oscillation with no repeating unit at some finite ambient temperature. Such *chaotic* responses exist over a finite range of experimental conditions and differ fundamentally from stochastic responses. Plotting the amplitude of one excursion against the amplitude of the next gives rise to a ‘next-maximum map’ (figure A3.14.8). This has a definite structure—a single-humped maximum—characteristic of a wide class of physical systems showing the period-doubling route to chaos.

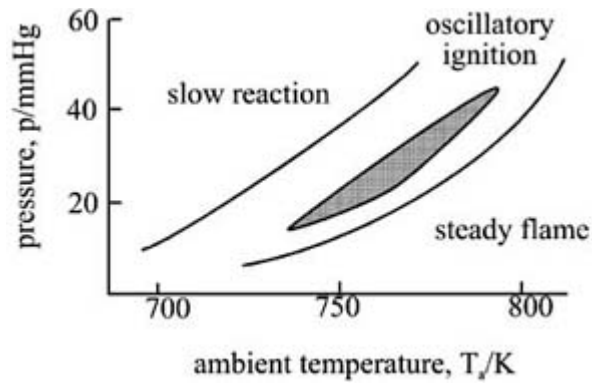


Figure A3.14.6. P - T_a ignition limit diagram for CO + O₂ system in a flow reactor showing location of ignition limits and regions of simple and complex (shaded area) oscillations.

-13-

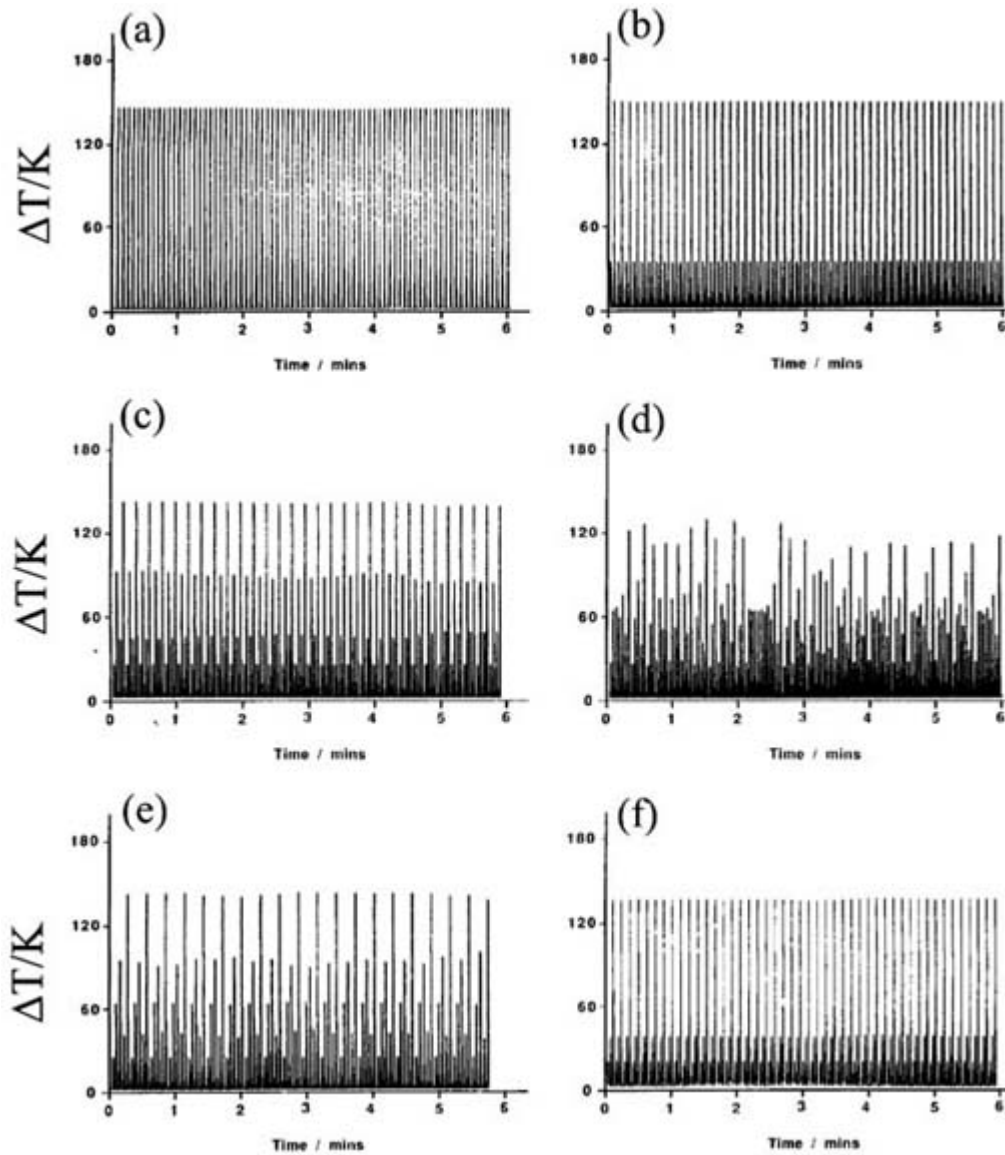


Figure A3.14.7. Example oscillatory time series for CO + O₂ reaction in a flow reactor corresponding to different P - T_a locations in [figure A3.14.6](#): (a) period-1; (b) period-2; (c) period-4; (d) aperiodic (chaotic) trace; (e) period-5; (f) period-3.

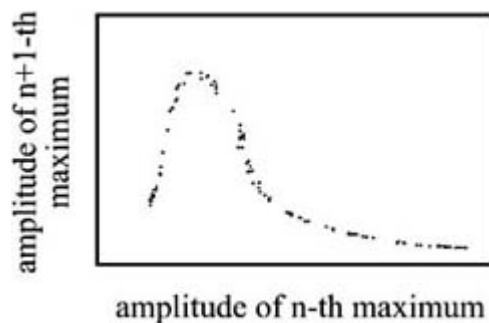


Figure A3.14.8. Next-maximum map obtained by plotting maximum temperature in one ignition against maximum in next ignition from trace (d) of [figure A3.14.7](#).

The mechanistic origin of simple and complex oscillation in the $\text{H}_2 + \text{O}_2$ system is well established. The basic oscillatory clockwork involves the self-acceleration of the reaction rate through the chain-branching cycle, [step \(1\)](#), [step \(2\)](#) and [step \(3\)](#) in [section A3.14.1.1](#) and the ‘self-inhibitory’ effect of H_2O production. Water is an inhibitor of the $\text{H}_2 + \text{O}_2$ system under these pressure and temperature conditions through its role in the main chain-termination step



where M is a ‘third-body’ species which removes energy, stabilizing the newly-formed HO_2 bond. H_2 and O_2 have third body efficiencies of 1 and 0.3 respectively (these are measured relative to H_2), but H_2O is substantially more effective, with an efficiency of ~ 6.3 relative to H_2 . Following an ignition, then, the rate of [step \(A3.14.13\)](#) is enhanced relative to the branching cycle, due to the now high concentration of H_2O , and further reaction is inhibited. The effect of the flow to the reactor, however, is to replace H_2O with H_2 and O_2 , thus lowering the overall third-body effectiveness of the mixture in the reactor. Eventually, the rate of [step \(A3.14.13\)](#) relative to the branching rate becomes sufficiently small for another ignition to occur. Complex oscillations require the further feedback associated with the self-heating accompanying ignition. A ‘minimal’ complex oscillator mechanism for this system has been determined [\[28\]](#).

Surprisingly, the origin of the complex oscillations and chaos in the $\text{CO} + \text{O}_2$ system (where trace quantities of H-containing species have a major influence on the reaction and, consequently, many of the reactions of the $\text{H}_2 + \text{O}_2$ system predominate) are far from established to date.

The ‘cool flames’ associated with the low-temperature oxidation of hydrocarbon have great relevance, being the fundamental cause of knock in internal combustion engines. Their mechanistic origin arises through a thermokinetic feedback [\[2\]](#). The crucial feature is the reaction through which O_2 reacts by addition to an alkyl radical R^\cdot



Under typical operating conditions, in the absence of self-heating from the reaction, the equilibrium for this step lies in favour of the product RO_2^\cdot . This species undergoes a series of intramolecular hydrogen-abstraction and further O_2 -addition steps before fragmentation of the carbon chain. This final step produces three radical

species, leading to a delayed, but overall branching of the radical chain ('degenerate branching'). This channel is overall an exothermic process, and the acceleration in rate associated with the branching leads to an increase in the gas temperature. This increase causes the equilibrium of [step \(A3.14.14\)](#) to shift to the left, in favour of the R' radical. The subsequent reaction channel for this species involves H-atom abstraction by O₂, producing the conjugate alkene. This is a significantly less exothermic channel, and the absence of branching means that the overall reaction rate and the rate of heat release fall. The temperature of the reacting mixture, consequently, decreases, causing a shift of the equilibrium back to the right. Complex oscillations have been observed in hydrocarbon oxidation in a flow reactor, although chaotic responses have not yet been reported.

A3.14.3.4 CONTROLLING CHAOS

The simple shape of the next-maximum map has been exploited in approaches to 'control' chaotic systems. The basic idea is that a system in a chaotic state is coexisting with an infinite number of unstable periodic states—indeed the chaotic 'strange attractor' is comprised of the period-1, period-2, period-4 and all other periodic solutions which have now become unstable. Control methods seek to select one of these unstable periodic solutions and to 'stabilize' it by applying appropriate but very small perturbations to the experimental operating conditions. Such control methods can also be adapted to allow an unstable state, such as the period-1 oscillation, to be 'tracked' through regions of operating conditions for which it would be naturally unstable. These techniques have been successfully employed for the BZ chaos as well as for chaos in lasers and various other physical and biological systems. For a full review and collection of papers see [\[29\]](#).

A3.14.4 TARGETS AND SPIRAL WAVES

The BZ and other batch oscillatory systems are capable of supporting an important class of reaction–diffusion structures. As mentioned earlier, clock reactions support one-off travelling wave fronts or flame, converting reactants to products. In an oscillatory system, the 'resetting' process can be expected to produce a 'wave back' following the front, giving rise to a propagating wave pulse. Furthermore, as the system is then returned more or less to its initial state, further wave initiation may be possible. A series of wave pulses travelling one after the other forms a wave train. If the solution is spread as a thin film, for example in a Petri dish, and initiation is from a point source, the natural geometry will be for a series of concentric, circular wave pulses—a *target 'pattern'* [\[16, 30, 31\]](#). An example of such reaction–diffusion structures in the BZ system is shown in [figure A3.14.9\(a\)](#). For such studies, the reactant solution is typically prepared with initial composition such that the system lies just outside the range for which it is spontaneously oscillatory (i.e. for f marginally in excess of $1 + \sqrt{2}$, see [section A3.14.3.1](#)) by increasing the initial malonic acid concentration relative to bromate. The system then sits in a stable steady state corresponding to the reduced form of the catalyst, and has the property of being *excitable*. An excitable system is characterized by (i) having a steady state; (ii) the steady state is stable to small perturbations and (iii) if the perturbation exceeds some critical or threshold value, the system responds by exhibiting an *excitation event*. For the BZ system, this excitation event is the oxidation of the redox catalyst, corresponding to process B with a local colour change in the vicinity of the perturbation (initiation) site.

This response is typically large compared with the critical stimulus, so the system acts as a 'nonlinear amplifier' of the perturbing signal. Following the excitation, the system eventually returns to the initial steady state and recovers its excitability. There is, however, a finite period, the *refractory period*, between the excitation and the recovery during which the system is unresponsive to further stimuli. These basic characteristics are summarized in [figure A3.14.10](#). Excitability is a feature not just of the BZ system, but is found widely throughout physical and, in particular, biological systems, with important examples in nerve signal transmission and co-ordinated muscle contraction [\[32\]](#).



Figure A3.14.9. Reaction–diffusion structures for an excitable BZ system showing (a) target and (b) spiral waves. (Courtesy of A F Taylor.)

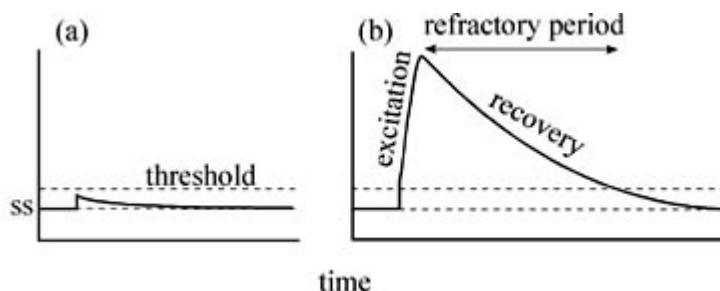


Figure A3.14.10. Schematic representation of important features of an excitable system (see the text for details).

The target structures shown in figure A3.14.9(a) reveal several levels of detail. Each ‘pacemaker site’ at the centre of a target typically corresponds to a position at which the system exhibits some heterogeneity, in some cases due to the presence of dust particles or defects of the dish surface. It is thought that these alter the local pH, so as to produce a composition in that vicinity such that the system is locally oscillatory, the spontaneous oscillations then serving as initiation events. Different sites have differing natural oscillatory frequencies, leading to the differing observed wavelengths of the various target structures. The speed of the waves also depends on the frequency of the pacemaker, through the so-called dispersion relation. The underlying cause of this for the BZ system is that the speed of a given front is dependent on the bromide ion (inhibitor) concentration into which it is propagating: the higher the pacemaker frequency, the less time the bromide ion concentration has to fall, so high-frequency (low-period) structures have lower propagation speeds.

-17-

If a wave pulse is broken, for example through mechanical disturbance, another characteristic feature of excitable media is that the two ‘ends’ then serve as sites around which the wave may develop into a pair of counter-rotating spirals, see figure A3.14.9(b). Once created, the spiral core is a persistent structure (in contrast to the target, in which case removal of the pacemaker heterogeneity prevents further initiation). The spiral structures have a wavelength determined by the composition of the bulk solution rather than the local properties at the core (although other features such as the *meandering* of the core may depend more crucially on local properties).

Targets and spirals have been observed in the CIMA/CDIMA system [13] and also in dilute flames (i.e. flames close to their lean flammability limits) in situations of enhanced heat loss [33]. In such systems, substantial fuel is left unburnt. Spiral waves have also been implicated in the onset of cardiac arrhythmia [32]: the normal contractive events occurring across the atria in the mammalian heart are, in some sense, equivalent to a wave pulse initiated from the sino-atrial node, which acts as a pacemaker. If this pulse becomes fragmented, perhaps by passing over a region of heart muscle tissue of lower excitability, then spiral structures (in 3D, these are *scroll waves*) or ‘re-entrant waves’ may develop. These have the incorrect

sequencing of contractions to squeeze blood from the atria to the ventricles and impair the operation of the heart. Similar waves have been observed in neuronal tissue and there are suggested links to pathological behaviour such as epilepsy and migraine [34]. Spirals and targets have also been observed accompanying the oxidation of CO on appropriate single-crystal catalysts, such as Pt(110), and in other heterogeneously catalysed systems of technological relevance [35] (see figure A3.14.11). The light-sensitive nature of the Ru (bipy)₂-catalysed BZ system has been exploited in many attempts to ‘control’ or influence spiral structures (for example to remove spirals). The excitable properties of the BZ system have also been used to develop generic methods for devising routes through complex mazes or to construct chemical equivalents of logic gates [36].



Figure A3.14.11. Spiral waves imaged by photoelectron electron microscopy for the oxidation of CO by O₂ on a Pt(110) single crystal under UHV conditions. (Reprinted with permission from [35], © The American Institute of Physics.)

A3.14.5 TURING PATTERNS AND OTHER STRUCTURES

A3.14.5.1 TURING PATTERNS

Diffusive processes normally operate in chemical systems so as to disperse concentration gradients. In a paper in 1952, the mathematician Alan Turing produced a remarkable prediction [37] that if *selective* diffusion were coupled with chemical feedback, the opposite situation may arise, with a spontaneous development of sustained spatial distributions of species concentrations from initially uniform systems. Turing’s paper was set in the context of the development of form (morphogenesis) in embryos, and has been adopted in some studies of animal coat markings. With the subsequent theoretical work at Brussels [1], it became clear that oscillatory chemical systems should provide a fertile ground for the search for experimental examples of these Turing patterns.

The basic requirements for a Turing pattern are:

- (i) the chemical reaction must exhibit feedback kinetics;

- (ii) the diffusivity of the feedback species must be less than those of the other species;
- (iii) for the patterns to be sustained, the system must be open to the inflow and outflow of reactants and products.

Requirement (i) is met particularly well by the BZ and CIMA/CDIMA reactions, although many chemical reactions with feedback are known. Requirements (ii) and (iii) were met almost simultaneously through the use of ‘continuous flow unstirred reactors’ (CFURs) in which the reaction is carried out in a dilute gel or membrane, with reactant free flows across the edges or faces of the gel. The incorporation of large indicator molecules such as starch into the gel is the key. This indicator is used with the CIMA/CDIMA system for which I_3^- is formed where the I^- concentration is high, and this binds as a complex to the starch to produce the characteristic blue colour. The complexed ion is temporarily immobilized compared with the free ion, thus reducing the effective diffusion coefficient in a kind of ‘reactive chromatography’ [24]. In this way the first laboratory examples of Turing patterns were produced in Bordeaux [38] and in Texas [39]: examples are shown in figure A3.14.12 and figure A3.14.13. Turing patterns have not been unambiguously observed in the BZ system as no similar method of reducing the diffusivity of the autocatalytic species $HBrO_2$ has been devised.

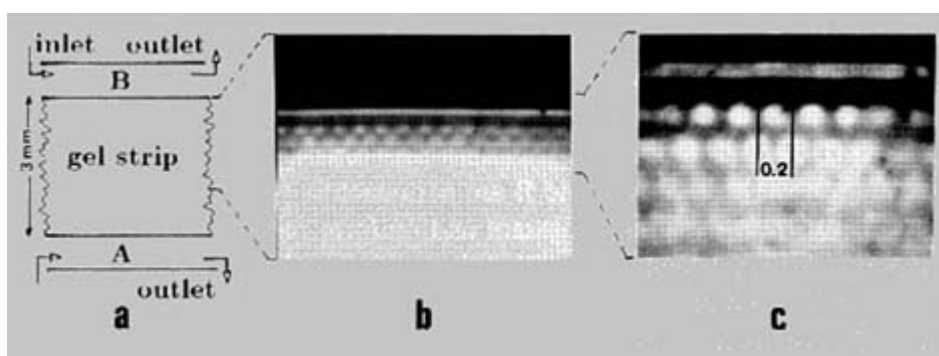


Figure A3.14.12. The first experimental observation of a Turing pattern in a gel strip reactor. Solutions containing separate components of the CIMA/CDIMA reaction are flowed along each edge of the strip and a spatial pattern along the horizontal axis develops for a range of experimental conditions. (Reprinted with permission from [38], © The American Physical Society.)

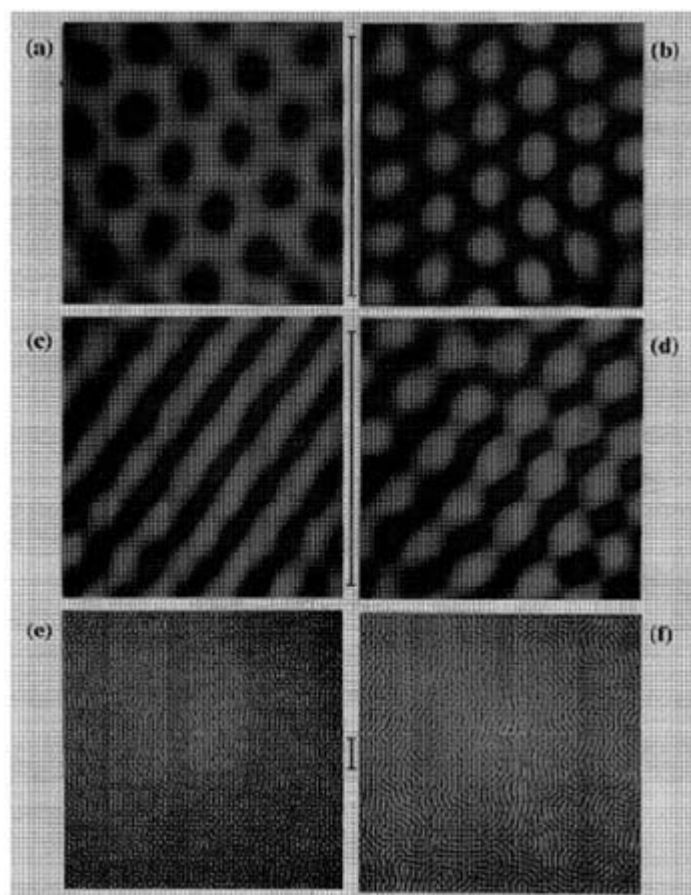


Figure A3.14.13. Further examples of the various Turing patterns observable in a 2D gel reactor. (a) and (b) spots, (c) and (d) stripes, (e) and (f): wider field of view showing long-range defects in basic structure. The scale bar alongside each figure represents 1 mm. (Reprinted with permission from [39], © The American Institute of Physics.)

A3.14.5.2 CELLULAR FLAMES

Such ‘diffusion-driven instabilities’ have been observed earlier in combustion systems. As early as 1892, Smithells reported the observation of ‘cellular flames’ in fuel-rich mixtures [40]. An example is shown in [figure A3.14.14](#). These were explained theoretically by Sivashinsky in terms of a ‘thermodiffusive’ mechanism [41]. The key feature here involves the role played by the Lewis number, Le , the ratio of the thermal to mass diffusivity. If $Le < 1$, which may arise with fuel-rich flame, for which H-atoms are the relevant species, of relatively low thermal conductivity (due to the high hydrocarbon content), a planar flame is unstable to spatial perturbations along the front. This mechanism has also been shown to operate for simple one-off chemical wave fronts, such as the iodate–arsenite system [42] and for various pH-driven fronts [43], if the diffusivity of I^- or H^+ are reduced via complexing strategies similar to that described above for the CIMA system.

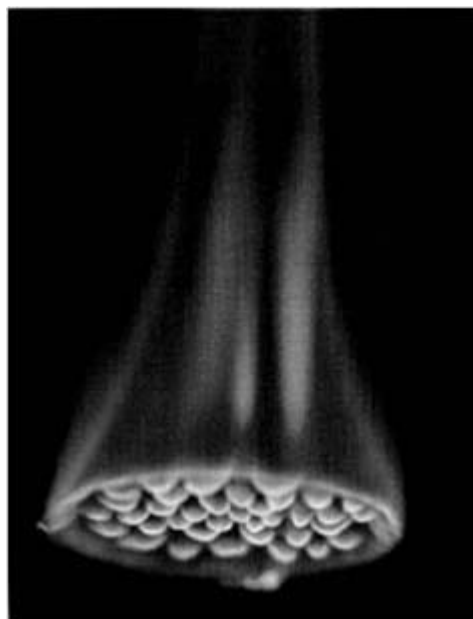


Figure A3.14.14. A cellular flame in butane oxidation on a burner. (Courtesy of A C McIntosh.)

A3.14.5.3 OTHER REACTION–DIFFUSION STRUCTURES

The search for Turing patterns led to the introduction of several new types of chemical reactor for studying reaction–diffusion events in feedback systems. Coupled with huge advances in imaging and data analysis capabilities, it is now possible to make detailed quantitative measurements on complex spatiotemporal behaviour. A few of the reactor configurations of interest will be mentioned here.

The Turing instability is specific in requiring the feedback species to be selectively immobilized. An related instability, the differential flow-induced chemical instability or DIFICI requires only that one active species be immobilized relative to the others [44]. The experimental configuration is simple: a column of ion exchange beads is loaded with one chemical component, for example, ferrion for the BZ system. The remaining species are prepared in solution and flowed through this column. Above some critical flow rate, a travelling spatial structure with narrow bands of oxidized reagent (in the BZ system) separated by a characteristic wavelength and propagating with a characteristic velocity (not equal to the liquid flow velocity) is established. This effect has been realized experimentally—see [figure A3.14.15](#).

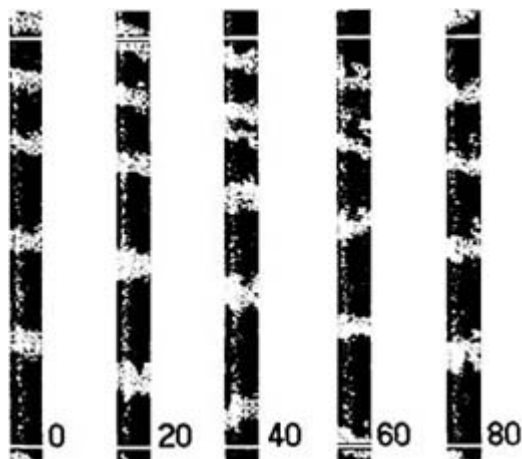


Figure A3.14.15. The differential flow-induced chemical instability (DIFICI) in the BZ reaction. (Reprinted with permission from [44], © The American Physical Society.)

If a fluid is placed between two concentric cylinders, and the inner cylinder rotated, a complex fluid dynamical motion known as Taylor–Couette flow is established. Mass transport is then by exchange between eddy vortices which can, under some conditions, be imagined as a substantially enhanced diffusivity (typically with ‘effective diffusion coefficients several orders of magnitude above molecular diffusion coefficients) that can be altered by varying the rotation rate, and with all species having the same diffusivity. Studies of the BZ and CIMA/CDIMA systems in such a Couette reactor [45] have revealed bifurcation through a complex sequence of front patterns, see [figure A3.14.16](#).

-22-

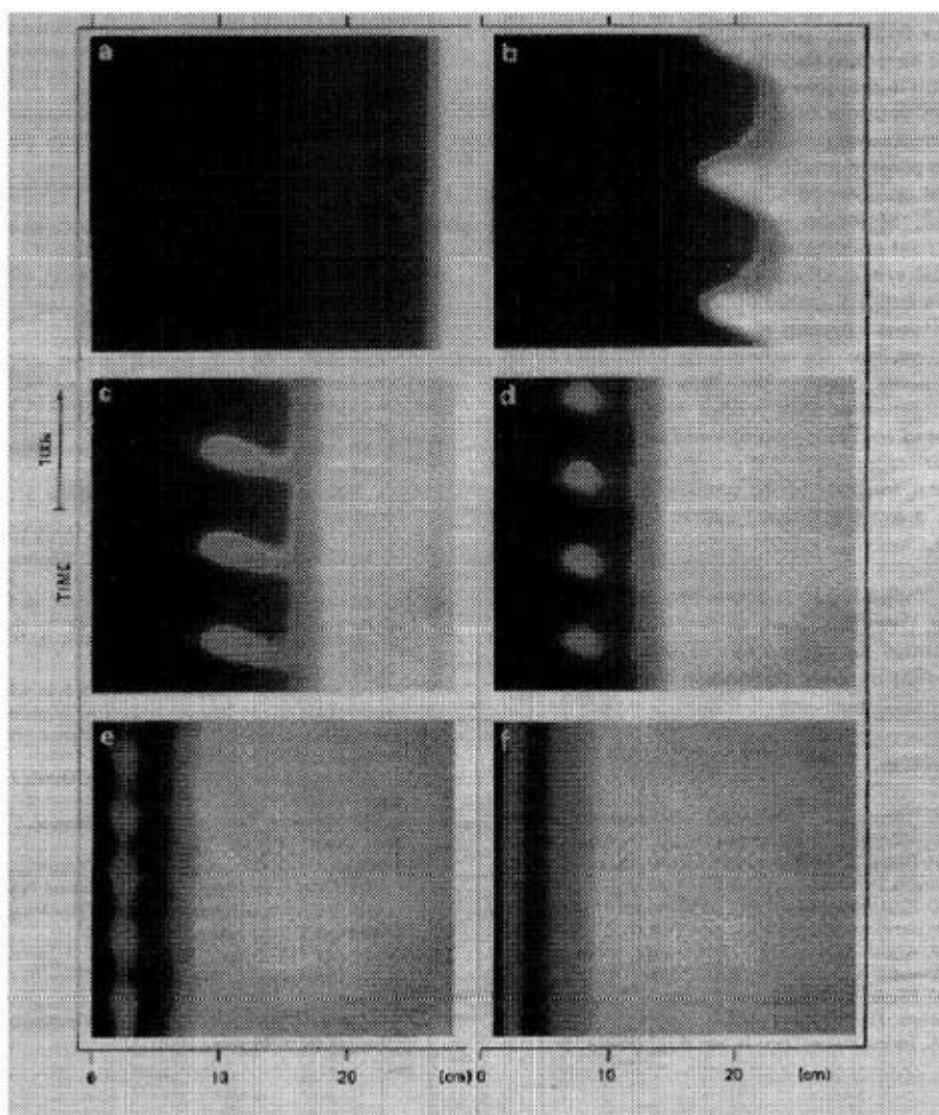


Figure A3.14.16. Spatiotemporal complexity in a Couette reactor: space–time plots showing the variation of position with time of fronts corresponding to high concentration gradients of IO_3^- in the CIMA/CDIMA reaction. (Reprinted with permission from Ouyang *et al* [45], © Elsevier Science Publishers 1989.)

The FIS reaction (section A3.14.3.2) has been studied in a CFUR and revealed a series of structures known as ‘serpentine patterns’; also, the birth, self-replication and death of ‘spots’, corresponding to regions of high concentration of particular species (see [figure A3.14.17](#) have been observed [46]).

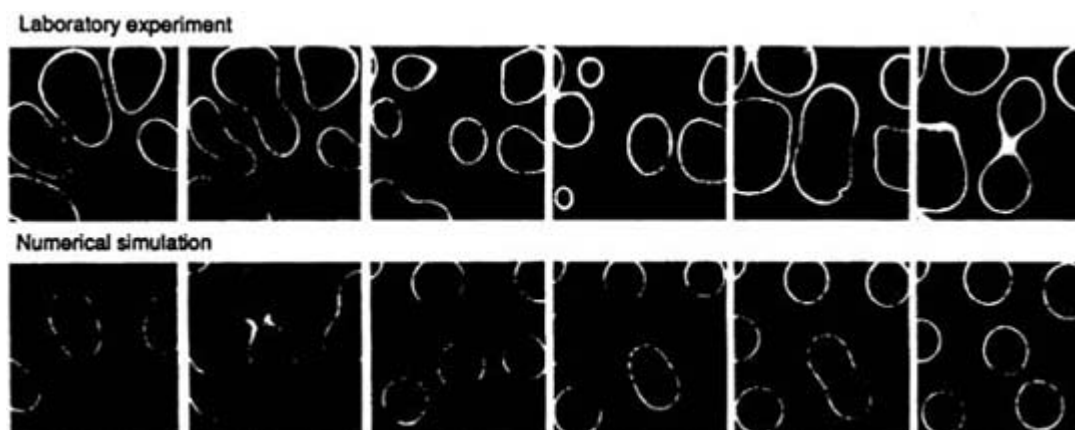


Figure A3.14.17. Self-replicating spots in the FIS reaction in a CFUR, comparing an experimental time sequence with numerical simulation based on a simple autocatalytic scheme. (Reprinted with permission from Lee *et al* [46], © Macmillan Magazines Ltd. 1994.)

A3.14.6 THEORETICAL METHODS

Much use has been, and continues to be, made of simplified model schemes representative of general classes of chemical or thermal feedback. The Oregonator and Lengyel–Epstein models for the BZ and CDIMA systems have been given earlier. Pre-eminent among the more abstracted caricature models is the Brusselator introduced by Prigogine and Lefever [47] which has the following form:



Here, A and B are regarded as ‘pool chemicals’, with concentrations regarded as imposed constants. The concentrations of the intermediate species X and Y are the variables, with D and E being product species whose concentrations do not influence the reaction rates. The reaction rate equations for [X] and [Y] can be written in the following dimensionless form:

$$\frac{dx}{dt} = A - Bx + yx^2 - x \quad \frac{dy}{dt} = Bx - yx^2.$$

Oscillations are found in this model if $B < B^* = 1 + A^2$.

A variation on this theme introduced by Gray and Scott, known as the ‘autocatalator’, is also widely exploited. This is often written in the form



so here A and B are equivalent to the Y and X in the brusselator and the main clockwork again involves the cubic autocatalytic [step \(15c\)](#) or step (16). The dimensionless equations here are

$$\frac{da}{dt} = \mu - \kappa a - ab^2 \quad \frac{db}{dt} = \kappa a + ab^2 - b \quad (\text{A3.14.17})$$

where μ is a scaled concentration of the reactant P and κ is a dimensionless rate coefficient for the ‘uncatalysed’ conversion of A to B in step (16). In this form, the model has oscillatory behaviour over a range of experimental conditions:

$$\mu_1^* < \mu < \mu_2^* \quad \text{with} \quad (\mu_{1,2}^*)^2 = \frac{1}{2}[1 - 2\kappa \pm (1 - 8\kappa)^{1/2}]. \quad (\text{A3.14.18})$$

Outside this range, the system approaches the steady state obtained by setting $da/dt = db/dt = 0$:

$$(a_{ss}, b_{ss}) = (\mu/(\mu^2 + \kappa), \mu). \quad (\text{A3.14.19})$$

The existence of an upper and a lower limit to the range of oscillatory behaviour is more typical of observed behaviour in chemical systems.

The autocatalytic driving [step \(16c\)](#) can also be taken on its own, or with the ‘decay’ [step \(16d\)](#) in models of open systems such as a CSTR, with an inflow of species A and, perhaps, of B. This system is then one of the simplest to show bistability and more complex steady-state loci of the type described in [section A3.14.2.3](#). Also, generic features of wave front propagation can be studied on the basis of this scheme [7]. A comprehensive account can be found in the book by Gray and Scott (see [Further Reading](#)). Essentially, this model has been used in the context of modelling the broad features of oscillations in glycolysis and, with some modification, for animal coat patterning through Turing-like mechanisms.

The main theoretical methods have in common the determination of the *stability* of steady-state or other

simple solutions to the appropriate form of the governing mass balance equations. Bifurcations from simple to more complex responses occur when such a solution loses its stability. Thus, the steady state given in [equation \(A3.14.19\)](#) does not cease to exist in the oscillatory region defined by [\(A3.14.18\)](#), but is now unstable, so that, if the system is perturbed, it departs from the steady state and moves to the (stable) oscillatory state which is also a solution of the reaction rate [equations \(A3.14.17\)](#).

In the generalized representation of the rate equations

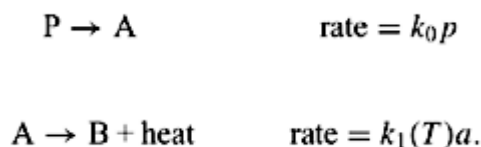
$$da/dt = f(a, b) \quad db/dt = g(a, b)$$

where f and g are functions of the species concentrations, a determining role is played by the eigenvalues of the Jacobian matrix J defined by

$$J = \begin{pmatrix} \partial f/\partial a & \partial f/\partial b \\ \partial g/\partial a & \partial g/\partial b \end{pmatrix}$$

evaluated with the steady-state concentrations. (This is readily generalized to n -variable systems, with J then being an $n \times n$ matrix.) Bifurcations corresponding to a turning or fold point in a steady-state locus (ignition or extinction point) occur if a real eigenvalue passes through zero. Equivalently, this arises if the determinant $\det J = 0$. This is known as a ‘saddle-node’ bifurcation. The oscillatory instability, or Hopf bifurcation, occurs if the real part of an imaginary pair of eigenvalues passes through zero (provided all other eigenvalues are negative or have negative real parts). For a two-variable system, this occurs if the trace $\text{Tr } J = 0$, and is the origin of the result in [equation \(A3.14.18\)](#).

The autocatalator model is in many ways closely related to the FONI system, which has a single first-order exothermic reaction step obeying an Arrhenius temperature dependence and for which the role of the autocatalyst is taken by the temperature of the system. An extension of this is the Sal’nikov model which supports ‘thermokinetic’ oscillations in combustion-like systems [48]. This has the form:



-26-

The reactant P is again taken as a pool chemical, so the first step has a constant rate. The rate of the second step depends on the concentration of the intermediate A and on the temperature T and this step is taken as exothermic. (In the simplest case, k_0 is taken to be independent of T and the first step is thermoneutral.) Again, the steady state is found to be unstable over a range of parameter values, with oscillations being observed.

The approach to investigating spatial structure is similar—usually some simple solutions, such as a spatially uniform steady state exists and the condition for instability to spatial perturbations is determined in terms of eigenvalues of an extended Jacobian matrix. For conditions marginally beyond a bifurcation point (whether the instability is temporal or spatial), amplitude equations, such as the complex Ginsburg–Landau equation, are exploited [49]. For conditions far from bifurcation points, however, recourse to numerical integration is generally required. Frequently these will involve reaction–diffusion (and perhaps advection) equations, although representations of such systems in terms of cellular automata or gas-lattice models can be advantageous [50].

REFERENCES

- [1] Nicolis G and Prigogine I 1977 *Self-organization in Nonequilibrium Systems* (New York: Wiley)
- [2] Griffiths J F 1986 The fundamentals of spontaneous ignition of gaseous hydrocarbons and related organic compounds *Adv. Chem. Phys.* **64** 203–303
- [3] Epstein I R and Pojman J A 1999 Overview: nonlinear dynamics related to polymeric systems *Chaos* **9** 255–9
- [4] Gray P and Lee P R 1967 Thermal explosion theory *Combust. Oxid. Rev.* **2** 1–183
- [5] Scott S K and Showalter K 1992 Simple and complex propagating reaction-diffusion fronts *J. Phys. Chem.* **96** 8702–11
- [6] Pojman J A, Curtis G and Ilyashenko V M 1996 Frontal polymerisation in solution *J. Am. Chem. Soc.* **118** 3783–4
- [7] Gray P and Scott S K 1983 Autocatalytic reactions in the isothermal continuous, stirred-tank reactor: isolas and other forms of multistability *Chem. Eng. Sci.* **38** 29–43
- [8] Johnson B R, Scott S K and Tinsley M R 1998 A reduced model for complex oscillatory responses in the mesosphere *J. Chem. Soc. Faraday Trans.* **94** 2709–16
- [9] Epstein I R and Showalter K 1996 Nonlinear chemical dynamics: oscillation, patterns and chaos *J. Phys. Chem.* **100** 13 132–47
- [10] Winfree A T 1984 The prehistory of the Belousov–Zhabotinsky oscillator *J. Chem. Educ.* **61** 661–3
- [11] Zhabotinsky A M 1991 A history of chemical oscillations and waves *Chaos* **1** 379–86
- [12] Tyson J J 1976 *The Belousov–Zhabotinsky Reaction (Lecture Notes in Biomathematics vol 10)* (Berlin: Springer)
- [13] De Kepper P, Boissonade J and Epstein I R 1990 Chlorite–iodide reaction: a versatile system for the study of nonlinear dynamical behaviour *J. Phys. Chem.* **94** 6525–36
- [14] Field R J, Koros E and Noyes R M 1972 Oscillations in chemical systems, part 2: thorough analysis of temporal oscillations in the bromate–cerium–malonic acid system *J. Am. Chem. Soc.* **94** 8649–64

- [15] Field R J and Noyes R M 1974 Oscillations in chemical systems, part 4: limit cycle behavior in a model of a real chemical reaction *J. Chem. Phys.* **60** 1877–84
- [16] Tyson J J 1979 Oscillations, bistability and echo waves in models of the Belousov–Zhabotinskii reaction *Ann. New York Acad. Sci.* **316** 279–95
- [17] Zhabotinsky A M, Buchholtz F, Kiyatin A B and Epstein I R 1993 Oscillations and waves in metal-ion catalysed bromate oscillating reaction in highly oxidised states *J. Phys. Chem.* **97** 7578–84
- [18] Wang J, Sorensen P G and Hynne F 1994 Transient period doublings, torus oscillations and chaos in a closed chemical system *J. Phys. Chem.* **98** 725–7
- [19] Schmitz R A, Graziani K R and Hudson J L 1977 Experimental evidence of chaotic states in Belousov–Zhabotinskii reaction *J. Chem. Phys.* **67** 4071–5
- [20] Degn H, Olsen L F and Perram J W 1979 Bistability, oscillations and chaos in an enzyme reaction *Ann. New York Acad. Sci.* **316** 623–37

- [21] Swinney H L, Argoul F, Arneodo A, Richetti P and Roux J-C 1987 Chemical chaos: from hints to confirmation *Acc. Chem. Res.* **20** 436–42
- [22] Gyorgyi L and Field R J 1992 A three-variable model of deterministic chaos in the Belousov–Zhabotinsky reaction *Nature* **355** 808–10
- [23] Lengyel I, Rabai G and Epstein I R Experimental and modelling study of oscillations in the chlorine dioxide–iodine–malonic acid reaction *J. Am. Chem. Soc.* **112** 9104–10
- [24] Lengyel I and Epstein I R 1992 A chemical approach to designing Turing patterns in reaction–diffusion systems *Proc. Natl Acad. Sci.* **89** 3977–9
- [25] Epstein I R, Kustin K, De Kepper P and Orban M 1983 Oscillatory chemical reactions 1983 *Sci. Am.* **248** 96–108
- [26] Imbihl R and Ertl G 1995 Oscillatory kinetics in heterogeneous catalysis *Chem. Rev.* **95** 697–733
- [27] Johnson B R and Scott S K 1990 Period doubling and chaos during the oscillatory ignition of the CO + O₂ reaction *J. Chem. Soc. Faraday Trans.* **86** 3701–5
- [28] Johnson B R and Scott S K 1997 Complex and non-periodic oscillations in the H₂ + O₂ reaction *J. Chem. Soc. Faraday Trans.* **93** 2997–3004
- [29] Ditto W L and Showalter K (eds) 1997 Control and synchronization of chaos: focus issue *Chaos* **7** 509–687
- [30] Zaikin A N and Zhabotinsky A M 1970 Concentration wave propagation in two-dimensional liquid-phase self-oscillating system *Nature* **225** 535–7
- [31] Winfree A T 1972 Spiral waves of chemical activity *Science* **175** 634–6
- [32] Winfree A T 1998 Evolving perspectives during 12 years of electrical turbulence *Chaos* **8** 1–19
- [33] Pearlman H 1997 Target and spiral wave patterns in premixed gas combustion *J. Chem. Soc. Faraday Trans.* **93** 2487–90
- [34] Larter R, Speelman B and Worth R M 1999 A coupled ordinary differential equation lattice model for the simulation of epileptic seizures *Chaos* **9** 795–804

- [35] Nettesheim S, von Oertzen A, Rotermund H H and Ertl G 1993 Reaction diffusion patterns in the catalytic CO-oxidation on Pt(110): front propagation and spiral waves *J. Chem. Phys.* **98** 9977–85
- [36] Toth A and Showalter K 1995 Logic gates in excitable media *J. Chem. Phys.* **103** 2058–66
- [37] Turing A M 1952 The chemical basis of morphogenesis *Phil. Trans. R. Soc. B* **641** 37–72
- [38] Castets V, Dulos E, Boissonade J and De Kepper P 1990 Experimental evidence of a sustained standing Turing-type nonequilibrium structure *Phys. Rev. Lett.* **64** 2953–6
- [39] Ouyang Q and Swinney H L 1991 Transition to chemical turbulence *Chaos* **1** 411–20
- [40] Smithells A and Ingle H 1892 The structure and chemistry of flames *J. Chem. Soc. Trans.* **61** 204–16
- [41] Sivashinsky G I 1983 Instabilities, pattern formation and turbulence in flames *Ann. Rev. Fluid Mech.* **15** 179–99
- [42] Horvath D and Showalter K 1995 Instabilities in propagating reaction–diffusion fronts of the iodate–arsenous acid reaction *J. Chem. Phys.* **102** 2471–8
- [43] Toth A, Lagzi I and Horvath D 1996 Pattern formation in reaction–diffusion systems: cellular acidity fronts *J. Phys. Chem.* **100** 14 837–9

- [44] Rovinsky A B and Menzinger M 1993 Self-organization induced by the differential flow of activator and inhibitor *Phys. Rev. Lett.* **70** 778–81
- [45] Ouyang Q, Boissonade J, Roux J C and De Kepper P 1989 Sustained reaction–diffusion structures in an open reactor 1989 *Phys. Lett. A* **134** 282–6
- [46] Lee K-J, McCormick W D, Pearson J E and Swinney H L 1994 Experimental observation of self-replicating spots in a reaction–diffusion system *Nature* **369** 215–8
- [47] Prigogine I and Lefever R 1968 Symmetry breaking instabilities in dissipative systems *J. Chem. Phys.* **48** 1695–700
- [48] Gray P, Kay S R and Scott S K Oscillations of simple exothermic reactions in closed systems *Proc. R. Soc. Lond. A* **416** 321–41
- [49] Borckmans P, Dewel G, De Wit A and Walgraef D Turing bifurcations and pattern selection *Chemical Waves and Patterns* eds R Kapral and K Showalter (Dordrecht: Kluwer) ch 10, pp 323–63
- [50] Kapral R and Wu X-G Internal noise, oscillations, chaos and chemical waves *Chemical Waves and Patterns* eds R Kapral and K Showalter (Dordrecht: Kluwer) ch 18, pp 609–34
-

FURTHER READING

Scott S K 1994 *Oscillations, Waves and Chaos in Chemical Kinetics* (Oxford: Oxford University Press)

A short, final-year undergraduate level introduction to the subject.

Epstein I R and Pojman J A 1998 *An Introduction to Nonlinear Chemical Dynamics: Oscillations, Waves, Patterns and Chaos* (Oxford: Oxford University Press)

-29-

Introductory text at undergraduate/postgraduate level.

Gray P and Scott S K 1994 *Chemical Oscillations and Instabilities* (Oxford: Oxford University Press)

Graduate-level introduction mainly to theoretical modelling of nonlinear reactions

Scott S K 1993 *Chemical Chaos* (Oxford: Oxford University Press)

Graduate-level text giving detailed summary of status of chaotic behaviour in chemical systems to 1990.

Field R J and Burger M (eds) 1984 *Oscillations and Travelling Waves in Chemical Systems* (New York: Wiley)

Multi-author survey of nonlinear kinetics field to 1984, still a valuable introduction to researchers in this area.

Kapral R and Showalter K (eds) 1995 *Chemical Waves and Patterns* (Dordrecht: Kluwer)

Multi-author volume surveying chemical wave and pattern formation, an up-to-date introduction for those entering the field.

B1.1 Electronic spectroscopy

S J Strickler

B1.1.1 INTRODUCTION

Optical spectroscopy is the study of the absorption and emission of light by atoms, molecules, or larger assemblies. Electronic spectroscopy is the branch of the field in which the change produced by the absorption or emission is a rearrangement of the electrons in the system. These changes are interpreted in terms of the quantum theory of electronic structure. To a first approximation, the rearrangements usually correspond to an electron being transferred from one orbital to another, and a transition will be described in terms of those orbitals. The wavelengths or frequencies of transitions help identify atoms and molecules and give information about their energy levels and hence their electronic structure and bonding. Intensities of absorption or emission give information about the nature of the electronic states and help to determine concentrations of species. In the case of molecules, along with the rearrangements of electrons there are usually changes in nuclear motions, and the vibrational and rotational structure of electronic bands give valuable insights into molecular structure and properties. For all these reasons, electronic spectroscopy is one of the most useful tools in chemistry and physics.

Most electronic transitions of interest fall into the visible and near-ultraviolet regions of the spectrum. This range of photon energies commonly corresponds to electrons being moved among valence orbitals. These orbitals are important to an understanding of bonding and structure, so are of particular interest in physical chemistry and chemical physics. For this reason, most of this chapter will concentrate on visible and near-UV spectroscopy, roughly the region between 200 and 700 nm, but there are no definite boundaries to the wavelengths of interest. Some of the valence orbitals will be so close in energy as to give spectra in the near-infrared region. Conversely, some valence transitions will be at high enough energy to lie in the vacuum ultraviolet, below about 200 nm, where air absorbs strongly and instrumentation must be evacuated to allow light to pass. In this region are also transitions of electrons to states of higher principal quantum number, known as Rydberg states. At still higher energies, in the x-ray region, are transitions of inner-shell electrons, and their spectroscopy has become an extremely useful tool, especially for studying solids and their surfaces. However, these other regions will not be covered in detail here.

Section B1.1.2 provides a brief summary of experimental methods and instrumentation, including definitions of some of the standard measured spectroscopic quantities. Section B1.1.3 reviews some of the theory of spectroscopic transitions, especially the relationships between transition moments calculated from wavefunctions and integrated absorption intensities or radiative rate constants. Because units can be so confusing, numerical factors with their units are included in some of the equations to make them easier to use. Vibrational effects, the Franck–Condon principle and selection rules are also discussed briefly. In the final section, B1.1.4, a few applications are mentioned to particular aspects of electronic spectroscopy.

B1.1.2 EXPERIMENTAL METHODS

B1.1.2.1 STANDARD INSTRUMENTATION

There are two fundamental types of spectroscopic studies: absorption and emission. In absorption spectroscopy an atom or molecule in a low-lying electronic state, usually the ground state, absorbs a photon to go to a higher state. In emission spectroscopy the atom or molecule is produced in a higher electronic state by some excitation process, and emits a photon in going to a lower state. In this section we will consider the traditional instrumentation for studying the resulting spectra. They define the quantities measured and set the standard for experimental data to be considered.

(A) EMISSION SPECTROSCOPY

Historically, emission spectroscopy was the first technique to be extensively developed. An electrical discharge will break up most substances into atoms and small molecules and their ions. It also excites these species into nearly all possible stable states, and the higher states emit light by undergoing transitions to lower electronic states. In typical classical instruments light from the sample enters through a slit, is collimated by a lens or mirror, is dispersed by a prism or grating so that different colours or wavelengths are travelling in different directions and is focused on a detector. Gratings are generally favoured over prisms. A concave grating may do the collimation, dispersion and focusing all in one step. The angle at which the light is reflected by a plane grating is simply related to the spacing, d , of the grooves ruled on the grating and the wavelength, λ , of the light:

$$\pm n\lambda = d(\sin \alpha + \sin \beta)$$

where α and β are the angles of incidence and reflection measured from the normal, and n is the order of the reflection, i.e. the number of cycles by which wave fronts from successive grooves differ for constructive interference. (The angles may be positive or negative depending on the experimental arrangement.) The first determinations of the absolute wavelengths of light and most of our knowledge of energies of excited states of atoms and molecules came from measuring these angles. In everyday use now the wavelength scale of an instrument can be calibrated using the wavelengths of known atomic lines. Grating instruments do have the disadvantage in comparison with prism spectrometers that different orders of light may overlap and need to be sorted out.

The earliest detector was the human eye observing the different colours. More versatile is a photographic plate, where each wavelength of light emitted shows up as a dark line (an image of the entrance slit) on the plate. It has the advantage that many wavelengths are measured simultaneously. Quantitative measurements are easier with a photomultiplier tube placed behind an exit slit. The spectrum is obtained as the different wavelengths are scanned across the slit by rotating the grating. The disadvantage is that measurements are made one wavelength at a time. With the advent of solid-state electronics, array detectors have become available that will measure many wavelengths at a time much like a photographic plate, but which can be read out quickly and quantitatively into a computer or other data system.

The design and use of spectrographs or spectrometers involves a compromise between resolution—how close in wavelength two lines can be and still be seen as separate—and sensitivity—how weak a light can be observed or how

long it takes to make a measurement. Books have been written about the design of such instruments [1], and the subject cannot be pursued in this work.

Larger molecules generally cannot be studied in quite the same way, as an electric discharge merely breaks them up into smaller molecules or atoms. In such a case excited states are usually produced by optical excitation using light of the same or higher energy. Many modern fluorimeters are made with two

monochromators, one to select an excitation wavelength from the spectrum of a suitable lamp, and the other to observe the emission as discussed above. Most studies of large molecules are done on solutions because vapour pressures are too low to allow gaseous spectra.

The fundamental measurements made in emission spectroscopy are the wavelengths or frequencies and the intensities of emission lines or bands. The problem with intensity measurements is that the efficiency of a dispersing and detecting system varies with wavelength. Relative intensities at a single wavelength are usually quite easily measured and can be used as a measure of concentration or excitation efficiency. Relative intensities of lines at nearly the same wavelength, say different rotational lines in a given band, can usually be obtained by assuming that the efficiency is the same for all. But the absolute intensity of a band or relative intensities of well separated bands require a calibration of the sensitivity of the instrument as a function of wavelength. In favourable cases this may be done by recording the spectrum of a standard lamp that has in turn been calibrated to give a known spectrum at a defined lamp current. For critical cases it may be necessary to distinguish between intensities measured in energy flow per unit time or in photons per unit time.

(B) ABSORPTION SPECTROSCOPY

Absorption spectroscopy is a common and well developed technique for studying electronic transitions between the ground state and excited states of atoms or molecules. A beam of light passes through a sample, and the amount of light that is absorbed during the passage is measured as a function of the wavelength or frequency of the light. The absorption is measured by comparing the intensity, I , of light leaving the sample with the intensity, I_0 , entering the sample. The transmittance, T , is defined as the ratio

$$T = I/I_0.$$

It is often quoted as a percentage. In measuring the spectra of gases or solutions contained in cells, I_0 is usually taken to be the light intensity passing through an empty cell or a cell of pure solvent. This corrects well for reflection at the surfaces, absorption by the solvent or light scattering, which are not usually the quantities of interest.

It is usually convenient to work with the decadic absorbance, A , defined by

$$A = \log(I_0/I) = -\log T.$$

The unmodified term absorbance usually means this quantity, though some authors use the Napierian absorbance $B = -\ln T$. The absorbance is so useful because it normally increases linearly with path length, l , through the sample and with the concentration, c , of the absorbing species within the sample. The relationship is usually called Beer's law:

-4-

$$A = \epsilon cl.$$

The quantity ϵ is called the absorption coefficient or extinction coefficient, more completely the molar decadic absorption coefficient; it is a characteristic of the substance and the wavelength and to a lesser extent the solvent and temperature. It is common to take path length in centimetres and concentration in moles per litre, so ϵ has units of $\text{l mol}^{-1} \text{cm}^{-1}$. The electronic absorption spectrum of a compound is usually shown as a plot of ϵ versus wavelength or frequency.

Another useful quantity related to extinction coefficient is the cross section, σ , defined for a single atom or molecule. It may be thought of as the effective area blocking the beam at a given wavelength, and the value

may be compared with the size of the molecule. The relationship is

$$\sigma = (\ln 10)\epsilon/N_A$$

where N_A is Avogadro's number. If ϵ is in $\text{l mol}^{-1} \text{cm}^{-1}$ and σ is desired in cm^2 the relationship may be written

$$\sigma = (3.8235 \times 10^{-21} \text{ cm}^3 \text{ mol l}^{-1})\epsilon.$$

The standard instrument for measuring an absorption spectrum is a double-beam spectrophotometer. A typical instrument uses a lamp with a continuous spectrum to supply the light, usually a tungsten lamp for the visible, near-infrared, and near-ultraviolet regions and a discharge lamp filled with hydrogen or deuterium for farther in the ultraviolet. Light from the source passes through a monochromator to give a narrow band of wavelengths, and is then split into two beams. One beam passes through a cell containing the sample, the other through an identical reference cell filled with solvent. These beams define I and I_0 , respectively. The beams are monitored by a detection system, usually a photomultiplier tube. An electronic circuit measures the ratio of the two intensities and displays the transmittance or absorbance. The wavelength is varied by scanning the monochromator, and the spectrum may be plotted on a chart recorder.

A different design of instrument called a diode array spectrometer has become popular in recent years. In this instrument the light from the lamp passes through the sample, then into a spectrometer to be dispersed, and then is focused onto an array of solid-state detectors arranged so that each detector element measures intensity in a narrow band of wavelengths—say one detector for each nanometre of the visible and ultraviolet regions. The output is digitized and the spectrum displayed on a screen, and it can be read out in digital form and processed with a computer. The complete spectrum can be recorded in a few seconds. This is not formally a double-beam instrument, but because a spectrum is taken so quickly and handled so easily, one can record the spectrum of a reference cell and the sample cell and then compare them in the computer, so it serves the same purpose. The available instruments do not give quite the resolution or versatility of the standard spectrophotometers, but they are far quicker and easier to use.

B1.1.2.2 SOME MODERN TECHNIQUES

The traditional instruments for measuring emission and absorption spectra described above set the standard for the types of information which can be obtained and used by spectroscopists. In the more recent past, several new

-5-

techniques have become available which have extended the range of spectroscopic measurements to higher resolution, lower concentrations of species, weaker transitions, shorter time scales, etc. Many studies in electronic spectroscopy as a branch of physical chemistry or chemical physics are now done using these new techniques. The purpose of this section is to discuss some of them.

(A) LASERS

The foremost of the modern techniques is the use of lasers as spectroscopic tools. Lasers are extremely versatile light sources. They can be designed with many useful properties (not all in the same instrument) such as high intensity, narrow frequency bandwidth with high-frequency stability, tunability over reasonable frequency ranges, low-divergence beams which can be focused into very small spots, or pulsed beams with

very short time durations. There are nearly as many different experimental arrangements as there are experimenters, and only a few examples will be mentioned here.

While a laser beam can be used for traditional absorption spectroscopy by measuring I and I_0 , the strength of laser spectroscopy lies in more specialized experiments which often do not lend themselves to such measurements. Other techniques are commonly used to detect the absorption of light from the laser beam. A common one is to observe fluorescence excited by the laser. The total fluorescence produced is normally proportional to the amount of light absorbed. It can be used as a measurement of concentration to detect species present in extremely small amounts. Or a measurement of the fluorescence intensity as the laser frequency is scanned can give an absorption spectrum. This may allow much higher resolution than is easily obtained with a traditional absorption spectrometer. In other experiments the fluorescence may be dispersed and its spectrum determined with a traditional spectrometer. In suitable cases this could be the emission from a single electronic–vibrational–rotational level of a molecule and the experimenter can study how the spectrum varies with level.

Other methods may also be useful for detecting the absorption of laser radiation. For example, the heat generated when radiation is absorbed can be detected in several ways. One way observes the defocusing of the laser beam when the medium is heated and its refractive index changes. Another way, called photoacoustic spectroscopy, detects sound waves or pressure pulses when light is absorbed from a pulsed laser. Still another method useful with high-intensity pulsed lasers is to measure light absorption by the excited states produced. This is often useful for studying the kinetics of the excited species as they decay or undergo reactions.

Another example of a technique for detecting absorption of laser radiation in gaseous samples is to use multiphoton ionization with intense pulses of light. Once a molecule has been electronically excited, the excited state may absorb one or more additional photons until it is ionized. The electrons can be measured as a current generated across the cell, or can be counted individually by an electron multiplier; this can be a very sensitive technique for detecting a small number of molecules excited.

(B) EXCITED-STATE LIFETIMES

Measurements of the decay rates of excited states are important, both for the fundamental spectroscopic information they can give, and for studies of other processes such as energy transfer or photochemistry. The techniques used vary greatly depending on the time scale of the processes being studied. For rather long time scales, say of the order of a millisecond or longer, it is rather simple to excite the molecules optically, cut off the exciting light, and watch the decay of emission or some other measurement of excited-state concentration.

-6-

For fluorescent compounds and for times in the range of a tenth of a nanosecond to a hundred microseconds, two very successful techniques have been used. One is the phase-shift technique. In this method the fluorescence is excited by light whose intensity is modulated sinusoidally at a frequency f , chosen so its period is not too different from the expected lifetime. The fluorescent light is then also modulated at the same frequency but with a time delay. If the fluorescence decays exponentially, its phase is shifted by an angle $\Delta\phi$ which is related to the mean life, τ , of the excited state. The relationship is

$$\tan \Delta\phi = 2\pi f\tau.$$

The phase shift is measured by comparing the phase of the fluorescence with the phase of light scattered by a cloudy but non-fluorescent solution.

The other common way of measuring nanosecond lifetimes is the time-correlated single-photon counting

technique [2]. In this method the sample is excited by a weak, rapidly repeating pulsed light source, which could be a flashlamp or a mode-locked laser with its intensity reduced. The fluorescence is monitored by a photomultiplier tube set up so that current pulses from individual photons can be counted. It is usually arranged so that at most one fluorescence photon is counted for each flash of the excitation source. A time-to-amplitude converter and a multichannel analyser (equipment developed for nuclear physics) are used to determine, for each photon, the time between the lamp flash and the photon pulse. A decay curve is built up by measuring thousands of photons and sorting them by time delay. The statistics of such counting experiments are well understood and very accurate lifetimes and their uncertainties can be determined by fitting the resulting decay curves.

One advantage of the photon counting technique over the phase-shift method is that any non-exponential decay is readily seen and studied. It is possible to detect non-exponential decay in the phase-shift method too by making measurements as a function of the modulation frequency, but it is more cumbersome.

At still shorter time scales other techniques can be used to determine excited-state lifetimes, but perhaps not as precisely. Streak cameras can be used to measure faster changes in light intensity. Probably the most useful techniques are pump-probe methods where one intense laser pulse is used to excite a sample and a weaker pulse, delayed by a known amount of time, is used to probe changes in absorption or other properties caused by the excitation. At short time scales the delay is readily adjusted by varying the path length travelled by the beams, letting the speed of light set the delay.

(C) PHOTOELECTRON SPECTROSCOPY

Only brief mention will be made here of photoelectron spectroscopy. This technique makes use of a beam of light whose energy is greater than the ionization energy of the species being studied. Transitions then occur in which one of the electrons of the molecule is ejected. Rather than an optical measurement, the kinetic energy of the ejected electron is determined. Some of the technology is described in [section B1.6](#) on electron energy-loss spectroscopy. The ionization energy of the molecule is determined from the difference between the photon energy and the kinetic energy of the ejected electron.

A useful light source is the helium resonance lamp which produces light of wavelength 58.4 nm or a photon energy of 21.2 eV, enough to ionize any neutral molecule. Often several peaks can be observed in the photoelectron spectrum

-7-

corresponding to the removal of electrons from different orbitals. The energies of the peaks give approximations to the orbital energies in the molecule. They are useful for comparison with theoretical calculations.

An interesting variation on the method is the use of a laser to photodetach an electron from a negative ion produced in a beam of ions. Since it is much easier to remove an electron from a negative ion than from a neutral molecule, this can be done with a visible or near-ultraviolet laser. The difference between the photon energy and the electron energy, in this case, gives the electron affinity of the neutral molecule remaining after the photodetachment, and may give useful energy levels of molecules not easily studied by traditional spectroscopy [3].

(D) OTHER TECHNIQUES

Some other extremely useful spectroscopic techniques will only be mentioned here. Probably the most important one is spectroscopy in free jet expansions. Small molecules have often been studied by gas-phase spectroscopy where sharp rotational and vibrational structure gives detailed information about molecular

states and geometries. The traditional techniques will often not work for large molecules because they must be heated to high temperatures to vaporize them and then the spectra become so broad and congested that detailed analysis is difficult or impossible. In jet spectroscopy the gaseous molecules are mixed with an inert gas and allowed to expand into a vacuum. The nearly adiabatic expansion may cool them rapidly to temperatures of a few degrees Kelvin while leaving them in the gas phase. The drastic simplification of the spectrum often allows much more information to be extracted from the spectrum.

Fourier-transform instruments can also be used for visible and ultraviolet spectroscopy. In this technique, instead of dispersing the light with a grating or prism, a wide region of the spectrum is detected simultaneously by splitting the beam into two components, one reflected from a stationary mirror and one from a movable mirror, and then recombining the two beams before they enter the detector. The detected intensity is measured as a function of the position of the movable mirror. Because of interference between the two beams, the resulting function is the Fourier transform of the normal spectrum as a function of wavelength. This offers some advantages in sensitivity and perhaps resolution because the whole spectrum is measured at once rather than one wavelength at a time. The technique is not too common in electronic spectroscopy, but is very widely used for the infrared. It is described more fully in the chapter on vibrational spectroscopy, [section B1.2](#). While the light sources and detectors are different for the visible and ultraviolet region, the principles of operation are the same.

B1.1.3 THEORY

The theory of absorption or emission of light of modest intensity has traditionally been treated by time-dependent perturbation theory [4]. Most commonly, the theory treats the effect of the oscillating electric field of the light wave acting on the electron cloud of the atom or molecule. The instantaneous electric field is assumed to be uniform over the molecule but it oscillates in magnitude and direction with a frequency ν . The energy of a system of charges in a uniform electric field, \mathbf{E} , depends on its dipole moment according to

$$E = -\boldsymbol{\mu} \cdot \mathbf{E}$$

-8-

where the dipole moment is defined by

$$\boldsymbol{\mu} = \sum_i q_i \mathbf{r}_i$$

and the q_i and \mathbf{r}_i are the charges and positions of the particles, i.e. the electrons and nuclei. The result of the time-dependent perturbation theory is that the transition probability for a transition between one quantum state i and another state j is proportional to the absolute value squared of the matrix element of the electric dipole operator between the two states

$$\mu_{ij} = \int \Psi_i^* \boldsymbol{\mu} \Psi_j \, d\tau$$

(B1.1.1)

transition probability $\propto |\mu_{ij}|^2$.

The transition occurs with significant probability only if the frequency of the light is very close to the familiar resonance condition, namely $h\nu = \Delta E$, where h is Planck's constant and ΔE is the difference in energy of the

two states. However, transitions always occur over a range of frequencies because of various broadening effects; if nothing else, as required by the uncertainty principle, the states will not have precisely defined energies if they have finite lifetimes.

B1.1.3.1 ABSORPTION SPECTROSCOPY

(A) INTEGRATED ABSORPTION INTENSITY

The relationship between the theoretical quantity μ_{ij} and the experimental parameter ϵ of absorption spectroscopy involves, not the value of ϵ at any one wavelength, but its integral over the absorption band. The relationship is

$$\int \epsilon \, d\tilde{\nu} = \frac{2\pi^2 N_A \tilde{\nu}}{3hc\epsilon_0 \ln 10} |\mu_{ij}|^2 = (2.512 \times 10^{19} \text{ l mol}^{-1} \text{ cm}^{-3}) \frac{\tilde{\nu}}{e^2} |\mu_{ij}|^2. \quad (\text{B1.1.2})$$

We will quote a numerical constant in some of these equations to help with actual calculations. The units can be very confusing because it is conventional to use non-SI units for several quantities. The wavenumber value, $\tilde{\nu}$, is usually taken to be in cm^{-1} . The extinction coefficient is conveniently taken in units of $\text{l mol}^{-1} \text{ cm}^{-1}$. We have inserted the factor e^2 into the equation because values of μ_{ij} are usually calculated with the charges measured in units of the electron charge. For the sake of consistency, we have quoted the numerical factor appropriate for μ/e taken in centimetres, but the values are easily converted to use other length units such as Ångstroms or atomic units. The value of $\tilde{\nu}$ in the right-hand side of the equation is to be interpreted as a suitable average frequency for the transition. This causes no difficulty unless the band is very broad. Some of the difficulties with different definitions of intensity terms have been discussed by Hilborn [5].

-9-

(B) OSCILLATOR STRENGTH

A related measure of the intensity often used for electronic spectroscopy is the oscillator strength, f . This is a dimensionless ratio of the transition intensity to that expected for an electron bound by Hooke's law forces so as to be an isotropic harmonic oscillator. It can be related either to the experimental integrated intensity or to the theoretical transition moment integral:

$$(\text{B1.1.3})$$

or

$$(\text{B1.1.4})$$

The harmonically bound electron is, in a sense, an ideal absorber since its harmonic motion can maintain a perfect phase relationship with the oscillating electric field of the light wave. Strong electronic transitions have oscillator strengths of the order of unity, but this is not, as sometimes stated, an upper limit to f . For example, some polyacetylenes have bands with oscillator strengths as high as 5 [6]. There is a theorem, the Kuhn–Thomas sum rule, stating that the sum of the oscillator strengths of all electronic transitions must be equal to the number of electrons in an atom or molecule [7].

In the above discussion we have used the electric dipole operator μ . It is also sometimes possible to observe electronic transitions occurring due to interaction with the magnetic field of the light wave. These are called magnetic dipole transitions. They are expected to be weaker than electric dipole transitions by several orders of magnitude. If account is taken of a variation of the field of the light wave over the size of the molecule it is possible to treat quadrupole or even higher multipole transitions. These are expected to be even weaker than typical magnetic dipole transitions. We will concentrate on the more commonly observed electric dipole

transitions.

[Equation \(B1.1.1\)](#) for the transition moment integral is rather simply interpreted in the case of an atom. The wavefunctions are simply functions of the electron positions relative to the nucleus, and the integration is over the electronic coordinates. The situation for molecules is more complicated and deserves discussion in some detail.

(C) TRANSITION MOMENTS FOR MOLECULES

Electronic spectra are almost always treated within the framework of the Born–Oppenheimer approximation [8] which states that the total wavefunction of a molecule can be expressed as a product of electronic, vibrational, and rotational wavefunctions (plus, of course, the translation of the centre of mass which can always be treated separately from the internal coordinates). The physical reason for the separation is that the nuclei are much heavier than the electrons and move much more slowly, so the electron cloud normally follows the instantaneous position of the nuclei quite well. The integral of [equation \(B1.1.1\)](#) is over all internal coordinates, both electronic and nuclear. Integration over the rotational wavefunctions gives rotational selection rules which determine the fine structure and band shapes of electronic transitions in gaseous molecules. Rotational selection rules will be discussed below. For molecules in condensed phases the rotational motion is suppressed and replaced by oscillatory and diffusional motions.

-10-

In this section we concentrate on the electronic and vibrational parts of the wavefunctions. It is convenient to treat the nuclear configuration in terms of normal coordinates describing the displacements from the equilibrium position. We call these nuclear normal coordinates Q_j and use the symbol Q without a subscript to designate the whole set. Similarly, the symbol x_i designates the coordinates of the i th electron and x the whole set of electronic coordinates. We also use subscripts l and u to designate the lower and upper electronic states of a transition, and subscripts a and b to number the vibrational states in the respective electronic states. The total wavefunction Ψ can be written

$$\Psi_{la}(x, Q) = \psi_l(x, Q)\phi_a(Q) \quad \Psi_{ub}(x, Q) = \psi_u(x, Q)\phi_b(Q).$$

Here each $\phi(Q)$ is a vibrational wavefunction, a function of the nuclear coordinates Q , in first approximation usually a product of harmonic oscillator wavefunctions for the various normal coordinates. Each $\psi(x, Q)$ is the electronic wavefunction describing how the electrons are distributed in the molecule. However, it has the nuclear coordinates within it as parameters because the electrons are always distributed around the nuclei and follow those nuclei whatever their position during a vibration. The integration of [equation \(B1.1.1\)](#) can be carried out in two steps—first an integration over the electronic coordinates x , and then integration over the nuclear coordinates Q . We define an electronic transition moment integral which is a function of nuclear position:

$$\mu_{lu}(Q) = \int \psi_l^*(x, Q)\mu\psi_u(x, Q) dx. \quad (\text{B1.1.5})$$

We then integrate this over the vibrational wavefunctions and coordinates:

$$\mu_{lu,ab} = \int \Psi_{la}^*\mu\Psi_{ub} d\tau = \int \phi_{la}^*(Q)\mu_{lu}(Q)\phi_{ub}(Q) dQ. \quad (\text{B1.1.6})$$

This last transition moment integral, if plugged into [equation \(B1.1.2\)](#), will give the integrated intensity of a vibronic band, i.e. of a transition starting from vibrational state a of electronic state l and ending on vibrational level b of electronic state u .

(D) THE FRANCK—CONDON PRINCIPLE

The electronic transition moment of [equation \(B1.1.5\)](#) is related to the intensity that the transition would have if the nuclei were fixed in configuration Q , but its value may vary with that configuration. It is often useful to expand $\mu_{lu}(Q)$ as a power series in the normal coordinates, Q_i :

$$\mu_{lu}(Q) = \mu_{lu}(0) + \sum_i \left(\frac{\partial \mu_{lu}}{\partial Q_i} \right)_0 Q_i + \dots \quad (\text{B1.1.7})$$

-11-

Here $\mu_{lu}(0)$ is the value at the equilibrium position of the initial electronic state.

In many cases the variation is not very strong for reasonable displacements from equilibrium, and it is sufficient to use only the zero-order term in the expansion. If this is inserted into [equation \(B1.1.6\)](#) we get

$$\mu_{la,ub} = \mu_{lu}(0) \int \phi_{la}^*(Q) \phi_{ub}(Q) dQ$$

and using this in [equation \(B1.1.2\)](#) for the integrated intensity of a vibronic band we get the relationship

$$\int \varepsilon d\tilde{\nu} = \frac{2\pi^2 N_A \tilde{\nu}_{la \rightarrow ub}}{3hc\varepsilon_0 \ln 10} |\mu_{lu}(0)|^2 \left| \int \phi_{la}^*(Q) \phi_{ub}(Q) dQ \right|^2. \quad (\text{B1.1.8})$$

The last factor, the square of the overlap integral between the initial and final vibrational wavefunctions, is called the Franck—Condon factor for this transition.

The Franck—Condon principle says that the intensities of the various vibrational bands of an electronic transition are proportional to these Franck—Condon factors. (Of course, the frequency factor must be included for accurate treatments.) The idea was first derived qualitatively by Franck through the picture that the rearrangement of the light electrons in the electronic transition would occur quickly relative to the period of motion of the heavy nuclei, so the position and momentum of the nuclei would not change much during the transition [9]. The quantum mechanical picture was given shortly afterwards by Condon, more or less as outlined above [10].

The effects of the principle are most easily visualized for diatomic molecules for which the vibrational potential can be represented by a potential energy curve. A typical absorption starts from the lowest vibrational level of the ground state (actually a thermal distribution of low-lying levels). A useful qualitative statement of the Franck—Condon principle is that vertical transitions should be favoured. [Figure B1.1.1\(a\)](#) illustrates the case where the potential curve for the excited state lies nearly directly above that for the ground state. Then by far the largest overlap of excited state wavefunctions with the lowest level of the ground state will be for the $v = 0$ level, and we expect most intensity to be in the so-called 0–0 band, i.e. from $v = 0$ in the lower state to $v = 0$ in the upper state. A case in point is the transition of the O_2 molecule at about 750 nm in the near-infrared. (This is actually a magnetic dipole transition rather than electric dipole, so it is very weak,

but the vibrational effects are the same.) Both ground and excited state have a $(\pi^*)^2$ electron configuration and nearly the same equilibrium bond length, only 0.02 Å different. The spectrum shows most of the intensity in the 0–0 band with less than one tenth as much in the 1–0 band [11].

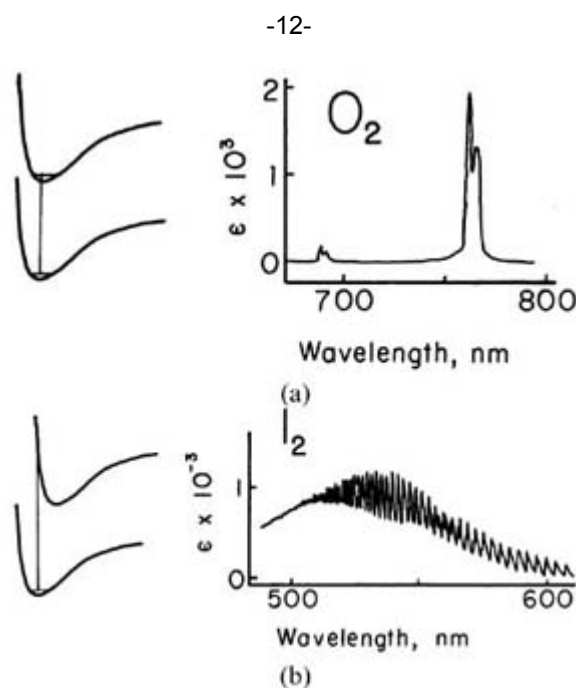


Figure B1.1.1. (a) Potential curves for two states with little or no difference in the equilibrium position of the upper and lower states. A transition of O_2 , with displacement only 0.02 Å, is shown as an example. Data taken from [11]. Most of the intensity is in the 0–0 vibrational band with a small intensity in the 1–0 band. (b) Potential curves for two states with a large difference in the equilibrium position of the two states. A transition in I_2 , with a displacement of 0.36 Å, is shown as an example. Many vibrational peaks are observed.

Figure B1.1.1(b) shows a contrasting case, where the potential curve for the excited state is displaced considerably relative to the ground-state curve. Then a vertical transition would go to a part of the excited-state curve well displaced from the bottom, and the maximum overlap and greatest intensity should occur for high-lying levels. There results a long progression of bands to various vibrational levels. The spectrum of I_2 is shown as an illustration; here the displacement between the two minima is about 0.36 Å. Many vibronic transitions are seen. One can observe the excited-state levels getting closer together and converging as the dissociation limit is approached, and part of the absorption goes to continuum states above the dissociation energy. (The long-wavelength part of the spectrum is complicated by transitions starting from thermally excited vibrational levels of the ground state.)

(E) BEYOND FRANCK—CONDON

There are cases where the variation of the electronic transition moment with nuclear configuration cannot be neglected. Then it is necessary to work with equation (B1.1.6) keeping the dependence of μ_{lu} on Q and integrating it over the vibrational wavefunctions. In most such cases it is adequate to use only the terms up to first-order in equation (B1.1.7). This results in ‘modified Franck–Condon factors’ for the vibrational intensities [12].

(F) TOTAL INTENSITY OF AN ELECTRONIC TRANSITION

Equation (B1.1.8) gives the intensity of one vibronic band in an absorption spectrum. It is also of interest to consider

the total integrated intensity of a whole electronic transition, i.e. the sum of all the vibronic bands corresponding to the one electronic change. In the most common absorption spectroscopy experiment we can assume that all transitions originate in the lowest vibrational level of the ground electronic state, which we can designate as level 10. The transitions can go to various levels ub of the upper electronic state. The total integrated intensity is then obtained by summing over the index b which numbers the excited state vibrational levels.

$$\int \varepsilon d\tilde{\nu} = \frac{2\pi^2 N_A}{3hc\varepsilon_0 \ln 10} \sum_b \tilde{\nu}_{10 \rightarrow ub} \left| \int \phi_{10}^*(Q) \mu_{lu}(Q) \phi_{ub}(Q) dQ \right|^2. \quad (\text{B1.1.9})$$

This equation can be simplified if the frequency term $\tilde{\nu}_{10 \rightarrow ub}$ is removed from the summation. One way to do this is to incorporate it into the integral on the left-hand side by writing $\int \varepsilon d \ln \tilde{\nu}$. The alternative is to use an appropriate average $\langle \tilde{\nu} \rangle$ outside the sum, choosing the proper average by making the expressions equal. Often it is enough to pick an average by eye, but if high accuracy is important the value to use is given by

$$\langle \tilde{\nu} \rangle = \frac{\int \varepsilon d\tilde{\nu}}{\int \varepsilon d \ln \tilde{\nu}}.$$

With the frequency removed from the sum, (B1.1.9) has just a sum over vibrational integrals. Because all the vibrational wavefunctions for a given potential surface will form a complete set, it is possible to apply a sum rule to simplify the resulting expression:

$$\sum_b \left| \int \phi_{10}^*(Q) \mu_{lu}(Q) \phi_{ub}(Q) dQ \right|^2 = \int \phi_{10}^*(Q) |\mu_{lu}(Q)|^2 \phi_{10}(Q) dQ$$

i.e. the sum is just the mean value of $|\mu_{lu}(Q)|^2$ in the initial vibrational state. Then the total integrated intensity of the electronic band is given by

$$\int \varepsilon d\tilde{\nu} = \frac{2\pi^2 N_A \langle \tilde{\nu} \rangle}{3hc\varepsilon_0 \ln 10} \int \phi_{10}^*(Q) |\mu_{lu}(Q)|^2 \phi_{10}(Q) dQ. \quad (\text{B1.1.10})$$

If we can get by with using only the zero-order term of (B1.1.7), we can take μ_{lu} out of the integral and use the fact that ϕ_{10} is normalized. The last equation then simplifies further to

$$\int \varepsilon d\tilde{\nu} = \frac{2\pi^2 N_A \langle \tilde{\nu} \rangle}{3hc\varepsilon_0 \ln 10} |\mu_{lu}(0)|^2 = (2.512 \times 10^{19} \text{ l mol}^{-1} \text{ cm}^{-3}) \frac{\langle \tilde{\nu} \rangle}{e^2} |\mu_{lu}(0)|^2. \quad (\text{B1.1.11})$$

Equation (B1.1.10) and equation (B1.1.11) are the critical ones for comparing observed intensities of electronic transitions with theoretical calculations using the electronic wavefunctions. The transition moment integral μ_{lu}

calculated from electronic wavefunctions is related to the absorption intensity integrated over the whole electronic transition. It is found that simple forms of electronic wavefunctions often do not give very good intensities, and high-quality wavefunctions are required for close agreement with experiment.

B1.1.3.2 EMISSION SPECTROSCOPY

The interpretation of emission spectra is somewhat different but similar to that of absorption spectra. The intensity observed in a typical emission spectrum is a complicated function of the excitation conditions which determine the number of excited states produced, quenching processes which compete with emission, and the efficiency of the detection system. The quantities of theoretical interest which replace the integrated intensity of absorption spectroscopy are the rate constant for spontaneous emission and the related excited-state lifetime.

(A) EMISSION RATE CONSTANT

Einstein derived the relationship between spontaneous emission rate and the absorption intensity or stimulated emission rate in 1917 using a thermodynamic argument [13]. Both absorption intensity and emission rate depend on the transition moment integral of equation (B1.1.1), so that gives us a way to relate them. The symbol A is often used for the rate constant for emission; it is sometimes called the Einstein A coefficient. For emission in the gas phase from a state i to a lower state j we can write

$$A_{i \rightarrow j} = \frac{16\pi^3}{3h\epsilon_0} \tilde{\nu}_{i \rightarrow j}^3 |\mu_{ij}|^2 = (7.235 \times 10^{10} \text{ cm s}^{-1}) \frac{\tilde{\nu}_{i \rightarrow j}^3}{e^2} |\mu_{ij}|^2. \quad (\text{B1.1.12})$$

(B) MOLECULAR EMISSION AND THE FRANCK—CONDON PRINCIPLE

For molecules we can use Born–Oppenheimer wavefunctions and talk about emission from one vibronic level to another. Equation (B1.1.5), equation (b1.1.6) and equation (b1.1.7) can be used just as they were for absorption. If we have an emission from vibronic state ub to the lower state la , the rate constant for emission would be given by

$$A_{ub \rightarrow la} = \frac{16\pi^3}{3h\epsilon_0} \tilde{\nu}_{ub \rightarrow la}^3 \left| \int \phi_{ub}^*(Q) \mu_{ul}(Q) \phi_{la}(Q) dQ \right|^2. \quad (\text{B1.1.13})$$

If we can use only the zero-order term in equation (B1.1.7) we can remove the transition moment from the integral and recover an equation involving a Franck–Condon factor:

$$A_{ub \rightarrow la} = \frac{16\pi^3}{3h\epsilon_0} |\mu_{ul}(0)|^2 \tilde{\nu}_{ub \rightarrow la}^3 \left| \int \phi_{ub}^*(Q) \phi_{la}(Q) dQ \right|^2. \quad (\text{B1.1.14})$$

Now the spectrum will show various transitions originating with state ub and ending on the various vibrational levels la of the lower electronic state. Equation (B1.1.14) (or B1.1.13) if we have to worry about variation of transition

moment) gives us a way of comparing the intensities of the bands. The intensities will be proportional to the $A_{ub \rightarrow \lambda, \alpha}$ provided that we measure the intensity in photons per unit time rather than the more conventional

units of energy per unit time. The first part of the expression in (B1.1.14) is the same for all the transitions. The part that varies between bands is the Franck–Condon factor multiplied by the cube of the frequency. Equation (B1.1.8) for absorption intensity also had a frequency factor, but the variation in frequency has more effect in emission spectroscopy because it appears to a higher power. Equation (B1.1.14) embodies the Franck–Condon principle for emission spectroscopy.

(C) EXCITED-STATE LIFETIME

We now discuss the lifetime of an excited electronic state of a molecule. To simplify the discussion we will consider a molecule in a high-pressure gas or in solution where vibrational relaxation occurs rapidly, we will assume that the molecule is in the lowest vibrational level of the upper electronic state, level u_0 , and we will further assume that we need only consider the zero-order term of equation (B1.1.7). A number of radiative transitions are possible, ending on the various vibrational levels l of the lower state, usually the ground state. The total rate constant for radiative decay, which we will call $A_{u_0 \rightarrow \lambda}$, is the sum of the rate constants, $A_{u_0 \rightarrow \lambda \alpha}$. By summing the terms in equation (B1.1.14) we can get an expression relating the radiative lifetime to the theoretical transition moment μ_{ul} . Further, by relating the transition moment to integrated absorption intensity we can get an expression for radiative rate constant involving only experimental quantities and not dependent on the quality of the electronic wavefunctions:

$$\begin{aligned} A_{u_0 \rightarrow l} &= \frac{16\pi^3 n^2}{3h\epsilon_0} \langle \tilde{\nu}_f^{-3} \rangle_{Av}^{-1} \frac{g_l}{g_u} |\mu_{ul}(0)|^2 \\ &= (7.235 \times 10^{10} \text{ cm s}^{-1}) n^2 \frac{g_l \langle \tilde{\nu}_f^{-3} \rangle_{Av}^{-1}}{g_u e^2} |\mu_{ul}(0)|^2 \end{aligned} \quad (\text{B1.1.15})$$

or

$$\begin{aligned} A_{u_0 \rightarrow l} &= \frac{8\pi c n^2 \ln 10}{N_A} \langle \tilde{\nu}_f^{-3} \rangle_{Av}^{-1} \frac{g_l}{g_u} \int \epsilon \, d \ln \tilde{\nu} \\ &= (2.881 \times 10^{-9} \text{ s}^{-1} \text{ l}^{-1} \text{ mol cm}^4) \langle \tilde{\nu}_f^{-3} \rangle_{Av}^{-1} n^2 \frac{g_l}{g_u} \int \epsilon \, d \ln \tilde{\nu}. \end{aligned} \quad (\text{B1.1.16})$$

These equations contain the peculiar average fluorescence frequency $\langle \tilde{\nu}_f^{-3} \rangle_{Av}^{-1}$, the reciprocal of the average value of $\tilde{\nu}^{-3}$ in the fluorescence spectrum. It arises because the fluorescence intensity measured in photons per unit time has a ν^3 dependence. For completeness we have added a term n^2 , the square of the refractive index, to be used for molecules in solution, and the term g_l / g_u , the ratio of the degeneracies of the lower and upper electronic states, to allow for degenerate cases [14]. It is also possible to correct for a variation of transition moment with nuclear configuration if that should be necessary [15].

$A_{u_0 \rightarrow \lambda}$ is the first-order rate constant for radiative decay by the molecule. It is the reciprocal of the intrinsic mean life of the excited state, τ_0 :

-16-

$$1/\tau_0 = A_{u_0 \rightarrow l}.$$

If there are no competing processes the experimental lifetime τ should equal τ_0 . Most commonly, other processes such as non-radiative decay to lower electronic states, quenching, photochemical reactions or

energy transfer may compete with fluorescence. They will reduce the actual lifetime. As long as all the processes are first-order in the concentration of excited molecules, the decay will remain exponential and the mean life τ will be reduced by a factor of the fluorescence quantum yield, Φ_f , the fraction of the excited molecules which emit:

$$\tau = \Phi_f \tau_0. \quad (\text{B1.1.17})$$

B1.1.3.3 SELECTION RULES

Transition intensities are determined by the wavefunctions of the initial and final states as described in the last sections. In many systems there are some pairs of states for which the transition moment integral vanishes while for other pairs it does not vanish. The term ‘selection rule’ refers to a summary of the conditions for non-vanishing transition moment integrals—hence observable transitions—or vanishing integrals so no observable transitions. We discuss some of these rules briefly in this section. Again, we concentrate on electric dipole transitions.

(A) ATOMS

The simplest case arises when the electronic motion can be considered in terms of just one electron: for example, in hydrogen or alkali metal atoms. That electron will have various values of orbital angular momentum described by a quantum number l . It also has a spin angular momentum described by a spin quantum number s of $\frac{1}{2}$, and a total angular momentum which is the vector sum of orbital and spin parts with quantum number j . In the presence of a magnetic field the component of the angular momentum in the field direction becomes important and is described by a quantum number m . The selection rules can be summarized as

$$\Delta l = \pm 1 \quad \Delta j = 0, \pm 1 \quad \Delta m = 0, \pm 1.$$

This means that one can see the electron undergo transitions from an s orbital to a p orbital, from a p orbital to s or d, from a d orbital to p or f, etc, but not s to s, p to p, s to d, or such. In terms of state designations, one can have transitions from $^2S_{1/2}$ to $^2P_{1/2}$ or to $^2P_{3/2}$, etc.

In more complex atoms there may be a strong coupling between the motion of different electrons. The states are usually described in terms of the total orbital angular momentum L and the total spin angular momentum S . These are coupled to each other by an interaction called spin-orbital coupling, which is quite weak in light atoms but gets rapidly stronger as the nuclear charge increases. The resultant angular momentum is given the symbol J . States are named using capital letters **S**, **P**, **D**, **F**, **G**, . . . to designate L values of 0, 1, 2, 3, 4, A left superscript gives the multiplicity ($2S + 1$), and a right subscript gives the value of J , for example 1S_0 , 3P_1 or $^2D_{5/2}$.

-17-

There is a strict selection rule for J :

$$\Delta J = 0, \pm 1 \quad \text{with the restriction that } J = 0 \text{ to } J = 0 \text{ is forbidden.}$$

There are approximate selection rules for L and S , namely

$$\Delta L = 0, \pm 1 \quad \text{and} \quad \Delta S = 0.$$

These hold quite well for light atoms but become less dependable with greater nuclear charge. The term ‘intercombination bands’ is used for spectra where the spin quantum number S changes: for example, singlet–triplet transitions. They are very weak in light atoms but quite easily observed in heavy ones.

(B) ELECTRONIC SELECTION RULES FOR MOLECULES

Atoms have complete spherical symmetry, and the angular momentum states can be considered as different symmetry classes of that spherical symmetry. The nuclear framework of a molecule has a much lower symmetry. Symmetry operations for the molecule are transformations such as rotations about an axis, reflection in a plane, or inversion through a point at the centre of the molecule, which leave the molecule in an equivalent configuration. Every molecule has one such operation, the identity operation, which just leaves the molecule alone. Many molecules have one or more additional operations. The set of operations for a molecule form a mathematical group, and the methods of group theory provide a way to classify electronic and vibrational states according to whatever symmetry does exist. That classification leads to selection rules for transitions between those states. A complete discussion of the methods is beyond the scope of this chapter, but we will consider a few illustrative examples. Additional details will also be found in [section A1.4](#) on molecular symmetry.

In the case of linear molecules there is still complete rotational symmetry about the internuclear axis. This leads to the conservation and quantization of the component of angular momentum in that direction. The quantum number for the component of orbital angular momentum along the axis (the analogue of L for an atom) is called Λ . States which have $\Lambda = 0, 1, 2, \dots$ are called $\Sigma, \Pi, \Delta, \dots$ (analogous to S, P, D, \dots of atoms). Σ states are non-degenerate while Π, Δ , and higher angular momentum states are always doubly degenerate because the angular momentum can be in either direction about the axis. Σ states need an additional symmetry designation. They are called Σ^+ or Σ^- according to whether the electronic wavefunction is symmetric or antisymmetric to the symmetry operation of a reflection in a plane containing the internuclear axis. If the molecule has a centre of symmetry like N_2 or CO_2 , there is an additional symmetry classification, g or u, depending on whether the wavefunction is symmetric or antisymmetric with respect to inversion through that centre. Symmetries of states are designated by symbols such as Π_g, Π_u, Σ_g^+ , etc. Finally, the electronic wavefunctions will have a spin multiplicity ($2S + 1$), referred to as singlet, doublet, triplet, etc. The conventional nomenclature for electronic states is as follows. The state is designated by its symmetry with a left superscript giving its multiplicity. An uppercase letter is placed before the symmetry symbol to indicate where it stands in order of energy: the ground state is designated X, higher states of the same multiplicity are designated A, B, C, \dots in order of increasing energy, and states of different multiplicity are designated a, b, c, \dots in order of increasing energy. (Sometimes, after a classification of states as A, B, C, etc has become well established, new states will be discovered lying between, say, the B and C states. Then the new states may be designated B', B'' and so on rather

-18-

than renaming all the states.) For example, the C_2 molecule has a singlet ground state designated as $X^1\Sigma_g^+$, a triplet state designated a $^3\Pi_u$ lying only 700 cm^{-1} above the ground state, another triplet 5700 cm^{-1} higher in energy designated $b^3\Sigma_g^-$, a singlet state designated A $^1\Pi_u$ lying 8400 cm^{-1} above the ground state and many other known states [16]. A transition between the ground state and the A state would be designated as $A^1\Pi_u-X^1\Sigma_g^+$. The convention is to list the upper state first regardless of whether the transition is being studied in absorption or emission.

The electronic selection rules for linear molecules are as follows. $\Delta\Lambda = 0, \pm 1$. $\Delta S = 0$. Again, these are really

valid only in the absence of spin-orbital coupling and are modified in heavy molecules. For transitions between Σ states there is an additional rule that Σ^+ combines only with Σ^+ and Σ^- combines only with Σ^- , so that transitions between Σ^+ states and Σ^- states are forbidden. If the molecule has a centre of symmetry then there is an additional rule requiring that $g \leftrightarrow u$, while transitions between two g states or between two u states are forbidden.

We now turn to electronic selection rules for symmetrical nonlinear molecules. The procedure here is to examine the structure of a molecule to determine what symmetry operations exist which will leave the molecular framework in an equivalent configuration. Then one looks at the various possible point groups to see what group would consist of those particular operations. The character table for that group will then permit one to classify electronic states by symmetry and to work out the selection rules. Character tables for all relevant groups can be found in many books on spectroscopy or group theory. Here we will only pick one very simple point group called C_{2v} and look at some simple examples to illustrate the method.

The C_{2v} group consists of four symmetry operations: an identity operation designated E , a rotation by one-half of a full rotation, i.e. by 180° , called a C_2 operation and two planes of reflection passing through the C_2 axis and called σ_v operations. Examples of molecules belonging to this point group are water, H_2O ; formaldehyde, H_2CO ; or pyridine, C_5H_5N . It is conventional to choose a molecule-fixed axis system with the z axis coinciding with the C_2 axis. If the molecule is planar, it is conventionally chosen to lie in the yz plane with the x axis perpendicular to the plane of the molecule [17]. For example, in H_2CO the C_2 or z axis lies along the C–O bond. One of the σ_v planes would be the plane of the molecule, the yz plane. The other reflection plane is the xz plane, perpendicular to the molecular plane.

Table B1.1.1 gives the character table for the C_{2v} point group as it is usually used in spectroscopy. Because each symmetry operation leaves the molecular framework and hence the potential energy unchanged, it should not change the electron density or nuclear position density: i.e. the square of an electronic or vibrational wavefunction should remain unchanged. In a group with no degeneracies like this one, that means that a wavefunction itself should either be unchanged or should change sign under each of the four symmetry operations. The result of group theory applied to such functions is that there are only four possibilities for how they change under the operations, and they correspond to the four irreducible representations designated as A_1 , A_2 , B_1 and B_2 . The characters may be taken to describe what happens in each symmetry. For example, a function classified as B_1 would be unchanged by the identity operation or by $\sigma_v(xz)$ but would be changed in sign by the C_2 or $\sigma_v(yz)$ operations. Every molecular orbital and every stationary state described by a many-electron wavefunction can be taken to belong to one of these symmetry classes. The same applies to vibrations and vibrational states.

-19-

Table B1.1.1 Character table for the C_{2v} point group.

C_{2v}	E	C_2	$\sigma_v(xz)$	$\sigma_v(yz)$		
A_1	1	1	1	1	z	x^2, y^2, z^2
A_2	1	1	-1	-1	R_z	xy
B_1	1	-1	1	-1	x, R_y	xz
B_2	1	-1	-1	1	y, R_x	yz

The last two columns of the character table give the transformation properties of translations along the x , y ,

and z directions, rotations about the three axes represented by R_x , etc, and products of two coordinates or two translations represented by x^2 , xy , etc. The information in these columns is very useful for working out selection rules.

Whenever a function can be written as a product of two or more functions, each of which belongs to one of the symmetry classes, the symmetry of the product function is the direct product of the symmetries of its constituents. This direct product is obtained in non-degenerate cases by taking the product of the characters for each symmetry operation. For example, the function xy will have a symmetry given by the direct product of the symmetries of x and of y ; this direct product is obtained by taking the product of the characters for each symmetry operation. In this example it may be seen that, for each operation, the product of the characters for B_1 and B_2 irreducible representations gives the character of the A_2 representation, so xy transforms as A_2 .

The applications to selection rules work as follows. Intensities depend on the values of the transition moment integral of equation (B1.1.1):

$$\mu_{ij} = \int \Psi_i^* \mu \Psi_j \, d\tau.$$

An integral like this must vanish by symmetry if the integrand is antisymmetric under any symmetry operation, i.e. it vanishes unless the integrand is totally symmetric. For C_{2v} molecules that means the integrand must have symmetry A_1 . The symmetry of the integrand is the direct product of the symmetries of the three components in the integral. The transition moment operator is a vector with three components, μ_x , μ_y and μ_z , which transform like x , y and z , respectively. To see if a transition between state i and state j is allowed, one determines the symmetries of the three products containing the three components of μ , i.e. $\Psi_i^* \mu_x \Psi_j$ and $\Psi_i^* \mu_z \Psi_j$. If any one of them is totally symmetrical, the transition is formally allowed. If none of the three is totally symmetrical the transition is forbidden. It should be noted that being allowed does not mean that a transition will be strong. The actual intensity depends on the matrix element μ_{ij} , whose value will depend on the details of the wavefunctions.

-20-

There is a further item of information in this procedure. If one of the three component integrals is non-zero, the molecule will absorb light polarized along the corresponding axis. For example, if $\mu_{y,ij} = \int \Psi_i^* \mu_y \Psi_j \, d\tau$ is non-zero, the transition will absorb light polarized along the y axis. One may be able to observe this polarization directly in the spectrum of a crystal containing the molecules in known orientations. Alternatively, in the gas phase one may be able to tell the direction of polarization in the molecular framework by looking at the intensity distribution among the rotational lines in a high-resolution spectrum.

Analogous considerations can be used for magnetic dipole and electric quadrupole selection rules. The magnetic dipole operator is a vector with three components that transform like R_x , R_y and R_z . The electric quadrupole operator is a tensor with components that transform like x^2 , y^2 , z^2 , xy , yz and xz . These latter symmetries are also used to get selection rules for Raman spectroscopy. Character tables for spectroscopic use usually show these symmetries to facilitate such calculations.

When spectroscopists speak of electronic selection rules, they generally mean consideration of the integral over only the electronic coordinates for wavefunctions calculated at the equilibrium nuclear configuration of the initial state, $Q = 0$,

$$\mu_{lu}(0) = \int \psi_l^*(x, 0) \mu \psi_u(x, 0) \, dx. \quad (\text{B1.1.18})$$

If one of the components of this electronic transition moment is non-zero, the electronic transition is said to be allowed; if all components are zero it is said to be forbidden. In the case of diatomic molecules, if the transition is forbidden it is usually not observed unless as a very weak band occurring by magnetic dipole or electric quadrupole interactions. In polyatomic molecules forbidden electronic transitions are still often observed, but they are usually weak in comparison with allowed transitions.

The reason they appear is that symmetric polyatomic molecules always have some non-totally-symmetric vibrations. When the nuclear framework is displaced from the equilibrium position along such a vibrational coordinate, its symmetry is reduced. It can then be thought of as belonging, for the instant, to a different point group of lower symmetry, and it is likely in that group that the transition will be formally allowed. Even though $\mu_{lu}(0)$ from equation (B1.1.18) is zero, $\mu_{lu}(Q)$ from equation (B1.1.5) will be non-zero for some configurations Q involving distortion along antisymmetric normal coordinates. The total integrated intensity of an electronic band is given by (b1.1.10). It involves the square of the electronic transition moment averaged over the initial vibrational state, including the configurations in which the transition moment is not zero. In suitable cases it may be possible to calculate the first-order terms of equation (B1.1.7) from electronic wavefunctions and use them in equation (B1.1.5) to calculate an integrated absorption intensity to compare with the observed integrated intensity or oscillator strength [18].

The spin selection rule for polyatomic molecules is again $\Delta S = 0$, no change in spin in the absence of spin-orbital coupling. This rule becomes less valid when heavy atoms are included in the molecule. Spin-changing transitions can be observed by suitable techniques even in hydrocarbons, but they are quite weak. When spin-orbital coupling is important it is possible to use the symmetries of the spin wavefunctions, assign symmetries to total orbital-plus-spin wavefunctions and use group theory as above to get the selection rules.

-21-

Most stable polyatomic molecules whose absorption intensities are easily studied have filled-shell, totally symmetric, singlet ground states. For absorption spectra starting from the ground state the electronic selection rules become simple: transitions are allowed to excited singlet states having symmetries the same as one of the coordinate axes, x , y or z . Other transitions should be relatively weak.

(C) VIBRONIC SELECTION RULES

Often it is possible to resolve vibrational structure of electronic transitions. In this section we will briefly review the symmetry selection rules and other factors controlling the intensity of individual vibronic bands.

In the Born–Oppenheimer approximation the vibronic wavefunction is a product of an electronic wavefunction and a vibrational wavefunction, and its symmetry is the direct product of the symmetries of the two components. We have just discussed the symmetries of the electronic states. We now consider the symmetry of a vibrational state. In the harmonic approximation vibrations are described as independent motions along normal modes Q_i and the total vibrational wavefunction is a product of functions, one wavefunction for each normal mode:

$$\phi(Q) = \varphi_{v_1}(Q_1)\varphi_{v_2}(Q_2)\varphi_{v_3}(Q_3)\dots \quad (\text{B1.1.19})$$

Each such normal mode can be assigned a symmetry in the point group of the molecule. The wavefunctions for non-degenerate modes have the following simple symmetry properties: the wavefunctions with an odd vibrational quantum number v_i have the same symmetry as their normal mode Q_i ; the ones with an even v_i are totally symmetric. The symmetry of the total vibrational wavefunction $\phi(Q)$ is then the direct product of the symmetries of its constituent normal coordinate functions $\varphi_{v_i}(Q_i)$. In particular, the lowest vibrational state,

with all $v_i = 0$, will be totally symmetric. The states with one quantum of excitation in one vibration and zero in all others will have the symmetry of that one vibration. Once the symmetry of the vibrational wavefunction has been established, the symmetry of the vibronic state is readily obtained from the direct product of the symmetries of the electronic state and the vibrational state. This procedure gives the correct vibronic symmetry even if the harmonic approximation or the Born–Oppenheimer approximation are not quite valid.

The selection rule for vibronic states is then straightforward. It is obtained by exactly the same procedure as described above for the electronic selection rules. In particular, the lowest vibrational level of the ground electronic state of most stable polyatomic molecules will be totally symmetric. Transitions originating in that vibronic level must go to an excited state vibronic level whose symmetry is the same as one of the coordinates, x , y , or z .

One of the consequences of this selection rule concerns forbidden electronic transitions. They cannot occur unless accompanied by a change in vibrational quantum number for some antisymmetric vibration. Forbidden electronic transitions are not observed in diatomic molecules (unless by magnetic dipole or other interactions) because their only vibration is totally symmetric; they have no antisymmetric vibrations to make the transitions allowed.

The symmetry selection rules discussed above tell us whether a particular vibronic transition is allowed or forbidden, but they give no information about the intensity of allowed bands. That is determined by [equation \(B1.1.9\)](#) for absorption or [\(B1.1.13\)](#) for emission. That usually means by the Franck–Condon principle if only the zero-order term in [equation \(B1.1.7\)](#) is needed. So we take note of some general principles for Franck–Condon factors (FCFs).

-22-

Usually the normal coordinates of the upper and lower states are quite similar. (When they are not it is called a Duschinsky effect [19] and the treatment becomes more complicated.) Because of the product form of the vibrational wavefunctions of [equation \(B1.1.19\)](#) the FCF is itself a product of FCFs for individual normal modes. If there is little or no change in the geometry of a given normal mode, the FCF for that mode will be large only if its vibrational quantum number does not change. But for modes for which there is a significant change in geometry, the FCFs may be large for a number of vibrational levels in the final state. The spectrum then shows a series of vibronic peaks differing in energy by the frequency of that vibration. Such a series is referred to as a progression in that mode.

Most commonly, the symmetry point group of the lower and upper states will be the same. Then only totally symmetric vibrations can change equilibrium positions—a change in a non-totally-symmetric mode would mean that the states have configurations belonging to different point groups. So one may expect to see progressions in one or more of the totally symmetric vibrations but not in antisymmetric vibrations. In symmetry-forbidden electronic transitions, however, one will see changes of one quantum (or possibly other odd numbers of quanta) in antisymmetric vibrations as required to let the transition appear.

An example of a single-absorption spectrum illustrating many of the effects discussed in this section is the spectrum of formaldehyde, H_2CO , shown in figure B1.1.2 [20]. This shows the region of the lowest singlet–singlet transition, the $A^1A_2-X^1A_1$ transition. This is called an $n \rightarrow \pi^*$ transition; the electronic change is the promotion of an electron from a non-bonding orbital (symmetry B_2) mostly localized on the oxygen atom into an antibonding π^* orbital (symmetry B_1) on the C–O bond. By the electronic selection rules, a transition from the totally symmetric ground state to a 1A_2 state is symmetry forbidden, so the transition is quite weak with an oscillator strength of 2.4×10^{-4} . The transition is appearing with easily measured intensity due to coupling with antisymmetric vibrations. Most of the intensity is induced by distortion along the out-of-plane coordinate, Q_4 . This means that in [equation \(B1.1.7\)](#) the most significant derivative of μ_{lu} is the one with respect to Q_4 . The first peak seen in figure B1.1.2, at 335 nm, has one quantum of vibration v_4 excited in the upperstate. The band is designated as which means that vibration number 4 has 1 quantum of excitation in the

upper state and 0 quanta in the lower state. The symmetry of Q_4 is B_1 , and combined with the A_2 symmetry of the electronic state it gives an upper state of vibronic symmetry B_2 , the direct product $A_2 \times B_1$. It absorbs light with its electric vector polarized in the y direction, i.e. in plane and perpendicular to the C–O bond.

Figure B1.1.2. Spectrum of formaldehyde with vibrational resolution. Several vibronic origins are marked. One progression in ν_2 starting from the origin is indicated on the line along the top. A similar progression is built on each vibronic origin. Reprinted with permission from [20]. Copyright 1982, American Chemical Society.

-23-

If the 0–0 band were observable in this spectrum it would be called the origin of the transition. The 4_0^1 band is referred to as a vibronic origin. Starting from it there is a progression in ν_2 , the C–O stretching mode, which gets significantly longer in the upper state because the presence of the antibonding π electron weakens the bond. Several of the peaks in the progression are marked along the line at the top of the figure.

Several other vibronic origins are also marked in this spectrum. The second major peak is the 4_0^3 band, with three quanta of ν_4 in the upper state. This upper state has the same symmetry as the state with one quantum. Normally, one would not expect much intensity in this peak, but it is quite strong because the excited state actually has a non-planar equilibrium geometry, i.e. it is distorted along Q_4 . Every vibronic origin including this one has a progression in ν_2 built on it. The intensity distribution in a progression is determined by the Franck–Condon principle and, as far as can be determined, all progressions in this spectrum are the same.

At 321 nm there is a vibronic origin marked 5_0^1 . This has one quantum of ν_5 , the antisymmetric C–H stretching mode, in the upper state. Its intensity is induced by a distortion along Q_5 . This state has B_2 vibrational symmetry. The direct product of B_2 and A_2 is B_1 , so it has B_1 vibronic symmetry and absorbs x-polarized light. One can also see a $4_0^2 6_0^1$ vibronic origin which has the same symmetry and intensity induced by distortion along Q_6 .

A very weak peak at 348 nm is the 4_0^2 origin. Since the upper state here has two quanta of ν_4 , its vibrational symmetry is A_1 and the vibronic symmetry is A_2 , so it is forbidden by electric dipole selection rules. It is actually observed here due to a magnetic dipole transition [21]. By magnetic dipole selection rules the 1A_2 – 1A_1 electronic transition is allowed for light with its magnetic field polarized in the z direction. It is seen here as having about 1% of the intensity of the symmetry-forbidden electric dipole transition made allowed by vibronic coupling, or an oscillator strength around 10^{-6} . This illustrates the weakness of magnetic dipole transitions.

(D) ROTATIONAL SELECTION RULES

If the experimental technique has sufficient resolution, and if the molecule is fairly light, the vibronic bands discussed above will be found to have a fine structure due to transitions among rotational levels in the two states. Even when the individual rotational lines cannot be resolved, the overall shape of the vibronic band will be related to the rotational structure and its analysis may help in identifying the vibronic symmetry. The analysis of the band appearance depends on calculation of the rotational energy levels and on the selection rules and relative intensity of different rotational transitions. These both come from the form of the rotational wavefunctions and are treated by angular momentum theory. It is not possible to do more than mention a simple example here.

The simplest case is a ${}^1\Sigma\text{--}{}^1\Sigma$ transition in a linear molecule. In this case there is no orbital or spin angular momentum. The total angular momentum, represented by the quantum number J , is entirely rotational angular momentum. The rotational energy levels of each state approximately fit a simple formula:

$$E_J = BJ(J + 1) - DJ^2(J + 1)^2.$$

The second term is used to allow for centrifugal stretching and is usually small but is needed for accurate work. The quantity B is called the rotation constant for the state. In a rigid rotator picture it would have the value

-24-

and is usually quoted in reciprocal centimetres. I is the moment of inertia. In an actual molecule which is vibrating, the formula for B must be averaged over the vibrational state, i.e. one must use an average value of $1/I$. As a result B varies somewhat with vibrational level. The values of B and the moments of inertia obtained from the spectra are used to get structural information about the molecule. The bonding and hence the structure will be different in the two states, so the B values will generally differ significantly. They are called B' and B'' . The convention is to designate quantities for the upper state with a single prime and quantities for the lower state with a double prime.

The rotational selection rule for a ${}^1\Sigma\text{--}{}^1\Sigma$ transition is $\Delta J = \pm 1$. Lines which have $J' = J'' - 1$ are called P lines and the set of them is called the P branch of the band. Lines for which $J' = J'' + 1$ are called R lines and the set of them the R branch. (Although not seen in a ${}^1\Sigma\text{--}{}^1\Sigma$ transition, a branch with $J' = J''$ would be called a Q branch, one with $J' = J'' - 2$ would be an O branch, or one with $J' = J'' + 2$ would be an S branch, etc.) Individual lines may be labeled by giving J'' in parentheses like R(1), P(2), etc. For lines with low values of J , the R lines get higher in energy as J increases while the P lines get lower in energy with increasing J . If B'' and B' are sufficiently different, which is the usual case in electronic spectra, the lines in one of the two branches will get closer together as J increases until they pile up on top of each other and then turn around and start to move in the other direction as J continues to increase. The point at which the lines pile up is called a band head. It is often the most prominent feature of the band. If $B'' > B'$, this will happen in the R branch and the band head will mark the high-energy or short-wavelength limit of each vibronic band. Such a band is said to be shaded to the red because absorption or emission intensity has a sharp onset on the high-energy side and then falls off gradually on the low-energy or red side. This is the most common situation where the lower state is bound more tightly and has a smaller moment of inertia than the upper state. But the opposite can occur as well. If $B'' < B'$ the band head will be formed in the P branch on the low-energy side of the vibronic band, and the band will be said to be shaded to the violet or shaded to the blue. Note that the terms red for the low-energy direction and violet or blue for the high-energy direction are used even for spectra in the ultraviolet or infrared regions where the actual visible red and blue colours would both be in the same direction.

The analysis of rotational structure and selection rules for transitions involving Π or Δ states becomes considerably more complicated. In general, Q branches will be allowed as well as P and R branches. The coupling between different types of angular momenta—orbital angular momentum of the electrons, spin angular momentum for states of higher multiplicity, the rotational angular momentum and even nuclear spin terms—is a complex subject which cannot be covered here: the reader is referred to the more specialized literature.

B1.1.3.4 PERTURBATIONS

Spectroscopists working with high-resolution spectra of small molecules commonly fit series of rotational lines to formulae involving the rotational constants, angular momentum coupling terms, etc. However, occasionally they find that some lines in the spectrum are displaced from their expected positions in a systematic way. Of course, a displacement of a line from its expected position means that the energy levels of one of the states are displaced from their expected energies. Typically, as J increases some lines will be seen to be displaced in one direction by increasing amounts up to a maximum at some particular J , then for the next J the line will be displaced in the opposite direction, and then as J increases further the lines will gradually approach their expected positions. These displacements of lines and of state energies are called perturbations [22].

-25-

They are caused by interactions between states, usually between two different electronic states. One hard and fast selection rule for perturbations is that, because angular momentum must be conserved, the two interacting states must have the same J . The interaction between two states may be treated by second-order perturbation theory which says that the displacement of a state is given by

$$\Delta E_1 = \frac{|H'_{12}|^2}{E_1^0 - E_2^0}$$

where H'_{12} is the matrix element between the two states of some small term H' in the Hamiltonian which is unimportant in determining the expected energies of the states. This interaction always has the effect of pushing the two states apart in energy by equal and opposite displacements inversely proportional to the zero-order separation of the two states. The perturbation is observed when the vibronic level of the state with the larger B value lies slightly lower in energy than a vibronic level of the other state. Then with increasing J the energy of the rotating level of the first state gets closer and closer to the energy of the second state and finally passes it and then gets farther away again. The maximum displacements occur at the J values where the two energies are the closest.

The spectral perturbations are observed in a transition involving one of the interacting states. Sometimes it is possible also to see an electronic transition involving the other of the interacting states, and then one should see equal but opposite displacements of rotational levels with the same J .

An interesting example occurs in the spectrum of the C_2 molecule. The usual rule of absorption spectroscopy is that the transitions originate in the ground electronic state because only it has sufficient population. However, in C_2 transitions were observed starting both from a $^3\Pi_u$ state and from a $^1\Sigma_g^+$ state, so it was not clear which was the ground state. The puzzle was solved by Ballik and Ramsay [23] who observed perturbations in a $^3\Sigma_g^- - ^3\Pi_u$ transition due to displacements of levels in the $^3\Sigma_g^-$ state. They then reinvestigated a $^1\Pi_u - ^1\Sigma_g^+$ transition known as the Phillips system, and they observed equal and opposite displacements of levels in the $^1\Sigma_g^+$ state, thus establishing that the $^1\Sigma_g^+$ and $^3\Sigma_g^-$ states were perturbing each other. For example, in the $\nu = 4$ vibrational level of the $^1\Sigma_g^+$ state, the $J = 40$ rotational level was displaced to lower energy by 0.26 cm^{-1} ; correspondingly, in the $\nu = 1$ vibrational level of the $^3\Sigma_g^-$ state, the $J = 40$ level was displaced upwards by 0.25 cm^{-1} . The values have an uncertainty of 0.02 cm^{-1} , so the displacements of the levels with the same J are equal and opposite within experimental error. Similarly, the $J = 42$ level in the $^1\Sigma_g^+$ state was displaced upwards by 0.17 cm^{-1} and the $J = 42$ level of the $^3\Sigma_g^-$ state displaced downwards by 0.15 cm^{-1} . These observations established that these particular levels were very close to each other in energy and the authors

were able to prove that the ${}^1\Sigma_g^+$ was lower by about 600 cm^{-1} than the ${}^3\Pi_u$ state and was the ground state. Absorption spectra of C_2 are typically observed in the vapour over graphite at high temperatures. At 2500 K the value of kT is about 1700 cm^{-1} , much greater than the energy separation of the two states. Since the ${}^1\Sigma_g^+$ state is non-degenerate and the ${}^3\Pi_u$ state has a sixfold degeneracy, most of the molecules are actually in the upper state. This accounts for the observation of absorptions starting from both states.

The perturbations in this case are between a singlet and a triplet state. The perturbation Hamiltonian, H' , of the second-order perturbation theory is spin-orbital coupling, which has the effect of mixing singlet and triplet states.

-26-

The magnitude of the perturbations can be calculated fairly quantitatively from high-quality electronic wavefunctions including configuration interaction [24].

B1.1.4 EXAMPLES

B1.1.4.1 PHOTOPHYSICS OF MOLECULES IN SOLUTION

To understand emission spectroscopy of molecules and/or their photochemistry it is essential to have a picture of the radiative and non-radiative processes among the electronic states. Most stable molecules other than transition metal complexes have all their electrons paired in the ground electronic state, making it a singlet state. Figure B1.1.3 gives a simple state energy diagram for reference. Singlet states are designated by the letter S and numbered in order of increasing energy. The ground state is called S_0 . Excited singlet states have configurations in which an electron has been promoted from one of the filled orbitals to one of the empty orbitals of the molecule. Such configurations with two singly occupied molecular orbitals will give rise to triplet states as well as singlet states. A triplet state results when one of the electrons changes its spin so that the two electrons have parallel spin. Each excited singlet state will have its corresponding triplet state. Because the electron-electron repulsion is less effective in the triplet state, it will normally be lower in energy than the corresponding singlet state.

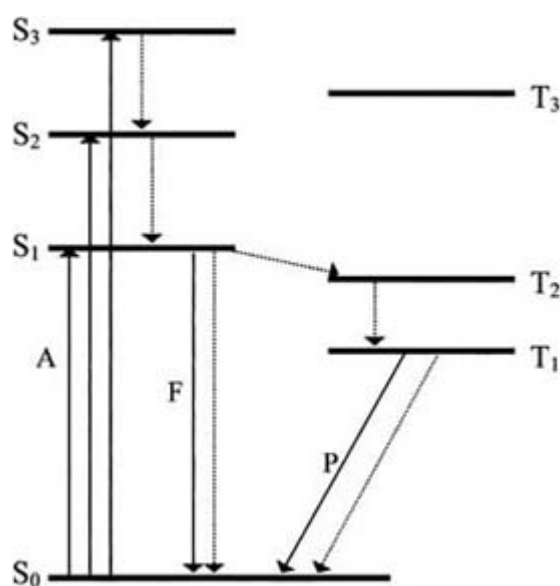


Figure B1.1.3. State energy diagram for a typical organic molecule. Solid arrows show radiative transitions; A: absorption, F: fluorescence, P: phosphorescence. Dotted arrows: non-radiative transitions.

Spectroscopists observed that molecules dissolved in rigid matrices gave both short-lived and long-lived emissions which were called fluorescence and phosphorescence, respectively. In 1944, Lewis and Kasha [25] proposed that molecular phosphorescence came from a triplet state and was long-lived because of the well known spin selection rule $\Delta S = 0$, i.e. interactions with a light wave or with the surroundings do not readily change the spin of the electrons.

-27-

Typical singlet lifetimes are measured in nanoseconds while triplet lifetimes of organic molecules in rigid solutions are usually measured in milliseconds or even seconds. In liquid media where diffusion is rapid the triplet states are usually quenched, often by the nearly ubiquitous molecular oxygen. Because of that, phosphorescence is seldom observed in liquid solutions. In the spectroscopy of molecules the term fluorescence is now usually used to refer to emission from an excited singlet state and phosphorescence to emission from a triplet state, regardless of the actual lifetimes.

If a light beam is used to excite one of the higher singlet states, say S_2 , a very rapid relaxation occurs to S_1 , the lowest excited singlet state. This non-radiative process just converts the difference in energy into heat in the surroundings. A radiationless transition between states of the same multiplicity is called internal conversion. Relaxation between states of the same multiplicity and not too far apart in energy is usually much faster than radiative decay, so fluorescence is seen only from the S_1 state. These radiationless processes in large molecules are the analogue of the perturbations observed in small molecules. They are caused by small terms in the Hamiltonian such as spin-orbital coupling or Born-Oppenheimer breakdown, which mix electronic states. The density of vibrational levels of large molecules can be very high and that makes these interactions into irreversible transitions to lower states.

Once the excited molecule reaches the S_1 state it can decay by emitting fluorescence or it can undergo a further radiationless transition to a triplet state. A radiationless transition between states of different multiplicity is called intersystem crossing. This is a spin-forbidden process. It is not as fast as internal conversion and often has a rate comparable to the radiative rate, so some S_1 molecules fluoresce and others produce triplet states. There may also be further internal conversion from S_1 to the ground state, though it is not easy to determine the extent to which that occurs. Photochemical reactions or energy transfer may also occur from S_1 .

Molecules which reach a triplet state will generally relax quickly to state T_1 . From there they can emit phosphorescence or decay by intersystem crossing back to the ground state. Both processes are spin forbidden and again often have comparable rates. The T_1 state is often also important for photochemistry because its lifetime is relatively long. Both phosphorescence and intersystem crossing are dependent on spin-orbital coupling and are enhanced by heavy atoms bound to the molecule or in the environment. They are also enhanced by the presence of species with unpaired electrons such as O_2 because electron exchange can effect such transitions without actually requiring the spin of an electron to be reversed. O_2 is found to quench both fluorescence and phosphorescence, and it is often necessary to remove oxygen from solutions for precise emission measurements.

B1.1.4.2 WIDTHS AND SHAPES OF SPECTRAL LINES

High-resolution spectroscopy used to observe hyperfine structure in the spectra of atoms or rotational structure in electronic spectra of gaseous molecules commonly must contend with the widths of the spectral lines and how that compares with the separations between lines. Three contributions to the linewidth will be mentioned here: the natural line width due to the finite lifetime of the excited state, collisional broadening of lines, and the Doppler effect.

The most fundamental limitation on sharpness of spectral lines is the so-called natural linewidth. Because an

excited state has a finite lifetime, the intensity of light it emits falls off exponentially as a function of time. A beam of light whose intensity varies with time cannot have a single frequency. Its spectral distribution is the Fourier transform of its temporal shape. For an exponential decay the spectral distribution will have the form

-28-

$$I(\nu) = I(\nu_0) \frac{(1/4\pi\tau)^2}{(\nu - \nu_0)^2 + (1/4\pi\tau)^2} \quad (\text{B1.1.20})$$

where ν_0 is the frequency of the centre of the band and τ is the mean life of the excited state. The same formula applies to lines in the absorption spectrum. This shape is called a Lorentzian shape. Its full width at half maximum (FWHM) is $1/(2\pi\tau)$. The shorter the lifetime, the broader the line. Another way to think about the width is to say that the energy of a state has an uncertainty related to its lifetime by the uncertainty principle. If the transition is coupling two states, both of which have finite lifetimes and widths, it is necessary to combine the effects.

Spectral lines are further broadened by collisions. To a first approximation, collisions can be thought of as just reducing the lifetime of the excited state. For example, collisions of molecules will commonly change the rotational state. That will reduce the lifetime of a given state. Even if the state is not changed, the collision will cause a phase shift in the light wave being absorbed or emitted and that will have a similar effect. The line shapes of collisionally broadened lines are similar to the natural line shape of equation (B1.1.20) with a lifetime related to the mean time between collisions. The details will depend on the nature of the intermolecular forces. We will not pursue the subject further here.

A third source of spectral broadening is the Doppler effect. Molecules moving toward the observer will emit or absorb light of a slightly higher frequency than the frequency for a stationary molecule; those moving away will emit or absorb a slightly lower frequency. The magnitude of the effect depends on the speed of the molecules. To first order the frequency shift is given by

$$\Delta\nu = \nu_0(v_x/c)$$

where v_x is the component of velocity in the direction of the observer and c is the speed of light.

For a sample at thermal equilibrium there is a distribution of speeds which depends on the mass of the molecules and on the temperature according to the Boltzmann distribution. This results in a line shape of the form

$$I(\nu) = I(\nu_0) \exp\left[\frac{-Mc^2}{2\nu_0^2 RT}(\nu - \nu_0)^2\right]$$

where M is the atomic or molecular mass and R the gas constant. This is a Gaussian line shape with a width given by

$$\text{FWHM} = \frac{2\nu_0}{c} \left(\frac{2RT \ln 2}{M}\right)^{1/2}.$$

The actual line shape in a spectrum is a convolution of the natural Lorentzian shape with the Doppler shape. It must be calculated for a given case as there is no simple formula for it. It is quite typical in electronic

spectroscopy to have the FWHM determined mainly by the Doppler width. However, the two shapes are quite different and the Lorentzian shape

-29-

does not fall off as rapidly at large $(\nu - \nu_0)$. It is likely that the intensity in the wings of the line will be determined by the natural line shape.

Collisional broadening is reduced by the obvious remedy of working at low pressures. Of course, this reduces the absorption and may require long path lengths for absorption spectroscopy. Doppler widths can be reduced by cooling the sample. For many samples this is not practical because the molecules will have too low a vapour pressure. Molecular beam methods and the newer technique of jet spectroscopy can be very effective by restricting the motion of the molecules to be at right angles to the light beam. Some other techniques for sub-Doppler spectroscopy have also been demonstrated using counter-propagating laser beams to compensate for motion along the direction of the beam. The natural linewidth, however, always remains when the other sources of broadening are removed.

B1.1.4.3 RYDBERG SPECTRA

The energies of transitions of a hydrogen atom starting from the ground state fit exactly the equation

where R is the Rydberg constant, E_1 is the ionization energy of the atom (which in hydrogen is equal to the Rydberg constant) and n is the principal quantum number of the electron in the upper state. The spectrum shows a series of lines of increasing n which converge to a limit at the ionization energy.

Other atoms and molecules also show similar series of lines, often in the vacuum ultraviolet region, which fit approximately a similar formula:

Such a series of lines is called a Rydberg series [26]. These lines also converge to the ionization energy of the atom or molecule, and fitting the lines to this formula can give a very accurate value for the ionization energy. In the case of molecules there may be resolvable vibrational and rotational structure on the lines as well.

The excited states of a Rydberg series have an electron in an orbital of higher principal quantum number, n , in which it spends most of its time far from the molecular framework. The idea is that the electron then feels mainly a Coulomb field due to the positive ion remaining behind at the centre, so its behaviour is much like that of the electron in the hydrogen atom. The constant δ is called the quantum defect and is a measure of the extent to which the electron interacts with the molecular framework. It has less influence on the energy levels as n gets larger, i.e. as the electron gets farther from the central ion. The size of δ will also depend on the angular momentum of the electron. States of lower angular momentum have more probability of penetrating the charge cloud of the central ion and so may have larger values of δ . Actual energy levels of Rydberg atoms and molecules can be subject to theoretical calculations [27]. Sometimes the higher states have orbitals so large that other molecules may fall within their volume, causing interesting effects [28].

-30-

B1.1.4.4 MULTIPHOTON SPECTROSCOPY

All the previous discussion in this chapter has been concerned with absorption or emission of a single photon. However, it is possible for an atom or molecule to absorb two or more photons simultaneously from a light beam to produce an excited state whose energy is the sum of the energies of the photons absorbed. This can happen even when there is no intermediate stationary state of the system at the energy of one of the photons. The possibility was first demonstrated theoretically by Maria Göppert-Mayer in 1931 [29], but experimental observations had to await the development of the laser. Multiphoton spectroscopy is now a useful technique [30, 31].

The transition probability for absorption of two photons can be described in terms of a two-photon cross section δ by

$$-dI = \delta I^2 N dl$$

where I is a photon flux in photons $\text{cm}^{-2} \text{s}^{-1}$, N is the number of molecules per cubic centimetre, and l is distance through the sample. Measured values of δ are of the order of $10^{-50} \text{ cm}^4 \text{ s photon}^{-1} \text{ molecule}^{-1}$ [32]. For molecules exposed to the intensity of sunlight at the earth's surface this would suggest that the molecule might be excited once in the age of the universe. However, the probability is proportional to the square of the light intensity. For a molecule exposed to a pulsed laser focused to a small spot, the probability of being excited by one pulse may be easily observable by fluorescence excitation or multiphoton ionization techniques.

One very important aspect of two-photon absorption is that the selection rules for atoms or symmetrical molecules are different from one-photon selection rules. In particular, for molecules with a centre of symmetry, two-photon absorption is allowed only for $g \leftrightarrow g$ or $u \leftrightarrow u$ transitions, while one-photon absorption requires $g \leftrightarrow u$ transitions. Therefore, a whole different set of electronic states becomes allowed for two-photon spectroscopy. The group-theoretical selection rules for two-photon spectra are obtained from the symmetries of the x^2 , xy , etc. terms in the character table. This is completely analogous to the selection rules for Raman spectroscopy, a different two-photon process.

A good example is the spectrum of naphthalene. The two lowest excited states have B_{2u} and B_{1u} symmetries and are allowed for one-photon transitions. A weak transition to one of these is observable in the two-photon spectrum [33], presumably made allowed by vibronic effects. Much stronger two-photon transitions are observable at somewhat higher energies to a B_{3g} and an A_g state lying quite close to the energies predicted by theory many years earlier [34].

An interesting aspect of two-photon spectroscopy is that some polarization information is obtainable even for randomly oriented molecules in solution by studying the effect of the relative polarization of the two photons. This is readily done by comparing linearly and circularly polarized light. Transitions to A_g states will absorb linearly polarized light more strongly than circularly polarized light. The reverse is true of transitions to B_{1g} , B_{2g} , or B_{3g} states. The physical picture is that the first photon induces an oscillating u-type polarization of the molecule in one direction. To get to a totally symmetric A_g state the second photon must reverse that polarization, so is favoured for a photon of the same polarization. However, to get to, say, a B_{3g} state, the second photon needs to act at right angles to the first, and in circularly polarized light that perpendicular polarization is always strong. Figure B1.1.4 shows the two-photon fluorescence excitation spectrum of naphthalene in the region of the g states [35]. One peak shows stronger absorption for circularly polarized, one for linearly polarized light. That confirms the identification as B_{3g} and A_g states respectively.

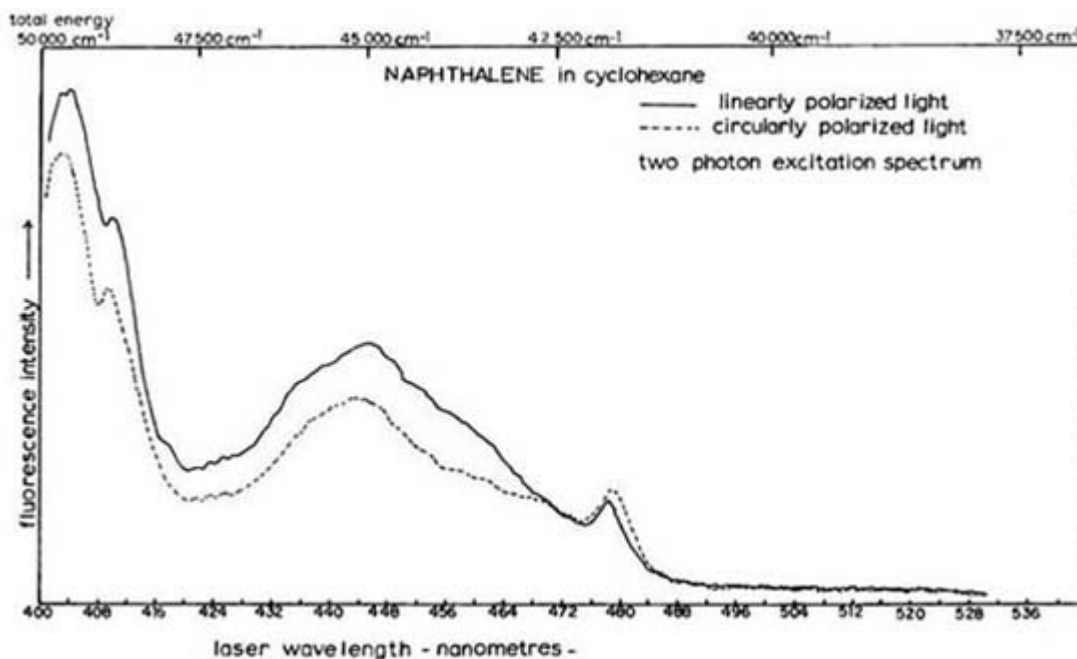


Figure B1.1.4. Two-photon fluorescence excitation spectrum of naphthalene. Reprinted from [35]. Courtesy, Tata McGraw-Hill Publishing Company Ltd, 7 West Patel Nagar, New Dehli, 110008, India.

Three-photon absorption has also been observed by multiphoton ionization, giving Rydberg states of atoms or molecules [36]. Such states usually require vacuum ultraviolet techniques for one-photon spectra, but can be done with a visible or near-ultraviolet laser by three-photon absorption.

B1.1.4.5 OTHER EXAMPLES

Many of the most interesting current developments in electronic spectroscopy are addressed in special chapters of their own in this encyclopedia. The reader is referred especially to [sections B2.1](#) on ultrafast spectroscopy, [C1.5](#) on single molecule spectroscopy, [C3.2](#) on electron transfer, and [C3.3](#) on energy transfer. Additional topics on electronic spectroscopy will also be found in many other chapters.

REFERENCES

- [1] Sawyer R A 1951 *Experimental Spectroscopy* 2nd edn (Englewood Cliffs, NJ: Prentice-Hall)
- [2] O'Conner D V and Phillips D 1984 *Time-Correlated Single Photon Counting* (London: Academic)
- [3] Wenthold P and Lineberger W C 1999 Negative ion photoelectron spectroscopy studies of organic reactive intermediates *Accts. Chem. Res.* **32** 597–604
- [4] Pauling L and Wilson E B 1935 *Introduction to Quantum Mechanics* (New York: McGraw-Hill) pp 294–314

- [5] Hilborn H 1982 Einstein coefficients, cross sections, f values, dipole moments and all that *Am. J. Phys.* **50** 982–6
- [6] Kuhn H 1958 Oscillator strength of absorption bands in dye molecules *J. Chem. Phys.* **29** 958–9
- [7] Kauzmann W 1957 *Quantum Chemistry* (New York: Academic) pp 651–3

- [8] Born M and Oppenheimer R 1927 Concerning the quantum theory of molecules *Ann. Phys., Lpz* **84** 457–84
- [9] Franck J 1925 Elementary processes of photochemical reactions *Trans. Faraday Soc.* **21** 536
- [10] Condon E U 1928 Nuclear motion associated with electron transitions in diatomic molecules *Phys. Rev.* **32** 858–72
 Condon E U 1947 The Franck–Condon principle and related topics *Am. J. Phys.* **15** 365–79
- [11] Greenblatt G D, Orlando J J, Burkholder J B and Ravishankara A R 1990 Absorption measurements of oxygen between 330 and 1140 nm *J. Geophys. Res.* **95** 18 577–82
- [12] Strickler S J and Vikesland J P 1974 ${}^3B_1-{}^1A_1$ transition of SO_2 gas. I. Franck–Condon treatment and transition moments *J. Chem. Phys.* **60** 660–3
- [13] Einstein A 1917 On the quantum theory of radiation *Phys. Z.* **18** 121–8
- [14] Strickler S J and Berg R A 1962 Relationship between absorption intensity and fluorescence lifetime of molecules *J. Chem. Phys.* **37** 814–22
- [15] Strickler S J, Vikesland J P and Bier H D 1974 ${}^3B_1-{}^1A_1$ transition of SO_2 gas. II. Radiative lifetime and radiationless processes *J. Chem. Phys.* **60** 664–7
- [16] Herzberg G, Lagerquist A and Malmberg C 1969 New electronic transitions of the C_2 molecule in absorption in the vacuum ultraviolet region *Can. J. Phys.* **47** 2735–43
- [17] Mulliken R S 1955 Report on notation for the spectra of polyatomic molecules *J. Chem. Phys.* **23** 1997–2011
- [18] Robey M J, Ross I G, Southwood-Jones R V and Strickler S J 1977 *A priori* calculations on vibronic coupling in the ${}^1B_{2u}-{}^1A_g$ (3200 Å) and higher transitions of naphthalene *Chem. Phys.* **23** 207–16
- [19] Duschinsky F 1937 On the interpretation of electronic spectra of polyatomic molecules. I. Concerning the Franck–Condon Principle *Acta Physicochimica URSS* **7** 551
- [20] Strickler S J and Barnhart R J 1982 Absolute vibronic intensities in the ${}^1A_2 \leftarrow {}^1A_1$ absorption spectrum of formaldehyde *J. Phys. Chem.* **86** 448–55
- [21] Callomon J H and Innes K K 1963 Magnetic dipole transition in the electronic spectrum of formaldehyde *J. Mol. Spectrosc.* **10** 166–81
- [22] Lefebvre-Brion H and Field R W 1986 *Perturbations in the Spectra of Diatomic Molecules* (Orlando: Academic)
- [23] Ballik E A and Ramsay D A 1963 The $A^3\Sigma_g^- - X^3\Pi_u$ band system of the C_2 molecule *Astrophys. J.* **137** 61–83
- [24] Langhoff S R, Sink M L, Pritchard R H, Kern C W, Strickler S J and Boyd M J 1977 *Ab initio* study of perturbations between the $X\Sigma_g^+$ and $b^3\Sigma_g^-$ states of the C_2 molecule *J. Chem. Phys.* **67** 1051–60
- [25] Lewis G N and Kasha M 1944 Phosphorescence and the triplet state *J. Am. Chem. Soc.* **66** 2100–16
- [26] Duncan A B F 1971 *Rydberg Series in Atoms and Molecules* (New York: Academic)
- [27] Sandorfy C (ed) 1999 *The Role of Rydberg States in Spectroscopy and Photochemistry* (London: Kluwer Academic)

- [28] Merkt F 1997 Molecules in high Rydberg states *Ann. Rev. Phys. Chem.* **48** 675–709
- [29] Göppert-Mayer M 1931 Concerning elementary processes with two quanta *Ann. Phys.* **9** 273–94
- [30] Ashfold M N R and Howe J D 1994 Multiphoton spectroscopy of molecular species *Ann. Rev. Phys. Chem.* **45** 57–82

- [31] Callis P R 1997 Two-photon induced fluorescence *Ann. Rev. Phys. Chem.* **48** 271–97
- [32] Monson P R and McClain W M 1970 Polarization dependence of the two-photon absorption of tumbling molecules with application to liquid 1-chloronaphthalene and benzene *J. Chem. Phys.* **53** 29–37
- [33] Mikami N and Ito M 1975 Two-photon excitation spectra of naphthalene and naphthalene-d₈ *Chem. Phys. Lett.* **31** 472–8
- [34] Pariser R 1956 Theory of the electronic spectra and structure of the polyacenes and of alternant hydrocarbons *J. Chem. Phys.* **24** 250–68
- [35] Strickler S J, Gilbert J V and McClanahan J E 1984 Two-photon absorption spectroscopy of molecules *Lasers and Applications* eds H D Bist and J S Goela (New Delhi: Tata McGraw-Hill) pp 351–61
- [36] Johnson P M 1976 The multiphoton ionization spectrum of benzene *J. Chem. Phys.* **64** 4143–8
-

FURTHER READING

The pre-eminent reference works in spectroscopy are the set of books by G Herzberg.

Herzberg G 1937 *Atomic Spectra and Atomic Structure* (New York: Prentice-Hall)

Herzberg G 1950 *Molecular Spectra and Molecular Structure I. Spectra of Diatomic Molecules* 2nd edn (Princeton, NJ: Van Nostrand)

Herzberg G 1945 *Molecular Spectra and Molecular Structure II. Infrared and Raman Spectra of Polyatomic Molecules* (Princeton, NJ: Van Nostrand)

Herzberg G 1966 *Molecular Spectra and Molecular Structure III. Electronic Spectra and Electronic Structure of Polyatomic Molecules* (Princeton, NJ: Van Nostrand)

Huber K P and Herzberg G 1979 *Molecular Spectra and Molecular Structure IV. Constants of Diatomic Molecules* (New York: Van Nostrand-Reinhold)

Herzberg G 1971 *The Spectra and Structures of Simple Free Radicals: An Introduction to Molecular Spectroscopy* (Ithaca, NY: Cornell University Press)

-1-

B1.2 Vibrational spectroscopy

Charles Schmittenmaer

B1.2.1 INTRODUCTION

B1.2.1.1 OVERVIEW

Vibrational spectroscopy provides detailed information on both structure and dynamics of molecular species. Infrared (IR) and Raman spectroscopy are the most commonly used methods, and will be covered in detail in this chapter. There exist other methods to obtain vibrational spectra, but those are somewhat more specialized and used less often. They are discussed in other chapters, and include: inelastic neutron scattering (INS), helium atom scattering, electron energy loss spectroscopy (EELS), photoelectron spectroscopy, among others.

Vibrational spectra span the frequency range 10–4000 cm⁻¹ (10 cm⁻¹ = 1.2 meV = 0.03 kcal mol⁻¹ = 0.11 kJ mol⁻¹, and 4000 cm⁻¹ = 496 meV = 10.3 kcal mol⁻¹ = 42.9 kJ mol⁻¹), depending on the strength of the bond

and the reduced mass of the vibrational mode. Very weakly bound species, such as clusters bound only by van der Waals forces, or condensed phase modes with very large effective reduced masses, such as intermolecular modes in liquids or solids, or large amplitude motions in proteins or polymers, have very low frequency vibrations of 10–100 cm^{-1} . Modes involving many atoms in moderately large molecules absorb in the range 300–1200 cm^{-1} . The region 600–1200 cm^{-1} is referred to as the fingerprint region because while many organic molecules will all have bands due to vibrations of C–H, C = O, O–H and so on, there will be low frequency bands unique to each molecule that involve complicated motions of many atoms. The region 1200–3500 cm^{-1} is where most functional groups are found to absorb. Thus, the presence or absence of absorption at a characteristic frequency helps to determine the identity of a compound. The frequency of the absorption can be justified in terms of the masses of the atoms participating, the type of motion involved (stretch versus bend) and the bond strengths. The H_2 molecule has a reasonably large force constant and the smallest reduced mass, which results in the highest vibrational frequency at 4400 cm^{-1} . The width and intensity of an absorption feature provide information in addition to the absorption frequency. In favourable cases in the gas phase the width is determined by the vibrational lifetime or even the lifetime of a molecule if the vibrational energy is greater than the bond strength. The intensity yields information on the nature of the vibrational motion, and can also be used to determine the temperature of the sample.

Infrared and Raman spectroscopy each probe vibrational motion, but respond to a different manifestation of it. Infrared spectroscopy is sensitive to a change in the dipole moment as a function of the vibrational motion, whereas Raman spectroscopy probes the change in polarizability as the molecule undergoes vibrations. Resonance Raman spectroscopy also couples to excited electronic states, and can yield further information regarding the identity of the vibration. Raman and IR spectroscopy are often complementary, both in the type of systems that can be studied, as well as the information obtained.

Vibrational spectroscopy is an enormously large subject area spanning many scientific disciplines. The methodology, both experimental and theoretical, was developed primarily by physical chemists and has branched far and wide over the last 50 years. This chapter will mainly focus on its importance with regard to physical chemistry.

-2-

B1.2.1.2 INFRARED SPECTROSCOPY

For many chemists, the most familiar IR spectrometer is the dual beam instrument that covers the region 900–3400 cm^{-1} ; it is used for routine analysis and compound identification. Typically, each of the functional groups of a molecule have unique frequencies, and different molecules have different combinations of functional groups. Thus, every molecule has a unique absorption spectrum. Of course, there can be situations where two molecules are similar enough that their spectra are indistinguishable on a system with moderate signal-to-noise ratio, or where there are strong background absorptions due to a solvent or matrix that obscures the molecular vibrations, so that it is not possible to distinguish all compounds under all circumstances; but it is usually quite reliable, particularly if one is comparing the spectrum of an unknown to reference spectra of a wide variety of compounds. Ease of implementation and reasonably unambiguous spectra have led to the widespread use of IR spectroscopy outside of physical chemistry.

Within physical chemistry, the long-lasting interest in IR spectroscopy lies in structural and dynamical characterization. High resolution vibration–rotation spectroscopy in the gas phase reveals bond lengths, bond angles, molecular symmetry and force constants. Time-resolved IR spectroscopy characterizes reaction kinetics, vibrational lifetimes and relaxation processes.

B1.2.1.3 RAMAN SPECTROSCOPY

Raman spectrometers are not as widespread as their IR counterparts. This is partially due to the more stringent

requirements on light source (laser) and monochromator. As is the case with IR spectroscopy, every molecule has a unique Raman spectrum. It is also true that there can be ambiguity because of molecular similarities or impurities in the sample. Resonance Raman spectroscopy allows interfering bands to be eliminated by selectively exciting only specific species by virtue of their electronic absorption, or coupling to a nearby chromophore. This is particularly helpful in discriminating against strong solvent bands. For example, the first excited electronic state of water is at about 7 eV (~175 nm excitation wavelength), whereas many larger molecules have electronic transitions at much lower photon energy. By using the resonant enhancement of the Raman signal from the excited electronic state, it is possible to obtain a factor of 10^6 enhancement of the dissolved molecule.

One of the well known advantages of resonance Raman spectroscopy is that samples dissolved in water can be studied since water is transparent in the visible region. Furthermore, many molecules of biophysical interest assume their native state in water. For this reason, resonance Raman spectroscopy has been particularly strongly embraced in the biophysical community.

B1.2.2 THEORY

B1.2.2.1 CLASSICAL DESCRIPTION

Both infrared and Raman spectroscopy provide information on the vibrational motion of molecules. The techniques employed differ, but the underlying molecular motion is the same. A qualitative description of IR and Raman spectroscopies is first presented. Then a slightly more rigorous development will be described. For both IR and Raman spectroscopy, the fundamental interaction is between a dipole moment and an electromagnetic field. Ultimately, the two

can only couple with each other if they have the same frequency, otherwise the time average of their interaction energy is zero.

The most important consideration for a vibration to be IR active is that its dipole moment *changes* upon vibration. That is to say, its dipole derivative must be nonzero. The time-dependence of the dipole moment for a heteronuclear diatomic is shown in figure B1.2.1. Classically, an oscillating dipole radiates energy at the oscillation frequency. In a sense, this is true for a vibrating molecule in that when it is in an excited vibrational state it can emit a photon and lose energy. However, there are two fundamental differences. First, it does not continuously radiate energy, as a classical oscillator would. Rather, it vibrates for long periods without radiating energy, and then there is an instantaneous jump to a lower energy level accompanied by the emission of a photon. It should be noted that vibrational lifetimes in the absence of external perturbations are quite long, on the millisecond timescale. The second difference from a classical oscillator is that when a molecule is in its ground vibrational state it cannot emit a photon, but is still oscillating. Thus, the dipole can oscillate for an indefinitely long period without radiating any energy. Therefore, a quantum mechanical description of vibration must be invoked to describe molecular vibrations at the most fundamental level.

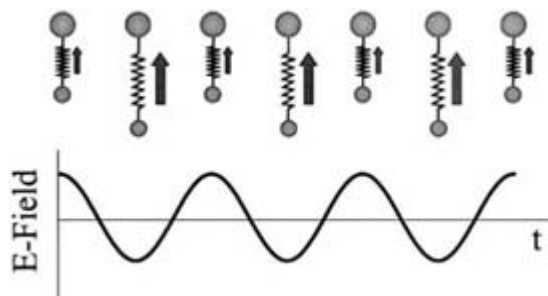


Figure B1.2.1. Schematic representation of the dependence of the dipole moment on the vibrational coordinate for a heteronuclear diatomic molecule. It can couple with electromagnetic radiation of the same frequency as the vibration, but at other frequencies the interaction will average to zero.

The qualitative description of Raman scattering is closely related. In this case, the primary criterion for a Raman active mode is that the polarizability of the molecule changes as a function of vibrational coordinate. An atom or molecule placed in an electric field will acquire a dipole moment that is proportional to the size of the applied field and the ease at which the charge distribution redistributes, that is, the polarizability. If the polarizability changes as a function of vibration, then there will be an induced dipole whose magnitude changes as the molecule vibrates, as depicted in [figure B1.2.2](#) and this can couple to the EM field. Of course, the applied field is oscillatory, not static, but as we will see below there will still be scattered radiation that is related to the vibrational frequency. In fact, the frequency of the scattered light will be the sum and difference of the laser frequency and vibrational frequency.

-4-



Figure B1.2.2. Schematic representation of the polarizability of a diatomic molecule as a function of vibrational coordinate. Because the polarizability *changes* during vibration, Raman scatter will occur in addition to Rayleigh scattering.

Before presenting the quantum mechanical description of a harmonic oscillator and selection rules, it is worthwhile presenting the energy level expressions that the reader is probably already familiar with. A vibrational mode ν , with an equilibrium frequency of $\tilde{\nu}_e$ (in wavenumbers) has energy levels (also in wavenumbers) given by $E_\nu = \tilde{\nu}_e(\nu + 1/2)$, where ν is the vibrational quantum number, and $\nu \geq 0$. The notation can become a bit confusing, so note that: ν (Greek letter nu) identifies the vibration, $\tilde{\nu}_e$ is the vibrational frequency (in wavenumbers), and ν (italic letter 'v') is the vibrational quantum number. It is trivial to extend this expression to a molecule with n uncoupled harmonic modes,

$$E_{\nu_1, \nu_2, \nu_3, \dots, \nu_n} = \sum_{i=1}^n \tilde{\nu}_i (\nu_i + 1/2) \quad (\text{B1.2.1})$$

where $\tilde{\nu}_i$ is the equilibrium vibrational frequency of the i th mode.

Of course, real molecules are not harmonic oscillators, and the energy level expression can be expanded in powers of $(v + 1/2)$. For a single mode we have

$$E_v = \tilde{\nu}_e(v + 1/2) - \tilde{x}_e \tilde{\nu}_e(v + 1/2)^2 + \dots$$

where \tilde{x}_e is the anharmonicity constant. This allows the spacing of the energy levels to decrease as a function of vibrational quantum number. Usually the expansion converges rapidly, and only the first two terms are needed. Harmonic oscillator and anharmonic oscillator potential energy curves with their respective energy levels are compared in [figure B1.2.3](#).

-5-

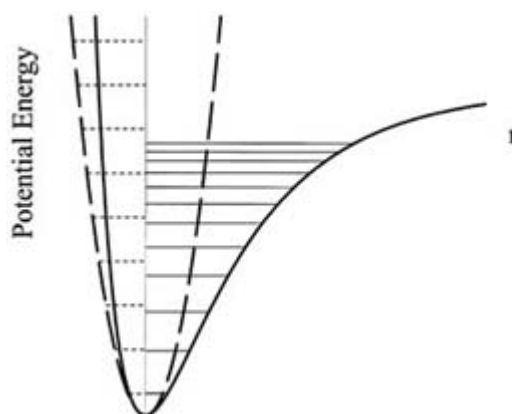


Figure B1.2.3. Comparison of the harmonic oscillator potential energy curve and energy levels (dashed lines) with those for an anharmonic oscillator. The harmonic oscillator is a fair representation of the true potential energy curve at the bottom of the well. Note that the energy levels become closer together with increasing vibrational energy for the anharmonic oscillator. The anharmonicity has been greatly exaggerated.

There usually is rotational motion accompanying the vibrational motion, and for a diatomic, the energy as a function of the rotational quantum number, J , is

$$E_J = B_e J(J + 1) - D_e [J(J + 1)]^2$$

where B_e and D_e are the equilibrium rotational constant and centrifugal distortion constant respectively. The rotational constant is related to the moment of inertia through $B_e = h/8\pi^2 I_e c$, where h is Planck's constant, I_e is the equilibrium moment of inertia, and c is the speed of light in vacuum. As the molecule rotates faster, it elongates due to centrifugal distortion. This increases the moment of inertia and causes the energy levels to become closer together. This is accounted for by including the second term with a negative sign. Overall, the vibration-rotation term energy is given by

$$E_{v,J} = \tilde{\nu}_e(v + 1/2) - \tilde{x}_e \tilde{\nu}_e(v + 1/2)^2 + B_e J(J + 1) - D_e [J(J + 1)]^2 - \alpha_e(v + 1/2)J(J + 1). \quad (\text{B1.2.2})$$

The only term in this expression that we have not already seen is α_e , the vibration-rotation coupling constant. It accounts for the fact that as the molecule vibrates, its bond length changes which in turn changes the moment of inertia. Equation B1.2.2 can be simplified by combining the vibration-rotation constant with the rotational constant, yielding a vibrational-level-dependent rotational constant,

$$B_v = B_e - \alpha_e(v + 1/2)$$

so the vibration–rotation term energy becomes

-6-

$$E_{v,J} = \tilde{\nu}_e(v + 1/2) - \tilde{x}_e \tilde{\nu}_e(v + 1/2)^2 + B_v J(J + 1) - D_e [J(J + 1)]^2.$$

B1.2.2.2 QUANTUM MECHANICAL DESCRIPTION

The quantum mechanical treatment of a harmonic oscillator is well known. Real vibrations are not harmonic, but the lowest few vibrational levels are often very well approximated as being harmonic, so that is a good place to start. The following description is similar to that found in many textbooks, such as McQuarrie (1983) [2]. The one-dimensional Schrödinger equation is

$$-\frac{\hbar^2}{2\mu} \frac{d^2\psi}{dx^2} + U(x)\psi(x) = E\psi(x) \quad (\text{B1.2.3})$$

where μ is the reduced mass, $U(x)$ is the potential energy, $\psi(x)$ is the wavefunction and E is the energy. The harmonic oscillator wavefunctions which solve this equation yielding the energies as given in [equation B1.2.1](#) are orthonormal, and are given by

$$\psi_v(x) = N_v H_v(\alpha^{1/2}x) e^{-\alpha x^2/2}$$

where $\alpha = (k\mu/\hbar^2)$, N_v are normalization constants given by

$$N_v = \frac{1}{(2^v v!)^{1/2}} \left(\frac{\alpha}{\pi}\right)^{1/4}$$

and H_v are the Hermite polynomials. The Hermite polynomials are defined as [1]

$$H_v(x) = (-1)^v e^{x^2} \frac{d^v}{dx^v} e^{-x^2}.$$

Upon inspection, the first three are seen to be

$$H_0(x) = 1 \quad H_1(x) = 2x \quad H_2(x) = 4x^2 - 2$$

and higher degree polynomials are obtained using the recursion relation

$$H_{v+1}(x) = 2xH_v(x) - 2vH_{v-1}(x). \quad (\text{B1.2.4})$$

-7-

The first few harmonic oscillator wavefunctions are plotted in figure B1.2.4.

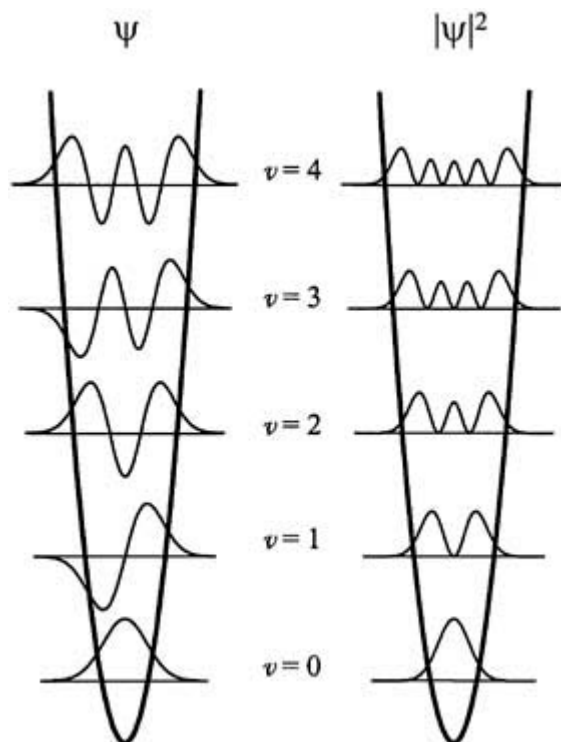


Figure B1.2.4. Lowest five harmonic oscillator wavefunctions ψ and probability densities $|\psi|^2$.

Upon solving the Schrödinger equation, the energy levels are $E_v = \tilde{\nu}_e(v + 1/2)$, where $\tilde{\nu}_e$ is related to the force constant k and reduced mass μ through

$$\tilde{\nu}_e = \frac{1}{2\pi c} \left(\frac{k}{\mu} \right)^{1/2} .$$

Since the reduced mass depends only on the masses of the atoms ($1/\mu = 1/m_1 + 1/m_2$ for a diatomic), which are known, measurement of the vibrational frequency allows the force constant to be determined.

The electric dipole selection rule for a harmonic oscillator is $\Delta v = \pm 1$. Because real molecules are not harmonic, transitions with $|\Delta v| > 1$ are weakly allowed, with $\Delta v = \pm 2$ being more allowed than $\Delta v = \pm 3$ and so on. There are other selection rules for quadrupole and magnetic dipole transitions, but those transitions are six to eight orders of magnitude weaker than electric dipole transitions, and we will therefore not concern ourselves with them.

The selection rules are derived through time-dependent perturbation theory [1, 2]. Two points will be made in the following material. First, the Bohr frequency condition states that the photon energy of absorption or emission is equal

to the energy level separation of the two states. Second, the importance of the transition dipole moment is shown and, furthermore, it is also shown that the transition dipole moment for a vibrational mode is in fact the change in dipole as a function of vibration, that is, the dipole derivative. The time-dependent Schrödinger equation is

$$\hat{H}\Psi = i\hbar \frac{\partial \Psi}{\partial t}. \quad (\text{B1.2.5})$$

The time- and coordinate-dependent wavefunction for any given state is

$$\Psi_v(x, t) = \psi_v(x) e^{-iE_v t/\hbar}$$

where $\psi_v(x)$ is a stationary state wavefunction obtained by solving the time-independent Schrödinger equation (B1.2.3). The Hamiltonian is broken down into two parts, $\hat{H} = \hat{H}_0 + \hat{H}^{(1)}$, where \hat{H}_0 is the Hamiltonian for the isolated molecule, and $\hat{H}^{(1)}$ describes the interaction Hamiltonian of the molecule with the electromagnetic field. In particular, the interaction energy between a dipole and a monochromatic field is

$$\hat{H}^{(1)} = -\mu E = -\mu E_0 \cos 2\pi \nu t.$$

Consider a two-state system, where

$$\Psi_1(x, t) = \psi_1(x) e^{-iE_1 t/\hbar} \quad \text{and} \quad \Psi_2(x, t) = \psi_2(x) e^{-iE_2 t/\hbar}.$$

These wavefunctions are orthogonal. Assume that the system is initially in state 1, and that the interaction begins at $t = 0$. Since there are only two states, at any later time the wavefunction for the system is

$$\Psi(t) = a_1(t)\Psi_1(t) + a_2(t)\Psi_2(t)$$

where the coefficients $a_1(t)$ and $a_2(t)$ are to be determined. We know that $a_1(0) = 1$ and $a_2(0) = 0$ from the initial conditions. By substitution into equation (B1.2.5), we have

$$a_1(t)\hat{H}^{(1)}\Psi_1 + a_2(t)\hat{H}^{(1)}\Psi_2 = i\hbar\Psi_1 \frac{da_1}{dt} + i\hbar\Psi_2 \frac{da_2}{dt}.$$

We can find the time-dependent coefficient for being in state 2 by multiplying from the left by ψ_2^* , and integrating over spatial coordinates:

-9-

$$a_1(t) \int \psi_2^* \hat{H}^{(1)} \Psi_1 d\tau + a_2(t) \int \psi_2^* \hat{H}^{(1)} \Psi_2 d\tau = i\hbar \frac{da_1}{dt} \int \psi_2^* \Psi_1 d\tau + i\hbar \frac{da_2}{dt} \int \psi_2^* \Psi_2 d\tau.$$

This expression can be simplified considerably since ψ_2 and ψ_1 are orthogonal which implies that Ψ_2 and Ψ_1 are orthogonal as well. Furthermore, $a_{11}(t) \approx a_1(0) = 1$ and $a_2(t) \approx a_2(0) = 0$ since $\hat{H}^{(1)}$ is a small perturbation:

(B1.2.6)

where we have used Dirac bracket notation for the integral, i.e.

$$\langle \psi_2 | \hat{H}^{(1)} | \psi_1 \rangle \equiv \int \psi_2^* \hat{H}^{(1)} \psi_1 d\tau.$$

In order to evaluate equation B1.2.6, we will consider the electric field to be in the z -direction, and express the interaction Hamiltonian as

$$\hat{H}^{(1)} = -\mu_z E_{0z} \cos 2\pi \nu t = -\frac{\mu_z E_{0z}}{2} (e^{i2\pi \nu t} + e^{-i2\pi \nu t}).$$

Before substituting everything back into equation B1.2.6, we define the *transition dipole moment* between states 1 and 2 to be the integral

$$(\mu_z)_{21} \equiv \langle \psi_2 | \mu_z | \psi_1 \rangle. \quad (\text{B1.2.7})$$

Now, we substitute it into equation B1.2.6 to get

$$\frac{da_2}{dt} \propto (\mu_z)_{21} E_{0z} \left\{ \exp \left[\frac{i(E_2 - E_1 + h\nu)t}{\hbar} \right] + \exp \left[\frac{i(E_2 - E_1 - h\nu)t}{\hbar} \right] \right\}$$

and then integrate from 0 to t to obtain

$$a_2(t) \propto (\mu_z)_{21} E_{0z} \left\{ \frac{1 - \exp[i(E_2 - E_1 + h\nu)t/\hbar]}{E_2 - E_1 + h\nu} + \frac{1 + \exp[i(E_2 - E_1 - h\nu)t/\hbar]}{E_2 - E_1 - h\nu} \right\}.$$

There are two important features of this result. The energy difference between states 1 and 2 is $\Delta E = E_2 - E_1$. When $\Delta E \approx h\nu$, the denominator of the second term becomes very small, and this term dominates. This is the well known

Bohr frequency condition. The second important feature is that $(\mu_z)_{21}$ must be nonzero for an allowed transition, which is how the selection rules are determined.

The molecular dipole moment (not the transition dipole moment) is given as a Taylor series expansion about the equilibrium position

$$\mu = \mu_0 + \left(\frac{d\mu}{dx} \right) x + \left(\frac{d^2\mu}{dx^2} \right) x^2 + \dots = \mu_0 + \mu_1 x + \mu_2 x^2 + \dots$$

where x is the displacement from equilibrium, μ_0 is the permanent dipole moment and μ_1 is the dipole derivative and so on. It is usually fine to truncate the expansion after the second term. Now we need to evaluate $(\mu_z)_{21}$ for the harmonic oscillator wavefunctions. Using [equation \(B1.2.7\)](#), we have

$$(\mu_z)_{j,i} = N_j N_i \int_{-\infty}^{\infty} H_{xj}(\alpha^{1/2}x) e^{-\alpha x^2/2} \mu H_{xi}(\alpha^{1/2}x) e^{-\alpha x^2/2} dx$$

which can be expanded by substituting $\mu = \mu_0 + \mu_1 x$:

$$\begin{aligned} (\mu_z)_{j,i} &= N_j N_i \mu_0 \int_{-\infty}^{\infty} H_{xj}(\alpha^{1/2}x) e^{-\alpha x^2/2} H_{xi}(\alpha^{1/2}x) e^{-\alpha x^2/2} dx \\ &+ N_j N_i \mu_1 \int_{-\infty}^{\infty} H_{xj}(\alpha^{1/2}x) e^{-\alpha x^2/2} x H_{xi}(\alpha^{1/2}x) e^{-\alpha x^2/2} dx. \end{aligned} \quad (\text{B1.2.8})$$

The first term is zero if $i \neq j$ due to the orthogonality of the Hermite polynomials. The recursion relation in [equation \(B1.2.4\)](#) is rearranged

$$x H_v(x) = v H_{v-1}(x) + \frac{1}{2} H_{v+1}(x)$$

and substituted into the second term in [equation \(B1.2.8\)](#),

$$(\mu_z)_{j,i} = \frac{N_j N_i}{\alpha} \mu_1 \int_{-\infty}^{\infty} H_{xj}(\xi) \left[xi H_{xi-1}(\xi) + \frac{1}{2} H_{xi+1}(\xi) \right] e^{-\xi^2} d\xi$$

where we have let $\xi = \alpha^{1/2}x$. Clearly, this integral will be nonzero only when $j = (i \pm 1)$, yielding the familiar $\Delta v = \pm 1$ harmonic oscillator selection rule. Furthermore, the overtone intensities for an anharmonic oscillator are obtained in a straightforward manner by determining the eigenfunctions of the energy levels in a harmonic oscillator basis set, and then summing the weighted contributions from the harmonic oscillator integrals.

-11-

B1.2.2.3 RAMAN SPECTROSCOPY

NORMAL RAMAN SPECTROSCOPY

Raman scattering has been discussed by many authors. As in the case of IR vibrational spectroscopy, the interaction is between the electromagnetic field and a dipole moment, however in this case the dipole moment is induced by the field itself. The induced dipole is $\mu_{\text{ind}} = \alpha E$, where α is the polarizability. It can be expressed in a Taylor series expansion in coordinate displacement

$$\alpha = \alpha_0 + \left(\frac{d\alpha}{dx} \right) x + \left(\frac{d^2\alpha}{dx^2} \right) x^2 + \dots = \alpha_0 + \alpha' x + \alpha'' x^2 + \dots$$

Here, α_0 is the static polarizability, α' is the change in polarizability as a function of the vibrational coordinate, α'' is the second derivative of the polarizability with respect to vibration and so on. As is usually the case, it is possible to truncate this series after the second term. As before, the electric field is $E = E_0 \cos 2\pi \nu_0 t$, where ν_0 is the frequency of the light field. Thus we have

$$\mu_{\text{ind}} = (\alpha_0 + \alpha' x) E_0 \cos 2\pi \nu_0 t. \quad (\text{B1.2.9})$$

The time dependence of the displacement coordinate for a mode undergoing harmonic oscillation is given by $x = x_m \cos 2\pi \nu_v t$, where x_m is the amplitude of vibration and ν_v is the vibrational frequency. Substitution into equation (B1.2.9) with use of Euler's half-angle formula yields

$$\begin{aligned}\mu_{\text{ind}} &= \alpha_0 E_0 \cos 2\pi \nu_0 t + \alpha' (x_m \cos 2\pi \nu_v t) E_0 \cos 2\pi \nu_0 t \\ &= \alpha_0 E_0 \cos 2\pi \nu_0 t + \frac{E_0 \alpha' x_m}{2} [\cos 2\pi (\nu_0 - \nu_v) t + \cos 2\pi (\nu_0 + \nu_v) t].\end{aligned}$$

The first term results in Rayleigh scattering which is at the same frequency as the exciting radiation. The second term describes Raman scattering. There will be scattered light at $(\nu_0 - \nu_v)$ and $(\nu_0 + \nu_v)$, that is at sum and difference frequencies of the excitation field and the vibrational frequency. Since $\alpha' x_m$ is about a factor of 10^6 smaller than α_0 , it is necessary to have a very efficient method for dispersing the scattered light.

The bands on the low frequency side of the excitation frequency $(\nu_0 - \nu_v)$ are referred to as the Stokes lines, consistent with the terminology used in fluorescence, whereas those on the high frequency side $(\nu_0 + \nu_v)$ are the anti-Stokes lines. It is a bit unfortunate that this terminology was chosen, since the Raman process is fundamentally different from fluorescence. In particular, fluorescence is the result of a molecule absorbing light, undergoing vibrational relaxation in the upper electronic state, and re-emitting a photon at a lower frequency. The timescale for fluorescence is typically of the order of nanoseconds. The Raman process, on the other hand, is an instantaneous scattering process that occurs on a femtosecond timescale. The photon is never absorbed by the molecule. It is usually clear whether fluorescence or Raman scattering is being observed, but there are situations where it is ambiguous. We shall not pursue the issue any further here, however.

-12-

It is well known that the intensity of scattered light varies as the fourth power of the frequency, and based on this alone one would predict the Stokes lines to be less intense than the anti-Stokes by a factor of

$$\frac{I_{\text{Stokes}}}{I_{\text{anti-Stokes}}} = \frac{(\nu_0 - \nu_v)^4}{(\nu_0 + \nu_v)^4}$$

which is 0.68 for a 1000 cm^{-1} vibration excited at 488 nm ($20\,492 \text{ cm}^{-1}$). In reality, the Stokes lines are *more* intense, typically by a factor of 2 to 10,000, as the vibrational frequency varies from 200 to 2000 cm^{-1} . This is easily justified when the Boltzmann population of initial states is taken into account. As seen in figure B1.2.5 the Stokes transitions correspond to the molecule being initially in a low energy state, usually $\nu = 0$, whereas it must be in an excited vibrational state if it is going to undergo an anti-Stokes process. The ratio of populations of two energy levels is given by

$$P_{12} = e^{-\Delta E_{12}/kT}$$

where ΔE_{12} is the energy difference between the two levels, k is Boltzmann's constant and T is the temperature in Kelvin. Since the Stokes lines are more intense than anti-Stokes, only the Stokes lines for a Raman spectrum are typically presented, and the abscissa is labelled with the frequency offset $(\nu_0 - \nu_v)$ rather than the actual frequency of scattered light.

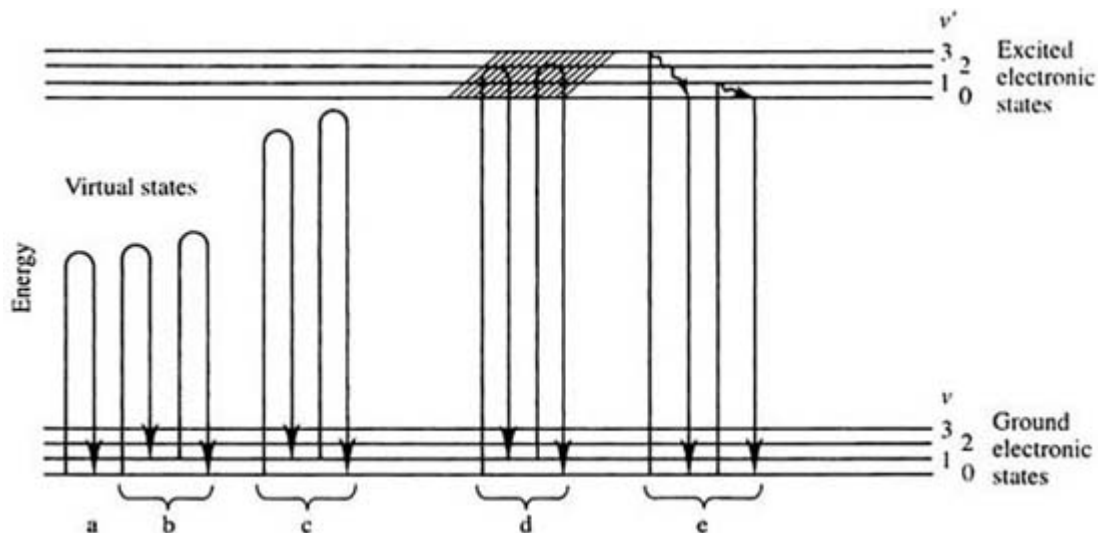


Figure B1.2.5. Comparison of several light scattering processes. (a) Rayleigh scattering, (b) Stokes and anti-Stokes Raman scattering, (c) pre-resonance Raman scattering, (d) resonance Raman scattering and (e) fluorescence where, unlike resonance Raman scattering, vibrational relaxation in the excited state takes place. From [3], used with permission.

-13-

Additional information about the vibration can be obtained through the depolarization ratio. This is the ratio of the intensity of scattered light that is polarized in a plane perpendicular to the incident radiation relative to that the scattered light that is polarized parallel to the incident polarization, $\rho = I_{\perp}/I_{\parallel}$. For totally symmetric modes, $\rho = 0$, while $0 < \rho < 3/4$ for non-totally symmetric modes [1, 3]. The polarization ratio can actually be greater than 3/4 for a resonantly enhanced Raman band [3].

Consistent with the notion that Raman scattering is due to a change in polarizability as a function of vibration, some of the general features of Raman spectroscopy [3] are:

- (1) it is more sensitive to stretching modes than bending modes, especially totally symmetric modes;
- (2) the intensity increases with bond order (i.e. double bond vibrations are more intense than single bond vibrations);
- (3) for modes involving only one bond, the intensity increases with increasing atomic number of the atoms;
- (4) for cyclic molecules, breathing modes are strongest.

RESONANCE RAMAN SPECTROSCOPY

Resonance Raman spectroscopy has been discussed by many authors [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 and 15]. If the excitation frequency is resonant with an excited electronic state, it is possible to dramatically increase the Raman cross-section. It is still a scattering process without absorption, but the light and polarizability can now couple much more efficiently. Resonant enhancements of 10^4 to 10^6 are achievable. A second important fact is that the vibrational modes that are coupled to the electronic transition are selectively enhanced, which greatly aids in structural determination. Another implication is that much lower concentrations can be used. A typical neat liquid might have a concentration of roughly 5 to 50 mol Γ^{-1} . It is possible for resonant enhanced Raman bands of a molecule to be as strong or stronger than the solvent bands even at a concentration of 10^{-6} M. This is useful because it is often desirable to maintain low enough concentrations such that the solute molecules do not interact with each other or aggregate. Finally, excited state vibrational dephasing rates can be determined.

The sum-over-states method for calculating the resonant enhancement begins with an expression for the resonance Raman intensity, $I_{i,f}$ for the transition from initial state i to final state f in the ground electronic state, and is given by [14]

$$I_{i,f} = \frac{2^7 \pi^5}{3^2 c^4} I_0 \sum_{\rho, \sigma} |(\alpha_{\rho\sigma})_{i,f}|^2 (\nu_L \pm \nu_v)^4$$

where I_0 is the incident intensity, ν_L is the laser frequency, ν_v is the vibrational frequency and $(\alpha_{\rho\sigma})_{i,f}$ is the $\rho\sigma$ th element of the Raman scattering tensor,

$$(\alpha_{\rho\sigma})_{i,f} = \sum_r \frac{(M_\rho)_{i,r} (M_\sigma)_{r,f}}{E_{r,i} - E_L + i\Gamma_r} + \frac{(M_\sigma)_{i,r} (M_\rho)_{r,f}}{E_{r,f} + E_L + i\Gamma_r}. \quad (\text{B1.2.10})$$

Here, $(M_\rho)_{a,b}$ is the ρ th component of the electronic transition moment from state a to b , $E_{r,i}$ and $E_{r,f}$ are the energy

-14-

differences between the resonant excited state and the initial and final states, respectively, $E_L - h\nu_L$ is the photon energy of the laser radiation, and Γ_r is the excited state vibrational dephasing rate. The Raman scattering tensor has two contributions, the so-called A and B terms:

The B -type enhancement is smaller than the A -type, and is usually, though not always, negligible. Therefore, we will concentrate on the A term [14],

where ρ and σ are pure electronic transition moments, and $\langle i | r \rangle$ and $\langle r | f \rangle$ are vibrational overlap integrals. In many cases, a single diagonal component of the scattering tensor dominates, which leads to a simplified expression,

(B1.2.11)

where $\langle e | M_z | g \rangle$ is the transition dipole moment for the electronic transition between ground state g and electronically excited state e , and r are the vibrational levels in the excited electronic state with vibrational energy $E_{i,r}$ relative to the initial vibrational level in the ground electronic state. From the form of this equation, we see that the enhancement depends on the overlap of the ground and excited state wavefunctions (Franck–Condon overlap), and is strongest when the laser frequency equals the energy level separation. Furthermore, systems with very large dephasing rates will not experience as much enhancement, all other factors being equal.

If there is only significant overlap with one excited vibrational state, equation (B1.2.11) simplifies further. In fact, if the initial vibrational state is $v_i = 0$, which is usually the case, and there is not significant distortion of the molecule in the excited electronic state, which may or may not hold true, then the intensity is given by

It is also possible to determine the resonant Raman intensities via a time-dependent method [16]. It has the

advantages that the vibrational eigenstates of the excited electronic state need not be known, and that it provides a physical picture of the dynamics in the excited state. Assuming only one ground vibrational state is occupied, the Raman cross section is [16, 17]

-15-

where ω_L is the laser frequency, ω_s is the frequency of the scattered light, (M_{ge}^0) is the electronic transition dipole, $\hbar\omega_i$ is the zero point vibrational energy of the initial state, $\langle f | i(t) \rangle$ is the overlap of the final state with the time-evolving state on the upper electronic state potential energy surface and the $\exp[-g(t)]$ term accounts for solvent-induced electronic dephasing. Unlike the sum-over-states method, the only excited state information needed is the potential energy surface in the immediate vicinity of where the initial state is projected onto the excited state.

B1.2.3 SPECTROMETERS

B1.2.3.1 INFRARED SPECTROMETERS

In the most general terms, an infrared spectrometer consists of a light source, a dispersing element, a sample compartment and a detector. Of course, there is tremendous variability depending on the application.

LIGHT SOURCES

Light sources can either be broadband, such as a Globar, a Nernst glower, an incandescent wire or mercury arc lamp; or they can be tunable, such as a laser or optical parametric oscillator (OPO). In the former case, a monochromator is needed to achieve spectral resolution. In the case of a tunable light source, the spectral resolution is determined by the linewidth of the source itself. In either case, the spectral coverage of the light source imposes limits on the vibrational frequencies that can be measured. Of course, limitations on the dispersing element and detector also affect the overall spectral response of the spectrometer.

Desirable characteristics of a broadband light source are stability, brightness and uniform intensity over as large a frequency range as possible. Desirable characteristics of a tunable light source are similar to those of a broadband light source. Furthermore, its wavelength should be as jitter-free as possible. Having a linewidth of 0.001 cm^{-1} is meaningless if the frequency fluctuates by 0.01 cm^{-1} on the timescale of scanning over an absorption feature.

The region $4500\text{--}2850 \text{ cm}^{-1}$ is covered nicely by f-centre lasers, and $2800\text{--}1000 \text{ cm}^{-1}$ by diode lasers (but there are gaps in coverage). Difference frequency crystals allow two visible beams whose frequencies differ by the wavelengths of interest to be mixed together. Using different laser combinations, coverage from greater than $5000\text{--}1000 \text{ cm}^{-1}$ has been demonstrated. The spectral range is limited only by the characteristics of the difference frequency crystal. The ultimate resolution of a laser spectrometer is dictated by the linewidth of the tunable light source. If the linewidth of the light source is broader than the absorption linewidth, then sensitivity is diminished, and the transition will appear broader than it actually is.

When extremely high resolution is not required, an attractive alternative is found in OPOs. An OPO is based on a nonlinear crystal that converts an input photon into two output photons whose energies add up to the input photon's energy. They can be rapidly tuned over a relatively large range, $5000\text{--}2200 \text{ cm}^{-1}$, depending on the nonlinear crystal. In the IR, commonly used crystals are LiNbO_3 and KDP (KH_2PO_4). The wavelengths of the two output beams are determined by the angle of the nonlinear crystal. That is, it will only

function when the index of refraction of the crystal in the direction of propagation is identical for all three beams. If narrow linewidths are required, it is necessary to seed the OPO with a weak beam from a diode laser. Thus, the parametric oscillation does not have to build up out of the noise, and is therefore more stable.

-16-

DISPERSING ELEMENTS

Dispersing elements must be used when broadband IR light sources are employed; either diffraction gratings or prisms can be used. Gratings are made by depositing a metal film, usually gold, on a ruled substrate. They can be used over a broad spectrum because the light never penetrates into the substrate. However, the reflectivity of the coating can be wavelength dependent. The useable range is determined by the pitch of the rulings. A grating suitable for use from 1000 to 3000 cm^{-1} would have 100–200 lines mm^{-1} . If there are too many lines per millimetre, then it acts like a mirror rather than a grating. If it has too few, then the efficiency is greatly reduced.

Prisms can also be used to disperse the light. They are much easier to make than gratings, and are therefore less expensive, but that is their only real advantage. Since the light has to pass through them, it is necessary to find materials that do not absorb. Greater dispersion is obtained by using materials with as high an index of refraction as possible. Unfortunately, materials with high index are also close to an absorption, which leads to a large nonlinear variation in index as a function of frequency.

When dispersing elements are used, the resolution of the spectrometer is determined by the entrance slit width, the exit slit width, the focal length and the dispersing element itself. Resolving power is defined as

$$R = \lambda / \Delta\lambda = \nu / \Delta\nu$$

that is, the central wavelength (or frequency) of the light exiting the spectrometer relative to its linewidth (or bandwidth). Higher resolving powers allow closely spaced lines to be distinguished.

DETECTORS

The detector chosen is just as important as the light source. If the sample is absorbing light, but the detector is not responding at that frequency, then changes in absorption will not be recorded. In fact, one of the primary limitations faced by spectroscopists working in the far-IR region of the spectrum (10–300 cm^{-1}) has been lack of highly sensitive detectors. The type of detector used depends on the frequency of the light. Of course, at any frequency, a bolometer can be used, assuming the detection element absorbs the wavelength of interest. A bolometer operates on the principle of a temperature-dependent resistance in the detector element. If the detector is held at a very low temperature, 2–4 K, and has a small heat capacity, then incoming energy will heat it, causing the resistance to change. While these detectors are general, they are susceptible to missing weak signals because they are swamped by thermal blackbody radiation, and are not as sensitive as detectors that have a response tied to the photon energy. At frequencies between 4000 and 1900 cm^{-1} , InSb photodiodes are used, HgCdTe detectors are favoured for frequencies between 2000 and 700 cm^{-1} , and copper-doped germanium (Cu:Ge) photoconductors are used for frequencies between 1000 and 300 cm^{-1} [18]. More recently, HgCdTe array detectors have become available that respond in the range 3500–900 cm^{-1} [19].

B1.2.3.2 RAMAN SPECTROMETERS

While Raman spectroscopy was first described in a paper by C V Raman and K S Krishnan in 1928, it has

only come into widespread use in the last three decades owing to the ready availability of intense monochromatic laser light

-17-

sources. Oddly enough, prior to 1945, Raman spectroscopy using Hg arc lamps was the method of choice for obtaining vibrational spectra since there was not yet widespread availability of IR spectrometers [5]. Just as with IR spectrometers, a Raman spectrometer consists of a light source, a dispersing element, a sample compartment and a detector. Again, there is tremendous variability depending on the application.

LIGHT SOURCES

The light source must be highly monochromatic so that the Raman scattering occurs at a well-defined frequency. An inhomogeneously broadened vibrational linewidth might be of the order of 20 cm^{-1} , therefore, if the excitation source has a wavelength of 500 nm, its linewidth must be less than 0.5 nm to ensure that it does not further broaden the line and decrease its intensity. This type of linewidth is trivial to achieve with a laser source, but more difficult with an arc lamp source.

Fixed frequency laser sources are most commonly used. For example, the Ar^+ laser has lines at 514, 496, 488, 476 and 458 nm. Sometimes a helium–neon laser is used (628 nm), or a doubled or tripled YAG (532 or 355 nm, respectively). Other wavelengths are generated by employing a Raman shifter with a variety of different gases. It is also desirable to have tunability, in order to carry out resonance Raman studies wherein one selectively measures the vibrations most strongly coupled to the electronic state being excited. Tunable lasers can be either line-tunable or continuously tunable. Thanks to the high sensitivity of photomultiplier tubes, the light source need only provide moderate power levels, $\sim 10\text{--}100 \text{ mW}$ for example. In fact, one must be careful not to use too much power and damage the sample.

DISPERSING ELEMENTS

Due to the rather stringent requirements placed on the monochromator, a double or triple monochromator is typically employed. Because the vibrational frequencies are only several hundred to several thousand cm^{-1} , and the linewidths are only tens of cm^{-1} , it is necessary to use a monochromator with reasonably high resolution. In addition to linewidth issues, it is necessary to suppress the very intense Rayleigh scattering. If a high resolution spectrum is not needed, however, then it is possible to use narrow-band interference filters to block the excitation line, and a low resolution monochromator to collect the spectrum. In fact, this is the approach taken with Fourier transform Raman spectrometers.

DETECTORS

Because the scattered light being detected is in the visible to near UV region, photomultiplier tubes (PMTs) are the detector of choice. By using a cooled PMT, background counts can be reduced to a level of only a few per second. For studies with higher signal levels, array detectors such as optical multichannel analysers (OMAs), or CCD arrays can be used. This allows a complete spectrum to be obtained without having to scan the monochromator.

RESOLUTION

The resolution of the Raman spectrum is determined by the monochromator. Furthermore, since the light being measured is in the visible region, usually around $20\,000 \text{ cm}^{-1}$, the resolution of the monochromator must be significantly better than that of its IR counterpart because the resolving power is described by $\Delta\nu/\nu$. That is, for

a 2000 cm^{-1} vibration a resolving power of 20,000 is needed to get the same resolution as that obtained in an IR spectrometer with a resolution of only 2000.

B1.2.3.3 FOURIER TRANSFORM TECHNIQUES

Fourier transform techniques do not change the underlying absorption or scattering mechanisms that couple light with matter. The data acquisition and processing are significantly different, however. In short, the difference is that the data are collected interferometrically in the time domain and then Fourier transformed into the frequency domain for subsequent analysis. These types of instruments are often used in experiments where broad spectral coverage with moderate sensitivity and frequency resolution is needed. This is often encountered when other aspects of the experiment are more difficult, such as surface studies. There is, however, ongoing research directed towards time-resolved FTIR with nanosecond time resolution [20, 21]. The basic requirements are a broadband light source, a beamsplitter, two delay lines (one fixed and one variable), a detector, and a computer to run the show and Fourier-transform the data.

The underlying concept behind an FTIR spectrometer can be understood when considering what happens when a beam of monochromatic light with wavelength λ is sent through a Michelson interferometer, shown schematically in [figure B1.2.6](#) [22]. If the pathlength difference between the two beams is zero, then they will constructively interfere, yielding an intensity at the detector of I_0 . Now consider what happens if the movable mirror is displaced by a distance $d = \lambda/4$. This will cause the optical path of that beam to change by an amount $\delta = \lambda/2$, and lead to destructive interference at the detector, with no power being measured. If the mirror is displaced by another $\lambda/4$, then the two beams will once again constructively interfere, leading to full intensity at the detector. The intensity as a function of mirror position is shown in [figure B1.2.7\(a\)](#). The intensity as a function of optical delay, δ , is described by

$$I(\delta) = \frac{I(\tilde{\nu})}{2} [1 + \cos(2\pi \tilde{\nu} \delta)]$$

where $\tilde{\nu}$ is the the frequency of interest and $I(\tilde{\nu})$ is its intensity. Similarly, [figure B1.2.7\(b\)](#) shows the intensity as a function of mirror position when two frequencies with equal amplitudes are present, and $f_1 = 1.2f_2$. [Figure B1.2.7\(c\)](#) depicts these same two frequencies, but with the amplitude of the lower frequency twice as large as that of the higher frequency. Finally, [figure B1.2.7\(d\)](#) shows the result for a Gaussian distribution of frequencies. For a discrete distribution of frequencies, the intensity as a function of optical delay is

$$I(\delta) = \frac{1}{2} \sum_{i=1}^n I(\tilde{\nu}_i) [1 + \cos(2\pi \tilde{\nu}_i \delta)]$$

and if there is a continuous distribution of frequencies, it is

$$I(\delta) = \frac{1}{2} \int_0^{\infty} I(\tilde{\nu}_i) [1 + \cos(2\pi \tilde{\nu}_i \delta)] d\tilde{\nu}.$$

While the data are collected in the time domain by scanning a delay line, they are most easily interpreted in the frequency domain. It is straightforward to connect the time and frequency domains through a Fourier transform

$$I(\tilde{\nu}) = 4 \int_0^{\infty} [I(\delta) - I(0)/2] \cos(2\pi \tilde{\nu} \delta) d\delta.$$

Two scans are required to obtain an absorption spectrum. First, a blank reference scan is taken that characterizes the broadband light source. Then a scan with the sample in place is recorded. The ratio of the sample power spectrum to the reference power spectrum is the transmission spectrum. If the source has stable output, then a single reference scan can be used with many sample scans.

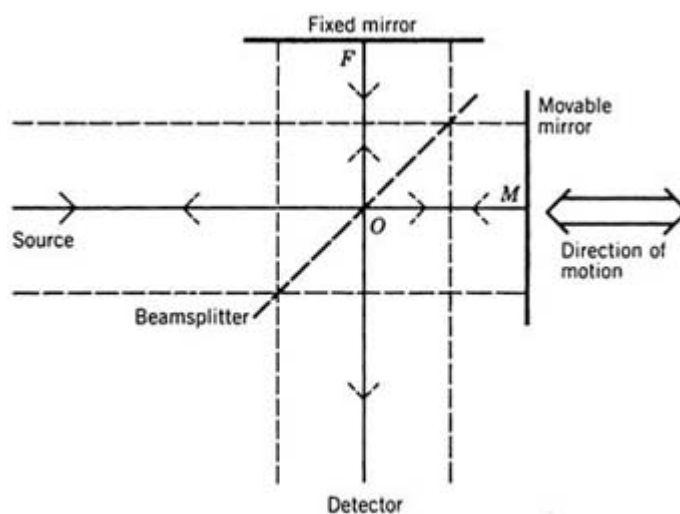


Figure B1.2.6. Schematic representation of a Michelson interferometer. From Griffiths P R and de Haseth J A 1986 Fourier transform infrared spectroscopy *Chemical Analysis* ed P J Elving and J D Winefordner (New York: Wiley). Reprinted by permission of John Wiley and Sons Inc.

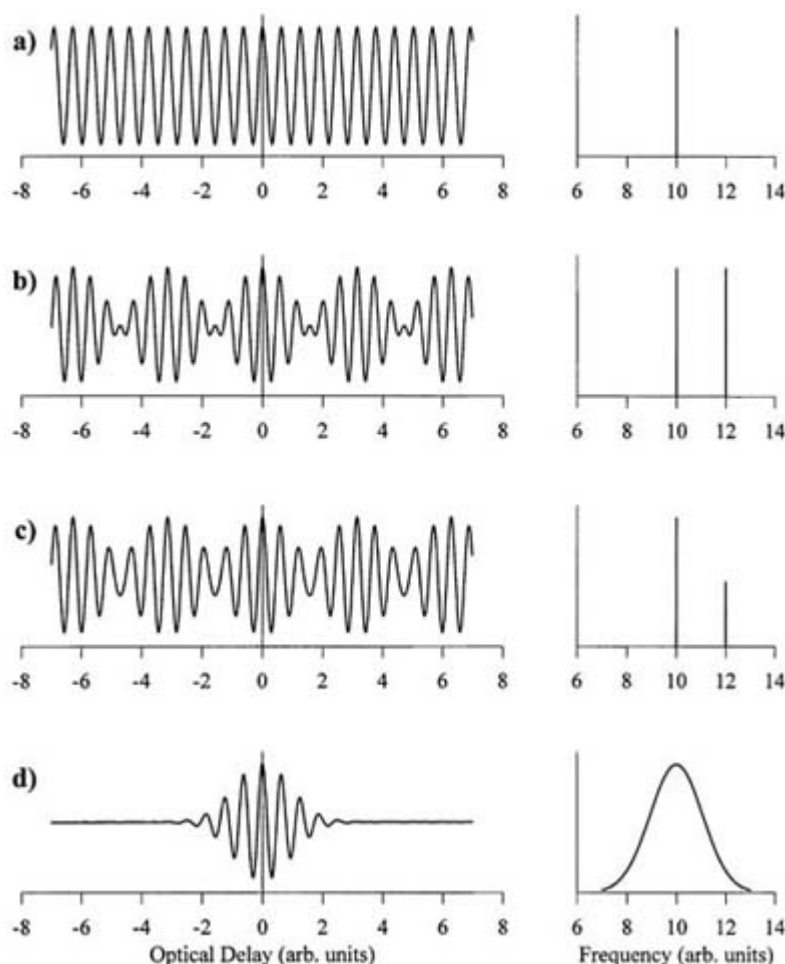


Figure B1.2.7. Time domain and frequency domain representations of several interferograms. (a) Single frequency, (b) two frequencies, one of which is 1.2 times greater than the other, (c) same as (b), except the high frequency component has only half the amplitude and (d) Gaussian distribution of frequencies.

The interferogram is obtained by continuously scanning the movable mirror and collecting the intensity on the detector at regular intervals. Thus, each point corresponds to a different mirror position, which in turn corresponds to a different optical delay. Several scans can be averaged together because the scanning mechanism is highly reproducible.

The spectral resolution of the Fourier transformed data is given by the wavelength corresponding to the maximum optical delay. We have $\lambda_{\max} = \delta_{\max}$, and therefore $\Delta\tilde{\nu} = 1/\delta_{\max}$. Therefore, if 0.1 cm^{-1} spectral resolution is desired, the optical delay must cover a distance of 10 cm. The high frequency limit for the transform is the Nyquist frequency which is determined by the wavelength that corresponds to two time steps: $\lambda_{\min} = 2\Delta\delta$, which leads to $\tilde{\nu}_{\max} = 1/(2\Delta\delta)$. The factor of two arises because a minimum of two data points per period are needed to sample a sinusoidal waveform. Naturally, the broadband light source will determine the actual content of the spectrum, but it is important that the step size be small enough to accommodate the highest frequency components of the source, otherwise they

will be folded into lower frequencies, which is known as ‘aliasing’. The step size for an FT-Raman instrument must be roughly ten times smaller than that in the IR, which is one of the reasons that FT-Raman studies are usually done with near-IR excitation.

There are several advantages of FT techniques [23]. One is the Jaquinot (or throughput) advantage. Since there are fewer optical elements and no slits, more power reaches the detector. Assuming that the noise source is detector noise, which is true in the IR, but not necessarily for visible and UV, the signal-to-noise (S/N) ratio will increase. A second advantage of FT techniques is that they benefit from the Fellgett (or multiplex) advantage. That is, when collecting the data, absorptions at all frequencies simultaneously contribute to the interferogram. This is contrasted with a grating spectrometer where the absorption is measured only at a single frequency at any given time. Theoretically, this results in an increase in the S/N ratio by a factor of

$$\sqrt{\frac{\tilde{\nu}_{\max}}{\Delta\tilde{\nu}}} = \sqrt{\frac{\delta_{\max}}{2\Delta\delta}}$$

This assumes that both spectra have the same resolution, and that it takes the same amount of time to collect the whole interferogram as is required to obtain one wavelength on the dispersive instrument (which is usually a reasonable assumption). Thus, $\tilde{\nu}_{\max}/\Delta\tilde{\nu}$ interferograms can be obtained and averaged together in the same amount of time it takes to scan the spectrum with the dispersive instrument. Since the S/N ratio scales with the square root of the number of scans averaged, the square root of this number is the actual increase in S/N ratio.

B1.2.4 TYPICAL EXAMPLES

There are thousands of scientists whose work can be classified as vibrational spectroscopy. The following examples are meant to show the breadth of the field, but cannot be expected to constitute a complete representation of all the fields where vibrational spectroscopy is important.

B1.2.4.1 LASER IR

For the highest resolution and sensitivity, laser-based spectrometers must be used. These have the advantage that the resolution depends on the linewidth of the laser, rather than the monochromator. Furthermore, at any given moment, all of the power is at the frequency of interest, rather than being spread out over the whole IR spectrum. Due to the fact that the emission from any given laser typically has only 100–1500 cm^{-1} of tunability, and there can be difficulty maintaining narrow linewidths, the spectral coverage can be limited.

High resolution spectroscopic measurements in the gas phase yield the most detailed structural information possible. For example, measurements of weakly-bound complexes in the far-IR [24, 25] and IR [26, 27] have provided the most exact information on their structure and steady-state dynamics. Of course, a much higher level of theory must be used than was presented in [section B1.2.2.1](#). Quite often the modes are so strongly coupled that all vibrational degrees of freedom must be treated simultaneously. Coriolis coupling and symmetry-allowed interactions among bands, i.e. Fermi resonances, are also quite significant, and must be treated explicitly. Direct measurement of the low frequency van der

Waals modes in weakly-bound complexes has been discussed in [section B1.4](#), *Microwave and Terahertz Spectroscopy*, and will not be repeated here.

A recent review of high-resolution, direct IR laser absorption spectroscopy in supersonic slit jets provides a prototypical example [26]. Figure B1.2.8 displays the experimental set-up. There are three different IR sources employed. The first utilizes difference frequency mixing of a single mode tunable ring dye laser with a single mode Ar^+ laser in a LiNbO_3 crystal to obtain light in the 2–4 μm region ($5000\text{--}2500\text{ cm}^{-1}$) with a frequency stability of 2 MHz ($6.7\times 10^{-5}\text{ cm}^{-1}$). The second uses cryogenically cooled, single mode, lead-salt

diodes to cover 4–10 μm ($2500\text{--}1000\text{ cm}^{-1}$) with a frequency resolution of 15 MHz ($5 \times 10^{-4}\text{ cm}^{-1}$). The third is a difference frequency scheme between a single mode dye laser and a single mode Nd:YAG laser which accesses wavelengths below 2 μm (frequencies greater than 5000 cm^{-1}). A long pathlength is achieved by combining a slit jet expansion with a multipass cell. This enhances the sensitivity by over two orders of magnitude.

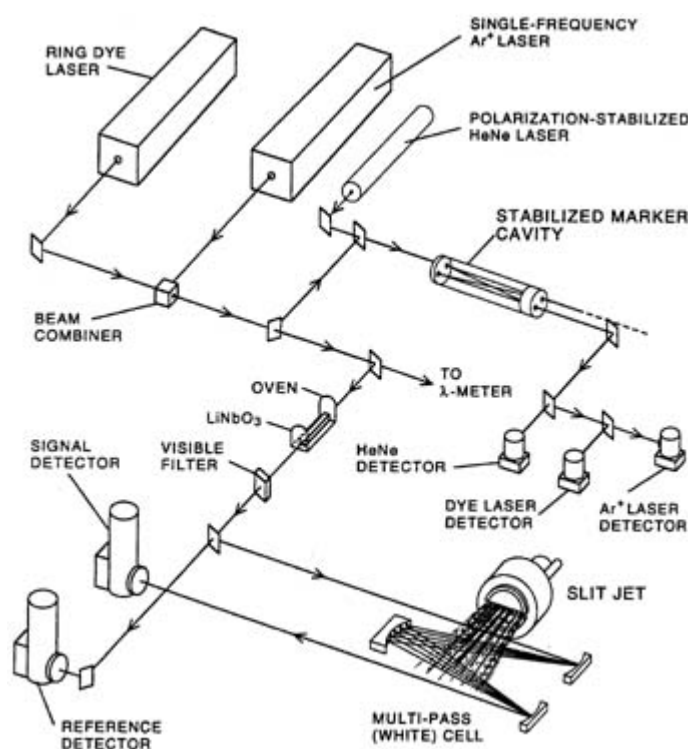


Figure B1.2.8. Schematic diagram of a slit jet spectrometer for high resolution IR studies of weakly-bound species. From [26], used with permission.

One of the systems studied is the Ar–HF van der Waals (vdW) dimer. This rare gas–molecule complex has provided a wealth of information regarding intermolecular interactions. The vdW complex can be thought of in terms of the monomer subunit being perturbed through its interaction with the rare gas. The rotational motion is hindered, and for certain modes can be thought of as extremely large amplitude bending motion. Furthermore, there is low frequency intermolecular stretching motion. In a weakly-bound complex, the bends, stretches and internal rotations are all coupled with each other. The large spectral coverage has allowed for the measurement of spectra of Ar–HF with $\nu_{\text{HF}} = 1$ and 2. By combining these measurements with those made in the far-IR [28], intermolecular potential energy

surfaces for $\nu_{\text{HF}} = 0, 1$ and 2 have been determined. Quite small amounts of internal energy ($\sim 150\text{ cm}^{-1}$) allow the complex to fully sample the angular degree of freedom. Interestingly, the barrier to internal rotation is about the same for $\nu_{\text{HF}} = 0$ or 1, but significantly larger when $\nu_{\text{HF}} = 2$.

In addition to the dependence of the intermolecular potential energy surface on monomer vibrational level, the red-shifting of the monomer absorption as a function of the number of rare gas atoms in the cluster has been studied. The band origin for the $\nu_{\text{HF}} = 1 \leftarrow 0$ vibration in a series of clusters $\text{Ar}_n\text{--HF}$, with $0 < n < 5$, was measured and compared to the HF vibrational frequency in an Ar matrix ($n = \infty$). The monomer vibrational frequency ν_{HF} red shifts monotonically, but highly nonlinearly, towards the matrix value as sequential Ar atoms are added. Indeed, roughly 50% of the shift is already accounted for by $n = 3$.

CAVITY RINGDOWN SPECTROSCOPY

The relatively new technique of cavity ringdown laser absorption spectroscopy, or CRLAS [29, 30, 31 and 32], has proven to be exceptionally sensitive in the visible region of the spectrum. Recently, it has been used in the IR to measure O–H and O–D vibrations in weakly-bound water and methanol clusters [33]. The concept of cavity ringdown is quite straightforward. If a pulse of light is injected into a cavity composed of two very highly reflective mirrors, only a very small fraction will escape upon reflection from either mirror. If mirrors with 99.996% reflectivity are used, then the photons can complete up to 15 000 round trip passes, depending on other loss factors in the cavity. This makes the effective pathlength up to 30,000 longer than the physical pathlength of the sample. Currently, dielectric coatings for IR wavelengths are not as efficient as those for the visible, and the highest reflectivity available is about 99.9–99.99%, which leads to sensitivity enhancements of several hundred to several thousand. It is best to use pulses with a coherence length that is less than the cavity dimensions in order to avoid destructive interference and cancellation of certain frequencies.

The light leaking out of the cavity will decay exponentially, with a time constant that reflects the round-trip losses. When an absorbing sample is placed in the cavity, there are additional losses and the exponential time constant will become shorter. More highly absorbing samples will affect the time constant to a larger extent, and the absolute absorption is determined. The experiment is shown schematically in figure B1.2.9. One of the most important attributes of CRLAS is that it is relatively insensitive to laser pulse intensity fluctuations since the ringdown time constant, not the transmitted intensity, is measured.

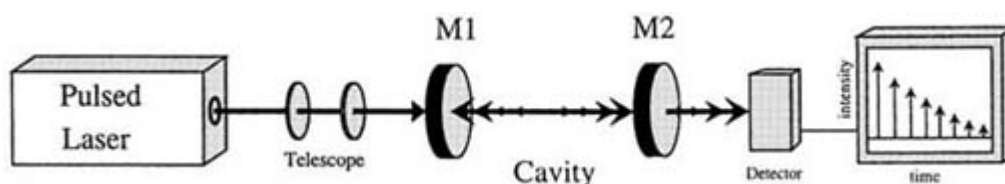


Figure B1.2.9. Schematic representation of the method used in cavity ringdown laser absorption spectroscopy. From [33], used with permission.

To date, the IR-CRLAS studies have concentrated on water clusters (both H₂O and D₂O), and methanol clusters. Most importantly, these studies have shown that it is in fact possible to carry out CRLAS in the IR. In one study, water cluster concentrations in the molecular beam source under a variety of expansion conditions were characterized [34]. In a second study OD stretching bands in (D₂O)_n clusters were measured [35]. These bands occur between 2300

-24-

and 2800 cm⁻¹, a spectral region that has been largely inaccessible with other techniques. Cooperative effects in the hydrogen-bonded network were measured, as manifested in red shifts of OD stretches for clusters up to (D₂O)₈. These data for additional isotopes of water are necessary to fully characterize the intermolecular interactions.

For methanol clusters [36], it was found that the dimer is linear, while clusters of 3 and 4 molecules exist as monocyclic ring structures. There also is evidence that there are two cyclic ring trimer conformers in the molecular beam.

B1.2.4.2 RESONANCE RAMAN SPECTROSCOPY

The advantages of resonance Raman spectroscopy have already been discussed in section B1.2.2.3. For these reasons it is rapidly becoming the method of choice for studying large molecules in solution. Here we will present one study that exemplifies its attributes. There are two complementary methods for studying proteins.

First, it is possible to excite a chromophore corresponding to the active site, and determine which modes interact with it. Second, by using UV excitation, the amino acids with phenyl rings (tryptophan and tyrosine, and a small contribution from phenylalanine) can be selectively excited [4]. The frequency shifts in the resonance Raman spectrum associated with them provide information on their environment.

There has been extensive work done on myoglobin, haemoglobin, Cytochrome-*c*, rhodopsin and bacteriorhodopsin. In fact, there are literally hundreds of articles on each of the above subjects. Here we will consider haemoglobin [12]. The first three of these examples are based on the protohaeme unit, shown in figure B1.2.10.

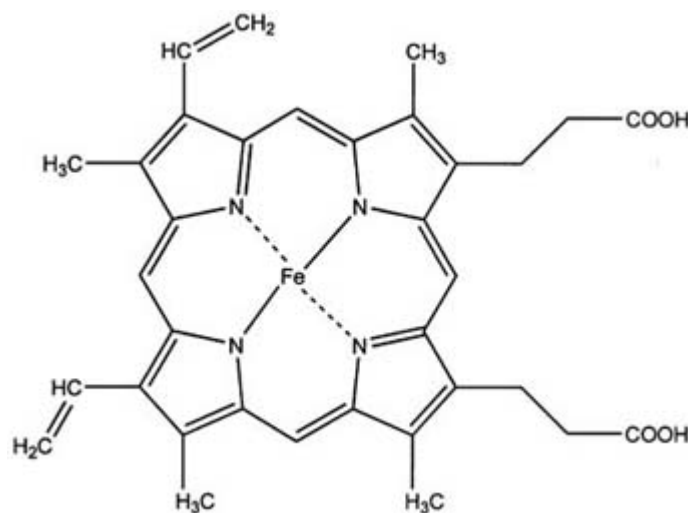


Figure B1.2.10. Structure of the protohaeme unit found in haemoglobin and myoglobin.

Haemoglobin is made up of four protohaeme subunits, two are designated α and two are designated β . The protohaeme unit with surrounding protein is shown in figure B1.2.11. The binding curve of O₂ suggests a two state model, where haemoglobin binds O₂ more strongly when the concentration is higher. Thus, it will strongly bind O₂ in the lungs where the concentration is high, and then release it in tissues where the O₂ concentration is low. The high affinity and low affinity states are referred to as the R state and T state, respectively [37]. The R and T states each

have a distinct structure, and have been characterized crystallographically. Therefore, there are three primary issues:

- (1) What is the behaviour of the haeme macrocycles?
- (2) What is the interaction between the iron and the nearby histidines?
- (3) What are the structural dynamics of the tetramer as a whole? The haemoglobin bound to O₂ (HbO₂) is not photoactive, so the CO adduct, HbCO, is used instead.

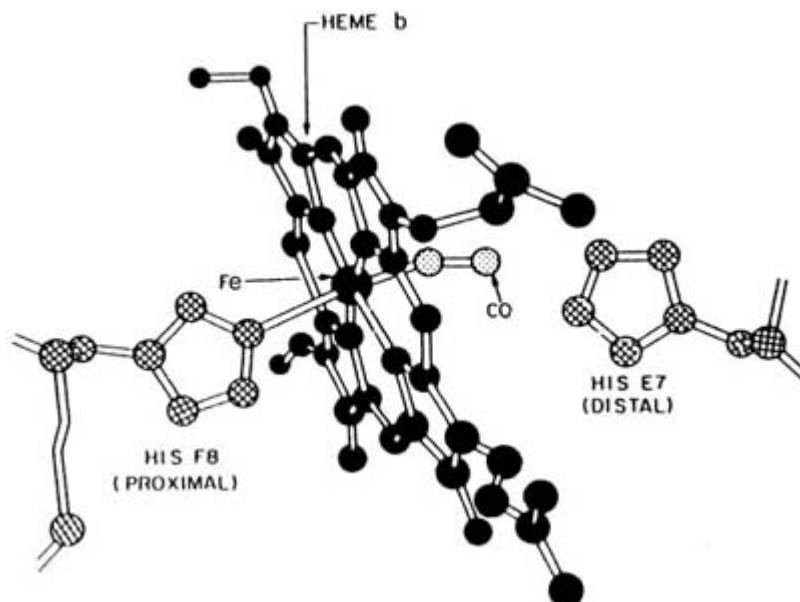


Figure B1.2.11. Biologically active centre in myoglobin or one of the subunits of haemoglobin. The bound CO molecule as well as the proximal and distal histidines are shown in addition to the protohaeme unit. From Rousseau D L and Friedman J M 1988 *Biological Applications of Raman Spectroscopy* vol 3, ed T G Spiro (New York: Wiley). Reprinted by permission of John Wiley and Sons Inc.

Information about the haeme macrocycle modes is obtained by comparing the resonance Raman spectra of deoxyHb with HbCO. The d–d transitions of the metal are too weak to produce large enhancement, so the Soret band of the macrocycle, a π to π^* transition, is excited instead. It has been found that the vinyl groups do not participate in the conjugated system [4]. This is based on the fact that the vinyl C=C stretch does not exhibit resonant enhancement when the π to π^* transition is excited. On the other hand, it is found that both totally symmetric and nontotally symmetric modes of the macrocycle are all at a lower frequency in the HbCO photoproduct spectra relative to deoxyHb. This is interpreted to mean that the photoproduct has a slightly expanded core relative to the deoxy structure [12, 13]. Given that there is a structural change in the haeme centre, it is expected that the interaction of the iron with the proximal histidine should also be affected.

The Fe–N^{His} mode is at 222 cm⁻¹ in the R state and 207 cm⁻¹ in the T state for the α subunits, but only shifted to 218 cm⁻¹ T state for the β subunits. This is consistent with the interpretation that the Fe–imidazole interactions are weakened more in the T state of the α subunits than β subunits. Time-resolved resonance Raman studies have shown that the R \rightarrow T switch is complete on a 10 μ s timescale [38]. Finally, UV excitation of the aromatic protein side chains yields

-26-

frequency shifts that indicate a change in the quaternary structure, and this occurs on the same timescale as the frequency shifts in the Fe–N^{His} modes.

B1.2.4.3 TIME-RESOLVED SPECTROSCOPY

Time-resolved spectroscopy has become an important field from x-rays to the far-IR. Both IR and Raman spectroscopies have been adapted to time-resolved studies. There have been a large number of studies using time-resolved Raman [39], time-resolved resonance Raman [7] and higher order two-dimensional Raman spectroscopy (which can provide coupling information analogous to two-dimensional NMR studies) [40]. Time-resolved IR has probed neutrals and ions in solution [41, 42], gas phase kinetics [43] and vibrational dynamics of molecules chemisorbed and physisorbed to surfaces [44]. Since vibrational frequencies are very sensitive to the chemical environment, pump-probe studies with IR probe pulses allow structural changes to

be monitored as a reaction proceeds.

As an illustrative example, consider the vibrational energy relaxation of the cyanide ion in water [45]. The mechanisms for relaxation are particularly difficult to assess when the solute is strongly coupled to the solvent, and the solvent itself is an associating liquid. Therefore, precise experimental measurements are extremely useful. By using a diatomic solute molecule, this system is free from complications due to coupling of vibrational modes in polyatomics. Furthermore, the relatively low frequency stretch of roughly 2000 cm^{-1} couples strongly to the internal modes of the solvent.

Infrared pulses of 200 fs duration with 150 cm^{-1} of bandwidth centred at 2000 cm^{-1} were used in this study. They were generated in a two-step procedure [46]. First, a $\beta\text{-BaB}_2\text{O}_4$ (BBO) OPO was used to convert the 800 nm photons from the Ti:sapphire amplifier system into signal and idler beams at 1379 and 1905 nm, respectively. These two pulses were sent through a difference frequency crystal (AgGaS_2) to yield pulses centred at 2000 cm^{-1} . A 32-element array detector was used to simultaneously detect the entire bandwidth of the pulse [45].

Two isotopic combinations of CN^- ($^{13}\text{C}^{15}\text{N}^-$ which has a stretching frequency of 2004 cm^{-1} and $^{12}\text{C}^{14}\text{N}^-$ with a stretching frequency of 2079 cm^{-1}) in both H_2O and D_2O yield a range of relaxation times. In particular, it is found that the vibrational relaxation time decreases from 120 to 71 ps in D_2O as the vibrational frequency increases from 2004 to 2079 cm^{-1} . However, in H_2O , the relaxation time is roughly 30 ps for both isotopomers. The vibrational relaxation rate is highly correlated to the IR absorption cross section of the solvent at the frequency of solute vibration, which indicates that Coulombic interactions have a dominant role in the vibrational relaxation of CN^- .

B1.2.4.4 ACTION SPECTROSCOPY

The term ‘action spectroscopy’ refers to those techniques that do not directly measure the absorption, but rather the consequence of photoabsorption. That is, there is some measurable change associated with the absorption process. There are several well known examples, such as photoionization spectroscopy [47], multi-photon ionization spectroscopy [48], photoacoustic spectroscopy [49], photoelectron spectroscopy [50, 51], vibrational predissociation spectroscopy [52] and optothermal spectroscopy [53, 54]. These techniques have all been applied to vibrational spectroscopy, but only the last one will be discussed here.

Optothermal spectroscopy is a bolometric method that monitors the energy in a stream of molecules rather than in the light beam. A well collimated molecular beam is directed toward a liquid helium cooled bolometer. There will be energy

deposited in the bolometer from the translational kinetic energy of the molecules as well as any internal energy they may have. A narrow linewidth (2 MHz) infrared laser illuminates the molecular beam, typically in a multipass geometry. As the laser frequency is scanned the molecules will absorb energy when the frequency corresponds to a transition frequency. At that point, the energy deposited in the bolometer will change.

The optothermal spectrum will faithfully represent the absorption spectrum provided that the molecules do not fluoresce prior to arrival at the detector. Fluorescence is not a problem because infrared fluorescence lifetimes are of the order of milliseconds. The transit time from the laser-molecular beam interaction region to the detector is tens of μs . Furthermore, it is possible to determine if the absorbing species is a stable monomer, or weakly bound cluster. When a stable monomer absorbs a photon, the amount of energy measured by the bolometer increases. When a weakly-bound species absorbs a photon greater than the dissociation energy, vibrational predissociation will take place and the dissociating fragments will not hit the bolometer element,

leading to a decrease in energy measured by the bolometer. It is also possible to place the bolometer off-axis from the collimated molecular beam so that only dissociating molecules will register a signal.

A nice example of this technique is the determination of vibrational predissociation lifetimes of $(\text{HF})_2$ [55]. The HF dimer has a nonlinear hydrogen bonded structure, with nonequivalent HF subunits. There is one ‘free’ HF stretch (ν_1), and one ‘bound’ HF stretch (ν_2), which rapidly interconvert. The vibrational predissociation lifetime was measured to be 24 ns when exciting the free HF stretch, but only 1 ns when exciting the bound HF stretch. This makes sense, as one would expect the bound HF vibration to be most strongly coupled to the weak intermolecular bond.

B1.2.4.5 MICROSCOPY

It is possible to incorporate a Raman or IR spectrometer within a confocal microscope. This allows the spatial resolution of the microscope and compound identification of vibrational spectroscopy to be realized simultaneously. One of the reasons that this is a relatively new development is because of the tremendous volume of data generated. For example, if a Raman microscope has roughly $1 \mu\text{m}$ spatial resolution, and an area of $100 \mu\text{m} \times 100 \mu\text{m}$ is to be imaged, and the frequency region from 800 cm^{-1} – 3400 cm^{-1} is covered with 4 cm^{-1} spectral resolution, then the data set has 6 million elements. Assuming each value is represented with a 4 byte number, the image would require 24 M Bytes of storage space. While this is not a problem for current computers, the capacity of a typical hard drive on a PC from around 1985 (IBM 8088) was only 20 M Byte. Also, rapid data transfer is needed to archive and retrieve images. Furthermore, in order to obtain the spectrum at any spatial position, array detectors (or FT methods) are required. A representative experimental set-up is shown in figure B1.2.12 [3].

-28-

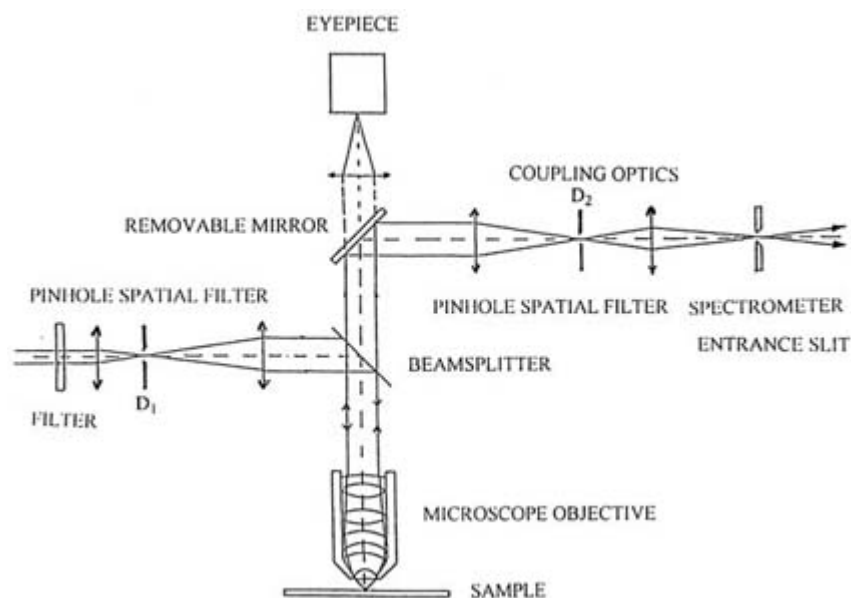


Figure B1.2.12. Schematic diagram of apparatus for confocal Raman microscopy. From [3], used with permission.

Raman microscopy is more developed than its IR counterpart. There are several reasons for this. First, the diffraction limit for focusing a visible beam is about 10 times smaller than an IR beam. Second, Raman spectroscopy can be done in a backscattering geometry, whereas IR is best done in transmission. A microscope is most easily adapted to a backscattering geometry, but it is possible to do it in transmission.

Raman microscopy is particularly adept at providing information on heterogeneous samples, where a

conventional spectrometer would average over many domains. It has found applications in materials science and catalysis; earth, planetary and environmental sciences; biological and medical sciences; and even in art history and forensic science [3]. For example, consider a hypothetical situation where someone owned what they believed to be a mediæ val manuscript from the 12th century. One could easily identify a small region of green colour to non-destructively analyse using Raman microscopy. If it happens that the dye is identified as phthalocyanine green, which was not discovered until 1938, then one can be certain that the manuscript is not authentic, or that it has undergone restoration relatively recently.

B1.2.5 CONCLUSIONS AND FUTURE PROSPECTS

Vibrational spectroscopy has been, and will continue to be, one of the most important techniques in physical chemistry. In fact, the vibrational absorption of a *single* acetylene molecule on a Cu(100) surface was recently reported [56]. Its endurance is due to the fact that it provides detailed information on structure, dynamics and environment. It is employed in a wide variety of circumstances, from routine analytical applications, to identifying novel (often transient) species, to providing some of the most important data for advancing the understanding of intramolecular and intermolecular interactions.

REFERENCES

- [1] Califano S 1976 *Vibrational States* (New York: Wiley)
- [2] McQuarrie D A 1983 *Quantum Chemistry* (Mill Valley, CA: University Science Books)
- [3] Turrell G and Corset J (eds) 1996 *Raman Microscopy: Developments and Applications* (New York: Academic)
- [4] Asher S A 1993 UV resonance Raman-spectroscopy for analytical, physical and biophysical chemistry 2 *Anal. Chem.* **65** A 201–10
- [5] Asher S A 1993 UV resonance Raman-spectroscopy for analytical, physical and biophysical chemistry 1 *Anal. Chem.* **65** A 59–66
- [6] Asher S A and Chi Z H 1998 UV resonance Raman studies of protein folding in myoglobin and other proteins *Biophys. J.* **74** A29
- [7] Bell S E J 1996 Time-resolved resonance Raman spectroscopy *Analyst* **121** R107–20
- [8] Biswas N and Umapathy S 1998 Resonance Raman spectroscopy and ultrafast chemical dynamics *Curr. Sci.* **74** 328–40
- [9] Chi Z H, Chen X G, Holtz J S W and Asher S A 1998 UV resonance Raman-selective amide vibrational enhancement: quantitative methodology for determining protein secondary structure *Biochemistry* **37** 2854–64
- [10] Hoskins L C 1984 Resonance Raman-spectroscopy of beta-carotene and lycopene—a physical-chemistry experiment *J. Chem. Educ.* **61** 460–2
- [11] Johnson B R, Kittrell C, Kelly P B and Kinsey J L 1996 Resonance Raman spectroscopy of dissociative polyatomic molecules *J. Chem. Educ.* **100** 7743–64
- [12] Kincaid J R 1995 Structure and dynamics of transient species using time-resolved resonance Raman-spectroscopy *Biochemical Spectroscopy Methods Enzymol.* vol 246, ed K Sauer (San Diego, CA: Academic) pp 460–501
- [13] Spiro T G and Czernuszewicz R S 1995 Resonance Raman-spectroscopy of metalloproteins *Biochemical Spectroscopy Methods Enzymol.* vol 246, ed K Sauer (San Diego, CA: Academic) pp 416–60
- [14] Strommen D P and Nakamoto K 1977 Resonance Raman-spectroscopy *J. Chem. Educ.* **54** 474–8
- [15] Mathies R A 1995 Biomolecular vibrational spectroscopy *Biochemical Spectroscopy Methods Enzymol.* vol 246, ed K Sauer (San Diego, CA: Academic) pp 377–89
- [16] Heller E J, Sundberg R L and Tannor D 1982 Simple aspects of Raman scattering *J. Phys. Chem.* **86** 1822–33
- [17] Zhong Y and McHale J L 1997 Resonance Raman study of solvent dynamics in electron transfer. II. Betaine-30 in

- [18] Gruebele M H W 1988 *Infrared Laser Spectroscopy of Molecular Ions and Clusters* (Berkeley: University of California)
- [19] Kidder L H, Levin I W, Lewis E N, Kleiman V D and Heilweil E J 1997 Mercury cadmium telluride focal-plane array detection for mid-infrared Fourier-transform spectroscopic imaging *Opt. Lett.* **22** 742–4
- [20] Pibel C D, Sirota E, Brenner J and Dai H L 1998 Nanosecond time-resolved FTIR emission spectroscopy: monitoring the energy distribution of highly vibrationally excited molecules during collisional deactivation *J. Chem. Phys.* **108** 1297–300
- [21] Leone S R 1989 Time-resolved FTIR emission studies of molecular photofragmentation *Accounts Chem. Res.* **22** 139–44
- [22] Elving P J and Winefordner J D (eds) 1986 *Fourier Transform Infrared Spectroscopy* (New York: Wiley)
- [23] Skoog D A, Holler F J and Nieman T A 1998 *Principles of Instrumental Analysis* 5th edn (Philadelphia: Harcourt Brace)
-

- [24] Saykally R J and Blake G A 1993 Molecular-interactions and hydrogen-bond tunneling dynamics—some new perspectives *Science* **259** 1570–5
- [25] Liu K, Brown M G and Saykally R J 1997 Terahertz laser vibration rotation tunneling spectroscopy and dipole moment of a cage form of the water hexamer *J. Phys. Chem. A* **101** 8995–9010
- [26] Nesbitt D J 1994 High-resolution, direct infrared-laser absorption-spectroscopy in slit supersonic jets—intermolecular forces and unimolecular vibrational dynamics in clusters *Ann. Rev. Phys. Chem.* **45** 367–99
- [27] Bacic Z and Miller R E 1996 Molecular clusters: structure and dynamics of weakly bound systems *J. Phys. Chem.* **100** 12,945–59
- [28] Dvorak M A, Reeve S W, Burns W A, Grushow A and Leopold K R 1991 Observation of three intermolecular vibrational-states of Ar-HF *Chem. Phys. Lett.* **185** 399–402
- [29] O’Keefe A and Deacon D A G 1988 Cavity ring-down optical spectrometer for absorption-measurements using pulsed laser sources *Rev. Sci. Instrum.* **59** 2544–51
- [30] Scherer J J, Paul J B, O’Keefe A and Saykally R J 1997 Cavity ringdown laser absorption spectroscopy: history, development, and application to pulsed molecular beams *Chem. Rev.* **97** 25–51
- [31] Zalicki P and Zare R N 1995 Cavity ring-down spectroscopy for quantitative absorption-measurements *J. Chem. Phys.* **102** 2708–17
- [32] Lehmann K K and Romanini D 1996 The superposition principle and cavity ring-down spectroscopy *J. Chem. Phys.* **105** 10,263–77
- [33] Scherer J J *et al* 1995 Infrared cavity ringdown laser-absorption spectroscopy (IR-CRLAS) *Chem. Phys. Lett.* **245** 273–80
- [34] Paul J B, Collier C P, Saykally R J, Scherer J J and O’Keefe A 1997 Direct measurement of water cluster concentrations by infrared cavity ringdown laser absorption spectroscopy *J. Phys. Chem. A* **101** 5211–14
- [35] Paul J B, Provencal R A and Saykally R J 1998 Characterization of the (D₂O)₂ hydrogen-bond-acceptor antisymmetric stretch by IR cavity ringdown laser absorption spectroscopy *J. Chem. Phys. A* **102** 3279–83
- [36] Provencal R A *et al* 1999 Infrared cavity ringdown spectroscopy of methanol clusters: single donor hydrogen bonding *J. Chem. Phys.* **110** 4258–67
- [37] Monod J, Wyman J and Changeux J P 1965 On the nature of allosteric transitions: a plausible model *J. Mol. Biol.* **12** 88–118
- [38] Scott T W and Friedman J M 1984 Tertiary-structure relaxation in haemoglobin—a transient Raman-study *J. Am. Chem. Soc.* **106** 5677–87
- [39] Shreve A P and Mathies R A 1995 Thermal effects in resonance Raman-scattering—analysis of the Raman intensities of rhodopsin and of the time-resolved Raman-scattering of bacteriorhodopsin *J. Phys. Chem.* **99** 7285–99
- [40] Tokmakoff A, Lang M J, Larsen D S, Fleming G R, Chernyak V and Mukamel S 1997 Two-dimensional Raman spectroscopy of vibrational interactions in liquids *Phys. Rev. Lett.* **79** 2702–5
- [41] Owrutsky J C, Raftery D and Hochstrasser R M 1994 Vibrational-relaxation dynamics in solutions *Ann. Rev. Phys. Chem.* **45** 519–55
- [42] Hamm P, Lim M, DeGrado W F and Hochstrasser R M 1999 The two-dimensional IR nonlinear spectroscopy of a

cyclic penta-peptide in relation to its three-dimensional structure *Proc. Natl Acad. Sci. USA* **96** 2036–41

- [43] Zheng Y F, Wang W H, Lin J G, She Y B and Fu K J 1992 Time-resolved infrared studies of gas-phase coordinatively unsaturated photofragments (Eta-5-C5h5)Mn(Co)X (X = 2 and 1) *J. Phys. Chem.* **96** 7650–6
- [44] Cavanagh R R, Heilweil E J and Stephenson J C 1994 Time-resolved measurements of energy-transfer at surfaces *Surf. Sci.* **300** 643–55
- [45] Hamm P, Lim M and Hochstrasser R M 1997 Vibrational energy relaxation of the cyanide ion in water *J. Chem. Phys.* **107** 10,523–31

-31-

- [46] Seifert F, Petrov V and Woerner M 1994 Solid-state laser system for the generation of midinfrared femtosecond pulses tunable from 3.3-Mu-M to 10-Mu-M *Opt. Lett.* **19** 2009–11
- [47] Wight C A and Armentrout P B 1993 Laser photoionization probes of ligand-binding effects in multiphoton dissociation of gas-phase transition-metal complexes *ACS Symposium Series* **530** 61–74
- [48] Belbruno J J 1995 Multiphoton ionization and chemical-dynamics *Int. Rev. Phys. Chem.* **14** 67–84
- [49] Crippa P R, Vecli A and Viappiani C 1994 Time-resolved photoacoustic-spectroscopy—new developments of an old idea *J. Photochem. Photobiol. B-Biol.* **24** 3–15
- [50] Muller-Dethlefs K and Schlag E W 1998 Chemical applications of zero kinetic energy (ZEKE) photoelectron spectroscopy *Angew. Chem.-Int. Edit.* **37** 1346–74
- [51] Bailey C G, Dessent C E H, Johnson M A and Bowen K H 1996 Vibronic effects in the photon energy-dependent photoelectron spectra of the CH₃CN⁻ dipole-bound anion *J. Chem. Phys.* **104** 6976–83
- [52] Ayotte P, Bailey C G, Weddle G H and Johnson M A 1998 Vibrational spectroscopy of small Br • (H₂O)_n and I • H₂O)_n clusters: infrared characterization of the ionic hydrogen bond *J. Phys. Chem. A* **102** 3067–71
- [53] Miller R E 1990 Vibrationally induced dynamics in hydrogen-bonded complexes *Accounts Chem. Res.* **23** 10–16
- [54] Lehmann K K, Scoles G and Pate B H 1994 Intramolecular dynamics from eigenstate-resolved infrared-spectra *Ann. Rev. Phys. Chem.* **45** 241–74
- [55] Huang Z S, Jucks K W and Miller R E 1986 The vibrational predissociation lifetime of the HF dimer upon exciting the free-H stretching vibration *J. Chem. Phys.* **85** 3338–41
- [56] Stipe B C, Rezaei M A and Ho W 1998 Single-molecule vibrational spectroscopy and microscopy *Science* **280** 1732–5

FURTHER READING

Wilson E B Jr, Decius J C and Cross P C 1955 *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra* (New York: Dover)

A classic text on molecular vibrations.

Herzberg G 1945 *Molecular Spectra and Molecular Structure II. Infrared and Raman Spectra of Polyatomic Molecules* (New York: Van Nostrand Reinhold)

Comprehensive treatment of vibrational spectroscopy, including data for a wide variety of molecules.

Califano S 1976 *Vibrational States* (New York: Wiley)

Similar to Wilson, Decius and Cross, but somewhat more modern. First three chapters provide a nice overview.

McQuarrie D A 1983 *Quantum Chemistry* (Mill Valley, CA: University Science Books)

The sections on vibrations and spectroscopy are somewhat more accessible mathematically than the previous three books.

Turrell G and Corset J (eds) 1996 *Raman Microscopy: Developments and Applications* (New York: Academic Press)

In addition to covering Raman microscopy, this book has a wealth of information on Raman instrumentation in general.

Elving P J and Winefordner J D (eds) 1986 *Fourier Transform Infrared Spectroscopy* (New York: Wiley)

Comprehensive coverage of all fundamental aspects of Fourier transform infrared spectroscopy.

B1.3 Raman spectroscopy

Darin J Ulness, Jason C Kirkwood and A C Albrecht

B1.3.1 INTRODUCTION

Light, made monochromatic and incident upon a sample, is scattered and transmitted at the incident frequency—a process known as Rayleigh scattering. Early in 1928 C V Raman and K S Krishnan reported [1, 2 and 3] the visual discovery of a new form of secondary radiation. Sunlight, passed through a blue–violet filter, was used as the light source incident upon many different organic liquids and even vapours. A green filter was placed between the sample and the viewer. The filters were sufficiently complementary to suppress the strong Rayleigh scattering and leave at longer wavelengths a feeble new kind of scattered radiation. Impurity fluorescence was discounted, for the signal was robust under purification and it also was strongly polarized. It was suggested that the incident photons had undergone inelastic scattering with the material—much as the Compton scattering of x-rays by electrons. Almost simultaneously, G Landsberg and L Mandelstam [4] reported a new kind of secondary radiation from crystalline quartz illuminated by the lines of the mercury vapour lamp. Spectrograms revealed a feeble satellite line to the red of each of the Rayleigh scattered mercury lines. In each case the displacement came close to a characteristic vibrational frequency of quartz at $\sim 480\text{ cm}^{-1}$. They wondered whether their new secondary radiation might not be of the same type as that seen by Krishnan and Raman. Such inelastic scattering of photons soon came to be called (spontaneous) Raman scattering or, given its quantized energy displacements, simply (spontaneous) Raman spectroscopy [5].

The 70 years since these first observations have witnessed dramatic developments in Raman spectroscopy, particularly with the advent of lasers. By now, a large variety of Raman spectroscopies have appeared, each with its own acronym. They all share the common trait of using high energy (‘optical’) light to probe small energy level spacings in matter.

This chapter is aimed at any scientist who wishes to become acquainted with this broad and interesting field. At the start we outline and present the principles and modern theoretical structure that unifies the many versions of Raman spectroscopies currently encountered. Then we sketch, briefly, individual examples from the contemporary literature of many of the Raman spectroscopies, indicating various applications. Though the theoretical structure is intended to stand on its own when discussing any one subject, there is no pretence of completeness, nor depth—this is not a review. But it is hoped that through the selected citations the interested reader is easily led into the broader literature on any given topic—including investigations yet to appear.

The study of small energy gaps in matter using the ‘optical’ spectral region (say the near-IR, visible and UV) offers many advantages over direct one-photon spectroscopies in the IR, far IR or even the microwave. First,

it is instrumentally convenient. Second, one can readily avoid the problem of absorbing solvents when studying dissolved solutes. Finally, in the optical region, additional strong resonances (usually involving electronic transitions) are available which can enhance the Raman scattered intensity by many orders of magnitude. This has created the very lively field of resonance Raman spectroscopy—arguably one of the most popular current Raman based spectroscopies. Resonance Raman spectroscopy not only provides greatly amplified signals from specific Raman resonances in the ground state, but it also exposes useful properties of the upper electronic potential energy hypersurface that is reached in the optical resonance.

-2-

In general, the coupling of light with matter has as its leading term the electric field component of the electromagnetic (EM) radiation—extending from the microwave into the vacuum ultraviolet. In the optical region (or near optical), where the Raman spectroscopies are found, light fields oscillate in the ‘petahertz’ range (of the order of 10^{15} cycles per second). So for technical reasons the signals are detected as photons (‘quadrature in the field’)—not as the oscillating field itself. As in any spectroscopy, Raman spectroscopies measure eigenvalue differences, dephasing times (through the bandwidths, or through time-resolved measurements) and quantitative details concerning the strength of the light/matter interaction—through the scattering cross-sections obtained from absolute Raman intensities. The small energy gap states of matter that are explored in Raman spectroscopy include phonon-like lattice modes (where Brillouin scattering can be the Raman scattering equivalent), molecular rotations (rotational Raman scattering), internal vibrations of molecules (vibrational Raman scattering) and even low lying electronic states (electronic Raman scattering). Also spin states may be probed through the magnetic component of the EM field. With the introduction of lasers, the Raman spectroscopies have been brought to a new level of sensitivity as powerful analytic tools for probing samples from the microscopic level (microprobes) to remote sensing. Not only have lasers inspired the development of new techniques and instrumentation, but they also have spawned more than 25 new kinds of Raman spectroscopy. As we shall see, this growth in experimental diversity has been accomplished by increasingly comprehensive theoretical understanding of all the spectroscopies.

The many Raman spectroscopies are found as well defined subgroups among the ‘electric field’ spectroscopies in general. (In this chapter, the magnetic field spectroscopies are mentioned only in passing.) First, except for very high intensities, the energy of interaction of light with matter is sufficiently weak to regard modern spectroscopies as classifiable according to perturbative orders in the electric field of the light. Thus any given spectroscopy is regarded as being *linear* or *nonlinear* with respect to the incident light fields. In another major classification, a given spectroscopy (linear or nonlinear) is said to be *active* or *passive* [6, 7]. The *active* spectroscopies are those in which the principal event is a change of state population in the material. In order to conserve energy this must be accompanied by an appropriate change of photon numbers in the light field. Thus net energy is transferred between light and matter in a manner that survives averaging over many cycles of the perturbing light waves. We call these the Class I spectroscopies. They constitute all of the well known absorption and emission spectroscopies—whether they are one-photon or multi-photon. The *passive* spectroscopies, called Class II, arise from the momentary exchange of energy between light and matter that induces a macroscopically coherent, oscillating, electrical polarization (an oscillating electric dipole density wave) in the material. As long as this coherence is sustained, such polarization can serve as a source term in the wave equation for the electric field. A new (EM) field (the signal field) is produced at the frequency of the oscillating polarization in the sample. Provided the polarization wave retains some coherence, and matches the signal field in direction and wavelength (a condition called ‘phase matching’), the new EM field can build up and escape the sample and ultimately be measured ‘in quadrature’ (as photons). In their extreme form, when no material resonances are operative, the Class II events (‘spectroscopies’ is a misnomer in the absence of resonances!) will alter only the states of the EM radiation and none of the material. In this case, the material acts *passively* while ‘catalysing’ alterations in the radiation. When resonances are present, Class II events become spectroscopies; some net energy may be transferred between light and matter, even as one focuses experimentally not on population changes in the material, but on alterations of the radiation. Class II events include all of the resonant and nonresonant, linear and nonlinear *dispersions*. Examples of Class II spectroscopies are classical diffraction and reflection (strongest at the linear level) and a whole array of light-scattering phenomena such as frequency summing (harmonic generation),

frequency differencing, free induction decay and optical echoes.

The Class II (passive) spectroscopies

-3-

- (i) may or may not contain resonances,
- (ii) can appear at all orders of the incident field (but only at odd order for isotropic media (gases, liquids, amorphous solids)),
- (iii) have a cross-section that is quadratic in concentration when the signal wave is homodyne detected and
- (iv) require phase-matching through experimental design, because the signal field must be allowed to build up from the induced polarization over macroscopic distances.

In [table B1.3.1](#) and [table B1.3.2](#), we assemble the more than twenty-five Raman spectroscopies and order them according to their degree of nonlinearity and by their class ([table B1.3.1](#) for Class I, [table B1.3.2](#) for Class II).

A diagrammatic approach that can unify the theory underlying these many spectroscopies is presented. The most complete theoretical treatment is achieved by applying statistical quantum mechanics in the form of the time evolution of the light/matter *density operator*. (It is recommended that anyone interested in advanced study of this topic should familiarize themselves with density operator formalism [[8](#), [9](#), [10](#), [11](#) and [12](#)]. Most books on nonlinear optics [[13](#), [14](#), [15](#), [16](#) and [17](#)] and nonlinear optical spectroscopy [[18](#), [19](#)] treat this in much detail.) Once the density operator is known at any time and position within a material, its matrix in the eigenstate basis set of the constituents (usually molecules) can be determined. The ensemble averaged electrical polarization, \mathbf{P} , is then obtained—the centrepiece of all spectroscopies based on the electric component of the EM field.

Following the section on theory, the chapter goes on to present examples of most of the Raman spectroscopies that are organized in [table B1.3.1](#) and [table B1.3.2](#).

The Class I (active) spectroscopies, both linear and nonlinear [[6](#), [7](#)],

- (i) always require resonances,
- (ii) appear only at odd order in the incident fields, which are acting ‘maximally in quadrature’. (That is, in polarizing the medium all but one of the fields act in pairs of Fourier components—also known as ‘in quadrature’ or as ‘conjugate pairs’. When this happens, the electrical polarization must carry the same frequency, wavelength, and wave vector as the odd, unpaired field. Since this odd order polarization acts conjugately with the odd incident field, the ‘photon’ picture of the spectroscopy survives.),
- (iii) have a cross-section that is linear in concentration of the resonant species and
- (iv) do not require phase matching by experimental design, for it is automatic.

B1.3.2 THEORY¹

The oscillating electric dipole density, \mathbf{P} (the polarization), that is induced by the total incident electric field, \mathbf{E} is the principal property that generates all of the spectroscopies, both linear and nonlinear. The energy contained in \mathbf{P} may be used in part (or altogether) to shift the population of energy states of the material, or it may in part (or altogether) reappear in the form of a new EM field oscillating at the same frequency. When the population changes, or energy loss or gain in the light is detected, one is engaged in a Class I spectroscopy. On the other hand, when properties of the new field are being measured—such as its frequency, direction (wavevector), state of polarization and amplitude (or intensity)—one has a Class II spectroscopy.

Normally the amplitude of the total incident field (or intensity of the incident light) is such that the light/matter coupling energies are sufficiently weak not to compete seriously with the ‘dark’ matter Hamiltonian. As already noted, when this is the case, the induced polarization, \mathbf{P} is treated perturbatively in orders of the total electric field. Thus one writes

$$\mathbf{P} = \overbrace{X^{(1)}\mathbf{E}}^{\mathbf{P}^{(1)}} + \overbrace{X^{(2)}\mathbf{E}\mathbf{E}}^{\mathbf{P}^{(2)}} + \overbrace{X^{(3)}\mathbf{E}\mathbf{E}\mathbf{E}}^{\mathbf{P}^{(3)}} + \dots \overbrace{X^{(s)}\mathbf{E}\dots\mathbf{E}}^{\mathbf{P}^{(s)}} + \dots \quad (\text{B1.3.1})$$

where the successive terms clearly appear with increasing order of nonlinearity in the total field, \mathbf{E} . At this point, all properties appearing in equation (B1.3.1) are mathematically pure real. The response function of the material to the electric field acting at s th order is the electrical susceptibility, $X^{(s)}$ (it is an $s + 1$ rank tensor). Each element of this tensor will carry $s + 1$ subscripts, a notation that is used, understandably, only when necessary. Furthermore, the events at, say the s th order, are sometimes referred to as ‘ $s + 1$ -wave-mixing’—the additional field being the new EM field derived from $\mathbf{P}^{(s)}$.

All *nonlinear* (electric field) spectroscopies are to be found in all terms of equation (B1.3.1) except for the first. The latter exclusively accounts for the standard linear spectroscopies—one-photon absorption and emission (Class I) and linear dispersion (Class II). For example, the term at third order contains by far the majority of the modern Raman spectroscopies ([table B1.3.1](#) and [table B1.3.2](#)).

It is useful to recognize that in the laboratory one normally configures an experiment to concentrate on one particular kind of spectroscopy, even while all possible light/matter events must be occurring in the sample. How can one isolate from equation (B1.3.1) a spectroscopy of interest? In particular, how does one uncover the Raman spectroscopies? We shall see how passage to the complex mathematical representation of the various properties is invaluable in this process. It is useful to start by addressing the issue of distinguishing Class I and Class II spectroscopies at any given order of nonlinearity.

B1.3.2.1 CLASS I AND CLASS II SPECTROSCOPIES AND THE COMPLEX SUSCEPTIBILITIES

All Class I spectroscopies at any order can be exposed by considering the long-term exchange of energy between light and matter as judged by the nonvanishing of the induced power density over many cycles of the field. The instantaneous power density at s th order is given by $\{\mathbf{E} \cdot \frac{\partial \mathbf{P}^{(s)}}{\partial t}\}$. For this product to survive over time, we ask that its normalized integral over a time T , which is much longer than the optical period of the field, should not vanish. This is called the cycle averaged, s th order power density. It is expressed as

$$W^{(s)} = \frac{1}{T} \int_0^T dt \left\{ \mathbf{E} \cdot \frac{\partial \mathbf{P}^{(s)}}{\partial t} \right\}. \quad (\text{B1.3.2})$$

For $W^{(s)} > 0$, one has absorption; for $W^{(s)} < 0$, emission. Multiphoton absorption and emission fall into this class. The Class I Raman spectroscopies clearly exhibit a net absorption of energy in Stokes scattering and a net emission of energy in anti-Stokes scattering. Though $\mathbf{P}^{(s)}$ involve s actions of the total field, the light/matter energy exchange is always in the language of photons. A net energy in the form of photons is destroyed (absorbed) as the quantum state population of the material moves upward in energy; a net energy in the form of photons is created (emitted) as the

population moves downward in energy. To survive the integration, the instantaneous power density should not oscillate rapidly (if at all), certainly not at optical frequencies. Since \mathbf{E} is intended to consist entirely of fields that oscillate at optical frequencies, the power density can have a non-oscillating term only when the field

appears altogether an *even* number of times. Since it appears s times in $\mathbf{P}^{(s)}$ (equation (B1.3.1)) and once in \mathbf{E} , all Class I spectroscopies exist only when $s + 1$ is even, or s is odd (see equation (B1.3.2)).

Furthermore, the non-oscillating component of the integrand can best be sorted out by going to the complex representation of the total field, the polarization, and the susceptibility. The mathematically pure real quantities in equation (B1.3.2) can be written in their complex representation as follows:

$$\mathbf{E} = \frac{1}{2}(\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^*) \quad (\text{B1.3.3})$$

$$\mathbf{P}^{(s)} = \frac{1}{2}(\mathbf{p}^{(s)} + \mathbf{p}^{(s)*}) \quad (\text{B1.3.4})$$

and

$$\mathbf{X}^{(s)} = \frac{1}{2}(\chi^{(s)} + \chi^{(s)*}) \quad (\text{B1.3.5})$$

in which $\boldsymbol{\varepsilon}$, $\mathbf{p}^{(s)}$ and $\chi^{(s)}$ are, in general, complex quantities whose real parts are given by \mathbf{E} , $\mathbf{P}^{(s)}$ and $\mathbf{X}^{(s)}$, respectively. Introducing equation (B1.3.4)–equation (B1.3.5) into equation (B1.3.2), and applying the cycle average theorem for the integral [20], one finds that for all spectroscopies at s th order involving long-term light/matter energy exchange (Class I, in particular), the signal measured as a net energy exchange, $S_1^{(s)}$, is proportional to the cycle averaged power density, or $S_1^{(s)} \propto W^{(s)} \propto \text{Im } \chi^{(s)}$. If the complex susceptibility, $\chi^{(s)}$, is pure real, there can be no long term energy exchange between light and matter. There can be no Class I spectroscopies based on a susceptibility component for which $\text{Im } \chi^{(s)} = 0$.

Consider an ensemble composed of N constituents (such as molecules) per unit volume. The (complex) density operator for this system is developed perturbatively in orders of the applied field, and at s th order is given by $\rho^{(s)}$. The (complex) s th order contribution to the ensemble averaged polarization is given by the trace over the eigenstate basis of the constituents of the product of the dipole operator, $N \boldsymbol{\mu}$ and $\rho^{(s)}$: $\mathbf{p}^{(s)} = \text{Tr}\{N \boldsymbol{\mu} \rho^{(s)}\}$. In turn, an expression for $\chi^{(s)}$ is obtained, which, in the frequency domain, consists of a numerator containing a product of $(s + 1)$ transition moment matrix elements and a denominator of s complex energy factors. These complex energies express light/matter resonances and allow $\chi^{(s)}$ to become a complex quantity, its $\text{Im } \chi^{(s)}$ part (pure real) being responsible for Class I spectroscopies. The light/matter resonances introduce the imaginary component to $\chi^{(s)}$ and permit a Class I spectroscopy to exist.

As noted, the Class II spectroscopies are based on detecting the new EM field that is derived from the induced polarization, $\mathbf{P}^{(s)}$, at s th order. Here $\mathbf{P}^{(s)}$, oscillating at optical frequencies, acts as the source term in Maxwell's equation to create the new optical field, \mathbf{E}_{new} , at the same frequency. Again, we recognize $\boldsymbol{\varepsilon}_{\text{new}} \propto \mathbf{p}^{(s)} \propto \chi^{(s)}$.

Since optical fields oscillate too quickly for direct detection, they are measured ‘in quadrature’—as photons (see below). There are two ways to achieve quadrature. One is *homodyne* detection in which the new field is measured at

-6-

its quadrature, $\boldsymbol{\varepsilon}_{\text{new}} \boldsymbol{\varepsilon}_{\text{new}}^* = |\boldsymbol{\varepsilon}_{\text{new}}|^2$. These signals must be proportional to $|\chi^{(s)}|^2$. Thus $S_{\text{I}}^{(s)}$ (homodyne) $\propto |\chi^{(s)}|^2$ and all phase information in $\chi^{(s)}$ is lost. Such is the case for almost all of the Class II spectroscopies, especially the Raman events at third order.

The second way to achieve quadrature is to introduce another field, \mathbf{E}_{lo} , (called a local oscillator) designed in frequency and wavevector to conjugate (go into quadrature) in its complex representation with the new field of interest. Thus in the heterodyne case, the signal photons are derived from $\boldsymbol{\varepsilon}_{\text{new}} \boldsymbol{\varepsilon}_{\text{lo}}^*$, or $S_{\text{II}}^{(s)}$ (heterodyne) $\propto \chi^{(s)}$.

In heterodyne detected $s + 1$ wave mixing, phase information is retained and one can take a full measure of the complex susceptibility, including its phase. The phase of the complex induced polarization, $\mathbf{p}^{(s)}$, determines how its energy will partition between Class I (absorbed or emitted) and Class II (a new EM wave is launched) spectroscopies.

Consider all of the spectroscopies at third order ($s = 3$). To be as general as possible, suppose the total incident field consists of the combination of three experimentally distinct fields ($j = 1, 2, 3$). These can differ in any combination of their frequency, polarization and direction of incidence (wavevector). Thus the total field is written as

$$\mathbf{E} = \sum_{j=1}^3 \mathbf{E}_j. \quad (\text{B1.3.6})$$

In using the complex representation (equation (B1.3.4)), the j th electric field is given as

$$\mathbf{E}_j = \frac{1}{2}(\epsilon_j + \epsilon_j^*) \quad (\text{B1.3.7})$$

where (using Euler's identity)

$$\epsilon_j = \mathbf{E}_j^0 e^{-i(\mathbf{k}_j \cdot \mathbf{r} - \omega_j t)} \quad (\text{B1.3.8})$$

$$\epsilon_j^* = (\mathbf{E}_j^0)^* e^{i(\mathbf{k}_j^* \cdot \mathbf{r} - \omega_j^* t)}. \quad (\text{B1.3.9})$$

Here \mathbf{E}_j^0 is the amplitude of the j th field and the real part of ω_j is its (circular) frequency or 'colour'. The real part of \mathbf{k}_j is the product of the unit vector of incidence inside the sample, \mathbf{e}_k , and its amplitude, $\frac{2\pi n_j}{\lambda_j}$. Here $\frac{\lambda_j}{n_j}$ is the wavelength of the j th field inside the sample— λ_j being the wavelength inside a vacuum and n_j being the (real) index of refraction of the sample at ω_j . As implied, all three properties may be complex: the amplitude because of an added phase to the field and/or a field that is elliptically (or circularly) polarized; the frequency because the field may be growing or decaying in time and the wavevector because the field may be decaying and growing according to its location within the sample.

The total field (equation (B1.3.6) with equation (B1.3.7)) is now

$$\mathbf{E} = \frac{1}{2} \sum_{j=1}^3 (\epsilon_j + \epsilon_j^*). \quad (\text{B1.3.10})$$

It is only a matter of inserting this 'hexanomial', equation (B1.3.10), into equation (B1.3.1) to organize all possible three-beam spectroscopies that might appear at any given order.

B1.3.2.2 THE 'GENERATORS' FOR ALL THIRD ORDER SPECTROSCOPIES FROM THE COMPLEX REPRESENTATION OF THE FIELD

In order to develop the theoretical structure that underlies each of the many Raman spectroscopies at third

order, we use the above complex representation of the incident fields to produce the ‘generators’ of all possible electric field spectroscopies at third order. After this exercise, it is a simple matter to isolate the subset that constitutes the entire family of Raman spectroscopies.

At third order, one must expand $\frac{1}{8}(\sum_{i=1}^3(\epsilon_i + \epsilon_i^*) \sum_{j=1}^3(\epsilon_j + \epsilon_j^*) \sum_{k=1}^3(\epsilon_k + \epsilon_k^*))$ to enumerate the ‘generators’ of all possible third order spectroscopies. In this case, any given generator consists of an ordered list of these complex fields, such as $\epsilon_i \epsilon_j \epsilon_k^*$. The ordering of the fields in each generator represents a time ordering of the actions of the applied fields. This can be of physical significance. Clearly this expansion must give 216 terms (6^3). These 216 terms or generators can be arranged into 108 pairs of mutually conjugate generators, since the total electric field is itself a quantity that is pure real. Of these 108 paired terms, exactly 27 are in the category of what is termed *nondegenerate* four wave mixing (ND4WM), where the signal frequency must be very far from any of the incident optical frequencies. These 27 pairs can generate only Class II spectroscopies and (it turns out) none of the Raman spectroscopies. The generic pair for these ND4WM processes is $\epsilon_i \epsilon_j \epsilon_k + \epsilon_i^* \epsilon_j^* \epsilon_k^*$, where each of i, j or k can be fields 1, 2 or 3. (Henceforth the factor of $\frac{1}{8}$ shall be suppressed since it is common to all 216 generators.) The simple algebra of exponents is applied to such a product using [equation \(B1.3.9\)](#) and [equation \(B1.3.10\)](#). Thus, one sees how the polarization wave generated from such a term must oscillate at a frequency much larger than any of the incident colours, namely at the (real part of) $\omega_p = \omega_i + \omega_j + \omega_k$. The polarization wave must have a wavevector given by

$$\text{Re}(\mathbf{k}_p) = \text{Re}(\mathbf{k}_i + \mathbf{k}_j + \mathbf{k}_k) = 2\pi \left(\left(\frac{n_i}{\lambda_i} \right) \mathbf{e}_{\mathbf{k}_i} + \left(\frac{n_j}{\lambda_j} \right) \mathbf{e}_{\mathbf{k}_j} + \left(\frac{n_k}{\lambda_k} \right) \mathbf{e}_{\mathbf{k}_k} \right).$$

The appropriate (complex) susceptibility tensor for this generator is $\chi^{(3)}(\omega_p = \omega_i + \omega_j + \omega_k)$.

When resonances, or near resonances, are present in the 4WM process, the ordering of the field actions in the perturbative treatment ([equation \(B1.3.1\)](#)), can be highly significant. Though the three-colour generators ($\epsilon_i \epsilon_j \epsilon_k, \epsilon_j \epsilon_k \epsilon_i, \epsilon_k \epsilon_i \epsilon_j, \dots$) have identical frequency and wavevector algebra, their associated susceptibility functions ($\chi^{(3)}(\omega_p = \omega_i + \omega_j + \omega_k), \chi^{(3)}(\omega_p = \omega_j + \omega_k + \omega_i), \chi^{(3)}(\omega_p = \omega_k + \omega_i + \omega_j), \dots$) are, in general, different. As a result of the different colour ordering, two of their three energy denominator factors must differ. For this reason, the field ordering in each generator, together with its own response function, must be regarded individually.

The fourth electromagnetic wave, \mathbf{E}_{new} , shall henceforth be called the signal wave, $\mathbf{E}_s = \frac{1}{2}(\epsilon_s + \epsilon_s^*)$. It always must carry the same frequency as that of the polarization wave ($\omega_p = \omega_s$). It is launched by the collapse of this wave, provided that the polarization extends coherently over at least a few wavelengths of the incident light and that $\mathbf{k}_p = \mathbf{k}_s$.

The latter condition corresponds to the phase matching requirement already mentioned—the wavelength and direction of the material polarization wave must match those of the new EM wave as closely as possible. However, for all Class I spectroscopies, this condition is automatically achieved because of quadrature. In fact, this is true for all ‘quadrature’ spectroscopies—the Class I spectroscopies being the principal such, but, as noted, it is a nontrivial requirement in the ‘nonquadrature’ Class II spectroscopies, particularly in optically dispersive media.

In the complex mathematical representation, ‘quadrature’ means that, at the $(s + 1)$ wave mixing level, the product of s input fields constituting the s th order generator and the signal field can be organized as a product of $(s + 1)/2$ conjugately paired fields. Such a pair for field i is given by One sees that the exponent algebra for

such a pair removes all dependence on the $\text{Re}(\omega_i)$ and $\text{Re}(\mathbf{k}_i)$, thus automatically removing all oscillations as well as satisfying the phase-matching requirement. A necessary (but not sufficient) requirement for full quadrature is for s to be odd and also that half of the $s + 1$ fields be found to act conjugately with respect to the other half. Thus, for $s + 1$ fields, whenever the number of nonconjugate fields differs from the number of conjugate fields, one can only have ‘nonquadrature’. Phase matching then must become an issue. This must always be the case for odd-wave mixing (s is even), and it is also true for the above set of 27 frequency summing generators (for ND4WM). Thus for the currently considered generator, $\epsilon_i \epsilon_j \epsilon_k$, all three fields (i, j and k) act nonconjugately, so quadrature at the four-wave level simply is not possible.

B1.3.2.3 THE FIELD GENERATORS FOR ALL THIRD ORDER RAMAN SPECTROSCOPIES

We are now prepared to uncover all of the third order Raman spectroscopies (table B1.3.1 and table B1.3.2) and, in doing so, indicate how one might proceed to reveal the Raman events at higher order as well. Of the 108 pairs of conjugate generators, the remaining 81, unlike the above 27, are characterized by having one of the three fields acting conjugately with respect to the remaining two. Nine of these 81 terms generate fully degenerate 4WM. They constitute the ‘one-colour’ terms: and with susceptibility $\chi^{(3)}(\omega_s = \omega_i)$, since $\omega_s = -\omega_i + \omega_i + \omega_i = \omega_i$ with wavevector $\mathbf{k}_s = -\mathbf{k}_i + \mathbf{k}_i + \mathbf{k}_i = \mathbf{k}_i$. Their full degeneracy is evident since all four fields carry the same frequency (apart from sign). Resonances appear in the electric susceptibilities when, by choice of incident colours and their signs, one or more of their energy denominators (s in number at s th order) approaches a very small value because the appropriate algebraic colour combination matches material energy gaps. All Raman spectroscopies must, by definition, contain at least one low frequency resonance. When using only optical frequencies, this can only be achieved by having two fields acting conjugately and possessing a difference frequency that matches the material resonance. Further, they must act in the first two steps along the path to the third order polarization of the sample. These first two steps together prepare the Raman resonant material coherence and can be referred to as the ‘doorway’ stage of the Raman 4WM event.

Suppose the incident colours are such that the required Raman resonance (ω_R) appears at $\omega_1 - \omega_2 \approx \omega_R$ (with $\omega_1 > \omega_2$). Thus the appropriate generators for the two-step doorway stage, common to all of the Raman spectroscopies, must be and , (and their conjugates, and). These differ only in the permutation of the ordering of the actions of fields 1 and 2. The usual algebra of exponents tells us that this doorway stage produces an intermediate polarization that oscillates at $\omega_1 - \omega_2 \approx \omega_R$. The resonantly produced Raman coherence having been established, it remains only to probe this intermediate polarization, that is, to convert it into an optical polarization from which the signal at optical frequencies is prepared. This is accomplished by the action of the third field which converts the low frequency Raman coherence into an optical one. This in turn, leads to the signal. This last two-step event can be referred to as the ‘window’ stage of the 4WM process. Obviously there are at most only six possible choices for this

third field: ϵ_j and ϵ_j^* with $j = 1, 2, 3$. In this way we have isolated from the original 108 4WM generators the 12 that are responsible for all of the Raman spectroscopies. The choice of the probe field determines the frequency of the signal as well as its wavevector. While all types of Raman signal must be present at some level of intensity in any given experiment, it is a matter of experimental design—the detected ω_s and the aperture selected \mathbf{k}_s —as to which one Raman spectroscopy is actually being studied.

This classification by field generators in the complex field representation goes a long way towards organizing the nonlinear spectroscopies that carry at least one resonance. However it must be remembered that, in emphasizing a field ordering that locates the essential Raman resonance, we have neglected any other possible resonant and all nonresonant contributions to the third order polarization. While any additional resonance(s) are important to both Class I and Class II spectroscopies, the nonresonant contributions play no role in Class I spectroscopies, but must not be ignored in Class II studies. If one starts with the generator $\epsilon_1 \epsilon_2^* \epsilon_3$ and then permutes the ordering of the three distinct complex field amplitudes, one arrives at 6 (3!) generators, only two

of which induce the Raman resonance at $\omega_1 - \omega_2 \approx \omega_R$ ($\epsilon_1 \epsilon_2^* \epsilon_3$ and $\epsilon_2^* \epsilon_1 \epsilon_3$). The remaining four can only polarize the material without this Raman resonance ($\omega_3 \neq \omega_1, \omega_2$), but, if otherwise resonant, it can interfere with the Raman lineshape. Also when the issue of time resolved spectroscopies arises, the time ordering of field actions is under experimental control, and the active field generators are limited not only by requirements of resonance but by their actual time ordering in the laboratory.

In any case, the polarizing action upon the material by a given generator must be followed in more detail. In density matrix evolution, each specified field action transforms either the ‘ket’ or the ‘bra’ side of the density matrix. Thus for any specified s th order generator, there can be 2^s detailed paths of evolution. In addition, evolution for each of the $s!$ generators corresponding to all possible field orderings must be considered. One then has altogether 2^s paths of evolution. For $s = 3$ there are 48—eight for each of six generators.

B1.3.2.4 TIME EVOLUTION OF THE THIRD ORDER POLARIZATION BY WAVE MIXING ENERGY LEVEL (WMEL) DIAGRAMS. THE RAMAN SPECTROSCOPIES CLASSIFIED

The general task is to trace the evolution of the third order polarization of the material created by each of the above 12 Raman field operators. For brevity, we choose to select only the subset of eight that is based on two colours only—a situation that is common to almost all of the Raman spectroscopies. Three-colour Raman studies are rather rare, but are most interesting, as demonstrated at both third and fifth order by the work in Wright’s laboratory [21, 22, 23 and 24]. That work anticipates variations that include infrared resonances and the birth of doubly resonant vibrational spectroscopy (DOVE) and its two-dimensional Fourier transform representations analogous to 2D NMR [25].

Interestingly, three-colour spectroscopies at third order can only be of Class II, since the generators cannot possibly contain any quadrature. Maximal quadrature is necessary for Class I.

For the two-colour spectroscopies, maximal quadrature is possible and both Class I and Class II events are accounted for. Thus we trace diagrammatically the evolution to produce a signal field caused by the eight generators: (i) the four $(\epsilon_1 \epsilon_2^*) \epsilon_1^*$, $(\epsilon_1 \epsilon_2^*) \epsilon_2^*$, $(\epsilon_1 \epsilon_2^*) \epsilon_1$ and $(\epsilon_1 \epsilon_2^*) \epsilon_2$ and (ii) the four with the first two fields permuted, $(\epsilon_2^* \epsilon_1) \epsilon_1^*$, $(\epsilon_2^* \epsilon_1) \epsilon_2^*$, $(\epsilon_2^* \epsilon_1) \epsilon_1$ and $(\epsilon_2^* \epsilon_1) \epsilon_2$.

In each case we have indicated the doorway stage using parentheses. We note how in each of the two groups (i) and

The appropriate (complex) susceptibility tensor for this generator is $\chi^{(3)}(\omega_p = \omega_i + \omega_j + \omega_k)$.

When resonances, or near resonances, are present in the 4WM process, the ordering of the field actions in the perturbative treatment (equation (B1.3.1)), can be highly significant. Though the three-colour generators ($\epsilon_i \epsilon_j \epsilon_k$, $\epsilon_j \epsilon_k \epsilon_i$, $\epsilon_k \epsilon_i \epsilon_j$, . . .) have identical frequency and wavevector algebra, their associated susceptibility functions ($\chi^{(3)}(\omega_p = \omega_i + \omega_j + \omega_k)$, $\chi^{(3)}(\omega_p = \omega_j + \omega_k + \omega_i)$, $\chi^{(3)}(\omega_p = \omega_k + \omega_i + \omega_j)$, . . .) are, in general, different. As a result of the different colour ordering, two of their three energy denominator factors must differ. For this reason, the field ordering in each generator, together with its own response function, must be regarded individually.

The fourth electromagnetic wave, E_{new} , shall henceforth be called the signal wave, $E_s = \frac{1}{2}(\epsilon_s + \epsilon_s^*)$. It always must carry the same frequency as that of the polarization wave ($\omega_p = \omega_s$). It is launched by the collapse of this wave, provided that the polarization extends coherently over at least a few wavelengths of the incident light and that $\mathbf{k}_n = \mathbf{k}_s$.

The latter condition corresponds to the phase matching requirement already mentioned—the wavelength and direction of the material polarization wave must match those of the new EM wave as closely as possible. However, for all Class I spectroscopies, this condition is automatically achieved because of quadrature. In fact, this is true for all ‘quadrature’ spectroscopies—the Class I spectroscopies being the principal such, but, as noted, it is a nontrivial requirement in the ‘nonquadrature’ Class II spectroscopies, particularly in optically dispersive media.

In the complex mathematical representation, ‘quadrature’ means that, at the $(s + 1)$ wave mixing level, the product of s input fields constituting the s th order generator and the signal field can be organized as a product of $(s + 1)/2$ conjugately paired fields. Such a pair for field i is given by $\epsilon_i \epsilon_i^* = |\epsilon_i|^2$. One sees that the exponent algebra for such a pair removes all dependence on the $\text{Re}(\omega_i)$ and $\text{Re}(\mathbf{k}_i)$, thus automatically removing all oscillations as well as satisfying the phase-matching requirement. A necessary (but not sufficient) requirement for full quadrature is for s to be odd and also that half of the $s + 1$ fields be found to act conjugately with respect to the other half. Thus, for $s + 1$ fields, whenever the number of nonconjugate fields differs from the number of conjugate fields, one can only have ‘nonquadrature’. Phase matching then must become an issue. This must always be the case for odd-wave mixing (s is even), and it is also true for the above set of 27 frequency summing generators (for ND4WM). Thus for the currently considered generator, $\epsilon_i \epsilon_j \epsilon_k$, all three fields (i, j and k) act nonconjugately, so quadrature at the four-wave level simply is not possible.

B1.3.2.3 THE FIELD GENERATORS FOR ALL THIRD ORDER RAMAN SPECTROSCOPIES

We are now prepared to uncover all of the third order Raman spectroscopies (table B1.3.1 and table B1.3.2) and, in doing so, indicate how one might proceed to reveal the Raman events at higher order as well. Of the 108 pairs of conjugate generators, the remaining 81, unlike the above 27, are characterized by having one of the three fields acting conjugately with respect to the remaining two. Nine of these 81 terms generate fully degenerate 4WM. They constitute the ‘one-colour’ terms: $\epsilon_i^* \epsilon_i \epsilon_i + \text{cc}$, $\epsilon_i \epsilon_i^* \epsilon_i + \text{cc}$ and

$\epsilon_i \epsilon_i \epsilon_i^* + \text{cc}$ ($i = 1, 2, \text{ or } 3$) with susceptibility $\chi^{(3)}(\omega_s = \omega_i)$, since $\omega_s = -\omega_i + \omega_i + \omega_i = \omega_i$ with wavevector $\mathbf{k}_s = -\mathbf{k}_i + \mathbf{k}_i + \mathbf{k}_i = \mathbf{k}_i$. Their full degeneracy is evident since all four fields carry the same frequency (apart from sign). Resonances appear in the electric susceptibilities when, by choice of incident colours and their signs, one or more of their energy denominators (s in number at s th order) approaches a very small value because the appropriate algebraic colour combination matches material energy gaps. All Raman spectroscopies must, by definition, contain at least one low frequency resonance. When using only optical frequencies,

chromophore). This matter–bath interaction is an important, usually dominant, source of damping of the macroscopic coherence. The radiative blackbody (bb) background is another ever-present bath that imposes radiative damping of excited states, also destroying coherences.

For $s = 3$, the time evolution of the system is tracked by following the stepwise changes in the bra state, $\langle j|$, or the ket state, $|k\rangle$, of the system caused by each of the three successive field interventions. This perturbative evolution of the density operator, or of the density matrix, is conveniently depicted diagrammatically using double sided Feynman diagrams or, equivalently, the WMEL diagrams. The latter are preferred since light/matter resonances are explicitly exposed. In WMEL diagrams, the energy levels of the constituents of the matter are laid out as solid horizontal lines to indicate the states (called ‘real’) that are active in a resonance, and as dashed horizontal lines (or no lines) when they serve as nonresonant (‘virtual’) states. The perturbative evolution of the density matrix is depicted using vertically oriented arrows for each of the field actions that appears in a given generator. These arrows are placed from left to right in the diagram in the same order as the corresponding field action in the generator. The arrow length is scaled to the frequency of the acting field. Solid arrows indicate evolution from the old ket (tail of arrow) to the new ket (head of arrow); dashed arrows indicate evolution from the old bra (tail of arrow) to the new bra (head of arrow). For a field acting nonconjugately, like ϵ_i , the frequency is positively signed, ω_i , and the arrow for a ket change points up

and that for a bra change points down. When the field acts conjugately, ω_1 , the frequency is negatively signed, $-\omega_1$, and a ket changing arrow points down, while a bra changing arrow points up. These rules allow one to depict diagrammatically any and all density matrix evolutions at any order. Given the option of a bra or a ket change at each field action, one sees how a given s th order generator leads to 2^s diagrams, or paths of evolution. Normally only some (if any) encounter resonances. A recipe has been published [6] that allows one to translate any WMEL diagram into the analytic expression for its corresponding electrical susceptibility. After s arrows have appeared (for an s th order evolution), the $(s + 1)$ th field is indicated for any WMEL diagram of the nonquadrature class by a vertical wavy *line segment* whose vertical length scales to the signal frequency. For the WMEL diagrams of the full quadrature sort, the $(s + 1)$ th field must be conjugate to one of the incident fields, so the wavy segment becomes a wavy arrow; either solid (ket-side action) or dashed (bra-side action).

Of the four possible WMEL diagrams for each the and doorway generators, only one encounters the Raman resonance in each case. We start with two parallel horizontal solid lines, together representing the energy gap of a Raman resonance. For ket evolution using ω_1 , we start on the left at the lowest solid line (the ground state, g) and draw a long solid arrow pointing up ($+\omega_1$), followed just to the right by a shorter solid arrow pointing down ($-\omega_2$) to reach the upper solid horizontal line, f . The head of the first arrow brings the ket to a virtual state, from which the second arrow carries the ket to the upper of the two levels of the Raman transition. Since the bra is until now unchanged, it remains in g ($\langle g|$); this doorway event leaves the density matrix at second order off-diagonal in which is not zero. Thus a Raman coherence has been established. Analogously, the doorway action on the ket side must be short solid arrow down ($-\omega_2$) from g to a virtual ket state, then long arrow up ($+\omega_1$) to f from the virtual state. This evolution also produces χ . Both doorway actions contain the same Raman resonance denominator, but differ in the denominator appearing at the first step; the downward action is inherently anti-resonant ('N' for nonresonant) in the first step, the upward action is potentially resonant ('R' for resonant) in the first step and is therefore stronger. Accordingly, we distinguish these two doorway events by labels D_N and D_R , respectively (see figure B1.3.2). In resonance Raman spectroscopy, this first step in D_R is fully resonant and overwhelms D_N . (The neglect of D_N is known as the rotating wave approximation.) It is unnecessary to explore the bra-side version of these doorway actions, for they would appear in the fully conjugate version of these doorway events. Each of the doorway steps, D_R

-12-

and D_N , may be followed by any one of eight window events. The WMEL diagrams for the window events consist of the arrow for the last step of the third order polarization and the wavy segment for the signal wave. There are eight such window diagrams since each of the two steps can involve two colours and either bra- or ket-side evolution. These eight window WMEL diagrams are shown in figure B1.3.2(a) and figure B1.3.2(b) and are identified alphabetically. These also carry potentially resonant and anti-resonant properties in the third energy denominator (the first window step) and accordingly are labelled W_R and W_N , where, as before, $W_R > W_N$. If the third step is completely resonant, $W_R \gg W_N$, and W_N may be completely neglected (as with $D_R \gg D_N$).

Figure B1.3.2. The separate WMEL diagrams for the doorway and window stages of the Raman spectroscopies. Solid and dashed vertical arrows correspond to ket- and bra-side light/matter interactions, respectively. The signal field is denoted by the vertical wavy line (arrow). The ground and final molecular levels (solid horizontal lines) are labelled g and f , while the virtual levels (dashed horizontal lines) are labelled j and k . The associated generators are given below each diagram. The doorway/window stages are classified as potentially resonant (D_R/W_R) or certainly nonresonant (D_N/W_N). In addition, the window stages are labelled alphabetically in order to distinguish the Raman techniques by their window stage WMEL diagram(s)

(as in table B1.3.1 and table B1.3.2 and figure B1.3.1). (a) The doorway and window stage WMEL diagrams for SR, SRS and RRS. (b) The doorway and window stage WMEL diagrams for CARS and CSRS.

-13-

It is now possible to label every one of the Raman spectroscopies listed in [table B1.3.1](#) and [table B1.3.2](#) according to its essential doorway/window WMEL diagram. This is shown in the third column of those tables. Again, the analytic form of the associated susceptibilities is obtained by recipe from the diagrams. When additional resonances are present, other WMEL diagrams must be included for both Class I and Class II spectroscopies. For the Class II spectroscopies, all of the nonresonant WMEL diagrams must be included as well.

B1.3.2.5 THE MICROSCOPIC HYPERPOLARIZABILITY TENSOR, ORIENTATIONAL AVERAGING, THE KRAMERS–HEISENBERG EXPRESSION AND DEPOLARIZATION RATIOS

As implied by the trace expression for the macroscopic optical polarization, the macroscopic electrical susceptibility tensor at any order can be written in terms of an ensemble average over the microscopic nonlinear polarizability tensors of the individual constituents.

(A) MICROSCOPIC HYPERPOLARIZABILITY AND ORIENTATIONAL AVERAGING

Consider an isotropic medium that consists of independent and identical microscopic chromophores (molecules) at number density N . At s th order, each element of the macroscopic susceptibility tensor, given in laboratory Cartesian coordinates A, B, C, D , must carry $s + 1$ (laboratory) Cartesian indices (X, Y or Z) and therefore number altogether $3^{(s+1)}$. Thus the third order susceptibility tensor contains 81 elements. Each tensor element of the macroscopic susceptibility is directly proportional to the sum over all elements of the corresponding microscopic, or molecular, hyperpolarizability tensor. The latter are expressed in terms of the four local (molecule based) Cartesian coordinates, a, b, c, d (each can be x, y , or z —accounting for all 81 elements of the *microscopic* tensor). To account for the contribution to the macroscopic susceptibility from each molecule, one must sum over all molecules in a unit volume, explicitly treating all possible orientations, since the projection of a microscopic induced dipole onto the X, Y and Z laboratory coordinates depends very much on the molecular orientation. This is accomplished by averaging the microscopic hyperpolarizability contribution to the susceptibility over a normalized distribution of orientations, and then simply multiplying by the number density, N . Let the orientational averaging be denoted by $\langle \dots \rangle$. For any macroscopic tensor element, $\chi_{ABCD}^{(3)}$, one finds

$$\chi_{ABCD}^{(3)} = \frac{L^3}{\hbar^3 \epsilon_0} N \sum_{a,b,c,d} \langle (A, a)(B, b)(C, c)(D, d) \rangle \gamma_{abcd} \quad (\text{B1.3.11})$$

where (A, a) etc are the direction cosines linking the specified local Cartesian axes with specified laboratory axes and L is a ‘local’ field factor. (The field experienced by the molecule is the incident field altered by polarization effects.) Often the \hbar^{-S} factor is absorbed into the definition of the s th order hyperpolarizability.

The microscopic polarization involves four molecular based transition moment vectors—the induced dipole is along a , the first index. The transition moments along b, c and d are coupled to the laboratory axes, B, C and D , respectively, along which the successive incident (or black-body) fields are polarized. The four transition moment unit vectors have been extracted and projected onto the laboratory axes: A —the direction of the induced macroscopic polarization, B —the polarization of the first acting field, C —the polarization of the second acting field and D —the polarization of the third acting field. The product of the four direction cosines is subjected to the orientational averaging process, as indicated. Each such average belongs to its

corresponding scalar tensor element γ_{abcd} . It is important to sum over all

-14-

local Cartesian indices, since all elements of γ can contribute to each tensor element of $\chi^{(3)}$. At third order, the averaging over the projection of microscopic unit vectors to macroscopic, such as those in equation (B1.3.11), is identical to that found in two-photon spectroscopy (another Class I spectroscopy at third order). This has been treated in a general way (including circularly polarized light) according to molecular symmetry considerations by Monson and McClain [26].

For an isotropic material, all orientations are equally probable and all such products that have an odd number of ‘like’ direction cosines will vanish upon averaging². This restricts the nonvanishing tensor elements to those such as $\chi_{AAAA}^{(3)}$, $\chi_{ABBA}^{(3)}$ etc. Similarly for the elements γ_{abcd} . Such orientational averaging is crucial in dictating how the signal field in any spectroscopy is polarized. In turn, polarization measurements can lead to important quantitative information about the elements of the macroscopic and microscopic tensors.

The passage from microscopic to macroscopic (equation (B1.3.11)) clearly exposes the additivity of the microscopic hyperpolarizabilities. Significantly, it is seen immediately why $\chi^{(3)}$ is linear in concentration, N . This brings out one of the major distinctions between the Class I and Class II spectroscopies (see item (iii) in section B1.3.1). The signals from Class I, being proportional to $\text{Im } \chi^{(3)}$, are *linear* in concentration. Those (homodyne signals) from Class II are proportional to $|\chi^{(3)}|^2$ and therefore must be *quadratic* in concentration. (However, Class II spectroscopies that are heterodyne detected are proportional to $\chi^{(3)}$ and are linear in N .)

(B) THE MICROSCOPIC HYPERPOLARIZABILITY IN TERMS OF THE LINEAR POLARIZABILITY: THE KRAMERS–HEISENBERG EQUATION AND PLACZEK LINEAR POLARIZABILITY THEORY OF THE RAMAN EFFECT

The original Placzek theory of Raman scattering [30] was in terms of the linear, or first order microscopic polarizability, α (a second rank tensor), not the third order hyperpolarizability, γ (a fourth rank tensor). The Dirac and Kramers–Heisenberg quantum theory for linear dispersion did account for Raman scattering. It turns out that this link of properties at third order to those at first order works well for the electronically nonresonant Raman processes, but it cannot hold rigorously for the fully (triply) resonant Raman spectroscopies. However, provided one discards the important line shaping phenomenon called ‘pure dephasing’, one can show how the third order susceptibility does reduce to the treatment based on the (linear) polarizability tensor [6, 27].

What is the phenomenon ‘pure dephasing’ that one cannot formally encounter in the linear polarizability theories of Raman spectroscopies? It arises when theory is obliged to treat the environment of the spectroscopically active entity as a ‘bath’ that statistically modulates its states. In simple terms, there are two mechanisms for the irreversible decay of a coherent state such as a macroscopically coherent polarization wave. One involves the destruction by the bath of the local induced dipoles that make up the wave (a lifetime effect); the other involves the bath induced randomization of these induced dipoles (without their destruction). This latter mechanism is called pure dephasing. Together, their action is responsible for dephasing of a pure coherence. In addition, if the system is inherently inhomogeneous in the distribution of the two-level energy gap of the coherence, the local coherences will oscillate at slightly different frequencies causing, as these walk off, the macroscopic coherence, and its signal, to decay—even while the individual local coherences might not. This is especially important for the Class II spectroscopies. Unlike the first two dephasing mechanisms, this third kind can be reversed by attending to signals generated by the appropriate Fourier components of the subsequent field actions. The original macroscopic coherence will be reassembled (at least partially) in due course, to produce a renewed signal, called an ‘echo’. For an electronic coherence made by a single optical field, this happens at the four-wave mixing level. However for the Raman spectroscopies, *two* (conjugately acting) optical fields are needed to create the vibrational coherence, and hence the true Raman echo appears at the eight-wave mixing level

where $\chi^{(7)}$ is the important susceptibility. (We shall see that a *quasi* Raman echo can be exploited at the $\chi^{(5)}$ level.)

The important and frequently ignored fact is that Raman theory based on the polarizability tensor cannot contain the randomization mechanism for dephasing (pure dephasing). This mechanism is especially important in electronically resonant Raman spectroscopy in the condensed phase. The absence of pure dephasing in linear polarizability theory arises simply because the perturbative treatment upon which it is based involves the *independent* evolution of the bra and the ket states of the system. Conversely, the third order susceptibility approach, based on the perturbative development of the density operator, links together the evolution of the bra and ket states and easily incorporates pure dephasing. Furthermore, in resonance Raman spectroscopy, it is the pure dephasing mechanism that governs the interesting competition between resonance fluorescence and resonance Raman scattering [6]. In the linear polarization theory these are fixed in their relation, and the true resonance fluorescence component becomes an indistinguishable part of the Raman line shape. (However, interestingly, if the exciting light itself is incoherent, or the exciting light consists of sufficiently short pulses, even the linear polarizability theory tells how the resonance fluorescence-like component becomes distinguishable from the Raman-like signal [20].)

At the linear level, the microscopic induced dipole vector on a single molecule in the local Cartesian coordinate system is simply written as $\mu^{(1)} = \alpha \mathbf{E}$ where \mathbf{E} is the applied field also expressed in the local Cartesian system. In full matrix language, in which the local second rank polarizability tensor is exposed, we can write:

If we neglect pure dephasing, the general tensor element of the third order hyperpolarizability relates to those of the first order polarizability tensor according to

$$(B1.3.12)$$

Here, the linear polarizability, $\alpha_{bc}(\omega_1, \omega_2)$, corresponds to the doorway stage of the 4WM process while to the window stage. We also see the (complex) Raman resonant energy denominator exposed. Of the three energy denominator factors required at third order, the remaining two appear, one each, in the two linear polarizability tensor elements.

In fact, each linear polarizability itself consists of a sum of two terms, one potentially *resonant* and the other *anti-resonant*, corresponding to the two doorway events, D_R and D_N , and the window events, W_R and W_N , described above. The hyperpolarizability chosen in equation (B1.3.12) happens to belong to the generator. As noted, such three-colour generators cannot produce Class I spectroscopies (full quadrature with three colours is not possible). Only the two-colour generators are able to create the Class I Raman spectroscopies and, in any case, only two colours are normally used for the Class II Raman spectroscopies as well.

For linear polarizability elements that are pure real, we see that (from equation (B1.3.12))

$$\text{Im}[\gamma_{abcd}] = \frac{-\gamma_R}{(\omega_1 - \omega_2 - \omega_R)^2 + \gamma_R^2} \alpha_{bc} \alpha_{ad}.$$

When $\omega_1 = \omega_2$ (two colours), this is relevant to the Class I Raman spectroscopies (see section B1.3.2.1). In

this case we expose a Lorentzian Raman lineshape with an HWHM of γ_R . At this point, the notation for the elements of the polarizability tensor suppresses the identity of the Raman transition, so it is now necessary to be more specific.

Consider Raman transitions between thermalized *molecular* eigenstate g (ground) and *molecular* eigenstate f (final). The quantum mechanical expression for α_{bc} responding to colours i and j is the famous (thermalized) *Kramers–Heisenberg equation* [29]

$$(\alpha_{bc}(\omega_i, \omega_j))_{gf} = \sum_n \left(\frac{\langle g|\mu_b|n\rangle\langle n|\mu_c|f\rangle}{\omega_{ng} - \omega_i} + \frac{\langle g|\mu_c|n\rangle\langle n|\mu_b|f\rangle}{\omega_{ng} + \omega_j} \right) \quad (\text{B1.3.13})$$

where the notation on the left-hand side recognizes the Raman transition between molecular eigenstates g and f . The sum on n is over all molecular eigenstates. One should be reminded that the ‘ground state’ should actually be a thermal distribution over Boltzmann weighted states. Thus at the hyperpolarizability level, one would write $\sum_g W_g (\gamma_{abcd})_{gf}$, where W_g is the appropriate Boltzmann weighting factor, $\exp[-\hbar\omega_g/kT]$. This detail is suppressed in what follows.

To branch into electronic, vibrational and rotational Raman spectroscopy, the Born–Oppenheimer (B–O) approximation must be introduced, as needed, to replace the molecular eigenstates as rovibronic products. For example, consider vibrational Raman scattering within the ground electronic state (or, analogously, within any other electronic state). For scattering between vibrational levels v and v' in the ground electronic state, we expand the molecular eigenstate notation to $|g\rangle \equiv |g\rangle|v\rangle$ and $|f\rangle \equiv |g\rangle|v'\rangle$ (the intermediate states, $|n\rangle$, may be left as molecular eigenstates). The curved bracket refers to the electronic eigenstate and the straight bracket to the vibrational states (where until now it referred to the molecular eigenstate). Now equation (B1.3.13) becomes

$$(\alpha_{bc}(\omega_i, \omega_j))_{vv'} = \langle v| \left[\sum_n \left(\frac{\langle g|\mu_b|n\rangle\langle n|\mu_c|g\rangle}{\omega_{ng} - \omega_i} + \frac{\langle g|\mu_c|n\rangle\langle n|\mu_b|g\rangle}{\omega_{ng} + \omega_j} \right) \right] |v'\rangle.$$

We note that the expression in brackets is just the $b c$ tensor element of the *electronic* polarizability in the ground electronic state, $\alpha_{bc}^{\text{el}}(\omega_i, \omega_j)$. Thus

$$(\alpha_{bc}(\omega_i, \omega_j))_{vv'} = \langle v|\alpha_{bc}^{\text{el}}(\omega_i, \omega_j)|v'\rangle. \quad (\text{B1.3.14})$$

Since the vibrational eigenstates of the ground electronic state constitute an orthonormal basis set, the off-diagonal matrix elements in equation (B1.3.14) will vanish unless the ground state electronic polarizability depends on nuclear coordinates. (This is the Raman analogue of the requirement in infrared spectroscopy that, to observe a transition, the electronic dipole moment in the ground electronic state must properly vary with nuclear displacements from

equilibrium.) Indeed such electronic properties do depend on nuclear coordinates, for in the B–O approximation electronic eigenstates are parametrized by the positional coordinates of the constituent nuclei. For matrix elements in vibrational space, these coordinates become variables and the electronic polarizability (or the electronic dipole moment) is expanded in a Taylor series in nuclear displacements. Usually the normal mode approximation is introduced into the vibrational space problem (though it need not be) and the expansion is in terms of the normal displacement coordinates, $\{\Delta Q_i\}$, of the molecule. Thus for the electronic

polarizability we have:

$$\alpha_{bc}^{\text{el}} = (\alpha_{bc}^{\text{el}})_0 + \sum_{i=1} \left(\frac{\partial \alpha_{bc}^{\text{el}}}{\partial Q_i} \right)_0 \Delta Q_i + \frac{1}{2} \sum_{i=1} \sum_{j=1} \left(\frac{\partial^2 \alpha_{bc}^{\text{el}}}{\partial Q_i \partial Q_j} \right)_0 \Delta Q_i \Delta Q_j + \dots \quad (\text{B1.3.15})$$

the leading term of which is unable to promote a vibrational transition because

$$\langle v | (\alpha_{bc}^{\text{el}})_0 | v' \rangle = (\alpha_{bc}^{\text{el}})_0 \langle v | v' \rangle = (\alpha_{bc}^{\text{el}})_0 \delta_{vv'}. \quad (\text{The } v = v' \text{ situation corresponds to Rayleigh scattering for which}$$

this leading term is the principal contributor.) The ΔQ_i in the second term is able to promote the scattering of fundamentals, since $\langle v | \Delta Q_i | v + 1 \rangle$ need not vanish. The $\Delta Q_i \Delta Q_j$ in the third set of terms can cause scattering of the first overtones ($i = j$) and combination states ($i \neq j$), etc, for the subsequent terms. As usual in spectroscopy, point group theory governs the selection rules for such matrix elements. As already noted these are identical to the two-photon selection rules [26], though here in vibrational space.

This linear polarizability theory of Raman scattering [30] forms the basis for bond polarizability theory of the Raman effect. Here the polarizability derivative is discussed in terms of its projection onto bonds of a molecule and the concept of additivity and the transferability of such bond specific polarizability derivatives can be discussed, and even semiquantitatively supported. Further, the vibronic (vibrational–electronic) theory of Raman scattering appears at this level. It introduces the Herzberg–Teller development for the nuclear coordinate dependence of electronic states, therefore that of the electronic transition moments and hence that of the electronic polarizability. This leads to the so-called ‘A’, ‘B’ and ‘C’ terms for Raman scattering, each having a different analytical form for the dispersive behaviour of the Raman cross-section as the exciting light moves from a nonresonant region towards an electronically resonant situation. An early review of these subjects can be found in [31].

For excitation at a wavenumber $\bar{\nu}_1$, ($\bar{\nu}_1 = \omega_1/2\pi c$), and the Raman wavenumber at $\bar{\nu}_R$, the total Raman cross-section for scattering in isotropic media^{3, 4} onto a spherical surface of 4π radians, for all analysing polarizations, for excitation with linearly polarized or unpolarized light, and integration over the Raman line, we have in terms of rotational invariants of the linear polarizability:

$$\sigma_R(\bar{\nu}_1) = \frac{32\pi^3}{9} \left(\frac{e^2}{4\pi\epsilon_0\hbar c} \right)^2 \bar{\nu}_1 (\bar{\nu}_1 \pm \bar{\nu}_R)^3 \{ \Sigma^0 + \Sigma^1 + \Sigma^2 \}. \quad (\text{B1.3.16})$$

The upper sign is for anti-Stokes scattering, the lower for Stokes scattering. The factor in the parentheses is just the fine-structure constant and Σ^0 , Σ^1 , Σ^2 are the three rotationally invariant tensor elements of the hyperpolarizability (or the linear polarizability when pure dephasing is ignored), which are given by:

$$\Sigma^0 = \frac{1}{3} \sum_{\rho, \sigma} \gamma_{\sigma\rho\rho\sigma} \approx \frac{1}{3\Omega} \left| \sum_{\rho} \alpha_{\rho\rho} \right|^2 \quad (\text{B1.3.17})$$

$$\Sigma^1 = \frac{1}{2} \sum_{\rho, \sigma} (\gamma_{\rho\rho\sigma\sigma} - \gamma_{\rho\sigma\rho\sigma}) \approx \frac{1}{4\Omega} \sum_{\rho, \sigma \neq \rho} |\alpha_{\rho\sigma} - \alpha_{\sigma\rho}|^2 \quad (\text{B1.3.18})$$

and

$$\Sigma^2 = \frac{1}{2} \sum_{\rho,\sigma} (\gamma_{\rho\rho\sigma\sigma} + \gamma_{\rho\sigma\rho\sigma}) - \frac{1}{3} \sum_{\rho,\sigma} \gamma_{\rho\sigma\sigma\rho} \approx \frac{1}{4\Omega} \sum_{\rho,\sigma \neq \rho} |\alpha_{\rho\sigma} + \alpha_{\sigma\rho}|^2 + \frac{1}{6\Omega} \sum_{\rho,\sigma} |\alpha_{\rho\rho} - \alpha_{\sigma\sigma}|^2 \quad (\text{B1.3.19})$$

where Ω is the Raman resonant energy denominator, $\omega_1 - \omega_2 - \omega_R + i\gamma_R$. With appropriate algebra, one finds that their sum is given by: $\Sigma^0 + \Sigma^1 + \Sigma^2 = \sum_{\rho,\sigma} \gamma_{\rho\rho\sigma\sigma}$, or in terms of the linear polarizability tensor elements: $\Sigma^0 + \Sigma^1 + \Sigma^2 = \frac{1}{\Omega} \sum_{\rho,\sigma} |\alpha_{\rho\sigma}|^2$.

Experimentally, it is these invariants (equation (B1.3.17), equation (B1.3.18) and equation (B1.3.19)) that can be obtained by scattering intensity measurements, though clearly not by measuring the total cross-section only.

Measurement of the total Raman cross-section is an experimental challenge. More common are reports of the differential Raman cross-section, $d\sigma_R/d\Omega$, which is proportional to the intensity of the scattered radiation that falls within the element of solid angle $d\Omega$ when viewing along a direction that is to be specified [15]. Its value depends on the design of the Raman scattering experiment.

In the appendix, we present the differential Raman scattering cross-section for viewing along any wavevector in the scattering sphere for both linearly and circularly polarized excitation. The more conventional geometries used for exciting and analysing Raman scattering are discussed next.

Suppose the exciting beam travels along X , and is linearly (l) polarized along Z . A popular experimental geometry is to view the scattered light along Y (at $\pi/2$ radians to the plane defined by the wavevector, and the polarization unit vector, e_Z , of the exciting light). One analyses the Z polarized component of the scattered light, called I_{\parallel} (I_Z), and the X polarized component, called I_{\perp} (I_X). (Careful work must properly correct for the finite solid angle of detection.) The two intensities are directly proportional to the differential cross-section given by

$$\left(\frac{d\sigma}{d\Omega}\right)_{\parallel} = 4\pi^2 \left(\frac{e^2}{4\pi\epsilon_0\hbar c}\right)^2 \bar{v}_1 (\bar{v}_1 \pm \bar{v}_R)^3 \left(\frac{5\Sigma^0 + 2\Sigma^2}{15}\right) \quad (\text{B1.3.20})$$

-19-

and

$$\left(\frac{d\sigma}{d\Omega}\right)_{\perp} = 4\pi^2 \left(\frac{e^2}{4\pi\epsilon_0\hbar c}\right)^2 \bar{v}_1 (\bar{v}_1 \pm \bar{v}_R)^3 \left(\frac{5\Sigma^1 + 3\Sigma^2}{30}\right). \quad (\text{B1.3.21})$$

The depolarization ratio is defined as

$$\rho_1 = \frac{I_{\perp}}{I_{\parallel}} \quad (\text{B1.3.22})$$

which for the present case of an orientationally averaged isotropic assembly of Raman scatterers reduces to:

$$\rho_1 = \frac{5\Sigma^1 + 3\Sigma^2}{10\Sigma^0 + 4\Sigma^2}. \quad (\text{B1.3.23})$$

This result is general, for it includes the case where the tensor elements are complex, regardless of whether or

not the hyperpolarizability tensor is built of the linear polarizability.

Another frequent experimental configuration uses naturally (n) polarized incident light, with the same viewing geometry and polarization analysis. Such light may be regarded as polarized equally along Z (as before) and along the viewing axis, Y . Given the I_{\perp} and I_{\parallel} as defined in the linearly polarized experiment, one can reason that now with naturally polarized excitation $I_Z \propto I_{\parallel} + I_{\perp}$ (where the additional I_{\perp} term along Z originates from the Y polarized excitation). Similarly we expect that $I_X \propto 2I_{\perp}$, one I_{\perp} from each of the two excitation polarizations. The differential Raman cross-sections for naturally polarized excitation are defined as

$$\left(\frac{d\sigma}{d\Omega}\right)_X = 4\pi^2 \left(\frac{e^2}{4\pi\epsilon_0\hbar c}\right)^2 \bar{\nu}_1(\bar{\nu}_1 \pm \bar{\nu}_R)^3 \left(\frac{5\Sigma^1 + 3\Sigma^2}{30}\right)$$

and

$$\left(\frac{d\sigma}{d\Omega}\right)_Z = 4\pi^2 \left(\frac{e^2}{4\pi\epsilon_0\hbar c}\right)^2 \bar{\nu}_1(\bar{\nu}_1 \pm \bar{\nu}_R)^3 \left(\frac{10\Sigma^0 + 5\Sigma^1 + 7\Sigma^2}{60}\right).$$

Thus one predicts that the depolarization ratio for excitation with natural light should be

$$\rho_n = \frac{I_X}{I_Z} = \frac{2I_{\perp}}{I_{\parallel} + I_{\perp}} = \frac{2\rho_1}{1 + \rho_1} = \frac{10\Sigma^1 + 6\Sigma^2}{10\Sigma^0 + 5\Sigma^1 + 7\Sigma^2} \quad (\text{B1.3.24})$$

-20-

or

$$\rho_1 = \frac{\rho_n}{2 - \rho_n}$$

two well known relations. Of course, if one does not measure the polarization of the scattered light for either experiment, the detector collects signal from the sum of the differential cross sections, $(\frac{d\sigma}{d\Omega})_{\parallel(Z)} + (\frac{d\sigma}{d\Omega})_{\perp(X)}$.

Similar reasoning shows that were one to view along the X , Y and Z axes and polarization analyse the signal each time, whether excited by linearly or by naturally polarized light, the total intensity should be given by $I_{\text{total}} \propto I_{\parallel} + 2I_{\perp}$. Given [equation \(B1.3.23\)](#), if we add its denominator to twice the numerator we find that $I_{\text{total}} \propto \{\Sigma^0 + \Sigma^1 + \Sigma^2\}$, a reassuring result.

Knowledge of the depolarization ratios allows one to classify easily the Raman modes of a molecule into symmetric and asymmetric vibrations. If a molecule is undergoing a totally symmetric vibration, the depolarization ratio, ρ_1 (ρ_n) will be less than 3/4 (6/7) and we say that the vibration is polarized (p). On the other hand, for asymmetric vibrations, the depolarization ratio will have a value close to 3/4 (6/7) and we say these vibrations are depolarized (dp) [34]. It should be stated that these values for ρ are only valid when the scattered radiation is collected at right angles to the direction of the incident light. If different geometry is used, ρ_1 and ρ_n are accordingly changed (see [the appendix](#)).

An interesting phenomenon called the ‘noncoincidence effect’ appears in the Raman spectroscopies. This is seen when a given Raman band shows a peak position and a bandwidth that differs (slightly) with the

polarization. It can be attributed to varying sensitivity of the different tensor elements to interchromophore interactions.

B1.3.3 RAMAN SPECTROSCOPY IN MODERN PHYSICS AND CHEMISTRY

Raman spectroscopy is pervasive and ever changing in modern physics and chemistry. In this section of the chapter, sources of up-to-date information are given followed by brief discussions of a number of currently employed Raman based techniques. It is impractical to discuss every possible technique and impossible to predict the many future novel uses of Raman scattering that are sure to come, but it is hoped that this section will provide a firm launching point into the modern uses of Raman spectroscopy for present and future readers.

B1.3.3.1 SOURCES OF UP-TO-DATE INFORMATION

There are three very important sources of up-to-date information on all aspects of Raman spectroscopy. Although papers dealing with Raman spectroscopy have appeared and will continue to appear in nearly every major chemical physics–physical chemistry based serial, *The Journal of Raman Spectroscopy* [35] is solely devoted to all aspects, both theoretical and experimental, of Raman spectroscopy. It originated in 1973 and continues to be a constant source of information on modern applications of Raman spectroscopy.

Advances in Infrared and Raman Spectroscopy [36] provides review articles, both fundamental and applied, in the fields

-21-

of both infrared and Raman spectroscopy. This series aims to review the progress in all areas of science and engineering in which application of these techniques has a significant impact. Thus it provides an up-to-date account of both the theory and practice of these two complementary spectroscopic techniques.

The third important source for information on modern Raman spectroscopy are the books cataloguing the proceedings of the *International Conference on Raman Spectroscopy (ICORS)* [37]. *ICORS* is held every two years at various international locations and features hundreds of contributions from leading research groups covering all areas of Raman spectroscopy. Although the published presentations are quite limited in length, they each contain references to the more substantial works and collectively provide an excellent overview of current trends in Raman spectroscopy. A ‘snapshot’ or brief summary of the 1998 conference appears at the end of this chapter.

Through these three serials, a researcher new to the field, or one working in a specialized area of Raman spectroscopy, can quickly gain access to its current status.

B1.3.3.2 SURVEY OF TECHNIQUES

With the theoretical background presented in the previous sections, it is now possible to examine specific Raman techniques. Of the list in [table B1.3.1](#), we briefly discuss and provide references to additional information for the Class I spectroscopies—spontaneous Raman scattering (SR), Fourier transform Raman scattering (FTRS), resonance Raman scattering (RRS), stimulated Raman scattering (SRS), and surface enhanced Raman scattering (SERS)—and in [table B1.3.2](#), the Class II spectroscopies—coherent Raman scattering (CRS), Raman induced Kerr-effect spectroscopy (RIKES), Raman scattering with noisy light, time resolved coherent Raman scattering (TRCRS), impulsive stimulated Raman scattering (ISRS) and higher

order and higher dimensional Raman scattering.

First we discuss some Class I spectroscopies.

(A) SPONTANEOUS RAMAN SCATTERING (SR)

Conventional spontaneous Raman scattering is the oldest and most widely used of the Raman based spectroscopic methods. It has served as a standard technique for the study of molecular vibrational and rotational levels in gases, and for both intra- and inter-molecular excitations in liquids and solids. (For example, a high resolution study of the vibrons and phonons at low temperatures in crystalline benzene has just appeared [38].)

In this earliest of Raman spectroscopies, there is only one incident field (originally sunlight or lines of the mercury lamp; today a single laser source). This is field 1 in the above language and it appears in quadrature in the two generators, and , relevant to SR. [Figure B1.3.2\(a\)](#) shows that these generators lead to four WMEL diagrams: $D_R W_R(A)$, $D_R W_N(B)$, $D_N W_R(A)$, $D_N W_N(B)$. The first is the strongest contributor (it is potentially resonant, R, in both the first and last steps). The last term is the weakest (being nonresonant, N, in both the first and last steps). We note that at the 4WM level, in all four terms, not only is field 1 in quadrature, but field 2 is likewise in quadrature (since for window events A and B, we have $\varepsilon_s = \varepsilon_2$, namely the signal field is conjugate to the action of field 2). Now, since quadrature means photons, the Raman scattering event has destroyed a photon at ω_1 , while it has created a new photon at ω_2 .

-22-

The unique feature in spontaneous Raman spectroscopy (SR) is that field 2 is not an incident field but (at room temperature and at optical frequencies) it is resonantly drawn into action from the zero-point field of the ubiquitous blackbody (bb) radiation. Its active frequency is spontaneously selected (from the infinite colours available in the blackbody) by the resonance with the Raman transition at $\omega_1 - \omega_2 = \omega_R$ in the material. The effective bb field intensity may be obtained from its energy density per unit circular frequency, $\frac{\hbar\omega_s^3}{\pi^2c^3}$, the

Einstein A coefficient at ω_2 . When the polarization field at frequency ω_s , $\mathbf{p}^{(3)}(\omega_s = \omega_1 - \omega_2 - \omega_1)$, produces an electromagnetic field which acts conjugately with this selected blackbody field (at $\omega_s = \omega_2$), the scattered Raman photon is created. Thus, one simply has growth of the blackbody radiation field at ω_2 , since full quadrature removes all oscillatory behaviour in time and all wavelike properties in space.

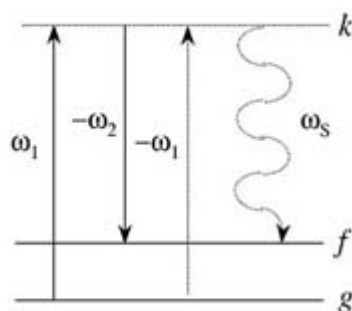
Unlike the typical laser source, the zero-point blackbody field is spectrally ‘white’, providing all colours, ω_2 , that seek out all $\omega_1 - \omega_2 = \omega_R$ resonances available in a given sample. Thus all possible Raman lines can be seen with a single incident source at ω_1 . Such multiplex capability is now found in the Class II spectroscopies where broadband excitation is obtained either by using modeless lasers, or a femtosecond pulse, which on first principles must be spectrally broad [32]. Another distinction between a coherent laser source and the blackbody radiation is that the zero-point field is spatially isotropic. By performing the simple wavevector algebra for SR, we find that the scattered radiation is isotropic as well. This concept of spatial incoherence will be used to explain a certain ‘stimulated’ Raman scattering event in a subsequent section.

For SR, a Class I spectroscopy, there must be a net transfer of energy between light and matter which survives averaging over many cycles of the optical field. Thus, the material must undergo a state population change such that the overall energy (light and matter) may be conserved. In Stokes vibrational Raman scattering ([figure B1.3.3\(a\)](#)), the chromophore is assumed to be in the ground vibrational state $|g\rangle$. The launching of the Stokes signal field creates a population shift from the ground state $|g\rangle$ to an excited vibrational state $|f\rangle$. Conversely, in anti-Stokes vibrational Raman scattering ([figure B1.3.3\(b\)](#)), the chromophore is assumed to be

initially in an excited vibrational state, $|f\rangle$. Thus, the launching of the anti-Stokes field leaves the chromophore in the ground vibrational state, $|g\rangle$. This process is typically weaker than the Stokes process since it requires that an excited vibrational population exist (usually $W_f \ll W_g$). In thermal equilibrium, the intensity of the anti-Stokes frequencies compared to the Stokes frequencies clearly is reduced by the Boltzmann factor, $W_f/W_g = \exp[-\hbar\omega_{fg}/kT]$ [17]. Let us now discuss the apparatus used for the production and detection of the Raman scattered radiation.

-23-

(a)



(b)

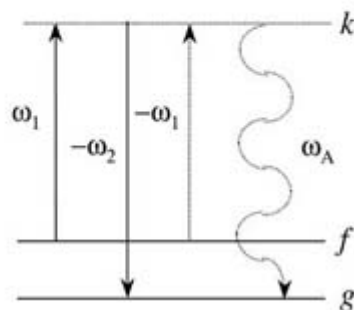


Figure B1.3.3. The full WMEL diagram (of the $D_R W_R$ sort) for spontaneous (or stimulated) (a) Stokes Raman scattering and (b) anti-Stokes Raman scattering. In Stokes scattering, the chromophore is initially in the ground vibrational state, g , and $\omega_1 > \omega_2$. In spontaneous anti-Stokes scattering, the chromophore must be initially in an excited vibrational state, f . Also note that in (b), ω_2 is (arbitrarily) defined as being greater than ω_1 .

-24-

(B) RAMAN INSTRUMENTATION

Dramatic advances in Raman instrumentation have occurred since 1928. At the beginning, various elemental

lamps served as the incident light source and photographic plates were used to detect the dispersed scattered light. Mercury arc lamps were primarily used since they had strong emission lines in the blue region of the visible spectrum (see [equation \(B1.3.16\)](#)). As in all spectroscopies, detection devices have moved from the photographic to the photoelectric.

Even while Raman spectrometers today incorporate modern technology, the fundamental components remain unchanged. Commercially, one still has an excitation source, sample illuminating optics, a scattered light collection system, a dispersive element and a detection system. Each is now briefly discussed.

Continuous wave (CW) lasers such as Ar⁺ and He–Ne are employed in commonplace Raman spectrometers. However laser sources for Raman spectroscopy now extend from the edge of the vacuum UV to the near infrared. Lasers serve as an energetic source which at the same time can be highly monochromatic, thus effectively supplying the single excitation frequency, $\bar{\nu}_1$. The beams have a small diameter which may be focused quite easily into the sample and are convenient for remote probing. Finally, almost all lasers are linearly polarized, which makes measurements of the depolarization ratio, [equation \(B1.3.22\)](#), relatively straightforward.

The laser beam is typically focused onto the sample with a simple optical lens system (though microscopes are also used). The resultant scattered radiation is often collected at 90° from the incident beam propagation and focused onto the entrance of a monochromator. The monochromator removes extraneous light from the Raman radiation by use of a series of mirrors and optical gratings. Two features of the monochromator for maximum resolution are the slit width and, for monochromatic detection, the scanning speed. The resolution of the Raman spectrum is optimized by adjusting the slit width and scanning rate, while still maintaining a strong signal intensity.

After passing through the monochromator, the signal is focused onto the detection device. These devices range from photomultiplier tubes (PMTs) in which the signal is recorded at each frequency and the spectrum is obtained by scanning over a selected frequency range, to multichannel devices, such as arrays of photodiodes and charge coupling devices, which simultaneously detect the signal over a full frequency range. One may choose a specific detection device based on the particulars of an experiment. Sensitive detection of excited vibrational states that are produced in the Class I Raman spectroscopies is an alternative that can include acoustic detection of the heat released and resonance enhanced multiphoton ionization (REMPI). For special applications microspectroscopic techniques and fibre optic probes ('optodes') are used.

This basic instrumentation, here described within the context of spontaneous Raman scattering, may be generalized to most of the other Raman processes that are discussed. Specific details can be found in the citations.

(C) FOURIER TRANSFORM RAMAN SCATTERING (FTRS)

Normal spontaneous Raman scattering suffers from lack of frequency precision and thus good spectral subtractions are not possible. Another limitation to this technique is that high resolution experiments are often difficult to perform [39]. These shortcomings have been circumvented by the development of Fourier transform (FT) Raman spectroscopy [40]. FT Raman spectroscopy employs a long wavelength laser to achieve viable interferometry,

typically a Nd:YAG operating at 1.064 μm . The laser radiation is focused into the sample where spontaneous Raman scattering occurs. The scattered light is filtered to remove backscatter and the Raman light is sent into a Michelson interferometer. An interferogram is collected and detected on a near-infrared detector (typically

an N₂ cooled Ge photoresist). The detected signal is then digitized and Fourier transformed to obtain a spectrum [41]. This technique offers many advantages over conventional Raman spectroscopy. The 1.064 μm wavelength of the incident laser is normally far from electronic transitions and reduces the likelihood of fluorescence interference. Also near-IR radiation decreases sample heating, thus higher powers can be tolerated [42]. Since the signal is obtained by interferometry, the FT instrument records the intensity of many wavelengths. Such simultaneous detection of a multitude of wavelengths is known as the multiplex advantage, and leads to improved resolution, spectral acquisition time and signal to noise ratio over ordinary spontaneous Raman scattering [39]. The improved wavelength precision gained by the use of an interferometer permits spectral subtraction, which is effective in removing background features [42]. The interferogram is then converted to a spectrum by Fourier transform techniques using computer programs. By interfacing the Raman setup with an FT-IR apparatus, one has both IR and Raman capabilities with one instrument.

Other than the obvious advantages of reduced fluorescence and high resolution, FT Raman is fast, safe and requires minimal skill, making it a popular analytic tool for the characterization of organic compounds, polymers, inorganic materials and surfaces and has been employed in many biological applications [41].

It should be noted that this technique is not without some disadvantages. The blackbody emission background in the near IR limits the upper temperature of the sample to about 200°C [43]. Then there is the ν^4 dependence of the Raman cross-section (equation (B1.3.16) and equation (B1.3.20)–equation (B1.3.21)) which calls for an order of magnitude greater excitation intensity when exciting in the near-IR rather than in the visible to produce the same signal intensity [39].

(D) RESONANCE RAMAN SCATTERING (RRS)

As the incident frequencies in any Raman spectroscopy approach an electronic transition in the material, the $D_R W_R$ term in Raman scattering is greatly enhanced. One then encounters an extremely fruitful and versatile branch of spectroscopy called resonance Raman scattering (RRS). In fact, it is fair to say that in recent years RRS (Class I or Class II) has become the most popular form of Raman based spectroscopy. On the one hand, it offers a diagnostic approach that is specific to those subsystems (even minority components) that exhibit the resonance, even the very electronic transition to which the experiment is tuned. On the other hand, it offers a powerful tool for exploring potential-energy hypersurfaces in polyatomic systems. It forms the basis for many time resolved resonant Raman spectroscopies (TRRRSs) that exploit the non-zero vibronic memory implied by an electronic resonance. It has inspired the time-domain theoretical picture of RRS which is formally the appropriate transform of the frequency domain picture [44, 45]. Here the physically appealing picture arises in which the two-step doorway event (D_R) prepares (vertically upward) a vibrational wave-packet that moves (propagates) on the upper electronic state potential energy hypersurface. In the window stage (W_R) of the 4WM event, this packet projects (vertically downward), accordingly to its lingering time on the upper surface, back onto the ground state to complete the third order induced optical polarization that leads to the new fourth wave.

RRS has also introduced the concept of a ‘Raman excitation profile’ (REP_{*j*} for the *j*th mode) [46, 47, 48, 49, 50 and 51]. An REP_{*j*} is obtained by measuring the resonance Raman scattering strength of the *j*th mode as a function of the excitation frequency [52, 53]. How does the scattering intensity for a given (the *j*th) Raman active vibration vary with excitation frequency within an electronic absorption band? In turn, this has led to transform theories that try to predict

the REP_{*j*} from the ordinary absorption band (ABS), or the reverse. Thus one has the so-called forward transform, ABS → REP_{*j*}, and the inverse transform, REP_{*j*} → ABS [54, 55 and 56]. The inverse transform is a formal method that transforms an observed REP_{*j*} into the electronic absorption band that is responsible for resonantly scattering mode *j*. This inverse transform raises theoretical issues concerning the frequently

encountered problem of phase recovery of a complex function (in this case the complex Raman susceptibility), knowing only its amplitude [57].

One group has successfully obtained information about potential energy surfaces without measuring REPs. Instead, easily measured second derivative absorption profiles are obtained and linked to the full RRS spectrum taken at a single incident frequency. In this way, the painstaking task of measuring a REP is replaced by carefully recording the second derivative of the electronic absorption spectrum of the resonant transition [58, 59].

The fitting parameters in the transform method are properties related to the two potential energy surfaces that define the electronic resonance. These curves are obtained when the two hypersurfaces are cut along the j th normal mode coordinate. In order of increasing theoretical sophistication these properties are: (i) the relative position of their minima (often called the displacement parameters), (ii) the force constant of the vibration (its frequency), (iii) nuclear coordinate dependence of the electronic transition moment and (iv) the issue of mode mixing upon excitation—known as the Duschinsky effect—requiring a multidimensional approach.

We have seen how, by definition, all Raman spectroscopies must feature the difference frequency resonance that appears following the two-step doorway stage of the 4WM process. Basically, RRS takes advantage of achieving additional resonances available in the two remaining energy denominator factors found at third order (and, of course, still more at higher order). The two remaining energy factors necessarily involve an algebraic sum of an odd number of optical frequencies (one for the first step in the doorway stage, and three for the initial step in the window event). Since the algebraic sum of an odd number of optical frequencies must itself be optical, these additional resonances must be at optical frequencies. Namely, they must correspond to electronic transitions, including (in molecules) their dense rotational– (or librational–) vibrational substructure. The literature is filled with a great many interesting RRS applications, extending from resonances in the near-IR (dyes and photosynthetic pigments for example) to the deep UV (where backbone electronic resonances in proteins and nucleic acids are studied). These increasingly include TRRRS in order to follow the folding/unfolding dynamics of substructures (through the chromophore specificity of RRS) in biologically important molecules [60, 61 and 62].

The reader must turn to the literature to amplify upon any of these topics. Here we return to the two-colour generator/WMEL scheme to see how it easily can be adapted to the RRS problem.

Let us consider RRS that contains both of the available additional resonances, as is normally the case (though careful choices of colours and their time sequence can isolate one or the other of these). First, we seek out the doorway events that contain not only the usual Raman resonance after the two fields, 1 and 2, have acted conjugately, but also the new resonance that appears after the first field has acted. The appropriate doorway generators remain $\epsilon_1\epsilon_2^*$ and $\epsilon_2^*\epsilon_1$, in order to retain the Raman resonance. There are now two fully resonant doorway WMEL diagrams, which we shall call $D_{RR}(A_{12*})$ and $D_{RR}(B_{12*})$. These diagrams are shown in [figure B1.3.4](#) in which the full manifold of sublevels for each electronic state is intimated. In doorway channel A, vibrational coherences are produced in the ground electronic state g , as usual, but now they are enhanced by the electronic resonance in step 1 (ϵ_1). However, in doorway channel B, vibrational coherences are produced in the excited electronic state, e , as well. Interestingly, since A is a ket–ket event and B is a ket–bra event (which differ by an overall sign), these two coherences must differ in

phase by π (180°). This may be important in any 4WM experiment that is phase sensitive (heterodyned Class II) and in which the window event does not reverse this phase difference.

Figure B1.3.4. The two fully resonant doorway stages for resonance Raman scattering (RRS), in which the manifold of vibrational sublevels for each electronic state is indicated. (a) Doorway stage $D_{RR}(A_{12*})$, in which a vibrational coherence is produced in the ground electronic state, g . It is a ket–ket evolution. (b) Doorway stage $D_{RR}(B_{12*})$, where a vibrational coherence is created in the excited electronic state, e . It is a ket–bra evolution. The coherences in both doorway stages are enhanced by the electronic resonance in their identical step 1 (generated by ϵ_1).

Frequently, femtosecond pulses are used in such electronically resonant spectroscopy. Such pulses usually have near-transform limited bandwidths and can spectrally embrace a fair range of vibrational coherences. Thus, even when a single central colour is chosen to define the femtosecond exciting pulse, actually a broad band of colours is available to provide a range for both ω_1 and ω_2 . Instead of preparing single well-defined vibrational coherences using sharply

-28-

defined colours, now vibrational wavepackets are made in both the ground (channel A) and excited states (channel B). These must evolve in time, for they are not eigenstates in either potential energy hypersurface. The nature of the 4WM signal is sensitive to the location of the wavepacket in either hypersurface at the time the window event takes place. Such wavepackets, prepared from a spectrum of colours contained coherently within a femtosecond pulse, are termed ‘impulsively’ prepared (see impulsive Raman scattering for further details).

Whenever coherences in the upper manifold are particularly short lived, doorway channel A will dominate the evolution of the polarization at least at later times. In that case, the fringes seen in the 4WM signal as the time between the doorway and window stages is altered (with a delay line) reflect those Raman frequencies in the ground state that can be spectrally embraced by the femtosecond pulse. Then the Fourier transform of the fringes leads to the conventional spontaneous RRS of the ground state. Indeed, in the absence of electronic resonance, channel B reverts to a purely nonresonant doorway event (D_N) and only channel A reveals Raman resonances—those in the electronic ground state [62].

Whatever the detection technique, the window stage of the 4WM event must convert these evolved vibrational wavepackets into the third order polarization field that oscillates at an ensemble distribution of optical frequencies. One must be alert to the possibility that the window event after doorway channel B may involve resonances from electronic state manifold e to some higher manifold, say r . Thus channel B followed by an ϵ_3 (ket) or a ϵ_3^* (bra) event might be enhanced by an e -to- r resonance. However, it is normal to confine the window event to the e -to- g resonances, but this is often simply for lack of substantive e -to- r information. Given that the third field action can be of any available colour, and considering only g -to- e resonances, one has, for any colour ω_3 , only two possibilities following each doorway channel. Channel A should be followed by an ϵ_3 (ket up) or an ϵ_3^* (bra up) event; channel B should be followed by a ϵ_3 (bra down) or a ϵ_3^* (ket down) event.

Before looking more closely at these, it is important to recognize another category of pump–probe Raman experiments. These are often referred to as ‘transient Raman’ pump–probe studies. In these, a given system is ‘pumped’ into a transient condition such as an excited vibronic state, or a photochemical event such as dissociation or radical formation [63, 64 and 65]. Such pumping can be achieved by any means—even by high energy radiation [66, 67 and 68]—though normally laser pumping is used. The product(s) formed by the pump step is then studied by a Raman probe (often simply spontaneous Raman, sometimes CARS). Since the transient state is normally at low concentrations, the Raman probing seeks out resonant enhancement, as we are describing, and also means must be taken to stay away from the luminescence background that is invariably caused by the pump event. Often, time gated Fourier transform Raman in the near-IR is employed

(that spectral region being relatively free of interfering fluorescence) and yet upper *e*-to-*r* type resonances may still be available for RRS. Since transient systems are ‘hot’ by their very nature, both anti-Stokes as well as Stokes spontaneous Raman scattering can be followed to time the vibrational relaxation in transient excited states (see [69, 70]).

Returning to the original pump–probe RRS, it is a simple matter to complete the 4WM WMEL diagrams for any proposed RRS. Usually RRS experiments are of the full quadrature sort, both spontaneous RRS as well as homodyne detected femtosecond RRS. The latter fit most pump–probe configurations.

Let us, for example, present the full WMEL diagrams for full quadrature RRS with two colours, 1 and 2. (Recall that three colours cannot lead to full *Q* at the 4WM level.) Given the $\epsilon_1 \epsilon_2^*$ doorway generator for channels A and B, the generator for the first step of the window event must either be ϵ_1^* or ϵ_2 , and the corresponding signal must conjugate

-29-

either with ϵ_2^* ($\epsilon_s = \epsilon_2$) or with ϵ_1 ($\epsilon_s^* = \epsilon_1^*$), respectively. In the former case, field 1 will have acted twice (and conjugately) to help produce the third order polarization and signal field directed along k_2 . In the latter case, field 2 has acted twice (and conjugately) to help produce a third order polarization along $-k_1$. Of the two fields, the most intense clearly is the candidate for the twice acting field. The weaker field, then, is examined for the signal.

In addition, we have asked that a third resonance exist in the window stage. For channel A this requires that the window event begin either with *ket up* using ϵ_2 , or with *bra up* using ϵ_1^* , while the window event following channel B should begin either with *bra down* using ϵ_2 or with *ket down* using ϵ_1^* . The corresponding four WMEL diagrams are shown in figure B1.3.5.

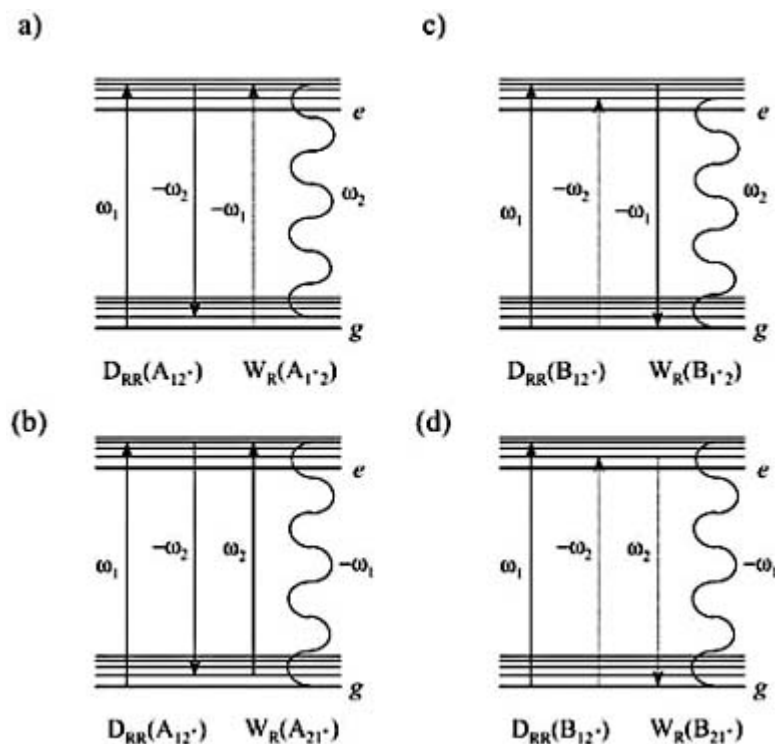


Figure B1.3.5. Four WMEL diagrams for fully resonant Raman scattering (RRS). Diagrams (a) and (b) both have doorway stage $D_{RR}(A_{12}^*)$ (Figure B1.3.4(a)), in which a vibrational coherence is created in the ground electronic state, *g*. For the window event in (a), field 1 promotes the bra from the ground electronic state, *g*, to

an excited electronic bra state, e . In this window stage $W_R(A_{1*2})$, the third field action helps produce the third order polarization, which in turn gives rise to a signal field with frequency ω_2 . For the window event $W_R(A_{21*})$ in (b), field 2 acts to promote the ket in the ground electronic state into one in the excited electronic state. Now a signal field with frequency $-\omega_1$ is created. Both diagrams (c) and (d) have doorway stage $D_{RR}(B_{12*})$ (figure B1.3.4(b)), in which a vibrational coherence is created in the excited electronic state, e . At the window stage $W_R(B_{1*2})$ (c) field 1 demotes the ket from one in electronic state e to one in the ground electronic state g . A consequent third order polarization leads to a signal field at frequency ω_2 . At the window stage $W_R(B_{21*})$ (d) the bra is demoted by field 2 to the ground electronic state to produce a signal field at frequency $-\omega_1$. Full quadrature is achieved in all four diagrams.

-30-

In these WMEL diagrams, the outcome of the collapse of the polarization, the second step of the window stage (the curvy line segment), depends on the phase difference between the induced s th order polarization and the new $(s + 1)$ th electromagnetic field [7]. If this difference is 0° , then the energy contained in the polarization is fully converted into population change in the medium (pure Class I spectroscopies). At third order, channel A populates vibrational levels in the ground electronic state; channel B populates vibrational levels in the excited electronic state. However if this phase difference is 90° , the energy is converted fully into the new $(s + 1)$ th electromagnetic field and the material is unchanged (pure Class II spectroscopies). If the phase difference lies between 0 and 90° , as is almost always the case, then both outcomes occur to some extent and the experimentalist may perform either a Class I or a Class II spectroscopy.

As we have already seen, in such a full quadrature situation, phase matching is automatic (the signal is collinear with either of the incident fields), so the experiment then measures changes in one or more properties of one of the incident fields—either the first appearing light pulse, or the second appearing light pulse. These are distinguished both in order of appearance and by their wavevector of incidence. At full quadrature, the obvious property to measure is simply the intensity (Class I) as one (or more) of the time parameters is changed. With this example, it should be a simple matter to explore WMEL diagrams for any other RRS spectroscopies, in particular the nonquadrature ones.

(E) STIMULATED RAMAN SCATTERING (SRS)

Since the inception of the laser, the phenomenon called stimulated Raman scattering (SRS) [71, 72, 73 and 74] has been observed while performing spontaneous Raman experiments. Stimulated Raman scattering is inelastic light scattering by molecules caused by presence of a light field that is stronger than the zero-point field of the blackbody. Thus SRS overtakes the spontaneous counterpart. SRS, like SR, is an active Class I spectroscopy. For the case of Stokes scattering, this light field may be a second laser beam at frequency ω_2 , or it may be from the polarization field that has built up from the spontaneous scattering event. Typically in an SRS experiment one also observes frequencies other than the Stokes frequency: those at $\omega_1 - 2\omega_R$, $\omega_1 - 3\omega_R$, $\omega_1 - 4\omega_R$ etc which are referred to as higher order Stokes stimulated Raman scattering and at $\omega_1 + \omega_R$ (anti-Stokes) and $\omega_1 + 2\omega_R$, $\omega_1 + 3\omega_R$, $\omega_1 + 4\omega_R$ (higher order anti-Stokes stimulated Raman scattering) [19]. This type of scattering may arise only after an appreciable amount of Stokes scattering is produced (unless the system is not at thermal equilibrium).

First order stimulated Stokes scattering experiences an exponential gain in intensity as the fields propagate through the scattering medium. This is given by the expression [75]

$$I_S(z) = I_S(0) e^{g_S I_L z}$$

where z is the path length, I_L is the intensity of the incident radiation and $I_S(0)$ is the intensity of the quantum noise associated with the vacuum state [15], related to the zero-point bb field. g_S is known as the stimulated

Raman gain coefficient, whose maximum occurs at exact resonance, $\omega_L - \omega_S = \omega_R$. For a Lorentzian lineshape, the maximum gain coefficient is given by

-31-

$$g_{S_{\max}} = \frac{8\pi^2 c^2 N}{\hbar \omega_L \omega_S^2 n_S n_L} \frac{d\sigma_R}{d\Omega} \frac{1}{\gamma_R}$$

where ω_S is the frequency of the Stokes radiation and γ_R is the HWHM of the Raman line (in units of circular frequency) [75]. This gain coefficient is seen to be proportional to the spontaneous differential Raman cross section ($d\sigma_R/d\Omega$), (the exact nature of which depends on experimental design (see [equation \(B1.3.20\)](#) and [equation \(B1.3.21\)](#)) and also to the number density of scatterers, N . This frequency dependent gain coefficient may also be written in terms of the third order nonlinear Raman susceptibility, $\chi^{(3)}(\omega_S)$. As with SR, only the imaginary part of $\chi^{(3)}(\omega_S)$ contributes to Stokes amplification (the real part accounts for intensity dependent (nonlinear) refractive indices). For the definition of $\chi^{(3)}$ used here ([equation \(B1.3.11\)](#) and [equation \(B1.3.12\)](#)), their relation is given by

$$g_S(\omega_S) = \frac{-2\omega_S}{\epsilon_0^2 n_S n_L c^2} \text{Im } \chi_{AAAA}^{(3)}(\omega_S)$$

where the magnitude of $\text{Im } \chi^{(3)}(\omega_S)$ is negative, thus leading to a positive gain coefficient [33]. For this expression, only the component of the scattered radiation parallel to the incident light is analysed.

Once an appreciable amount of Stokes radiation is generated, enough scatterers are left in an excited vibrational state for the generation of anti-Stokes radiation. Also, the Stokes radiation produced may now act as incident radiation in further stimulated Raman processes to generate the higher order Stokes fields. Although the Stokes field is spatially isotropic, scattered radiation in the forward and backward directions with respect to the incident light traverses the longest interaction length and thus experiences a significantly larger gain (typically several orders of magnitude larger than in the other directions [33]). Thus the first and higher order Stokes frequencies lie along the direction of the incident beam.

This is not the case for stimulated anti-Stokes radiation. There are two sources of polarization for anti-Stokes radiation [17]. The first is analogous to that in [figure B1.3.3\(b\)](#) where the action of the blackbody ($-\omega_2$) is replaced by the action of a previously produced anti-Stokes wave, with frequency ω_A . This radiation actually experiences an attenuation since the value of $\text{Im } \chi^{(3)}(\omega_A)$ is positive (leading to a negative ‘gain’ coefficient). This is known as the stimulated Raman loss (SRL) spectroscopy [76]. However the second source of anti-Stokes polarization relies on the presence of Stokes radiation [17]. This anti-Stokes radiation will emerge from the sample in a direction given by the wavevector algebra: $\mathbf{k}_A = 2\mathbf{k}_l - \mathbf{k}_S$. Since the Stokes radiation is isotropic ($-\mathbf{k}_S$), the anti-Stokes radiation (and subsequent higher order radiation) is emitted in the form of concentric rings.

(F) SURFACE ENHANCED RAMAN SCATTERING (SERS)

We have seen that the strength of Raman scattered radiation is directly related to the Raman scattering cross-section (σ_R). The fact that this cross-section for Raman scattering is typically much weaker than that for absorption ($\sigma_R \ll \sigma_{\text{abs}}$) limits conventional SR as a sensitive analytical tool compared to (linear) absorption techniques. The complication of fluorescence in the usual Raman techniques of course tends to decrease the signal-to-noise ratio.

It was first reported in 1974 that the Raman spectrum of pyridine is enhanced by many orders of magnitude when

-32-

the pyridine is adsorbed on silver metal [77]. This dramatic increase in the apparent Raman cross-section in the pyridine/silver system was subsequently studied in more detail [78, 79]. The enhancement of the Raman scattering intensity from molecules adsorbed to surfaces (usually, though not exclusively, noble metals) has come to be called the surface enhanced Raman spectroscopy (SERS). Since these early discoveries, SERS has been intensively studied for a wide variety of adsorbate/substrate systems [80, 81, 82 and 83]. The pyridine/silver system, while already thoroughly studied, still remains a popular choice among investigators both for elucidating enhancement mechanisms and for analytical purposes. The fluorescence of the adsorbed molecules does not experience similarly strong enhancement, and often is actually quenched. So, since the signal is dominated by the adsorbate molecules, fluorescence contamination is relatively suppressed.

The metal substrate evidently affords a huge ($\sim 10^{10}$ and even as high as 10^{14} [84, 85]) increase in the cross-section for Raman scattering of the adsorbate. There are two broad classes of mechanisms which are said to contribute to this enhancement [86, 87 and 88]. The first is based on electromagnetic effects and the second on 'chemical' effects. Of these two classes the former is better understood and, for the most part, the specific mechanisms are agreed upon; the latter is more complicated and is less well understood. SERS enhancement can take place in either physisorbed or chemisorbed situations, with the chemisorbed case typically characterized by larger Raman frequency shifts from the bulk phase.

The substrate is, of course, a necessary component of any SERS experiment. A wide variety of substrate surfaces have been prepared for SERS studies by an equally wide range of techniques [87]. Two important substrates are electrochemically prepared electrodes and colloidal surfaces (either deposited or in solution).

A side from the presence of a substrate, the SERS experiment is fundamentally similar to the standard conventional Raman scattering experiment. Often a continuous wave laser, such as an argon ion laser, is used as the excitation source, but pulsed lasers can also be used to achieve time resolved SERS. Also, as in conventional Raman scattering, one can utilize pre-resonant or resonant conditions to perform resonant SERS (often denoted SERRS for surface enhanced resonant Raman scattering). SERRS combines the cross-section enhancement of SERS with the electronic resonance enhancement of resonance Raman scattering. In fact, through SERRS, one can achieve extraordinary sensitivity, with reports appearing of near-single-molecule-based signals [84, 85, 89].

We now move on to some Class II spectroscopies.

(G) COHERENT RAMAN SCATTERING (CRS)

The major Class II Raman spectroscopy is coherent Raman scattering (CRS) [90, 91, 92 and 93]. It is an extremely important class of nearly degenerate four-wave mixing spectroscopies in which the fourth wave (or signal field) is a result of the coherent stimulated Raman scattering. There are two important kinds of CRS distinguished by whether the signal is anti-Stokes shifted (to the blue) or Stokes shifted (to the red). The former is called CARS (coherent anti-Stokes Raman scattering) and the latter is called CSRS (coherent Stokes Raman scattering). Both CARS and CSRS involve the use of two distinct incident laser frequencies, ω_1 and ω_2 ($\omega_1 > \omega_2$). In the typical experiment ω_1 is held fixed while ω_2 is scanned. When $\omega_1 - \omega_2$ matches a Raman frequency of the sample a resonant condition results and there is a strong gain in the CARS or CSRS signal intensity. The complete scan of ω_2 then traces out the CARS or CSRS spectrum of the sample. (Figure B1.3.2 (b)) shows representative WMEL diagrams for the CARS and CSRS processes.) There are, in actuality, 48 WMEL diagrams (including the nonresonant contributions) that one must

consider for either of these two processes. These have been displayed in the literature ([98] (CARS) and [99] (CSRS)). For both processes, a pair of field–matter interactions produces a vibrational coherence between states $|g\rangle$ and $|f\rangle$ (see D_R and D_N of [figure B1.3.2\(b\)](#)). For the CARS process, the third field, having frequency ω_1 , acts in phase (the same Fourier component) with the first action of ω_1 to produce a polarization that is anti-Stokes shifted from ω_1 (see $W_R(E)$ and $W_N(F)$ of [figure B1.3.2\(b\)](#)). For the case of CSRS the third field action has frequency ω_2 and acts in phase with the earlier action of ω_2 ($W_R(C)$ and $W_N(D)$ of [figure B1.3.2\(b\)](#)). Unlike the Class I spectroscopies, no fields in CARS or CSRS (or any homodyne detected Class II spectroscopies) are in quadrature at the polarization level. Since homodyne detected CRS is governed by the modulus square of $\chi^{(3)}$ ($I_{\text{CRS}} \propto |\chi^{(3)}|^2$), its lineshape is not a symmetric lineshape like those in the Class I spectroscopies, but it depends on both the resonant and nonresonant components of $\chi^{(3)}$, $\chi_R^{(3)}$ and $\chi_N^{(3)}$, respectively. Thus

$$|\chi^{(3)}|^2 = |\chi_R^{(3)} + \chi_N^{(3)}|^2 = |\chi_R^{(3)}|^2 + \chi_N^{(3)}(\chi_R^{(3)} + \chi_R^{*(3)}) + (\chi_N^{(3)})^2$$

and one is faced with both an absorptive component $|\chi_R^{(3)}|^2$ and a dispersive component, $\chi_N^{(3)}(\chi_R^{(3)} + \chi_R^{*(3)})$. $\chi_N^{(3)}$ can, to a very good approximation, be taken to be a (pure real) constant over the width of the Raman line.) As a result, the CRS lineshape is asymmetric and more complicated due to this nonresonant background interference.

The primary advantages of CARS and CSRS include an inherently stronger signal than spontaneous Raman scattering (the incident fields are stronger than the zero-point blackbody fields) and one that is directional (phase matched). These characteristics combine to give the technique a much lower vulnerability to sample fluorescence and also an advantage in remote sensing. For CARS, fluorescence is especially avoided since the signal emerges to the blue of the incident laser frequencies and fluorescence must be absent (unless it were biphotonically induced). The primary disadvantage of CARS and CSRS is the interference of the nonresonant part of $\chi^{(3)}$ in the form of the dispersive cross-term. A class of techniques called polarization CRS utilizes the control over the polarization of the input beams to suppress the nonresonant background interference [94, 95]. On the other hand, the background interference necessarily carries information about the nonresonant component of the electric susceptibility which is sometimes a sought after quantity.

(H) RAMAN INDUCED KERR EFFECT SPECTROSCOPY (RIKES)

The nonresonant background prevalent in CARS experiments (discussed above), although much weaker than the signals due to strong Raman modes, can often obscure weaker modes. Another technique which can suppress the nonresonant background signal is Raman induced Kerr-effect spectroscopy or RIKES [96, 97].

A RIKES experiment is essentially identical to that of CW CARS, except the probe laser need not be tunable. The probe beam is linearly polarized at 0° (\rightarrow), while the polarization of the tunable pump beam is controlled by a linear polarizer and a quarter waveplate. The pump and probe beams, whose frequency difference must match the Raman frequency, are overlapped in the sample (just as in CARS). The strong pump beam propagating through a nonlinear medium induces an anisotropic change in the refractive indices seen by the weaker probe wave, which alters the polarization of a probe beam [96]. The signal field is polarized orthogonally to the probe laser and any altered polarization may be detected as an increase in intensity transmitted through a crossed polarizer. When the pump beam is linearly polarized at 45° (\nearrow), contributions from the nonlinear background susceptibility exist ($\chi_{\text{eff}}^{(3)} = 3[\chi_{\text{AABB}}^{(3)} + \chi_{\text{ABAB}}^{(3)}]$). If the quarter-wave plate is adjusted to give circularly polarized light (\odot), the nonresonant background will disappear

, provided [19].

A unique feature of this Class II spectroscopy is that it occurs in full quadrature, and thus the phase-matching condition is automatically fulfilled for every propagation direction and frequency combination (for isotropic media). Characteristic WMEL diagrams for this process are given by diagrams B and G of [figure B1.3.2\(a\)](#). From these diagrams, one may notice that the frequency of the signal field is identical to that of one of the incident fields, thus one must carefully align the crossed polarizer to eliminate contamination by the probe beam.

A common technique used to enhance the signal-to-noise ratio for weak modes is to inject a local oscillator field polarized parallel to the RIKE field at the detector. This local oscillator field is derived from the probe laser and will add coherently to the RIKE field [96]. The relative phase of the local oscillator and the RIKE field is an important parameter in describing the optical heterodyne detected (OHD)-RIKES spectrum. If the local oscillator at the detector is in phase with the probe wave, the heterodyne intensity is proportional to $\text{Re}\{\chi_{\text{eff}}^{(3)}\}$. If the local oscillator is in phase quadrature with the probe field, the heterodyne intensity becomes proportional to $\text{Im}\{\chi_{\text{eff}}^{(3)}\}$. Thus, in addition in to signal-to-noise improvements, OHD-RIKES, being a heterodyne method, demonstrates a phase sensitivity not possible with more conventional homodyne techniques.

Still another spectroscopic technique used to suppress the nonresonant background is ASTERISK. The setup is identical to a conventional CARS experiment, except three independent input fields of frequencies, ω_1 , ω_2 and ω_3 are used. The relative polarization configuration (not wavevectors) for the *three* incident fields and the analyser (e_s) is shown in [figure B1.3.6](#) where the signal generated at ω_s will be polarized in the x -direction. (Both θ and ϕ are defined to be positive angles, as denoted in [figure B1.3.6](#).) The recorded spectra will be relatively free of the nonresonant background (for $\theta = \phi \simeq 45^\circ$, the detected intensity will be proportional to $|\chi_{\text{ABAB}}^{(3)} - \chi_{\text{ABBA}}^{(3)}|^2$; however one must satisfy the phase matching condition: $\Delta\mathbf{k} = \mathbf{k}_1 - \mathbf{k}_2 + \mathbf{k}_3 - \mathbf{k}_s$. In transparent materials, a greater than three-orders-of-magnitude reduction of the nonresonant background occurs as compared to its 25–100-fold suppression by RIKES [96].

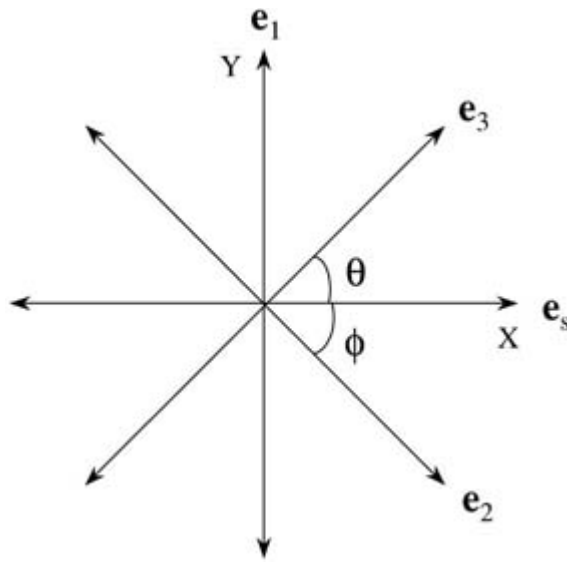


Figure B1.3.6. The configuration of the unit polarization vectors \mathbf{e}_1 , \mathbf{e}_2 , \mathbf{e}_3 and \mathbf{e}_s in the laboratory Cartesian basis as found in the ASTERISK technique.

(i) RAMAN SCATTERING WITH NOISY LIGHT

In the early 1980s, it was shown that noisy light offers a time resolution of the order of the noise correlation time, τ_c , of the light (typically tens to several hundreds of femtoseconds)—many orders of magnitude faster than the temporal profile of the light (which is often several nanoseconds, but in principle can be CW) [100, 101 and 102]. A critical review of many applications of noisy light (including CRS) is given by Kobayashi [103]. A more recent review by Kummrow and Lau [104] contains an extensive listing of references.

A typical noisy light based CRS experiment involves the splitting of a noisy beam (short autocorrelation time, broadband) into identical twin beams, B and B', through the use of a Michelson interferometer. One arm of the interferometer is computer controlled to introduce a relative delay, τ , between B and B'. The twin beams exit the interferometer and are joined by a narrowband field, M, to produce the CRS-type third order polarization in the sample ($\omega_B - \omega_M \approx \omega_R$). The delay between B and B' is then scanned and the frequency-resolved signal of interest is detected as a function of τ to produce an *interferogram*. As an interferometric spectroscopy, it has come to be called $I^{(2)}$ CRS ($I^{(2)}$ CARS and $I^{(2)}$ CSRS), in which the ' $I^{(2)}$ ' refers to the two, twin incoherent beams that are interferometrically treated [105]. The theory of $I^{(2)}$ CRS [106, 107] predicts that the so-called radiation difference oscillations (RDOs) should appear in the 'monochromatically' detected $I^{(2)}$ CRS interferogram with a frequency of $\Delta \equiv \omega_M + 2\omega_R - \omega_D$, where ω_D is the detected frequency, ω_M is the narrowband frequency and ω_R the Raman (vibrational) frequency. Since ω_D and ω_M are known, ω_R may be extracted from the experimentally measured RDOs. Furthermore, the dephasing rate constant, γ_R , is determined from the observed decay rate constant, γ , of the $I^{(2)}$ CRS interferogram. Typically for the $I^{(2)}$ CRS signal $\omega_D \approx \omega_M + 2\omega_R$ and thus $\Delta \approx 0$. That is, the RDOs represent strongly *down-converted* (even to zero frequency) Raman frequencies. This down-conversion is one of the chief advantages of the $I^{(2)}$ CRS technique, because it allows for the characterization of vibrations using optical fields but with a much smaller interferometric sampling rate than is needed in FT-Raman or in FT-IR. More explicitly, the Nyquist sampling rate criterion for the

RDOs is much smaller than that for the vibration itself, not to mention that for the near-IR FT-Raman technique already discussed. This is particularly striking for high energy modes such as the C–H vibrations [108]. Modern applications of $I^{(2)}$ CRS now utilize a 'two-dimensional' time–frequency detection scheme

which involves the use of a CCD camera to detect an entire $I^{(2)}$ CARS spectrum at every delay time [109]. These are called Raman spectrograms and allow for a greatly enhanced level of precision in the extraction of the Raman parameters—a precision that considerably exceeds the instrumental uncertainties.

The understanding of the underlying physical processes behind $I^{(2)}$ CRS (and noisy light spectroscopies in general) has been aided by the recent development of a diagrammatic technique called factorized time correlation (FTC) diagram analysis for properly averaging over the noise components in the incident light [110, 111 and 112] in any noisy light based spectroscopy (linear or nonlinear).

(J) TIME RESOLVED COHERENT RAMAN SCATTERING (TRCRS)

With the advent of short pulsed lasers, investigators were able to perform time resolved coherent Raman scattering. In contrast to using femtosecond pulses whose spectral width provides the two colours needed to produce Raman coherences, discussed above, here we consider pulses having two distinct centre frequencies whose difference drives the coherence. Since the 1970s, picosecond lasers have been employed for this purpose [113, 114], and since the late 1980s femtosecond pulses have also been used [115]. Here we shall briefly focus on the two-colour femtosecond pulsed experiments since they and the picosecond experiments are very similar in concept.

The TR-CRS experiment requires a femtosecond scale light source (originally a rhodamine 6G ring dye laser [115, 116]) and a second longer pulsed (typically several picoseconds) laser operating at a different frequency. The femtosecond source at one colour is split into two pulses having a relative and controllable delay, τ , between them. Each of these two pulses acts once and with the same Fourier component, one in the doorway stage, the other in the window stage. The third, longer pulsed field at the second colour and in a conjugate manner participates with one of the femtosecond pulses in the doorway event to produce the Raman coherence. This polarization then launches the TR-CRS signal field which can be either homodyne or heterodyne detected. This signal must decay with increasing τ as the Raman coherence is given time to decay before the window event takes place.

For homodyne detection, the TR-CRS intensity (for Lorentzian Raman lines) is of the form [115]

$$I_{\text{TR-CRS}} \propto \left| \sum_j e^{-2\gamma_j \tau} e^{-i\omega_j \tau + i\phi_j} \right|^2 \quad (\text{B1.3.25})$$

where j runs over all Raman active modes contained within the bandwidth of the femtosecond scale pulse. The parameters γ_j , ω_j and ϕ_j are the dephasing rate constant, the Raman frequency and phase for the j th mode. One can see from equation (B1.3.25) that for a single mode $I_{\text{TR-CRS}}$ is a simple exponential decay, but when more modes are involved $I_{\text{TR-CRS}}$ will reveal a more complicated beat pattern due to the cross-terms.

TR-CRS has been used to study many molecules from benzene [115, 116, 117 and 118] to betacarotene [119].

(K) IMPULSIVE STIMULATED RAMAN SCATTERING (ISRS)

In discussing RRS above, mention is made of the ‘impulsive’ preparation of wavepackets in both the excited electronic potential surface and the ground state surface. In the absence of electronic resonance only the latter channel is operative and ground state wavepackets can be prepared in transparent materials using the spectral width of femtosecond light to provide the necessary colours. Such impulsive stimulated Raman scattering (ISRS) was first performed by Nelson *et al* on a variety of systems including acoustic modes in glasses [120]

and librational and intramolecular vibrations in liquids [121].

To date, there are two types of configuration employed in ISRS: three-pulse [121] and two-pulse [121]. In both cases, (an) excitation pulse(s) provide(s) the necessary frequencies to create a vibrational wavepacket which proceeds to move within the potential surface. After a delay time τ , a probe pulse having the same central frequency enters the sample and converts the wavepacket into an optical polarization and the coherent fourth wave is detected. For the case of two-pulse ISRS, the transmitted intensity along the probe pulse is followed. In three-pulse ISRS (defined by three wavevectors), the coherently scattered radiation is detected along its unique wavevector. The intensity of the scattered (transmitted) pulse as a function of τ shows damped oscillations at the frequency of the Raman mode (roughly the reciprocal of the recurrence time as the packet oscillates between the two walls of the potential curve). If the pulse durations are longer than the vibrational period (the spectral width is too small to embrace the resonance), no such oscillations can occur. Since in ISRS the spectral width of each pulse is comparable to the Raman frequency, each pulse contains spectral components that produce Stokes and anti-Stokes scattering. These oscillations occur due to the interference between the Stokes and anti-Stokes scattering processes [122]. These processes differ in phase by 180° (the WMEL rules can show this). This expected phase difference has been demonstrated when heterodyne detection is used (the optical Kerr effect probed by an E_{10}) and the signal is frequency resolved [123].

As already mentioned, electronically resonant, two-pulse impulsive Raman scattering (RISRS) has recently been performed on a number of dyes [124]. The main difference between resonant and nonresonant ISRS is that the beats occur in the absorption of the probe rather than the spectral redistribution of the probe pulse energy [124]. These beats are $\frac{\pi}{2}$ out of phase with respect to the beats that occur in nonresonant ISRS (cosine-like rather than sinelike). RISRS has also been shown to have the phase of oscillation depend on the detuning from electronic resonance and it has been shown to be sensitive to the vibrational dynamics in both the ground and excited electronic states [122, 124].

(L) HIGHER ORDER AND HIGHER DIMENSIONAL TIME RESOLVED TECHNIQUES

Of great interest to physical chemists and chemical physicists are the broadening mechanisms of Raman lines in the condensed phase. Characterization of these mechanisms provides information about the microscopic dynamical behaviour of material. The line broadening is due to the interaction between the Raman active chromophore and its environment.

It has been shown that spontaneous or even coherent Raman scattering cannot be used to distinguish between the fast homogeneous and the slow inhomogeneous broadening mechanisms in vibrational transitions [18, 125]. One must use higher order (at least fifth order) techniques if one wishes to resolve the nature of the broadening mechanism. The ability of these higher order techniques to make this distinction is based on the echo phenomena very well known for NMR and mentioned above for D4WM with electronic resonances. The true Raman echo experiment is a time resolved seventh order technique which has recently been reported by Berg *et al* [126, 127, 128 and 129]. It is thus an 8WM

process in which two fields are needed for each step in the normal 4WM echo. A Raman echo WMEL diagram is shown in figure B1.3.7. It is seen that, as in CRS, the first two pulsed field actions create a vibrational coherence. This dephases until the second pair of field actions creates a vibrational population. This is followed by two field actions which again create a vibrational coherence but, now, with opposite phase to the first coherence. Hence one obtains a partial rephasing, or echo, of the macroscopic polarization. The final field action creates the seventh order optical polarization which launches the signal field (the eighth field). Just as for the spin echo in NMR or the electronic echo in 4WM, the degree of rephasing (the

magnitude of the echo) is determined by the amount of slow time scale (inhomogeneous) broadening of the two-level system that is present. Spectral diffusion (the exploration of the inhomogeneity by each chromophore) destroys the echo.

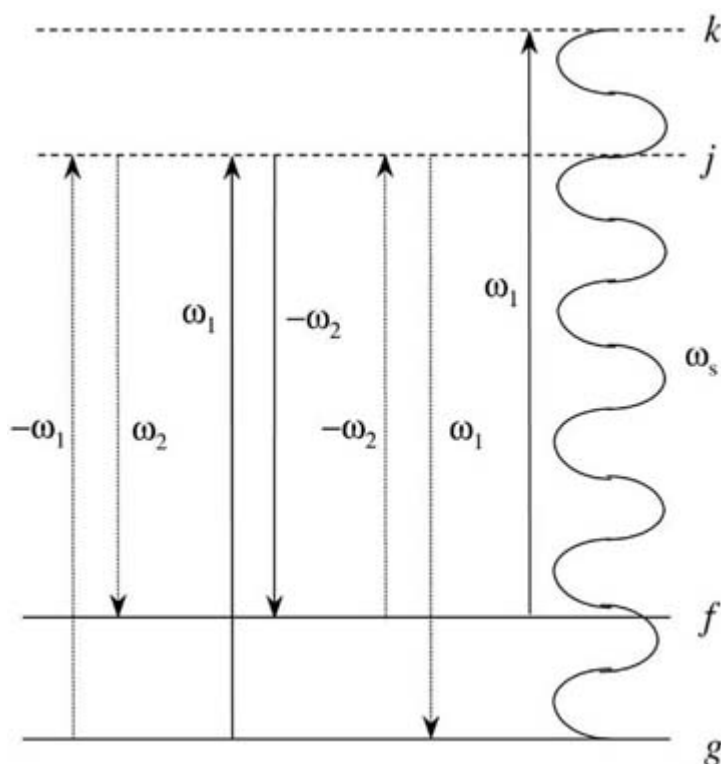


Figure B1.3.7. A WMEF diagram for the seventh order Raman echo. The first two field actions create the usual Raman vibrational coherence which dephases and, to the extent that inhomogeneity is present, also weakens as the coherence from different chromophores ‘walks off’. Then such dephasing is stopped when a second pair of field actions converts this coherence into a population of the excited vibrational state *f*. This is followed by yet another pair of field actions which reconvert the population into a vibrational coherence, but now one with phase opposite to the first. Now, with time, the ‘walked-off’ component of the original coherence can reassemble into a polarization peak that produces the Raman echo at frequency $\omega_s = 2\omega_1 - \omega_2 = \omega_1 + \omega_{fg} = \omega_2 + 2\omega_{fg}$.

An alternative fifth order Raman quasi-echo experiment can also be performed [130, 131, 132, 133 and 134]. Unlike the true Raman echo which involves only two vibrational levels, this process requires the presence of three very nearly evenly spaced levels. A WMEF diagram for the Raman quasi-echo process is shown in figure B1.3.8. Here again the first two field actions create a vibrational coherence which is allowed to dephase. This is followed by a second pair of

field actions, which, instead of creating a population, creates a different vibrational coherence which is of opposite phase and roughly the same order of magnitude as the initial coherence. This serves to allow a rephasing of sorts, the quasi-echo, provided the levels are in part inhomogeneously spread. The final field action creates the fifth order optical polarization that launches the signal field (the sixth field in this overall six-wave mixing process).

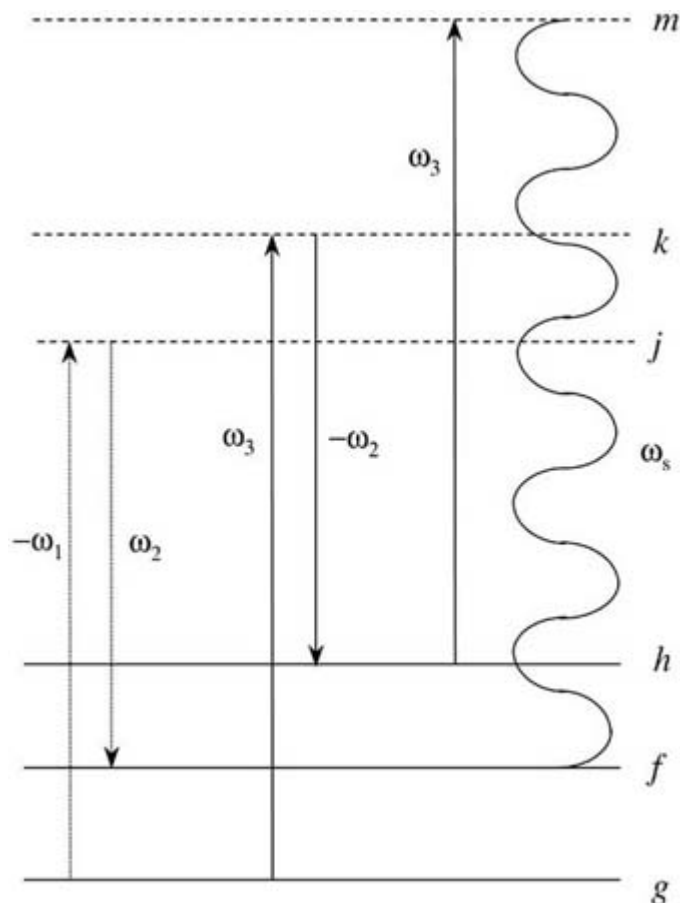


Figure B1.3.8. A W MEL diagram for the three-colour fifth order 'quasi-Raman echo'. As usual, the first pair of field actions creates the fg Raman coherence which is allowed both to dephase and 'walk off' with time. This is followed by a second pair of field actions, which creates a different but oppositely phased Raman coherence (now hf) to the first. Its frequency is at $\omega_3 - \omega_1 = \omega_{hf}$. Provided that ω_{hf} frequencies are identified with an inhomogeneous distribution that is similar to those of the ω_{fg} frequencies, then a quasi-rephasing is possible. The fifth field action converts the newly rephased Raman polarization into the quasi-echo at $\omega_s = 2\omega_3 - \omega_1 = \omega_3 + \omega_{hf}$.

As one goes to higher orders, there are many other processes that can and do occur. Some are true fifth or seventh order processes and others are 'cascaded' events arising from the sequential actions of lower order process [135]. Many of these cascaded sources of polarization interfere with the echo and quasi-echo signal and must be handled theoretically and experimentally.

The key for optimally extracting information from these higher order Raman experiments is to use two time dimensions. This is completely analogous to standard two-dimensional NMR [136] or two-dimensional 4WM echoes. As in NMR, the extra dimension gives information on coherence transfer and the coupling between Raman modes (as opposed to spins in NMR).

With the wealth of information contained in such two-dimensional data sets and with the continued improvements in technology, the Raman echo and quasi-echo techniques will be the basis for much activity and will undoubtedly provide very exciting new insights into condensed phase dynamics in simple molecular materials to systems of biological interest.

This survey of Raman spectroscopies, with direct or implicit use of WMEL diagrams, has by no means touched upon all of the Raman methods ([table B1.3.1](#) and [table B1.3.2](#)). We conclude by mentioning a few additional methods without detailed discussion but with citation. The first is Raman microscopy/imaging [[137](#)]. This technique combines microscopy and imaging with Raman spectroscopy to add an extra dimension to the optical imaging process and hence to provide additional insight into sample surface morphology. (Microscopy is the subject of another chapter in this encyclopedia.) The second is Raman optical activity [[138](#), [139](#)]. This technique discriminates small differences in Raman scattering intensity between left and right circularly polarized light. Such a technique is useful for the study of chiral molecules. The third technique is Raman-SPM [[140](#)]. Here Raman spectroscopy is combined with scanning probe microscopy (SPM) (a subject of another chapter in this encyclopedia) to form a complementary and powerful tool for studying surfaces and interfaces. The fourth is photoacoustic Raman spectroscopy (PARS) which combines CRS with photoacoustic absorption spectroscopy. Class I Raman scattering produces excited vibrational or rotational states in a gas whose energies are converted to bulk translational heating. A pressure wave is produced that is detected as an acoustic signal [[141](#), [142](#)]. The fifth is a novel 5WM process ($\chi^{(4)}$) which can occur in noncentrosymmetric isotropic solutions of chiral macromolecules [[143](#)]. This technique has been given the acronym BioCARS since it has the potential to selectively record background free vibrational spectra of biological molecules [[143](#), [144](#)]. It could be generalized to BioCRS (to include both BioCSRS and BioCARS). The signal is quite weak and can be enhanced with electronic resonance. Finally, we touch upon the general class of spectroscopies known as hyper-Raman spectroscopies. The spontaneous version is called simply hyper-Raman scattering, HRS, a Class I spectroscopy, but there is also the coherent Class II version called CHRS (CAHRS and CSHRS) [[145](#), [146](#) and [147](#)]. These 6WM spectroscopies depend on $\chi^{(5)}$ (Im $\chi^{(5)}$ for HRS) and obey the three-photon selection rules. Their signals are always to the blue of the incident beam(s), thus avoiding fluorescence problems. The selection rules allow one to probe, with optical frequencies, the usual IR spectrum (one photon), not the conventional Raman active vibrations (two photon), but also new vibrations that are symmetry forbidden in both IR and conventional Raman methods.

Although the fifth order hyper- (H-) Raman analogues exist for most of the Raman spectroscopies at third order (HRS [[148](#)], RHRS [[149](#), [150](#)], CHRS [[145](#), [146](#) and [147](#)], SEHRS [[151](#), [152](#) and [153](#)], SERHRS [[154](#)]), let us illustrate the hyper-Raman effect with HRS as the example. In this three-photon process, the scattered radiation contains the frequencies $2\omega_1 \pm \omega_R$ (Stokes and anti-Stokes hyper-Raman scattering), at almost twice the frequency of the incident light. The WMEL diagrams are identical to those of SR, except each single action of the incident laser field (ω_1) must be replaced by two simultaneous actions of the laser field. (The WMEL diagrams for any other 'hyper' process may be obtained in a similar manner.)

Experimentally, this phenomenon is difficult to observe ($I_{\text{HRS}}/I_{\text{SR}} 10^{-5}$); however again electronic resonance enhancement is seen to greatly increase the signal intensity [[148](#)].

B1.3.4 APPLICATIONS

To emphasize the versatility of Raman spectroscopy we discuss just a few selected applications of Raman based spectroscopy to problems in chemical physics and physical chemistry.

B1.3.4.1 APPLICATIONS IN SURFACE PHYSICS

In addition to the many applications of SERS, Raman spectroscopy is, in general, a useful analytical tool having many applications in surface science. One interesting example is that of carbon surfaces which do not support SERS. Raman spectroscopy of carbon surfaces provides insight into two important aspects. First, Raman spectral features correlate with the electrochemical reactivity of carbon surfaces; this allows one to study surface oxidation [[155](#)]. Second, Raman spectroscopy can probe species at carbon surfaces which may account for the highly variable behaviour of carbon materials [[155](#)]. Another application to surfaces is the use

of Raman microscopy in the nondestructive assessment of the quality of ceramic coatings [156]. Finally, an interesting type of surface which does allow for SERS are Mellfs (metal liquid-like films) [157]. Mellfs form at organic–aqueous interfaces when colloids of organic and aqueous metal sols are made. Comparisons with resonance Raman spectra of the bulk solution can give insight into the molecule–surface interaction and adsorption [157].

B1.3.4.2 APPLICATIONS IN COMBUSTION CHEMISTRY

Laser Raman diagnostic techniques offer remote, nonintrusive, nonperturbing measurements with high spatial and temporal resolution [158]. This is particularly advantageous in the area of combustion chemistry. Physical probes for temperature and concentration measurements can be debatable in many combustion systems, such as furnaces, internal combustors etc., since they may disturb the medium or, even worse, not withstand the hostile environments [159]. Laser Raman techniques are employed since two of the dominant molecules associated with air-fed combustion are O₂ and N₂. Homonuclear diatomic molecules unable to have a nuclear coordinate-dependent dipole moment cannot be diagnosed by infrared spectroscopy. Other combustion species include CH₄, CO₂, H₂O and H₂ [160]. These molecules are probed by Raman spectroscopy to determine the temperature profile and species concentration in various combustion processes.

For most practical applications involving turbulent flames and combustion engines, CRS is employed. Temperatures are derived from the spectral distribution of the CRS radiation. This may either be determined by scanning the Stokes frequency through the spectral region of interest or by exciting the transition in a single laser shot with a broadband Stokes beam, thus accessing all Raman resonances in a broad spectral region (multiplexing) [161]. The spectrum may then be observed by a broadband detector such as an optical multichannel analyser [162]. This broadband approach leads to weaker signal intensities, but the entire CRS spectrum is generated with each pulse, permitting instantaneous measurements [163]. Concentration measurements can be carried out in certain ranges (0.5–30% [161]) by using the nonresonant susceptibility as an *in situ* reference standard [158]. Thus fractional concentration measurements are obtained from the spectral profile.

evenly spaced levels. A WMEL diagram for the Raman quasi-echo process is shown in figure B1.3.8 Here again the first two field actions create a vibrational coherence which is allowed to dephase. This is followed by a second pair of field actions, which, instead of creating a population, creates a different vibrational coherence which is of opposite phase and roughly the same order of magnitude as the initial coherence. This serves to allow a rephasing of sorts, the quasi-echo, provided the levels are in part inhomogeneously spread. The final field action creates the fifth order optical polarization that launches the signal field (the sixth field in this overall six-wave mixing process).

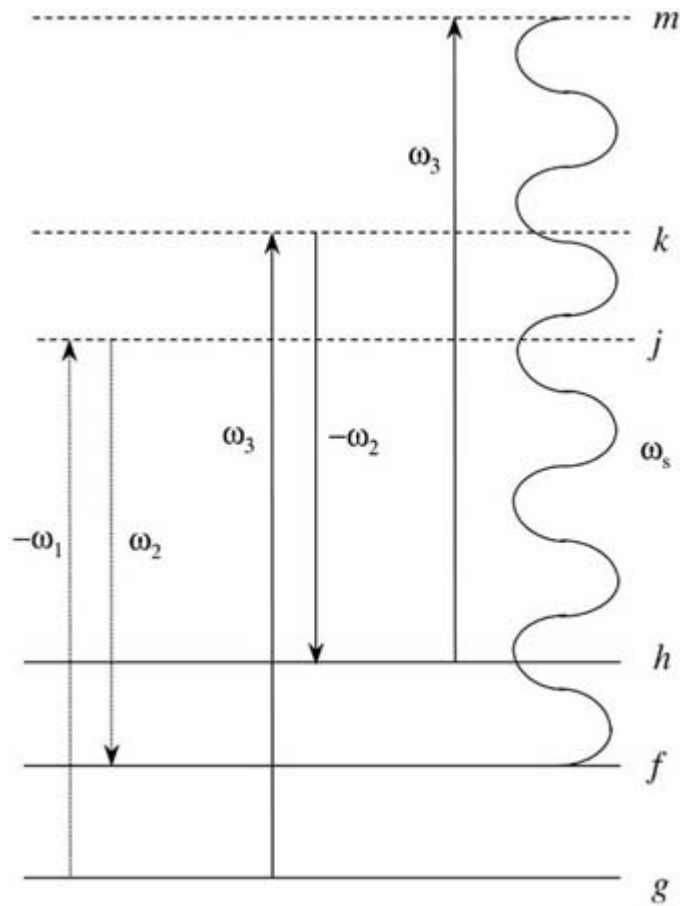


Figure B1.3.8. A W MEL diagram for the three-colour fifth order ‘quasi-Raman echo’. As usual, the first pair of field actions creates the fg Raman coherence which is allowed both to dephase and ‘walk off’ with time. This is followed by a second pair of field actions, which creates a different but oppositely phased Raman coherence (now hf) to the first. Its frequency is at $\omega_3 - \omega_1 = \omega_{hf}$. Provided that ω_{hf} frequencies are identified with an inhomogeneous distribution that is similar to those of the ω_{fg} frequencies, then a quasi-rephasing is possible. The fifth field action converts the newly rephased Raman polarization into the quasi-echo at $\omega_s = 2\omega_3 - \omega_1 = \omega_3 + \omega_{hf}$.

As one goes to higher orders, there are many other processes that can and do occur. Some are true fifth or seventh order processes and others are ‘cascaded’ events arising from the sequential actions of lower order process [135]. Many of these cascaded sources of polarization interfere with the echo and quasi-echo signal and must be handled theoretically and experimentally.

The diagnostic themes in art and archeology are to identify pigments and materials that are used in a given sample, to date them (and determine changes with time) and to authenticate the origins of a given specimen. Diagnosis for forensic purposes may also be considered. In medicine, the principal aim is to provide fast (tens of seconds or faster for *in vivo* studies), noninvasive, nondestructive Raman optical methods for the earliest detection of disease, malignancies being those of immediate interest. In a recent review [173], reference is made to the large variety of subjects for Raman probing of diseased tissues and cells. These include plaques in human arteries and malignancies in breasts, lungs, brain, skin and the intestine. Hair and nails have been studied for signs of disease—metabolic or toxic. Teeth, as well as implants and prostheses, have been examined. Foreign inclusions and chemical migration from surgical implants have been studied.

Another application to biomedicine is to use Raman probing to study DNA biotargets and to identify sequence genes. In fact SERS has been applied to such problems [174].

B1.3.4.5 ANALYTICAL/INDUSTRIAL APPLICATIONS

Examples that use Raman spectroscopy in the quantitative analysis of materials are enormous. Technology that takes Raman based techniques outside the basic research laboratory has made these spectroscopies also available to industry and engineering. It is not possible here to recite even a small portion of applications. Instead we simply sketch one specific example.

Undeniably, one of the most important technological achievements in the last half of this century is the microelectronics industry, the computer being one of its outstanding products. Essential to current and future advances is the quality of the semiconductor materials used to construct vital electronic components. For example, ultra-clean silicon wafers are needed. Raman spectroscopy contributes to this task as a monitor, in real time, of the composition of the standard SC-1 cleaning solution (a mixture of water, H₂O₂ and NH₄OH) [175] that is essential to preparing the ultra-clean wafers.

B1.3.5 A SNAPSHOT OF RAMAN ACTIVITY IN 1998

In conclusion, we attempt to provide a ‘snapshot’ of current research in Raman spectroscopy. Since any choice of topics must be necessarily incomplete, and certainly would reflect our own scientific bias, we choose, instead, an arbitrary approach (at least one not that is not biased by our own specialization). Thus an abbreviated summary of the topics just presented in the keynote/plenary lectures at *ICORS XVI* in Cape Town, South Africa, is presented. Each of the 22 lectures appears in the *Proceedings (and Supplement) of the 16th International Conference on Raman Spectroscopy (1998)*, edited by A M Heyns (Chichester: Wiley) in a four-page format, almost all containing a short list of references. Rather than ourselves searching for seminal citations, we instead give the e-mail address of the principal author, when available. Though the intent in this procedure is to expose the wide scope of current Raman activity, it is hoped that the reader who is looking for more details will not hesitate to seek out the author in this fashion, and that the authors will not feel put upon by this manner of directing people to their work. To relate to [table B1.3.1](#) and [table B1.3.2](#), the acronym is given for the principal Raman spectroscopy that is used for each entry.

Keynote lecture. T G Spiro, e-mail address: spiro@princeton.edu (RRS and TRRRS). Review of protein dynamics followed by TRRRS selective to specific structural and prosthetic elements.

-44-

Plenary 1. D A Long, tel.: + 44 - 1943 608 472. Historical review of the first 70 years of Raman spectroscopy

Plenary 2. S A Asher *et al*, e-mail address: asher@vms.cis.pitt.edu/asher+ (RRS, TRRRS). UV RRS is used to probe methodically the secondary structure of proteins and to follow unfolding dynamics. Developing a library based approach to generalize the method to any protein.

Plenary 3. Ronald E Hester *et al*, e-mail address: reh@york.ac.uk (SERS). Use of dioxane envelope to bring water insoluble chromophores (chlorophylls) into contact with aqueous silver colloids for SERS enhancement. PSERRS—‘protected surface-enhanced resonance Raman spectroscopy’.

Plenary 4. George J Thomas Jr *et al*, e-mail address: thomasgj@cctr.umkc.edu (RS). Protein folding and assembly into superstructures. (Slow) time resolved RS probing of virus construction via protein assembly into an icosahedral (capsid) shell.

Plenary 5. Manuel Cardona, e-mail address: cardona@cardix.mpi-stuttgart.de (RS). Studies of high T_c superconductors. These offer all possible Raman transitions—phonons, magnons, free carrier excitations, pair

breaking excitations and mixed modes.

Plenary 6. Shu-Lin Zhang *et al*, e-mail address: slzhang@pku.edu.cn (RS). Studies of phonon modes of nanoscale one-dimensional materials. Confinement and defect induced Raman transitions.

Plenary 7. S Lefrante, e-mail address: lefrant@cnrs-imn.fr (RRS and SERS). Raman studies of electronic organic materials from conjugated polymers to carbon nanotubes. New insight into chain length distribution, charge transfer and diameter distribution in carbon nanotubes offered by Raman probing.

Plenary 8. J Greve *et al*, e-mail address: J.Greve@tn.utwente.nl (RS). Confocal direct imaging Raman microscope (CDIRM) for probing of the human eye lens. High spatial resolution of the distribution of water and cholesterol in lenses.

Plenary 9. J W Nibler *et al*, e-mail address: niblerj@chem.orst.edu (CARS and SRS). High resolution studies of high lying vibration–rotational transitions in molecules excited in electrical discharges and low density monomers and clusters in free jet expansions. Ionization detected (REMPI) SRS or IDSRS. Detect Raman lines having an FWHM of 30 MHz (10^{-3} cm^{-1}), possibly the sharpest lines yet recorded in RS. Line broadening due to saturation and the ac Stark effect is demonstrated.

Plenary 10. Hiro-o Hamaguchi, e-mail address: hhama@chem.s.u-tokyo.ac.jp (time and polarization resolved multiplex 2D-CARS). Two-dimensional (time and frequency) CARS using broadband dye source and streak camera timing. Studies dynamic behaviour of excited (pumped) electronic states. Follows energy flow within excited molecules. Polarization control of phase of signal (NR background suppression).

Plenary 11. W Kiefer *et al*, e-mail address: wolfgang.kiefer@mail.uni-wue.de (TR CARS). Ultrafast impulsive preparation of ground state and excited state wavepackets by impulsive CARS with REMPI detection in potassium and iodine dimers.

-45-

Plenary 12. Soo-Y Lee, e-mail address: scileesy@nus.edu.sg (RRS). Addresses fundamental theoretical questions in the phase recovery problem in the inverse transform (REP to ABS). See above.

Plenary 13. Andreas Otto, e-mail address: otto@rz.uni-duesseldorf.de (SERS). A survey of problems and models that underlie the SERS effect, now two decades old. Understanding the role of surface roughness in the enhancement.

Plenary 14. A K Ramdas *et al*, e-mail address: akr@physics.purdue.edu (RS). Electronic RS studies of doped diamond as potential semiconducting materials. A Raman active $1s(p_{3/2})-1s(p_{1/2})$ transition of a hole trapped on a boron impurity both in natural and ^{13}C diamond. A striking sensitivity of the transition energy to the isotopic composition of the host lattice.

Plenary 15. B Schrader *et al*, e-mail address: bernhard.schrader@uni-essen.de (NIR-FTRS). A review of the use of Raman spectroscopy in medical diagnostics. Its possibilities, limitations and expectations. Emphasizes the need for a library of reference spectra and the applications of advanced analysis (chemometry) for comparing patient/library spectra.

Plenary 16. N I Koroteev *et al*, e-mail address: Koroteev@nik.phys.msu.su (CARS/CSRS, CAHRS, BioCARS). A survey of the many applications of what we call the Class II spectroscopies from third order and beyond. 2D and 3D Raman imaging. Coherence as stored information, quantum information (the ‘qubit’). Uses terms CARS/CSRS regardless of order. BioCARS is fourth order in optically active solutions.

Plenary 17. P M Champion *et al*, e-mail address: champ@neu.edu (TRRRS). Femtosecond impulsive preparation and timing of ground and excited state Raman coherences in heme proteins. Discovery of coherence transfer along a de-ligation coordinate. See above for further comment.

Plenary 18. Robin J H Clark, e-mail address: r.j.h.clark@ucl.ac.uk (RS). Reports on recent diagnostic probing of art works ranging from illuminated manuscripts, paintings and pottery to papyri and icons. Nondestructive NIR microscopic RS is now realistic using CCD detection. Optimistic about new developments.

Plenary 19. H G M Edwards, e-mail address: h.g.m.edwards@bradford.ac.uk (NIR-FTRS). A review of recent applications of RS to archeology—characterizing ancient pigments, human skin, bone, ivories, teeth, resins, waxes and gums. Aging effects and dating possibilities. Emphasizes use of microscopic Raman.

Plenary 20. M Grimsditch, e-mail address: marcos_grimsditch@qmgate.anl.gov (magnetic field based RS). Low frequency Raman scattering from acoustic phonons is known as Brillouin scattering (BS). However any kind of small quantum Raman scattering is likewise called BS. Ferromagnetic materials offer spin that precesses coherently in the presence of an applied field. Such spin waves or magnons can undergo quantum jumps by the inelastic scattering of light. Experiments (and energy level spacing theory) involving surface magnons in very thin multilayer slabs (such as Fe/Cr/Fe) are discussed. The energy spacing (the Brillouin spectrum) depends on the applied magnetic field, and the RS (or BS) theory is driven by the magnetic component of the electromagnetic field, not the electric (as discussed exclusively in the present chapter).

-46-

Plenary 21A. Alian Wang *et al*, e-mail address: alianw@levee.wustl.edu (RS). (Unable to attend ICORS, but abstract is available in proceedings.) With technological advances, Raman spectroscopy now has become a field tool for geologists. Mineral characterization for terrestrial field work is feasible and a Raman instrument is being designed for the next rover to Mars, scheduled for 2003.

Plenary 21B. A C Albrecht *et al*, e-mail address: aca7@cornell.edu ($I^{(2)}$ CRS) (substituting for plenary 21A). Discusses four new applications using a ‘third’ approach to the Class II spectroscopies (see above). Raman spectrograms from $I^{(2)}$ CARS and $I^{(3)}$ CARS are seen to (i) decisively discriminate between proposed mechanisms for dephasing of the ring breathing mode in liquid benzene, (ii) detect the presence of memory in the Brownian oscillator model of dephasing, (iii) determine with very high accuracy Raman frequency shifts and bandwidths with changing composition in binary mixtures. Moreover these are successfully related to the partial pressures as they change with composition and (iv) to provide a new, definitive, way to discriminate between two competing processes at fifth order (6WM)—cascaded third order or true fifth order.

Clearly the broad survey of current activity in Raman spectroscopy revealed by this simple snapshot promises an exciting future that is likely to find surprising new applications, even as present methods and applications become refined.

APPENDIX

Here we examine the viewing angle dependence of the differential Raman cross-section for the cases of linearly polarized and circularly polarized incident light⁵. The angles used in such experiments are shown in figure B1.3.A.9. Experiments involving circularly polarized light are entirely defined in terms of the scattering angle, ζ , the angle between the wavevectors of the incident (\mathbf{k}_i) and scattered (\mathbf{k}_s) light. For experiment involving linearly polarized light, two angles are needed: ξ , the angle between \mathbf{k}_s and the unit vector along the direction of polarization of the incident light (\mathbf{e}_i) and η , the polar angle of \mathbf{k}_s in the plane perpendicular to \mathbf{e}_i . The polarization of the scattered light is analysed along the two axes \mathbf{e}_a and \mathbf{e}_b , where \mathbf{e}_b is chosen to be perpendicular to \mathbf{e}_i . The differential Raman cross-sections for the two analysing directions are

[176]

$$\left(\frac{d\sigma}{d\Omega}\right)_a = 4\pi^2 \left(\frac{e^2}{4\pi\epsilon_0\hbar c}\right)^2 \bar{v}_1 (\bar{v}_1 \pm \bar{v}_R)^3 \left(\frac{\cos^2 \xi (5\Sigma^1 + 3\Sigma^2)}{30} + \frac{\sin^2 \xi (10\Sigma^0 + 4\Sigma^2)}{30} \right) \quad (\text{B1.3.A1})$$

and

$$\left(\frac{d\sigma}{d\Omega}\right)_b = 4\pi^2 \left(\frac{e^2}{4\pi\epsilon_0\hbar c}\right)^2 \bar{v}_1 (\bar{v}_1 \pm \bar{v}_R)^3 \left(\frac{5\Sigma^1 + 3\Sigma^2}{30}\right). \quad (\text{B1.3.A2})$$

-47-

The total differential cross-section (equation (B1.3.A1) + equation (B1.3.A2)) is then

$$\left(\frac{d\sigma}{d\Omega}\right)_{\xi,\eta} = \left(\frac{d\sigma}{d\Omega}\right) \left(1 - \frac{1 - \rho_1}{1 + \rho_1} \cos^2 \xi\right) \quad (\text{B1.3.A3})$$

and the depolarization ratio

$$\rho(\xi, \eta) = \frac{\rho_1}{1 - (1 - \rho_1) \cos^2 \xi} \quad (\text{B1.3.A4})$$

where ρ_1 is given by [equation \(B1.3.23\)](#) and $(d\sigma/d\Omega)$ is the total differential cross section at 90° , $(d\sigma/d\Omega)_\parallel + (d\sigma/d\Omega)_\perp$, ([equation \(B1.3.20\)](#) and [equation \(B1.3.21\)](#)). From expression (B1.3.A3) and expression (B1.3.A4), one can see that no new information is gained from a linearly polarized light scattering experiment performed at more than one angle, since the two measurables at an angle (ξ, η) are given in terms of the corresponding quantities at 90° .

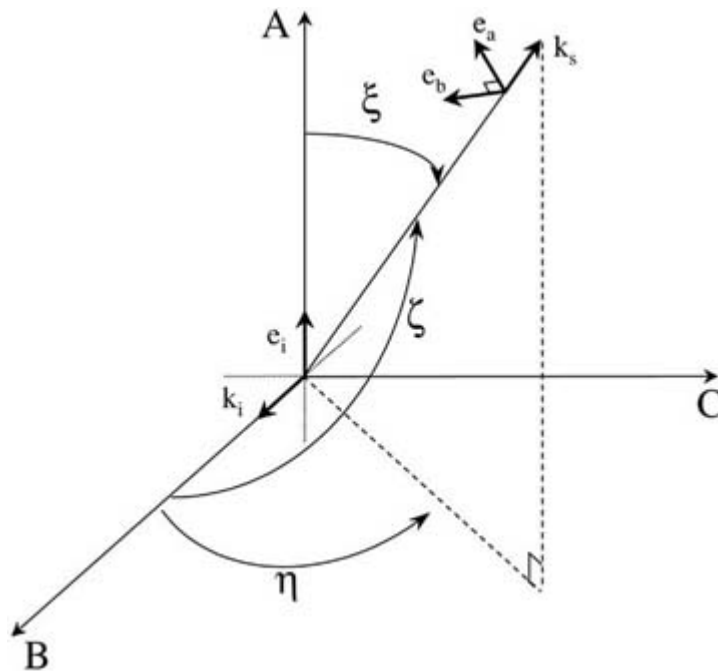


Figure B1.3.A.9. Diagram depicting the angles used in scattering experiments employing linearly and circularly polarized light. The subscripts i and s refer to the incident and scattered beam respectively.

For a circularly polarized light experiment, one can measure the cross sections for either right (r) or left (l) polarized scattered light. Suppose that right polarized light is made incident on a Raman active sample. The general expressions for the Raman cross sections are [176]

-48-

$$\left(\frac{d\sigma}{d\Omega}\right)_r = 4\pi^2 \left(\frac{e^2}{4\pi\epsilon_0\hbar c}\right)^2 \bar{\nu}_1 (\bar{\nu}_1 \pm \bar{\nu}_R)^3 \left(\begin{array}{l} \frac{1}{12}(1 + \cos \zeta)^2 \Sigma^0 \\ + \frac{1}{24}[(1 + \cos \zeta)^2 + 2 \sin^2 \zeta] \Sigma^1 \\ + \frac{1}{120}[(1 + \cos \zeta)^2 + 12(1 - \cos \zeta)] \Sigma^2 \end{array} \right)$$

and

In analogy with the depolarization ratio for linearly polarized light, the ratio of the two above quantities is known as the reversal coefficient, $R(\zeta)$, given by

$$R(\zeta) \equiv \frac{\left(\frac{d\sigma}{d\Omega}\right)_l}{\left(\frac{d\sigma}{d\Omega}\right)_r} = \frac{1 - \frac{1-\rho_1}{2(1+\rho_1)} \sin^2 \zeta - \frac{1-R(0)}{1+R(0)} \cos \zeta}{1 - \frac{1-\rho_1}{2(1+\rho_1)} \sin^2 \zeta + \frac{1-R(0)}{1+R(0)} \cos \zeta} \quad (\text{B1.3.A5})$$

where the zero angle reversal coefficient, $R(0)$, is $6\Sigma^2/(10\Sigma^0 + 5\Sigma^1 + \Sigma^2)$. Measurement of the reversal coefficient (equation (B1.3.A5)) at two appropriate scattering angles permits one to determine both ρ_1 and $R(0)$. Thus, only with circularly polarized light is one able to quantify all three rotational tensor invariants [176].

ACKNOWLEDGMENTS

One of us, ACA, also wishes to thank Dr David Klug for his provision of space, time, and a fine scientific atmosphere during a two month sabbatical leave at Imperial College, which turned out to involve considerable work on this material. The most helpful editing efforts of Professor R J H Clark have been very much appreciated. Finally, gratitude goes to Dr Gia Maisuradze for a careful reading of the manuscript and to Ms Kelly Case for her most attentive and effective reading of the proofs.

¹ A version of this material appears in a special issue of the *Journal of Physical Chemistry* dedicated to the Proceedings of the International Conference on Time-Resolved Vibrational Spectroscopy (TRVS IX), May 16–22 1999, Tucson, Arizona. See: Kirkwood J C, Ulness D J and Albrecht A C 2000 On the classification of the electric field spectroscopies: applications to Raman scattering *J. Phys. Chem. A* **104** 4167–73.

²In fact averaging over an odd number of direction cosines need not always vanish for an isotropic system. This is the case for solutions containing chiral centres which may exhibit even order signals such as 'BioCARS' in [table B1.3.2](#).

RS (or BS) theory is driven by the magnetic component of the electromagnetic field, not the electric (as discussed exclusively in the present chapter).

Plenary 21A. Alian Wang *et al*, e-mail address: alianw@levee.wustl.edu (RS). (Unable to attend ICORS, but abstract is available in proceedings.) With technological advances, Raman spectroscopy now has become a field tool for geologists. Mineral characterization for terrestrial field work is feasible and a Raman instrument is being designed for the next rover to Mars, scheduled for 2003.

Plenary 21B. A C Albrecht *et al*, e-mail address: aca7@cornell.edu (I⁽²⁾CRS) (substituting for plenary 21A). Discusses four new applications using a ‘third’ approach to the Class II spectroscopies (see above). Raman spectrograms from I⁽²⁾CARS and I⁽³⁾CARS are seen to (i) decisively discriminate between proposed mechanisms for dephasing of the ring breathing mode in liquid benzene, (ii) detect the presence of memory in the Brownian oscillator model of dephasing, (iii) determine with very high accuracy Raman frequency shifts and bandwidths with changing composition in binary mixtures. Moreover these are successfully related to the partial pressures as they change with composition and (iv) to provide a new, definitive, way to discriminate between two competing processes at fifth order (6WM)—cascaded third order or true fifth order.

Clearly the broad survey of current activity in Raman spectroscopy revealed by this simple snapshot promises an exciting future that is likely to find surprising new applications, even as present methods and applications become refined.

APPENDIX

Here we examine the viewing angle dependence of the differential Raman cross-section for the cases of linearly polarized and circularly polarized incident light⁵. The angles used in such experiments are shown in [figure B1.3.A.9](#) Experiments involving circularly polarized light are entirely defined in terms of the scattering angle, ζ , the angle between the wavevectors of the incident (\mathbf{k}_i) and scattered (\mathbf{k}_s) light. For experiment involving linearly polarized light, two angles are needed: ξ , the angle between \mathbf{k}_s and the unit vector along the direction of polarization of the incident light (\mathbf{e}_i) and η , the polar angle of \mathbf{k}_s in the plane perpendicular to \mathbf{e}_i . The polarization of the scattered light is analysed along the two axes \mathbf{e}_a and \mathbf{e}_b , where \mathbf{e}_b is chosen to be perpendicular to \mathbf{e}_i . The differential Raman cross-sections for the two analysing directions are [176]

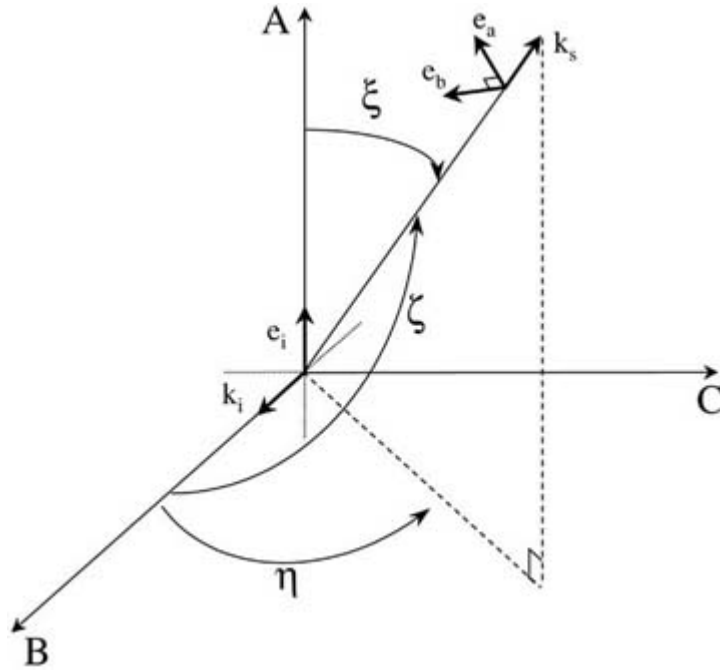


Figure B1.3.A.9. Diagram depicting the angles used in scattering experiments employing linearly and circularly polarized light. The subscripts i and s refer to the incident and scattered beam respectively.

$$\left(\frac{d\sigma}{d\Omega}\right)_a = 4\pi^2 \left(\frac{e^2}{4\pi\epsilon_0\hbar c}\right)^2 \bar{v}_1 (\bar{v}_1 \pm \bar{v}_R)^3 \left(\frac{\cos^2 \xi (5\Sigma^1 + 3\Sigma^2)}{30} + \frac{\sin^2 \xi (10\Sigma^0 + 4\Sigma^2)}{30} \right) \quad (\text{B1.3.A1})$$

and

$$\left(\frac{d\sigma}{d\Omega}\right)_b = 4\pi^2 \left(\frac{e^2}{4\pi\epsilon_0\hbar c}\right)^2 \bar{v}_1 (\bar{v}_1 \pm \bar{v}_R)^3 \left(\frac{5\Sigma^1 + 3\Sigma^2}{30}\right). \quad (\text{B1.3.A2})$$

The total differential cross-section (equation (B1.3.A1) + equation (B1.3.A2)) is then

$$\left(\frac{d\sigma}{d\Omega}\right)_{\xi,\eta} = \left(\frac{d\sigma}{d\Omega}\right) \left(1 - \frac{1 - \rho_1}{1 + \rho_1} \cos^2 \xi\right) \quad (\text{B1.3.A3})$$

and the depolarization ratio

$$\rho(\xi, \eta) = \frac{\rho_1}{1 - (1 - \rho_1) \cos^2 \xi} \quad (\text{B1.3.A4})$$

where ρ_1 is given by [equation \(B1.3.23\)](#) and $(d\sigma/d\Omega)$ is the total differential cross section at 90° , $(d\sigma/d\Omega)_\parallel + (d\sigma/d\Omega)_\perp$, ([equation \(B1.3.20\)](#) and [equation \(B1.3.21\)](#)). From [expression \(B1.3.A3\)](#) and [expression \(B1.3.A4\)](#), one can see that no new information is gained from a linearly polarized light scattering experiment performed at more than one angle, since the two measurables at an angle (ξ, η) are given in terms of the corresponding quantities at 90° .

For a circularly polarized light experiment, one can measure the cross sections for either right (r) or left (l) polarized scattered light. Suppose that right polarized light is made incident on a Raman active sample. The general expressions for the Raman cross sections are [176]

$$\left(\frac{d\sigma}{d\Omega}\right)_r = 4\pi^2 \left(\frac{e^2}{4\pi\epsilon_0\hbar c}\right)^2 \bar{\nu}_1(\bar{\nu}_1 \pm \bar{\nu}_R)^3 \begin{pmatrix} \frac{1}{12}(1 + \cos \zeta)^2 \Sigma^0 \\ + \frac{1}{24}[(1 + \cos \zeta)^2 + 2 \sin^2 \zeta] \Sigma^1 \\ + \frac{1}{120}[(1 + \cos \zeta)^2 + 12(1 - \cos \zeta)] \Sigma^2 \end{pmatrix}$$

and

$$\left(\frac{d\sigma}{d\Omega}\right)_l = 4\pi^2 \left(\frac{e^2}{4\pi\epsilon_0\hbar c}\right)^2 \bar{\nu}_1(\bar{\nu}_1 \pm \bar{\nu}_R)^3 \begin{pmatrix} \frac{1}{12}(-1 + \cos \zeta)^2 \Sigma^0 \\ + \frac{1}{24}[(-1 + \cos \zeta)^2 + 2 \sin^2 \zeta] \Sigma^1 \\ + \frac{1}{120}[(-1 + \cos \zeta)^2 + 12(1 + \cos \zeta)] \Sigma^2 \end{pmatrix}.$$

In analogy with the depolarization ratio for linearly polarized light, the ratio of the two above quantities is known as the reversal coefficient, $R(\zeta)$, given by

$$R(\zeta) \equiv \frac{\left(\frac{d\sigma}{d\Omega}\right)_l}{\left(\frac{d\sigma}{d\Omega}\right)_r} = \frac{1 - \frac{1-\rho_1}{2(1+\rho_1)} \sin^2 \zeta - \frac{1-R(0)}{1+R(0)} \cos \zeta}{1 - \frac{1-\rho_1}{2(1+\rho_1)} \sin^2 \zeta + \frac{1-R(0)}{1+R(0)} \cos \zeta} \quad (\text{B1.3.A5})$$

where the zero angle reversal coefficient, $R(0)$, is $6\Sigma^2/(10\Sigma^0 + 5\Sigma^1 + \Sigma^2)$. Measurement of the reversal coefficient (equation (B1.3.A5)) at two appropriate scattering angles permits one to determine both ρ_1 and $R(0)$. Thus, only with circularly polarized light is one able to quantify all three rotational tensor invariants [176].

ACKNOWLEDGMENTS

One of us, ACA, also wishes to thank Dr David Klug for his provision of space, time, and a fine scientific atmosphere during a two month sabbatical leave at Imperial College, which turned out to involve considerable work on this

material. The most helpful editing efforts of Professor R J H Clark have been very much appreciated. Finally, gratitude goes to Dr Gia Maisuradze for a careful reading of the manuscript and to Ms Kelly Case for her most attentive and effective reading of the proofs.

¹ A version of this material appears in a special issue of the *Journal of Physical Chemistry* dedicated to the Proceedings of the International Conference on Time-Resolved Vibrational Spectroscopy (TRVS IX), May 16–22 1999, Tucson, Arizona. See: Kirkwood J C, Ulness D J and Albrecht A C 2000 On the classification of the electric field spectroscopies: applications to Raman scattering *J. Phys. Chem. A* **104** 4167–73.

²In fact averaging over an odd number of direction cosines need not always vanish for an isotropic system. This is the case for solutions containing chiral centres which may exhibit even order signals such as 'BioCARS' in [table B1.3.2](#).

³ Here, we have averaged over all possible orientations of the molecules. (See [26].)

⁴ Raman cross-sections, based on the linear polarizability, are now routinely subject to quantum chemical calculations. These may be found as options in commercial packages such as 'Gaussian 98' (Gaussian Inc., Pittsburgh, PA).

⁵ This treatment is essentially that given in [176].

REFERENCES

- [1] Raman C V and Krishnan K S 1928 A new type of secondary radiation *Nature* **121** 501–2
- [2] Raman C V 1928 A new radiation *Indian J. Phys.* **2** 387–98
- [3] Raman C V and Krishnan K S 1928 A new class of spectra due to secondary radiation. Part I *Indian J. Phys.* **2** 399–419
- [4] Landsberg G and Mandelstam L 1928 Eine neue Erscheinung bei der Lichtzerstreuung in Krystallen *Naturwissenschaften* **16** 557–8
- [5] Long D A 1988 Early history of the Raman effect *Int. Rev. Phys. Chem.* **7** 314–49
- [6] Lee D and Albrecht A C 1985 A unified view of Raman, resonance Raman, and fluorescence spectroscopy (and their analogues in two-photon absorption) *Advances in Infrared and Raman Spectroscopy* vol 12, ed R J H Clark and R E Hester (New York: Wiley) pp 179–213
- [7] Lee D and Albrecht A C 1993 On global energy conservation in nonlinear light matter interaction: the nonlinear spectroscopies, active and passive *Adv. Phys. Chem.* **83** 43–87
- [8] Fano U 1957 Description of states in quantum mechanics by density matrix and operator techniques *Rev. Mod. Phys.* **29** 74–93
- [9] ter Haar D 1961 Theory and applications of the density matrix *Rep. Prog. Phys.* **24** 304–62
- [10] Cohen-Tannoudji C, Diu B and Laloë F 1977 *Quantum Mechanics* (New York: Wiley)

- [11] Sakurai J J 1994 *Modern Quantum Mechanics* revised edn (Reading, MA: Addison-Wesley)
- [12] Louisell W H 1973 *Quantum Statistical Properties of Radiation* (New York: Wiley)
- [13] Shen Y R 1984 *The Principles of Nonlinear Optics* (New York: Wiley)
- [14] Butcher P N and Cotter D 1990 *The Elements of Nonlinear Optics* (Cambridge: Cambridge University Press)
- [15] Boyd R W 1992 *Nonlinear Optics* (San Diego, CA: Academic)
- [16] Schubert M and Wilhiemi B 1986 *Nonlinear Optics and Quantum Electronics* (New York: Wiley)
- [17] Yariv A 1989 *Quantum Electronics* (New York: Wiley)
- [18] Mukamel S 1995 *Principles of Nonlinear Optical Spectroscopy* (New York: Oxford University Press)
- [19] Levenson M D 1982 *Introduction to Nonlinear Laser Spectroscopy* (New York: Academic)
- [20] Loudon R 1983 *The Quantum Theory of Light* (New York: Oxford University Press)
- [21] Wright J C, Labuda M J, Zilian A, Chen P C and Hamilton J P 1997 New selective nonlinear vibrational spectroscopies *J. Luminesc.* **72–74** 799–801
- [22] Labuda M J and Wright J C 1997 Measurement of vibrationally resonant $\chi^{(3)}$ and the feasibility of new vibrational spectroscopies *Phys. Rev. Lett.* **79** 2446–9
- [23] Wright J C, Chen P C, Hamilton J P, Zilian A and Labuda M J 1997 Theoretical foundations for a new

family of infrared four wave mixing spectroscopies *Appl. Spectrosc.* **51** 949–58

- [24] Chen P C, Hamilton J P, Zilian A, Labuda M J and Wright J C 1998 Experimental studies for a new family of infrared four wave mixing spectroscopies *Appl. Spectrosc.* **52** 380–92
- [25] Tokmakoff A, Lang M J, Larson D S and Fleming G R 1997 Intrinsic optical heterodyne detection of a two-dimensional fifth order Raman response *Chem. Phys. Lett.* **272** 48–54
- [26] Monson P R and McClain W 1970 Polarization dependence of the two-photon absorption of tumbling molecules with application to liquid 1-chloronaphthalene and benzene *J. Chem. Phys.* **53** 29–37
- [27] Lee S Y 1983 Placzek-type polarizability tensors for Raman and resonance Raman scattering *J. Chem. Phys.* **78** 723–34
- [28] Melinger J S and Albrecht A C 1986 Theory of time- and frequency-resolved resonance secondary radiation from a three-level system *J. Chem. Phys.* **84** 1247–58
- [29] Kramers H A and Heisenberg W 1925 Über die Streuung von Strahlung durch Atome *Z. Phys.* **31** 681–708
- [30] Placzek G 1934 The Rayleigh and Raman scattering *Handbuch der Radiologie* ed E Marx (Leipzig: Akademische) (Engl. transl. UCRL 526(L) 1962 Clearinghouse USAEC transl. A Werbin)
- [31] Tang J and Albrecht A C 1970 Developments in the theories of vibrational Raman intensities *Raman Spectroscopy: Theory and Practice* vol 2, ed H A Szymanski (New York: Plenum) pp 33–68
- [32] Toleutaev B N, Tahara T and Hamaguchi H 1994 Broadband (1000 cm^{-1}) multiplex CARS spectroscopy: application to polarization sensitive and time-resolved measurements *Appl. Phys.* **59** 369–75
- [33] Laubereau A 1982 Stimulated Raman scattering *Non-Linear Raman Spectroscopy and its Chemical Applications* ed W Kiefer and D A Long (Dordrecht: Reidel)

- [34] Woodward L A 1967 General introduction *Raman Spectroscopy: Theory and Practice* vol 1, ed H A Szymanski (New York: Plenum)
- [35] Long D A (ed) *J. Raman Spectrosc.* (Chichester: Wiley)
- [36] Clark R J H and Hester R E (eds) *Adv. Infrared Raman Spectros.* (London: Heyden)
- [37] *ICORS '96: XVth Int. Conf. on Raman Spectroscopy* (New York: Wiley)
- [38] Pinan J P, Ouillon R, Ranson P, Becucci M and Califano S 1998 High resolution Raman study of phonon and vibron bandwidths in isotropically pure and natural benzene crystal *J. Chem. Phys.* **109** 1–12
- [39] Ferraro J R and Nakamoto K 1994 *Introductory Raman Spectroscopy* (San Diego: Academic)
- [40] Hirschfeld T and Chase B 1986 FT-Raman spectroscopy: development and justification *Appl. Spectrosc.* **40** 133–9
- [41] McCreery R L 1996 Instrumentation for dispersive Raman spectroscopy *Modern Techniques in Raman Spectroscopy* ed J J Laserna (New York: Wiley)
- [42] Hendra P J 1996 Fourier transform Raman spectroscopy *Modern Techniques in Raman Spectroscopy* ed J J Laserna (New York: Wiley)
- [43] Hendra P J, Jones C and Warnes G 1991 *Fourier Transform Raman Spectroscopy: Instrumentation and Chemical Applications* (New York: Ellis Horwood)
- [44] Heller E J 1981 The semiclassical way to molecular spectroscopy *Accounts Chem. Res.* **14** 368–75
- [45] Myers A B 1997 'Time dependent' resonance Raman theory *J. Raman Spectrosc.* **28** 389–401
- [46] Hizhnyakov V and Tehver I 1988 Transform method in resonance Raman scattering with quadratic Franck–Condon and Herzberg–Teller interactions *J. Raman Spectrosc.* **19** 383–8
- [47] Page J B and Tonks D L 1981 On the separation of resonance Raman scattering into orders in the time correlator theory *J. Chem. Phys.* **75** 5694–708

- [48] Champion P M and Albrecht A C 1981 On the modeling of absorption band shapes and resonance Raman excitation profiles *Chem. Phys. Lett.* **82** 410–13
- [49] Cable J R and Albrecht A C 1986 The inverse transform in resonance Raman scattering *Conf. sponsored by the University of Oregon* ed W L Peticolas and B Hudson
- [50] Remacle F and Levine R D 1993 Time domain information from resonant Raman excitation profiles: a direct inversion by maximum entropy *J. Chem. Phys.* **99** 4908–25
- [51] Albrecht A C, Clark R J H, Oprescu D, Owens S J R and Svensen C 1994 Overtone resonance Raman scattering beyond the Condon approximation: transform theory and vibronic properties *J. Chem. Phys.* **101** 1890–903
- [52] Weber A (ed) 1979 *Raman Spectroscopy of Gases and Liquids* (New York: Springer)
- [53] Page J B 1991 Many-body problem to the theory of resonance Raman scattering by vibronic systems *Top. Appl. Phys.* **116** 17–72
- [54] Cable J R and Albrecht A C 1986 A direct inverse transform for resonance Raman scattering *J. Chem. Phys.* **84** 4745–54
- [55] Joo T and Albrecht A C 1993 Inverse transform in resonance Raman scattering: an iterative approach *J. Phys. Chem.* **97** 1262–4
-

-55-

- [56] Lee S-Y 1998 Forward and inverse transforms between the absorption lineshape and Raman excitation profiles *XVth Int. Conf. on Raman Spectroscopy* ed A M Heyns (New York: Wiley) pp 48–51
- [57] Lee S-Y and Feng Z W 1996 Reply to the comment on the inversion of Raman excitation profiles *Chem. Phys. Lett.* **260** 511–13
- [58] Marzocchi M P, Mantini A R, Casu M and Smulevich G 1997 Intramolecular hydrogen bonding and excited state proton transfer in hydroxyanthraquinones as studied by electronic spectra, resonance Raman scattering, and transform analysis *J. Chem. Phys.* **108** 1–16
- [59] Mantini A R, Marzocchi M P and Smulevich G 1989 Raman excitation profiles and second-derivative absorption spectra of beta-carotene *J. Chem. Phys.* **91** 85–91
- [60] Asher S A, Chi Z, Holtz J S W, Lednev I K, Karnoup A S and Sparrow M C 1998 UV resonance Raman studies of protein structure and dynamics *XVth Int. Conf. on Raman Spectroscopy* ed A M Heyns (New York: Wiley) pp 11–14
- [61] Zhu I, Widom A and Champion P M 1997 A multidimensional Landau–Zener description of chemical reaction dynamics and vibrational coherence *J. Chem. Phys.* **107** 2859–71
- [62] Champion P M, Rosca F, Chang W, Kumar A, Christian J and Demidov A 1998 Femtosecond coherence spectroscopy of heme proteins *XVth Int. Conf. on Raman Spectroscopy* ed A M Heyns (New York: Wiley) pp 73–6
- [63] Johnson B R, Kittrell C, Kelly P B and Kinsey J L 1996 Resonance Raman spectroscopy of dissociative polyatomic molecules *J. Phys. Chem.* **100** 7743–64
- [64] Kung C Y, Chang B-Y, Kittrell C, Johnson B R and Kinsey J L 1993 Continuously scanned resonant Raman excitation profiles for iodobenzene excited in the B continuum *J. Phys. Chem.* **97** 2228–35
- [65] Galica G E, Johnson B R, Kinsey J L and Hale M O 1991 Incident frequency dependence and polarization properties of the CH₃I Raman spectrum *J. Phys. Chem.* **95** 7994–8004
- [66] Tripathi G N R and Schuler R H 1984 The resonance Raman spectrum of phenoxy radical *J. Chem. Phys.* **81** 113–21
- [67] Tripathi G N R and Schuler R H 1982 Time-resolved resonance Raman scattering of transient radicals: the p-aminophenoxy radical *J. Chem. Phys.* **76** 4289–90
- [68] Qin L, Tripathi G N R and Schuler R H 1987 Radiolytic oxidation of 1,2,4-benzenetriol: an application of time-resolved resonance Raman spectroscopy to kinetic studies of reaction intermediates *J. Chem. Phys.*

- [69] Okamoto H, Nakabayashi T and Tasumi M 1997 Analysis of anti-Stokes RRS excitation profiles as a method for studying vibrationally excited molecules *J. Phys. Chem.* **101** 3488–93
- [70] Sakamoto A, Okamoto H and Tasumi M 1998 Observation of picosecond transient Raman spectra by asynchronous Fourier transform Raman spectroscopy 1998 *Appl. Spectrosc.* **52** 76–81
- [71] Bloembergen N 1967 The stimulated Raman effect *Am. J. Phys.* **35** 989–1023
- [72] Wang C-S 1975 The stimulated Raman process *Quantum Electronics* vol 1A, ed H Rabin and C L Tang (New York: Academic) pp 447–72
- [73] Kaiser W and Maier M 1972 Stimulated Rayleigh, Brillouin and Raman spectroscopy *Laser Handbook* vol 2, ed F T Arcchi and E O Schult-Dubois (Amsterdam: North-Holland) pp 1077–150
- [74] Wang C-S 1969 Theory of stimulated Raman scattering *Phys. Rev.* **182** 482–94
-

- [75] Maier M, Kaiser W and Giordmaine J A 1969 Backward stimulated Raman scattering *Phys. Rev.* **177** 580–99
- [76] Jones W T and Stoicheff B P 1964 Inverse Raman spectra: induced absorption at optical frequencies *Phys. Rev. Lett.* **13** 657–9
- [77] Fleischmann M, Hendra P J and McQuillan A J 1974 Raman spectra of pyridine adsorbed at a silver electrode *Chem. Phys. Lett.* **26** 163–6
- [78] Albrecht M G and Creighton J A 1977 Anomalously intense Raman spectra of pyridine at a silver electrode *J. Am. Chem. Soc.* **99** 5215–17
- [79] Jeanmaire D L and Van Duyne R P 1977 Part I: heterocyclic, aromatic and aliphatic amines adsorbed on the anodized silver electrode *J. Electroanal. Chem.* **84** 1–20
- [80] See articles in the special issue on Raman spectroscopy and surface phenomena 1991 *J. Raman Spectrosc.* **22** 727–839
- [81] See articles in the special issue on Raman spectroscopy in surface science 1985 *Surf. Sci.* **158** 1–693
- [82] Freunsholtz P, Van Duyne R P and Schneider S 1997 Surface-enhanced Raman spectroscopy of trans-stilbene adsorbed on platinum- or self-assembled monolayer-modified silver film over nanosphere surfaces *Chem. Phys. Lett.* **281** 372–8
- [83] Yang W H, Hulstee J C, Schatz G C and Van Duyne R P 1996 A surface-enhanced hyper-Raman and surface-enhanced Raman scattering study of trans-1,2-bis(4-pyridyl)ethylene adsorbed onto silver film over nanosphere electrodes. Vibrational assignments: experiments and theory *J. Chem. Phys.* **104** 4313–26
- [84] Nie S and Emory S R 1997 Probing single molecules and single nanoparticles by surface-enhanced Raman scattering *Science* **275** 1102–6
- [85] Nie S and Emory S R 1997 Near-field surface-enhanced Raman spectroscopy on single silver nanoparticles *Anal. Chem.* **69** 2631–5
- [86] Furtak T E and Reyes J 1980 A critical analysis of the theoretical models for the giant Raman effect from adsorbed molecules *Surf. Sci.* **93** 351–82
- [87] Rupérez A and Laserna J J 1996 Surface-enhanced Raman spectroscopy *Modern Techniques in Raman Spectroscopy* ed J J Laserna (New York: Wiley) pp 227–64
- [88] Otto A 1991 Surface-enhanced Raman scattering of adsorbates *J. Raman Spectrosc.* **22** 743–52
- [89] Kneipp K, Wang Y, Kneipp H, Itzkan I, Dasari R R and Feld M S 1996 Approach to single molecule detection using surface-enhanced Raman scattering *ICORS '98: XVth Int. Conf. on Raman Spectroscopy* ed S A Asher and P B Stein (New York: Wiley) pp 636–7
- [90] Eesley G L 1981 *Coherent Raman Spectroscopy* (Oxford: Pergamon)

- [91] Marowsky G and Smirnov V N (eds) 1992 *Coherent Raman Spectroscopy: Recent Advances* (Berlin: Springer)
- [92] Gomez J S 1996 Coherent Raman spectroscopy *Modern Techniques in Raman Spectroscopy* ed J J Laserna (New York: Wiley) pp 305–42
- [93] Castellucci E M, Righini R and Foggi P (eds) 1993 *Coherent Raman Spectroscopy* (Singapore: World Scientific)
- [94] Oudar J-L, Smith R W and Shen Y R 1979 Polarization-sensitive coherent anti-Stokes Raman spectroscopy *Appl. Phys. Lett.* **34** 758–60
-

-57-

- [95] Schaertel S A, Lee D and Albrecht A C 1995 Study of polarization CRS and polarization I⁽²⁾CRS with applications *J. Raman Spectrosc.* **26** 889–99
- [96] Eesley G L 1978 Coherent Raman spectroscopy *J. Quant. Spectrosc. Radiat. Transfer* **22** 507–76
- [97] Heiman D, Hellwarth R W, Levenson M D and Martin G 1976 Raman-induced Kerr effect *Phys. Rev. Lett.* **36** 189–92
- [98] Schaertel S A, Albrecht A C, Lau A and Kummrow A 1994 Interferometric coherent Raman spectroscopy with incoherent light: some applications *Appl. Phys. B* **59** 377–87
- [99] Schaertel S A and Albrecht A C 1994 Interferometric coherent Raman spectroscopy: resonant and non-resonant contributions *J. Raman Spectrosc.* **25** 545–55
- [100] Morita N and Yajima T 1984 Ultrafast-time-resolved coherent transient spectroscopy with incoherent light *Phys. Rev. A* **30** 2525–36
- [101] Asaka S, Nakatsuka H, Fujiwara M and Matsuoka M 1984 Accumulated photon echoes with incoherent light in Nd³⁺-doped silicate glass *Phys. Rev. A* **29** 2286–9
- [102] Beech R and Hartmann S R 1984 Incoherent photon echoes *Phys. Rev. Lett.* **53** 663–6
- [103] Kobayashi T 1994 Measurement of femtosecond dynamics of nonlinear optical responses *Modern Nonlinear Optics* part 3, ed M Evans and S Kielich *Adv. Chem. Phys.* **85** 55–104
- [104] Kummrow A and Lau A 1996 Dynamics in condensed molecular systems studied by incoherent light *Appl. Phys. B* **63** 209–23
- [105] Albrecht A C, Smith S P, Tan D, Schaertel S A and DeMott D 1995 Ultrasharp spectra and ultrafast timing from noisy coherence in four wave mixing *Laser Phys.* **5** 667–75
- [106] Dugan M A, Melinger J S and Albrecht A C 1988 Terahertz oscillations from molecular liquids in CSRS/CARS spectroscopy with incoherent light *Chem. Phys. Lett.* **147** 411–19
- [107] Dugan M A and Albrecht A C 1991 Radiation–matter oscillations and spectral line narrowing in field-correlated four-wave mixing I: theory *Phys. Rev. A* **43** 3877–921
- [108] Ulness D J, Stimson M J, Kirkwood J C and Albrecht A C 1997 Interferometric downconversion of high frequency molecular vibrations with time–frequency-resolved coherent Raman scattering using quasi-cw noisy laser light: C–H stretching modes of chloroform and benzene *J. Phys. Chem. A* **101** 4587–91
- [109] Stimson M J, Ulness D J and Albrecht A C 1996 Frequency and time resolved coherent Raman scattering in CS₂ using incoherent light *Chem. Phys. Lett.* **263** 185–90
- [110] Ulness D J and Albrecht A C 1996 Four-wave mixing in a Bloch two-level system with incoherent laser light having a Lorentzian spectral density: analytic solution and a diagrammatic approach *Phys. Rev. A* **53** 1081–95
- [111] Ulness D J and Albrecht A C 1997 A theory of time resolved coherent Raman scattering with spectrally tailored noisy light *J. Raman Spectrosc.* **28** 571–8
- [112] Ulness D J, Kirkwood J C, Stimson M J and Albrecht A C 1997 Theory of coherent Raman scattering with

quasi-cw noisy light for a general lineshape function *J. Chem. Phys.* **107** 7127–37

- [113] Laubereau A and Kaiser W 1978 Vibrational dynamics of liquids and solids investigated by picosecond light pulses *Rev. Mod. Phys.* **50** 607–65
-

-58-

- [114] Laubereau A and Kaiser W 1978 Coherent picosecond interactions *Coherent Nonlinear Optics* ed M S Feld and V S Letokov (Berlin: Springer) pp 271–92
- [115] Leonhardt R, Holzappel W, Zinth W and Kaiser W 1987 Terahertz quantum beats in molecular liquids *Chem. Phys. Lett.* **133** 373–7
- [116] Okamoto H and Yoshihara K 1990 Femtosecond time-resolved coherent Raman scattering under various polarization and resonance conditions *J. Opt. Soc. B* **7** 1702–8
- [117] Joo T, Dugan M A and Albrecht A C 1991 Time resolved coherent Stokes Raman spectroscopy (CSRS) of benzene *Chem. Phys. Lett.* **177** 4–10
- [118] Joo T and Albrecht A C 1993 Femtosecond time-resolved coherent anti-Stokes Raman spectroscopy of liquid benzene: a Kubo relaxation function analysis *J. Chem. Phys.* **99** 3244–51
- [119] Okamoto H and Yoshihara K 1991 Femtosecond time-resolved coherent Raman scattering from β -carotene in solution. Ultrahigh frequency (11 THz) beating phenomenon and sub-picosecond vibrational relaxation *Chem. Phys. Lett.* **177** 568–71
- [120] Yan Y X, Gamble E B and Nelson K A 1985 Impulsive stimulated Raman scattering: general importance in femtosecond laser pulse interactions with matter, and spectroscopic applications *J. Chem. Phys.* **83** 5391–9
- [121] Ruhman S, Joly A G and Nelson K A 1987 Time-resolved observations of coherent molecular vibrational motion and the general occurrence of impulsive stimulated scattering *J. Chem. Phys.* **86** 6563–5
- [122] Walsh A M and Loring R F 1989 Theory of resonant and nonresonant impulsive stimulated Raman scattering *Chem. Phys. Lett.* **160** 299–304
- [123] Constantine S, Zhou Y, Morais J and Ziegler L D 1997 Dispersed optical heterodyne birefringence and dichroism of transparent liquids *J. Phys. Chem. A* **101** 5456–62
- [124] Walmsley I A, Wise F W and Tang C L 1989 On the difference between quantum beats in impulsive stimulated Raman scattering and resonance Raman scattering *Chem. Phys. Lett.* **154** 315–20
- [125] Loring R F and Mukamel S 1985 Selectivity in coherent transient Raman measurements of vibrational dephasing in liquids *J. Chem. Phys.* **83** 2116–28
- [126] Vanden Bout D and Berg M 1995 Ultrafast Raman echo experiments in liquids *J. Raman Spectrosc.* **26** 503–11
- [127] Muller L J, Vanden Bout D and Berg M 1993 Broadening of vibrational lines by attractive forces: ultrafast Raman echo experiments in a $\text{CH}_3\text{I}:\text{CDCl}_3$ mixture *J. Chem. Phys.* **99** 810–19
- [128] Vanden Bout D, Fretas J E and Berg M 1994 Rapid, homogeneous vibrational dephasing in ethanol at low temperatures determined by Raman echo measurements *Chem. Phys. Lett.* **229** 87–92
- [129] Muller L J, Vanden Bout D and Berg M 1991 Ultrafast Raman echoes in liquid acetonitrile *Phys. Rev. Lett.* **67** 3700–3
- [130] Tokmakoff A and Fleming G R 1997 Two-dimensional Raman spectroscopy of the intermolecular modes of liquid CS_2 *J. Chem. Phys.* **106** 2569–82
- [131] Tokmakoff A, Lang M J, Larson D S, Fleming G R, Chernyak V and Mukamel S 1997 Two-dimensional Raman spectroscopy of vibrational interactions in liquids *Phys. Rev. Lett.* **79** 2702–5
- [132] Steffen T and Duppen K 1996 Time resolved four- and six-wave mixing in liquids I. Theory *J. Chem. Phys.* **105** 7364–82
-

- [133] Steffen T and Duppen K 1996 Time resolved four- and six-wave mixing in liquids II. *Experiment J. Chem. Phys.* **106** 3854–64
- [134] Khidekel V and Mukamel S 1995 High-order echoes in vibrational spectroscopy of liquids *Chem. Phys. Lett.* **240** 304–14
- [135] Ivanecky J E III and Wright J C 1993 An investigation of the origins and efficiencies of higher order nonlinear spectroscopic processes *Chem. Phys. Lett.* **206** 437–44
- [136] Ernst R R, Bodenhausen G and Wokaun A 1987 *Principles of Nuclear Magnetic Resonance in One and Two Dimensions* (Oxford: Clarendon)
- [137] Turrell G and Dhamelincourt P 1996 Micro-Raman spectroscopy *Modern Techniques in Raman Spectroscopy* ed J J Laserna (New York: Wiley) pp 109–42
- [138] Barron L D, Hecht L, Bell A F and Wilson G 1996 Raman optical activity: an incisive probe of chirality and biomolecular structure and dynamics *ICORS '96: XVth Int. Conf. on Raman Spectroscopy* ed S A Asher and P B Stein (New York: Wiley) pp 1212–15
- [139] Hecht L and Barron L D 1996 Raman optical activity *Modern Techniques in Raman Spectroscopy* ed J J Laserna (New York: Wiley) pp 265–342
- [140] Ren B, Li W H, Mao B W, Gao J S and Tian Z Q 1996 Optical fiber Raman spectroscopy combined with scattering tunneling microscopy for simultaneous measurements *ICORS '96: XVth Int. Conf. on Raman Spectroscopy* ed S A Asher and P B Stein (New York: Wiley) pp 1220–1
- [141] Barrett J J and Berry M J 1979 Photoacoustic Raman spectroscopy (PARS) using cw laser sources *Appl. Phys. Lett.* **34** 144–6
- [142] Siebert D R, West G A and Barrett J J 1980 Gaseous trace analysis using pulsed photoacoustic Raman spectroscopy *Appl. Opt.* **19** 53–60
- [143] Koroteev N I 1995 BioCARS—a novel nonlinear optical technique to study vibrational spectra of chiral biological molecules in solution *Biospectroscopy* **1** 341–50
- [144] Koroteev N I 1996 Optical rectification, circular photogalvanic effect and five-wave mixing in optically active solutions *Proc. SPIE* **2796** 227–38
- [145] Akhmanov S A and Koroteev N I 1981 *Methods of Nonlinear Optics in Light Scattering Spectroscopy* (Moscow: Nauka) (in Russian)
- [146] Cho M 1997 Off-resonant coherent hyper-Raman scattering spectroscopy *J. Chem. Phys.* **106** 7550–7
- [147] Yang M, Kim J, Jung Y and Cho M 1998 Six-wave mixing spectroscopy: resonant coherent hyper-Raman scattering *J. Chem. Phys.* **108** 4013–20
- [148] Ziegler L D 1990 Hyper-Raman spectroscopy *J. Raman Spectrosc.* **71** 769–79
- [149] Ziegler L D and Roebber J L 1987 Resonance hyper-Raman scattering of ammonia *Chem. Phys. Lett.* **136** 377–82
- [150] Chung Y C and Ziegler L D 1988 The vibronic theory of resonance hyper-Raman scattering *J. Chem. Phys.* **88** 7287–94
- [151] Golab J T, Sprague J R, Carron K T, Schatz G C and Van Duyne R P 1988 A surface enhanced hyper-Raman scattering study of pyridine adsorbed onto silver: experiment and theory *J. Chem. Phys.* **88** 7942–51

- [152] Kneipp K, Kneipp H and Seifert F 1994 Near-infrared excitation profile study of surface-enhanced hyper-Raman scattering and surface-enhanced Raman scattering by means of tunable mode-locked

Ti:sapphire laser excitation *Chem. Phys. Lett.* **233** 519–24

- [153] Yu N-T, Nie S and Lipscomb L 1990 Surface-enhanced hyper-Raman spectroscopy with a picosecond laser. New vibrational information for non-centrosymmetric carbocyanine molecules adsorbed on colloidal silver *J. Raman Spectrosc.* **21** 797–802
- [154] Baranov A V, Bobovich Y S and Petrov V I 1993 Surface-enhanced resonance hyper-Raman (SERHR) spectroscopy of photochromic molecules *J. Raman Spectrosc.* **24** 695–7
- [155] McCreery R L, Liu Y-C, Kagen M, Chen P and Fryling M 1996 Resonance and normal Raman spectroscopy of carbon surfaces: relationships of surface structure and reactivity *ICORS '96: XVth Int. Conf. on Raman Spectroscopy* ed S A Asher and P B Stein (New York: Wiley) pp 566–7
- [156] Evans R, Smith I, Münz W D, Williams K J P and Yarwood J 1996 Raman microscopic studies of ceramic coatings based on titanium aluminum nitride *ICORS '96: XVth Int. Conf. on Raman Spectroscopy* ed S A Asher and P B Stein (New York: Wiley) pp 596–7
- [157] Al-Obaidi A H R, Rigby S J, Hegarty J N M, Bell S E J and McGarvey J J 1996 Direct formation of silver and gold metal liquid-like films (MELLFS) from thiols and sols without organic solvents: SERS and AFM studies *ICORS '96: XVth Int. Conf. on Raman Spectroscopy* ed S A Asher and P B Stein (New York: Wiley) pp 590–1
- [158] Hall R J and Boedeker L R 1984 CARS thermometry in fuel-rich combustion zones *Appl. Opt.* **23** 1340–6
- [159] Stenhouse I A, Williams D R, Cole J B and Swords M D 1979 CARS measurements in an internal combustion engine *Appl. Opt.* **18** 3819–25
- [160] Bechtel J H and Chraplyvy A R 1982 Laser diagnostics of flame, combustion products and sprays *Proc. IEEE* **70** 658–77
- [161] Eckbreth A C, Dobbs G M, Stufflebeam J H and Tellex P A 1984 CARS temperature and species measurements in augmented jet engine exhausts *Appl. Opt.* **23** 1328–39
- [162] Gross L P, Trump D D, MacDonald B G and Switzer G L 1983 10-Hz coherent anti-Stokes Raman spectroscopy apparatus for turbulent combustion studies *Rev. Sci. Instrum.* **54** 563–71
- [163] Eckbreth A C 1988 Nonlinear Raman spectroscopy for combustion diagnostics *J. Quant. Spectrosc. Radiat. Transfer* **40** 369–83
- [164] Walrafen G E 1967 Raman spectral studies of the effects of temperature on water structure *J. Chem. Phys.* **47** 114–26
- [165] De Santis A, Sampoli M, Mazzacurati V and Ricci M A 1987 Raman spectra of water in the translational region *Chem. Phys. Lett.* **133** 381–4
- [166] Nardone M, Ricci M A and Benassi P 1992 Brillouin and Raman scattering from liquid water *J. Mol. Struct.* **270** 287–99
- [167] Carey D M and Korenowski G M 1998 Measurement of the Raman spectrum of liquid water *J. Chem. Phys.* **108** 2669–75
- [168] Shim M G and Wilson B C 1997 Development of an in vivo Raman spectroscopic system for diagnostic applications *J. Raman Spectrosc.* **28** 131–42

-1-

B1.4 Microwave and terahertz spectroscopy

Geoffrey A Blake

B1.4.1 INTRODUCTION

Spectroscopy, or the study of the interaction of light with matter, has become one of the major tools of the natural and physical sciences during this century. As the wavelength of the radiation is varied across the electromagnetic spectrum, characteristic properties of atoms, molecules, liquids and solids are probed. In the

optical and ultraviolet regions ($\lambda \sim 1 \mu\text{m}$ up to 100 nm) it is the electronic structure of the material that is investigated, while at infrared wavelengths ($\sim 1\text{--}30 \mu\text{m}$) the vibrational degrees of freedom dominate.

Microwave spectroscopy began in 1934 with the observation of the $\sim 20 \text{ GHz}$ absorption spectrum of ammonia by Cleeton and Williams. Here we will consider the microwave region of the electromagnetic spectrum to cover the 1 to $100 \times 10^9 \text{ Hz}$, or 1 to 100 GHz ($\lambda \sim 30 \text{ cm}$ down to 3 mm), range. While the ammonia microwave spectrum probes the inversion motion of this unique pyramidal molecule, more typically microwave spectroscopy is associated with the pure rotational motion of gas phase species.

The section of the electromagnetic spectrum extending roughly from 0.1 to $10 \times 10^{12} \text{ Hz}$ ($0.1\text{--}10 \text{ THz}$, $3\text{--}300 \text{ cm}^{-1}$) is commonly known as the far-infrared (FIR), submillimetre or terahertz (THz) region, and therefore lies between the microwave and infrared windows. Accordingly, THz spectroscopy shares both scientific and technological characteristics with its longer- and shorter-wavelength neighbours. While rich in scientific information, the FIR or THz region of the spectrum has, until recently, been notoriously lacking in good radiation sources—earning the dubious nickname ‘the gap in the electromagnetic spectrum’. At its high-frequency boundary, most coherent photonic devices (e.g. diode lasers) cease to radiate due to the long lifetimes associated with spontaneous emission at these wavelengths, while at its low-frequency boundary parasitic losses reduce the oscillatory output from most electronic devices to insignificant levels. As a result, existing coherent sources suffer from a number of limitations. This situation is unfortunate since many scientific disciplines—including chemical physics, astrophysics, cosmochemistry and planetary/atmospheric science to name but a few—rely on *high-resolution* THz spectroscopy (both in a spectral and temporal sense). In addition, technological applications such as ultrafast signal processing and massive data transmission would derive tremendous enhancements in rate and volume throughput from frequency-agile THz synthesizers.

In general, THz frequencies are suitable for probing low-energy light–matter interactions, such as rotational transitions in molecules, phonons in solids, plasma dynamics, electronic fine structure in atoms, thermal imaging of cold sources and vibrational–rotation–tunnelling behaviour in weakly bound clusters. Within the laboratory, THz spectroscopy of a variety of molecules, clusters and condensed phases provides results that are critical to a proper interpretation of the data acquired on natural sources, and also leads to a better understanding of important materials—particularly hydrogen-bonded liquids, solids and polymers that participate in a variety of essential (bio)chemical processes.

For remote sensing, spectroscopy at THz frequencies holds the key to our ability to remotely sense environments as diverse as primaeval galaxies, star and planet-forming molecular cloud cores, comets and planetary atmospheres.

-2-

In the dense interstellar medium characteristic of sites of star formation, for example, scattering of visible/UV light by sub-micron-sized dust grains makes molecular clouds optically opaque and lowers their internal temperature to only a few tens of Kelvin. The thermal radiation from such objects therefore peaks in the FIR and only becomes optically thin at even longer wavelengths. Rotational motions of small molecules and rovibrational transitions of larger species and clusters thus provide, in many cases, the only or the most powerful probes of the dense, cold gas and dust of the interstellar medium.

Since the major drivers of THz technology have been scientists, particularly physicists and astrophysicists seeking to carry out fundamental research, and not commercial interests, a strong coupling of technology development efforts for remote sensing with laboratory studies has long characterized spectroscopy at microwave and THz frequencies. In many respects the field is still in its infancy, and so this chapter will present both an overview of the fundamentals of microwave and THz spectroscopy as well as an assessment of the current technological state of the art and the potential for the future. We will begin with a brief overview of the general characteristics of THz spectrometers and the role of incoherent sources and detection

strategies in the THz region, before turning to a more detailed description of the various coherent THz sources developed over the past decade and their applications to both remote sensing and laboratory studies.

B1.4.2 INCOHERENT THZ SOURCES AND BROADBAND SPECTROSCOPY

B1.4.2.1 PRINCIPLES AND INSTRUMENTATION

Like most other fields of spectroscopy, research at THz frequencies in the first half of the twentieth century was carried out with either dispersive (i.e. grating-based) or Fourier transform spectrometers. The much higher throughput of Fourier transform spectrometers compared to those based on diffraction gratings has made THz Fourier transform spectroscopy, or THz FTS, the most popular incoherent technique for acquiring data over large regions of the THz spectrum. This is especially true for molecular line work where THz FTS resolutions of order 50–100 MHz or better have been obtained [1]. With large-format detector arrays, such as those available at optical through near- to mid-infrared wavelengths, grating- or Fabry–Perot-based instruments can provide superior sensitivity [2], but have not yet been widely utilized at THz frequencies due to the great difficulty of fabricating arrays of THz detectors.

The components of THz spectrometers can be grouped into three main categories: sources (e.g. lasers, Gunn oscillators, mercury-discharge lamps), propagating components (e.g. lenses, sample cells, filters) and detectors (e.g. bolometers, pyroelectric detectors, photoacoustic cells). Propagating components in the FIR are well established (see [3] for an excellent overview of technical information). In the area of detectors, recent progress has placed them ahead of source technology. For example, spider-web Si bolometers developed by Bock *et al* [4] have an electrical NEP (noise-equivalent power) of $4 \times 10^{17} \text{ W Hz}^{-1/2}$ when cooled to 300 mK. For those who desire less exotic cryogenic options, commercially available Si-composite bolometers offer an electrical NEP of $1 \times 10^{-13} \text{ W Hz}^{-1/2}$, operating at liquid helium temperature (4.2 K), and an electrical NEP of $3 \times 10^{-15} \text{ W Hz}^{-1/2}$, operating at 1.2 K (pumped L_{He}). Combining these into large-format arrays remains a considerable technological challenge, although arrays of several tens of pixels on a side are now beginning to make their way into various telescopes such as the Caltech Submillimeter Observatory and the James Clerk Maxwell Telescope [5, 6]. In addition, their electrical bandwidths are typically only between a few hundred hertz and 1 kHz. Fast-modulation schemes cannot therefore be used, and careful attention must be paid to $1/f$ noise in experiments with Si bolometers. Hot-electron bolometers based on InSb offer electrical

-3-

bandwidths of 1 MHz, but without cyclotron-assisted resonance, InSb THz bolometers cannot be used above $\nu = 25\text{--}30 \text{ cm}^{-1}$ [7]. Similarly, photoconductors based on Ga:Ge offer high electrical speed and good quantum efficiency, but due to the bandgap of the material are unusable below $50\text{--}60 \text{ cm}^{-1}$ [8].

In practice, the NEP of a room-temperature THz spectrometer is usually limited by fluctuations (shot-noise) in the ambient blackbody radiation. Using an optical bandwidth $\Delta\nu = 3 \text{ THz}$ (limited by, for example, a polyethylene/diamond dust window), a field of view (at normal incidence) $\theta = 9^\circ$ and a detecting diameter (using a so-called Winston cone, which condenses the incident radiation onto the detecting element) $d = 1.1 \text{ cm}$, values that are typical for many laboratory applications, the background-limited NEP of a bolometer is given by

(B1.4.1)

where k is the Boltzmann constant, $T = 300 \text{ K}$, Δf (the electronic amplified bandwidth) = 1 Hz, and λ (the band-centre wavelength) $\approx 200 \text{ }\mu\text{m}$. The equation above uses the Rayleigh–Jeans law, which is valid for $\nu \gg 17$

THz. Therefore, for laboratory absorption experiments, a typical FIR detector provides an estimated detection limit (NEP/source power) of 10^{-4} with a source output of 20 nW. In general, high-sensitivity bolometers saturate at an incident-power level of $\approx 1 \mu\text{W}$ or less, resulting in an ultimate detection limit of 10^{-7} . For yet higher dynamic range, a filter element (e.g. cold grating, prism, or etalon) must be placed before the detector to reduce background noise, or the background temperature must be lowered. Note the $(\Delta\nu)^{1/2}$ dependence in equation (B1.4.1), which means that the optical bandwidth must be reduced to ~ 30 GHz to drop the NEP by a factor of ten. *Thus, unlike shorter wavelength regions of the electromagnetic spectrum, due to the high background luminosity in the THz, spectroscopic sensitivity in the laboratory is limited by the source power, in comparison to the background power, incident on the detector—not by shot-noise of the available spectroscopic light sources.* As higher-power light sources are developed at THz frequencies, lower-NEP detectors can be utilized that are less prone to saturation, and shot-noise will become the limiting factor, as it is in other regions of the electromagnetic spectrum.

For broadband THz FTS instruments this large background actually leads to a ‘multiplex disadvantage’ in that the room-temperature laboratory background can easily saturate the sensitive THz detectors that are needed to detect the feeble output of incoherent THz blackbody sources, which drop rapidly as the wavelength increases. The resulting sensitivity is such that signal-to-noise ratios in excess of 100 are difficult to generate at the highest feasible resolutions of 50–100 MHz [1], which is still quite large compared to the 1–2 MHz Doppler-limited line widths at low pressure. For low-resolution work on condensed phases, or for the acquisition of survey spectra, however, THz FTS remains a popular technique. Beyond wavelengths of ~ 1 mm, the sensitivity of FTS is so low that the technique is no longer competitive with the coherent approaches described below.

B1.4.2.2 THZ FTS STUDIES OF PLANETARY ATMOSPHERES

Rather different circumstances are encountered when considering THz remote sensing of extraterrestrial sources. The major source of THz opacity in the Earth’s atmosphere is water vapour, and from either high, dry mountain sites or from space there are windows in which the background becomes very small. Incoherent instruments which detect the faint emission from astronomical sources can therefore be considerably more sensitive than their laboratory

-4-

counterparts. Again, grating- or etalon-based and FTS implementations can be considered, with the former being preferred if somewhat coarse spectral resolution is desired or if large-format detector arrays are available.

In planetary atmospheres, a distinct advantage of THz studies over those at optical and infrared wavelengths is the ability to carry out spectroscopy without a background or input source such as the sun. Global maps of a wide variety of species can therefore be obtained at any time of day or night and, when taken at high enough spectral resolution, the shapes of the spectral lines themselves also contain additional information about vertical abundance variations and can be used to estimate the atmospheric temperature profile. The ‘limb sounding’ geometry, in which microwave and THz emission from the limb of a planetary atmosphere is imaged by an orbiting spacecraft or a balloon, is particularly powerful in this regard, and excellent reviews are available on this subject [9]. The observing geometry is illustrated in [figure B1.4.1](#) which also presents a portion of the Earth’s stratospheric emission spectrum near 118 cm^{-1} obtained by the balloon-borne Smithsonian Astrophysical Observatory limb-sounding THz FTS [10].

B1.4.3 COHERENT THZ SOURCES AND HETERODYNE SPECTROSCOPY

B1.4.3.1 PRINCIPLES AND INSTRUMENTATION

The narrow cores of atmospheric transitions shown in [figure B1.4.1](#) can be used, among other things, to trace the wind patterns of the upper atmosphere. For such work, or for astronomical remote-sensing efforts, resolutions of the order of $30\text{--}300\text{ m s}^{-1}$ are needed to obtain pressure broadening or kinematic information, which correspond to spectral resolutions of $(\nu/\Delta\nu) \sim 1\text{--}10 \times 10^6$. Neither FTS nor grating-based spectrometers can provide resolution at this level and so other techniques based on coherent radiation sources must be used. The most important of these is called heterodyne spectroscopy. Heterodyne spectroscopy uses nonlinear detectors called mixers, in order to downconvert the high-frequency THz radiation into radiofrequency or microwave signals that can be processed using commercial instrumentation. An outline of a heterodyne receiver is presented in [figure B1.4.2](#). In such a receiver, an antenna (dish) collects radiation from space, and this radiation is focused onto a detector operating in heterodyne mode, which means that the incoming signal is mixed with the output of a coherent source (called the local oscillator, or LO). Now, if a device can be constructed that responds quadratically, rather than linearly, to the two input beams, the output, $S(t)$, of such a device is given by

$$\begin{aligned}
 S(t) &\propto e(t)^2 \\
 S(t) &\propto 1/2(E_1^2 + E_2^2) + 1/2[E_1^2 \cos(2\omega_1 t) + E_2^2 \cos(2\omega_2 t)] \\
 &\quad + E_1 E_2 \cos[(\omega_1 t + \omega_2 t) + (\phi_1 + \phi_2)] \\
 &\quad + \underbrace{E_1 E_2 \cos[(\omega_1 - \omega_2)t + (\phi_1 - \phi_2)]}_{\text{DIFFERENCE-FREQUENCY FIELD}}
 \end{aligned}
 \tag{B1.4.2}$$

where the identities $2\cos\alpha\cos\beta = \cos(\alpha+\beta) + \cos(\alpha-\beta)$ and $\cos^2\alpha = 1/2(1 + \cos 2\alpha)$ are invoked. The last term is a field oscillating at a frequency equal to the difference between the two incidental fields—representing the beat-note between ν_1 and ν_2 .

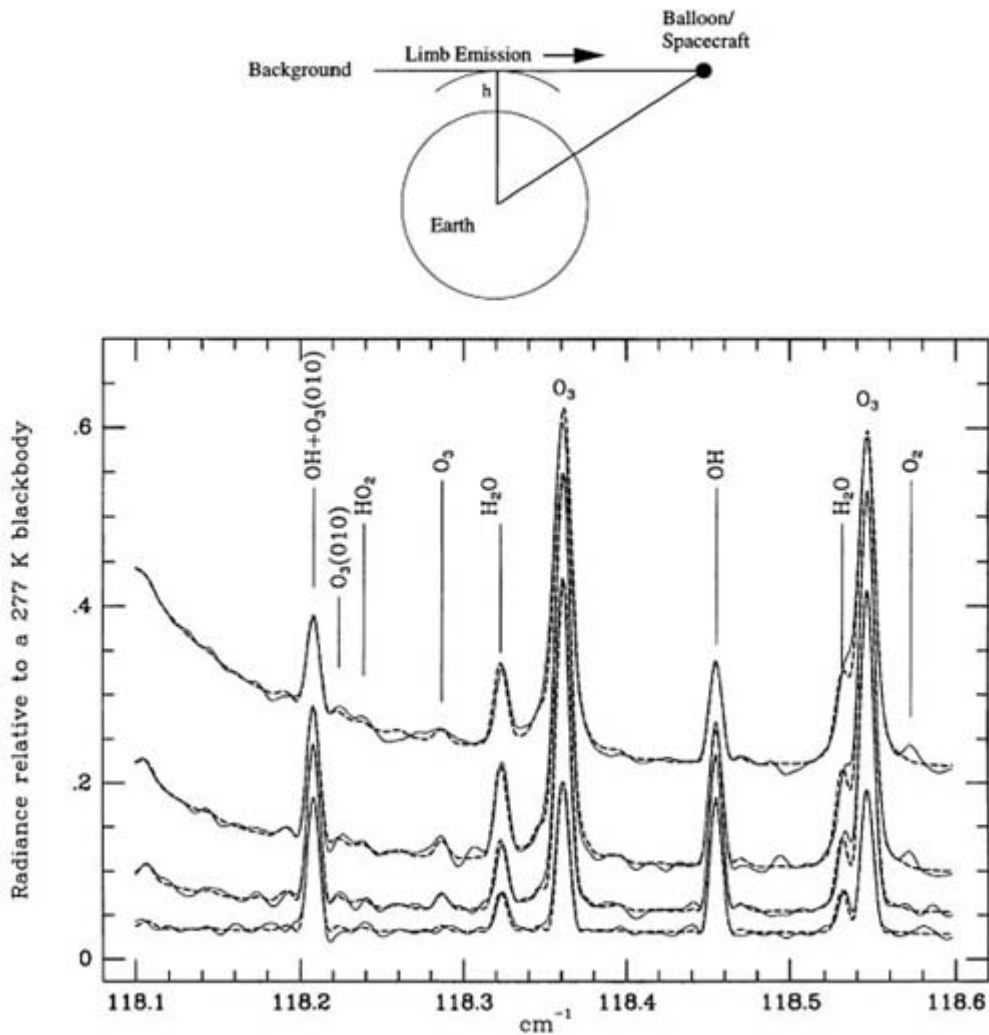


Figure B1.4.1. Top: schematic illustration of the observing geometry used for limb sounding of the Earth's atmosphere. Bottom: illustrative stratospheric OH emission spectra acquired by the SAO FIRS-2 far-infrared balloon-borne FTS in autumn 1989. The spectra are from a range of tangent heights (h = tangent height in the drawing), increasing toward the bottom, where the data are represented by solid curves; nonlinear least-square fits to the measurements, based on a combination of laboratory data, the physical structure of the stratosphere and a detailed radiative transfer calculation, are included as dashed curves. The OH lines are $F_1({}^2\Pi_{3/2}, 7/2^- \rightarrow 5/2^+$ and $7/2^+ \rightarrow 5/2^-$ (the hyperfine structure is unresolved in these measurements). Other major contributing lines are also identified [10].

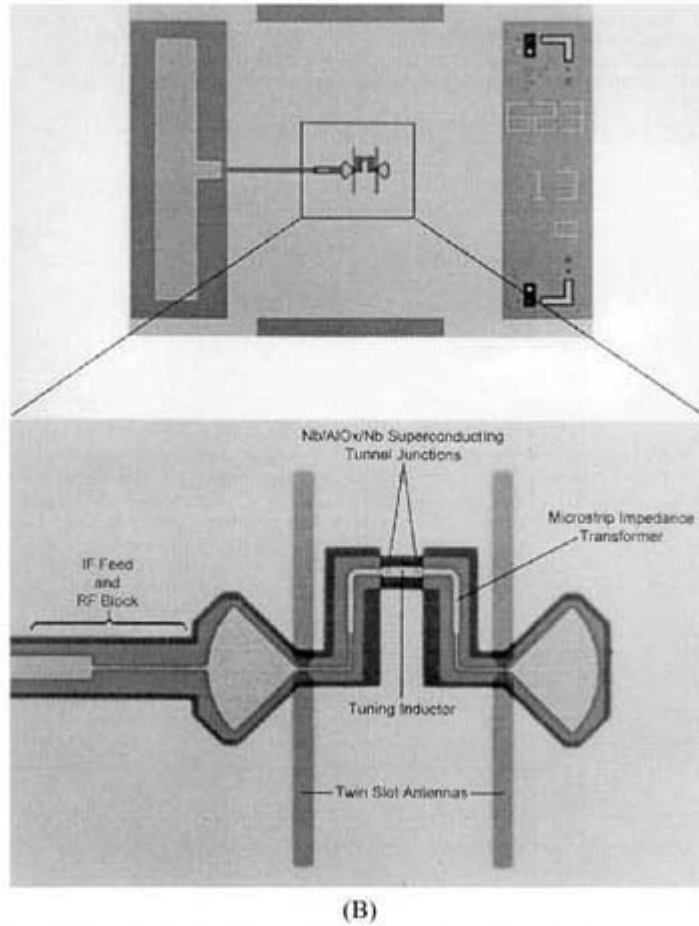
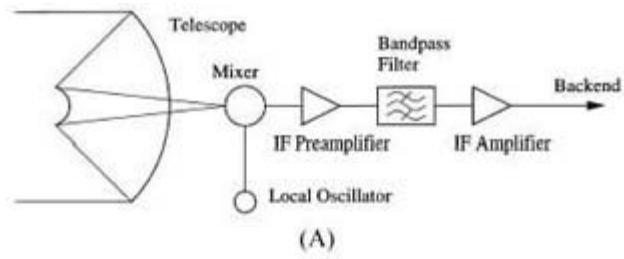


figure continued on next page.

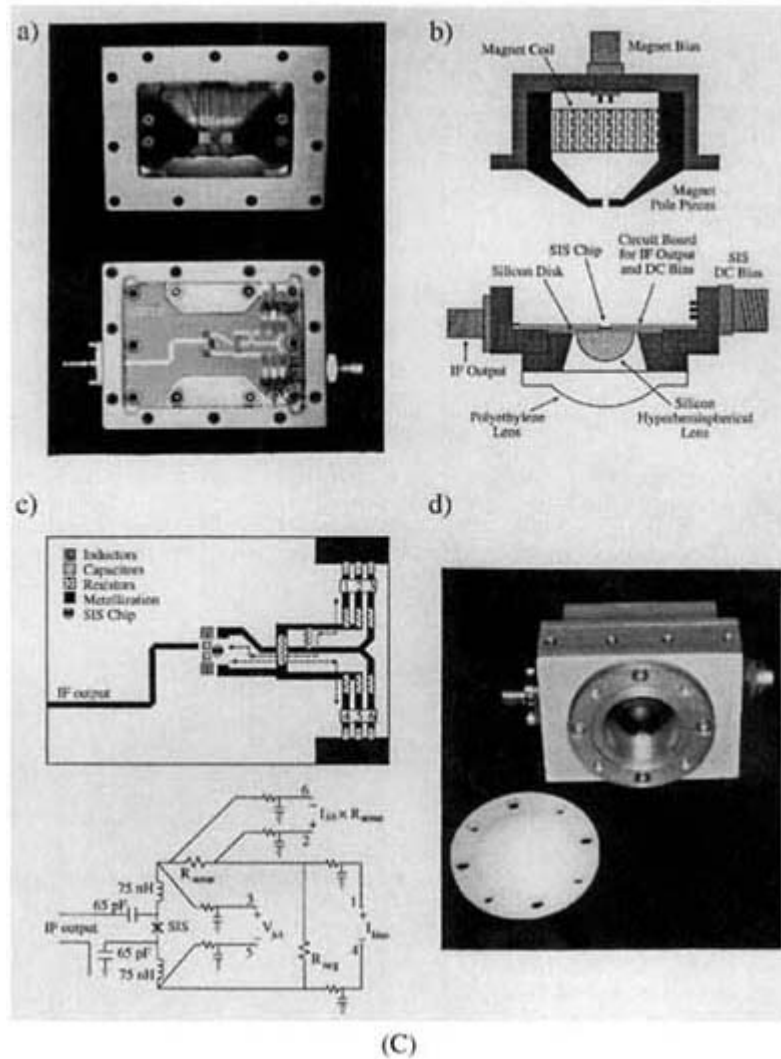


Figure B1.4.2. (A) Basic components of an astronomical heterodyne receiver. The photomicrograph in (B) presents the heart of a quasi-optical SIS mixer and its associated superconducting tuning circuits, while the image in (C) shows the fully assembled mixer, as it would be incorporated into a low-temperature cryostat (J Zmuidzinas, private communication).

If $\nu_1 = \nu_{LO}$ and $\nu_2 = \nu_{SKY} \approx \nu_{LO}$, the sum frequency lies at THz frequencies, while the difference lies at radio or microwave frequencies, and is called the intermediate frequency, or IF. The IF can be amplified and recorded (for example, on a spectrum analyser or by a digital correlator or filterbank) across a range of frequencies simultaneously. The fact that the IF power is proportional to the product of the remote-signal power and the LO power results in two main advantages: (a) the signal-to-noise ratio is enhanced by using an LO power that is much higher than that of the remote signal and (b) the spectral resolution is set by the linewidth of the LO, which can be as narrow as desired.

Among the first THz mixers to be constructed were those based on room-temperature Schottky diodes [11]. Over the past decade, new mixers based on superconducting tunnel junctions have been developed that have effective noise levels only a few times the quantum limit of $T_{\text{mixer}} = h\nu/k$ [12]. However, certain conditions must be met in order to exploit these advantages. For example, while the LO power should be strong compared to the remote signal and to the noise, it must not be so strong as to saturate the detecting element. Roughly speaking, the optical coupled LO power level is about 1 μW for SIS (superconductor–insulator–superconductor) mixers and 100 nW for superconductive Nb HEBs (hot-electron bolometers, or transition

edge bolometers, TEBs). In addition, since the overlap between the remote signal and the LO is important, the spatial distribution of the LO output must be well coupled to the receiver's antenna mode. Although this requirement imposes experimental complexity, it also provides excellent rejection of ambient background radiation.

As noted above, at THz frequencies the Rayleigh–Jeans approximation is a good one, and it is typical to report line intensities and detector sensitivities in terms of the Rayleigh–Jeans equivalent temperatures. In frequencies range where the atmospheric transmission is good, or from airborne or space-borne platforms, the effective background temperature is only a few tens of Kelvin. Under such conditions, SIS mixers based on Nb, a particular implementation of which is pictured in [figure B1.4.2](#) can now perform up to 1.0 THz with $T_{\text{mixer}} = 130$ K [13]. The earliest SIS microwave and millimetre-wave receivers utilized waveguide components, but as the operating frequencies have been pushed into the THz region, quasi-optical designs such as those shown in [figure B1.4.2](#) become attractive. Such designs may also be easier to incorporate into THz receiver arrays. Recently, alternative superconducting Nb hot-electron mixers that rely on a diffusion-based relaxation mechanism have been demonstrated with $T_{\text{mixer}} = 750$ K between 1 and 2.5 THz [14]. These devices are expected to operate up to at least several tens of THz—if coherent sources are available as LOs. Thus, THz source technology plays a key role in setting the spectroscopic sensitivity for both laboratory and remote-sensing experiments.

B1.4.3.2 THZ REMOTE SENSING WITH HETERODYNE RECEIVERS

Heterodyne spectroscopy has been particularly critical to the study of the Earth's stratosphere, where the improved resolution and sensitivity compared to the FTS spectra shown in [figure B1.4.1](#) have led to the collection of global maps of species important to ozone chemistry and atmospheric dynamics (O_3 , ClO, SO_2 , H_2O , O_2 , etc: see [9] for an extended overview), and of the dense interstellar medium. Although human beings have been systematically observing astronomical objects for thousands of years, until the advent of radioastronomy in the 1960s we possessed little knowledge of what, if anything, exists in the space between stars. Optical observations revealed only stars, galaxies and nebulae; if matter existed in the vast, dark interstellar medium, it was not detectable. However, the discovery of the first polyatomic microwave in the interstellar medium, water (H_2O), ammonia (NH_3) and formaldehyde (H_2CO), by microwave remote sensing (in 1968 [15]) set off an exciting era of discovery.

To date, researchers have identified more than 100 different molecules, composed of up to 13 atoms, in the interstellar medium [16]. Most were initially detected at microwave and (sub)millimetre frequencies, and the discoveries have reached far beyond the mere existence of molecules. Newly discovered entities such as diffuse interstellar clouds, dense (or dark) molecular clouds and giant molecular cloud complexes were characterized for the first time. Indeed, radioastronomy (which includes observations ranging from radio to submillimetre frequencies) has dramatically changed our perception of the composition of the universe. Radioastronomy has shown that most of the mass in the interstellar medium is contained in so-called dense molecular clouds, which have tremendous sizes of 1–100 light years, average gas densities of 10^2 – 10^3 cm^{-3} , and temperatures in the range of 10–600 K. An overview of the THz emission from a cold, dense interstellar cloud is presented in [figure B1.4.3](#) [17].

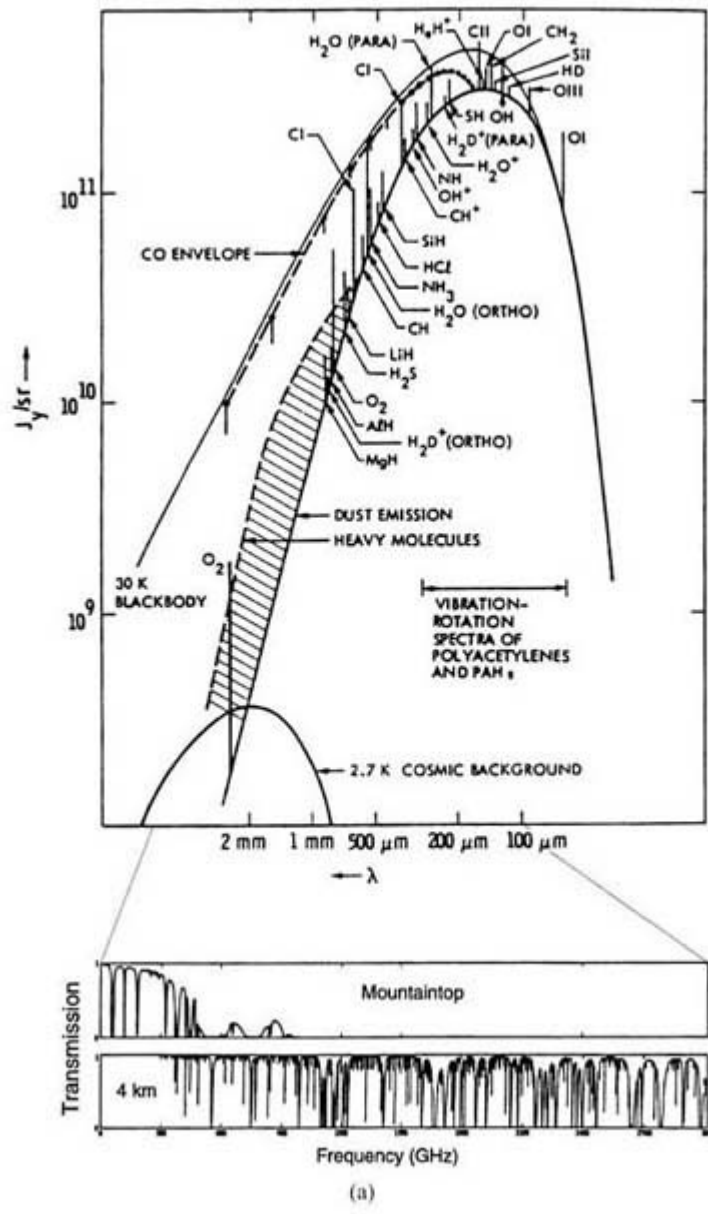


figure continued on next page.

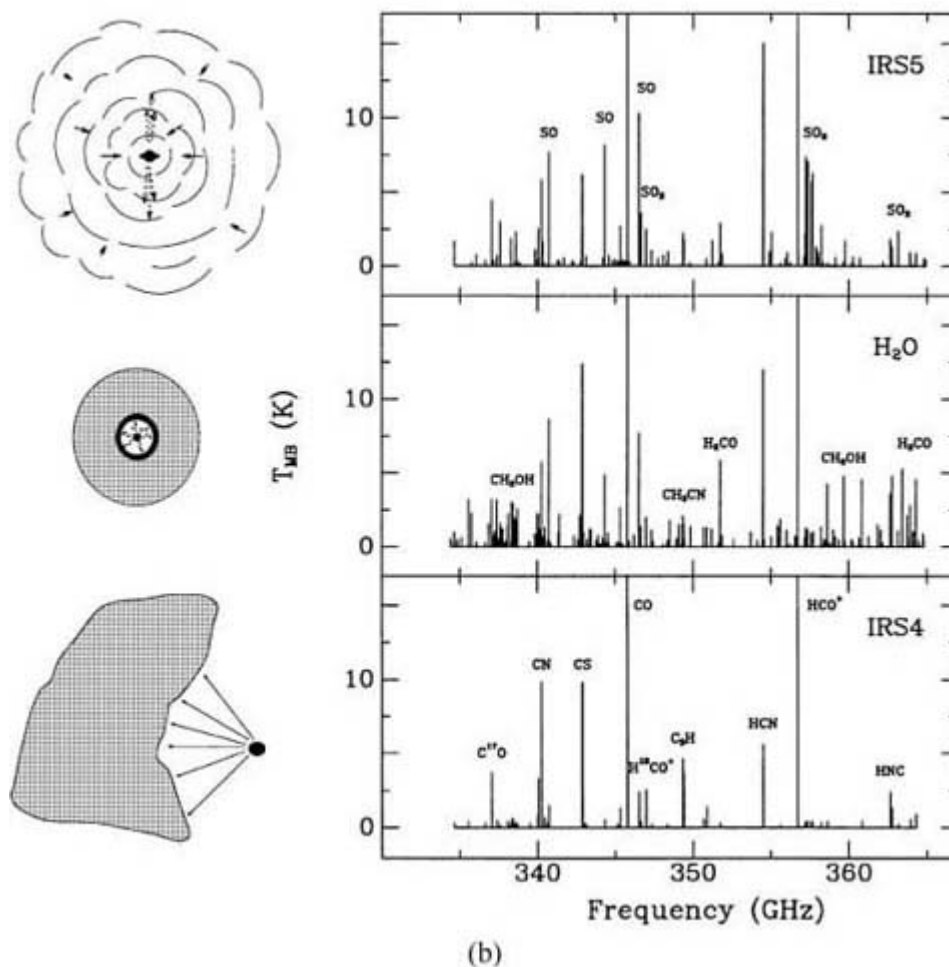


Figure B1.4.3. (a) A schematic illustration of the THz emission spectrum of a dense molecular cloud core at 30 K and the atmospheric transmission from ground and airborne altitudes (adapted, with permission, from [17]). (b) The results of 345 GHz molecular line surveys of three cores in the W3 molecular cloud; the graphics at left depict the evolutionary state of the dense cores inferred from the molecular line data [21].

In addition to striking differences from one cloud to another, each dense molecular cloud is inhomogeneous, containing clumps, or cores, of higher-density material situated within envelopes of somewhat lower density. Many of these higher-density cores are active sites of star formation, with the youngest stars being detectable only in the IR or FIR. Star formation is of major interest in astrophysics, and it contains a wealth of interesting chemical reactions and physical phenomena (for excellent reviews, see [18, 19 and 20]). Optical observations are unable to characterize interstellar clouds due to absorption and scattering, both of which have an inverse wavelength dependence, by the pervasive dust particles inside these clouds. Thus, microwave and THz spectroscopy is responsible for identifying most of the hundred or so interstellar molecules to date, and continues to dominate the fields of molecular astrophysics and interstellar chemistry. The power of heterodyne spectroscopy in examining the differences between dense clouds undergoing star (and presumably planet) formation is shown in the right panel of [figure B1.4.3](#) which depicts the

345 GHz spectral line surveys of three regions of the W3 giant molecular cloud complex [21]. From such studies, which reveal dramatic differences in the THz spectrum of various objects, molecular astrophysicists hope to classify the evolutionary state of the cloud, just as optical spectra are used to classify stars.

High angular resolution studies with modern THz telescopes and interferometric arrays can even probe the

material destined to become part of planetary systems. In the accretion discs around young stars, for example, a range of simple organic species have now been detected at (sub)millimetre wavelengths [22]. Such accretion disks are the assembly zones of planets, and the first steps in THz imaging their outer regions and in understanding the means by which they evolve have been taken [23]. The physical and chemical conditions in these objects can now be compared to that observed in primitive solar system objects such as comets and icy satellites. The recent apparitions of comets Hyakutake and Hale–Bopp, for example, have provided a wealth of new observations at IR, THz and microwave frequencies that have led to a much improved understanding of the origin and evolution of planetesimals in the outer solar system [24]. Future work in high-resolution THz imaging will be dramatically enhanced by the Atacama Large Millimeter Array (ALMA), which will operate over the 1 cm to 350 μm interval at an altitude of 16 000 ft in the Atacama desert of northern Chile [25].

Ultimately, studies from ground-based observatories are limited by absorption in the Earth's atmosphere. For example, no studies above ≈ 1 THz are possible even from mountain-top observatories. Two major instruments, SOFIA and FIRST, are poised to change this situation dramatically. SOFIA (for Stratospheric Observatory For Infrared Astronomy [26]) will carry a 2.7 m telescope in a 747SP aircraft to altitudes of 41 000–45 000 feet. At these altitudes, nearly 60–70% of the THz spectrum up to the mid-IR is accessible (figure B1.4.3), and both incoherent and heterodyne spectrometers are being constructed as part of the initial instrument suite. SOFIA will become operational in 2002. On somewhat longer timescales, FIRST (the Far-InfraRed Space Telescope [27]) will carry 0.4–1.2 THz SIS and 1.2–1.9/2.4–2.7 THz antenna coupled HEB receivers at the focal plane of a 4 m telescope into space. Like SOFIA, FIRST will also include incoherent spectrometers and imagers to take advantage of the low background flux and to survey larger regions of the THz spectrum and sky.

B1.4.4 SPECTROSCOPY WITH TUNABLE MICROWAVE AND THZ SOURCES

B1.4.4.1 PRINCIPLES AND BACKGROUND

Long before the invention of the laser, coherent radiation sources such as electron beam tubes (e.g. klystrons and backward wave oscillators, or BWOs) were in use by microwave spectroscopists to examine the direct absorption rotational spectroscopy of molecules. Being the interface between microwave spectroscopy (generally associated with low-energy rotational transitions of molecules) and mid-IR spectroscopy (generally associated with vibrational transitions of molecules), THz spectra such as those outlined in figure B1.4.3 probe the high-frequency rotations of molecules and certain large-amplitude vibrational motions. As the rotational energy levels of a molecule depend largely on its moment of inertia, which is determined by the molecular structure, high-resolution spectroscopy of gas phase molecules provides the most precise information available on molecular geometries.

Rotational transition frequencies acquired in the THz region expand upon and complement those acquired in the microwave. Two types of molecules undergo rotational transitions that fall in the FIR: molecules with rotation about an axis having a small moment of inertia, and molecules in high- J states. FIR spectra of the first type of molecules are

important for determining their equilibrium geometry, as many light molecules (H_2O , NH_3 , HF, etc) only have transitions in the submillimetre and FIR regions. Due to the high rotational energy in the second type of species (high- J molecules), interactions between the vibrational and rotational motions, namely, centrifugal distortion and Coriolis perturbations, become important. Given high enough spectral resolution and accuracy ($\Delta\nu/\nu \leq 10^{-5}$), shifts in rotational frequencies and changes in selection rules resulting from these interactions

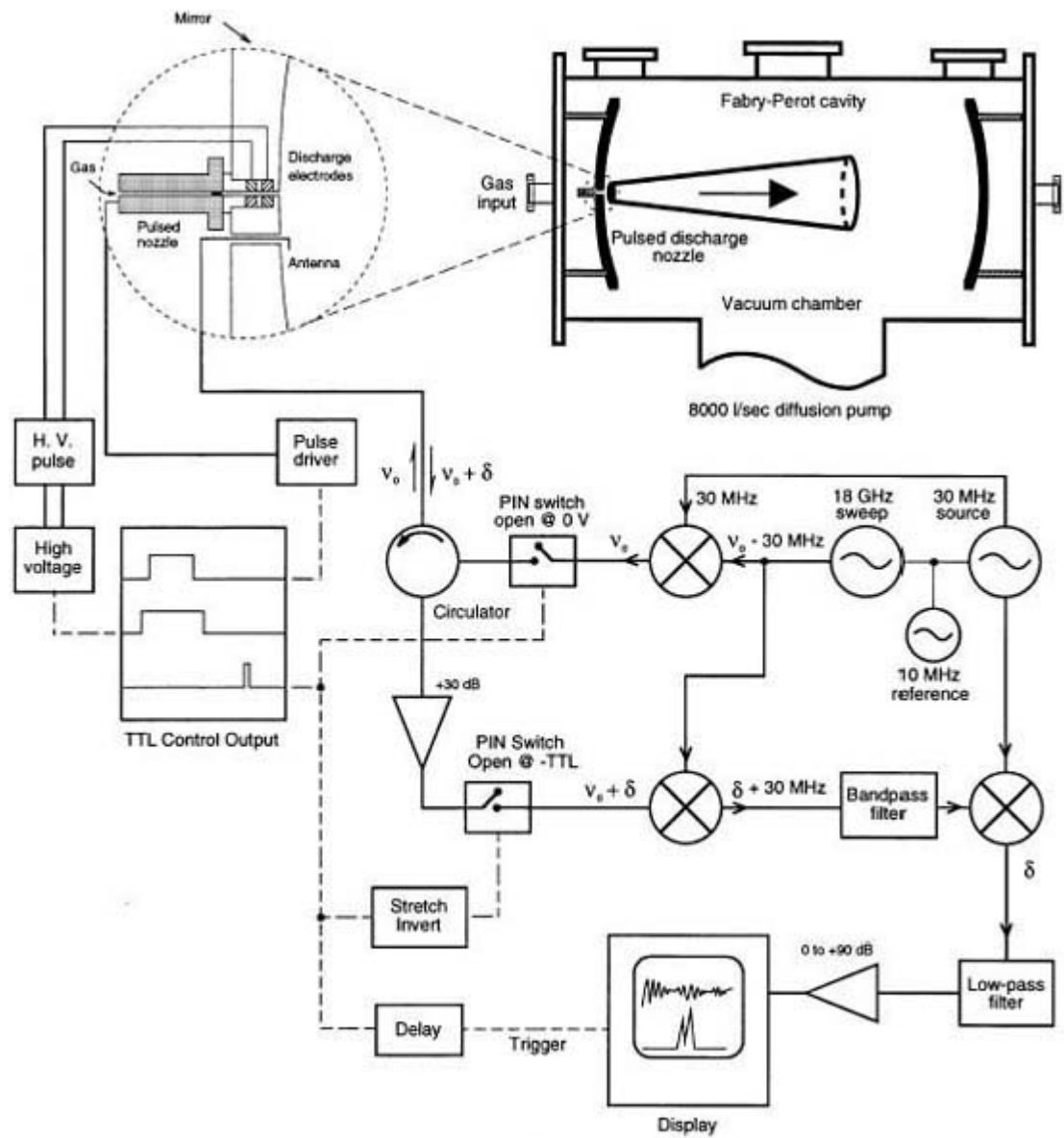
become significant. Thus, FIR spectroscopy of high- J transitions enables detailed characterization of molecular Hamiltonians far beyond the rigid rotor approximation, giving more accurate zero-point rotational constants and rough estimates of the shapes of potential energy surfaces. Finally, for very large molecules or weakly bound clusters, the softest vibrational degrees of freedom can be probed at THz frequencies, as is outlined in greater detail below.

B1.4.4.2 FOURIER TRANSFORM MICROWAVE SPECTROSCOPY

At microwave frequencies, direct absorption techniques become less sensitive than those in the THz region due to the steep dependence of the transition intensities with frequency. A variant of heterodyne spectroscopy, pioneered by Flygare and Balle [28], has proven to be much more sensitive. In this approach, molecules are seeded into or generated by a pulsed molecular beam which expands into a high- Q microwave cavity. The adiabatic expansion cools the rotational and translation degrees of freedom to temperatures near 1–10 K, and thus greatly simplifies the rotational spectra of large molecules. In addition, the low-energy collisional environment of the jet can lead to the growth of clusters held together by weak intermolecular forces.

A microwave pulse from a tunable oscillator is injected into the cavity by an antenna, and creates a coherent superposition of rotational states. In the absence of collisions, this superposition emits a free-induction decay signal, which is detected with an antenna-coupled microwave mixer similar to those used in molecular astrophysics. The data are collected in the time domain and Fourier transformed to yield the spectrum whose bandwidth is determined by the quality factor of the cavity. Hence, such instruments are called Fourier transform microwave (FTMW) spectrometers (or Flygare–Balle spectrometers, after the inventors). FTMW instruments are extraordinarily sensitive, and can be used to examine a wide range of stable molecules as well as highly transient or reactive species such as hydrogen-bonded or refractory clusters [29, 30].

An outline of an FTMW instrument used in the study of large, polar, carbonaceous species is shown in [figure B1.4.4](#). In this instrument, the FTMW cavity is mated to a pulsed electric discharge/supersonic expansion nozzle [31, 32]. Long-chain carbon species, up to that of HC₁₇N, as shown in [figure B1.4.4](#), can be studied with this technique, as can a wide variety of other molecules and clusters. With the jet directed along the axis of the cavity, the resolution is highly sub-Doppler, with the slight complication that a Doppler doublet is formed by the difference between the laboratory and molecular beam reference frames. Studies of the rotational spectra of hydrogen-bonded clusters have also been carried out by several groups using FTMW instruments, a topic we shall return to later. FT-THz instruments can, in principle, be built using the highly sensitive THz SIS or HEB mixers outlined above, and would have extraordinary sensitivities. In order to saturate the rotational or rovibrational transitions, however, high-power THz oscillators are needed, but are not yet available.



(a)

figure continued on next page.

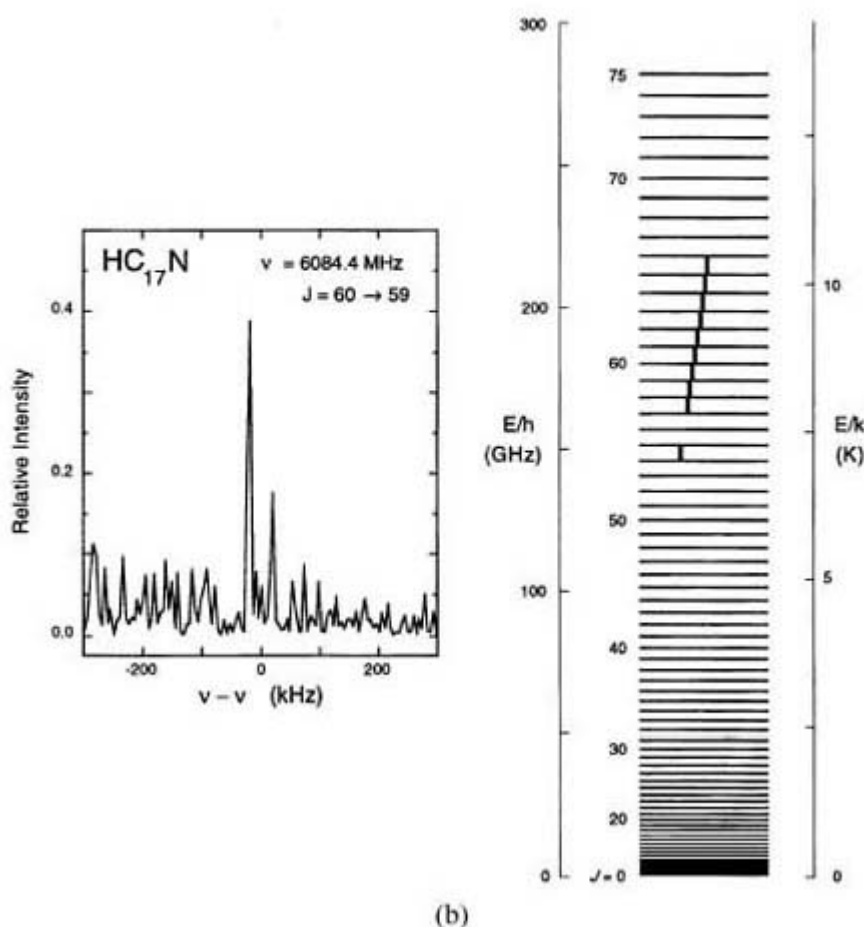


Figure B1.4.4. (a) An outline of the Harvard University electric discharge supersonic nozzle/Fourier transform microwave spectrometer. (b) The rotational states of HC_{17}N observed with this apparatus [31].

B1.4.4.3 CW THZ SOURCES AND MOLECULAR SPECTROSCOPY

At frequencies up to ~ 150 – 200 GHz, solid-state sources such as YIG-tuned oscillators or Gunn diode oscillators are now available with power outputs of up to 100 mW. The harmonic generation of such millimetre-wave sources is relatively efficient for doubling and tripling (≥ 10 – 15%), but for higher harmonics the power drops rapidly ($P_{\text{out}}(1 \text{ THz}) \leq 0.1$ – $10 \mu\text{W}$). Nevertheless, harmonic generation was used as early as the 1950s to record the submillimetre wave spectra of stable molecules [33]. Harmonics from optimized solid-state millimetre-wave sources are now used to drive astronomical heterodyne receivers up to 900–1100 GHz [34], and the prospects for operation up to 2–3 THz are promising.

Even higher output power (~ 1 – 10 mW) is available from rapidly tunable BWOs up to 1–15 THz. BWOs are capable laboratory sources where they operate, and offer wide tunability and excellent spectral purity, especially when phase locked to the harmonics of lower-frequency microwave or millimetre-wave oscillators [35]. The high output power of BWOs and the relatively strong intrinsic strengths of pure rotational transitions of polar molecules gives BWO spectrometers very high sensitivity, and also enables them to utilize nonlinear methods such as Lamb dip spectroscopy. An example for the CO molecule is presented in figure B1.4.5 [36]. The resulting resolution is truly exceptional and leads to among the most precise molecular

constants ever determined. Pioneered in the former Soviet Union, THz BWOs are finding increased applications in a number of laboratories.

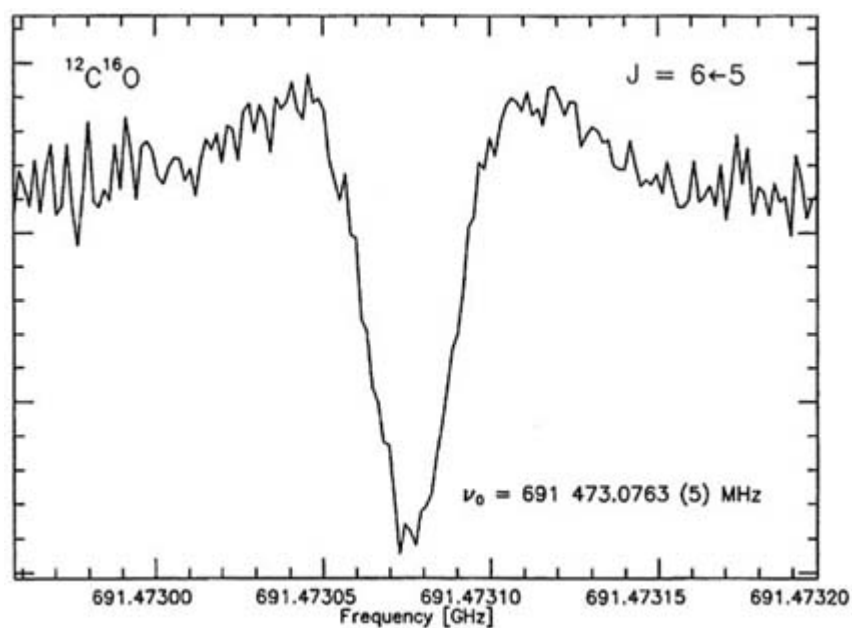


Figure B1.4.5. The Lamb dip spectrum of the CO 6–5 transition obtained with the Cologne THz BWO spectrometer. The dip is of order 30–40 kHz in width and the transition frequency is determined to 0.5 kHz [36].

BWOs are typically placed in highly overmoded waveguides, and the extension of this technology to higher frequencies will, by necessity, require a number of innovative solutions to the very small-scale structures that must be fabricated. Their size and weight also preclude them from space-based applications (e.g. FIRST). Thus, electronic oscillators are unlikely to cover all THz spectroscopy and/or remote sensing applications over the full range of 1–10 THz in the short term, and a number of alternative THz radiation sources are therefore under investigation. Among the most promising of these, over the long term, are engineered materials such as quantum wells that either possess high nonlinearity for THz mixing experiments [37] or that can be tailored to provide direct emission in the FIR [38]. Tunable laser sources are the ultimate goal of such development programmes, but while THz spontaneous emission has been observed, laser action is still some time in the future.

A number of mixing experiments have therefore been used to generate both pulses and CW THz radiation. Among these, diode-based mixers used as upconvertors (that is, heterodyne spectroscopy ‘in reverse’) have been the workhorse FIR instruments. Two such techniques have produced the bulk of the spectroscopic results:

(1) GaAs-based diode mixers that generate tunable sidebands on line-tunable FIR molecular gas lasers [39, 40], and (2) CO_2 laser-based THz difference frequency generation in ultrafast metal–insulator–metal (MIM) diodes [41, 42]. Both types of mixers have sufficient instantaneous bandwidth to place any desired millimetre-wave frequency on the carrier radiation, and respond well at THz frequencies. Having been used for many years in astronomical receiver applications, GaAs mixer technology is more mature than that for MIM diodes, and its conversion, noise and coupling mechanisms are better understood at present. In addition, their conversion efficiencies are good up to at least 4–5 THz, and several to several tens of microwatts are available from GaAs Schottky diode laser sideband generators. It is thus possible to construct THz spectrometers based on laser sideband generation that operate at or near the shot-noise limit, and this sensitivity has been used to

investigate a wide range of interesting reactive and/or transient species, as is described in [section \(B1.4.4.6\)](#).

While the conversion efficiency of MIM diodes is not as good as that of their GaAs counterparts, they are considerably faster, having also been used at IR and even visible wavelengths! Thus, MIM-based THz spectrometers work over wider frequency ranges than do GaAs FIR laser sideband generators, but with less output power. As described above, this translates directly into spectrometer sensitivity due to the high laboratory background in the THz region. Thus, where intense FIR gas laser lines are available, sideband generators are to be preferred, but at present only MIM-diode spectrometers can access the spectroscopically important region above 200 cm^{-1} [42]. Also, because it is easy to block CO_2 laser radiation with a variety of reststrahlen solid-state filters, there is no fixed-frequency FIR gas laser carrier to reject in MIM spectrometers, and this simplifies the overall experimental design. MIM spectrometers that perform third-order mixing of two CO_2 laser lines and a tunable microwave source have also been constructed. This approach leads to very wide tunability and eliminates the need to scan the CO_2 lasers, and only decreases the output power by a small amount. Thus, it is possible to phase lock the CO_2 and microwave sources, leading to a direct synthesis approach to THz radiation to very high frequencies indeed. This is not feasible for the FIR laser sideband generators, and so the MIM approach has provided extremely accurate THz frequency standards for calibration gases such as CO, HF and HCl, which can then be used as secondary standards in a number of other techniques [41].

While extremely useful in the laboratory, the size and power requirements of both FIR laser sideband generators and MIM-based CO_2 laser spectrometers are excessive for space-borne applications. Research on other THz generation approaches by mixing has therefore continued. One particularly interesting approach from a technology and miniaturization point of view is optical heterodyne conversion, in which optical radiation is converted to THz light by semiconducting materials pumped above their band gaps. The use of optical or near-IR lasers to drive the process results in wide tunability and spectral coverage of the THz spectrum. Such approaches, described next, also have the considerable advantage of leveraging the rapid technological innovations in diode-pumped lasers and fast optoelectronic devices required for emerging industries such as optical telecommunications and optical computing. They therefore ‘break the mould’ of traditional THz LO development by small groups focused on scientific problems, and rapid developments can be expected with little or no investment by the THz community. Finally, as described next, both ultrafast time-resolved and CW high-resolution spectroscopies can be carried out using these approaches.

B1.4.4.4 TIME DOMAIN THZ SPECTROSCOPY

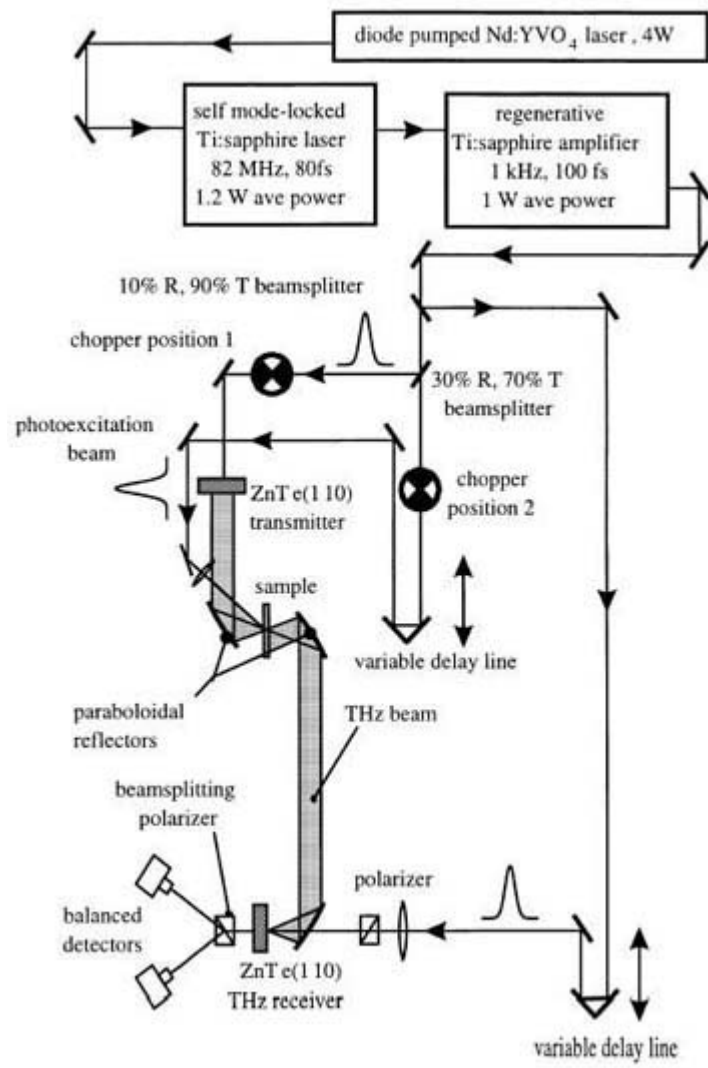
Free-electron lasers have long enabled the generation of extremely intense, sub-picosecond THz pulses that have been used to characterize a wide variety of materials and ultrafast processes [43]. Due to their massive size and great expense, however, only a few research groups have been able to operate them. Other approaches to the generation of sub-picosecond THz pulses have therefore been sought, and one of the earliest and most successful involved semiconducting materials. In a photoconductive semiconductor, carriers (for n-type material, electrons)

in the valence band absorb the incident radiative power (if $h\nu_0 \geq E_{\text{bandgap}}$) and are injected into the conduction band. Once they are in the conduction band, the electrons become mobile, and, if there is an applied bias, they begin to drift toward the photoconductor electrodes. Some of the electrons will reach the electrodes while some will encounter sites of ionic impurities. The latter electrons are trapped by the impurity sites and removed from the conduction band. As early as two decades ago, pulsed optical lasers were used by Auston and co-workers to generate and detect electrical pulses in DC biased voltage transmission lines [44]. For the earliest used materials such as GaAs and Si, the pulse widths were of order nanoseconds due to their long recombination times. The discovery of materials such as radiation-damaged silicon-on-sapphire and of low-temperature-grown (LTG) GaAs changed this situation dramatically.

Especially with LTG GaAs, materials became available that were nearly ideal for time-resolved THz spectroscopy. Due to the low growth temperature and the slight As excess incorporated, clusters are formed which act as recombination sites for the excited carriers, leading to lifetimes of ≤ 250 fs [45]. With such recombination lifetimes, THz radiators such as dipole antennae or log-periodic spirals placed onto optoelectronic substrates and pumped with ultrafast lasers can be used to generate sub-picosecond pulses with optical bandwidths of 2–4 THz. Moreover, *coherent* sub-picosecond detection is possible, which enables both the real and imaginary refractive indices of materials to be measured. The overall sensitivity is $>10^4$, and a variety of solid-state and gas phase THz spectra have been acquired with such systems [46, 47], an excellent overview of which may be found in [48].

Recently, it has been shown that both the detection and generation of ultrafast THz pulses can be carried out using the electro-optic effect in thin films of materials such as ZnTe, GaAs and InP that are pumped in the near-IR [49]. The generation efficiency is similar to that of the photoconducting antenna approach, but the electro-optic scheme offers two extremely significant advantages. First, the detection bandwidth can be extremely large, up to 30–40 THz under optimum conditions [49]. Second, it is possible to directly *image* the THz field with such spectrometers. Such approaches therefore make possible the THz imaging of optically opaque materials with a compact, all solid-state, room-temperature system [50]!

The great sensitivity and bandwidth of electro-optic approaches to optical–THz conversion also enable a variety of new experiments in condensed matter physics and chemistry to be conducted, as is outlined in [figure B1.4.6](#). The left-hand side of this figure outlines the experimental approach used to generate ultrafast optical and THz pulses with variable time delays between them [51]. A mode-locked Ti:sapphire laser is amplified to provide approximately 1 W of 100 fs near-IR pulses at a repetition rate of 1 kHz. The ~ 850 nm light is divided into three beams, two of which are used to generate and detect the THz pulses, and the third of which is used to optically excite the sample with a suitable temporal delay. The right-hand panel presents the measured relaxation of an optically excited TBNC molecule in liquid toluene. In such molecules, the charge distribution changes markedly in the ground and electronically excited states. In TBNC, for example, the excess negative charge on the central porphyrin ring becomes more delocalized in the excited state. The altered charge distribution must be accommodated by changes in the surrounding solvent. This so-called solvent reorganization could only be indirectly probed by Stokes shifts in previous optical–optical pump–probe experiments, but the optical–THz approach enables the solvent response to be *directly* investigated. In this case, at least three distinct temporal response patterns of the toluene solvent can be seen that span several temporal decades [51]. For solid-state spectroscopy, ultrafast THz studies have enabled the investigation of coherent oscillation dynamics in the collective (phonon) modes of a wide variety of materials for the first time [49].



(a)

figure continued on next page.

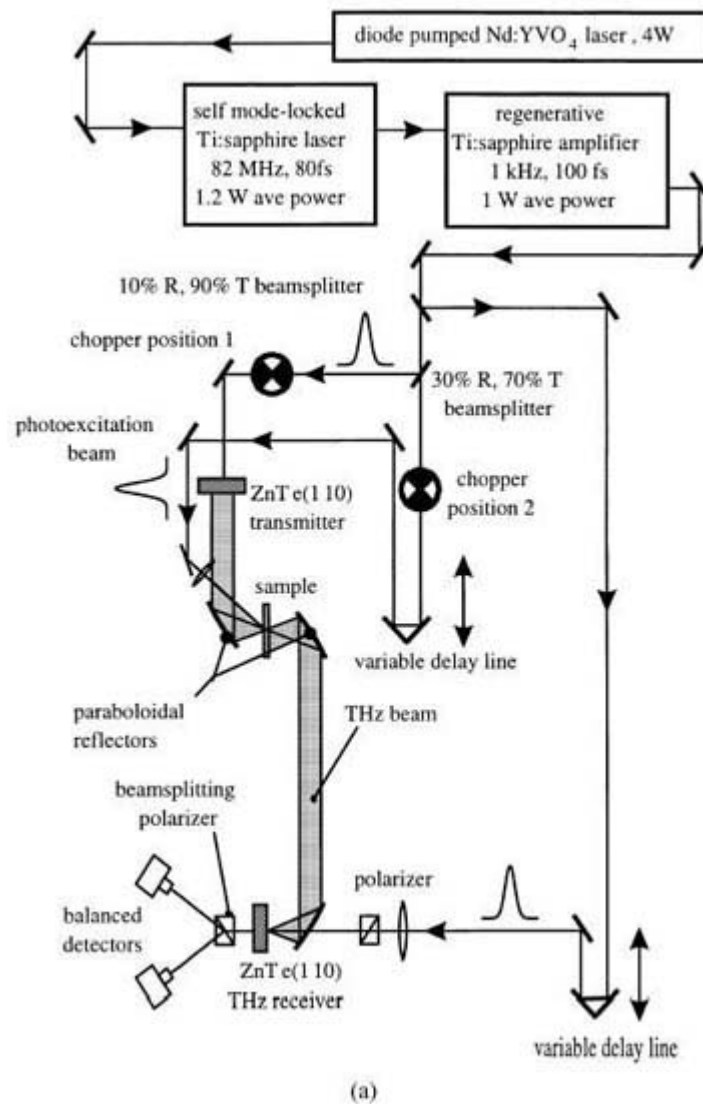


Figure B1.4.6. Left: an experimental optical THz pump–probe set-up using sub-picosecond THz pulse generation and detection by the electro-optic effect. Right: the application of such pulses to the relaxation of optically excited TBNC in toluene. The THz electric field used for these experiments is shown in the upper-right inset. Three exponential decay terms, of order 2, 50 and 700 ps, are required to fit the observed temporal relaxation of the solvent [51].

B1.4.4.5 THZ OPTICAL-HETERODYNE CONVERSION IN PHOTOCONDUCTORS

For CW applications of optical-heterodyne conversion, two laser fields are applied to the optoelectronic material. The non-linear nature of the electro-optic effect strongly suppresses continuous emission relative to ultrashort pulse excitation, and so most of the CW research carried out to date has used photoconductive antennae. The CW mixing process is characterized by the average drift velocity \bar{v} and carrier lifetime τ_0 of the mixing material, typically

LTG GaAs. If $\tau_0 \bar{v} \geq \delta$, the electrode spacing, then a significant amount of current will be generated by the photo-excitation. That is

$$i(t) = \frac{N_c(t)e\bar{v}}{\delta} \quad (\text{B1.4.3})$$

where N_c is the number of carriers. The rate equation for the photo-excitation-recombination process can be written as

$$\frac{dN_c(t)}{dt} = aP_0 - \frac{N_c(t)}{\tau_0} \quad (\text{B1.4.4})$$

where a is a proportionality constant and P_0 is the average power of the incident field over a few optical periods. The time-average is of interest here because τ_0 in a semiconductor is always longer than an optical cycle and, therefore, the output current will not respond directly to the optical oscillations. Since the THz waves are generated with optical light, $\omega \equiv (\omega_1 - \omega_2) \ll \omega_1 \approx \omega_2$. Thus, integrating the right-hand side of the expression above over a few cycles of $\omega_1 \approx \omega_2$ yields

$$P_0 = E_1^2 + E_2^2 + 2E_1E_2 \cos(\omega t + \phi_1 - \phi_2). \quad (\text{B1.4.5})$$

Substitution yields

$$\frac{dN_c}{dt} = a[E_1^2 + E_2^2 + 2E_1E_2 \cos(\omega t + \phi_1 - \phi_2)] - \frac{N_c}{\tau_0}. \quad (\text{B1.4.6})$$

By solving the differential equation above and using the single-pump case ($P_2 = 0$) to determine a , it can be shown that the photo-current is

$$i(t) = \frac{e\eta}{h\nu_0} \frac{\tau_0\bar{v}}{\delta} \left[P_1 + P_2 + 2\sqrt{\frac{P_1P_2}{1 + \omega^2\tau_0^2}} \cos(\omega t - \phi) \right] \quad (\text{B1.4.7})$$

where $\nu_0 \equiv \nu_1 \approx \nu_2$, P_1 and P_2 are the incident optical powers, $\phi = \tan^{-1}(\omega\tau_0)$ and η , the quantum efficiency, is the number of carriers excited per incident photon. Recognizing that $\bar{v} = \mu E$ (where μ is the carrier mobility and E is the electric field),

$$i(t) = \frac{e\eta}{h\nu_0} \frac{\tau_0\mu E}{\delta} \left[P_1 + P_2 + 2\sqrt{\frac{P_1P_2}{1 + \omega^2\tau_0^2}} \cos(\omega t - \phi) \right]. \quad (\text{B1.4.8})$$

Separating the DC and oscillatory parts of the above equation gives

$$i_{dc} = \frac{e\eta}{h\nu_0} \frac{\tau_0\mu E}{\delta} (P_1 + P_2) \quad (\text{B1.4.9})$$

$$(\text{B1.4.10})$$

$$i_{\text{THz}}(t) = \frac{e\eta}{h\nu_0} \frac{\tau_0 \mu E}{\delta} 2 \sqrt{\frac{P_1 P_2}{1 + \omega^2 \tau_0^2}} \cos(\omega t - \phi).$$

Thus, the beating of the two incident optical fields generates a modulation in the photo-excitation of the carriers, which in turn results in an oscillating electrical signal. Initial microwave experiments using an interdigitated electrode geometry by Brown *et al* [52] showed a flat frequency response up to 25 GHz with a conversion efficiency of 0.14%, in agreement with the signal level predicted by the theoretical analysis outlined above. At THz frequencies, the power decays rapidly from such structures due to the parasitic capacitance of the electrode structure and the finite carrier lifetime. Free-space radiation is generated by coupling the electrodes to a planar THz antenna. At 1 THz, the observed conversion efficiency is roughly 3×10^{-6} , and the damage threshold is of order $1 \text{ mW } \mu\text{m}^{-2}$. To alleviate these limitations, travelling-wave structures have now been developed that eliminate the capacitive roll-off and allow large-device active areas to be pumped. Powers in excess of $1 \text{ } \mu\text{W}$ can now be achieved above 2 THz for input drive levels of 300–400 mW [53].

This power level is sufficient for laboratory spectroscopy or for use as a THz local oscillator, and such travelling-wave structures can be used over at least a decade of frequency (0.3–3 THz, for example) without moving parts. Further, compact, all solid-state spectrometers, such as that outlined in [figure B1.4.7](#) can now be constructed using CW diode laser and optical tapered amplifier technology [54]. The major challenge in working with diode lasers is, in fact, their instantaneous line widths ($\geq 15 \text{ MHz}$) and long-term frequency stability ($\sim 100\text{--}200 \text{ MHz}$), both of which need considerable improvement to be useful as THz LOs. The main source of both instabilities is the notorious susceptibility of diode lasers to optical feedback. However, this susceptibility can be used to one's advantage by sending a small fraction of the laser output into a high-finesse ($F \geq 60$) optical cavity such that the diode laser 'sees' optical feedback only at cavity resonances. By locking the diode lasers to different longitudinal modes of an ultrastable reference cavity, it is possible to construct direct synthesis spectrometers that can be absolutely calibrated to $\Delta\nu/\nu < 10^{-8}$ or better. To demonstrate the continuous tunability and frequency stability of such an instrument, the lower panel of [figure B1.4.7](#) presents a submillimetre spectrum of acetonitrile in which the transition frequencies are measured to better than 50 kHz. Future improvements to such systems should allow similar measurements on both stable and transient species up to at least 5–6 THz.

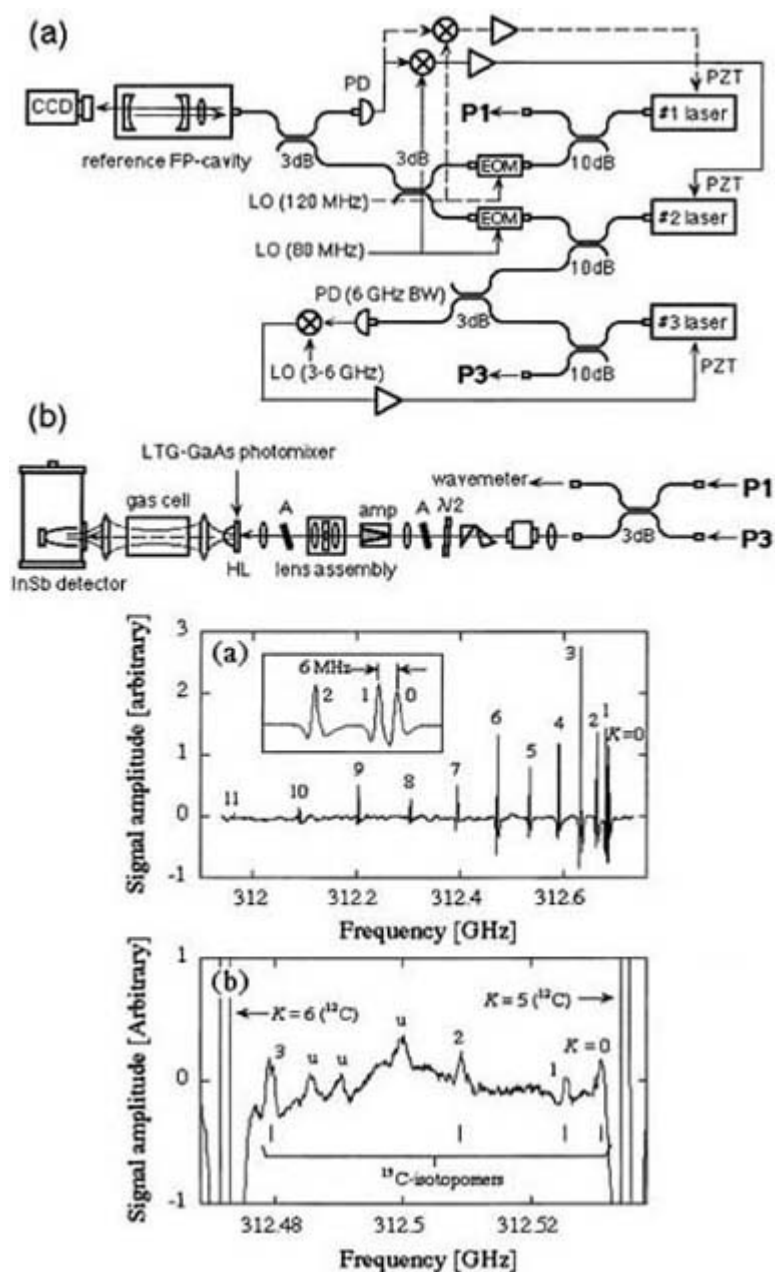


Figure B1.4.7. Top: THz generation by optical-heterodyne conversion in low-temperature GaAs. (a) The three DBR laser system that synthesizes a precise difference frequency for the THz photomixer spectrometer, (b) the MOPA system and the set-up for spectroscopy. Bottom: second-derivative absorption spectrum of the $\text{CH}_3\text{CN } J_K = 16_K \rightarrow 17_K$ rotational transitions near 312 GHz. (a) The spectrum for ordinary $^{12}\text{CH}_3$, ^{12}CN . The inset is an expanded view of the $K = 0-2$ lines. (b) The $K = 0-3$ lines of CH_2 , ^{13}CN [54].

B1.4.4.6 THZ SPECTROSCOPY OF HYDROGEN-BONDED AND REFRACTORY CLUSTERS

Among the most interesting transient species that can be studied at THz frequencies are those involving collections of molecules held together by van der Waals or hydrogen bonding forces. In no small measure this is true because hydrogen bonds are ubiquitous in nature. From the icy mantles covering interstellar dust to the nuclei of living cells, hydrogen bonds play crucial roles in the regulation and evolution of both inorganic and living systems. Accurate, fully anisotropic, descriptions of the intermolecular forces involved in these and other weak interactions are therefore assuming an increasingly pivotal role in modern molecular science,

particularly in molecular biology [54]. Within chemical physics, the anisotropy of intermolecular forces plays a central role in understanding the dynamics associated with photoinitiated reactions in clusters [55], to name but one example.

Over the past century, much of the data underlying current descriptions of van der Waals and hydrogen bonds were obtained from measurements of second virial coefficients, pressure broadening, and other classical properties. Experimental advances during the past three decades have led to many techniques capable of interrogating intermolecular forces, most notably scattering experiments and the spectroscopic study of isolated clusters. Despite this long-standing interest, however, truly quantitative, microscopic models of these forces have only become available in recent years as advances in *ab initio* theory, high-resolution spectroscopy and eigenvalue generation on complex potential energy surfaces have converged, to enable the fitting of fully anisotropic force fields to experimental data for systems with two, three and four degrees of freedom [57].

The weak interactions in clusters are mathematically modelled by means of a multi-dimensional intermolecular potential energy surface, or IPS. Microwave spectroscopy, carried out primarily with the elegant, extraordinarily sensitive and sub-Doppler resolution FTMW technique outlined previously, probes the very lowest region of the ground-state surface; visible and IR spectroscopy probes states above the dissociation energy of the adduct (the latter of which are described elsewhere in this encyclopaedia). None are, in themselves, direct probes of the total ground-state IPS. Indeed, while microwave, IR and UV/Vis instruments have produced structural parameters and dynamical lifetimes for literally dozens of binary (and larger) weakly bound complexes (WBCs) over the past two decades [58, 59, 60, and 61], recent calculations which explicitly allow coupling between all the degrees of freedom present in the cluster reveal that structural parameters alone are not sufficient to accurately characterize the IPS [62].

Weak interactions are characterized by binding energies of at most a few kcal/mole and by IPSs with a very rich and complex topology connected by barriers of at most a few hundred cm^{-12} . Rotational, tunnelling and intermolecular vibrational states can therefore become quite strongly mixed, hence the general term of vibration–rotation–tunnelling (VRT) spectroscopy for the study of eigenvalues supported by an IPS [63, 64]. The VRT states in nearly all systems lie close to or above the tunnelling barriers, and therefore sample *large* regions of the potential surface. In addition, as they become spectroscopically observable, the number, spacings and intensities of the tunnelling splittings are intimately related to the nature of the tunnelling *paths* over the potential surface.

Thus, by measuring the intermolecular vibrations of a WBC, ultimately with resolution of the rotational, tunnelling and hyperfine structure, the most sensitive measure of the IPS is accessed directly. The difficulty of measuring these VRT spectra is the fact that they lie nearly exclusively at THz frequencies. As expected, the ‘stiffer’ the interaction, the higher in frequency these modes are found. In general, the total 0.3–30 THz interval must be accessed, although for the softest or heaviest species the modes rarely lie above 10–15 THz.

For WBCs composed of stable molecules, planar jet expansions produce sufficiently high concentrations that direct absorption THz studies can be pursued for clusters containing $\lesssim 6$ small molecules using FIR laser sidebands. Research on water clusters has been particularly productive, and has been used to investigate the structures and large-amplitude dynamics of the clusters outlined in the top panel of figure B1.4.8 [65, 66, 67 and 68]. In addition, as the bottom panel illustrates, not only are the VRT modes directly sampled by such work, but the available spectral resolution of $\lesssim 1$ MHz enables full rotational resolution along with a detailed investigation of the VRT and hyperfine splittings. The high resolution is also essential in untangling the often overlapping bands from the many different clusters formed in the supersonic expansion.

For clusters beyond the dimer, each of the monomers can both accept and donate hydrogen bonds, which leads to a rich suite of large-amplitude motions. Their spectroscopic manifestations are illustrated for the water trimer in [figure B1.4.9](#). The most facile motion in this system is the ‘flipping’ of one of the non-bonded hydrogen atoms through the plane of the oxygen atoms. This motion is sufficiently fast that it produces symmetric top rovibrational spectra even though at any one instant the molecule is always asymmetric. Six of these flipping motions lead to the same structure, a process known as ‘pseudorotation’, and leads to the manifold of states produced in the bottom panel of [figure B1.4.9](#). The exchange of bound *versus* free hydrogen atoms in a monomer leads to the hyperfine splittings of the individual transitions, as is illustrated in the top panel of the same figure. From a comparison of such spectra with detailed calculations, a variety of IPS properties can be extracted to experimental precision.

Molecules like those presented in [figure B1.4.4](#) form another interesting suite of targets from a THz perspective. Such chains can be treated as rigid rods, and as they get longer their lowest bending frequencies move rapidly into the FIR. For example, the lowest frequencies of a variety of chains are as follows:

Cyanopolyynes, $\text{HC}_3\text{N} \rightarrow \text{HC}_{25}\text{N}$	222 \rightarrow 8 cm^{-1}
Polyacetylenes, $\text{HC}_4\text{N} \rightarrow \text{HC}_{20}\text{H}$	220 \rightarrow 19 cm^{-1}
Carbon clusters, $\text{C}_3 \rightarrow \text{C}_{20}$	63 \rightarrow 17 cm^{-1}
C_nN radicals, $\text{C}_3\text{N} \rightarrow \text{C}_{19}\text{N}$	144 \rightarrow 4 cm^{-1} .

A real advantage of working in the FIR is that both polar and non-polar chains may be searched for. Indeed, the lowest bending frequency of C_3 has been studied in the laboratory [69], and tentatively detected toward the galactic centre source Sgr B2 [70]. Other large molecules such as polycyclic aromatic hydrocarbons (anthracene, pyrene, perylene, etc) or ‘biomolecules’ such as glycine or uracil also possess low-frequency FIR vibrations, and can be produced in sizable quantities in supersonic expansions through heated planar nozzles [71]. The study of such species is important cosmochemically, but is quite difficult at microwave frequencies where the rotational spectra are weak, and nearly impossible at IR or optical wavelengths due to the extinction present in dense molecular clouds and young stellar objects.

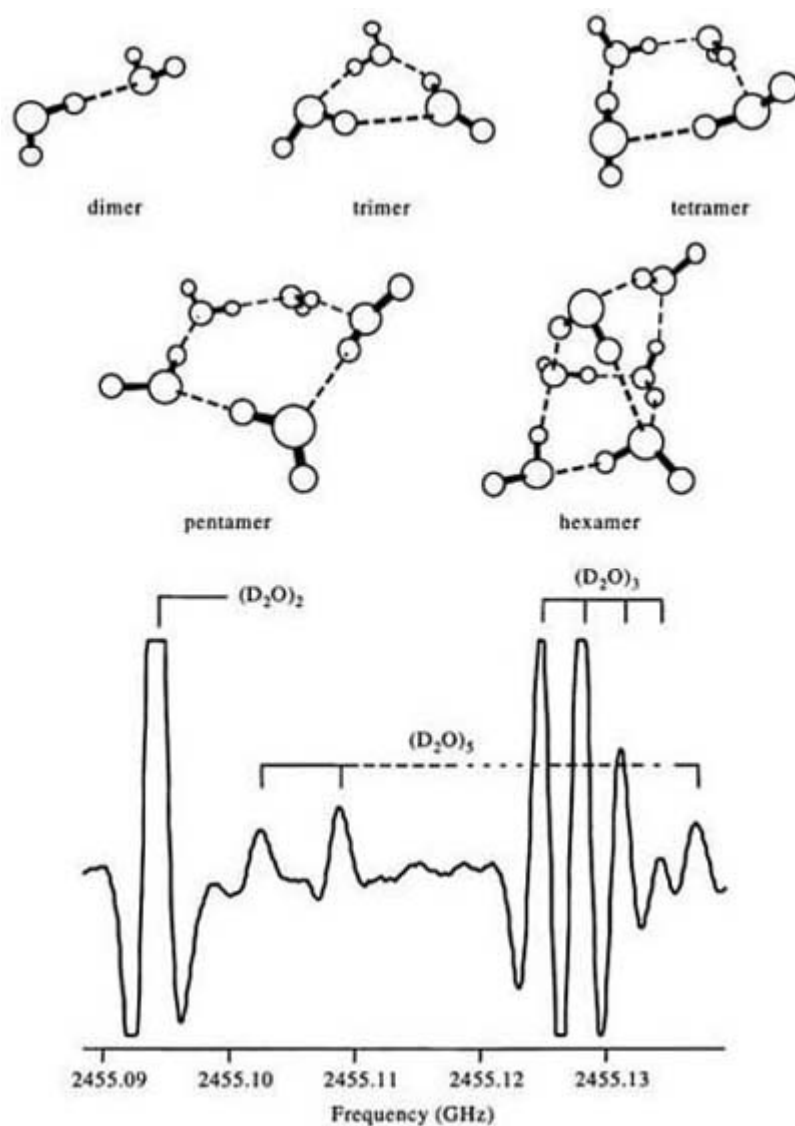


Figure B1.4.8. Top: the lowest-energy structures of water clusters, $(\text{H}_2\text{O})_n$, from $n = 2 - 6$. Bottom: a sample ~ 2.5 THz spectrum of such clusters formed in a pulsed planar supersonic expansion [65].

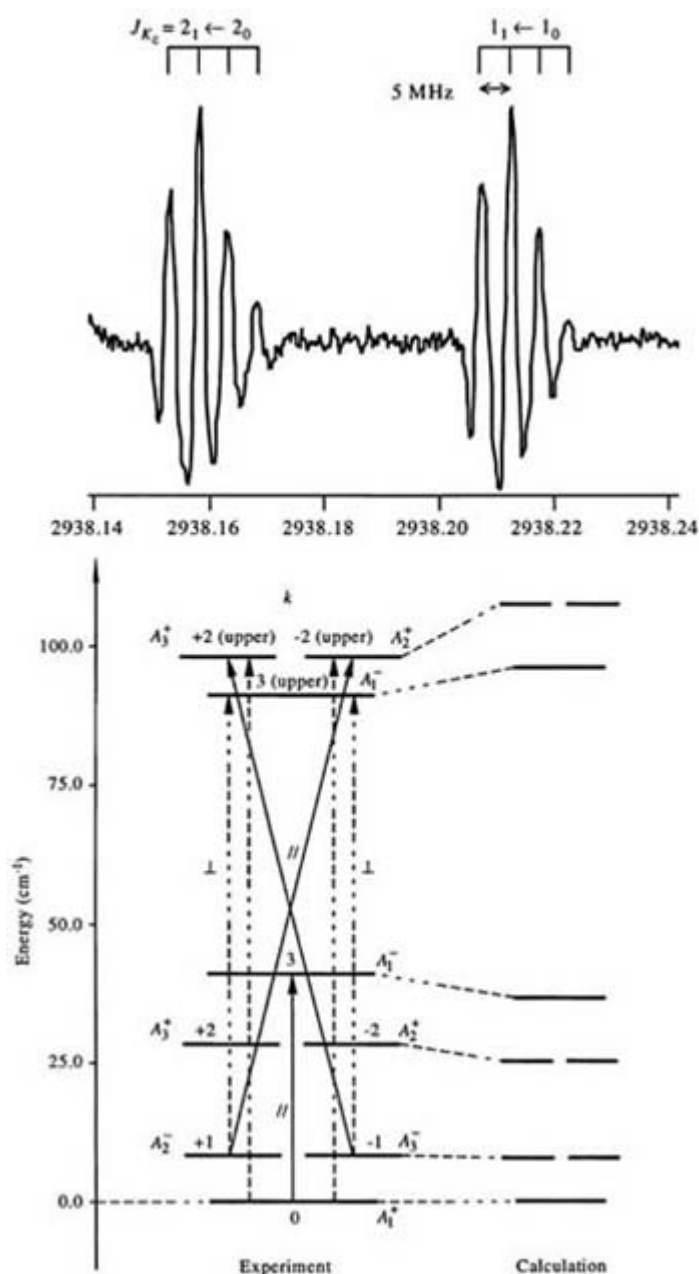


Figure B1.4.9. Top: rotation–tunnelling hyperfine structure in one of the ‘flipping’ modes of $(\text{D}_2\text{O})_3$ near 3 THz. The small splittings seen in the Q -branch transitions are induced by the bound–free hydrogen atom tunnelling by the water monomers. Bottom: the low-frequency torsional mode structure of the water dimer spectrum, including a detailed comparison of theoretical calculations of the dynamics with those observed experimentally [65]. The symbols next to the arrows depict the parallel ($\Delta k = 0$) versus perpendicular ($\Delta K = \pm 1$) nature of the selection rules in the pseudorotation manifold.

B1.4.5 OUTLOOK

Technology developments are revolutionizing the spectroscopic capabilities at THz frequencies. While no one technique is ideal for all applications, both CW and pulsed spectrometers operating at or near the fundamental limits imposed by quantum mechanics are now within reach. Compact, all-solid-state implementations will soon allow such spectrometers to move out of the laboratory and into a wealth of field and remote-sensing applications. From the study of the rotational motions of light molecules to the large-amplitude vibrations of

clusters and the collective motions of condensed phases, microwave and THz spectroscopy opens up new windows to a wealth of scientifically and technologically important fields. Over the coming decade, truly user-friendly and extraordinarily capable instruments should become cost affordable and widely available, enabling this critical region of the electromagnetic spectrum to be fully exploited for the first time.

REFERENCES

- [1] Johns J W C 1985 High resolution far-infrared ($20\text{--}350\text{ cm}^{-1}$) spectra of several isotopic species of H_2O *J. Opt. Soc. Am. B* **2** 1340–54
- [2] McLean I S 1995 Infrared array detectors—performance and prospects *IAU Symp.* **167** 69–78
- [3] Kimmitt M F 1970 *Far-Infrared Techniques* (London: Pion)
- [4] Bock J J, Chen D, Mauskopf P D and Lange A E 1995 A novel bolometer for infrared and millimeter-wave astrophysics *Space Sci. Rev.* **74** 229–35
- [5] Lis D C, Serabyn E, Keene J, Dowell C D, Benford D J, Phillips T G, Hunter T R and Wang N 1998 350 micro continuum imaging of the Orion A molecular cloud with the submillimeter high angular resolution camera (SHARC) *Astrophys. J.* **509** 299–308
- [6] Ivison R J, Smail I, Le Borgne J F, Blain A W, Kneib J P, Bezecourt J, Kerr T H and Davies J K 1997 A hyperluminous galaxy at $z = 2.8$ found in a deep submillimetre survey *Mon. Not. R. Astron. Soc.* **298** 583–93
- [7] Putley E H 1964 The ultimate sensitivity of sub-mm detectors *Infra. Phys.* **4** 1–35
- [8] Beeman J W and Haller E E 1994 Ga:Ge photoconductor arrays—design considerations and quantitative analysis of prototype single pixels *Infra. Phys. Technology* **35** 827–36
- [9] Waters J W 1993 Microwave limb sounding *Atmospheric Remote Sensing by Microwave Radiometry* ed M A Janssen (New York: Wiley) pp 383–496
- [10] Chance K, Traub W A, Johnson D G, Jucks K W, Ciarpallini P, Stachnik R A, Salawitch R J and Michelsen H A 1996 Simultaneous measurements of stratospheric HO_x , NO_x , and Cl_x : comparison with a photochemical model *J. Geophys. Res.* **101** 9031–43
- [11] Winnewisser G 1994 Submillimeter and infrared astronomy: recent scientific and technical developments *Infra. Phys. Tech.* **35** 551–67
- [12] Carlström J and Zmuidzinas J 1996 Millimeter and submillimeter techniques *Review of Radio Science 1993–1996* ed W Ross Stone (Oxford: Oxford University Press) pp 839–92

- [13] Bin M, Gaidis M C, Zmuidzinas J and Phillips T G 1997 Quasi-optical SIS mixers with normal tuning structures *IEEE Trans. Appl. Supercond.* **7** 3584–8
- [14] Karasik B S, Gaidis M C, McGrath W R, Bumble B and LeDuc H G 1997 A low noise 2.5 THz superconducting NB hot-electron mixer *IEEE Trans. Appl. Supercond.* **7** 3580–3
- [15] Barrett A H 1983 The beginnings of molecular radio astronomy *Serendipitous Discoveries in Radio Astronomy* ed K Kellerman and B Sheets (Green Bank, WV: NRAO)
- [16] Ohishi M 1997 Observations of hot cores *IAU Symp.* **178** 61–74
- [17] Phillips T G and Keene J 1992 Submillimeter astronomy *Proc. IEEE* **80** 1662–78

- [18] Hartquist T W and Williams D A (eds) 1998 *The Molecular Astrophysics of Stars and Galaxies* (Oxford: Oxford University Press)
- [19] Herbst E 1995 Chemistry in the interstellar medium *Ann. Rev. Phys. Chem.* **46** 27–53
- [20] van Dishoeck E F and Blake G A 1998 Chemical evolution of star forming regions *Ann. Rev. Astron. Ap.* **36** 317–68
- [21] van Dishoeck E F 1997 The importance of high resolution far-infrared spectroscopy of the interstellar medium *Proc. ESA Symp.* **SP-401** 81–90
- [22] Dutrey A, Guilloteau S and Guelin M 1997 Chemistry of protosolar-like nebulae: the molecular content of the DM Tau and GG Tau disks *Astron. Astrophys.* **317** L55–8
- [23] Sargent A I 1997 Protostellar and protoplanetary disks *IAU Symp.* **170** 151–8
- [24] Biver N *et al* 1997 Evolution of the outgassing of comet Hale–Bopp (C/1995 O1) from radio observations *Science* **275** 1915–18
- [25] For an overview of ALMA, see <http://www.nrao.edu>
- [26] Erickson E F 1995 SOFIA—the next generation airborne observatory *Space Sci. Rev.* **74** 91–100
- [27] Batchelor M, Adler D and Trogus W 1996 New plans for FIRST *Missions to the Moon & Exploring the Cold Universe* **18** 185–8
- [28] Balle T J and Flygare W H 1981 A Fourier transform microwave spectrometer *Rev. Sci. Instrum.* **52** 33–45
- [29] Munrow M R, Pringle W C and Novick S E 1999 Determination of the structure of the argon cyclobutanone van der Waals complex *J. Chem. Phys.* **103** 2256–61
- [30] Lovas F J, Suenram R D, Ogata T and Yamamoto S 1992 Microwave spectra and electric dipole moments for low-*J* levels of interstellar radicals—SO, CCS, CCCS, c-H₃, CH₂CC and c-C₃H₂ *Astrophys. J.* **399** 325–9
- [31] Thaddeus P, McCarthy M C, Travers M, Gottlieb C and Chen W 1998 New carbon chains in the laboratory and in interstellar space *Faraday Soc. Discuss.* **109** 121–36
- [32] McCarthy M C, Travers M J, Kovacs A, Chen W, Novick S E, Gottlieb C A and Thaddeus P 1997 *Science* **275** 518–20
- [33] Gordy W 1960 *Proc. Symp. MM Waves* (Brooklyn: Polytechnic Press)
- [34] Rothermel H, Phillips T G and Keene J 1989 *Int. J. Infra. MM Waves* **10** 83–100
- [35] Lewen F, Gendriesch R, Pak I, Paveliev D G, Hepp M, Schider R and Winnewisser G 1998 Phase locked backward wave oscillator pulsed beam spectrometer in the submillimeter wave range *Rev. Sci. Instrum.* **69** 32–9

- [36] Winnewisser G, Belov S P, Klaus T and Schieder R 1997 Sub-Doppler measurements on the rotational transitions of carbon monoxide *J. Mol. Spectrosc.* **184** 468–72
- [37] Maranowski K D, Gossard A C, Unterrainer K and Gornik E 1996 Far-infrared emission from parabolically graded quantum wells *Appl. Phys. Lett.* **69** 3522–4
- [38] Xu B, Hu Q and Melloch M R 1997 Electrically pumped tunable terahertz emitter based on inter-subband transitions *Appl. Phys. Lett.* **71** 440–2
- [39] Bicanic D D, Zuiberg B F J and Dymanus 1978 *Appl. Phys. Lett.* **32** 367–9
- [40] Blake G A, Laughlin K B, Cohen R C, Busarow K L, Gwo D H, Schmuttenmaer C A, Steyert D W and Saykally R J

1991 Tunable far-infrared spectrometers *Rev. Sci. Instrum.* **62** 1693–700

- [41] Varberg T D and Evenson K M 1992 Accurate far-infrared rotational frequencies of carbon monoxide *Astrophys. J.* **385** 763–5
- [42] Odashima H, Zink L R and Evenson K M 1999 Tunable far-infrared spectroscopy extended to 9.1 THz *Opt. Lett.* **24** 406–7
- [43] Kono J, Su M Y, Inoshita T, Noda T, Sherwin M S, Allen S J and Sakaki H 1997 Resonant THz optical sideband generation from confined magnetoexcitons *Phys. Rev. Lett.* **79** 1758–61
- [44] For a historical review see Lee C H 1984 *Picosecond Optoelectronic Devices* (New York: Academic)
- [45] Gupta S, Frankel M Y, Valdmanis J A, Whitaker J F, Mourou G A, Smith F W and Calawa A R 1991 Subpicosecond carrier lifetime in GaAs grown by MBE at low temperatures *Appl. Phys. Lett.* **59** 3276–8
- [46] Jeon T I and Grischkowsky D 1998 Characterization of optically dense, doped semiconductors by reflection THz time domain spectroscopy *Appl. Phys. Lett.* **72** 3032–4
- [47] Chevillon R A and Grischkowsky D 1999 Far-infrared foreign and self-broadened rotational linewidths of high temperature water vapor *J. Opt. Soc. Am. B* **16** 317–22
- [48] Nuss M C and Orenstein J 1998 Terahertz time domain spectroscopy *Millimeter Submillimeter Wave Spectrosc. Solids* **74** 7–50
- [49] Han P Y, Cho G C and Zhang X C 1999 Mid-IR THz beam sensors: exploration and application for phonon spectroscopy, ultrafast phenomena in semiconductors III *Proc. SPIE* **3624** 224–33
- [50] Koch M, Hunsche S, Schuacher P, Nuss M C, Feldmann J and Fromm J 1998 THz-imaging: a new method for density mapping of wood *Wood Sci. Technol.* **32** 421–7
- [51] Venables D S and Schmuttenmaer C A 1998 Far-infrared spectra and associated dynamics in acetonitrile-water mixtures measured with femtosecond THz pulse spectroscopy *J. Chem. Phys.* **108** 4935–44
- [52] Brown E R, McIntosh K A, Smith F W, Manfra M J and Dennis C L 1993 Measurements of optical-heterodyne conversion in low-temperature grown GaAs *Appl. Phys. Lett.* **62** 1206–8
- [53] Mastuura S, Blake G A, Wyss R, Pearson J, Kadow C, Jackson A and Gossard A C 1999 A travelling-wave photomixer based on angle-tuned phase matching *Appl. Phys. Lett.* **74** 2872–4
- [54] Matsuura S, Chen P, Blake G A, Pearson J and Pickett H M 1999 A tunable, cavity-locked diode laser system for terahertz photomixing *IEEE Micro. Theory Technol.* **48** 380–7
- [55] Stone A J 1996 *The Theory of Intermolecular Forces* (Oxford: Oxford University Press)
- [56] Ionov S I, Ionov P I and Wittig C 1994 Time resolved studies of photoinitiated reactions in binary and larger (N₂O)_m(HI)_n complexes *Discuss. Faraday Soc.* **97** 391–400

- [57] Cohen R C and Saykally R J 1991 Multidimensional intermolecular potential surfaces from VRT spectra of van der Waals complexes *Ann. Rev. Phys. Chem.* **42** 369–92
- [58] Novick S, Leopold K and Klemperer W 1990 *Atomic and Molecular Clusters* ed E Bernstein (New York: Elsevier) p 359–91 (the clusters listing presented here is now maintained electronically by S Novick)
- [59] Miller R E 1988 *Science* **240** 447–52
- [60] Nesbitt D J 1990 *Dynamics of Polyatomic van der Waals Complexes* ed N Halberstadt and K C Janda (New York: Plenum) pp 461–70

- [61] Chuang C, Andrews P M and Lester M I 1995 Intermolecular vibrations and spin-orbit predissociation dynamics of NeOH *J. Chem. Phys.* **103** 3418–29
- [62] Leforestier C, Braly L B, Liu K, Elrod M J and Saykally R J 1997 Fully coupled 6-dimensional calculations of the water dimer VRT states with a split Wigner pseudo-spectral approach *J. Chem. Phys.* **106** 8527–44
- [63] Saykally R J and Blake G A 1993 Molecular interactions and hydrogen bond tunneling dynamics: some new perspectives *Science* **259** 1570–5
- [64] Cotti G, Linnartz H, Meerts W L, van der Avoird A and Olthof E 1996 Stark effect and dipole moments of (NH₃)₂ in different vibration–rotation–tunneling states *J. Chem. Phys.* **104** 3898–906
- [65] Liu K 1997 *PhD Thesis* University of California at Berkeley
- [66] Viant M R, Cruzan J D, Lucas D D, Brown M G, Liu K and Saykally R J 1997 Pseudorotation in water trimer isotopomers using terahertz laser spectroscopy *J. Phys. Chem.* **101** 9032–41
- [67] Cruzan J D, Viant M R, Brown M G, Lucas D D, Liu K and Saykally R J 1998 Terahertz laser vibration–rotation–tunneling spectrum of the water pentamer-d(10). Constraints on the bifurcation tunneling dynamics *Chem. Phys. Lett.* **292** 667–76
- [68] Liu K, Brown M G and Saykally R J 1997 Terahertz laser vibration rotation tunneling spectroscopy and dipole moment of a cage form of the water hexamer *J. Phys. Chem.* **101** 8995–9010
- [69] Schmuttenmaer C A, Cohen R C, Pugliano N, Heath J R, Cooksy A L, Busarow K L and Saykally R J 1990 Tunable far-IR laser spectroscopy of jet-cooled carbon clusters—the ν_2 bending vibration of C₃ *Science* **249** 897–900
- [70] Heath J R, van Orden A, Hwang H J, Kuo E W, Tanaka K and Saykally R J 1994 Toward the detection of pure carbon clusters in the ISM *Adv. Space Res.* **15** 25–33
- [71] Liu K, Fellers R S, Viant M R, McLaughlin R P, Brown M G and Saykally R J 1996 A long pathlength pulsed slit valve appropriate for high temperature operation—infrared spectroscopy of jet-cooled large water clusters and nucleotide bases *Rev. Sci. Instrum.* **67** 410–16

FURTHER READING

Kimmitt M F 1970 *Far-Infrared Techniques* (London: Pion)

An excellent overview of optical approaches to THz spectrometers.

-31-

Janssen M A (ed) 1993 *Atmospheric Remote Sensing by Microwave Radiometry* (New York: Wiley)

The most complete guide to microwave and THz atmospheric sensing.

Carlstrom J and Zmuidzinas J 1996 Millimeter and submillimeter techniques *Review of Radio Science 1993–1996* ed W Ross Stone (Oxford: Oxford University Press) pp 839–82

A quite readable summary of heterodyne detection strategies.

Tsen K T (ed) 1999 Ultrafast phenomena in semiconductors III *Proc. SPIE* **624** 298

An overview of recent progress in this explosive field.

Gordy W and Cook R J 1991 *Microwave Molecular Spectra* (New York: Wiley)

The most complete textbook available, suitable for graduate students and researchers.

A thorough, advanced tutorial on the nature of clusters held together by intermolecular forces, and the theories that can be used to analyse them.

-1-

B1.5 Nonlinear optical spectroscopy of surfaces and interfaces

Jerry I Dadap and Tony F Heinz

B1.5.1 INTRODUCTION

B1.5.1.1 NONLINEAR OPTICS AND SPECTROSCOPY

Nonlinear optics is the study of the interaction between intense electromagnetic radiation and matter. It describes phenomena arising when the response of a medium to the electric field of light leaves the linear regime associated with the familiar and ubiquitous effects, such as reflection, refraction and absorption, comprising classical optics. In the presence of a sufficiently intense light source, the approximation of linearity breaks down. A new and much broader class of optical phenomena may be observed. Prototypical among these nonlinear optical effects is the production of light at new frequencies. Indeed, nonlinear optics is generally considered to have begun in 1961 when Franken and coworkers demonstrated optical second-harmonic generation (SHG) by insertion of a quartz crystal along the path of a laser beam [1]. In addition to the generation of new frequencies from excitation of a monochromatic source, nonlinear optical effects lead to the coupling between beams of identical and disparate frequencies, as well as to the action of a beam of light on itself.

Given the complexity of materials, it is perhaps surprising that a linear response to an applied optical field should be so common. This situation reflects the fact that the strength of electric fields for light encountered under conventional conditions is minute compared to that of the electric fields binding atoms and solids together. The latter may, for example, be estimated as $E_a \sim 1 \text{ V \AA}^{-1} = 10^8 \text{ V cm}^{-1}$. Since the irradiance of a light beam required to reproduce this electric field strength is $\sim 10^{13} \text{ W cm}^{-2}$, we may understand why a linear approximation of the material response is adequate for conventional light sources. With the advent of the laser, with its capability for producing high optical power and a high degree of coherence, this situation has changed. Under laser radiation, nonlinear optical effects are readily observed and widely exploited.

Over the past decades, nonlinear optics has come to have a broad impact on science and technology, playing a role in areas as diverse as telecommunications, materials processing and medicine. Within the context of chemical science, nonlinear optics is significant in providing new sources of coherent radiation, in permitting chemical processes to be induced under intense electromagnetic fields and in allowing matter to be probed by many powerful spectroscopic techniques. In this chapter, we shall be concerned only with the spectroscopic implications of nonlinear optics. In particular, our attention will be restricted to the narrowed range of spectroscopic techniques and applications related to probing surfaces and interfaces. This subject is a significant one. Surfaces and interfaces have been, and remain, areas of enormous scientific and technological importance. Sensitive and flexible methods of interface characterization are consequently of great value. As the advances discussed in this chapter reveal, nonlinear optics offers unique capabilities to address surface and interface analysis.

-2-

B1.5.1.2 PROBING SURFACES AND INTERFACES

The distinctive chemical and physical properties of surfaces and interfaces typically are dominated by the nature of one or two atomic or molecular layers [2, 3]. Consequently, useful surface probes require a very high degree of sensitivity. How can this sensitivity be achieved? For many of the valuable traditional probes of surfaces, the answer lies in the use of particles that have a short penetration depth through matter. These particles include electrons, atoms and ions, of appropriate energies. Some of the most familiar probes of solid surfaces, such as Auger electron spectroscopy (AES), low-energy electron diffraction (LEED), electron energy loss spectroscopy (EELS) and secondary ion mass spectroscopy (SIMS), exploit massive particles both approaching and leaving the surface. Other techniques, such as photoemission spectroscopy and inverse photoemission spectroscopy, rely on electrons for only half of the probing process, with photons serving for the other half. These approaches are complemented by those that directly involve the adsorbate of interest, such as molecular beam techniques and temperature programmed desorption (TPD). While these methods are extremely powerful, they are generally restricted to—or perform best for—probing materials under high vacuum conditions. This is a significant limitation, since many important systems are intrinsically incompatible with high vacuum (such as the surfaces of most liquids) or involve interfaces between two dense media. Scanning tunnelling microscopy (STM) is perhaps the electron-based probe best suited for investigations of a broader class of interfaces. In this approach, the physical proximity of the tip and the probe permits the method to be applied at certain interfaces between dense media.

Against this backdrop, the interest in purely optical probes of surfaces and interfaces can be easily understood. Since photons can penetrate an appreciable amount of material, photon-based methods are inherently appropriate to probing a very wide class of systems. In addition, photons, particularly those in the optical and infrared part of the spectrum, can be produced with exquisite control. Both the spatial and temporal properties of light beams can be tailored to the application. Particularly noteworthy is the possibility afforded by laser radiation of having either highly monochromatic radiation or radiation in the form of ultrafast (femtosecond) pulses. In addition, sources of high brightness are available, together with excellent detectors. As a consequence, within the optical spectral range several surface and interface probes have been developed. (Complementary approaches also exist within the x-ray part of the spectrum.) For optical techniques, one of the principal issues that must be addressed is the question of surface sensitivity. The ability of light to penetrate through condensed media was highlighted earlier as an attractive feature of optics. It also represents a potential problem in achieving the desired surface sensitivity, since one expects the bulk contribution to dominate that from the smaller region of the surface or interface.

Depending on the situation at hand, various approaches to achieving surface sensitivity may be appropriate for an optical probe. Some schemes rely on the presence of distinctive spectroscopic features in the surface region that may be distinguished from the response of the bulk media. This situation typically prevails in surface infrared spectroscopy [4] and surface-enhanced Raman scattering (SERS) [5, 6 and 7]. These techniques provide very valuable information about surface vibrational spectra, although the range of materials is somewhat restricted in the latter case. Ellipsometry [8, 9 and 10] permits a remarkably precise determination of the reflectivity of an interface through measurements of the polarization properties of light. It is a powerful tool for the analysis of thin films. Under appropriate conditions, it can be pushed to the limit of monolayer sensitivity. Since the method has, however, no inherent surface specificity, such applications generally require accurate knowledge and control of the relevant bulk media. A relatively recent addition to the set of optical probes is reflection difference absorption spectroscopy (RDAS) [10, 11]. In this scheme, the lowered symmetry of certain surfaces of crystalline materials is exploited in a differential measurement of reflectivity that cancels out the optical response of the bulk media.

In this chapter, we present a discussion of the nonlinear spectroscopic methods of second-harmonic generation (SHG) and sum-frequency generation (SFG) for probing surfaces and interfaces. While we have previously described the relative ease of observing nonlinear optical effects with laser techniques, it is still clear that linear optical methods will always be more straightforward to apply. Why then is nonlinear spectroscopy an attractive option for probing surfaces and interfaces? First, we may note that the method retains all of the advantages associated with optical methods. In addition, however, for a broad class of material systems the technique provides an *intrinsic* sensitivity to interfaces on the level of a single atomic layer. This is a very desirable feature that is lacking in linear optical probes. The relevant class of materials for which the nonlinear approach is inherently surface sensitive is quite broad. It consists of interfaces between all pairs of centrosymmetric materials. (Centrosymmetric materials, which include most liquids, gases, amorphous solids and elemental crystals, are those that remain unchanged when the position of every point is inverted through an appropriate origin.) The surface sensitivity of the SHG and SFG processes for these systems arises from a simple symmetry property: the second-order nonlinear optical processes of SHG and SFG are forbidden in centrosymmetric media. Thus, the bulk of the material does not exhibit a significant nonlinear optical response. On the other hand, the interfacial region, which necessarily breaks the inversion symmetry of the bulk, provides the desired nonlinear optical response.

Because of the generality of the symmetry principle that underlies the nonlinear optical spectroscopy of surfaces and interfaces, the approach has found application to a remarkably wide range of material systems. These include not only the conventional case of solid surfaces in ultrahigh vacuum, but also gas/solid, liquid/solid, gas/liquid and liquid/liquid interfaces. The information attainable from the measurements ranges from adsorbate coverage and orientation to interface vibrational and electronic spectroscopy to surface dynamics on the femtosecond time scale.

B1.5.1.3 SCOPE OF THE CHAPTER

In view of the diversity of material systems to which the SHG/SFG method has been applied and the range of the information that the method has yielded, we cannot give a comprehensive account of the technique in this chapter. For such accounts, we must refer the reader to the literature, particularly as summarized in various review articles [12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26 and 27] and monographs [10, 28, 29]. Our aim here is only to present an overview of the subject in which we attempt to describe the basic principles. The chapter is organized in the following fashion. We first outline basic theoretical considerations relevant to the technique, both in a brief general discussion of nonlinear optics and in a specific description of the nonlinear response of interfaces. After a few words about experimental techniques for surface SHG/SFG measurement, we devote the remainder of the chapter to describing the type of information that may be extracted from the nonlinear measurements. We have attempted at least to mention the different classes of information that have been obtained, such as adsorbate coverage or vibrational spectroscopy. In most cases, the corresponding approach has been widely and fruitfully applied in many experimental studies. Although we offer some representative examples, space does not permit us to discuss these diverse applications in any systematic way.

B1.5.2 THEORETICAL CONSIDERATIONS

B1.5.2.1 GENERAL BACKGROUND ON NONLINEAR OPTICS

(A) ANHARMONIC OSCILLATOR MODEL

In order to illustrate some of the basic aspects of the nonlinear optical response of materials, we first discuss the anharmonic oscillator model. This treatment may be viewed as the extension of the classical Lorentz model of the response of an atom or molecule to include nonlinear effects. In such models, the medium is treated as a collection of electrons bound about ion cores. Under the influence of the electric field associated with an optical wave, the ion cores move in the direction of the applied field, while the electrons are displaced in the opposite direction. These motions induce an oscillating dipole moment, which then couples back to the radiation fields. Since the ions are significantly more massive than the electrons, their motion is of secondary importance for optical frequencies and is neglected.

While the Lorentz model only allows for a restoring force that is linear in the displacement of an electron from its equilibrium position, the anharmonic oscillator model includes the more general case of a force that varies in a nonlinear fashion with displacement. This is relevant when the displacement of the electron becomes significant under strong driving fields, the regime of nonlinear optics. Treating this problem in one dimension, we may write an appropriate classical equation of motion for the displacement, x , of the electron from equilibrium as

$$m \left[\frac{d^2 x}{dt^2} + 2\gamma \frac{dx}{dt} + \omega_0^2 x - (b^{(2)} x^2 + b^{(3)} x^3 + \dots) \right] = -eE(t). \quad (\text{B1.5.1})$$

Here $E(t)$ denotes the applied optical field, and $-e$ and m represent, respectively, the electronic charge and mass. The (angular) frequency ω_0 defines the resonance of the harmonic component of the response, and γ represents a phenomenological damping rate for the oscillator. The nonlinear restoring force has been written in a Taylor expansion; the terms $m(b^{(2)} x^2 + b^{(3)} x^3 + \dots)$ correspond to the corrections to the harmonic restoring force, with parameters $b^{(2)}$, $b^{(3)}$, ... taken as material-dependent constants. In this equation, we have recognized that the excursion of the electron is typically small compared to the optical wavelength and have omitted any dependence of the driving term $-eE$ on the position of the electron.

Here we consider the response of the system to a monochromatic pump beam at a frequency ω ,

$$E(t) = E(\omega) \exp(-i\omega t) + \text{c.c.} \quad (\text{B1.5.2})$$

where the complex conjugate (c.c.) is included to yield an electric field $E(t)$ that is a real quantity. We use the symbol E to represent the electric field in both the time and frequency domain; the different arguments should make the interpretation clear. Note also that $E(\omega)$ and analogous quantities introduced later may be complex. As a first

approximation to the solution of [equation B1.5.1](#), we neglect the anharmonic terms to obtain the steady-state motion of the electron

$$x(t) = \frac{-eE(\omega)}{m} \frac{\exp(-i\omega t)}{\omega_0^2 - 2i\gamma\omega - \omega^2} + \text{c.c.} \quad (\text{B1.5.3})$$

This solution is appropriate for the regime of a weak driving field $E(t)$. If we now treat the material as a collection of non-interacting oscillators, we may write the induced polarization as a sum of the individual

dipole moments over a unit volume, i.e. $P(t) = -Nex(t)$, where N denotes the density of dipoles and local-field effects have been omitted. Following [equation B1.5.2](#), we express the $P(t)$ in the frequency domain as

$$P(t) = P(\omega) \exp(-i\omega t) + \text{c.c.} \quad (\text{B1.5.4})$$

We may then write the amplitude for the harmonically varying polarization as proportional to the corresponding quantity for the driving field, $E(\omega)$:

$$P(\omega) = \chi^{(1)}(\omega) E(\omega). \quad (\text{B1.5.5})$$

The constant of proportionality,

$$\chi^{(1)}(\omega) = \frac{Ne^2}{m} \frac{1}{\omega_0^2 - 2i\gamma\omega - \omega^2} \quad (\text{B1.5.6})$$

represents the linear susceptibility of the material. It is related to the dielectric constant $\varepsilon(\omega)$ by $\varepsilon(\omega) = 1 + 4\pi\chi^{(1)}(\omega)$.

Up to this point, we have calculated the linear response of the medium, a polarization oscillating at the frequency ω of the applied field. This polarization produces its own radiation field that interferes with the applied optical field. Two familiar effects result: a change in the speed of the light wave and its attenuation as it propagates. These properties may be related directly to the linear susceptibility $\chi^{(1)}(\omega)$. The index of refraction, $n = \text{Re}[\sqrt{1 + 4\pi\chi^{(1)}}]$, is associated primarily with the real part of $\chi^{(1)}(\omega)$. It describes the reduced (phase) velocity, c/n , of the optical wave travelling through the medium compared to its speed, c , in vacuum. The imaginary part, $\text{Im}[\chi^{(1)}(\omega)]$, on the other hand, gives rise to absorption of the radiation in the medium. The frequency dependence of the quantities $\text{Re}[\chi^{(1)}(\omega)]$ and $\text{Im}[\chi^{(1)}(\omega)]$ are illustrated in [figure B1.5.1](#). They exhibit a resonant response for optical frequencies ω near ω_0 , and show the expected dispersive and absorptive lineshapes.

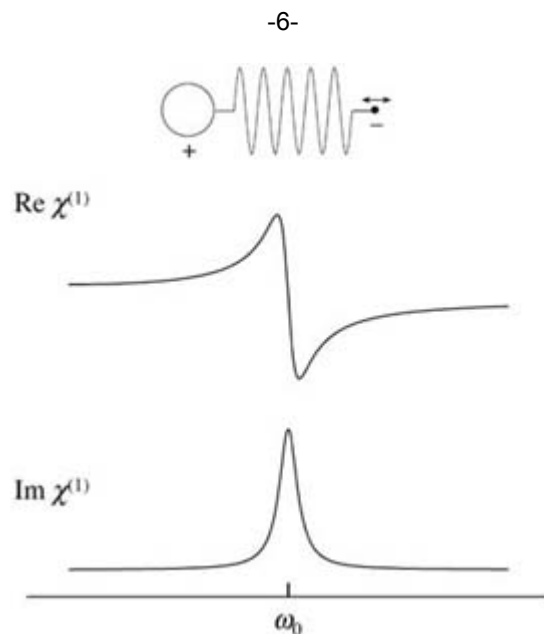


Figure B1.5.1 Anharmonic oscillator model: the real and imaginary parts of the linear susceptibility $\chi^{(1)}$ are

plotted as a function of angular frequency ω in the vicinity of the resonant frequency ω_0 .

If we now include the anharmonic terms in [equation B1.5.1](#), an exact solution is no longer possible. Let us, however, consider a regime in which we do not drive the oscillator too strongly, and the anharmonic terms remain small compared to the harmonic ones. In this case, we may solve the problem perturbatively. For our discussion, let us assume that only the second-order term in the nonlinearity is significant, i.e. $b^{(2)} \neq 0$ and $b^{(i)} = 0$ for $i > 2$ in [equation B1.5.1](#). To develop a perturbational expansion formally, we replace $E(t)$ by $\lambda E(t)$, where λ is the expansion parameter characterizing the strength of the field E . Thus, [equation B1.5.1](#) becomes

$$\ddot{x} + 2\gamma\dot{x} + \omega_0^2 x - b^{(2)}x^2 = -\lambda eE(t)/m. \quad (\text{B1.5.7})$$

We then write the solution of equation B1.5.7 as a power series expansion in terms of the strength λ of the perturbation:

$$x = \lambda x^{(1)} + \lambda^2 x^{(2)} + \lambda^3 x^{(3)} + \dots. \quad (\text{B1.5.8})$$

If we substitute the expression for B1.5.8 back into B1.5.7 and require that the terms proportional to λ and λ^2 on both sides of the resulting equation are equal, we obtain the equations

$$\ddot{x}^{(1)} + 2\gamma\dot{x}^{(1)} + \omega_0^2 x^{(1)} = -eE(t)/m \quad (\text{B1.5.9})$$

$$\ddot{x}^{(2)} + 2\gamma\dot{x}^{(2)} + \omega_0^2 x^{(2)} - b^{(2)}(x^{(1)})^2 = 0. \quad (\text{B1.5.10})$$

We immediately observe that the solution, $x^{(1)}$, to [equation B1.5.9](#) is simply that of the original harmonic oscillator problem given by [equation B1.5.3](#). Substituting this result for $x^{(1)}$ into the last term of [equation B1.5.10](#) and solving for the second-order term $x^{(2)}$, we obtain two solutions: one oscillating at twice the frequency of the applied field and a static part at a frequency of $\omega = 0$. This behaviour arises from the fact that the square of the term $x^{(1)}(t)$, which now acts as a source term in [equation B1.5.10](#), possesses frequency components at frequencies 2ω and zero. The material response at the frequency 2ω corresponds to SHG, while the response at frequency zero corresponds to the phenomenon of optical rectification. The effect of a linear and nonlinear relation between the driving field and the material response is illustrated in figure B1.5.2.

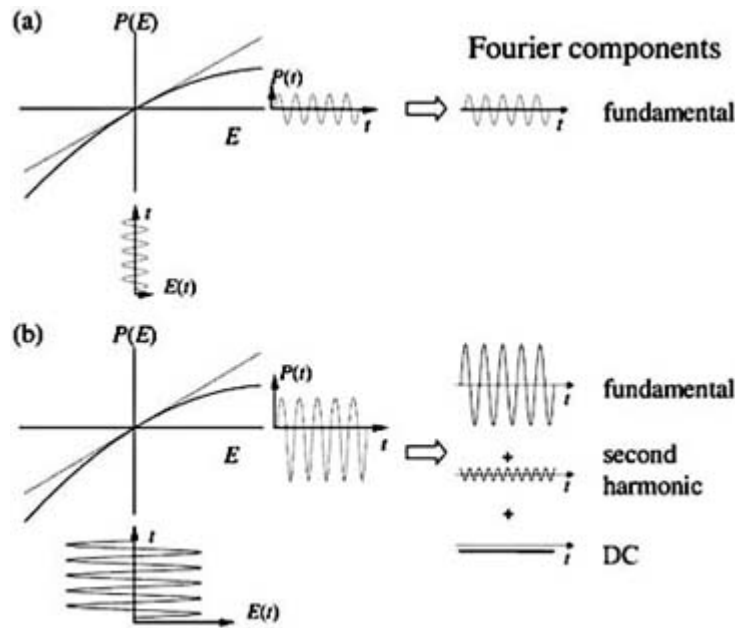


Figure B1.5.2 Nonlinear dependence of the polarization P on the electric field E . (a) For small sinusoidal input fields, P depends linearly on E ; hence its harmonic content is mainly that of E . (b) For a stronger driving electric field E , the polarization waveform becomes distorted, giving rise to new harmonic components. The second-harmonic and DC components are shown.

In analogy to [equation B1.5.3](#), we can write the steady-state solution to [equation B1.5.10](#) for the SHG process as

$$x^{(2)}(t) = x^{(2)}(2\omega) \exp(-2i\omega t) + \text{c.c.} \quad (\text{B1.5.11})$$

The amplitude of the response, $x^{(2)}(2\omega)$, is given by the steady-state solution of [equation B1.5.10](#) as

$$x^{(2)}(2\omega) = \frac{(e/m)^2 b^{(2)} E(\omega)^2}{D(2\omega)D^2(\omega)} \quad (\text{B1.5.12})$$

-8-

where the quantity $D(\omega)$ is defined as $D(\omega) \equiv \omega_0^2 - \omega^2 - 2i\gamma\omega$. Similarly, the amplitude of the response at frequency zero for the optical rectification process is given by

$$x^{(2)}(0) = \frac{2(e/m)^2 b^{(2)} E(\omega)E(\omega)^*}{D(0)D(\omega)D(-\omega)}. \quad (\text{B1.5.13})$$

Following the derivation of the linear susceptibility, we may now readily deduce the second-order susceptibility $\chi^{(2)}(2\omega = \omega + \omega)$ for SHG, as well as $\chi^{(2)}(0 = \omega - \omega)$ for the optical rectification process. Defining the second-order nonlinear susceptibility for SHG as the relation between the square of the relevant components of the driving fields and the nonlinear source polarization,

$$(\text{B1.5.14})$$

$$P(2\omega) = \chi^{(2)}(2\omega)E(\omega)E(\omega)$$

we obtain

$$\chi^{(2)}(2\omega = \omega + \omega) = \frac{-(e^2/m^2)Nb^{(2)}}{D(2\omega)D^2(\omega)}. \quad (\text{B1.5.15})$$

As we shall discuss later in a detailed fashion, the nonlinear polarization associated with the nonlinear susceptibility of a medium acts as a source term for radiation at the second harmonic (SH) frequency 2ω . Since there is a definite phase relation between the fundamental pump radiation and the nonlinear source term, coherent SH radiation is emitted in well-defined directions. From the quadratic variation of $P(2\omega)$ with $E(\omega)$, we expect that the SH intensity $I_{2\omega}$ will also vary quadratically with the pump intensity I_ω .

If we compare the nonlinear response of $\chi^{(2)}(2\omega = \omega + \omega)$ with the linear material response of $\chi^{(1)}(\omega)$, we find both similarities and differences. As is apparent from equation B1.5.15 and shown pictorially in [figure B1.5.3](#) $\chi^{(2)}(2\omega = \omega + \omega)$ exhibits a resonant enhancement for frequencies ω near ω_0 , just as in the case for $\chi^{(1)}(\omega)$. However, in addition to this so-called one-photon resonance, $\chi^{(2)}(2\omega = \omega + \omega)$ also displays a resonant response when 2ω is near ω_0 or $\omega \approx \omega_0/2$. This feature is termed a two-photon resonance and has no analogue in linear spectroscopy. Despite these differences between $\chi^{(1)}$ and $\chi^{(2)}$, one can see that both types of response provide spectroscopic information about the material system. A further important difference concerns symmetry characteristics. The linear response $\chi^{(1)}$ may be expected to be present in any material. The second-order nonlinear response $\chi^{(2)}$ requires the material to exhibit a non-centrosymmetric character, i.e. in the one-dimensional model, the $+x$ and $-x$ directions must be distinguishable. To understand this property, consider the potential energy associated with the restoring force on the electron. This potential may be written in the notation previously introduced as $V(x) = \frac{1}{2}m\omega_0^2x^2 - \frac{1}{3}mb^{(2)}x^3 + \dots$. The first term is allowed for all materials; the second term, however, cannot be present in centrosymmetric materials since it clearly differentiates between a displacement of $+x$ and $-x$. Thus $b^{(2)} = 0$ except in non-centrosymmetric materials. This symmetry distinction is the basis for the remarkable surface and interface sensitivity of SHG and SFG for bulk materials with inversion symmetry.

-9-

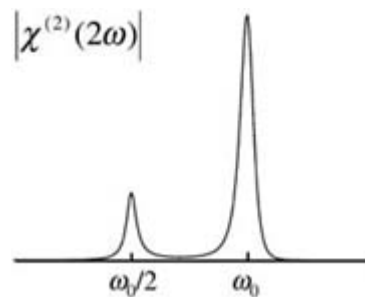


Figure B1.5.3 Magnitude of the second-order nonlinear susceptibility $\chi^{(2)}$ versus frequency ω , obtained from the anharmonic oscillator model, in the vicinity of the single- and two-photon resonances at frequencies ω_0 and $\omega_0/2$, respectively.

(B) MAXWELL'S EQUATIONS

We now embark on a more formal description of nonlinear optical phenomena. A natural starting point for this discussion is the set of Maxwell equations, which are just as valid for nonlinear optics as for linear optics.

In the absence of free charges and current densities, we have in cgs units:

$$\nabla \times \mathbf{H} = \frac{1}{c} \frac{\partial \mathbf{D}}{\partial t} \quad (\text{B1.5.16})$$

$$\nabla \times \mathbf{E} = -\frac{1}{c} \frac{\partial \mathbf{B}}{\partial t} \quad (\text{B1.5.17})$$

$$\nabla \cdot \mathbf{B} = 0 \quad (\text{B1.5.18})$$

$$\nabla \cdot \mathbf{D} = 0 \quad (\text{B1.5.19})$$

where \mathbf{E} and \mathbf{H} are the electric and magnetic field intensities, respectively; \mathbf{D} and \mathbf{B} are the electric displacement and magnetic induction, respectively. In the optical regime, we generally neglect the magnetic response of the material and take $\mathbf{B} = \mathbf{H}$. The material response is then incorporated into the Maxwell equations through the displacement vector \mathbf{D} . This quantity is related to the electric field \mathbf{E} and the polarization \mathbf{P} (the electric dipole moment per unit volume) by

$$\mathbf{D} = \mathbf{E} + 4\pi \mathbf{P}. \quad (\text{B1.5.20})$$

-10-

The polarization \mathbf{P} is given in terms of \mathbf{E} by the constitutive relation of the material. For the present discussion, we assume that the polarization $\mathbf{P}(\mathbf{r})$ depends only on the field \mathbf{E} evaluated at the same position \mathbf{r} . This is the so-called dipole approximation. In later discussions, however, we will consider, in some specific cases, the contribution of a polarization that has a non-local spatial dependence on the optical field. Once we have augmented the system of [equation B1.5.16](#), [equation B1.5.17](#), [equation B1.5.18](#), [equation B1.5.19](#) and [equation B1.5.20](#) with the constitutive relation for the dependence of \mathbf{P} on \mathbf{E} , we may solve for the radiation fields. This relation is generally characterized through the use of linear and nonlinear susceptibility tensors, the subject to which we now turn.

(C) NONLINEAR OPTICAL SUSCEPTIBILITIES

If the polarization of a given point in space and time (\mathbf{r}, t) depends only on the driving electric field at the same coordinates, we may write the polarization as $\mathbf{P} = \mathbf{P}(\mathbf{E})$. In this case, we may develop the polarization in power series as $\mathbf{P} = \mathbf{P}_L + \mathbf{P}_{NL} = \mathbf{P}^{(1)} + \mathbf{P}^{(2)} + \mathbf{P}^{(3)} + \dots$, where the linear term is $\mathbf{P}_i^{(1)} = \sum_j \chi_{ij}^{(1)} E_j$ and the nonlinear terms include the second-order response $\mathbf{P}_i^{(2)} = \sum_{jk} \chi_{ijk}^{(2)} E_j E_k$, the third-order response $\mathbf{P}_i^{(3)} = \sum_{jkl} \chi_{ijkl}^{(3)} E_j E_k E_l$, and so forth. The coefficients $\chi_{ij}^{(1)}$, $\chi_{ijk}^{(2)}$, and $\chi_{ijkl}^{(3)}$, are, respectively, the linear, second-order nonlinear and the third-order nonlinear susceptibilities of the material. The quantity $\chi_{ijk\dots}^{(n)}$, it should be noted, is a tensor of rank $n + 1$, and $ijkl$ refer to indices of Cartesian coordinates. The simple formulation just presented does not allow for the variation of the optical response with frequency, a behaviour of critical importance for spectroscopy. We now briefly discuss how to incorporate frequency-dependent behaviour into the polarization response. To treat the frequency response of the material, we consider an excitation electric field of the form of a superposition of monochromatic fields

$$\mathbf{E}(t) = \sum_m \mathbf{E}(\omega_m) e^{-i\omega_m t} \quad (\text{B1.5.21})$$

where the summation extends over all positive and negative frequency components. Since $\mathbf{E}(t)$ represents a physical field, it is constrained to be real and $\mathbf{E}(-\omega_m) = \mathbf{E}(\omega_m)^*$. In the same manner, we can write the polarization \mathbf{P} as

$$\mathbf{P}(t) = \sum_n \mathbf{P}(\omega_n) e^{-i\omega_n t}. \quad (\text{B1.5.22})$$

Here the collection of frequencies in the summation may include new frequencies ω_n in addition to those in summation of equation B1.5.21 for the applied field. The total polarization can be separated into linear, \mathbf{P}_L , and nonlinear, \mathbf{P}_{NL} , parts:

$$\mathbf{P}(t) = \mathbf{P}_L(t) + \mathbf{P}_{NL}(t) = \mathbf{P}^{(1)}(t) + \mathbf{P}^{(2)}(t) + \mathbf{P}^{(3)}(t) + \dots \quad (\text{B1.5.23})$$

where $\mathbf{P}_L(t) = \mathbf{P}^{(1)}(t)$ and $\mathbf{P}_{NL}(t) = \mathbf{P}^{(2)}(t) + \mathbf{P}^{(3)}(t) + \dots$, and the terms in \mathbf{P} correspond to an expansion in powers of the field \mathbf{E} .

-11-

The linear susceptibility, $\chi_{ij}^{(1)}$, is the factor that relates the induced linear polarization to the applied field:

$$P_i^{(1)}(\omega_n) = \sum_j \chi_{ij}^{(1)}(\omega_n) E_j(\omega_n). \quad (\text{B1.5.24})$$

$\chi_{ij}^{(1)}(\omega_n)$ in this formulation gives rise, as one expects, to a polarization oscillating as the applied frequency ω_n , but may now incorporate a strength that varies with ω_n , as illustrated earlier in the harmonic oscillator model. Similarly, we can define the corresponding frequency-dependent second-order, $\chi_{ijk}^{(2)}$, and third-order, $\chi_{ijkl}^{(3)}$, susceptibility tensors by

$$P_i^{(2)}(\omega_q + \omega_r) = p \sum_{jk} \chi_{ijk}^{(2)}(\omega_q + \omega_r, \omega_q, \omega_r) E_j(\omega_q) E_k(\omega_r) \quad (\text{B1.5.25})$$

$$P_i^{(3)}(\omega_q + \omega_r + \omega_s) = p \sum_{jkl} \chi_{ijkl}^{(3)}(\omega_q + \omega_r + \omega_s, \omega_q, \omega_r, \omega_s) E_j(\omega_q) E_k(\omega_r) E_l(\omega_s) \quad (\text{B1.5.26})$$

where the quantity p is called the degeneracy factor and is equal to the number of distinct permutations of the applied frequencies $\{\omega_q, \omega_r\}$ and $\{\omega_q, \omega_r, \omega_s\}$ for the second- and third-order processes, respectively. The inclusion of the degeneracy factor p ensures that the nonlinear susceptibility is not discontinuous when two of the fields become degenerate, e.g. the nonlinear susceptibility $\chi_{ijk}^{(2)}(\omega_1 + \omega_2, \omega_1, \omega_2)$ approaches the value $\chi_{ijk}^{(2)}(2\omega_1, \omega_1, \omega_1)$ as ω_2 approaches the value ω_1 [32]. As can be seen from equation B1.5.25 and equation B1.5.26, the first frequency argument of the nonlinear susceptibility is equal to the sum of the rest of its

frequency arguments. Note also that these frequencies are not constrained to positive values and that the complete material response involves both the positive and negative frequency components. We now consider some of the processes described by the nonlinear susceptibilities. For the case of the second-order nonlinear optical effects (equation B1.5.25), three processes can occur when the frequencies ω_1 and ω_2 are distinct. This can be seen by expanding equation B1.5.25:

$$P_i^{(2)}(\omega_3) = 2 \sum_{jk} \chi_{ijk}^{(2)}(\omega_3, \omega_1, \omega_2) E_j(\omega_1) E_k(\omega_2) \quad (\text{B1.5.27})$$

$$P_i^{(2)}(2\omega_\alpha) = \sum_{jk} \chi_{ijk}^{(2)}(2\omega_\alpha, \omega_\alpha, \omega_\alpha) E_j(\omega_\alpha) E_k(\omega_\alpha) \quad \alpha = 1, 2 \quad (\text{B1.5.28})$$

$$P_i^{(2)}(\mathbf{0}) = 2 \left[\sum_{jk} \chi_{ijk}^{(2)}(\mathbf{0}, \omega_1, -\omega_1) E_j(\omega_1) E_k^*(\omega_1) + \sum_{jk} \chi_{ijk}^{(2)}(\mathbf{0}, \omega_2, -\omega_2) E_j(\omega_2) E_k^*(\omega_2) \right]. \quad (\text{B1.5.29})$$

-12-

These effects correspond, respectively, to the processes of sum-frequency generation (SFG), SHG and optical rectification.

For the case of third-order nonlinear optical effects (equation B1.5.26), a wide variety of processes are described by the different possible combinations of applied frequencies. We shall not attempt to catalogue them here. The most intuitive case is that of third-harmonic generation ($3\omega = \omega + \omega + \omega$), corresponding to addition of three equal frequencies. When one of the frequencies is zero (i.e. a DC field) one obtains the so-called electric-field induced SHG (EFISH) process ($2\omega = \omega + \omega + 0$). Several third-order effects have found significant use in both frequency and time-domain spectroscopy. These include notably coherent anti-Stokes Raman scattering, stimulated Raman scattering, general and degenerate four-wave mixing and two-photon absorption [30, 31].

(D) SYMMETRY PROPERTIES

The second-order nonlinear susceptibility tensor $\chi^{(2)}(\omega_3, \omega_2, \omega_1)$ introduced earlier will, in general, consist of 27 distinct elements, each displaying its own dependence on the frequencies ω_1 , ω_2 and $\omega_3 (= \pm \omega_1 \pm \omega_2)$. There are, however, constraints associated with spatial and time-reversal symmetry that may reduce the complexity of $\chi^{(2)}$ for a given material [32, 33 and 34]. Here we examine the role of spatial symmetry.

The most significant symmetry property for the second-order nonlinear optics is inversion symmetry. A material possessing inversion symmetry (or centrosymmetry) is one that, for an appropriate origin, remains unchanged when all spatial coordinates are inverted via $\mathbf{r} \rightarrow -\mathbf{r}$. For such materials, the second-order nonlinear response vanishes. This fact is of sufficient importance that we shall explain its origin briefly. For a centrosymmetric material, $\chi^{(2)}$ should remain unchanged under an inversion operation, since the material by hypothesis does not change. On the other hand, the nature of the physical quantities \mathbf{E} and \mathbf{P} implies that they must obey $\mathbf{E} \rightarrow -\mathbf{E}$ and $\mathbf{P} \rightarrow -\mathbf{P}$ under the inversion operation [35]. The relation $P_i^{(2)} = \sum_{jk} \chi_{ijk}^{(2)} E_j E_k$ then yields $\sum_{jk} \chi_{ijk}^{(2)} E_j E_k = -\sum_{jk} \chi_{ijk}^{(2)} E_j E_k$, whence $\chi_{ijk}^{(2)} = \mathbf{0}$. A further useful symmetry relation applies specifically to SHG process. Since the incident fields $E_j(\omega)$ and $E_k(\omega)$ are identical, we may take

$\chi_{ijk}^{(2)} = \chi_{ikj}^{(2)}$ without loss of information.

For materials that exhibit other classes of spatial symmetry not including centrosymmetry, we expect that $\chi_{ijk}^{(2)}$ will be non-vanishing but will display a simplified form [37]. To see how this might work, consider a crystal in which the x and y directions are equivalent. The nonlinear response of the medium would consequently be equivalent for an applied optical field polarized along either the x or y direction. It then follows, for example, that the nonlinear susceptibilities $\chi_{zxx}^{(2)}$ and $\chi_{zyy}^{(2)}$ are equal, as are other elements in which the indices x and y are exchanged. This reduces the complexity of $\chi^{(2)}$ significantly. Table B1.5.1 presents the form of the second-order nonlinear susceptibility relevant for the classes of symmetry encountered at isotropic and crystalline surfaces and interfaces.

(E) QUANTUM MECHANICAL DESCRIPTION

Having now developed some of the basic notions for the macroscopic theory of nonlinear optics, we would like to discuss how the microscopic treatment of the nonlinear response of a material is handled. While the classical nonlinear

-13-

oscillator model provides us with a qualitative feeling for the phenomenon, a quantitative theory must generally begin with a quantum mechanical description. As in the case of linear optics, quantum mechanical calculations in nonlinear optics are conveniently described by perturbation theory and the density matrix formalism [36, 37]. The heart of this microscopic description is the interaction Hamiltonian $H_{\text{int}} = -\mathbf{\mu} \cdot \mathbf{E}(t)$, which characterizes the interaction of the system with the radiation and is treated as a perturbation. Here $\mathbf{\mu} = -e\mathbf{r}$ is the electric dipole operator and $\mathbf{E}(t)$ is the applied optical field. Applying this formalism with first-order perturbation theory to calculate the induced dipole moment $\mathbf{\mu}$, we obtain a linear susceptibility that is proportional to the product of two matrix elements of $\mathbf{\mu}$, $(\mu_i)_{gn}(\mu_j)_{ng}$, where $(\mu_i)_{gn} \equiv \langle g | \mu_i | n \rangle$. This product may be viewed as arising from a process in which a photon is first destroyed in a real or virtual transition from a populated energy eigenstate $|g\rangle$ to an empty state $|n\rangle$; a photon is then emitted in the transition back to the initial state $|g\rangle$.

The second-order nonlinear optical processes of SHG and SFG are described correspondingly by second-order perturbation theory. In this case, two photons at the driving frequency or frequencies are destroyed and a photon at the SH or SF is created. This is accomplished through a succession of three real or virtual transitions, as shown in figure B1.5.4. These transitions start from an occupied initial energy eigenstate $|g\rangle$, pass through intermediate states $|n'\rangle$ and $|n\rangle$ and return to the initial state $|g\rangle$. A full calculation of the second-order response for the case of SFG yields [37]

$$\chi_{ijk}^{(2)}(\omega_3, \omega_2, \omega_1) = -\frac{N}{\hbar^2} \sum_{g,n,n'} \frac{(\mu_i)_{gn}(\mu_j)_{nn'}(\mu_k)_{n'g}}{(\omega_3 - \omega_{ng} + i\Gamma_{ng})(\omega_2 - \omega_{n'g} + i\Gamma_{n'g})} \rho_g^{(0)} \quad (\text{B1.5.30})$$

+ seven similar terms.

As for the linear response, the transitions occur through the electric-dipole operator $\mathbf{\mu}$ and are characterized by the matrix elements $(\mu_i)_{gn}$. In equation B1.5.30, the energy denominators involve the energy differences $\hbar\omega_{ng} \equiv E_n - E_g$ and widths $\hbar\Gamma_{ng}$ for transitions between eigenstates $|n\rangle$ and $|g\rangle$. The formula includes a sum over different possible ground states weighted by the factor $\rho_g^{(0)}$ representing the probability that state $|g\rangle$ is occupied. It is assumed that the material can be treated as having localized electronic states of density N per unit volume and that their interaction and local-field effects may be neglected. Corresponding expressions

result for delocalized electrons in crystalline solids [36, 37].

The frequency denominators in the eight terms of equation B1.5.30 introduce a resonant enhancement in the nonlinearity when any of the three frequencies ($\omega_1, \omega_2, \omega_3$) coincides with a transition from the ground state $|g\rangle$ to one of the intermediate states $|n'\rangle$ and $|n\rangle$. The numerator of each term, which consists of the product of the three dipole matrix elements $(\mu_i)_{gn}(\mu_j)_{n'n}(\mu_k)_{n'g}$, reflects, through its tensor character, the structural properties of the material, as well as the details of the character of the relevant energy eigenstates.

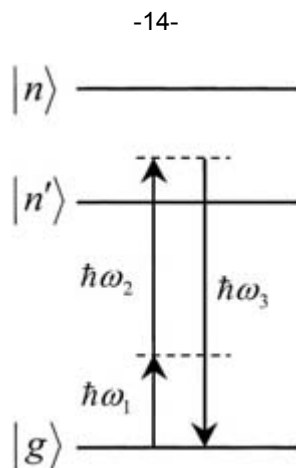


Figure B1.5.4 Quantum mechanical scheme for the SFG process with ground state $|g\rangle$ and excited states $|n'\rangle$ and $|n\rangle$.

B1.5.2.2 NONLINEAR OPTICS OF THE INTERFACE

The focus of the present chapter is the application of second-order nonlinear optics to probe surfaces and interfaces. In this section, we outline the phenomenological or macroscopic theory of SHG and SFG at the interface of centrosymmetric media. This situation corresponds, as discussed previously, to one in which the relevant nonlinear response is forbidden in the bulk media, but allowed at the interface.

(A) INTERFACIAL CONTRIBUTION

In order to describe the second-order nonlinear response from the interface of two centrosymmetric media, the material system may be divided into three regions: the interface and the two bulk media. The interface is defined to be the transitional zone where the material properties—such as the electronic structure or molecular orientation of adsorbates—or the electromagnetic fields differ appreciably from the two bulk media. For most systems, this region occurs over a length scale of only a few Ångströms. With respect to the optical radiation, we can thus treat the nonlinearity of the interface as localized to a sheet of polarization. Formally, we can describe this sheet by a nonlinear dipole moment per unit area, $\mathbf{P}_s^{(2)}$, which is related to a second-order bulk polarization $\mathbf{P}^{(2)}$ by $\mathbf{P}^{(2)}$ by $\mathbf{P}^{(2)}(x, y, z, t) = \mathbf{P}_s^{(2)}(x, y, t)\delta(z)$. Here z is the surface normal direction, and the x and y axes represent the in-plane coordinates (figure B1.5.5).

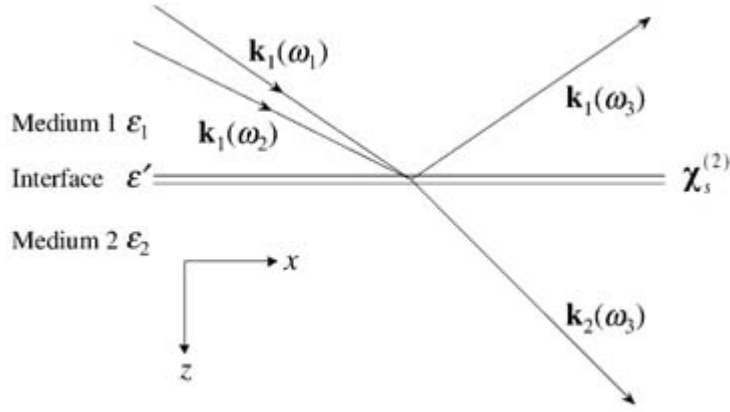


Figure B1.5.5 Schematic representation of the phenomenological model for second-order nonlinear optical effects at the interface between two centrosymmetric media. Input waves at frequencies ω_1 and ω_2 , with corresponding wavevectors $\mathbf{k}_1(\omega_1)$ and $\mathbf{k}_1(\omega_2)$, are approaching the interface from medium 1. Nonlinear radiation at frequency ω_3 is emitted in directions described by the wavevectors $\mathbf{k}_1(\omega_3)$ (reflected in medium 1) and $\mathbf{k}_2(\omega_3)$ (transmitted in medium 2). The linear dielectric constants of media 1, 2 and the interface are denoted by ε_1 , ε_2 , and ε' , respectively. The figure shows the xz -plane (the plane of incidence) with z increasing from top to bottom and $z = 0$ defining the interface.

The nonlinear response of the interface may then be characterized in terms of a surface (or interface) nonlinear susceptibility tensor $\chi_s^{(2)}$. This quantity relates the applied electromagnetic fields to the induced surface nonlinear polarization $\mathbf{P}_s^{(2)}$:

$$\mathbf{P}_s^{(2)}(\omega_3) = \chi_s^{(2)}(\omega_3 = \omega_1 + \omega_2) : \mathbf{E}(\omega_1)\mathbf{E}(\omega_2). \quad (\text{B1.5.31})$$

In this equation as well as in the succeeding discussions, we have suppressed, for notational simplicity, the permutation or degeneracy factor of two, required for SFG.

To define this model fully, we must specify the linear dielectric response in the vicinity of our surface nonlinear susceptibility. We do this in a general fashion by introducing a (frequency-dependent) linear dielectric response of the interfacial region ε' , which is bounded by the bulk media with dielectric constants ε_1 and ε_2 . For simplicity, we consider all of these quantities to be scalar, corresponding to isotropic linear optical properties. The phenomenological model for second-order nonlinear optical effects is summarized in figure B1.5.5 (An alternative convention [38, 39] for defining the surface nonlinear susceptibility is one in which the fundamental fields $\mathbf{E}(\omega_1)$ and $\mathbf{E}(\omega_2)$ are taken as their value in the lower medium and the polarized sheet is treated as radiating in the upper medium. This convention corresponds in our model to the assignments of $\varepsilon'(\omega_1) = \varepsilon_2(\omega_1)$, $\varepsilon'(\omega_2) = \varepsilon_2(\omega_2)$, and $\varepsilon'(\omega_3) = \varepsilon_1(\omega_3)$.)

From the point of view of tensor properties, the surface nonlinear susceptibility $\chi_s^{(2)}$ is quite analogous to the bulk nonlinear response $\chi_{ijk}^{(2)}$ in a non-centrosymmetric medium. Consequently, in the absence of any symmetry constraints, $\chi_s^{(2)}$ will exhibit 27 independent elements for SFG and, because of symmetry for the last two indices, 18 independent elements for SHG. If the surface exhibits certain in-plane symmetry properties, then the form of $\chi_s^{(2)}$ will be simplified

correspondingly. For the common situation of an isotropic surface, for example, the allowed elements of

$\chi_s^{(2)}$ may be denoted as $\chi_{s,\perp\perp\perp}^{(2)}$, $\chi_{s,\perp\parallel\parallel}^{(2)}$, $\chi_{s,\parallel\perp\parallel}^{(2)}$, and $\chi_{s,\parallel\parallel\perp}^{(2)}$ where \perp corresponds to the z direction and \parallel refers to either x or y . For the case of SHG, where the fundamental frequencies are equal, $\omega_1 = \omega_2 \equiv \omega$, the tensor elements $\chi_{s,\perp\perp\perp}^{(2)}$ and $\chi_{s,\parallel\parallel\perp}^{(2)}$ are likewise equivalent. The non-vanishing elements of $\chi_s^{(2)}$ for other commonly encountered surface symmetries are summarized in table B1.5.1.

Table B1.5.1 Independent non-vanishing elements of the nonlinear susceptibility, $\chi_s^{(2)}$ for an interface in the xy -plane for various symmetry classes. When mirror planes are present, at least one of them is perpendicular to the y -axis. For SHG, elements related by the permutation of the last two elements are omitted. For SFG, these elements are generally distinct; any symmetry constraints are indicated in parentheses. The terms enclosed in parentheses are antisymmetric elements present only for SFG. (After [71])

Symmetry class	Independent non-vanishing elements
1	$xxx, xxy, xyy, xyz, xxz, xzz, yxx, yxy, yxz, yyy, yyz, yzz, zxx,$ zxy, zxz, zyy, zyz, zzz
[3pt] m	$xxx, xyy, xzx, xzz, yxy, yyz, zxx, zxz, zyy, zzz$ $xzx, xyz, yxz, yzy, zxx, zyy, zxy, zzz$
$2mm$	xzx, yzy, zxx, zyy, zzz
3	$xxx=-xyy=-yyx(=-yxy)$ $yyy=-yxx=-xyx(=-xxy), yzy=xzx,$ $zxx=zyy, xyz=-yxz$ $zzz, (zxy=-zyx)$
$3m$	$xxx=-xyy=-yxy(=-yyx)$ $yzy=xzx, zxx=zyy, zzz$
4,6, ∞	$xxz=yyz, zxx=zyy, xyz=-yxz, zzz, (zxy=-zyx)$
$4mm, 6mm, \infty m$	$xxz=yyz, zxx=zyy, zzz$

The linear and nonlinear optical responses for this problem are defined by ϵ_1 , ϵ_2 , ϵ' and $\chi_s^{(2)}$, respectively, as indicated in figure B1.5.5. In order to determine the nonlinear radiation, we need to introduce appropriate pump radiation fields $E(\omega_1)$ and $E(\omega_2)$. If these pump beams are well-collimated, they will give rise to well-collimated radiation emitted through the surface nonlinear response. Because the nonlinear response is present only in a thin layer, phase matching [37] considerations are unimportant and nonlinear emission will be present in both transmitted and reflected directions.

-17-

Here we model the pump beams associated with fields $E(\omega_1)$ and $E(\omega_2)$ as plane waves with wavevectors $\mathbf{k}_1 = \hat{\mathbf{k}}_1 \omega_1 \sqrt{\epsilon_1(\omega_1)}/c$ and $\mathbf{k}_2 = \hat{\mathbf{k}}_2 \omega_2 \sqrt{\epsilon_1(\omega_2)}/c$. The directions of the reflected and transmitted beams can then be obtained simply through conservation of the in-plane component of the wavevector, i.e. $k_{1x}(\omega_1) + k_{1x}(\omega_2) = k_{1x}(\omega_3) = k_{2x}(\omega_3)$. This is the nonlinear optical analogue of Snell's law. For the case of SHG, this equation may be reduced to $n(\omega) \sin \theta_\omega = n(2\omega) \sin \theta_{2\omega}$ for the angle of the incident pump radiation, θ_ω , and the angle of the emitted nonlinear beams, $\theta_{2\omega}$. The refractive indices in this equation correspond to those of the relevant bulk medium through which the beams propagate. For reflection in a non-dispersive medium, we obtain simply $\theta_\omega = \theta_{2\omega}$, as for the law of reflection. For the transmitted beam, the relation in the absence of dispersion reduces to the usual Snell's law for refraction.

A full solution of the nonlinear radiation follows from the Maxwell equations. The general case of radiation from a second-order nonlinear material of finite thickness was solved by Bloembergen and Pershan in 1962 [40]. That problem reduces to the present one if we let the interfacial thickness approach zero. Other equivalent solutions involved the application of the boundary conditions for a polarization sheet [14] or the

use of a Green's function formalism for the surface [38, 39].

From such a treatment, we may derive explicit expressions for the nonlinear radiation in terms of the linear and nonlinear response and the excitation conditions. For the case of nonlinear reflection, we obtain an irradiance for the radiation emitted at the nonlinear frequency ω_3 of

$$I(\omega_3) = \frac{8\pi^3 \omega_3^2 \sec^2 \theta |e'(\omega_3) \cdot \chi_s^{(2)} : e'(\omega_1) e'(\omega_2)|^2 I(\omega_1) I(\omega_2)}{c^3 [\varepsilon_1(\omega_3) \varepsilon_1(\omega_1) \varepsilon_1(\omega_2)]^{1/2}} \quad (\text{B1.5.32})$$

where $I(\omega_1)$ and $I(\omega_2)$ denote the intensities of the pump beams at frequencies ω_1 and ω_2 incident from medium 1; c is the speed of light in vacuum; θ is the angle of propagation direction of the nonlinear radiation relative to the surface normal. The vectors $e'(\omega_1)$, $e'(\omega_2)$ and $e'(\omega_3)$ represent the *unit* polarization vectors $\hat{e}_1(\omega_1)$, $\hat{e}_1(\omega_2)$, and $\hat{e}_1(\omega_3)$, respectively, adjusted to account for the linear optical propagation of the waves. More specifically, we may write

$$e'(\omega) = F_{1 \rightarrow 2} \hat{e}_1(\omega). \quad (\text{B1.5.33})$$

Here the 'Fresnel transformation' $F_{1 \rightarrow 2}$ describes the relationship between the electric field $E\hat{e}_1$ in medium 1 (propagating towards medium 2) and the resulting field Ee' at the interface. For light incident in the x - z plane as shown in [figure B1.5.5](#) $F_{1 \rightarrow 2}$ is a diagonal matrix whose elements are

$$F_{1 \rightarrow 2}^{xx} = 2\varepsilon_1 k_{2,z} / (\varepsilon_2 k_{1,z} + \varepsilon_1 k_{2,z}) \quad (\text{B1.5.34})$$

$$F_{1 \rightarrow 2}^{yy} = 2k_{1,z} / (k_{1,z} + k_{2,z}) \quad (\text{B1.5.35})$$

$$F_{1 \rightarrow 2}^{zz} = 2(\varepsilon_1 \varepsilon_2 / \varepsilon') k_{1,z} / (\varepsilon_2 k_{1,z} + \varepsilon_1 k_{2,z}) \quad (\text{B1.5.36})$$

-18-

where the quantity $k_{i,z}$ denotes the magnitude of the z -component of the wavevector in medium i at the relevant wavelength (ω_1 , ω_2 or ω_3).

The treatment of this section has been based on an assumed nonlinear surface response $\chi_s^{(2)}$ and has dealt entirely with electromagnetic considerations of excitation and radiation from the interface. A complete theoretical picture, however, includes developing a microscopic description of the surface nonlinear susceptibility. In the discussion in [section B1.5.4](#), we will introduce some simplified models. In this context, an important first approximation for many systems of chemical interest may be obtained by treating the surface nonlinearity as arising from the composite of individual molecular contributions. The molecular response is typically assumed to be that of the isolated molecule, but in the summation for the surface nonlinear response, we take into account the orientational distribution appropriate for the surface or interface, as we discuss later. Local-field corrections may also be included [41, 42]. Such analyses may then draw on the large and well-developed literature concerning the second-order nonlinearity of molecules [43, 44]. If we are concerned with the response of the surface of a clean solid, we must typically adopt a different approach: one based on delocalized electrons. This is a challenging undertaking, as a proper treatment of the *linear* optical properties of surfaces of solids is already difficult [45]. Nonetheless, in recent years significant progress has been made in developing a fundamental theory of the nonlinear response of surfaces of both metals [46, 47,

48, 49, 50 and 51] and semiconductors [52, 53, 54 and 55].

(B) BULK CONTRIBUTION

For centrosymmetric media the spatially local contribution to the second-order nonlinear response vanishes, as we have previously argued, providing the interface specificity of the method. This spatially local contribution, which arises in the quantum mechanical picture from the electric-dipole terms, represents the dominant response of the medium. However, if we consider the problem of probing interfaces closely, we recognize that we are comparing the nonlinear signal originating from an interfacial region of monolayer thickness with that of the bulk media. In the bulk media, the signal can build up over a thickness on the scale of the optical wavelength, as dictated by absorption and phase-matching considerations. Thus, a bulk nonlinear polarization that is much weaker than that of the dipole-allowed contribution present at the interface may still prove to be significant because of the larger volume contributing to the emission. Let us examine this point in a somewhat more quantitative fashion.

The higher-order bulk contribution to the nonlinear response arises, as just mentioned, from a spatially non-local response in which the induced nonlinear polarization does not depend solely on the value of the fundamental electric field at the same point. To leading order, we may represent these non-local terms as being proportional to a nonlinear response incorporating a first spatial derivative of the fundamental electric field. Such terms correspond in the microscopic theory to the inclusion of electric-quadrupole and magnetic-dipole contributions. The form of these bulk contributions may be derived on the basis of symmetry considerations. As an example of a frequently encountered situation, we indicate here the non-local polarization for SHG in a cubic material excited by a plane wave $\mathbf{E}(\omega)$:

$$P_{b,i}^{(2)}(2\omega) = \gamma \nabla_i [\mathbf{E}(\omega) \cdot \mathbf{E}(\omega)] + \zeta E_i(\omega) \nabla_i E_i(\omega). \quad (\text{B1.5.37})$$

The two coefficients γ and ζ describe the material response and the Cartesian coordinate i must be chosen as a principal axis of the material.

-19-

From consideration of the quantum mechanical expression of such a non-local response, one may argue that the dipole-forbidden bulk nonlinear polarization will have a strength reduced from that of the dipole-allowed response by a factor of the order of (a/λ) , with a denoting a typical atomic dimension and λ representing the wavelength of light. On the other hand, the relevant volume for the bulk contribution typically exceeds that of the interface by a factor of the order of (λ/a) . Consequently, one estimates that the net bulk and surface contributions to the nonlinear radiation may be roughly comparable in strength. In practice, the interfacial contribution often dominates that of the bulk. Nonetheless, one should not neglect *a priori* the possible role of the bulk nonlinear response. This situation of a possible bulk background signal comparable to that of the interface should be contrasted to the expected behaviour for a conventional optical probe lacking interface specificity. In the latter case, the bulk contribution would be expected to dominate that of the interface by several orders of magnitude.

(C) OTHER SOURCES

As we have discussed earlier in the context of surfaces and interfaces, the breaking of the inversion symmetry strongly alters the SHG from a centrosymmetric medium. Surfaces and interfaces are not the only means of breaking the inversion symmetry of a centrosymmetric material. Another important perturbation is that induced by (static) electric fields. Such electric fields may be applied externally or may arise internally from a depletion layer at the interface of a semiconductor or from a double-charge layer at the interface of a liquid.

Since the electric field is a polar vector, it acts to break the inversion symmetry and gives rise to dipole-allowed sources of nonlinear polarization in the bulk of a centrosymmetric medium. Assuming that the DC field, E_{DC} , is sufficiently weak to be treated in a leading-order perturbation expansion, the response may be written as

$$P_{\text{DC}}^{(3)}(2\omega) = \chi^{(3)} : E(\omega)E(\omega)E_{\text{DC}} \quad (\text{B1.5.38})$$

where $\chi^{(3)}$ is the effective third-order response. This process is called electric-field-induced SHG or EFISH.

A different type of external perturbation is the application of a magnetic field. In contrast to the case of an electric field, an applied magnetic field does not lift the inversion symmetry of a centrosymmetric medium. Hence, it does not give rise to a dipole-allowed bulk polarization for SHG or SFG [23, 56]. A magnetic field can, however, modify the form and strength of the interfacial nonlinear response, as well as the bulk quadrupole nonlinear susceptibilities. This process is termed magnetization-induced SHG or MSHG. Experiments exploiting both EFISH and MSHG phenomena are discussed in [section B1.5.4.7](#).

B1.5.3 EXPERIMENTAL CONSIDERATIONS

In this section, we provide a brief overview of some experimental issues relevant in performing surface SHG and SFG measurements.

B1.5.3.1 EXPERIMENTAL GEOMETRY

The main panel of [figure B1.5.6](#) portrays a typical setup for SHG. A laser source of frequency ω is directed to the sample, with several optical stages typically being introduced for additional control and filtering. The combination of a

halfwave plate and a polarizer is used to specify the orientation of the polarization of the pump beam. It can also serve as a variable attenuator to adjust the intensity of the incoming beam. A lens then focuses the beam onto the sample. A low-pass filter is generally needed along the path of the fundamental radiation prior to the sample to remove any unwanted radiation at the frequency of the nonlinear radiation. This radiation may arise from previous optical components, including the laser source itself.

Figure B1.5.6 Experimental geometry for typical SHG and SFG measurements.

The reflected radiation consists of a strong beam at the fundamental frequency ω and a weak signal at the SH frequency. Consequently, a high-pass filter is introduced, which transmits the nonlinear radiation, but blocks the fundamental radiation. By inserting this filter immediately after the sample, we minimize the generation of other nonlinear optical signals from succeeding optical components by the fundamental light reflected from the sample. After this initial filtering stage, a lens typically recollimates the beam and an analyser is used to select the desired polarization. Although not always essential, a monochromator or bandpass filter is often desirable to ensure that only the SH signal is measured. Background signals near the SH frequency, but not associated with the SHG process, may arise from multiphoton fluorescence, hyper-Raman scattering and other nonlinear processes described by higher-order nonlinear susceptibilities. Detection is usually accomplished through a photomultiplier tube. Depending on the nature of the laser source, various sensitive schemes for

electronic detection of the photomultiplier may be employed, such as photon counting and gated integration. In conjunction with an optical chopper in the beam path, lock-in amplification techniques may also be advantageous.

One of the key factors for performing surface SHG/SFG measurements is to reduce all sources of background signals, since the desired nonlinear signal is always relatively weak. This goal is accomplished most effectively by exploiting the well-defined spectral, spatial and temporal characteristics of the nonlinear radiation. The first of these is achieved by spectral filtering, as previously discussed. The second may be achieved through the use of appropriate apertures; and the last, particularly for low-repetition rate systems, can be incorporated into the electronic detection scheme. When

-21-

the excitation is derived from two non-collinear beams, the nonlinear emission will generally travel in a distinct direction ([figure B1.5.6](#) inset). In this case, one can also exploit spatial filters to enhance spectral selectivity, since the reflected pump beams will travel in different directions. This property is particularly useful for SFG experiments in which the frequency of the visible beam and the SF signal are relatively similar.

For some experiments, it may be helpful to obtain a reference signal to correct for fluctuations and long-term drift in the pump laser. This correction is best accomplished by performing simultaneous measurements of the SHG or SFG from a medium that has a strong $\chi^{(2)}$ response in a separate detection arm. By this means, one may fully compensate for variations not only in pulse energy, but also in the temporal and spatial substructure of the laser pulses. Some experiments may require measurement of the phase of the nonlinear signal [57]. Such phase measurements rely on interference with radiation from a reference nonlinear source. The required interference can be achieved by placing a reference nonlinear crystal along the path of the laser beam immediately either before or after the sample [58]. For effective interference, we must control both the amplitude and polarization of the reference signal. This may be achieved by appropriate focusing conditions and crystal alignment. The phase of the reference signal must also be adjustable. Phase control may be obtained simply by translating the reference sample along the path of the laser, making use of the dispersion of the air (or other medium) through which the beams propagate.

B1.5.3.2 LASER SOURCES

In order to achieve a reasonable signal strength from the nonlinear response of approximately one atomic monolayer at an interface, a laser source with high peak power is generally required. Common sources include Q-switched (~ 10 ns pulsewidth) and mode-locked (~ 100 ps) Nd:YAG lasers, and mode-locked (~ 10 fs–1 ps) Ti:sapphire lasers. Broadly tunable sources have traditionally been based on dye lasers. More recently, optical parametric oscillator/amplifier (OPO/OPA) systems are coming into widespread use for tunable sources of both visible and infrared radiation.

In typical experiments, the laser fluence, or the energy per unit area, is limited to the sample's damage threshold. This generally lies in the range $\ll 1$ J cm⁻² and constrains our ability to increase signal strength by increasing the pump energy. Frequently, the use of femtosecond pulses is advantageous, as one may obtain a higher intensity (and, hence, higher nonlinear conversion efficiency) at lower fluence. In addition, such sources generally permit one to employ lower average intensity, which reduces average heating of the sample and other undesired effects [59]. Independently of these considerations, femtosecond lasers are, of course, also attractive for the possibilities that they offer for measurements of ultrafast dynamics.

B1.5.3.3 SIGNAL STRENGTHS

We now consider the signal strengths from surface SHG/SFG measurements. For this purpose, we may recast

expression [B1.5.32](#) for the reflected nonlinear radiation in terms of the number of emitted photon per unit time as

$$S = \frac{8\pi^3 \omega_3}{\hbar c^3 [\epsilon_1(\omega_3) \epsilon_1(\omega_2) \epsilon_1(\omega_1)]^{1/2}} \frac{\sec^2 \theta P_{\text{avg}}(\omega_1) P_{\text{avg}}(\omega_2)}{t_p R_{\text{rep}} A} |e'(\omega_3) \cdot \chi_s^{(2)} : e'(\omega_1) e'(\omega_2)|^2. \quad (\text{B1.5.39})$$

The quantities in this formula are defined as in [equation B1.5.32](#), but with the laser parameters translated into more convenient terms: P_{avg} is the average power at the indicated frequency; t_p is the laser pulse duration; R_{rep} is the pulse

-22-

repetition rate; and A is the irradiated area at the interface. The last three defined quantities are assumed to be equal for both excitation beams in an SFG measurement. If this is not the case, then t_p , R_{rep} , A , as well as the average power $P_{\text{avg}}(\omega_i)$, have to be replaced by the corresponding quantities within the window of spatial and temporal overlap.

From this expression, we may estimate typical signals for a surface SHG measurement. We assume the following as representative parameters: $\chi^{(2)} = 10^{-15}$ esu, $\epsilon_1 = 1$, and $\sec^2 \theta = 4$. For typical optical frequencies, we then obtain $S \approx 10^{-2} (P_{\text{avg}})^2 / (t_p R_{\text{rep}} A)$, where P_{avg} , A , R_{rep} and t_p are expressed, respectively, in W, cm^2 , Hz and s. Many recent SHG studies have been performed with mode-locked Ti:sapphire lasers. For typical laser parameters of $P_{\text{avg}} = 100$ mW, $A = 10^{-4}$ cm^2 , $R_{\text{rep}} = 100$ MHz, and $t_p = 100$ fs, one then obtains $S \approx 10^5$ counts per second as a representative nonlinear signal.

B1.5.3.4 DETERMINATION OF NONLINEAR SUSCEPTIBILITY ELEMENTS

The basic physical quantities that define the material for SHG or SFG processes are the nonlinear susceptibility elements $\chi_{s,ijk}^{(2)}$. Here we consider how one may determine these quantities experimentally. For simplicity, we treat the case of SHG and assume that the surface is isotropic. From symmetry considerations, we know that $\chi_{s,ijk}^{(2)}$ has three independent and non-vanishing elements: $\chi_{s,\perp\perp\perp}^{(2)}$, $\chi_{s,\perp\parallel\parallel}^{(2)}$, and $\chi_{s,\parallel\perp\parallel}^{(2)} = \chi_{s,\parallel\parallel\perp}^{(2)}$. The individual elements $\chi_{s,\parallel\perp\perp}^{(2)}$ and $\chi_{s,\parallel\perp\parallel}^{(2)}$ can be extracted directly by an appropriate choice of input and output polarizations. Response from the $\chi_{s,\perp\parallel\parallel}^{(2)}$ element requires s-polarized pump radiation and produces p-polarized SH emission; excitation of the $\chi_{s,\perp\perp\perp}^{(2)}$ element requires a mixed-polarized pump radiation and can be isolated by detection of s-polarized radiation. The measurement of $\chi_{s,\perp\perp\perp}^{(2)}$ is bit more complicated: to isolate it would require a pump electric field aligned normal to the surface, thus implying a pump beam travelling parallel to the surface (the limit of grazing incidence).

An alternative scheme for extracting all three isotropic nonlinear susceptibilities can be formulated by examining [equation B1.5.39](#). By choosing an appropriate configuration and the orientation of the polarization of the SH radiation $e'(2\omega)$ such that the SHG signal vanishes, one obtains, assuming only surface contribution with real elements $\chi_{s,ijk}^{(2)}$,

$$e'(2\omega) \cdot \chi_s^{(2)} : e'(\omega) e'(\omega) = \sum_{ijk} e'_i(2\omega) \chi_{s,ijk}^{(2)} e'_j(\omega) e'_k(\omega) = 0.$$

Expanding this equation, we deduce that

$$e'_\perp(2\omega)[\chi_{s,\perp\perp\perp}^{(2)}e'_\perp(\omega)e'_\perp(\omega) + \chi_{s,\perp\parallel\parallel}^{(2)}e'_\parallel(\omega)e'_\parallel(\omega)] + 2e'_\parallel(2\omega)\chi_{s,\perp\perp\parallel}^{(2)}e'_\perp(\omega)e'_\parallel(\omega) = 0. \quad (\text{B1.5.40})$$

The magnitudes of e'_i ($i = \perp, \parallel$) contain the Fresnel factors from [equation B1.5.34](#), [equation B1.5.35](#) and [equation B1.5.36](#), which depend on the incident, reflected and polarization angles. Experimentally, one approach is to fix the input polarization and adjust the analyser to obtain a null in the SH signal [60]. By choosing distinct configurations such that the corresponding three equations from [equation B1.5.40](#) are linearly independent, the relative values of $\chi_{s,\perp\perp\perp}^{(2)}$, $\chi_{s,\perp\parallel\parallel}^{(2)}$, and $\chi_{s,\perp\perp\parallel}^{(2)} = \chi_{s,\parallel\parallel\perp}^{(2)}$ may be inferred. This method has been implemented, for example, in determining the three susceptibility elements for the air/water interface [61]. The procedure just described is suitable for

-23-

ascertaining the relative magnitudes of the allowed elements of $\chi_s^{(2)}$.

A determination of the absolute magnitude of the elements of $\chi_s^{(2)}$ may also be of value. In principle, this might be accomplished by careful measurements of signal strengths from [equation \(B1.5.32\)](#) or [equation \(B1.5.39\)](#). In practice, such an approach is very difficult, as it would require precise calibration of the parameters of the laser radiation and absolute detection sensitivity. The preferred method is consequently to compare the surface SHG or SFG response to that of a reference crystal with a known bulk nonlinearity inserted in place of the sample. Since the expected signal can be calculated for this reference material, we may then infer the absolute calibration of $\chi_s^{(2)}$ by comparison. It should be further noted that the phase of elements of $\chi_s^{(2)}$ may also be established by interference measurements as indicated in [section B1.5.3.1](#).

B1.5.4 APPLICATIONS

The discussion of applications of the SHG and SFG methods in this section is directed towards an exposition of the different types of information that may be obtained from such measurements. The topics have been arranged accordingly into seven general categories that arise chiefly from the properties of the nonlinear susceptibility: surface symmetry and order, adsorbate coverage, molecular orientation, spectroscopy, dynamics, spatial resolution and perturbations induced by electric and magnetic fields. Although we have included some illustrative examples, a comprehensive description of the broad range of materials probed by these methods, and what has been learned about them, is clearly beyond the scope of this chapter.

B1.5.4.1 SURFACE SYMMETRY AND ORDER

Spatial symmetry is one of the basic properties of a surface or interface. If the symmetry of the surface is known *a priori*, then this knowledge may be used to simplify the form of the surface nonlinear susceptibility $\chi_s^{(2)}$, as discussed in [section B1.5.2.2](#). Conversely, in the absence of knowledge of the surface symmetry, we may characterize the form of $\chi_s^{(2)}$ experimentally and then make inferences about the symmetry of the surface or interface. This provides some useful information about the material system under study, as we shall illustrate in this section. Before doing so, we should remark that the spatial properties being probed are averaged over a length scale that depends on the precise experimental geometry, but exceeds the scale of an optical wavelength. In the following paragraphs, we consider two of the interesting cases of surface symmetry: that corresponding to the surface of a crystalline material and that of a surface with chiral character.

(A) CRYSTALLINE SURFACES

All of the symmetry classes compatible with the long-range periodic arrangement of atoms comprising crystalline surfaces and interfaces have been enumerated in [table B1.5.1](#). For each of these symmetries, we indicate the corresponding form of the surface nonlinear susceptibility $\chi_s^{(2)}$. With the exception of surfaces with four-fold or six-fold rotational symmetries, all of these symmetry classes give rise to a $\chi_s^{(2)}$ that may be distinguished from that of an isotropic surface with mirror symmetry, the highest possible surface symmetry.

-24-

An experimental analysis of the surface symmetry may be carried out in various ways. For a fixed crystal orientation, the surface symmetry may be probed by modifying the angle of incidence and polarization of the input and output beams. This approach is often employed for samples, such as those in ultrahigh vacuum, that are difficult to manipulate. An attractive alternative method is to probe the rotational anisotropy simply by recording the change in the nonlinear signal as the sample is rotated about its surface normal. The resulting data reflect directly the symmetry of the surface or interface. Thus, the rotational pattern for the (111) surface of a cubic centrosymmetric crystal with $3m$ symmetry will also have three-fold symmetry with mirror planes. A surface with four-fold symmetry, as in the case of the (001) surface of a cubic material, will give rise to a rotational anisotropy pattern that obeys four-fold symmetry. From [table B1.5.1](#), we see that it will, however, do so in a trivial fashion by giving a response equivalent to that of an isotropic material, i.e. lacking any variation with rotation of the crystal surface.

As an illustration, we consider the case of SHG from the (111) surface of a cubic material ($3m$ symmetry). More general treatments of rotational anisotropy in centrosymmetric crystals may be found in the literature [[62](#), [63](#) and [64](#)]. For the case at hand, we may determine the anisotropy of the radiated SH field from [equation B1.5.32](#) in conjunction with the form of $\chi_s^{(2)}$ from [table B1.5.1](#). We find, for example, for the p-in/p-out and s-in/s-out polarization configurations:

$$E_{\text{p-in/p-out}}^{(2\omega)} = a_0 + a_3 \cos(3\psi) \quad (\text{B1.5.41})$$

$$E_{\text{s-in/s-out}}^{(2\omega)} = b_3 \sin(3\psi) \quad (\text{B1.5.42})$$

where a_0 , a_3 and b_3 are constants. The angle ψ corresponds to the rotation of the sample about its surface normal and is measured between the plane of incidence and the $[11\bar{2}]$ direction in the surface plane.

[Figure B1.5.7](#) displays results of a measurement of the rotational anisotropy for an oxidized Si(111) surface [[65](#)]. For the case shown in the top panel, the results conform to the predictions of [equation B1.5.42](#) (with $I(2\omega) \propto |E^{(2\omega)}|^2$) for ideal $3m$ symmetry. The data clearly illustrate the strong influence of anisotropy on measured nonlinear signals. The lower panels of [figure B1.5.7](#) are perturbed rotational anisotropy patterns. They correspond to data for vicinal Si(111) surfaces cut at 3° and 5° , respectively, away from the true (111) orientation. The full lines fitting these data are obtained from an analysis in which the lowered symmetry of these surfaces is taken into account. The results show the sensitivity of this method to slight changes in the surface symmetry.

-25-

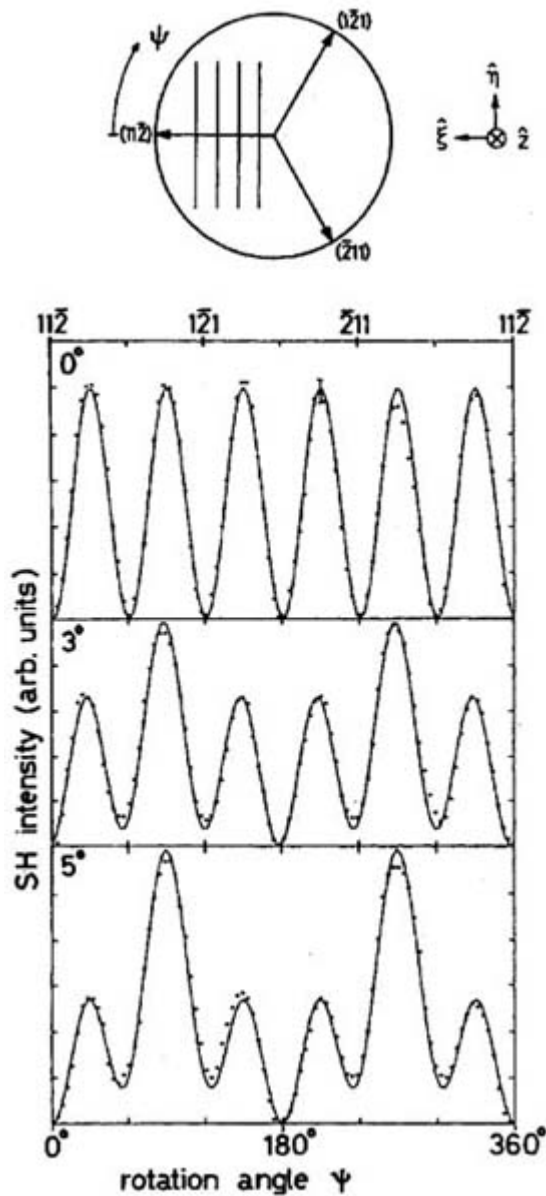


Figure B1.5.7 Rotational anisotropy of the SH intensity from oxidized Si(111) surfaces. The samples have either ideal orientation or small offset angles of 3° and 5° toward the $[11\bar{2}]$ direction. Top panel illustrates the step structure. The points correspond to experimental data and the full lines to the prediction of a symmetry analysis. (From [65].)

SH anisotropy measurements of this kind are of use in establishing the orientation of a crystal face of a material, as suggested by figure B1.5.7. The method is also of value for monitoring and study of crystal growth processes [66]. Consider, for example, the growth of Si on Si(111) surface. The crystalline surface exhibits strong rotational anisotropy, corresponding to the $3m$ symmetry of the surface. This will also be the case when a crystalline Si layer is grown on the sample. If, however, the overlayer is grown in an amorphous state, as would occur for Si deposition at

room temperature, then the anisotropy will be reduced: the disordered overlayer will exhibit isotropic symmetry on the length scale of the optical wavelength. A further application of rotational anisotropy measurements has been found in the characterization of surface roughness [67].

(B) CHIRAL INTERFACES

An important distinction among surfaces and interfaces is whether or not they exhibit mirror symmetry about a plane normal to the surface. This symmetry is particularly relevant for the case of isotropic surfaces (∞ -symmetry), i.e. ones that are equivalent in every azimuthal direction. Those surfaces that fail to exhibit mirror symmetry may be termed chiral surfaces. They would be expected, for example, at the boundary of a liquid comprised of chiral molecules. Magnetized surfaces of isotropic media may also exhibit this symmetry. (For a review of SHG studies of chiral interfaces, the reader is referred to [68].)

Given the interest and importance of chiral molecules, there has been considerable activity in investigating the corresponding chiral surfaces [68, 69 and 70]. From the point of view of performing surface and interface spectroscopy with nonlinear optics, we must first examine the nonlinear response of the bulk liquid. Clearly, a chiral liquid lacks inversion symmetry. As such, it may be expected to have a strong (dipole-allowed) second-order nonlinear response. This is indeed true in the general case of SFG [71]. For SHG, however, the permutation symmetry for the last two indices of the nonlinear susceptibility tensor combined with the requirement of isotropic symmetry in three dimensions implies that $\chi^{(2)} = 0$. Thus, for the case of SHG, the surface/interface specificity of the technique is preserved even for chiral liquids.

A schematic diagram of the surface of a liquid of non-chiral (a) and chiral molecules (b) is shown in [figure B1.5.8](#). Case (a) corresponds to ∞m -symmetry (isotropic with a mirror plane) and case (b) to ∞ -symmetry (isotropic). For the ∞m -symmetry, the SH signal for the polarization configurations of s-in/s-out and p-in/s-out vanish. From [table B1.5.1](#), we find, however, that for the ∞ -symmetry, an extra independent nonlinear susceptibility element, $\chi_{s,xyz}^{(2)} = -\chi_{s,yxz}^{(2)}$, is present for SHG. Because of this extra element, the SH signal for p-in/s-out configuration is no longer forbidden, and consequently, the SH polarization must no longer be strictly p-polarized. [figure B1.5.8\(c\)](#) shows the SH signal passing through an analyser as a function of its orientation for a racemic mixture (squares) and for a non-racemic mixture (circle) of molecules in a Langmuir–Blodgett film [70]. For the racemic mixture (squares), which contains equal amounts of both enantiomers, the effective symmetry is ∞m . Hence, the p-in/s-out signal vanishes and the response curve of [figure B1.5.8](#) is centred at 90° . For the case of the non-racemic mixture, the effective symmetry is ∞ . A p-in/s-out SH signal is present and leads, as shown in [figure B1.5.8\(c\)](#), to a displacement in the curve of the SH response versus analyser setting.

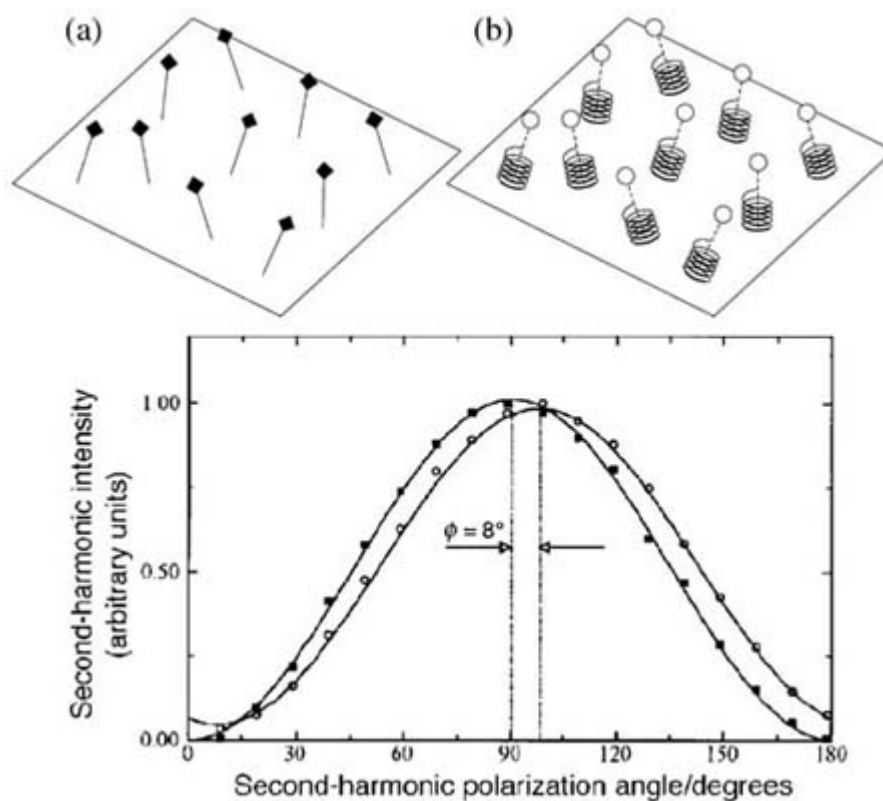


Figure B1.5.8 Random distribution of (a) non-chiral adsorbates that gives rise to a surface having effective ∞ m-symmetry; (b) chiral molecules that gives rise to effective ∞ -symmetry. (c) SH intensity versus the angle of an analyser for a racemic (squares) and a non-racemic (open circles) monolayer of chiral molecules. The pump beam was p-polarized; the SH polarization angles of 0° and 90° correspond to s- and p-polarization, respectively. (From [70].)

This effect of a change in the SH output polarization depending on the enantiomer or mixture of enantiomer is somewhat analogous to the linear optical phenomenon of optical rotary dispersion (ORD) in bulk chiral liquids. As such, the process for SH radiation is termed SHG-ORD [70]. In general, chiral surfaces will also exhibit distinct radiation characteristics for left- and right-polarized pump beams. Again, by analogy with the linear optical process of circular dichroism (CD), this effect has been termed SHG-CD [69].

B1.5.4.2 ADSORBATE COVERAGE

A quantity of interest in many studies of surfaces and interfaces is the concentration of adsorbed atomic or molecular species. The SHG/SFG technique has been found to be a useful probe of adsorbate density for a wide range of interfaces. The surface sensitivity afforded by the method is illustrated by the results of [figure B1.5.9](#) [72]. These data show the dramatic change in SH response from a clean surface of silicon upon adsorption of a fraction of a monolayer of atomic hydrogen.

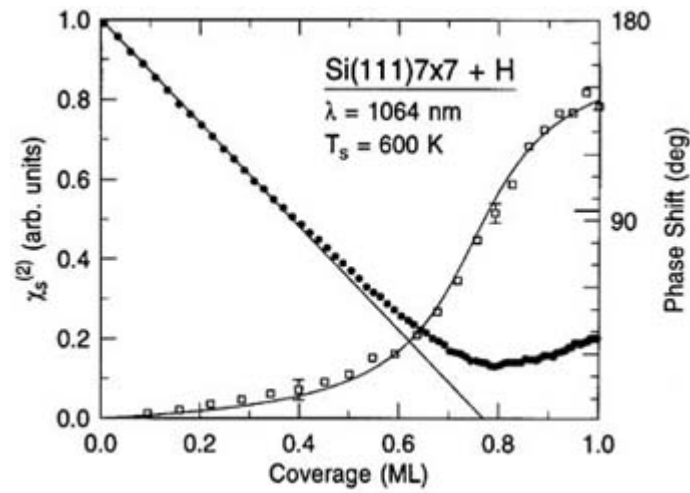


Figure B1.5.9 Dependence of the magnitude (full circles) and the phase (open squares) of the nonlinear susceptibility $\chi_s^{(2)}$ of Si(111) 7×7 on the coverage of adsorbed atomic hydrogen for an excitation wavelength of 1064 nm. (From [72].)

We now consider how one extracts quantitative information about the surface or interface adsorbate coverage from such SHG data. In many circumstances, it is possible to adopt a purely phenomenological approach: one calibrates the nonlinear response as a function of surface coverage in a preliminary set of experiments and then makes use of this calibration in subsequent investigations. Such an approach may, for example, be appropriate for studies of adsorption kinetics where the interest lies in the temporal evolution of the surface adsorbate density N_s .

For other purposes, obtaining a measure of the adsorbate surface density directly from the experiment is desirable. From this perspective, we introduce a simple model for the variation of the surface nonlinear susceptibility with adsorbate coverage. An approximation that has been found suitable for many systems is

$$\chi_s^{(2)}(N_s) = \chi_{s,0}^{(2)} + N_s \alpha^{(2)}. \quad (\text{B1.5.43})$$

From a purely phenomenological perspective, this relationship describes a constant rate of change in the nonlinear susceptibility of the surface with increasing adsorbate surface density N_s . Within a picture of adsorbed molecules, $\alpha^{(2)}$ may be interpreted as the nonlinear polarizability of the adsorbed species. The quantity $\chi_{s,0}^{(2)}$ represents the nonlinear response in the absence of the adsorbed species.

If we consider the optical response of a molecular monolayer of increasing surface density, the form of equation B1.5.43 is justified in the limit of relatively low density where local-field interactions between the adsorbed species may be neglected. It is difficult to produce any rule for the range of validity of this approximation, as it depends strongly on the system under study, as well as on the desired level of accuracy for the measurement. The relevant corrections, which may be viewed as analogous to the Clausius–Mossotti corrections in linear optics, have been the

subject of some discussion in the literature [41, 42]. In addition to the local-field effects, the simple proportionality of variation in $\chi_s^{(2)}(N_s)$ with N_s frequently breaks down for reasons related to the physical and chemical nature of the surface or interface. In particular, inhomogeneous surfaces in which differing binding sites fill in different proportions may give rise to a variation in $\chi_s^{(2)}(N_s)$ that would, to leading order, vary with

relative populations of the different sites. Also, inter-adsorbate interactions that lead to shifts in energy levels or molecular orientation, as will be discussed later, will influence the nonlinear response in a manner beyond that captured in [equation B1.5.43](#).

Despite these caveats in the application of [equation B1.5.43](#), one finds that it provides reasonable accuracy in many experimental situations. The SH response for the H/Si system of [figure B1.5.9](#) for example, is seen to obey the simple linear variation of $\chi_s^{(2)}(N_s)$ with N_s of [equation B1.5.43](#) rather well up to an adsorbate coverage of about 0.5 monolayers. These data are also interesting because they show how destructive interference between the terms $\chi_{s,0}^{(2)}$ and $N_s\alpha^{(2)}$ can cause the SH signal to decrease with increasing N_s . In this particular example, the physical interpretation of this effect is based on the strong nonlinear response of the bare surface from the Si dangling bonds. With increasing hydrogen coverage, the concentration of dangling bonds is reduced and the surface nonlinearity decreases. For a system where $\chi_{s,0}^{(2)}$ is relatively small and the nonlinear response of the adsorbed species is significant, just the opposite trend for the variation of $\chi_s^{(2)}(N_s)$ with N_s would occur.

The applications of this simple measure of surface adsorbate coverage have been quite widespread and diverse. It has been possible, for example, to measure adsorption isotherms in many systems. From these measurements, one may obtain important information such as the adsorption free energy, $\Delta G^\circ = -RT\ln(K_{eq})$ [21]. One can also monitor the kinetics of adsorption and desorption to obtain rates. In conjunction with temperature-dependent data, one may further infer activation energies and pre-exponential factors [73, 74]. Knowledge of such kinetic parameters is useful for technological applications, such as semiconductor growth and synthesis of chemical compounds [75]. Second-order nonlinear optics may also play a role in the investigation of physical kinetics, such as the rates and mechanisms of transport processes across interfaces [76].

Before leaving this topic, we would like to touch on two related points. The first concerns the possibility of an absolute determination of the surface adsorbate density. [Equation B1.5.43](#) would suggest that one might use knowledge, either experimental or theoretical, of $\alpha^{(2)}$ and an experimental determination of $\chi_s^{(2)}(N_s)$ and $\chi_{s,0}^{(2)}$ to infer N_s in absolute terms. In practice, this is problematic. One experimental issue is that a correct measurement of $\chi_s^{(2)}$ in absolute terms may be difficult. However, through appropriate comparison with the response of a calibrated nonlinear reference material, we may usually accomplish this task. More problematic is obtaining knowledge of $\alpha^{(2)}$ for the adsorbed species. The determination of $\alpha^{(2)}$ in the gas or liquid phase is already difficult. In addition, the perturbation induced by the surface or interface is typically significant. Moreover, as discussed later, molecular orientation is a critical factor in determining the surface nonlinear response. For these reasons, absolute surface densities can generally be found from surface SHG/SFH measurements only if we can calibrate the surface nonlinear response at two or more coverages determined by other means. This situation, it should be noted, is not dissimilar to that encountered for many other common surface probes.

The second issue concerns molecular specificity. For a simple measurement of SHG at an arbitrary laser frequency, one cannot expect to extract information of the behaviour of a system with several possible adsorbed species. To make the technique appropriate for such cases, one needs to rely on spectroscopic information. In the simplest implementation, one chooses a frequency for which the nonlinear response of the species of interest is large or dominant. As will

be discussed in [section B1.5.4.4](#), this capability is significantly enhanced with SFG and the selection of a frequency corresponding to a vibrational resonance.

B1.5.4.3 MOLECULAR ORIENTATION

The nonlinear response of an individual molecule depends on the orientation of the molecule with respect to the polarization of the applied and detected electric fields. The same situation prevails for an ensemble of molecules at an interface. It follows that we may garner information about molecular orientation at surfaces and interfaces by appropriate measurements of the polarization dependence of the nonlinear response, taken together with a model for the nonlinear response of the relevant molecule in a standard orientation.

We now consider this issue in a more rigorous fashion. The inference of molecular orientation can be explained most readily from the following relation between the surface nonlinear susceptibility tensor and the molecular nonlinear polarizability $\alpha^{(2)}$:

$$(B1.5.44)$$

Here the ijk coordinate system represents the laboratory reference frame; the primed coordinate system $i'j'k'$ corresponds to coordinates in the molecular system. The quantities $T_{ii'}$ are the matrices describing the coordinate transformation between the molecular and laboratory systems. In this relationship, we have neglected local-field effects and expressed the in a form equivalent to summing the molecular response over all the molecules in a unit surface area (with surface density N_s). (For simplicity, we have omitted any contribution to not attributable to the dipolar response of the molecules. In many cases, however, it is important to measure and account for the background nonlinear response not arising from the dipolar contributions from the molecules of interest.) In equation B1.5.44, we allow for a distribution of molecular orientations and have denoted by $\langle \rangle$ the corresponding ensemble average:

$$(B1.5.45)$$

Here $f(\theta, \varphi, \psi)$ is the probability distribution of finding a molecule oriented at (θ, φ, ψ) within an element $d\Omega$ of solid angle with the molecular orientation defined in terms of the usual Euler angles ([figure B1.5.10](#)).

-31-

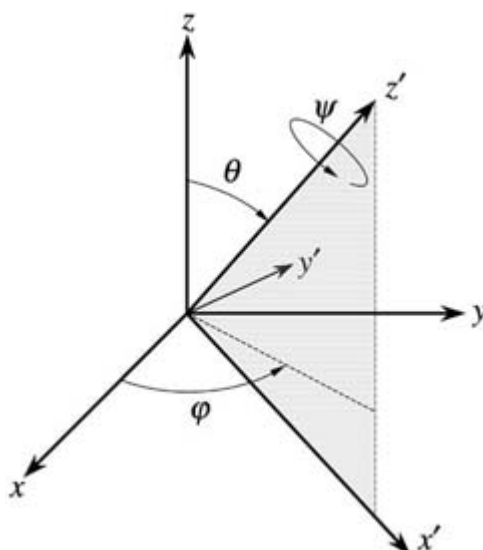


Figure B1.5.10 Euler angles and reference frames for the discussion of molecular orientation: laboratory frame (x, y, z) and molecular frame (x', y', z') .

Equation B1.5.44 indicates that if we know $\chi_{s,ijk}^{(2)}$ and $\alpha_{i'j'k'}^{(2)}$, we may infer information about the third-order orientational moments $\langle T_{ii'} T_{jj'} T_{kk'} \rangle$. Since calibration of absolute magnitudes is difficult, we are generally concerned with a comparison of the relative magnitudes of the appropriate molecular ($\alpha^{(2)}$) and macroscopic ($\chi^{(2)}$)

$\chi_s^{(2)}$) quantities. In practice, the complexity of the general relationship between $\alpha^{(2)}$ and $\chi_s^{(2)}$ means that progress requires the introduction of certain simplifying assumptions. These usually follow from symmetry considerations or, for the case of $\alpha^{(2)}$, from previous experimental or theoretical insight into the nature of the expected molecular response.

The approach may be illustrated for molecules with a nonlinear polarizability $\alpha^{(2)}$ dominated by a single axial component $\alpha_{z'z'z'}^{(2)}$, corresponding to a dominant nonlinear response from transitions along a particular molecular axis. Let us further assume that all in-plane direction of the surface or interface are equivalent. This would naturally be the case for a liquid or amorphous solid, but would not necessarily apply to the surface of a crystal. One then obtains from [equation B1.5.44](#), the following relations between the molecular quantities and surface nonlinear susceptibility:

$$\chi_{s,\perp\perp\perp}^{(2)} = \chi_{s,zzz}^{(2)} = N_s \langle \cos^3 \theta \rangle \alpha_{z'z'z'}^{(2)} \quad (\text{B1.5.46})$$

$$\chi_{s,\perp\parallel\parallel}^{(2)} = \chi_{s,zzx}^{(2)} = \frac{1}{2} N_s \langle \cos^3 \theta \sin^2 \theta \rangle \alpha_{z'z'z'}^{(2)} \quad (\text{B1.5.47})$$

$$\chi_{s,\parallel\perp\parallel}^{(2)} = \chi_{s,\parallel\parallel\perp}^{(2)} = \chi_{s,xzx}^{(2)} = \frac{1}{2} N_s \langle \cos^3 \theta \sin^2 \theta \rangle \alpha_{z'z'z'}^{(2)}. \quad (\text{B1.5.48})$$

-32-

Notice that $\chi_{s,\perp\parallel\parallel}^{(2)} = \chi_{s,\parallel\perp\parallel}^{(2)} = \chi_{s,\parallel\parallel\perp}^{(2)}$, so that only two of the three nonlinear susceptibility tensor elements allowed for an isotropic surface are independent. From [equation B1.5.46](#), [equation B1.5.47](#) and [equation B1.5.48](#), we may form the ratio

$$\frac{2\chi_{s,\parallel\perp\parallel}^{(2)} + \chi_{s,\perp\perp\perp}^{(2)}}{\chi_{s,\perp\perp\perp}^{(2)}} = \frac{\langle \cos \theta \rangle}{\langle \cos^3 \theta \rangle}. \quad (\text{B1.5.49})$$

Thus, a well-defined measure of molecular orientation is inferred from the measurement of the macroscopic quantities $\chi_{s,ijk}^{(2)}$. For the case of a narrow and isotropic distribution, i.e. $f(\theta) = \delta(\theta - \theta_0)$, the left-hand side term of [equation B1.5.49](#) becomes $\langle \cos \theta \rangle / \langle \cos^3 \theta \rangle = \sec^2 \theta_0$, for which the mean orientation θ_0 is directly obtained. For a broad distribution, one may extract the mean orientation from such an expression for an assumed functional form.

As an example, the model described earlier for a molecule having a dominant nonlinear polarizability element $\alpha_{z'z'z'}^{(2)}$ has been applied to the determination of the molecular inclination θ between the molecular axis of a surfactant molecule, sodium-dodecylsulfonate (SDNS) and the surface normal at the air/water interface [77]. This tilt angle θ , shown in [figure B1.5.11](#) was determined according to [equation B1.5.46](#), [equation B1.5.47](#) and [equation B1.5.48](#) under the assumption of a narrow orientational distribution. As the [figure](#) shows, the mean molecular orientation changes with increasing surface pressure π as the molecules are forced into a more nearly vertical orientation.

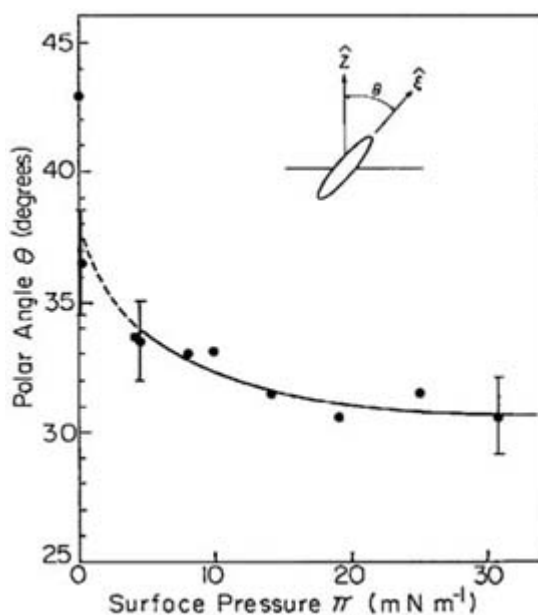


Figure B1.5.11 Tilt angle θ between the molecular axis of sodium-dodecylnaphthalene-sulphonate (SDNS) and the surface normal as a function of the surface pressure π at the air/water interface. (From [77].)

-33-

In the literature, the interested reader may find treatment of molecules with differing, and more complex, behaviour for $\alpha^{(2)}$ [14]. Also of importance is the role of SFG for orientational analysis. Surface SFG may provide more orientational information, since there are additional independent elements of the surface nonlinear susceptibility. For example, an isotropic surface is characterized by three independent elements for SHG, but is described by four independent elements for SFG. The principal advantage of SFG over SHG, however, lies in its molecular specificity. As will be discussed in section B1.5.4.4, we may enhance the response of a molecular species of interest by choosing the appropriate infrared frequency for the SFG process. This behaviour helps to eliminate background signal and is useful in more complex systems with two or more molecular species present. Equally important, the excitation of a given vibration helps to define the form, and reduce the generality, of the molecular response $\alpha^{(2)}$. Further, the method may be applied for different vibrational resonances to deduce the orientation of different moieties of larger molecules [78].

B1.5.4.4 SPECTROSCOPY

The second-order nonlinear susceptibility describing a surface or interface, as indicated by the microscopic form of [equation B1.5.30](#), is resonantly enhanced whenever an input or output photon energy matches a transition energy in the material system. Thus, by scanning the frequency or frequencies involved in the surface nonlinear process, we may perform surface-specific spectroscopy. This method has been successfully applied to probe both electronic transitions and vibrational transitions at interfaces.

For studying electronic transitions at surfaces and interfaces, both SHG and SFG have been employed in a variety of systems. One particular example is that of the buried $\text{CaF}_2/\text{Si}(111)$ interface [79]. [Figure B1.5.12\(a\)](#) displays the experimental SH signal as a function of the photon energy of a tunable pump laser. An interface resonance for this system is found to occur for a photon energy near 2.4 eV. This value is markedly different from that of the energies of transitions in either of the bulk materials and clearly illustrates the capability of nonlinear spectroscopy to probe distinct electronic excitations of the interfacial region. The sharp feature appearing at 2.26 eV has been attributed to the formation of a two-dimensional exciton. It is important to point out that the measurement of the SHG signal alone does not directly show whether an observed resonance corresponds to a single- or a two-photon transition. To verify that the resonance enhancement does, in fact, correspond to a transition energy of 2.4 eV, a separate SF measurement ([Figure B1.5.12\(b\)](#)) was

performed. In this measurement, the tunable laser photon was mixed with another photon at a fixed photon energy (1.17 eV). By comparing the two sets of data, one finds that the resonance must indeed lie at the fundamental frequency of the tunable laser for this system.

-34-

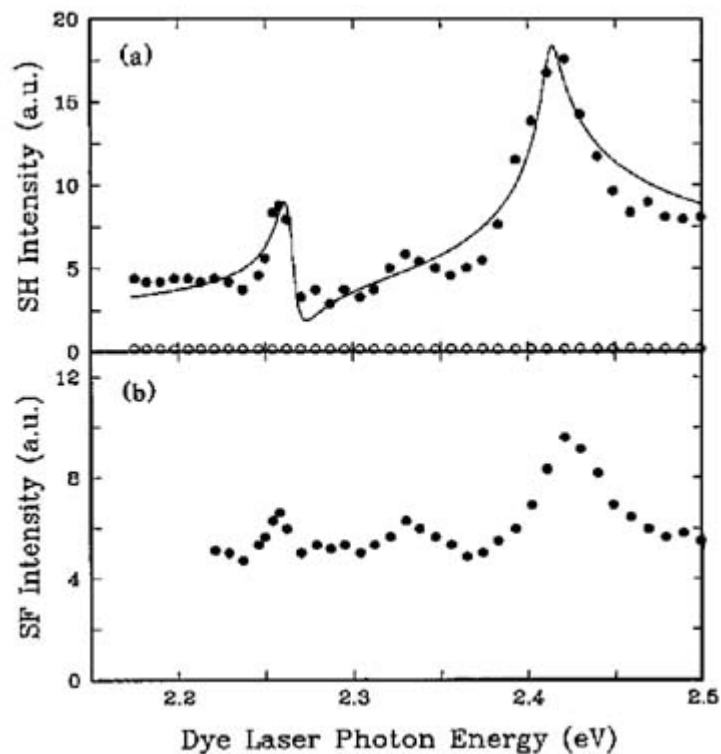


Figure B1.5.12 SH and SF spectra (full dots) for the $\text{CaF}_2/\text{Si}(111)$ interface: (a) SH intensity as a function of the photon energy of the tunable laser; (b) SF intensity obtained by mixing the tunable laser with radiation at a fixed photon energy of 1.17 eV. For comparison, the open circles in (a) are signals obtained for a native-oxide covered Si(111). The full line is a fit to the theory as discussed in [79].

The SHG/SFG technique is not restricted to interface spectroscopy of the delocalized electronic states of solids. It is also a powerful tool for spectroscopy of electronic transitions in molecules. [Figure B1.5.13](#) presents such an example for a monolayer of the R-enantiomer of the molecule 2,2'-dihydroxyl-1,1'-binaphthyl, (R)-BN, at the air/water interface [80]. The spectra reveal two-photon resonance features near wavelengths of 332 and 340 nm that are assigned to the two lowest exciton-split transitions in the naphth-2-ol monomer of BN. An increase in signal at higher photon energies is also seen as a resonance as the 1B_p state of the molecules is approached. The spectra in [figure B1.5.13](#) have been obtained for differing polarization configurations. The arrangements of p-in/p-out and s-in/p-out will yield SH signals for any isotropic surface. In this case, however, signal is also observed for the p-in/s-out configuration. This response arises from the $\chi_{s,xy}^{(2)}$ element of the surface nonlinear response that is present because of the chiral character of the molecules under study, as previously discussed in [section B1.5.4.1](#).

-35-

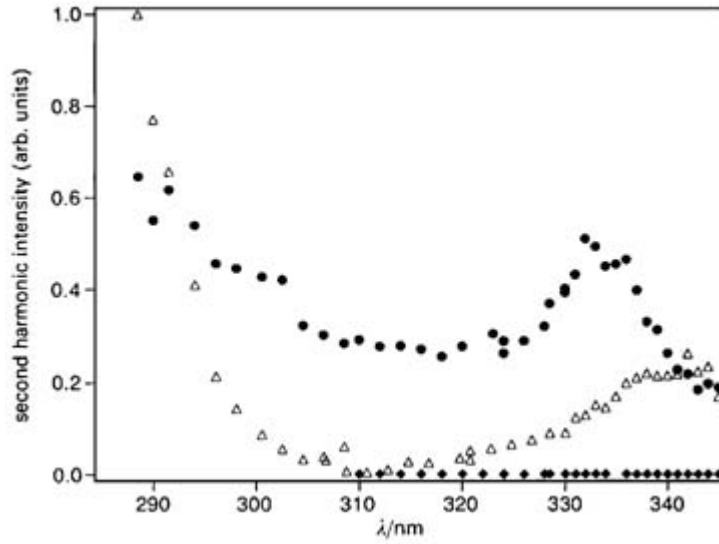


Figure B1.5.13 Spectra of the various non-chiral [p-in/p-out (filled circles) and s-in/p-out (filled diamonds)] and chiral [p-in/s-out (triangle)] SHG signals of (R)-BN molecules adsorbed at the air/water interface. (From [80].)

In addition to probing electronic transitions, second-order nonlinear optics can be used to probe vibrational resonances. This capability is of obvious importance and value for identifying chemical species at interfaces and probing their local environment. In contrast to conventional spectroscopy of vibrational transitions, which can also be applied to surface problems [4], nonlinear optics provides intrinsic surface specificity and is of particular utility in problems where the same or similar vibrational transitions occur at the interface as in the bulk media. In order to access the infrared region corresponding to vibrational transitions while maintaining an easily detectable signal, Shen and coworkers developed the technique of the infrared-visible sum-frequency generation [81]. In this scheme, a tunable IR source is mixed on the surface with visible light at a fixed frequency to produce readily detectable visible radiation. As the IR frequency is tuned through the frequency of a vibrational transition, the SF signal is resonantly enhanced and the surface vibration spectrum is recorded. In order to examine the IR-visible SFG process more closely, let us consider an appropriate formula for the surface nonlinear susceptibility when the IR frequency $\omega_1 = \omega_{\text{IR}}$ is near a single vibrational resonance and the visible frequency $\omega_2 = \omega_{\text{vis}}$ is not resonant with an electronic transition. We may then write [81, 82]

$$\chi_{s,ijk}^{(2)}(\omega_{\text{IR}}) = \chi_{s,ijk}^{(2)\text{NR}} + \chi_{s,ijk}^{(2)\text{R}}(\omega_{\text{IR}}) = \chi_{s,ijk}^{(2)\text{NR}} + \sum_l \frac{A_{l,ijk}}{\omega_{\text{IR}} - \omega_l + i\Gamma_l} \quad (\text{B1.5.50})$$

where $\chi_{s,ijk}^{(2)\text{NR}}$ and $\chi_{s,ijk}^{(2)\text{R}}$ are the non-resonant and resonant contributions to the signal, respectively; $A_{l,ijk}$, ω_l and Γ_l are the strength, resonant frequency, linewidth of the l th vibrational mode. The quantity A_l is proportional to the product of the first derivatives of the molecular dipole moment μ_i and of the electronic polarizability α_{jk} with respect to the l th normal coordinate Q_l :

$$A_{l,ijk} \propto \frac{\partial \mu_i}{\partial Q_l} \frac{\partial \alpha_{jk}}{\partial Q_l}. \quad (\text{B1.5.51})$$

Consequently, in order for a vibrational mode to be observed in infrared-visible SFG, the molecule in its adsorbed state has to be both IR [$(d\mu_i/dQ_l) \neq 0$] and Raman [$(d\alpha_{jk}/dQ_l) \neq 0$] active.

The form of [equation B1.5.50](#) also allows us to make some remarks about the measured lineshapes in surface nonlinear spectroscopy. From this point of view, we may regard [equation B1.5.50](#) as being representative of the surface nonlinear response typically encountered near any resonance: It has a strongly varying resonant contribution together with a spectrally flat non-resonant background. The interesting aspect of this situation arises from the fact that we generally detect the *intensity* of the SH or SF signal, which is proportional to $|\chi_{s,ijk}^{(2)}|^2$. Consequently, interference between the resonant and non-resonant contributions is expected.

Depending on the relative phase difference between these terms, one may observe various experimental spectra, as illustrated in figure B1.5.14. This type of behaviour, while potentially a source of confusion, is familiar for other types of nonlinear spectroscopy, such as CARS (coherent anti-Stokes Raman scattering) [[30](#), [31](#)] and can be readily incorporated into modelling of measured spectral features.

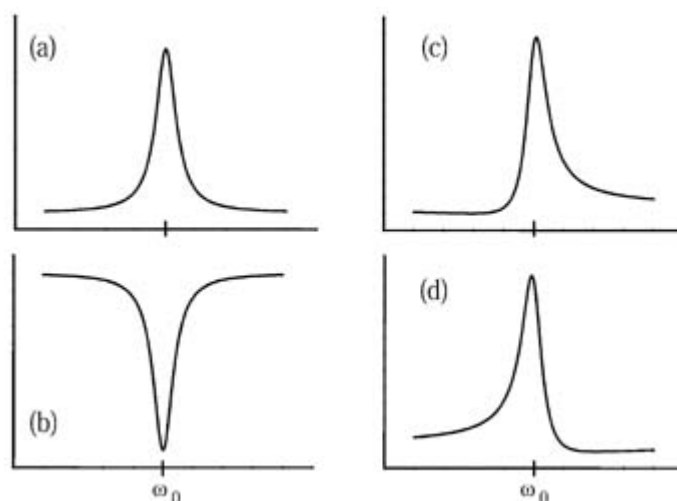


Figure B1.5.14 Possible lineshapes for an SFG resonance as a function of the infrared frequency ω_{IR} . The measured SFG signal is proportional to $|\chi^{\text{NR}} + A/(\omega_{\text{IR}} - \omega_0 + i\Gamma)|^2$. Assuming both χ^{NR} and Γ are real and positive, we obtain the lineshapes for various cases: (a) $\chi^{\text{NR}} \ll A/\Gamma$ (b) A is purely imaginary and negative with $|\chi^{\text{NR}}\Gamma/A| > \frac{1}{2}$ (c) A is real and positive; and (d) A is real and negative. Note the apparent blue and red shifts of the peaks in cases (c) and (d), respectively.

We now present one of the many examples of interfacial vibrational spectroscopy using SFG. [Figure B1.5.15](#) shows the surface vibrational spectrum of the water/air interface at a temperature of 40 °C [[83](#)]. Notice that the spectrum exhibits peaks at 3680, 3400 and 3200 cm^{-1} . These features arise from the OH stretching mode of water molecules in different environments. The highest frequency peak is assigned to free OH groups, the next peak to water molecules with hydrogen-bonding to neighbours in a relatively disordered structure and the lowest frequency peak to water molecules in a well-ordered tetrahedrally bonded (ice-like) structure. In addition to the analogy of these assignments to water molecules in different bulk environments, the assignments are compatible with the measured temperature dependence of the spectra. The strong and narrow peak at 3680 cm^{-1} provides interesting new information about the water surface. It indicates that a substantial fraction of the surface water molecules have unbonded OH groups protruding from the surface of the water into the vapour phase. This study exemplifies the unique capabilities of

surface SFG, as there is no other technique that could probe the liquid/vapour interface in the presence of strong features of the vibrational modes of the bulk water molecules.

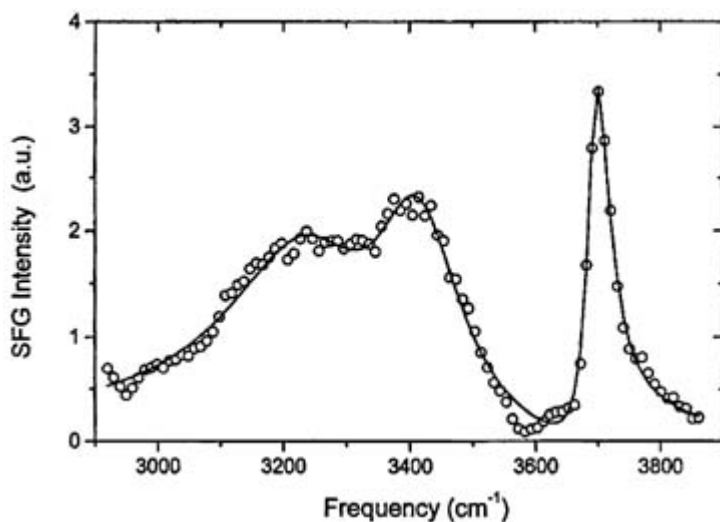


Figure B1.5.15 SFG spectrum for the water/air interface at 40 °C using the ssp polarization combination (s-, s- and p-polarized sum-frequency signal, visible input and infrared input beams, respectively). The peaks correspond to OH stretching modes. (After [83].)

An important consideration in spectroscopic measurements concerns the bandwidth of the laser sources. In order to resolve the vibrational resonances in a conventional approach, one needs, in the conventional scheme, a tunable source that has a narrow bandwidth compared to the resonance being studied. For typical resolutions, this requirement implies, by uncertainty principle, that IR pulses of picosecond or longer duration must be used. On the other hand, ultrafast pulsed IR sources with broad bandwidths are quite attractive from the experimental standpoint. In order to make use of these sources, two types of new experimental techniques have been introduced. One technique involves mixing the broadband IR source ($\sim 300 \text{ cm}^{-1}$) with a narrowband visible input ($\sim 5 \text{ cm}^{-1}$). By spectrally resolving the SF output, we may then obtain resolution of the IR spectrum limited only by the linewidth of the visible source [84, 85]. This result follows from the fact that $\omega_{\text{IR}} = \omega_{\text{SF}} - \omega_{\text{vis}}$ must be satisfied for the SFG process. The second new approach involves the application of a Fourier transform scheme [86]. This is accomplished by passing the IR pulses through an interferometer and then mixing these pairs of pulses with visible radiation at the surface.

B1.5.4.5 DYNAMICS

Many of the fundamental physical and chemical processes at surfaces and interfaces occur on extremely fast time scales. For example, atomic and molecular motions take place on time scales as short as 100 fs, while surface electronic states may have lifetimes as short as 10 fs. With the dramatic recent advances in laser technology, however, such time scales have become increasingly accessible. Surface nonlinear optics provides an attractive approach to capture such events directly in the time domain. Some examples of application of the method include probing the dynamics of melting on the time scale of phonon vibrations [87], photoisomerization of molecules [88], molecular dynamics of adsorbates [89, 90], interfacial solvent dynamics [91], transient band-flattening in semiconductors [92] and laser-induced desorption [93]. A review article discussing such time-resolved studies in metals can be found in

[94]. The SHG and SFG techniques are also suitable for studying dynamical processes occurring on slower time scales. Indeed, many valuable studies of adsorption, desorption, diffusion and other surface processes have been performed on time scales of milliseconds to seconds.

In a typical time-resolved SHG (SFG) experiment using femtosecond to picosecond laser systems, two (three) input laser beams are necessary. The pulse from one of the lasers, usually called the pump laser, induces the

reaction or surface modification. This defines the starting point ($t = 0$). A second pulse (or a second set of synchronized pulses, for the case of SFG) delayed relative to the first pulse by a specified time Δt is used to probe the reaction as it evolves. By varying this time delay Δt , the temporal evolution of the reaction can be followed. In order to preserve the inherent time resolution of an ultrafast laser, the relevant pulses are generally derived from a common source. For instance, in a basic time-resolved SHG experiment, where both the pump and probe pulses are of the same frequency, one simply divides the laser beam into two sets of pulses with a beam splitter. One of these pulses travels a fixed distance to the sample, while the other passes through a variable delay line to the sample. This approach provides a means of timing with sub-femtosecond accuracy, if desired. In some cases, at least one of the input beams has a different frequency from the others. Such pulses can be produced through processes such as harmonic generation or optical parametric generation from the main laser pulse.

As an example of this class of experiment, we consider an experimental study of the dynamics of molecular orientational relaxation at the air/water interface [90]. Such investigations are of interest as a gauge of the local environment at the surface of water. The measurements were performed with time-resolved SHG using Coumarin 314 dye molecules as the probe. In order to examine orientational motion, an anisotropic orientational distribution of molecules must first be produced. This is accomplished through a photoselection process in which the interface is irradiated by a linearly polarized laser pulse that is resonant with an electronic transition in the dye molecules. Those molecules that are oriented with their transition dipole moments parallel to the polarization of the pump beam are preferentially excited, producing an orientational anisotropy in the ground- and excited-state population. Subsequently, these anisotropic orientational distributions relax to the equilibrium configuration. The time evolution of the rotational anisotropy was followed by detecting the SH of a probe laser pulse as a function of the delay time, as shown in [figure B1.5.16](#). Through a comparison of the results for different initial anisotropic distributions (produced by two orthogonal linearly-polarized pump beams, as shown in the figure, as well as by circularly-polarized pump radiation), one may deduce rates for both in-plane and out-of-plane orientational relaxation. The study yielded the interesting result that the orientational relaxation times at the liquid/vapour interface significantly exceeded those for the Coumarin molecules in the bulk of water. This finding was interpreted as reflecting the increased friction encountered in the surface region where the water molecules are more highly ordered than in the bulk liquid.

-39-

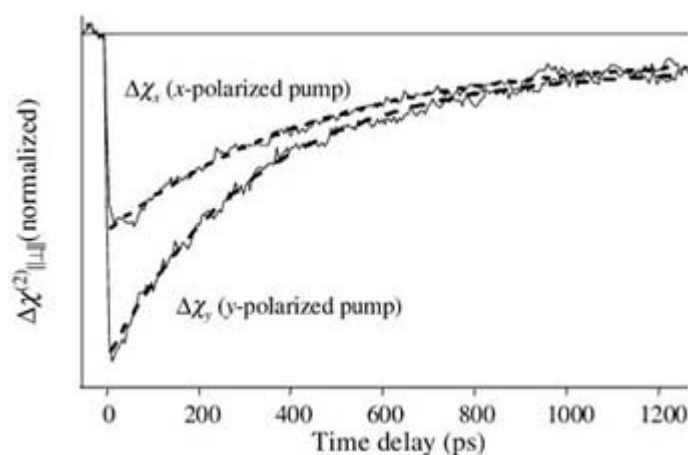


Figure B1.5.16 Rotational relaxation of Coumarin 314 molecules at the air/water interface. The change in the SH signal is recorded as a function of the time delay between the pump and probe pulses. Anisotropy in the orientational distribution is created by linearly polarized pump radiation in two orthogonal directions in the surface. (After [90].)

B1.5.4.6 SPATIAL RESOLUTION

Another application of surface SHG or SFG involves the exploitation of the lateral resolution afforded by these optical processes. While the dimension of the optical wavelength obviously precludes direct access to the length scale of atoms and molecules, one can examine the micrometre and submicrometre length scale that is important in many surface and interface processes. Spatial resolution may be achieved simply by detecting the nonlinear response with a focused laser beam that is scanned across the surface [95]. Alternatively, one may illuminate a large area of the surface and image the emitted nonlinear radiation [96]. Applications of this imaging capability have included probing of magnetic domains [97] and spatially varying electric fields [98]. The application of near-field techniques may permit, as in linear optics, the attainment of spatial resolution below the diffraction limit. In a recent work [99], submicrometre spatial resolution was indeed reported by collecting the emitted SH radiation for excitation with a near-field fibre probe.

Diffraction measurements offer a complementary approach to the real-space imaging described earlier. In such schemes, periodically modulated surfaces are utilized to produce well-defined SH (or SF) radiation at discrete angles, as dictated by the conservation of the in-plane component of the wavevector. As an example of this approach, a grating in the surface adsorbate density may be produced through laser-induced desorption in the field of two interfering beams. This monolayer grating will readily produce diffracted SH beams in addition to the usual reflected beam. In addition to their intrinsic interest, such structures have permitted precise measurements of surface diffusion. One may accomplish this by observing the temporal evolution of SH diffraction efficiency, which falls as surface diffusion causes the modulation depth of the adsorbate grating to decrease. This technique has been applied to examine diffusion of adsorbates on the surface of metals [100] and semiconductors [101].

B1.5.4.7 ELECTRIC AND MAGNETIC FIELD PERTURBATION

Probing electric and magnetic fields, and the effects induced by them, is of obvious interest in many areas of science and technology. We considered earlier the influence of such perturbations in a general fashion in [section B1.5.2.2](#). Here we describe some related experimental measurements and applications. Electric fields act to break inversion and thus

-40-

may yield bulk SHG and SFG signals from centrosymmetric media surrounding the interface, in addition to any field-dependent contribution of the interface itself. The high sensitivity of SHG toward applied electric fields was first demonstrated in calcite by Terhune *et al* as early as 1962 [102]. Subsequent investigations of EFISH from centrosymmetric media have involved semiconductor/electrolyte and metal/electrolyte interfaces [103, 104 and 105], as well as metal-oxide–semiconductor interfaces [106, 107].

The generality of the EFISH process has led to a variety of applications. These include probing the surface potential at interfaces involving liquids. Such measurements rely on the fact that the EFISH field is proportional to the voltage drop across the polarized layer provided, as is generally the case, that this region is thin compared to the scale of an optical wavelength [108]. This effect also serves as a basis for probing surface reactions involving changes of charge state, such as acid/base equilibria [109]. Another related set of applications involves probing electric fields in semiconductors, notably in centrosymmetric material silicon [98, 110, 111]. These studies have demonstrated the capability for spatial resolution, vector analysis of the electric field and, significantly, ultrafast (subpicosecond) time resolution. These capabilities of SHG complement other optical schemes, such as electro-optical and photoconductive sampling, for probing the dynamics of electric fields on very fast time scales.

The influence of an applied magnetic field, as introduced in [section B1.5.2.2](#), is quite different from that of an applied electric field. A magnetic field may perturb the interfacial nonlinear response (and that of the weak bulk terms), but it does not lead to any dipole-allowed bulk nonlinear response. Thus, in the presence of magnetic fields and magnetization, SHG remains a probe that is highly specific to surfaces and interfaces. It

may be viewed as the interface-sensitive analogue of linear magneto-optical effects. The first demonstration of the influence of magnetization on SHG was performed on an Fe(110) surface [112]. Subsequent applications have included examination of other materials for which both the bulk and surface exhibit magnetization. For these systems, surface specificity is of key importance. In addition, the technique has been applied to examine buried magnetic interfaces [113]. Excellent review articles on this subject matter are presented in [23] and [114].

B1.5.4.8 RECENT DEVELOPMENTS

Up to this point, our discussion of surface SHG and SFG has implicitly assumed that we are examining a smooth planar surface. This type of interface leads to well-defined and highly collimated transmitted and reflected beams. On the other hand, many material systems of interest in probing surfaces or interfaces are not planar in character. From the point of view of symmetry, the surface sensitivity for interfaces of centrosymmetric media should apply equally well for such non-planar interfaces, although the nature of the electromagnetic wave propagation may be modified to a significant degree. One case of particular interest concerns appropriately roughened surfaces of noble metals. These were shown as early as 1974 [115] to give rise to strong enhancements in Raman scattering of adsorbed species and led to extensive investigation of the phenomenon of surface-enhanced Raman scattering or SERS [5]. Significant enhancements in the SHG signals from such surfaces have also been found [116]. The resulting SH radiation is diffuse, but has been shown to preserve a high degree of surface sensitivity. Carrying this progression from planar surfaces one step further, researchers have recently demonstrated the possibility of probing the surfaces of small particles by SHG.

Experimental investigations of the model system of dye molecules adsorbed onto surfaces of polystyrene spheres have firmly established the sensitivity and surface specificity of the SHG method even for particles of micrometre size [117]. The surface sensitivity of the SHG process has been exploited for probing molecular transport across the bilayer in liposomes [118], for measurement of electrostatic potentials at the surface of small particles [119] and for imaging

-41-

membranes in living cells [120]. The corresponding theoretical description of SHG from the surfaces of small spheres has been examined recently using the type of formalism presented earlier in this chapter [121]. Within this framework, the leading-order contributions to the SH radiation arise from the non-local excitation of the dipole and the local excitation of the quadrupole moments. This situation stands in contrast to linear optical (Rayleigh) scattering, which arises from the local excitation of the dipole moment.

B1.5.5 CONCLUSION

In this brief chapter, we have attempted to describe some of the underlying principles of second-order nonlinear optics for the study of surfaces and interfaces. The fact that the technique relies on a basic symmetry consideration to obtain surface specificity gives the method a high degree of generality. As a consequence, our review of some of the applications of the method has necessarily been quite incomplete. Still, we hope that the reader will gain some appreciation for the flexibility and power of the method. Over the last few years, many noteworthy applications of the method have been demonstrated. Further advances may be anticipated from on-going development of the microscopic theory, as well as from adaptation of the macroscopic theory to new experimental conditions and geometries. At the same time, we see continual progress in the range and ease of use of the technique afforded by the impressive improvement of the performance and reliability of high-power laser sources.

REFERENCES

- [1] Franken P A, Hill A E, Peters C W and Weinreich G 1961 Generation of optical harmonics *Phys. Rev. Lett.* **7** 118
 - [2] Somorjai G A 1981 *Chemistry in Two Dimensions* (Ithaca, NY: Cornell University Press)
 - [3] Duke C B (ed) 1994 Surface science: the first thirty years *Surf. Sci.* **299/300** 1–1054
 - [4] Dumas P, Weldon M K, Chabal Y J and Williams G P 1999 Molecules at surfaces and interfaces studied using vibrational spectroscopies and related techniques *Surf. Rev. Lett.* **6** 225–55
 - [5] Chang R K and Furtak T E 1982 *Surface Enhanced Raman Scattering* (New York: Plenum)
 - [6] Moskovits M 1985 Surface-enhanced spectroscopy *Rev. Mod. Phys.* **57** 783–826
 - [7] Campion A and Kambhampati P 1998 Surface-enhanced Raman scattering *Chem. Soc. Rev.* **27** 241–50
 - [8] Aspnes D E 1993 New developments in spectroellipsometry—the challenge of surfaces *Thin Solid films* **233** 1–8
 - [9] Azzam R M A and Bashara N M 1977 *Ellipsometry and Polarized Light* (Amsterdam: North-Holland)
 - [10] McGilp J F, Patterson C H and Weaire D L (ed) 1995 *Epioptics: Linear and Nonlinear Optical Spectroscopy of Surfaces and Interfaces* (Berlin: Springer)
 - [11] Aspnes D E 1985 Above-bandgap optical anisotropies in cubic semiconductors: a visible–near ultraviolet probe of surfaces *J. Vac. Sci. Technol. B* **3** 1498–506
-

-42-

- [12] Richmond G L, Robinson J M and Shannon V L 1988 Second harmonic generation studies of interfacial structure and dynamics *Prog. Surf. Sci.* **28** 1–70
- [13] Shen Y R 1989 Surface-properties probed by second-harmonic and sum-frequency generation *Nature* **337** 519–25
- [14] Shen Y R 1989 Optical second harmonic-generation at interfaces *Ann. Rev. Phys. Chem.* **40** 327–50
- [15] Heinz T F 1991 Second-order nonlinear optical effects at surfaces and interfaces *Nonlinear Surface Electromagnetic Phenomena* ed H-E Ponath and G I Stegeman (Amsterdam: North-Holland) pp 353–416
- [16] Eienthal K B 1992 Equilibrium and dynamic processes at interfaces by second harmonic and sum frequency generation *Ann. Rev. Phys. Chem.* **43** 627–61
- [17] Corn R M and Higgins D A 1994 Optical second-harmonic generation as a probe of surface-chemistry *Chem. Rev.* **94** 107–25
- [18] McGilp J F 1995 Optical characterisation of semiconductor surfaces and interfaces *Prog. Surf. Sci.* **49** 1–106
- [19] Reider G A and Heinz T F 1995 Second-order nonlinear optical effects at surfaces and interfaces: recent advances *Photonic Probes of Surfaces* ed P Halevi (Amsterdam: Elsevier) pp 413–78
- [20] Bain C D 1995 Sum-frequency vibrational spectroscopy of the solid–liquid interface *J. Chem. Soc. Faraday Trans.* **91** 1281–96
- [21] Eienthal K B 1996 Liquid interfaces probed by second-harmonic and sum-frequency spectroscopy *Chem. Rev.* **96** 1343–60
- [22] Richmond G L 1997 Vibrational spectroscopy of molecules at liquid/liquid interfaces *Anal. Chem.* **69** A536–43
- [23] Rasing Th 1998 Nonlinear magneto-optical studies of ultrathin films and multilayers *Nonlinear Optics in Metals* ed

K H Bennemann (Oxford: Clarendon) pp 132–218

- [24] Lüpke G 1999 Characterization of semiconductor interfaces by second-harmonic generation *Surf. Sci. Rep.* **35** 75–161
 - [25] McGilp J F 1999 Second-harmonic generation at semiconductor and metal surfaces *Surf. Rev. Lett.* **6** 529–58
 - [26] Miranda P B and Shen Y R 1999 Liquid interfaces: a study by sum-frequency vibrational spectroscopy *J. Phys. Chem. B* **103** 3292–307
 - [27] Shultz M J, Schnitzer C, Simonelli D and Baldelli S 2000 Sum-frequency generation spectroscopy of the aqueous interface: ionic and soluble molecular solutions *Int. Rev. Phys. Chem.* **19** 123–53
 - [28] Brevet P F 1997 *Surface Second Harmonic Generation* (Lausanne: Presses Polytechniques et Universitaires Romandes)
 - [29] Bennemann K H (ed) 1998 *Nonlinear Optics in Metals* (Oxford: Clarendon)
 - [30] Levenson M D and Kano S S 1988 *Introduction to Nonlinear Laser Spectroscopy* (Boston: Academic)
 - [31] Mukamel S 1995 *Principles of Nonlinear Optical Spectroscopy* (Oxford: Oxford University Press)
 - [32] Flytzanis C 1975 Theory of nonlinear optical susceptibilities *Quantum Electronics* vol 1A (New York: Academic)
 - [33] Butcher P N and Cotter D 1990 *The Elements of Nonlinear Optics* (Cambridge: Cambridge University Press)
 - [34] Boyd R W 1992 *Nonlinear Optics* (New York: Academic)
-

-43-

- [35] Jackson J D 1975 *Classical Electrodynamics* (New York: Wiley)
- [36] Bloembergen N 1965 *Nonlinear Optics* (New York: Benjamin)
- [37] Shen Y R 1984 *The Principles of Nonlinear Optics* (New York: Wiley)
- [38] Sipe J E 1987 New Green-function formalism for surface optics *J. Opt. Soc. Am. B* **4** 481–9
- [39] Mizrahi V and Sipe J E 1988 Phenomenological treatment of surface second-harmonic generation *J. Opt. Soc. Am. B* **5** 660–7
- [40] Bloembergen N and Pershan P S 1962 Light waves at the boundary of nonlinear media *Phys. Rev.* **128** 606–22
- [41] Ye P and Shen Y R 1983 Local-field effect on linear and nonlinear optical properties of adsorbed molecules *Phys. Rev. B* **28** 4288–94
- [42] Hayden L M 1988 Local-field effects in Langmuir–Blodgett films of hemicyanine and behenic acid mixtures *Phys. Rev. B* **38** 3718–21
- [43] Zyss J (ed) 1994 *Molecular Nonlinear Optics: Materials, Physics, and Devices* (Boston: Academic)
- [44] Prasad P N and Williams D J 1991 *Introduction to Nonlinear Optical Effects in Molecules and Polymers* (New York: Wiley)
- [45] Feibelman P J 1982 Surface electromagnetic fields *Prog. Surf. Sci.* **12** 287–407
- [46] Rudnick J and Stern E A 1971 Second-harmonic radiation from metal surfaces *Phys. Rev. B* **4** 4274–90
- [47] Keller O 1986 Random-phase-approximation study of the response function describing optical second-harmonic generation from a metal selvedge *Phys. Rev. B* **33** 990–1009

- [48] Liebsch A 1997 *Electronic Excitations at Metal Surfaces* (New York: Plenum)
- [49] Petukhov A V 1995 Sum-frequency generation on isotropic surfaces: general phenomenology and microscopic theory for jellium surfaces *Phys. Rev. B* **52** 16 901–11
- [50] Luce T A and Bennemann K H 1998 Nonlinear optical response of noble metals determined from first-principles electronic structures and wave functions: calculation of transition matrix elements *Phys. Rev. B* **58** 15 821–6
- [51] Schaich W L 2000 Calculations of second-harmonic generation for a jellium metal surface *Phys. Rev. B* **61** 10 478–83
- [52] Ghahramani E, Moss D J and Sipe J E 1990 Second-harmonic generation in odd-period, strained, $(\text{Si})_n/(\text{Ge})_n/\text{Si}$ superlattices and at Si/Ge interfaces *Phys. Rev. Lett.* **64** 2815–18
- [53] Gavrilenko V I and Rebrost F 1995 Nonlinear optical susceptibility of the surfaces of silicon and diamond *Surf. Sci. B* **331–3** 1355–60
- [54] Mendoza B S, Gaggiotti A and Del Sole R 1998 Microscopic theory of second harmonic generation at Si(100) surfaces *Phys. Rev. Lett.* **81** 3781–4
- [55] Lim D, Downer M C, Ekerdt J G, Arzate N, Mendoza B S, Gavrilenko V I and Wu R Q 2000 Optical second harmonic spectroscopy of boron-reconstructed Si(001) *Phys. Rev. Lett.* **84** 3406–9
- [56] Pan R P, Wei H D and Shen Y R 1989 Optical second-harmonic generation from magnetized surfaces *Phys. Rev. B* **39** 1229–34
- [57] Stolle R, Marowsky G, Schwarzberg E and Berkovic G 1996 Phase measurements in nonlinear optics *Appl. Phys. B* **63** 491–8

- [58] Kemnitz K, Bhattacharyya K, Hicks J M, Pinto G R, Eisenthal K B and Heinz T F 1986 The phase of second-harmonic light generated at an interface and its relation to absolute molecular orientation *Chem. Phys. Lett.* **131** 285–90
- [59] Dadap J I, Hu X F, Russell N M, Ekerdt J G, Lowell J K and Downer M C 1995 Analysis of second-harmonic generation by unamplified, high-repetition-rate, ultrashort laser pulses at Si(001) interfaces *IEEE J. Selected Topics Quantum Electron* **1** 1145–55
- [60] Heinz T F, Tom H W K and Shen Y R 1983 Determination of molecular orientation of monolayer adsorbates by optical second-harmonic generation *Phys. Rev. A* **28** 1883–5
- [61] Goh M C, Hicks J M, Kemnitz K, Pinto G R, Bhattacharyya K, Heinz T F and Eisenthal K B 1988 Absolute orientation of water-molecules at the neat water-surface *J. Phys. Chem.* **92** 5074–5
- [62] Tom H W K, Heinz T F and Shen Y R 1983 Second-harmonic reflection from silicon surfaces and its relation to structural symmetry *Phys. Rev. Lett.* **51** 1983
- [63] Aktsipetrov O A, Baranova I M and Il'inskii Y A 1986 Surface contribution to the generation of reflected second-harmonic light for centrosymmetric semiconductors *Zh. Eksp. Teor. Fiz.* **91** 287–97 (Engl. transl. 1986 *Sov. Phys. JETP* **64** 167–73)
- [64] Sipe J E, Moss D J and van Driel H M 1987 Phenomenological theory of optical second- and third-harmonic generation from cubic centrosymmetric crystals *Phys. Rev. B* **35** 1129–41
- [65] van Hasselt C W, Verheijen M A and Rasing Th 1990 Vicinal Si(111) surfaces studied by optical second-harmonic generation: step-induced anisotropy and surface-bulk discrimination *Phys. Rev. B* **42** 9263–6
- [66] Heinz T F, Loy M M T and Iyer S S 1987 Nonlinear optical study of Si epitaxy *Mater. Res. Soc. Symp. Proc.* **55** 697
- [67] Dadap J I, Doris B, Deng Q, Downer M C, Lowell J K and Diebold A C 1994 Randomly oriented Ångstrom-scale

microroughness at the Si(100)/SiO₂ interface probed by optical second harmonic generation *Appl. Phys. Lett.* **64** 2139–41

- [68] Verbiest T, Kauranen M and Persoons A 1999 Second-order nonlinear optical properties of chiral thin films *J. Mater. Chem.* **9** 2005–12
- [69] Petralli-Mallow T, Wong T M, Byers J D, Yee H I and Hicks J M 1993 Circular dichroism spectroscopy at interfaces—a surface second harmonic-generation study *J. Phys. Chem.* **97** 1383–8
- [70] Byers J D, Yee H I and Hicks J M 1994 A second harmonic generation analog of optical rotary dispersion for the study of chiral monolayers *J. Chem. Phys.* **101** 6233–41
- [71] Giordmaine J A 1965 Nonlinear optical properties of liquids *Phys. Rev. A* **138** 1599
- [72] Höfer U 1996 Nonlinear optical investigations of the dynamics of hydrogen interaction with silicon surfaces *Appl. Phys. A* **63** 533–47
- [73] Reider G A, Höfer U and Heinz T F 1991 Desorption-kinetics of hydrogen from the Si(111)7 × 7 surface *J. Chem. Phys.* **94** 4080–3
- [74] Höfer U, Li L P and Heinz T F 1992 Desorption of hydrogen from Si(100)2 × 1 at low coverages—the influence of π-bonded dimers on the kinetics *Phys. Rev. B* **45** 9485–8
- [75] Dadap J I, Xu Z, Hu X F, Downer M C, Russell N M, Ekerdt J G and Aktsiperov O A 1997 Second-harmonic spectroscopy of a Si(001) surface during calibrated variations in temperature and hydrogen coverage *Phys. Rev. B* **56** 13 367–79
- [76] Crawford M J, Frey J G, VanderNoot T J and Zhao Y G 1996 Investigation of transport across an immiscible liquid/liquid interface—electrochemical and second harmonic generation studies *J. Chem. Soc. Faraday Trans.* **92** 1369–73

-45-

- [77] Rasing Th, Shen Y R, Kim M W, Valint P Jr and Bock J 1985 Orientation of surfactant molecules at a liquid-air interface measured by optical second-harmonic generation *Phys. Rev. A* **31** 537–9
- [78] Zhuang X, Miranda P B, Kim D and Shen Y R 1999 Mapping molecular orientation and conformation at interfaces by surface nonlinear optics *Phys. Rev. B* **59** 12 632–40
- [79] Heinz T F, Himpel F J, Palange E and Burstein E 1989 Electronic transitions at the CaF₂/Si(111) interface probed by resonant three-wave-mixing spectroscopy *Phys. Rev. Lett.* **63** 644–7
- [80] Hicks J M, Petralli-Mallow T and Byers J D 1994 Consequences of chirality in second-order nonlinear spectroscopy at interfaces *Faraday Disc.* **99** 341–57
- [81] Zhu X D, Suhr H and Shen Y R 1987 Surface vibrational spectroscopy by infrared-visible sum frequency generation *Phys. Rev. B* **35** 3047–59
- [82] Lin S H and Villaeys A A 1994 Theoretical description of steady-state sum-frequency generation in molecular absorbates *Phys. Rev. A* **50** 5134–44
- [83] Du Q, Superfine R, Freysz E and Shen Y R 1993 Vibrational spectroscopy of water at the vapor–water interface *Phys. Rev. Lett.* **70** 2313–16
- [84] Richter L T, Petralli-Mallow T P and Stephenson J C 1998 Vibrationally resolved sum-frequency generation with broad-bandwidth infrared pulses *Opt. Lett.* **23** 1594–6
- [85] van der Ham E W M, Vreken Q H F and Eliel E R 1996 Self-dispersive sum-frequency generation at interfaces *Opt. Lett.* **21** 1448–50
- [86] McGuire J A, Beck W, Wei X and Shen Y R 1999 Fourier-transform sum-frequency surface vibrational

spectroscopy with femtosecond pulses *Opt. Lett.* **24** 1877–9

- [87] Shank C V, Yen R and Hirlimann C 1983 Femtosecond-time-resolved surface structural dynamics of optically excited silicon *Phys. Rev. Lett.* **51** 900–2
- [88] Sitzmann E V and Eisenthal K B 1988 Picosecond dynamics of a chemical-reaction at the air–water interface studied by surface second-harmonic generation *J. Phys. Chem.* **92** 4579–80
- [89] Castro A, Sitzmann E V, Zhang D and Eisenthal K B 1991 Rotational relaxation at the air–water interface by time-resolved second-harmonic generation *J. Phys. Chem.* **95** 6752–3
- [90] Zimdars D, Dadap J I, Eisenthal K B and Heinz T F 1999 Anisotropic orientational motion of molecular adsorbates at the air–water interface *J. Chem. Phys.* **103** 3425–33
- [91] Zimdars D, Dadap J I, Eisenthal K B and Heinz T F 1999 Femtosecond dynamics of solvation at the air/water interface *Chem. Phys. Lett.* **301** 112–20
- [92] Lantz J M and Corn R M 1994 Time-resolved optical second harmonic generation measurements of picosecond band flattening processes at single crystal TiO₂ electrodes *J. Phys. Chem.* **98** 9387–90
- [93] Prybyla J A, Tom H W K and Aumiller G D 1992 Femtosecond time-resolved surface reaction: desorption of CO from Cu(111) in < 325 fs *Phys. Rev. Lett.* **68** 503–6
- [94] Hohlfeld J, Conrad U, Müller Wellershoff S S and Matthias E 1998 Femtosecond time-resolved linear and second-order reflectivity of metals *Nonlinear Optics in Metals* ed K H Bennemann (Oxford: Clarendon) pp 219–67

- [95] Boyd G T, Shen Y R and Hansch T W 1986 Continuous-wave second-harmonic generation as a surface microprobe *Opt. Lett.* **11** 97–9
- [96] Schultz K A and Seebauer E G 1992 Surface diffusion of Sb on Ge(111) monitored quantitatively with optical second harmonic microscopy *J. Chem. Phys.* **97** 6958–67
- [97] Kirilyuk V, Kirilyuk A and Rasing Th 1997 A combined nonlinear and linear magneto-optical microscopy *Appl. Phys. Lett.* **70** 2306–8
- [98] Dadap J I, Shan J, Weling A S, Misewich J A, Nahata A and Heinz T F 1999 Measurement of the vector character of electric fields by optical second-harmonic generation *Opt. Lett.* **24** 1059–61
- [99] Smolyaninov I P, Zayats A V and Davis C C 1997 Near-field second-harmonic imaging of ferromagnetic and ferroelectric materials *Opt. Lett.* **22** 1592–4
- [100] Zhu X D, Rasing T H and Shen Y R 1988 Surface diffusion of CO on Ni(111) studied by diffraction of optical second-harmonic generation off a monolayer grating *Phys. Rev. Lett.* **61** 2883–5
- [101] Reider G A, Höfer U and Heinz T F 1991 Surface diffusion of hydrogen on Si(111) 7*7 *Phys. Rev. Lett.* **66** 1994–7
- [102] Terhune R W, Maker P D and Savage C M 1962 Optical harmonic generation in calcite *Phys. Rev. Lett.* **8** 404
- [103] Lee C H, Chang R K and Bloembergen N 1967 Nonlinear electroreflectance in silicon and silver *Phys. Rev. Lett.* **18** 167–70
- [104] Aktsipetrov O A and Mishina E D 1984 Nonlinear optical electroreflection in germanium and silicon *Dokl. Akad. Nauk SSSR* **274** 62–5
- [105] Fischer P R, Daschbach J L and Richmond G L 1994 Surface second harmonic studies of Si(111)/electrolyte and Si(111)/SiO₂/electrolyte interfaces *Chem. Phys. Lett.* **218** 200–5
- [106] Aktsipetrov O A, Fedyanin A A, Golovkina V N and Murzina T V 1994 Optical second-harmonic generation

induced by a DC electric field at the Si–SiO₂ interface *Opt. Lett.* **19** 1450–2

- [107] Dadap J I, Hu X F, Anderson M H, Downer M C, Lowell J K and Aktsiperov O A 1996 Optical second-harmonic electroreflectance spectroscopy of a Si(0001) metal–oxide–semiconductor structure *Phys. Rev. B* **53** R7607–9
- [108] Zhao X L, Ong S W and Eisenthal K B 1993 Polarization of water-molecules at a charged interface. Second harmonic studies of charged monolayers at the air/water interface *Chem. Phys. Lett.* **202** 513–20
- [109] Ong S W, Zhao X L and Eisenthal K B 1992 Polarization of water-molecules at a charged interface: second harmonic studies of the silica water interface *Chem. Phys. Lett.* **191** 327–35
- [110] Nahata A, Heinz T F and Misewich J A 1996 High-speed electrical sampling using optical second-harmonic generation *Appl. Phys. Lett.* **69** 746–8
- [111] Ohlhoff C, Lupke G, Meyer C and Kurz H 1997 Static and high-frequency electric fields in silicon MOS and MS structures probed by optical second-harmonic generation *Phys. Rev. B* **55** 4596–606
- [112] Reif J, Zink J C, Schneider C-M and Kirschner J 1991 Effects of surface magnetism on optical second harmonic generation *Phys. Rev. Lett.* **67** 2878–81
- [113] Spierings G, Koutsos V, Wierenga H A, Prins M W J, Abraham D and Rasing Th 1993 Optical second harmonic generation study of interface magnetism *Surf. Sci.* **287–8** 747–9

-47-

- Spierings G, Koutsos V, Wierenga H A, Prins M W J, Abraham D and Rasing Th 1993 Interface magnetism studied by optical second harmonic generation *J. Magn. Magn. Mater.* **121** 109–11
- [114] Vollmer R 1998 Magnetization-induced second harmonic generation from surfaces and ultrathin films *Nonlinear Optics in Metals* ed K H Bennemann (Oxford: Clarendon) pp 42–131
 - [115] Fleischmann M, Hendra P J and McQuillan A J 1974 Raman-spectra of pyridine adsorbed at a silver electrode *Chem. Phys. Lett.* **26** 163–6
 - [116] Chen C K, de Castro A R B and Shen Y R 1981 Surface enhanced second-harmonic generation *Phys. Rev. Lett.* **46** 145–8
 - [117] Wang H, Yan E C Y, Borguet E and Eisenthal K B 1996 Second harmonic generation from the surface of centrosymmetric particles in bulk solution *Chem. Phys. Lett.* **259** 15–20
 - [118] Srivastava A and Eisenthal K B 1998 Kinetics of molecular transport across a liposome bilayer *Chem. Phys. Lett.* **292** 345–51
 - [119] Yan E C Y, Liu Y and Eisenthal K B 1998 New method for determination of surface potential of microscopic particles by second harmonic generation *J. Phys. Chem. B* **102** 6331–6
 - [120] Campagnola P J, Wei M D, Lewis A and Loew L M 1999 High-resolution nonlinear optical imaging of live cells by second harmonic generation *Biophys. J.* **77** 3341–9
 - [121] Dadap J I, Shan J, Eisenthal K B and Heinz T F 1999 Second-harmonic Rayleigh scattering from a sphere of centrosymmetric material *Phys. Rev. Lett.* **83** 4045–8

FURTHER READING

Further General texts on nonlinear optics

Boyd R W 1992 *Nonlinear Optics* (New York: Academic)

Butcher P N and Cotter D 1990 *The Elements of Nonlinear Optics* (Cambridge: Cambridge University Press)

Shen Y R 1984 *The Principles of Nonlinear Optics* (New York: Wiley)

Yariv A 1989 *Quantum Electronics* 3rd edn (New York: Wiley)

General texts on nonlinear optical spectroscopy

Demtröder W 1996 *Laser Spectroscopy: Basic Concepts and Instrumentation* 2nd edn (Berlin: Springer)

Levenson M D and Kano S S 1988 *Introduction to Nonlinear Laser Spectroscopy* (Boston, MA: Academic)

Mukamel S 1995 *Principles of Nonlinear Optical Spectroscopy* (Oxford: Oxford University Press)

General texts on surface nonlinear optics

-48-

Brevet P F 1997 *Surface Second Harmonic Generation* (Lausanne: Presses Polytechniques et Universitaires Romandes)

Bennemann K H (ed) 1998 *Nonlinear Optics in Metals* (Oxford: Clarendon)

McGilp J F, Patterson C H and Weaire D L (eds) 1995 *Epioptics: Linear and Nonlinear Optical Spectroscopy of Surfaces and Interfaces* (Berlin: Springer)

-1-

B1.6 Electron-impact spectroscopy

John H Moore

B1.6.0 INTRODUCTION

When one speaks of a ‘spectrum’, the dispersed array of colours from a luminous body comes to mind; however, in the most general sense, a spectrum is a record of the energy and probability of transitions between states of a substance. In electron spectroscopy the ‘spectrum’ takes the form of the energy distribution of electrons emanating from a sample. Electron spectroscopies are classified according to the phenomena giving rise to these electrons; historically, each technique has acquired an acronym until today one finds a veritable alphabet soup of electron spectroscopies in the scientific literature. For example, PES refers to photoelectron spectroscopy, a technique in which the detected electrons are emitted after the absorption of a photon induces transitions into the continuum beyond the first ionization potential of the sample. Electron-impact spectroscopies, the subject of this entry, entail the excitation of a transition by an electron impinging upon a sample with the subsequent measurement of the energy of the scattered electron. The spectrum is the scattered-electron intensity as a function of the difference between the incident- and scattered-electron energies—the *energy loss*.

The technologies of the various electron spectroscopies are similar in many ways. The techniques for measuring electron energies and the devices used to detect electrons are the same. All electron spectrometers must be housed in an evacuated container at pressures less than about 10^{-6} mbar since electrons cannot be transported through the atmosphere. Stray fields that perturb electron trajectories are a potential problem. To function correctly, an electron spectrometer must be shielded from the earth’s magnetic field.

Electron-impact energy-loss spectroscopy (EELS) differs from other electron spectroscopies in that it is possible to observe transitions to states below the first ionization edge; electronic transitions to excited states of the neutral, vibrational and even rotational transitions can be observed. This is a consequence of the detected electrons not originating in the sample. Conversely, there is a problem when electron impact induces an ionizing transition. For each such event there are two outgoing electrons. To precisely account for the energy deposited in the target, the two electrons must be measured in coincidence.

In comparison to optical spectroscopy, electron-impact spectroscopy offers a number of advantages. Some of these are purely technological while others are a result of physical differences in the excitation mechanism. The energy of an electron can be varied simply and smoothly by scanning the voltage applied to, and hence the potential difference between, electrodes in the spectrometer. The same technology is applicable to electrons with energies, and energy losses, in the millielectronvolt (meV) range as in the kiloelectronvolt (keV) range. At least in principle, measurements analogous to IR spectroscopy can be carried out in the same instrument as measurements akin to x-ray spectroscopy. Unlike an optical instrument, the source intensity and the transmission of an electron spectrometer are nearly independent of energy, making the electron instrument more suitable for absolute intensity measurements; however, an electron spectrometer cannot always provide resolution comparable to that of an optical instrument. Studies of rotational and vibrational excitation, particularly in surface adsorbates, are routinely carried out by electron spectroscopy with a resolution of 2 to 5 meV (16 to 40 cm^{-1}). A resolution of 5 to 30 meV is typically obtained for

-2-

electron-impact excitation of valence-electron transitions in atoms and molecules in the gas phase. Analogous studies by vis-UV spectroscopy easily provide 1 cm^{-1} resolution. For inner-shell-electron excitation, electron spectroscopy provides resolution comparable or superior to x-ray spectroscopy with discrete-line-source x-ray tubes. X-ray synchrotron sources now becoming available will provide better resolution and intensity than can be achieved with an electron spectrometer, but it must be borne in mind that an electron spectrometer is a relatively inexpensive, table-top device, whereas a synchrotron is a remote, multimillion-dollar facility. In many applications, electron-scattering experiments are more sensitive than optical experiments. This is in part due to the superior sensitivity of electron detectors. An electron multiplier has essentially unit efficiency (100%); a photomultiplier or photodiode may have an efficiency of only a few per cent. For surface analysis, electron spectroscopies have a special advantage over optical techniques owing to the short range of electrons in solids.

The mechanism by which a transition is induced by electron impact depends on the nature of the coupling between the projectile electron and the target; this in turn is influenced by the velocity and closeness of approach of the projectile to the target. There is a wide range of possibilities. A high-energy projectile electron may pass quickly by, delivering only a photon-like electric-field pulse to the target at the instant of closest approach. Less probable are hard, billiard-ball-like collisions between the projectile and one target electron. At low energies, slower, more intimate collisions are characterized by many-electron interactions. Depending upon the mechanism, the momentum transferred from projectile to target can vary from the minimum necessary to account for the transition energy to many times more. The interaction influences the type of transition that can be induced and the way in which the projectile is scattered. It is even possible for the projectile electron to be exchanged for a target electron, thus allowing for electron-spin-changing transitions. This state of affairs is a contrast to optical excitation where the momentum transfer is a constant and only 'dipole-allowed' transitions occur with significant probability.

B1.6.1.1 CROSS SECTION AND SIGNAL INTENSITY

The quantities to be measured in electron-impact spectroscopy are the probability of an electron impact's inducing a transition and the corresponding transition energy. The energy for the transition is taken from the kinetic energy of the projectile electron. Unlike the situation in optical spectroscopy, the exciting particle is not annihilated, but is scattered from the target at some angle to its initial direction. The scattering angle is a measure of the momentum transferred to the target, and, as such, is also an important variable.

The probability of a collision between an electron and a target depends upon the *impact parameter*, b , which is the perpendicular distance between the line of travel of the electron and the centre of force exerted by the target on the electron. The impact parameter is equivalent to the distance of closest approach if no potential is present between the electron and the target. For a hard-sphere collision between an infinitesimally small projectile and a target of radius r , the impact parameter must be less than r for a collision to occur, and, from simple geometry, the scattering angle $\theta = 2 \arccos(b/r)$. The scattering angle is large for small impact parameters, while for 'grazing' collisions, where b approaches r , the scattering angle is small. The probability of a collision is proportional to the cross sectional area of the target, πr^2 . Real collisions between an electron and an atom involve at least central-force-field potentials and, frequently, higher-multipole potentials, but the billiard-ball scattering model is so pervasive that collision probabilities are almost always expressed as cross sections, often denoted by the symbol σ with units of the order of the cross

-3-

section of an atom, such as 10^{-16} cm^2 or square Ångströms (Å^2). The atomic unit of cross section is the Bohr radius squared ($r_0^2 = 0.28 \times 10^{-16} \text{ cm}^2$).

In a very simple form of electron spectroscopy, known as *electron transmission spectroscopy*, the attenuation of an essentially monoenergetic beam of electrons is measured after passage through a sample. If the target is very thin or of such low density that most electrons pass through unscattered, the attenuation is small and the transmitted current, I (in units of electrons per unit time, s^{-1}), compared to the incident current, I_0 (s^{-1}), is given by

$$\frac{I}{I_0} = e^{-\sigma n \ell}$$

where n is the number density of particles in the target (typically in units of cm^{-3}), ℓ (cm) is the thickness of the target and σ is the *total electron scattering cross section*. The cross section depends upon the energy, E_0 , of the incident electrons: $\sigma = \sigma(E_0)$. The total electron scattering spectrum presented as I/I_0 as a function of E_0 bears an inverse relation to the cross section, the transmitted current decreasing as the cross section increases.

An electron-energy-loss spectrometer consists of an 'electron gun' that directs a collimated beam of electrons upon a sample, and an 'analyser' that collects electrons scattered in a particular direction (specified by θ and ϕ in spherical coordinates) and transmits to a detector those electrons with energy E . The electron-energy-loss spectrum is a plot of the scattered-electron current, I_s , arriving at the detector as a function of the energy loss, $(E_0 - E)$. The cross section for the inelastic scattering process giving rise to the observed signal depends upon the scattering angle: it is a *differential cross section*, $d\sigma/d\Omega$, where $d\Omega = \sin\theta \, d\theta \, d\phi$ in spherical coordinates. The magnitude of the cross section depends upon the incident electron energy: there will be a threshold below which the cross section is zero when the incident electron has insufficient energy to excite the transition, and there will be an incident electron energy for which the coupling between projectile and target is greatest and the cross section passes through a maximum. The instrument parameters, as well as the cross section, determine the actual signal level:

$$I_s = I_0 n \ell \left(\frac{d\sigma}{d\Omega} \right) \Delta\Omega$$

where $\Delta\Omega$ is the solid-angle field of view of the scattered-electron analyser.

The foregoing description of an electron energy-loss spectrometer assumes a monoenergetic incident electron beam and the excitation of a transition of negligible energy width. It is often the case that the transition intensity spans a range in energy, and, in addition, the incident beam has some energy spread and the analyser a finite bandpass. One must consider a cross section differential in both angle and energy, $d^2\sigma/d\Omega dE$. The signal intensity is then

$$I_s = I_0 n \ell \left(\frac{d^2\sigma}{d\Omega dE} \right) \Delta\Omega \Delta E$$

where ΔE represents a convolution of the energy spread of the source and the passband of the analyser.

-4-

B1.6.1.2 ELECTRON OPTICS

The two essential elements of an electron spectrometer are the electrodes that accelerate electrons and focus them into a beam and the dispersive elements that sort electrons according to their energies. These serve the functions of lenses and prisms in an optical spectrometer. The same parameters are used to describe these elements in an electron spectrometer as in an optical spectrometer; the technology is referred to as *electron optics*.

(A) ELECTRON LENSES

The typical electron-optical lens consists of a closely spaced pair of coaxial cylindrical tubes biased at different electrical potentials. The equipotential surfaces in the gap between the tubes assume shapes similar to those of optical lenses as illustrated in figure B1.6.1. An electron passing across these surfaces will be accelerated or decelerated, and its path will be curved to produce focusing. The main difference between the electron lens and an optical lens is that the quantity analogous to the refractive index, namely the electron velocity, varies continuously across an electrostatic lens, whereas a discontinuous change of refractive index occurs at the surface of an optical lens. Electron lenses are 'thick' lenses, meaning that their axial dimensions are comparable to their focal lengths. An important consequence is that the principal planes employed in ray tracing are separated from the midplane of the lens and lie to the low-velocity side of the lens gap, as shown in the figure. The design of these lenses is facilitated by tables of electron lens optical properties [1], and by computer programs that calculate the potential array for an arbitrary arrangement of electrodes, and trace electron trajectories through the resultant field [2]. In addition to cylindrical electrodes, electron lenses are sometimes created by closely spaced planar electrodes with circular apertures or slits. Shaped magnetic fields are also used to focus electrons, especially for electron energies much in excess of 10 keV where electrostatic focusing requires inconveniently high voltages.

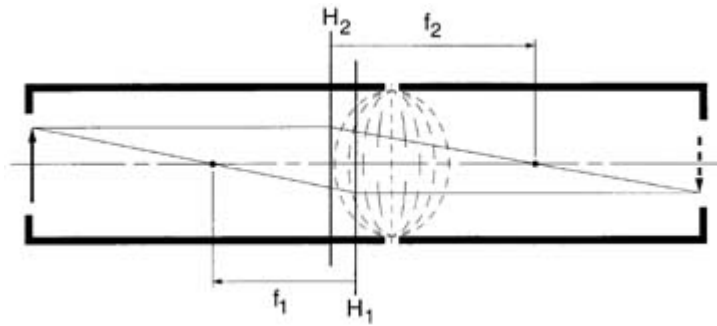


Figure B1.6.1 Equipotential surfaces have the shape of lenses in the field between two cylinders biased at different voltages. The focusing properties of the electron optical lens are specified by focal points located at focal lengths f_1 and f_2 , measured relative to the principal planes, H_1 and H_2 . The two principal rays emanating from an object on the left and focused to an image on the right are shown.

(B) ELECTRON ANALYSERS

An electron ‘prism’, known as an *analyser* or *monochromator*, is created by the field between the plates of a capacitor. The plates may be planar, simple curved, spherical, or toroidal as shown in [figure B1.6.2](#). The trajectory of an electron entering the gap between the plates is curved as the electron is attracted to the positively biased (inner) plate and

repelled by the negatively biased (outer) plate. The curvature of the trajectory is a function of the electron’s kinetic energy so that the electrons in a beam projected between the plates are dispersed in energy. These devices are not only dispersive, but focusing; electrons of the same energy originating from a point source are brought to a point on a focal plane at the output side of the analyser. The energy passband, or *resolution*, of an electrostatic analyser is the range of energies, ΔE , of electrons which, entering through a slit, are transmitted through the analyser to an exit slit. This quantity depends upon the width of the slits as well as the physical dimensions of the analyser. Fixing the slit widths and analyser dimension fixes the relative resolution, $\Delta E/E$, where E is the nominal energy of transmitted electrons. The *resolving power* of an analyser is specified as $E/\Delta E$, the inverse of the relative resolution. For each type of analyser there is a simple relation between analyser dimensions and resolution; for example, the analyser with hemispherical plates has relative resolution $\Delta E/E = w/2R$, where w is the diameter of the entrance and exit apertures and R is the mean radius of the plates.

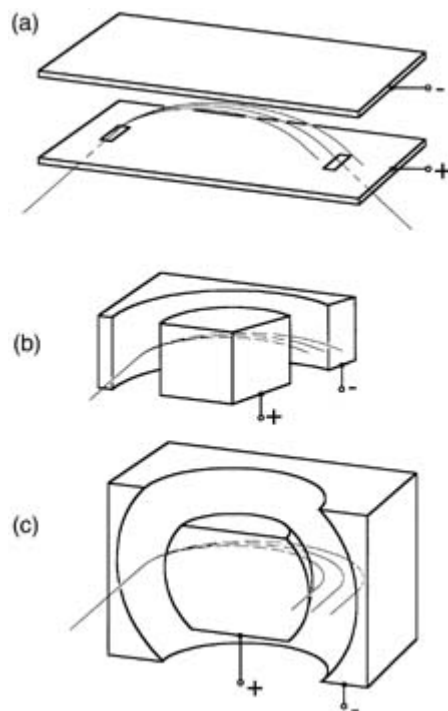


Figure B1.6.2 Electron analysers consisting of a pair of capacitor plates of various configurations: (a) the parallel-plate analyser, (b) the 127° cylindrical analyser and (c) the 180° spherical analyser. Trajectories for electrons of different energies are shown.

For an analyser of fixed dimensions, the absolute resolution, ΔE , can be improved, that is, made smaller, by reducing the pass energy, E . This is accomplished by decelerating the electrons to be analysed with a decelerating lens system at the input to the analyser. By this means, an absolute resolution as small as 2 meV has been achieved. For most practical analysers, $\Delta E/E$ is of the order of 0.01, and the pass energy in the highest-resolution spectrometers is of the order of 1 eV. The transmission of electrons of energy less than about 1 eV is generally not practical since unavoidable stray electric and magnetic fields produce unpredictable deflection of electrons of lower energies; even spectrometers of

modest resolution require magnetic shielding to reduce the magnetic field of the earth by two to three orders of magnitude.

Magnetic fields are employed in several electron-energy analysers and filters ([figure B1.6.3](#)). For very-low-energy electrons (0 to 10 eV), the ‘trochoidal analyser’ has proven quite useful. This device employs a magnetic field aligned to the direction of the incident electrons and an electric field perpendicular to this direction. The trajectory of an electron injected into this analyser describes a spiral and the guiding centre of the spiral drifts in the remaining perpendicular direction. The drift rate depends upon the electron energy so that a beam of electrons entering the device is dispersed in energy at the exit. The projection of the trajectory on a plane perpendicular to the electric field direction is a trochoid, hence the name trochoidal analyser. The Wien filter is similar in that it uses crossed electric and magnetic fields; however, the fields are perpendicular to one another and both are perpendicular to the injected electron beam direction. The Coulomb force induced by the electric field, E , deflects electrons in one direction and the Lorentz force associated with the magnetic field, \mathbf{B} , tends to deflect electrons in the opposite direction. The forces balance for one velocity, $v = |E|/|B|$, and electrons of this velocity are transmitted straight through the filter to the exit aperture. A magnetic field alone, perpendicular to the direction of an electron beam, will disperse electrons in energy. Sector magnets

such as those used in mass spectrometers are used in electron spectrometers for very-high-energy electrons, the advantage over electrostatic deflectors being that large electrical potentials are not required. Another advantage is that deflection is in the direction parallel to the magnet pole face making it possible to view the entire dispersed spectrum at one time. By contrast, energy dispersion of an electron beam in an electrostatic device results in a significant portion of the dispersed electrons striking one or the other electrode.

-7-

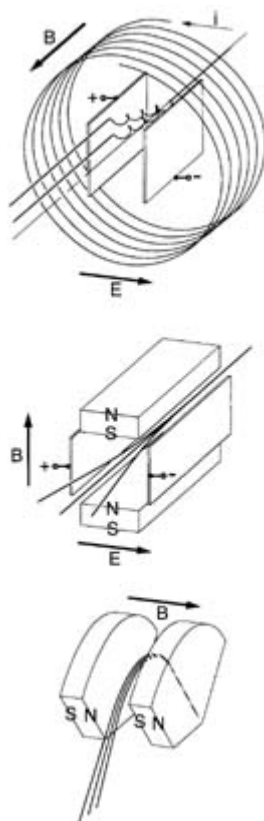


Figure B1.6.3 Electron energy analysers that use magnetic fields: (a) the trochoidal analyser employing an electromagnet, (b) the Wien filter and (c) the sector magnet analyser. Trajectories for electrons of different energies are shown.

(C) ELECTRON GUNS

Thermionic emission from an electrically heated filament is the usual source of electrons in an electron spectrometer. Field emission induced by a large electric-field gradient at a sharply pointed electrode may be used when fine spatial resolution is required. For very special applications, laser-induced photoemission has been used to produce nearly monoenergetic electrons and electrons with spin polarization. The filament in a thermionic source is a wire or ribbon of tungsten or some other refractory metal, sometimes coated or impregnated with thoria to reduce the work function. The passage of an electrical current of a few amps heats the filament to 1500 to 2500°C. As shown in [figure B1.6.4](#) the filament is typically mounted in a diode arrangement, protruding through a cathode and a small distance from an anode with an aperture through which electrons are extracted. The cathode is biased 30 to 50 V negative with respect to the source and the anode 50 to 100 V positive. An approximately Maxwellian energy distribution is produced. Depending on the filament temperature and the potential drop across the hot portion of the filament, this distribution is 0.3 to 0.7 eV wide. For most applications, the electron beam extracted from a thermionic source is passed through a monochromator to select a narrower band of energies from that emanating from the source.

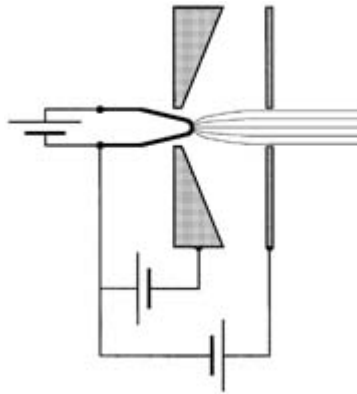


Figure B1.6.4 Diode electron source.

(D) ELECTRON SPECTROMETERS

A typical electron energy-loss spectrometer is shown in figure B1.6.5. The major components are an electron source, a premonochromator, a target, an analyser and an electron detector. For gaseous samples, the target may be a gas jet or the target may be a gas confined in a cell with small apertures for the incident beam and for the scattered electrons. The target may be a thin film to be viewed in transmission or a solid surface to be viewed in reflection. The analyser may be rotatable about the scattering centre so the angularly differential scattering cross section can be measured. Most often the detector is an electron multiplier that permits scattered electrons to be counted and facilitates digital processing of the scattered-electron spectrum. In low-resolution instruments, the scattered-electron intensity may be sufficient to be measured with a sensitive electrometer as an electron current captured in a 'Faraday cup'.

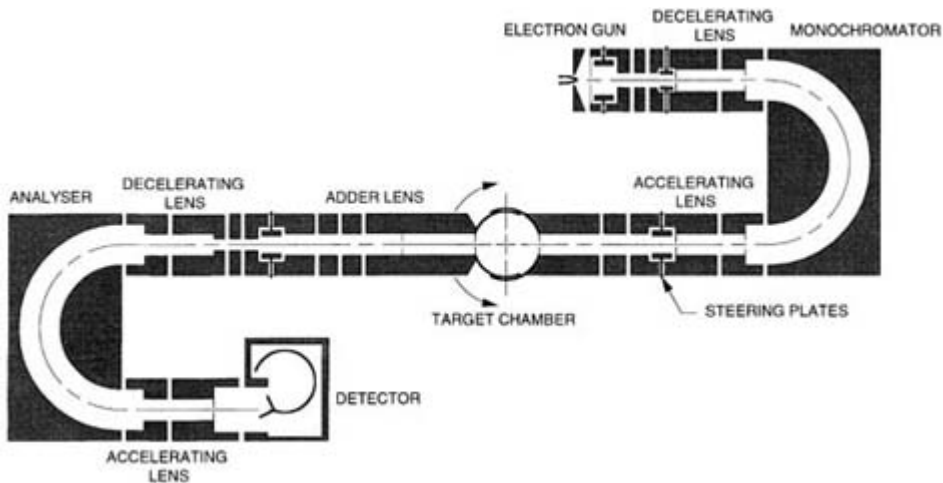


Figure B1.6.5 Typical electron energy-loss spectrometer.

Electron lens systems between each component serve a number of functions. A lens following the source focuses electrons on the entrance aperture of the premonochromator and decelerates these electrons to the pass energy required

to obtain the desired resolution. The lens following focuses electrons on the target and provides a variable amount of acceleration to permit experiments with different incident energies. The lens system at the input to the analyser again decelerates electrons and focuses on the entrance aperture of the analyser; however, this lens system has an additional function. The scattered-electron energy spectrum must be scanned. This is accomplished with an ‘adder lens’ that progressively adds back energy to the scattered electrons. The analyser is set to transmit electrons which have lost no energy; the energy-loss spectrum is then a plot of the detector signal as a function of the energy added by the adder lens. The alternative method of scanning is to vary the pass energy of the analyser; this has the disadvantage of changing the analyser resolution and transmission as the spectrum is scanned. The lens following the analyser accelerates transmitted electrons to an energy at which the detector is most sensitive, typically several hundred eV.

THEORY B1.6.2

Inelastic electron collisions can be roughly divided into two regimes: those in which the kinetic energy of the projectile electron greatly exceeds the energy of the target atom or molecule’s electrons excited by the collision, and those in which the projectile and target electron energies are comparable. In the higher-energy region the target electrons are little disturbed by the approach and departure of the projectile; the excitation occurs suddenly when the projectile is very close to the target. In the lower-energy region, the interaction proceeds on a time scale comparable to the orbital period of the target electrons; both projectile and target electrons make significant adjustments to one another’s presence. In some such cases, it may even make sense to consider the electron–target complex as a transient negative ion.

B1.6.2.1 BETHE–BORN THEORY FOR HIGH-ENERGY ELECTRON SCATTERING

Bethe provided the theoretical basis for understanding the scattering of fast electrons by atoms and molecules [3, 4]. We give below an outline of the quantum-mechanical approach to calculating the scattering cross section.

The Schrödinger equation for the projectile–target system is

$$\left[-\frac{\hbar^2}{2m} \nabla_r^2 - \frac{\hbar^2}{2m} \sum_j \nabla_{r_j}^2 + V(\mathbf{r}, \mathbf{r}_j, \mathbf{r}_N) \right] \phi(\mathbf{r}, \mathbf{r}_j) = \left[\varepsilon_0 + \frac{\hbar^2 k_0^2}{2m} \right] \phi(\mathbf{r}, \mathbf{r}_j) \quad \text{B1.6.1}$$

where r gives the position of the projectile electron and the r_j are the coordinates of the electrons in the target and the r_N are the coordinates of the nuclei; $V(\mathbf{r}, \mathbf{r}_j, \mathbf{r}_N)$ is the potential energy of interaction between the projectile and the particles (electrons and nuclei) that make up the target, as well as interactions between particles in the target; ε_0 is the energy of the target in its ground state; and $\hbar^2 k_0^2 / 2m$ is the initial kinetic energy of the projectile with wavevector k_0 and momentum $\hbar k_0$. The wave equation (B1.6.1) is inseparable because of terms in the potential-energy operator that go as $|\mathbf{r} - \mathbf{r}_j|^{-1}$. It is thus impossible to obtain an analytic solution for the wave function of the scattered electron. An approximate solution can be obtained by expanding the wave function, $\phi(\mathbf{r}, \mathbf{r}_j)$, in the complete set of eigenfunctions of the target, $\chi_m(\mathbf{r}_j)$, and of the projectile, $\psi(\mathbf{r})$. This separates the wave equation into a set of coupled differential equations each of which manifests a discrete interaction coupling two states (n and m) of the target. The interaction is

described by a matrix element:

$$V_{mn} = \int \chi_n^*(\mathbf{r}_j) V \chi_m(\mathbf{r}_j) d\mathbf{r}_j.$$

Approximate methods may be employed in solving this set of equations for the $\psi(r)$; however, the asymptotic form of the solutions are obvious. For the case of elastic scattering

$$\Psi_0(\mathbf{r}) \rightarrow e^{ik_0z} + \frac{e^{ik_0r}}{r} f_0(\theta, \phi)$$

the first term representing an incident plane wave moving in the z -direction in a spherical coordinate system and the second term an outgoing spherical wave modulated by a *scattering amplitude*, $f(\theta, \phi)$. For inelastic scattering, the solutions describe an outgoing wave with momentum $\hbar\mathbf{k}$,

$$\Psi(\mathbf{r}) \rightarrow \frac{e^{i\mathbf{k}\cdot\mathbf{r}}}{r} f(\theta, \phi).$$

In this case the projectile has imparted energy $\hbar^2(k_0^2 - k^2)/2m$ to the target. Assuming the target is initially in its ground state ($m = 0$), the collision has excited the target to a state of energy $E_n = \hbar^2(k_0^2 - k^2)/2m$.

The cross section for scattering into the differential solid angle $d\Omega$ centred in the direction (θ, ϕ) , is proportional to the square of the scattering amplitude:

$$\frac{d\sigma}{d\Omega} = \frac{k}{k_0} |f(\theta, \phi)|^2$$

where the ratio $k/k_0 = v/v_0$ accounts for the fact that, all other things being equal, the incident and scattered flux differ owing to the difference in velocity, v_0 , of the incident electron compared to the velocity, v , of the scattered electron. As a consequence of the expansion of the total wave function, the scattering amplitude can also be decomposed into terms each of which refers to an interaction coupling specific states of the target:

$$f_{mn}(\theta, \phi) = \frac{m}{2\pi\hbar^2} \int e^{-i\mathbf{k}\cdot\mathbf{r}} V_{mn} e^{i\mathbf{k}_0\cdot\mathbf{r}} d\mathbf{r} = \frac{m}{2\pi\hbar^2} \int V_{mn} e^{i\mathbf{K}\cdot\mathbf{r}} d\mathbf{r} \quad \text{B1.6.2}$$

where $\hbar\mathbf{k} = \hbar(\mathbf{k}_0 - \mathbf{k})$ is the momentum transfer in the collision.

In the high-energy regime it is appropriate to employ the Born approximation. There are three assumptions: (i) The incident wave is undistorted by the target. For a target in its ground state (specified by $m = 0$) this is equivalent to setting $V_{00} = 0$. (ii) There is no interaction between the outgoing electron and the excited target. For inelastic scattering with excitation of the final state n , this is equivalent to setting $V_{nn} = 0$. (iii) The excitation is a direct process with no involvement of intermediate states, thus $V_{mn} = 0$ unless $m = 0$. The scattering amplitude (equation (B1.6.2)) thus contains but one term:

$$f_{0n}(\theta, \phi) = \frac{m}{2\pi\hbar^2} \int e^{i\mathbf{K}\cdot\mathbf{r}} \chi_n^*(\mathbf{r}_j) V \chi_0(\mathbf{r}_j) d\mathbf{r}_j d\mathbf{r}. \quad \text{B1.6.3}$$

When the potential consists of electron–electron and electron–nucleus Coulombic interactions,

$$V = \sum_j \frac{e^2}{|\mathbf{r} - \mathbf{r}_j|} - \sum_N \frac{Z_N e^2}{|\mathbf{r} - \mathbf{r}_N|}$$

substitution in (B1.6.3) yields

$$f_{0n}(\theta, \phi) = \frac{me^2}{2\pi\hbar^2} \int \chi_n^*(\mathbf{r}_j) \sum_j \frac{e^{i\mathbf{K}\cdot\mathbf{r}}}{|\mathbf{r} - \mathbf{r}_j|} \chi_0(\mathbf{r}_j) d\mathbf{r}_j d\mathbf{r}$$

the electron–nucleus terms having been lost owing to the orthonormality of the target wave functions. The important physical implication of the Born approximation becomes clear if one first performs the integration with respect to \mathbf{r} , taking advantage of the transformation

$$\int \frac{e^{i\mathbf{K}\cdot\mathbf{r}}}{|\mathbf{r} - \mathbf{r}_j|} d\mathbf{r} = \frac{4\pi}{K^2} e^{i\mathbf{K}\cdot\mathbf{r}_j}.$$

The differential cross section for inelastic collisions exciting the n th state of the target then takes the form

$$\left(\frac{d\sigma}{d\Omega}\right)_{0n} = \frac{4e^2 m^2 k}{\hbar^4 K^4 k_0} \left| \int \chi_n^*(\mathbf{r}_j) \sum_j e^{i\mathbf{K}\cdot\mathbf{r}_j} \chi_0(\mathbf{r}_j) d\mathbf{r}_j \right|^2 = \frac{4e^4 m^2 k}{\hbar^4 K^4 k_0} |\varepsilon_n(\mathbf{K})|^2. \quad \text{B1.6.4}$$

In this expression, factors that describe the incident and scattered projectile are separated from the square modulus of an integral that describes the role of the target in determining the differential cross section. The term preceding the

integral, $4e^4 m^2 / \hbar^4 K^4$, with units of area, is the Rutherford cross section for electron–electron scattering. The integral, represented by the quantity $\varepsilon_n(\mathbf{K})$, is known as the *inelastic scattering form factor*.

In the discussion above, scattering from molecules is treated as a superposition of noninteracting electron

waves scattered from each atomic centre. In fact, there is a weak but observable interference between these waves giving rise to phase shifts associated with the different positions of the atoms in a molecule. This diffraction phenomenon produces oscillations in the differential cross section from which molecular structure information can be derived.

The interaction of the target with the incoming and outgoing electron wave must be considered at lower impact energies. This is achieved in the *distorted-wave approximation* by including V_{00} and V_{nn} in the calculation of the scattering amplitude. Higher-level calculations must also account for electron spin since spin exchange becomes important as the collision energy decreases.

B1.6.2.3 THE GENERALIZED OSCILLATOR STRENGTH

The Born approximation for the differential cross section provides the basis for the interpretation of many experimental observations. The discussion is often couched in terms of the *generalized oscillator strength*,

$$f_n(\mathbf{K}) = \frac{2m}{\hbar^2} \frac{E_n}{K^2} \left| \int \chi_n^*(\mathbf{r}_j) \sum_j e^{i\mathbf{K}\cdot\mathbf{r}_j} \chi_0(\mathbf{r}_j) d\mathbf{r}_j \right|^2 = \frac{2m}{\hbar^2} \frac{E_n}{K^2} |\varepsilon_n(\mathbf{K})|^2. \quad \text{B1.6.5}$$

Assuming the validity of the Born approximation, an ‘effective’ generalized oscillator strength can be derived in terms of experimentally accessible quantities:

$$f_n(\mathbf{K}) = \frac{\hbar^2}{2e^4 m} \frac{k_0}{k} E_n K^2 \left(\frac{d\sigma}{d\Omega} \right)_{0n}. \quad \text{B1.6.6}$$

All the quantities on the right can be measured (k_0 , k and K calculated from measurements of the incident energy, the energy-loss, and the scattering angle). For inelastic collisions resulting in transitions into the continuum beyond the first ionization potential, the cross section is measured per unit energy loss and the generalized oscillator strength *density* is determined:

$$\frac{df(\mathbf{K})}{dE} = \frac{\hbar^2}{2e^4 m} \frac{k_0}{k} E_n K^2 \frac{d^2\sigma}{d\Omega dE}. \quad \text{B1.6.7}$$

The generalized oscillator strength provides the basis of comparison between electron energy-loss spectra and optical spectra; however, there is a problem with the determination of the absolute value of the generalized oscillator strength since measurements of the differential cross section can rarely be made on an absolute basis owing to the difficulty of accurately determining the target dimension and density. The problem is overcome by a normalization of

experimentally determined generalized oscillator strengths according to the *Bethe sum rule* which requires that the sum of the $f_n(\mathbf{K})$ (equation B1.6.6) for all discrete transitions plus the integral of $df(\mathbf{K})/dE$ (equation B1.6.7) over the continuum adds up to the number of electrons in the target.

A particularly important property of the generalized oscillator strength is that, for high-energy, small-angle

scattering, the generalized oscillator strength is approximately equal to the *optical oscillator strength*, f_n^{opt} , for electric-dipole transitions induced by photon absorption. That this is so can be seen from a power-series expansion of the form factor that appears in the expression for the generalized oscillator strength ([equation B1.6.5](#)):

$$\varepsilon_n(\mathbf{K}) = \sum_{k=1}^{\infty} \int \chi_n^*(\mathbf{r}_j) \sum_j \frac{(i\mathbf{K}\cdot\mathbf{r}_j)^k}{k!} \chi_0(\mathbf{r}_j) d\mathbf{r}_j.$$

The operator in the first term goes as r_j , and is thus proportional to the optical *dipole transition moment*

$$M_{0n} = \int \chi_n^*(\mathbf{r}_j) \sum_j e\mathbf{r}_j \chi_0(\mathbf{r}_j) d\mathbf{r}_j.$$

The second term is proportional to the optical quadrupole transition moment, and so on. For small values of momentum transfer, only the first term is significant, thus

$$\lim_{k \rightarrow 0} f_n(K) = f_n^{opt}.$$

The result is that the small-angle scattering intensity as a function of energy loss (the energy-loss spectrum) looks like the optical absorption spectrum. In fact, oscillator strengths are frequently more accurately and conveniently measured from electron-impact energy-loss spectra than from optical spectra, especially for higher-energy transitions, since the source intensity, transmission and detector sensitivity for an electron scattering spectrometer are much more nearly constant than in an optical spectrometer. The proportionality of the generalized oscillator strength and the optical dipole oscillator strength appears to be valid even for incident-electron energies as low as perhaps 200 eV, but it is strictly limited to small-angle, forward scattering that minimizes momentum transfer in a collision [5]. Of course $K = 0$ is inaccessible in an inelastic collision; there must be at least sufficient momentum transferred to account for the kinetic energy lost by the projectile in exciting the target. For very-high-energy, small-angle scattering, the minimum momentum transfer is relatively small and can be ignored. At lower energies, an extrapolation technique has been employed in very accurate work.

On the other hand, there are unique advantages to electron-scattering measurements in the lower-energy, larger-scattering-angle regime in which the momentum transfer is larger than the minimum required to induce a transition. In this case, higher-order multipoles in the transition moment become significant, with the result that the cross section for the excitation of optically forbidden transitions increases relative to that for dipole-allowed transitions. The ability to vary the momentum transfer in electron-energy-loss spectroscopy yields a spectrum much richer than the optical spectrum.

B1.6.2.4 THE BETHE SURFACE: BINARY VERSUS DIPOLE COLLISIONS

An important feature of electron-impact spectroscopy in comparison to photoabsorption is that the momentum transfer can be varied. As the scattering angle increases and the incident energy decreases, higher-order terms in the expansion of $\varepsilon_n(\mathbf{K})$ become relatively more important. Large-angle, high-momentum-transfer scattering results from small impact parameters. In this case the target experiences a nonuniform electric field as the electron passes by. Significant amplitude of higher-order multipoles in a nonuniform field permits the field to

couple with higher-order multipoles of the target. Thus, for example, optically forbidden electric-quadrupole transitions are a significant feature of the low-energy, large-scattering-angle electron energy-loss spectrum. Selection rules for electronic excitation by electron impact have been treated in detail in the review by Hall and Read [6].

From the discussion above it appears that small-angle scattering events might better not be thought of as collisions at all. The excitation, which is photon-like, appears to be a consequence of the high-frequency electric pulse delivered to the target as the projectile electron passes rapidly by. These energetic, but glancing collisions are referred to as *dipole collisions*, in contradistinction to the larger-angle scattering regime of *binary collisions* where the projectile electron appears to undergo a hard collision with one of the target electrons.

A succinct picture of the nature of high-energy electron scattering is provided by the *Bethe surface* [4], a three-dimensional plot of the generalized oscillator strength as a function of the logarithm of the square of the momentum transfer, (K^2) and the energy-loss, E_n . To see how this works, consider the form of the Bethe surface for a 3D billiards game with a ‘projectile’ cue ball incident on a stationary billiard ball. This is a two-body problem so the energy-loss is uniquely determined by the momentum transfer: $E_n = \hbar^2 K^2 / 2m$. For each value of K , all the oscillator strength appears at a single value of E_n . The Bethe surface displays a sharp ridge extending from low values of K^2 and E_n (corresponding to large-impact-parameter, glancing collisions) to high values (corresponding to near ‘head on’ collisions).

A schematic of the Bethe surface for electron scattering from an atom or molecule is shown in [figure B1.6.6](#) superimposed on the surface for the two-body problem with a stationary target. The sharp ridge is broadened in the region of large momentum transfer and energy loss. These are hard collisions in which the projectile ejects an electron from the target. This is at least a three-body problem: because the recoil momentum of the ionic core is not accounted for, or, alternatively, since the target electron is not stationary, K and E_n are not uniquely related. The breadth and the shape of the ridge in the Bethe surface in the high-momentum region are a reflection of the momentum distribution of the target electron. Electron Compton scattering experiments [7, 8] and (e, 2e) experiments [9] are carried out in the high-momentum-transfer, large-energy-loss, region for the purpose of investigating the electron momentum distribution in atoms and molecules.

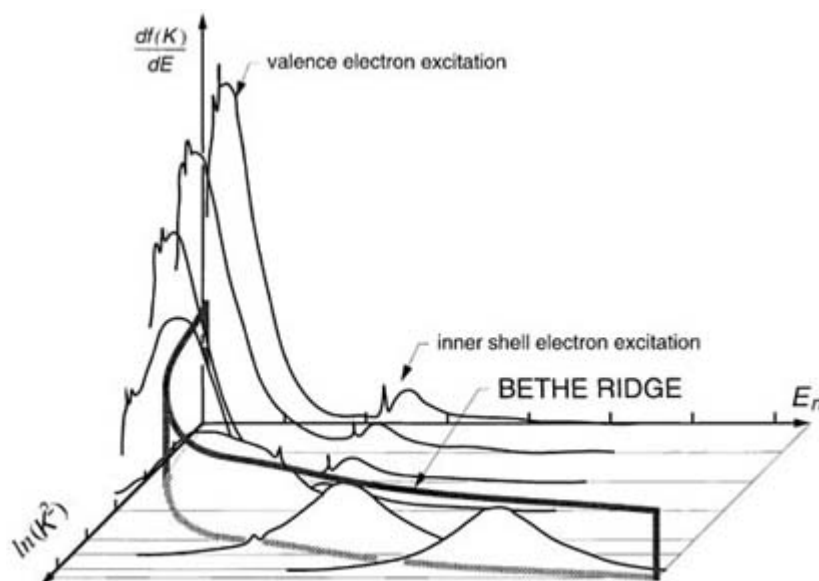


Figure B1.6.6 The Bethe surface. The sharp ridge corresponds to scattering from a single stationary target electron; the broadened ridge to scattering from the electrons in an atom or molecule.

The dipole region is approached as one proceeds to the low-momentum-transfer portion of the Bethe surface. For sufficiently small momentum transfer, where $2m|\hbar\mathbf{K}|^2$ approaches the binding energy of valence electrons in the target, a section through the Bethe surface has the appearance of the optical absorption spectrum. For values of the energy loss less than the first ionization energy of the target, sharp structure appears corresponding to the excitation of discrete, dipole-allowed transitions in the target. The ionization continuum extends from the first ionization potential, but, as in the photoabsorption spectrum, one typically sees sharp ‘edges’ as successive ionization channels become energetically possible. Also, as in the photoabsorption spectrum, resonance structures may appear corresponding to the excitation of metastable states imbedded in the continuum.

B1.6.2.5 LOW-ENERGY ELECTRON SCATTERING

Theoretically, the asymptotic form of the solution for the electron wave function is the same for low-energy projectiles as it is at high energy; however, one must account for the protracted period of interaction between projectile and target at the intermediate stages of the process. The usual procedure is to separate the incident-electron wave function into *partial waves*

$$e^{i\mathbf{k}_0 \cdot \mathbf{r}} = \sum_{l=0}^{\infty} (2l+1) i^l P_l(\cos \theta) j_l(k_0 r)$$

(j_l a spherical Bessel function; P_l a Legendre polynomial) and expand the wave function of the target in some complete basis set, typically the complete set of eigenfunctions for the unperturbed target. This approach allows for any distortion of the incoming and scattered electron waves, as well as any perturbation of the target caused by the

approaching charge of the projectile. Each term in the expansion for the incident wave represents an angular momentum component: $l=0$, the s-wave component; $l=1$, the p-wave component; and so on. The scattering of the incident wave is treated component by component, the interaction with each giving rise to a phase shift in that component of the outgoing wave.

To see how this works, consider elastic scattering in a situation where the electron–target interaction can be described by a simple central-force-field potential, $V(r)$, that does not fall off faster than r^{-1} at large r . In this case the wave equation for the projectile electron can be separated from the Schrödinger equation for the total electron–target system given above (equation (B1.6.1)). The wave equation for the projectile is

$$\left[-\frac{\hbar^2}{2m} \nabla_r^2 + V(r) \right] \Psi(\mathbf{r}) = \frac{\hbar^2 k_0^2}{2m} \Psi(\mathbf{r}).$$

Since the potential depends only upon the scalar r , this equation, in spherical coordinates, can be separated into two equations, one depending only on r and one depending on θ and ϕ . The wave equation for the r -dependent part of the solution, $R(r)$, is

$$\left[-\frac{\hbar^2}{2m} \frac{1}{r} \frac{\partial^2}{\partial r^2} r + \frac{\hbar^2}{2m} \frac{l(l+1)}{r^2} + V(r) \right] R(r) = \frac{\hbar^2 k_0^2}{2m} R(r) \quad \text{B1.6.8}$$

where $\hbar\sqrt{l(l+1)}/r$ is the orbital angular momentum associated with the l th partial wave. The solutions have the asymptotic form

$$R(r) \rightarrow \frac{1}{k_0 r} \sin \left(k_0 r - \frac{l\pi}{2} + \eta_l \right) \quad \text{B1.6.9}$$

and the calculation of the cross section is reduced to calculating the phase shift, η_l , for each partial wave. The phase shift is a measure of the strength of the interaction of a partial wave in the field of the target, as well as a measure of the time period of the interaction. High-angular-momentum components correspond to large impact parameters for which the interaction can generally be expected to be relatively weak. The exceptions are for the cases of long-range potentials, as when treating scattering from highly polarizable targets or from molecules with large dipole moments. In any event, only a limited number of partial waves need be considered in calculating the cross section—sometimes only one or two.

B1.6.2.6 RESONANCES

The partial wave decomposition of the incident-electron wave provides the basis of an especially appealing picture of strong, low-energy resonant scattering wherein the projectile electron spends a sufficient period of time in the vicinity

-17-

of the target that the electron–target complex is describable as a temporary negative ion. With the radial wave equation (B1.6.8) for the projectile in a central-force field as a starting point, define a fictitious potential

$$V'(r) = \frac{\hbar^2}{2m} \frac{l(l+1)}{r^2} + V(r).$$

This is a fictitious potential because it includes not only the true potential, $V(r)$, that contains the screened Coulomb potential and the polarization potential of the target, but also the term $(\hbar^2/2m)l(l+1)/r^2$ that arises from the centrifugal force acting on the projectile, a fictitious force associated with curvilinear motion. As shown in figure B1.6.7 this repulsive term may give rise to an *angular momentum barrier*. Some part of the incident-electron-wave amplitude may tunnel through the barrier to impinge upon the repulsive part of the true potential from which it is reflected to tunnel back out to join the incident wave. The superposition of these two waves produces the phase shift in the scattered wave (see equation (B1.6.9)). More interestingly, as shown in the figure, there are special incident electron energies for which the width of the well behind the barrier is equal to some integral multiple of the electron wavelength. A standing wave then persists, corresponding to an electron being temporarily trapped in the field of the target. This model describes the resonant formation of a metastable negative ion, or, more simply, a ‘resonance’.

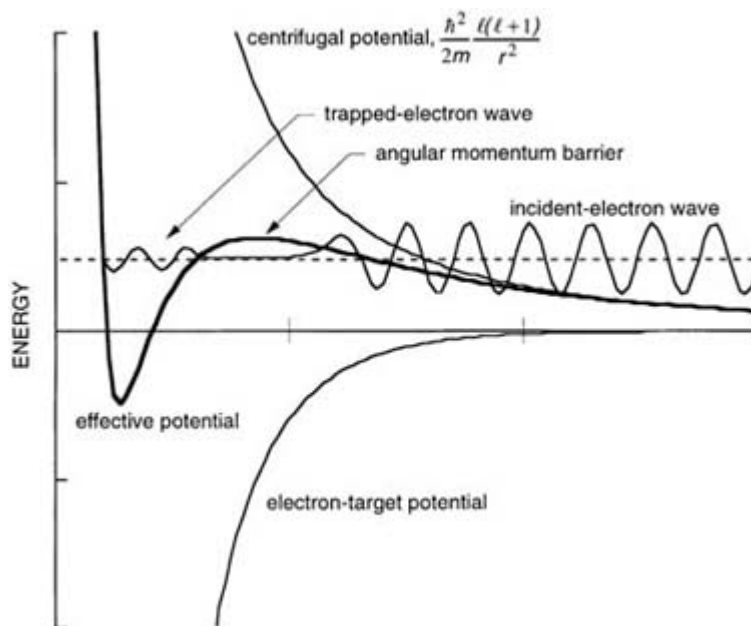


Figure B1.6.7 An angular momentum barrier created by the addition of the centrifugal potential to the electron–atom potential.

Finally, it must be recognized that all the above discussion assumes an isolated atomic or molecular target. To describe electron scattering in a complex target such as a solid, one must consider the extended nature of the valence-electron

density that constitutes a kind of electron gas enveloping the array of positively charged atomic cores. Most calculations employ a so-called ‘jellium’ model in which the mobile electrons move in the field of a positive charge smeared out into a homogeneous neutralizing density.

B1.6.3 APPLICATIONS

Electron energy-loss spectroscopy is used for obtaining spectroscopic data as a convenient substitute for optical spectroscopy, and, taking advantage of differences in selection rules, as an adjunct to optical spectroscopy. In addition, electron spectroscopy has many applications to chemical and structural analysis of samples in the gas phase, in the solid phase, and at the solid–gas interface.

B1.6.3.1 VALENCE-SHELL-ELECTRON SPECTROSCOPY

Electronic transitions within the valence shell of atoms and molecules appear in the energy-loss spectrum from a few electron volts up to, and somewhat beyond, the first ionization energy. Valence-shell electron spectroscopy employs incident electron energies from the threshold required for excitation up to many kiloelectron volts. The energy resolution is usually sufficient to observe vibrational structure within the Franck–Condon envelope of an electronic transition. The sample in valence-shell electron energy-loss spectroscopy is most often in the gas phase at a sufficiently low pressure to avoid multiple scattering of the projectile electrons, typically about 10^{-3} mbar. Recently, electronic excitation in surface adsorbates has been observed in the energy-loss spectrum of electrons reflected from metallic substrates. When the measurements

are carried out with relatively high incident-electron energies (many times the excitation and ionization energies of the target electrons) and with the scattered electrons detected in the forward direction (0° scattering angle), the energy-loss spectrum is essentially identical to the optical spectrum. As described above, this is the arrangement employed to determine oscillator strengths since forward scattering corresponds to collisions with the lowest momentum transfer. If the incident energy is reduced to about 100 eV (just a few times the target electron energies), symmetry-forbidden transitions can be uncovered and distinguished from optically allowed transitions by measuring the energy-loss spectrum at different scattering angles. An example is shown in [figure B1.6.8](#). As the incident energy approaches threshold, it becomes possible to detect electron-spin-changing transitions.

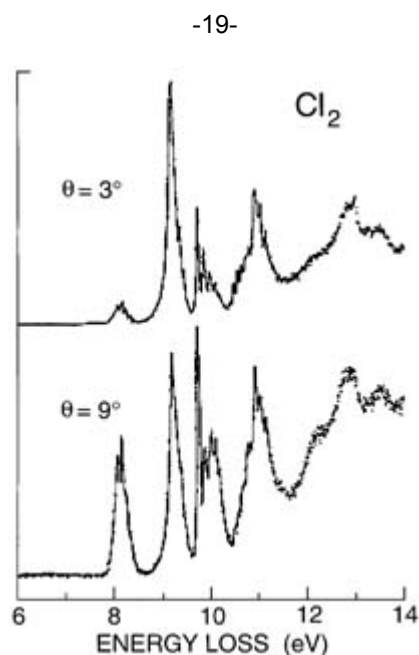


Figure B1.6.8 Energy-loss spectra of 200 eV electrons scattered from chlorine at scattering angles of 3° and 9° [10]. Optically forbidden transitions are responsible for the intensity in the 9° spectrum that does not appear in the 3° spectrum.

B1.6.3.2 INNER-SHELL-ELECTRON ENERGY-LOSS SPECTROSCOPY

Inner-shell-electron energy-loss spectroscopy (ISEELS) refers to measurements of the energy lost by projectile electrons that have promoted inner-shell electrons into unfilled valence orbitals or into the ionization continuum beyond the valence shell. Inner-shell excitation and ionization energies fall in the region between 100 eV and several kiloelectron volts. The corresponding features in the energy-loss spectrum tend to be broad and diffuse. Inner-shell-hole states of neutral atoms and molecules are very short lived; the transition energies to these states are correspondingly uncertain. The energy of a transition into the continuum is completely variable. Inner-shell-electron ionization energies cannot be uniquely determined from a measurement of the energy lost by the scattered electron since an unknown amount of energy is carried away by the undetected ejected electron. To be more precise, the electron-impact ionization cross section depends upon the energy (E_s) and angle (Ω_s) of the scattered electron as well as the energy (E_e) and angle (Ω_e) of the ejected electron; it is a fourfold differential cross section: $d^4\sigma/d\Omega_s dE_s d\Omega_e dE_e$. The intensity in the inner-shell energy-loss spectrum is proportional to this cross section integrated over E_e and Ω_e . For collisions in which the scattered electron is detected in the forward direction, the momentum transfer to the target is small and it is highly probable that the ejected-electron energy is very small. As a consequence, the ionization cross section for forward scattering is largest at the energy-loss threshold for ionization of each inner-shell electron and decreases monotonically, approximately as the inverse square of the energy loss. The basic appearance of

each feature in the inner-shell electron energy-loss spectrum is that of a sawtooth that rises sharply at the low-energy 'edge' and falls slowly over many tens of electron volts. In addition, sharp structures may appear near the edge (figure B1.6.9).

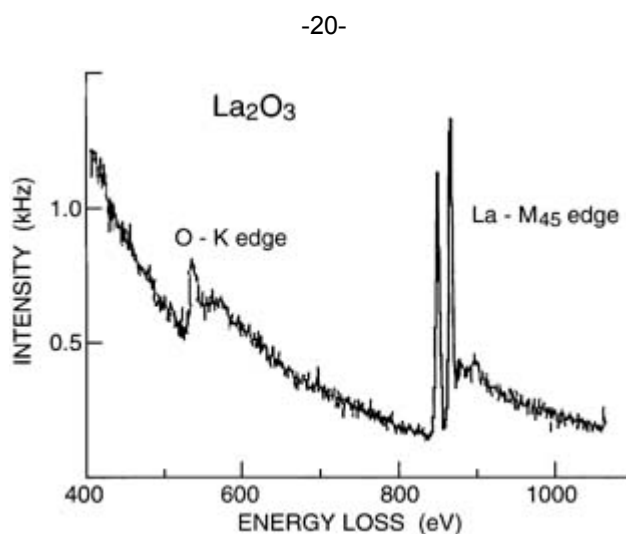


Figure B1.6.9 Energy-loss spectrum of La_2O_3 showing O K and La M_{45} ionization edges with prominent 'white line' resonances at the La edge [11].

In contrast to the broadly distributed valence-shell electron density in molecules and solids, inner-shell electron density is localized on a single atom. Molecular configuration and solid-state crystal structure have little effect upon inner-shell ionization energies. Each ionization edge in the energy-loss spectrum is characteristic of a particular type of atom; consequently, ISEELS has become an important technique for qualitative and quantitative elemental analysis. An especially useful elaboration of this technique is carried out in the transmission electron microscope where energy-loss analysis of electrons passing through a thin sample makes elemental analysis possible with the spatial resolution of the microscope [12]. A precision of the order of 100 ppm with nanometre spatial resolution has been achieved; this is close to single-atom sensitivity. A difficulty with ISEELS for analytical purposes is that each ionization edge is superimposed on the continuum associated with lower-energy ionizations of all the atoms in a sample. This continuum comprises an intense background signal which must be subtracted if the intensity at a characteristic edge is to be used as a quantitative measure of the concentration of a particular element in a sample.

Electrons arising from near-threshold inner-shell ionization have very little kinetic energy as they pass through the valence shell and may become trapped behind an angular momentum barrier in the exterior atomic potential, much as do low-energy incident electrons. This phenomenon, a wave-mechanical resonance as described above (section B1.6.2.6), gives rise to structure in the vicinity of an ionization edge in the energy-loss spectrum. For isolated molecular targets in the gas phase, the energies of near-edge resonances (relative to threshold) can be correlated with the eigenenergies of low-lying, unoccupied molecular orbitals (relative to the first ionization energy). Resonances in solid-state targets are especially prominent for fourth- and fifth-period elements where sharp threshold peaks known as 'white lines' are associated with electrons being trapped in vacant d and f bands (see figure B1.6.9). Resonance peaks have the effect of concentrating transition intensity into a narrow band of energies, thereby increasing the analytical sensitivity for these elements. Near-edge structure, being essentially a valence-shell or valence-band phenomenon, can provide important spectroscopic information about the chemical environment of the atoms in a sample.

B1.6.3.3 REFLECTED-ELECTRON ENERGY-LOSS SPECTROSCOPY

Vibrational spectroscopy of atoms and molecules near or on the surface of a solid has become an essential tool for the microscopic description of surface processes such as catalysis and corrosion. The effect of a surface on bonding is sensitively reflected by the frequencies of vibrational motions. Furthermore, since vibrational selection rules are determined by molecular symmetry that in turn is profoundly modified by the presence of a surface, it is frequently possible to describe with great accuracy the orientation of molecular adsorbates and the symmetry of absorption sites from a comparison of spectral intensities for surface-bound molecules to those for free molecules [13]. Electron spectroscopy has an advantage over optical methods for studying surfaces since electrons with energies up to several hundred electron volts penetrate only one or two atomic layers in a solid before being reflected, while the dimension probed by photons is of the order of the wavelength. Reflected-electron energy-loss spectroscopy (REELS) applied to the study of vibrational motion on surfaces represents the most highly developed technology of electron spectroscopy [14]. Incident electron energies are typically between 1 and 10 eV and sensitivity as low as a few per cent of a monolayer is routinely achieved (figure B1.6.10).

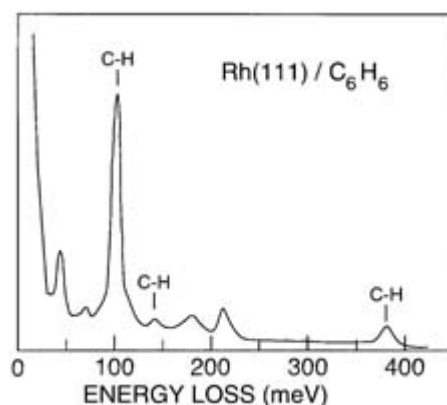


Figure B1.6.10 Energy-loss spectrum of 3.5 eV electrons specularly reflected from benzene adsorbed on the rhenium(111) surface [15]. Excitation of C–H vibrational modes appears at 100, 140 and 372 meV. Only modes with a changing electric dipole perpendicular to the surface are allowed for excitation in specular reflection. The great intensity of the out-of-plane C–H bending mode at 100 meV confirms that the plane of the molecule is parallel to the metal surface. Transitions at 43, 68 and 176 meV are associated with Rh–C and C–C vibrations.

B1.6.3.4 ELECTRON TRANSMISSION SPECTROSCOPY

An important feature of low-energy electron scattering is the formation of temporary negative ions by the resonant capture of incident electrons (see B1.6.2.6, above). These processes lead to sharp enhancements of the elastic-scattering cross section and often dominate the behaviour of the cross section for inelastic processes with thresholds lying close to the energy of a resonance [16]. Elastic-electron-scattering resonances are observed by electron transmission spectroscopy. Two types of resonance are distinguished: shape resonances and Feshbach resonances. *Shape resonances* arise when an electron is temporarily trapped in a well created in the ‘shape’ of the electron–target potential by a centrifugal barrier (figure B1.6.7). *Feshbach resonances* involve the simultaneous trapping of the projectile

π -electron systems. In order to emphasize the abrupt change in cross section characteristic of a resonance, the spectra are presented as the first derivative of the transmitted current as a function of incident-electron energy. This is accomplished by modulating the incident-electron energy and detecting the modulated component of the transmitted current. An example is shown in figure B1.6.11. Feshbach resonances fall in the 0 to about 30 eV range. They are found in electron scattering from atoms, but rarely for molecules. The study of these resonances has contributed to the understanding of optically inaccessible excited states of atoms and ions.

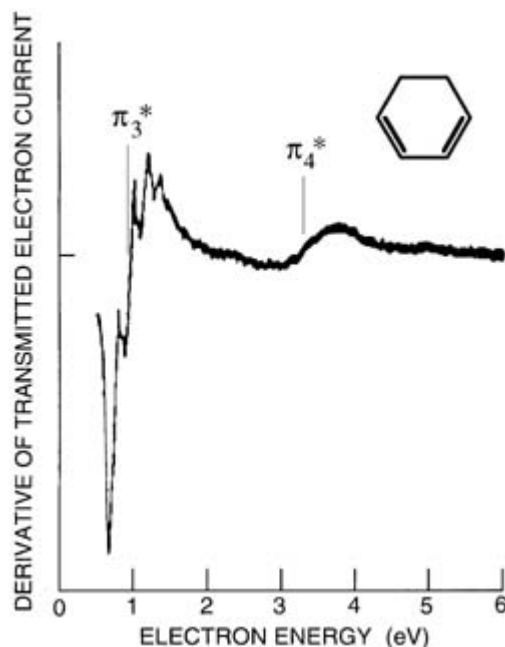


Figure B1.6.11 Electron transmission spectrum of 1,3-cyclohexadiene presented as the derivative of transmitted electron current as a function of the incident electron energy [17]. The prominent resonances correspond to electron capture into the two unoccupied, antibonding π^* -orbitals. The π_3^* negative ion state is sufficiently long lived that discrete vibronic components can be resolved.

B1.6.3.5 DIPOLE (E, 2E) SPECTROSCOPY

The information from energy-loss measurements of transitions into the continuum, that is, ionizing excitations, is significantly diminished because the energy of the ionized electron is not known. The problem can be overcome by

measuring simultaneously the energies of the scattered and ejected electrons. This is known as the (e, 2e) technique—the nomenclature is borrowed from nuclear physics to refer to a reaction with one free electron in the initial state and two in the final state. For spectroscopic purposes the experiment is carried out in the dipole scattering regime (see section B1.6.2.4). Two analyser/detector systems are used: one in the forward direction detects fast scattered electrons and the second detects slow electrons ejected at a large angle to the incident-electron direction (typically at the ‘magic angle’ of 54.7°). In order to ensure that pairs of electrons originate from the same ionizing collision, the electronics are arranged to record only those events in which a scattered and ejected electron are detected in coincidence (see B.1.11, ‘coincidence techniques’). The ionization energy, or binding energy, is unambiguously given by the difference between the incident electron energy and the sum of the energies of the scattered and ejected electrons detected in coincidence. Dipole (e, 2e) spectra (figure B1.6.12) are analogous to photoabsorption or photoelectron spectra obtained with tunable

UV or x-ray sources.

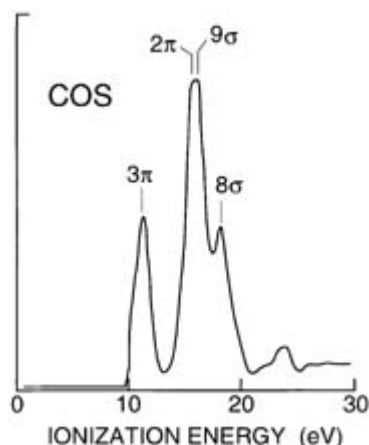


Figure B1.6.12 Ionization-energy spectrum of carbonyl sulphide obtained by dipole ($e, 2e$) spectroscopy [18]. The incident-electron energy was 3.5 keV, the scattered incident electron was detected in the forward direction and the ejected (ionized) electron detected in coincidence at 54.7° (angular anisotropies cancel at this ‘magic angle’). The energy of the two outgoing electrons was scanned keeping the net energy loss fixed at 40 eV so that the spectrum is essentially identical to the 40 eV photoabsorption spectrum. Peaks are identified with ionization of valence electrons from the indicated molecular orbitals.

REFERENCES

- [1] Harting E and Read F H 1976 *Electrostatic Lenses* (Amsterdam: Elsevier)
- [2] MacSimion C, McGilvery D C and Morrison R J S Montech Pty. Ltd., Monash University, Clayton, Victoria 3168, Australia Simion 3D version 6, Dahl D A, ms 2208, Idaho National Engineering Laboratory, PO Box 1625, Idaho Falls, ID 83415, USA Simion 3D version 6.0 for Windows, Princeton Electronic Systems, Inc, PO Box 8627, Princeton, NJ 08543, USA CPO-3D, RB Consultants Ltd, c/o Integrated Sensors Ltd, PO Box 88, Sackville Street, Manchester M60 1QD, UK. Fax: (UK)-61-200-4781.
- [3] Bethe H 1930 AP **5** 325

- [4] Inokuti M 1971 *Rev. Mod. Phys.* **43** 297
- [5] Lassetre E N and Skerbele A and Dillon M A 1969 *J. Chem. Phys.* **50** 1829 and references therein to other work of Lassetre
- [6] Hall R I and Read F H 1984 *Electron–Molecule Collisions* ed I Shimamura and K Takayanagi (New York: Plenum)
- [7] Williams B (ed) 1977 *Compton Scattering* (New York: McGraw-Hill)
- [8] Bonham R A and Fink M 1974 *High Energy Electron Scattering (ACS Monograph 169)* (New York: Van Nostrand Reinhold) ch 5
- [9] Coplan M A, Moore J H and Doering J P 1994 *Rev. Mod. Phys.* **66** 985
- [10] Spence D, Huebner R H, Tanaka H, Dillon M A and Wang R-G 1984 *J. Chem. Phys.* **80** 2989
- [11] Manoubi T, Colliex C and Rez P 1990 *J. Electron. Spectros. Relat. Phenom.* **50** 1
- [12] Leapman R D and Newbury D E 1993 *Anal. Chem.* **65** 2409

- [13] Richardson N V and Bradshaw A M 1981 *Electron Spectroscopy: Theory, Techniques and Applications* vol 4, ed C R Brundle and A D Baker (London: Academic)
- [14] Ibach H and Mills D L 1982 *Electron Energy Loss Spectroscopy and Surface Vibrations* (New York: Academic) Ibach H 1991 *Electron Energy Loss Spectrometers: the Technology of High Performance* (Berlin: Springer)
- [15] Koel B E and Somorjai G A 1983 *J. Electron. Spectrosc. Relat. Phenom.* **29** 287
- [16] Schulz G J 1973 *Rev. Mod. Phys.* **45** 378 Schulz G J 1973 *Rev. Mod. Phys.* **45** 423
- [17] Giordan J C, McMillan M R, Moore J H and Staley S W 1980 *J. Am. Chem. Soc.* **102** 4870
- [18] Cook J P D, White M G, Brion C E, Schirmer J, Cederbaum L S and Von Niessen W 1981 *J. Electron Spectrosc. Relat. Phenom.* **22** 261
-

FURTHER READING

Egerton R F 1986 *Electron Energy-Loss Spectroscopy in the Electron Microscope* (New York: Plenum)

This text covers quantitative analysis by electron energy-loss spectroscopy in the electron microscope along with instrumentation and applicable electron-scattering theory.

Joy D C 1986 The basic principles of EELS *Principles of Analytical Electron Microscopy* ed D C Joy, A D Romig Jr and J I Goldstein (New York: Plenum)

Good 20-page synopsis.

Moore J H, Davis C C and Coplan M A 1989 *Building Scientific Apparatus* 2nd edn (Redwood City, CA: Addison-Wesley) ch 5

The fundamentals of electron optical design.

-1-

B1.7 Mass spectrometry

Paul M Mayer

B1.7.1 INTRODUCTION

Mass spectrometry is one of the most versatile methods discussed in this encyclopedia. Ask a chemist involved in synthesis about mass spectrometry and they will answer that it is one of their most useful tools for identifying reaction products. An analytical chemist will indicate that mass spectrometry is one of the most sensitive detectors available for quantitative and qualitative analysis and is especially powerful when coupled to a separation technique such as gas chromatography. A physicist may note that high resolution mass spectrometry has been responsible for the accurate determination of the atomic masses listed in the periodic table. Biologists use mass spectrometry to identify high molecular weight proteins and nucleic acids and even for sequencing peptides. Materials scientists use mass spectrometry for characterizing the composition and properties of polymers and metal surfaces.

The mass spectrometer tends to be a passive instrument in these applications, used to record mass spectra. In chemical physics and physical chemistry, however, the mass spectrometer takes on a dynamic function as a

tool for the investigation of the physico-chemical properties of atoms, molecules and ions. It is this latter application that is the subject of this chapter, and it is hoped that it will bring the reader to a new understanding of the utility of mass spectrometry in their research.

The chapter is divided into sections, one for each general class of mass spectrometer: magnetic sector, quadrupole, time-of-flight and ion cyclotron resonance. The experiments performed by each are quite often unique and so have been discussed separately under each heading.

B1.7.2 ION SOURCES

A common feature of all mass spectrometers is the need to generate ions. Over the years a variety of ion sources have been developed. The physical chemistry and chemical physics communities have generally worked on gaseous and/or relatively volatile samples and thus have relied extensively on the two traditional ionization methods, electron ionization (EI) and photoionization (PI). Other ionization sources, developed principally for analytical work, have recently started to be used in physical chemistry research. These include fast-atom bombardment (FAB), matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ES).

B1.7.1.2 ELECTRON IONIZATION (EI)

A schematic diagram of an electron ionization (EI) ion source is shown in [figure B1.7.1](#). A typical source will consist of a block, filament, trap electrode, repeller electrode, acceleration region and a focusing lens. Sample vapour, introduced into the ion source (held at the operating potential of the instrument) through a variable leak valve or capillary interface, is ionized by electrons that have been accelerated towards the block by a potential gradient and collected at the trap electrode. A repeller electrode nudges the newly formed ions out of the source through an exit slit,

-2-

and they are accelerated to the operating kinetic energy of the instrument. A series of ion lenses is used to focus the ion beam onto the entrance aperture of the mass spectrometer.

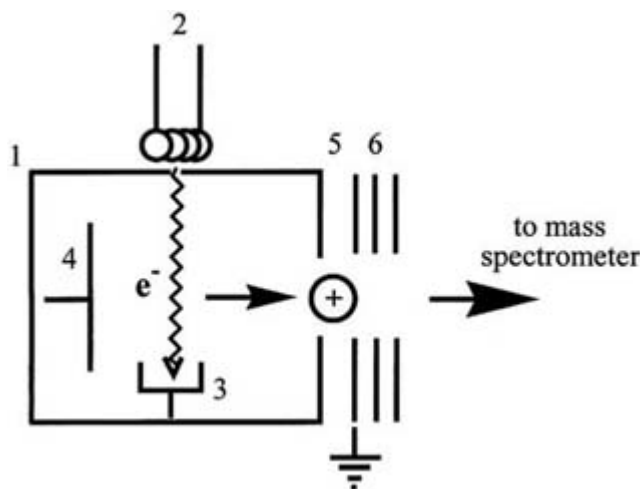


Figure B1.7.1. Schematic diagram of an electron ionization ion source: source block (1); filament (2); trap electrode (3); repeller electrode (4); acceleration region (5); focusing lens (6).

Ionization with energetic electrons does not deposit a fixed amount of energy into a molecule. Rather, a 10 eV

beam of electrons can deposit anywhere from 0 to 10 eV of energy. For this reason, most instruments relying on electron ionization (often referred to as ‘electron impact’ ionization, a misleading expression as no actual collision between a molecule and an electron takes place) use electron beams of fairly high energy. Most analytical instruments employ electron energies of ~70 eV as it has been found that a maximum in the ion yield for most organic molecules occurs around this value. The resulting ion internal energies can be described by the Wannier threshold law [1], but at an electron energy of 70 eV this corresponds almost exactly to the photoelectron spectrum. Superimposed on this is the internal energy distribution of the neutral molecules prior to ionization. Since most ion sources operate at very low pressure ($\sim 10^{-7}$ to 10^{-6} Torr), the resulting ion population has a non-Boltzmann distribution of internal energies and thus it is difficult to discuss the resulting ion chemistry in terms of a thermodynamic temperature.

It is fairly difficult to obtain energy-selected beams of electrons (see [chapter B1.7](#)). Thus, electron beams employed in most mass spectrometers have broad energy distributions. One advantage of EI over photoionization, though, is that it is relatively simple to produce high energy electrons. All that is required is the appropriate potential drop between the filament and the ion source block. This potential drop is also continuously adjustable and the resulting electron flux often independent of energy.

B1.7.2.2 PHOTOIONIZATION (PI)

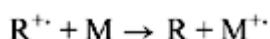
Photoionization with photons of selected energies can be a precise method for generating ions with known internal energies, but because it is bound to a continuum process (the ejected electron can take on any energy), ions are usually generated in a distribution of internal energy states, according to the equation $h\nu + E_{\text{therm}} = IE_{\text{AB}} + E_{\text{AB}^+} + E_{\text{e}}$, where $h\nu$ is the photon energy, E_{therm} is the average thermal energy of the molecule AB, IE is the ionization energy of the

molecule, E_{AB^+} is the ion internal energy and E_{e} is the kinetic energy of the departing electron. The deposition of energy into the ion typically follows the photoelectron spectrum up to the photon energy. Superimposed on this is the internal energy distribution of the original neutral molecules.

There are three basic light sources used in mass spectrometry: the discharge lamp, the laser and the synchrotron light source. Since ionization of an organic molecule typically requires more than 9 or 10 eV, light sources for photoionization must generate photons in the vacuum-ultraviolet region of the electromagnetic spectrum. A common experimental difficulty with any of these methods is that there can be no optical windows or lenses, the light source being directly connected to the vacuum chamber holding the ion source and mass spectrometer. This produces a need for large capacity vacuum pumping to keep the mass spectrometer at operating pressures. Multiphoton ionization with laser light in the visible region of the spectrum overcomes this difficulty.

B1.7.2.3 CHEMICAL IONIZATION

A third method for generating ions in mass spectrometers that has been used extensively in physical chemistry is chemical ionization (CI) [2]. Chemical ionization can involve the transfer of an electron (charge transfer), proton (or other positively charged ion) or hydride anion (or other anion).



The above CI reactions will occur if they are exothermic. In order for these reactions to occur with high efficiency, the pressure in the ion source must be raised to the milliTorr level. Also, the reagent species are often introduced in large excess so that they are preferentially ionized by the electron beam.

B1.7.2.4 OTHER IONIZATION METHODS

One feature common to all of the above ionization methods is the need to thermally volatilize liquid and solid samples into the ion source. This presents a problem for large and/or involatile samples which may decompose upon heating. Ionization techniques that have been developed to get around this problem include fast-atom bombardment (FAB) [3], matrix-assisted laser desorption ionization (MALDI) [4] and electrospray ionization (ESI) [5] (figure B1.7.2). FAB involves bombarding a sample that has been dissolved in a matrix such as glycerol with a high energy beam of atoms. Sample molecules that have been protonated by the glycerol matrix are sputtered off the probe tip, resulting in gas-phase ions. If high energy ions are used to desorb the sample, the technique is called SIMS (secondary ion mass spectrometry). MALDI involves ablating a sample with a laser. A matrix absorbs the laser light, resulting in a plume of ejected material, usually containing molecular ions or protonated molecules. In electrospray, ions are formed in solution by adding protons or other ions to molecules. The solution is sprayed through a fine capillary held at a high potential relative to ground (several keV are common). The sprayed solution consists of tiny droplets that evaporate, leaving gas-phase adduct ions which are then introduced into a mass spectrometer for analysis.

-4-

Figure B1.7.2. Schematic representations of alternative ionization methods to EI and PI: (a) fast-atom bombardment in which a beam of keV atoms desorbs solute from a matrix (b) matrix-assisted laser desorption ionization and (c) electrospray ionization.

B1.7.2.5 MOLECULAR BEAM SOURCES

Sample can be introduced into the ion source in the form of a molecular beam [6, 7] (figure B1.7.3). Molecular beams are most often coupled to time-of-flight instruments for reasons that are discussed in section (B1.7.5). The important advantage that molecular beams have over the other methods discussed in this section is their ability to cool the internal degrees of freedom of the sample. Collisions between a carrier gas (such as helium or argon) and the sample molecule in the rapidly expanding gas mixture results in rotational and vibrational cooling. Using this approach, the effective internal ‘temperature’ of the sample can be significantly less than ambient. One example of the benefits of using molecular beams is in photoionization. The photon energy can be more readily equated to the ion internal energy if the initial internal energy distribution of the neutral molecule is close to 0 K. This cooling also allows weakly bound species such as neutral clusters to be generated and their resulting PI or EI mass spectrum obtained.

-5-

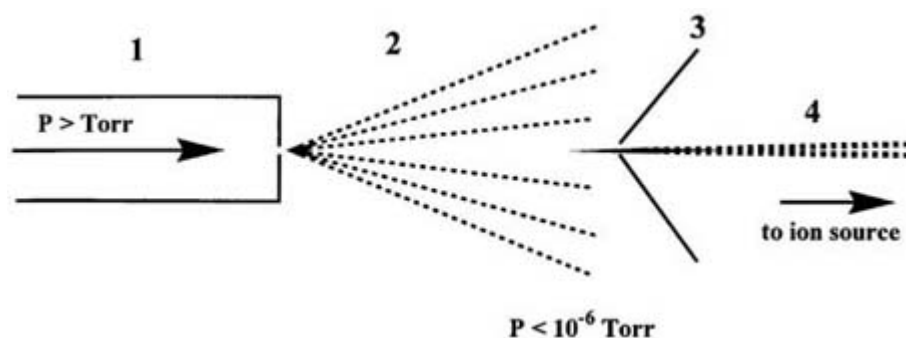


Figure B1.7.3. Schematic diagram of a molecular beam generator: nozzle (1); expansion region (2); skimmer (3) and molecular beam (4).

B1.7.2.6 HIGH PRESSURE SOURCES

There are two significant differences between a high pressure ion source and a conventional mass spectrometer ion source. High pressure sources typically have only two very small orifices (aside from the sample inlet), one to permit ionizing electrons into the source and one to permit ions to leave. Total ion source pressures of up to 5–10 Torr can be obtained allowing the sample vapour to reach thermal equilibrium with the walls of the source. Because the pressures in the source are large, a 70 eV electron beam is insufficient to effectively penetrate and ionize the sample. Rather, electron guns are used to generate electron translational energies of up to 1–2 keV.

B1.7.3 MAGNETIC SECTOR INSTRUMENTS

The first mass spectrometers to be widely used as both analytical and physical chemistry instruments were based on the deflection of a beam of ions by a magnetic field, a method first employed by J J Thomson in 1913 [8] for separating isotopes of noble gas ions. Modern magnetic sector mass spectrometers usually consist of both magnetic and electrostatic sectors, providing both momentum and kinetic energy selection. The term ‘double-focusing’ mass spectrometer refers to such a configuration and relates to the fact that the ion beam is focused at two places between the ion source of the instrument and the detector. It is also possible to add sectors to make three-, four, five- and even six-sector instruments, though the larger of these are typically used for large molecule analysis. One of the staple instruments used in physical chemistry has been the reverse-geometry tandem sector mass spectrometer (‘BE’ configuration), which will be described below. The basic principles apply to any magnetic sector instrument configuration.

B1.7.3.1 INSTRUMENTATION

A schematic diagram of a reverse geometry mass spectrometer is shown in figure B1.7.4.

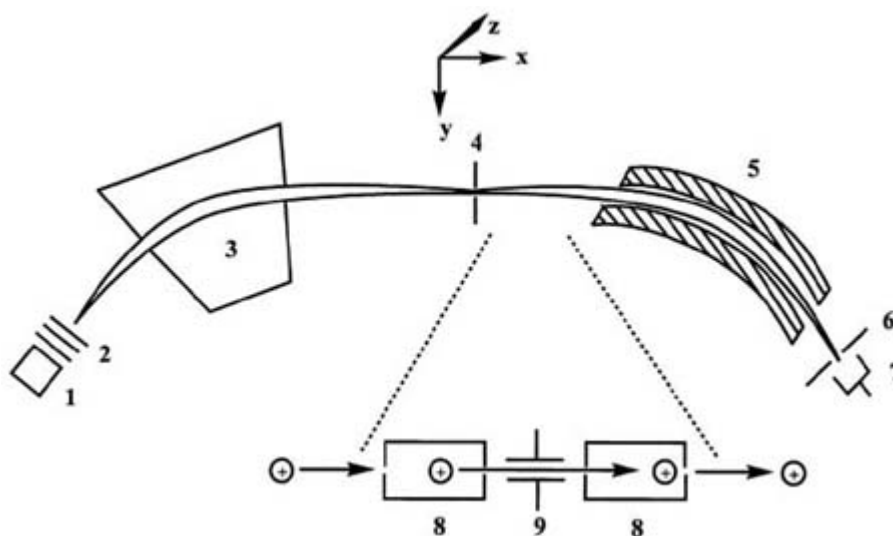


Figure B1.7.4. Schematic diagram of a reverse geometry (BE) magnetic sector mass spectrometer: ion source (1); focusing lens (2); magnetic sector (3); field-free region (4); beam resolving slits (5); electrostatic sector (6); electron multiplier detector (7). Second field-free region components: collision cells (8) and beam deflection electrodes (9).

(A) THE MAGNETIC SECTOR

The magnetic sector consists of two parallel electromagnets surrounding an iron core. The ion beam travels through the flight tube perpendicular to the direction of the imposed magnetic field. The path of an ion travelling orthogonal to a magnetic field is described by a simple mathematical relationship:

$$r = \frac{mv}{Bze} \tag{B1.7.1}$$

where r is the radius of curvature of the path of the ion, m is the ion's mass, v is the velocity, B is the magnetic field strength, z is the number of charges on the ion and e is the unit of elementary charge. Rearranged,

$$mv = rBze \tag{B1.7.2}$$

Most instruments are configured with a fixed value for the radius of curvature, r , so changing the value of B selectively passes ions of particular values of momentum, mv , through the magnetic sector. Thus, it is really the momentum that is selected by a magnetic sector, not mass. We can convert this expression to one involving the accelerating potential.

Magnetic sector instruments typically operate with ion sources held at a potential of between 6 and 10 kV. This results in ions with keV translational kinetic energies. The ion kinetic energy can be written as $zeV = \frac{1}{2}mv^2$ and thus the ion velocity is given by the relationship

$$v = \left(\frac{2zeV}{m} \right)^{1/2}$$

and [equation \(B1.7.2\)](#) becomes

$$m/z = \frac{B^2 r^2 e}{2V}. \quad (\text{B1.7.3})$$

In other words, ions with a particular mass-to-charge ratio, m/z , can be selectively passed through the magnetic sector by appropriate choice of a value of V and B (though normally V is held constant and only B is varied).

Magnetic sectors can be used on their own, or in conjunction with energy analysers to form a tandem mass spectrometer. The unique features of the reverse geometry instrument are presented from this point.

(B) THE FIELD-FREE REGION

The momentum-selected ion beam passes through the field-free region (FFR) of the instrument on its way to the electrostatic sector. The FFR is the main experimental region of the magnetic sector mass spectrometer. Significant features of the FFR can be collision cells and ion beam deflection electrodes. One particular arrangement is shown in [figure B1.7.4](#). A collision cell consists of a 2–3 cm long block of steel with a groove to pass the ion beam. A collision (target) gas can be introduced into the groove, prompting projectile–target gas collisions. The beam deflecting electrode assembly allows the ion beam to be deflected out of the beam path by the application of a potential difference across the assembly (see [section \(B1.7.3.2\)](#)).

(C) THE ELECTROSTATIC SECTOR (ESA)

The electrostatic sector consists of two curved parallel plates between which is applied a potential difference producing an electric field of strength E . Transmission of an ion through the sector is governed by the following relationship

$$\frac{1}{2}mv^2 = zeV = \frac{1}{2}zeEr.$$

This relationship gives an expression similar to [equation \(B1.7.1\)](#):

$$r = \frac{2V}{E}. \quad (\text{B1.7.4})$$

Adjusting the potential across the ESA plates allows ions of selected translational kinetic energy to pass through and be focused, at which point a detector assembly is present to monitor the ion flux. Note that in [equation \(B1.7.4\)](#) neither the mass nor charge of the ion is present. So, isobaric ions with one, two, three etc. charges, accelerated by a potential drop, V , will pass through the ESA at a common value of E . An ion with +1 charge accelerated across a potential difference of 8000 V will have 8 keV translational kinetic energy and be transmitted through the ESA by a field E_1 . An ion with +2 charges will have 16 keV translational energy

but will experience a field strength equivalent to $2E_1$, and so forth.

B1.7.3.2 EXPERIMENTS USING MAGNETIC SECTOR INSTRUMENTS

A single magnetic sector can be used as a mass filter for other apparatus. However, much more information of the simple mass spectrum of a species can be obtained using the tandem mass spectrometer.

(A) MASS-ANALYSED ION KINETIC ENERGY SPECTROMETRY (MIKES)

Ions accelerated out of the ion source with keV translational kinetic energies (and m/z selected with the magnetic sector) will arrive in the FFR of the instrument in several microseconds. Ions dissociating on this timescale (with unimolecular decay rate constants between 10^2 and 10^5 s⁻¹, depending on the physical geometry of the instrument) have been given the name ‘metastable ions’ [9].

In the FFR of the sector mass spectrometer, the unimolecular decomposition fragments, A⁺ and B, of the mass selected metastable ion AB⁺ will, by the conservation of energy and momentum, have lower translational kinetic energy, T , than their precursor:

$$zT_{A^+} = \frac{1}{2}m_{A^+}v^2 \quad zT_B = \frac{1}{2}m_Bv^2 \quad zT_{AB^+} = \frac{1}{2}m_{AB^+}v^2.$$

Thus we find

$$T_{A^+} = \frac{m_{A^+}}{m_{AB^+}}T_{AB^+}.$$

By scanning the ESA to pass ions with lower translational energies, the fragment ions will sequentially pass through to the detector (this is the so-called MS/MS, or MS², experiment). The final ion abundance kinetic energy spectrum ([figure B1.7.5\(a\)](#)) is converted to an ion abundance fragment m/z spectrum by the above relationships. The MIKE spectrum is the end result of all low energy unimolecular processes of the selected ions, including isomerization. Thus, isomeric ions which interconvert on the μ s timescale often have closely related, if not identical, MIKE spectra. There

are several characteristic peak shapes expected in a MIKE spectrum that are summarized in [figure B1.7.6](#).

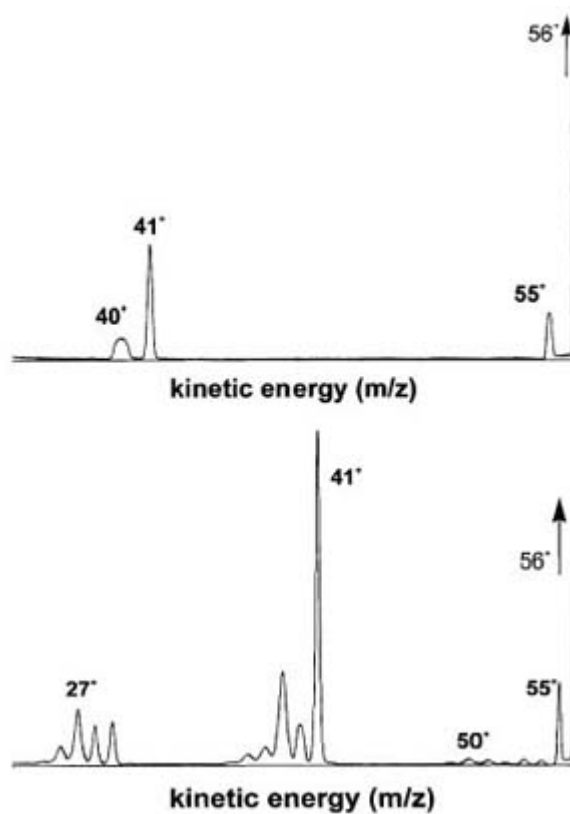


Figure B1.7.5. (a) MIKE spectrum of the unimolecular decomposition of 1-butene ions (m/z 56). This spectrum was obtained in the second field-free region of a reverse geometry magnetic sector mass

spectrometer (VG ZAB-2HF). (b) Collision-induced dissociation mass spectrum of 1-butene ions. Helium target gas was used to achieve 10% beam reduction (single collision conditions) in the second field-free region of a reverse geometry magnetic sector mass spectrometer (VG ZAB-2HF).

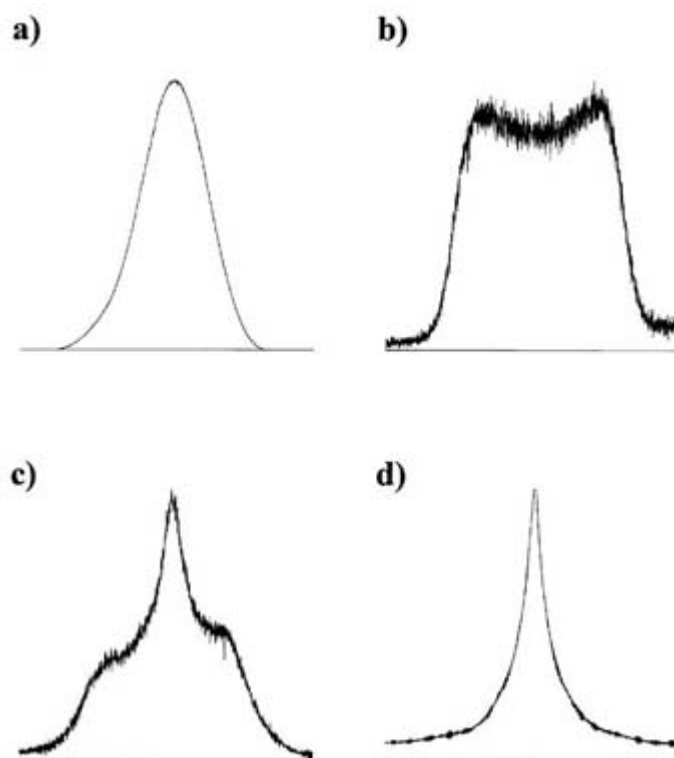


Figure B1.7.6. Fragment ion peak shapes expected in MIKE spectra: (a) typical Gaussian energy profile; (b) large average kinetic energy release causing z -axial discrimination of the fragment ions, resulting in a ‘dished-top’ peak; (c) competing fragmentation channels, each with its own distinct kinetic energy release, producing a ‘composite’ peak; (d) fragmentation occurring from a dissociative excited state.

(B) COLLISION-INDUCED DISSOCIATION (CID) MASS SPECTROMETRY

A collision-induced dissociation (CID) mass spectrum [10, 11] of mass selected ions is obtained by introducing a target gas into a collision cell in one of the field-free regions. The resulting high energy (keV) CID mass spectrum, obtained and analysed in the same way as a MIKE spectrum, contains peaks due to ions formed in virtually all possible unimolecular dissociation processes of the precursor ion (figure B1.7.5(b)). The timescale of the collision-induced fragmentation reactions is quite different from the MIKE experiment, ranging from the time of the collision event ($t \approx 10^{-15}$ s) to the time the ions exit the FFR. For this reason, isomerization reactions tend not to play a significant role in collision-induced reactions and thus the CID mass spectra are often characteristic of ion connectivity.

In collisional excitation, translational energy of the projectile ion is converted into internal energy. Since the excited states of the ions are quantized, so will the translational energy loss be. Under conditions of high energy resolution, it is

possible to obtain a translational energy spectrum of the precursor ions exhibiting peaks that correspond to the formation for discrete excited states (translational energy spectroscopy [12]).

It is possible in a sector instrument to perform a variety of other experiments on the projectile ions. Many involve examining the products of charge exchange with the target gas, while others allow neutral species to be studied. Some of the more common experiments are summarized in figure B1.7.7. All of the experiments

have been described for projectile cations, but anions can be studied in analogous manners.

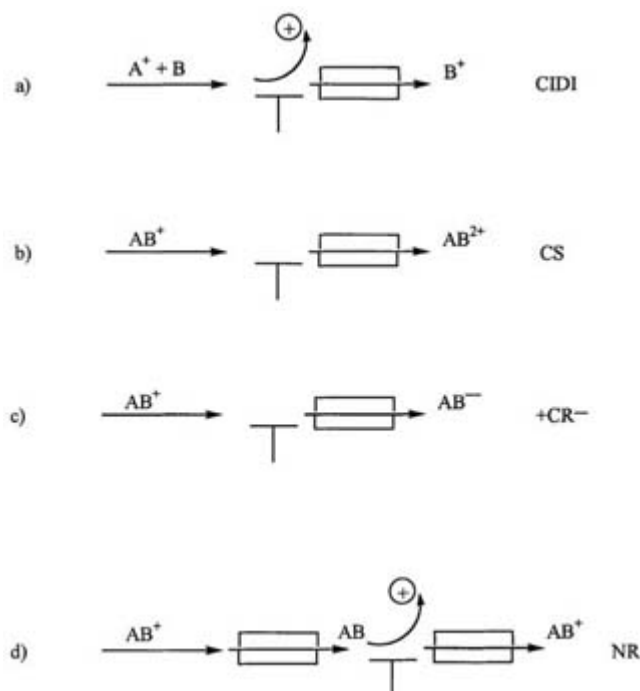


Figure B1.7.7. Summary of the other collision based experiments possible with magnetic sector instruments: (a) collision-induced dissociation ionization (CIDI) records the CID mass spectrum of the neutral fragments accompanying unimolecular dissociation; (b) charge stripping (CS) of the incident ion beam can be observed; (c) charge reversal (CR) requires the ESA polarity to be opposite that of the magnet; (d) neutralization–reionization (NR) probes the stability of transient neutrals formed when ions are neutralized by collisions in the first collision cell. Neutrals surviving to be collisionally reionized in the second cell are recorded as ‘recovery’ ions in the NR mass spectrum.

(C) KINETIC ENERGY RELEASE (KER) MEASUREMENTS

In a unimolecular dissociation, excess product energy is typically distributed among the translational, rotational and vibrational modes in a statistical fashion. The experimentally observed phenomenon is the distribution of translational kinetic energies of the departing fragment ions (the kinetic energy release, KER) [9]. In magnetic sector instruments, the result of x -axial (i.e. along the beam path) KER is the observation of fragment ion peaks in the MIKE or CID spectra which have a broader kinetic energy spread than the precursor ion peak. In a CID, however, this spread is complicated by collisional scattering and so KER is most often discussed for peaks in MIKE spectra. If the mass

spectrometer is operated under conditions of high resolution, obtained by narrowing the y -axis beam collimating slits throughout the instrument, the widths of the fragment ion peaks are indicative of this kinetic energy release. The measured value for the KER is typically expressed as the value at half-height of the fragment ion peak, $T_{0.5}$, and is calculated with the following equation

$$T_{0.5}(\text{meV}) = \frac{m_{\text{AB}^+}^2 (\Delta V_{0.5, \text{A}^+}^2 - \Delta V_{0.5, \text{AB}^+}^2)}{16(8)m_{\text{A}^+}m_{\text{B}}}$$

where m is the mass of the various species, $\Delta V_{0.5}$ is the full width energy spread of the fragment and precursor ion peaks at half height and (8) represents the typically 8 keV translational kinetic energy of the precursor [9]. The resulting $T_{0.5}$ is in meV. Note that since knowledge of the internal energy distribution of the dissociating ions is lacking, the relationship between $T_{0.5}$ and the average KER is strictly qualitative, i.e. a large $T_{0.5}$ indicates a large average KER value. How statistical the distribution of product excess energies is will depend on the dynamics of the dissociation.

B1.7.4 QUADRUPOLE MASS FILTERS, QUADRUPOLE ION TRAPS AND THEIR APPLICATIONS

Another approach to mass analysis is based on stable ion trajectories in quadrupole fields. The two most prominent members of this family of mass spectrometers are the quadrupole mass filter and the quadrupole ion trap. Quadrupole mass filters are one of the most common mass spectrometers, being extensively used as detectors in analytical instruments, especially gas chromatographs. The quadrupole ion trap (which also goes by the name ‘quadrupole ion store, QUISTOR’, Paul trap, or just ion trap) is fairly new to the physical chemistry laboratory. Its early development was due to its use as an inexpensive alternative to tandem magnetic sector and quadrupole filter instruments for analytical analysis. It has, however, started to be used more in the chemical physics and physical chemistry domains, and so it will be described in some detail in this section.

The principles of operation of quadrupole mass spectrometers were first described in the late 1950s by Wolfgang Paul who shared the 1989 Nobel Prize in Physics for this development. The equations governing the motion of an ion in a quadrupole field are quite complex and it is not the scope of the present article to provide the reader with a complete treatment. Rather, the basic principles of operation will be described, the reader being referred to several excellent sources for more complete information [13, 14 and 15].

B1.7.4.1 THE QUADRUPOLE MASS FILTER

A schematic diagram of a quadrupole mass filter is shown in [figure B1.7.8](#). In an ideal, three-dimensional, quadrupole field, the potential ϕ at any point (x, y, z) within the field is described by [equation \(B1.7.5\)](#):

-13-

$$\phi = \frac{\phi_0}{r_0^2} (ax^2 + by^2 + cz^2) \quad (\text{B1.7.5})$$

where ϕ_0 is the applied potential, r_0 is half the distance between the hyperbolic rods and a , b and c are coefficients. The applied potential is a combination of a radio-frequency (RF) potential, $V \cos \omega t$, and direct current (DC) potential, U . The two can be expressed in the following relationship:

$$\phi_0 = U + V \cos \omega t$$

where ω is the angular frequency of the RF field (in rad s^{-1}) and is 2π times the frequency in Hertz. A potential applied across the rods (see [figure B1.7.8](#)) is flipped at radio-frequencies.

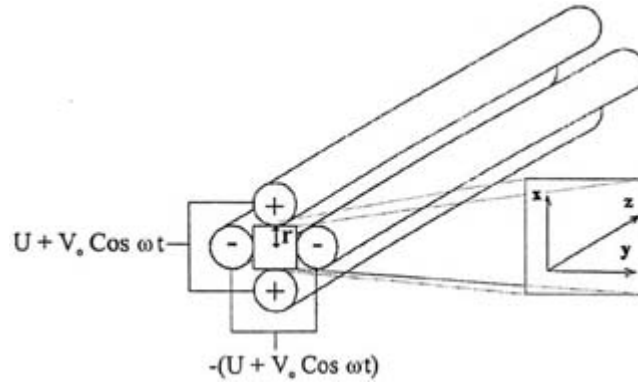


Figure B1.7.8. Quadrupole mass filter consisting of four cylindrical rods, spaced by a radius r (reproduced with permission of Professor R March, Trent University, Peterborough, ON, Canada).

Ideally, the rods in a quadrupole mass filter should have a hyperbolic geometry, but more common is a set of four cylindrical rods separated by a distance $2r$, where $r \approx 1.16r_0$ (figure B1.7.8). This arrangement provides for an acceptable field geometry along the axis of the mass filter. Since there is no field along the axis of the instrument, equation (B1.7.5) simplifies to $\phi = (\phi_0/r_0^2)(ax^2 + by^2)$.

The values of a and b that satisfy this relationship are $a = 1$ and $b = -1$, so the potential inside the quadrupole mass filter takes on the form:

$$\phi = \frac{\phi_0}{r_0^2}(x^2 - y^2).$$

The equations of motion within such a field are given by:

$$\frac{d^2x}{dt^2} + \frac{e}{mr_0^2}(U - V \cos \omega t)x = 0$$

$$\frac{d^2y}{dt^2} + \frac{e}{mr_0^2}(U - V \cos \omega t)y = 0.$$

Now, we can make three useful substitutions

$$a_x = -a_y = \frac{8eU}{m\omega^2 r_0^2} \quad q_x = -q_y = \frac{8eV}{m\omega^2 r_0^2} \quad \xi = \frac{\omega t}{2}$$

and the result is in the general form of the Mathieu equation [16]:

$$\frac{d^2u}{d\xi^2} + (a_u - 2q_u \cos 2\xi)u = 0. \tag{B1.7.6}$$

Equation (B1.7.6) describes the ion trajectories in the quadrupole field (where u can be either x or y). The stable, bounded solutions to these equations represent conditions of stable, bounded trajectories in the

quadrupole mass filter. A diagram representing the stable solutions to the equations for both the x - and y -axes (really the intersection of two sets of stability diagrams, one for the x -axis, one for the y -axis) is shown in [figure B1.7.9\(a\)](#). This figure represents the stability region closest to the axis of the instrument and is the most appropriate for the operation of the quadrupole. This stability region is also unique for a given m/z ratio. [figure B1.7.9\(b\)](#) represents the stability regions (transformed into axes of the applied DC and RF potential) for a series of ions with different m/z values. The line running through the apex of each region is called the operating line and represents the conditions (U and V) for the selective filtering of different mass ions through the instrument. The ratio of U to V is a constant along the operating line. It is apparent from [figure B1.7.9\(b\)](#) that the resolution of the quadrupole mass filter can be altered by changing the slope of the operating line. A greater slope means there is greater separation of the ions, but at the expense of sensitivity (fewer ions will have stable trajectories).

-15-

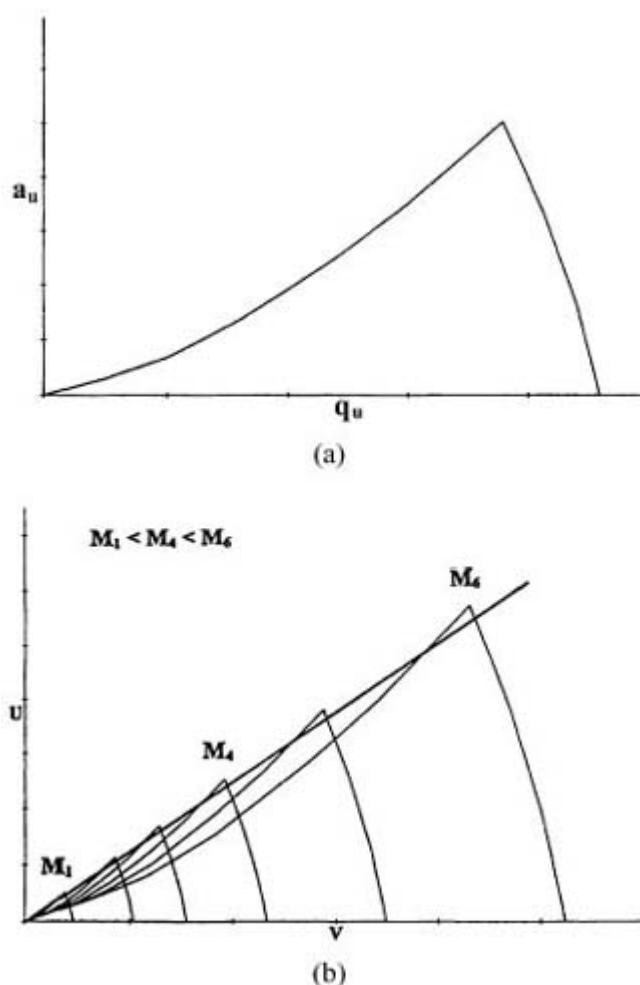


Figure B1.7.9. (a) Stability diagram for ions near the central axis of a quadrupole mass filter. Stable trajectories occur only if the a_u and q_u values lie beneath the curve. (b) Stability diagram (now as a function of U and V) for six ions with different masses. The straight line running through the apex of each set of curves is the ‘operating’ line, and corresponds to values of U/V that will produce mass resolution (reproduced with permission of Professor R March, Trent University, Peterborough, ON, Canada).

To be effective, it is necessary for the ions traversing the instrument to experience several RF cycles. Thus, unlike magnetic sector instruments, the ions formed in the ion source of a quadrupole mass filter apparatus are accelerated to only a few eV kinetic energy (typically 5–10 eV). The timescale of the experiment is therefore

much longer than for magnetic sectors, ions having to survive milliseconds rather than microseconds to be detected.

B1.7.4.2 EXPERIMENTS USING QUADRUPOLE MASS FILTERS

Aside from the single mass filter, the most common configuration for quadrupole mass spectrometers is the triple-quadrupole instrument. This is the simplest tandem mass spectrometer using quadrupole mass filters. Typically, the

-16-

first and last quadrupole are operated in mass selective mode as described above. The central quadrupole is usually an RF-only quadrupole. The lack of a DC voltage means that ions of all m/z values will have stable trajectories through the filter.

(A) COLLISION-INDUCED DISSOCIATION

Adding a collision gas to the RF-only quadrupole of a triple-quadrupole instrument permits collision-induced dissociation experiments to be performed. Unlike magnetic sector instruments, the low accelerating potential in quadrupole instruments means that low energy collisions occur. In addition, the time taken by the ions to traverse the RF only quadrupole results in many of these low energy collisions taking place. So, collisional excitation occurs in a multi-step process, rather than in a single, high-energy, process as in magnetic sector instruments.

(B) REACTIVE COLLISIONS

Since ions analysed with a quadrupole instrument have low translational kinetic energies, it is possible for them to undergo bimolecular reactions with species inside an RF-only quadrupole. These bimolecular reactions are often useful for the structural characterization of isomeric species. An example of this is the work of Harrison and co-workers [17]. They probed the reactions of CH_3NH_2^+ ions with isomeric butenes and pentenes in the RF-only quadrupole collision cell of a hybrid BEqQ instrument. The mass spectra of the products of the ion–molecule reactions were distinct for the various isomers probed. Addition of the amine to the olefin followed by fragmentation produced characteristic iminium ions only for terminal olefins without substitution at the olefinic carbons (figure B1.7.10).

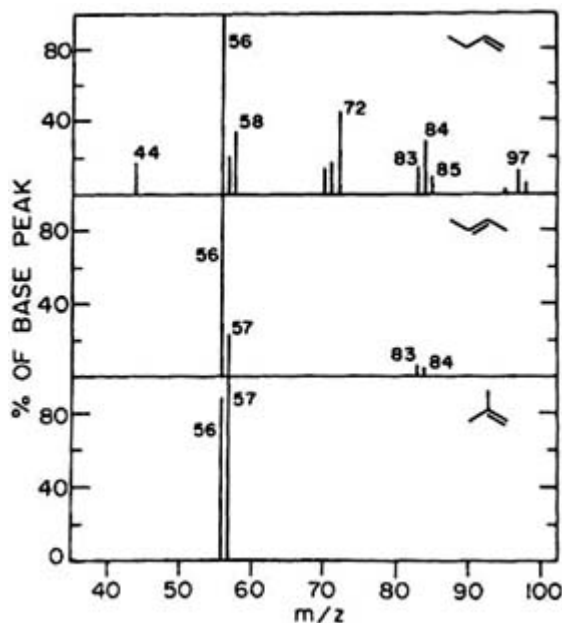
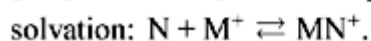
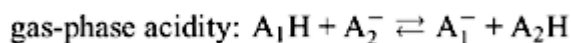


Figure B1.7.10. Three mass spectra showing the results of reactive collisions between a projectile ion CH_3NH_2^+ and three isomeric butenes. (Taken from Uspychuk L L, Harrison A G and Wang J 1992 Reactive collisions in quadrupole cells. Part 1. Reaction of $[\text{CH}_3\text{NH}_2]^+$ with isomeric butenes and pentenes *Org. Mass Spectrom.* **27** 777–82. Copyright John Wiley & Sons Limited. Reproduced with permission.)

-17-

(C) EQUILIBRIUM MEASUREMENTS

It is possible to determine the equilibrium constant, K , for the bimolecular reaction involving gas-phase ions and neutral molecules in the ion source of a mass spectrometer [18]. These measurements have generally focused on three properties, proton affinity (or gas-phase basicity) [19, 20], gas-phase acidity [21] and solvation enthalpies (and free energies) [22, 23]:



A common approach has been to measure the equilibrium constant, K , for these reactions as a function of temperature with the use of a variable temperature high pressure ion source (see section (B1.7.2)). The ion concentrations are approximated by their abundance in the mass spectrum, while the neutral concentrations are known from the sample inlet pressure. A van't Hoff plot of $\ln K$ versus $1/T$ should yield a straight line with slope equal to the reaction enthalpy (figure B1.7.11). Combining the PA with a value for $\Delta_{\text{basicity}} G^0$ at one temperature yields a value for ΔS^0 for the half-reaction involving addition of a proton to a species. While quadrupoles have been the instruments of choice for many of these studies, other mass spectrometers can act as suitable detectors [19, 20].

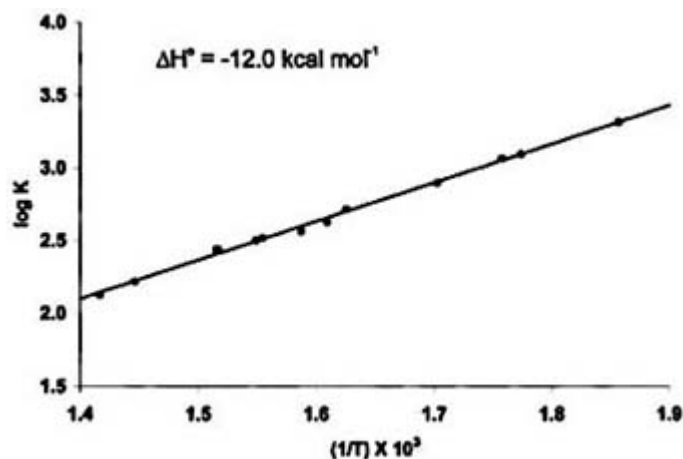


Figure B1.7.11. Van't Hoff plot for equilibrium data obtained for the reaction of isobutene with ammonia in a high pressure ion source (reproduced from data in [19]).

(D) SELECTED ION FLOW TUBE

Another example of the use of quadrupole filters in studying reactive collisions of gaseous ions is the selected ion flow tube (SIFT) [24]. This has been perhaps the most widely employed instrument for studying the kinetics of bimolecular reactions involving ions. Its development by N G Adams and D Smith sprang from the utility of the flowing afterglow (FA) technique (developed in the early 1960s by Ferguson, Fehsenfeld and Schmeltekopf [25, 26]) in the study of atmospheric reactions [27].

A schematic diagram of a SIFT apparatus is shown in figure B1.7.12. The instrument consists of five basic regions, the ion source, initial quadrupole mass filter, flow tube, second mass filter and finally the detector. The heart of the instrument is the flow tube, which is a steel tube approximately 1 m long and 10 cm in diameter. The pressure in the flow tube is kept of the order of 0.5 Torr, resulting in carrier gas flow rates of $\sim 100 \text{ m s}^{-1}$. Along the flow tube there are orifices that are used to introduce neutral reagents into the flow stream. Product ions arriving at the end of the flow tube are skimmed through a small orifice and mass analysed with a second quadrupole filter before being detected. The reactions occurring in the flow tube can be monitored as a function of carrier gas flow rate, and hence timescale. A detailed description of the extraction of rate constants from SIFT experiments is given by Smith and Adams [24]. Examples of the type of information obtained with the SIFT technique can be found in a recent series of articles by D Smith and co-workers [28, 29 and 30].

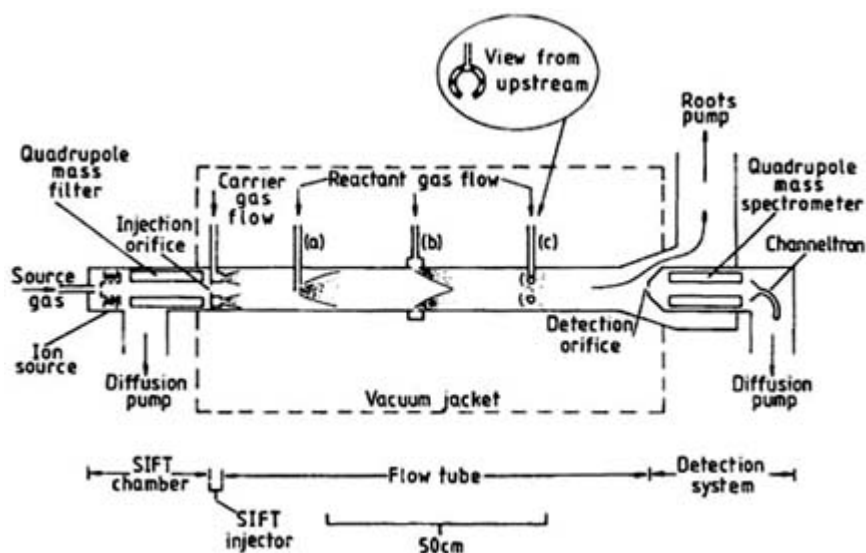


Figure B1.7.12. A schematic diagram of a typical selected-ion flow (SIFT) apparatus. (Smith D and Adams N G 1988 The selected ion flow tube (SIFT): studies of ion–neutral reactions *Advances in Atomic and Molecular Physics* vol 24, ed D Bates and B Bederson p 4. Copyright Academic Press, Inc. Reproduced with permission.)

(E) ION-GUIDE INSTRUMENTS

Another instrument used in physical chemistry research that employs quadrupole mass filters is the guided ion beam mass spectrometer [31]. A schematic diagram of an example of this type of instrument is shown in [figure B1.7.13. A](#)

mass selected beam of ions is introduced into an ion guide, in which their translational energy can be precisely controlled down to a few fractions of an electron volt. Normally at these energies, divergence of the ion beam would preclude the observation of any reaction products. To overcome this, the ions are trapped in the beam path with either a quadrupole or octapole filter operated in RF-only mode (see above). The trapping characteristics (i.e., how efficiently the ions are ‘guided’ along the flight path) of the octapole filter are superior to the quadrupole filter and so octapoles are often employed in this type of apparatus. The operational principles of an octapole mass filter are analogous to those of the quadrupole mass filter described above.

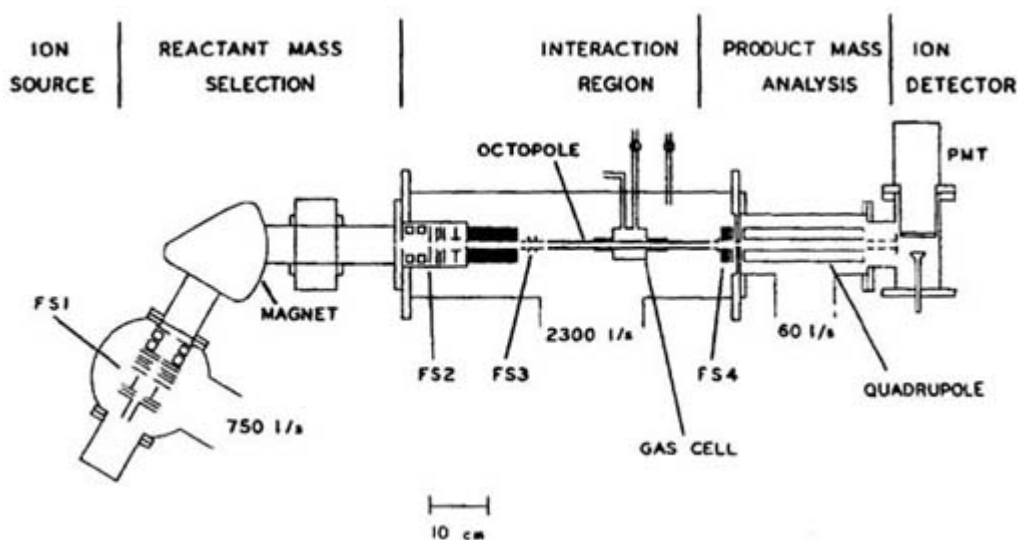


Figure B1.7.13. A schematic diagram of an ion-guide mass spectrometer. (Ervin K M and Armentrout P B 1985 Translational energy dependence of $\text{Ar}^+ + \text{XY} \rightarrow \text{ArX}^+ + \text{Y}$ from thermal to 30 eV c.m. *J. Chem. Phys.* **83** 166–89. Copyright American Institute of Physics Publishing. Reproduced with permission.)

Using a guided ion beam instrument the translational energy dependent reaction cross sections of endothermic fragmentation processes can be determined [32]. Modelling these cross sections ultimately yields their energy thresholds and a great deal of valuable thermochemical information has been derived with this technique. Precision of ± 0.2 eV can be obtained for reaction thresholds. Bimolecular reactions can also be studied and reaction enthalpies derived from the analysis of the cross section data.

B1.7.4.3 THE QUADRUPOLE ION TRAP

The quadrupole ion trap is the three dimensional equivalent to the quadrupole mass filter. A typical geometry consists of two hyperbolic endcap electrodes and a single ring electrode (figure B1.7.14). Unlike the quadrupole mass filter, however, the ion trap can be used both as a mass selective device or as an ion storage device. It is this latter ability that has led to the popularity and versatility of the ion trap. The theoretical treatment of ion trajectories inside the ion trap is similar to that presented above for the mass filter, except that now the field is no longer zero in the z -axis. It is convenient to use cylindrical coordinates rather than Cartesian coordinates and the resulting relationship describing the

-20-

potential inside the ion trap is given as:

$$\phi = \frac{\phi_0}{r_0^2} (r_0^2 - 2z_0^2)$$

where r_0 and z_0 are defined as in figure B1.7.14. The relationship $r_0^2 = 2z_0^2$ has usually governed the geometric arrangement of the electrodes. The equations of motion for an ion in the ion trap are analogous to those for the quadrupole mass filter:

$$\frac{d^2r}{dt^2} + \frac{2e}{mr_0^2}(U - V \cos \omega t)r = 0$$

$$\frac{d^2z}{dt^2} + \frac{4e}{mr_0^2}(U - V \cos \omega t)z = 0$$

and with the analogous substitutions,

$$a_z = -2a_r = \frac{-16eU}{m\omega^2(r_0^2 + 2z_0^2)}$$

$$q_z = -2q_r = \frac{-8eV}{m\omega^2(r_0^2 + 2z_0^2)}$$

$$\xi = \frac{\omega t}{2}.$$

the Mathieu equation is obtained.

-21-

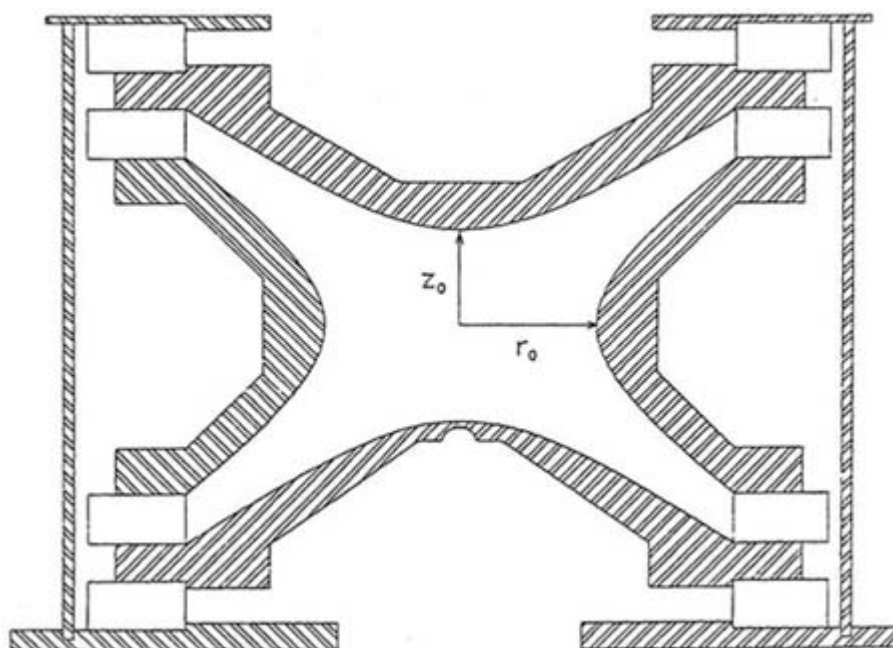


Figure B1.7.14. Schematic cross-sectional diagram of a quadrupole ion trap mass spectrometer. The distance between the two endcap electrodes is $2z_0$, while the radius of the ring electrode is r_0 (reproduced with permission of Professor R March, Trent University, Peterborough, ON, Canada).

The Mathieu equation for the quadrupole ion trap again has stable, bounded solutions corresponding to stable, bounded trajectories inside the trap. The stability diagram for the ion trap is quite complex, but a subsection of the diagram, corresponding to stable trajectories near the physical centre of the trap, is shown in [figure B1.7.15](#). The interpretation of the diagram is similar to that for the quadrupole mass filter.

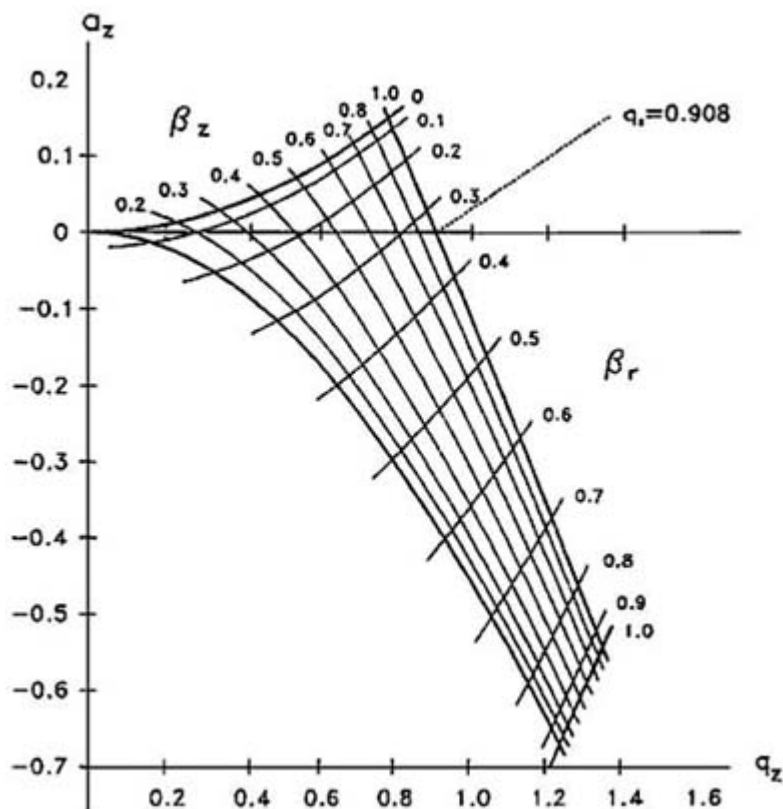


Figure B1.7.15. Stability diagram for ions near the centre of a quadrupole ion trap mass spectrometer. The enclosed area reflects values for a_z and q_z that result in stable trapping trajectories (reproduced with permission of Professor R March, Trent University, Peterborough, ON, Canada).

B1.7.4.4 EXPERIMENTS USING QUADRUPOLE ION TRAPS

The ion trap has three basic modes in which it can be operated. The first is as a mass filter. By adjusting U and V to reside near the apex of the stability diagram in figure B1.7.15 only ions of a particular m/z ratio will be selected by the trap. An operating line that intersects the stability diagram near the apex (as was done for the mass filter) with a fixed U/V ratio describes the operation of the ion trap in a mass scanning mode. A second mode of operation is with the potential ϕ_0 applied only to the ring electrode, the endcaps being grounded. This allows ions to be selectively stored. The application of an extraction pulse to the endcap electrodes ejects ions out of the trap for detection. A third mode is the addition of an endcap potential, $-U$. This mode permits mass selective storage in the trap, followed by storage of all ions. This latter method permits the ion trap to be used as a tandem mass spectrometer.

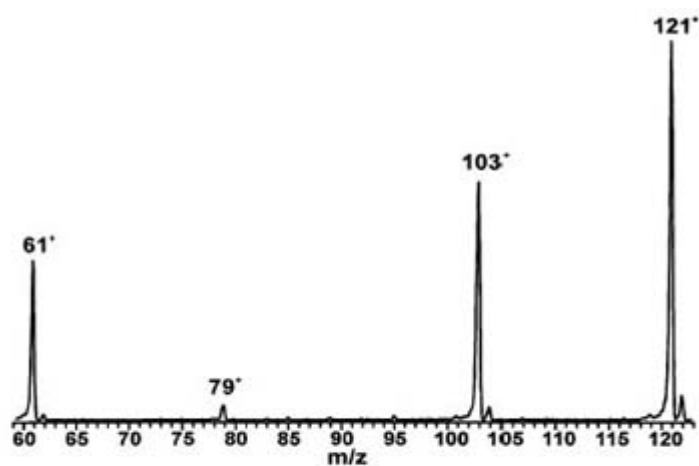
(A) ION ISOLATION

One of the principle uses of the ion trap is as a tandem-in-time mass spectrometer. Ions with a particular m/z ratio formed in the ion trap, or injected into the trap from an external source, can be isolated by resonantly ejecting all other

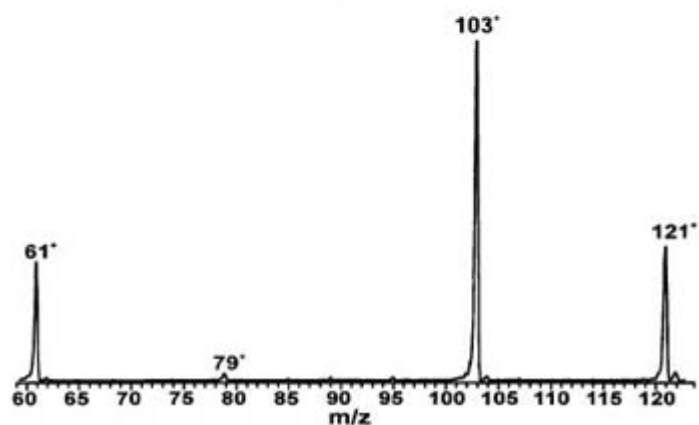
ions. This can be accomplished in a variety of ways. One method involves applying a broad-band noise field between the endcap electrodes to resonantly excite (in the axial direction) and eject all ions. The secular frequency of the ions to be stored is notched out of this noise field, leaving them in the trap. In another method, ions with lower and higher m/z ratio can be ejected by adjusting the amplitude of the RF and DC potentials so that the (a_z, q_z) value for the ion of interest lies just below the apex in [figure B1.7.15](#).

(B) COLLISION-INDUCED DISSOCIATION

A unique aspect of the ion trap is that the trapping efficiency is significantly improved by the presence of helium damping gas. Typical pressures of helium in the trap are ~ 1 milliTorr. Collisions between the trapped ions and helium gas effectively cause the ions to migrate towards the centre of the trap where the trapping field is most perfect. This causes significant improvements in sensitivity in analytical instruments. The presence of the damping gas also permits collision-induced dissociation to be performed. In the manner described above, ions with a particular m/z ratio can be isolated in the trap. The axial component of the ion motion is then excited resonantly by applying a potential across the endcap electrodes. The amplitude of this potential is controlled to prevent resonant ejection of the ions, but otherwise this is a similar experiment to that described above for mass selective ejection. The potential (often called the 'tickle' potential), increases the kinetic energy of the selected ions which in turn increases the centre-of-mass collision energy with the helium damping gas. These collisions now excite the mass selected ions causing them to dissociate. After the 'tickle' period, the trap is returned to a state where all ions can be trapped and thus the mass spectrum of the fragmentation product ions can be obtained. The CID mass spectra that are obtained in this manner are similar to those obtained with triple quadrupole instruments, in that they are the result of many low energy collisions ([figure B.1.7.16\(a\)](#)). There have been attempts to measure unimolecular reaction kinetics by probing fragment ion intensities as a function of time after excitation [[33](#)].



(a)



(b)

Figure B1.7.16. Mass spectra obtained with a Finnigan GCQ quadrupole ion trap mass spectrometer. (a) Collision-induced dissociation mass spectrum of the proton-bound dimer of isopropanol $[(\text{CH}_3)_2\text{CHOH}]_2\text{H}^+$. The m/z 121 ions were first isolated in the trap, followed by resonant excitation of their trajectories to produce CID. Fragment ions include water loss (m/z 103), loss of isopropanol (m/z 61) and loss of 42 amu (m/z 79). (b) Ion-molecule reactions in an ion trap. In this example the m/z 103 ion was first isolated and then resonantly excited in the trap. Endothermic reaction with water inside the trap produces the proton-bound cluster at m/z 121, while CID produces the fragment with m/z 61.

With the right software controlling the instrument, it is possible for the above process to be repeated n times, i.e., the ion trap is theoretically capable of MS^n experiments, though the ion concentration in the trap is always the limiting factor in these experiments.

(C) BIMOLECULAR REACTIONS

The same procedure as outlined above can be used to study ion-molecule reactions [15, 34]. Mass-selected ions will react with neutral species inside the trap. The presence of the damping gas means that stable (thermodynamic and

kinetic) complexes may be formed. Allowing the mass-selected ions to react in the trap, while storing all

reaction products, allows the course of the reactions to be followed ([figure B1.7.16\(b\)](#)). Changing the storage time allows the relative abundance of reactant and product ions to be monitored as a function of time. This introduces the possibility of measuring ion–molecule reaction rate constants with the ion trap. Since the ion internal energy distribution in such an experiment is described by a temperature near the ambient temperature of the damping gas, ion–molecule reactions can be probed as a function of temperature by raising the trap temperature. Resonant excitation of the mass-selected ions (see CID section) effectively raises their internal energy, allowing endothermic reactions with neutral species to take place (there will be a limit on the endothermicity extending from the limit on internal excitation occurring in the resonant excitation process). The degree of axial resonant excitation can be controlled and there have been attempts to relate this to an effective ‘temperature’ [35].

B1.7.5 TIME-OF-FLIGHT MASS SPECTROMETERS

Probably the simplest mass spectrometer is the time-of-flight (TOF) instrument [36]. Aside from magnetic deflection instruments, these were among the first mass spectrometers developed. The mass range is theoretically infinite, though in practice there are upper limits that are governed by electronics and ion source considerations. In chemical physics and physical chemistry, TOF instruments often are operated at lower resolving power than analytical instruments. Because of their simplicity, they have been used in many spectroscopic apparatus as detectors for electrons and ions. Many of these techniques are included as chapters unto themselves in this book, and they will only be briefly described here.

B1.7.5.1 TIME-OF-FLIGHT EQUATIONS

The basic principle behind TOF mass spectrometry [36] is the equation for kinetic energy, $zeV = \frac{1}{2}mv^2$, where the translational kinetic energy of an ion accelerated out of the ion source by a potential drop, V is zeV . If ions of mass m are given zeV kinetic energy, then the time, t_d , the ions take to travel a distance d is given by:

$$t_d = d \frac{m}{2zeV}. \quad (\text{B1.7.7})$$

In the simplest form, t_d reflects the time of flight of the ions from the ion source to the detector. This time is proportional to the square root of the mass, i.e., as the masses of the ions increase, they become closer together in flight time. This is a limiting parameter when considering the mass resolution of the TOF instrument.

The ion time of flight, as given by equation (B1.7.7), is oversimplified, however. There are a number of factors which change the final measured TOF. These are considered below.

(A) ION SOURCE RESIDENCE TIME

A schematic diagram of a simple TOF instrument is shown in [figure B1.7.17\(a\)](#). Since the ion source region of any instrument has a finite size, the ions will spend a certain amount of time in the source while they are accelerating. If the

initial velocity of the ions is v_0 , the time spent in the source, t_s , is given by

$$t_s = \frac{v_0 m}{zeV/d_s}$$

where v is the velocity of the ion, d_s is the width of the ion source and v/d_s represents the electric field strength inside the source. One obvious way to minimize this effect is to make the field strength as large as possible (by increasing V or decreasing d_s).

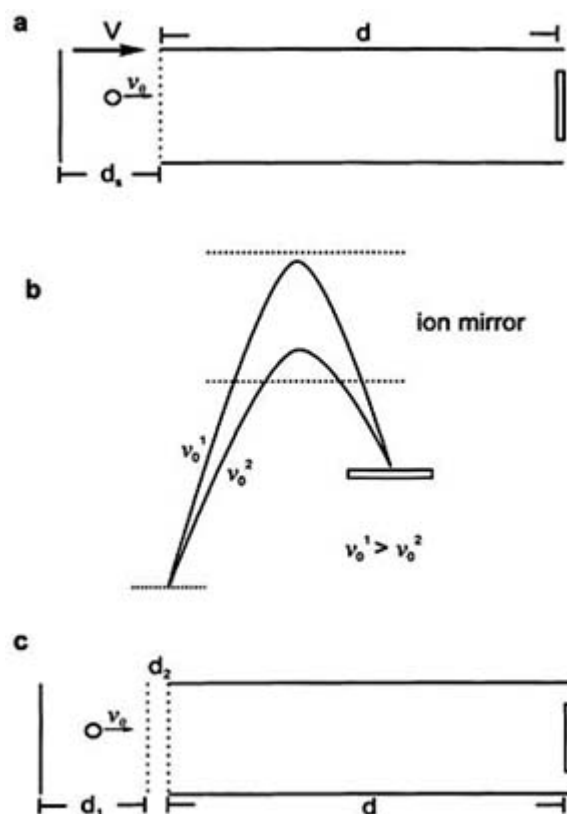


Figure B1.7.17. (a) Schematic diagram of a single acceleration zone time-of-flight mass spectrometer. (b) Schematic diagram showing the time focusing of ions with different initial velocities (and hence initial kinetic energies) onto the detector by the use of a reflecting ion mirror. (c) Wiley–McLaren type two stage acceleration zone time-of-flight mass spectrometer.

Ions generated in the ion source region of the instrument may have initial velocities isotropically distributed in three dimensions (for gaseous samples, this initial velocity is the predicted Maxwell–Boltzmann distribution at the sample temperature). The time the ions spend in the source will now depend on the direction of their initial velocity. At one extreme, the ions may have a velocity v_0 in the direction of the extraction grid. The time spent in the source will be

shorter than those with no component of initial velocity in this direction:

$$t_s = \frac{(v - v_0)m}{zeV/d_s}$$

At the other extreme, ions with initial velocities in the direction opposite to the accelerating potential must first be turned around and brought back to their initial position. From this point their behaviour is the same as described above. The time taken to turn around in the ion source and return to the initial position (t_r) is given by:

$$t_r = \frac{2v_0m}{zeV/d_s}$$

The final velocity of these two ions will be the same, but their final flight times will differ by the above turn-around time, t_r . This results in a broadening of the TOF distributions for each ion mass, and is another limiting factor when considering the mass (time) resolution of the instrument.

The final total ion time of flight in the TOF mass spectrometer with a single accelerating region can be written in a single equation, taking all of the above factors into account.

$$t_{\text{TOF}} = \frac{(2m)^{1/2}[(U_0 + zeV/d_s)^{1/2} \pm U_0^{1/2}]}{zeV/d_s} + \frac{(2m)^{1/2}d}{2(U_0 + zeV)^{1/2}}$$

where now the initial velocity has been replaced by the initial translational energy, U_0 . This is the equation published in 1955 by Wiley and McLaren [37] in their seminal paper on TOF mass spectrometry.

(B) ENERGY FOCUSING AND THE REFLECTRON TOF INSTRUMENT

The resolution of the TOF instrument can be improved by applying energy focusing conditions that serve to overcome the above stated spread in initial translational energies of the generated ions. While there have been several methods developed, the most successful and the most commonly used method is the reflectron. The reflectron is an ion mirror positioned at the end of the drift tube that retards the ions and reverses their direction. Ions with a higher kinetic energy penetrate into the mirror to a greater extent than those with lower kinetic energies. The result is a focusing (in time) at the detector of ions having an initial spread of kinetic energies (figure B1.7.17(b)). The mirror also has the effect of increasing the drift length without increasing the physical length of the instrument.

(C) SPATIAL FOCUSING

Another consideration when gaseous samples are ionized is the variation in where the ions are formed in the source. The above arguments assumed that the ions were all formed at a common initial position, but in practice they may be formed anywhere in the acceleration zone. The result is an additional spread in the final TOF distributions, since ions

made in different locations in the source experience different accelerating potentials and thus spend different

times in the source and drift tube.

Spatial focusing in a single acceleration zone linear TOF instrument naturally occurs at a distance of $2d_s$ along the drift tube, which is seldom a practical distance for a detector. Wiley and McLaren described a two stage accelerating zone that allows spatial focusing to be moved to longer distances. An example of such an instrument is shown in [figure B1.7.17\(c\)](#). The main requirement is for the initial acceleration region to have a much weaker field than the second. The equation relating the instrumental parameters in [figure B1.7.17\(c\)](#) is:

$$d = d_1 k^{3/2} \left(1 - \frac{1}{k + k^{1/2}} \frac{2d_2}{d_1} \right)$$

where $k = (\frac{1}{2}d_1E_1 + d_2E_2)/(\frac{1}{2}d_1E_1)$, and E_1 and E_2 are the field strengths in regions 1 and 2. So, if the physical dimensions of the instrument d_1 , d_2 and d are fixed, a solution can be obtained for the relative field strengths necessary for spatial focusing.

(D) OTHER IONIZATION SOURCES

Other methods of sample introduction that are commonly coupled to TOF mass spectrometers are MALDI, SIMS/FAB and molecular beams (see [section \(B1.7.2\)](#)). In many ways, the ablation of sample from a surface simplifies the TOF mass spectrometer since all ions originate in a narrow space above the sample surface. This reduces many of the complications arising from the need for spatial focusing. Also, the initial velocity of ions generated are invariably in the TOF direction.

Molecular beam sample introduction (described in [section \(B1.7.2\)](#)), followed by the orthogonal extraction of ions, results in improved resolution in TOF instruments over effusive sources. The particles in the molecular beam typically have translational temperatures orthogonal to the beam path of only a few Kelvin. Thus, there is less concern with both the initial velocity of the ions once they are generated and with where in the ion source they are formed (since the particles are originally confined to the beam path).

B1.7.5.2 EXPERIMENTS USING TOF MASS SPECTROMETERS

Time-of-flight mass spectrometers have been used as detectors in a wider variety of experiments than any other mass spectrometer. This is especially true of spectroscopic applications, many of which are discussed in this encyclopedia. Unlike the other instruments described in this chapter, the TOF mass spectrometer is usually used for one purpose, to acquire the mass spectrum of a compound. They cannot generally be used for the kinds of ion–molecule chemistry discussed in this chapter, or structural characterization experiments such as collision-induced dissociation. However, they are easily used as detectors for spectroscopic applications such as multi-photoionization (for the spectroscopy of molecular excited states) [38], zero kinetic energy electron spectroscopy [39] (ZEKE, for the precise measurement of ionization energies) and coincidence measurements (such as photoelectron–photoion coincidence spectroscopy [40] for the measurement of ion fragmentation breakdown diagrams).

Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometry is another in the class of trapping mass spectrometers and, as such is related to the quadrupole ion trap. The progenitor of FT-ICR, the ICR mass spectrometer, originated just after the Second World War when the cyclotron accelerator was developed into a means for selectively detecting ions other than protons. At the heart of ICR is the presence of a magnetic field that confines ions into orbital trajectories about their flight axis. Early ICR experiments mainly took advantage of this trapping and were focused on ion–molecule reactions. The addition of the three-dimensional trapping cell by McIver in 1970 [41, 42] led to improved storage of ions. In 1974 Comisarow and Marshall introduced the Fourier transform detection scheme that paved the way for FT-ICR [43, 44] which is now employed in virtually all areas in physical chemistry and chemical physics that use mass spectrometry.

B1.7.6.1 ION MOTION IN MAGNETIC AND ELECTRIC FIELDS

Figure B1.7.18(a) shows a typical FT-ICR mass spectrometer cubic trapping cell. The principal axes are shown in the diagram, along with the direction of the imposed magnetic field. To understand the trajectory of an ion in such a field, the electrostatic and magnetic forces acting on the ion must be described [45]. If only a magnetic field is present, the field acts on the ions such that they take up circular orbits with a frequency defined by the ion mass:

$$\omega_c = \frac{zeB}{m}$$

where ze is the charge on the ion, B is the magnetic field strength (in tesla), m is the ion mass and ω_c is the ion's cyclotron frequency (in rad s^{-1}). In modern FT-ICR instruments, magnetic fields of 6 T or more are common, with the latest being upwards of 20 T. These high field magnets are usually superconducting magnets, not unlike modern NMR instruments. Without trapping electrodes, the ions would describe a spiral trajectory and be lost from the trap.

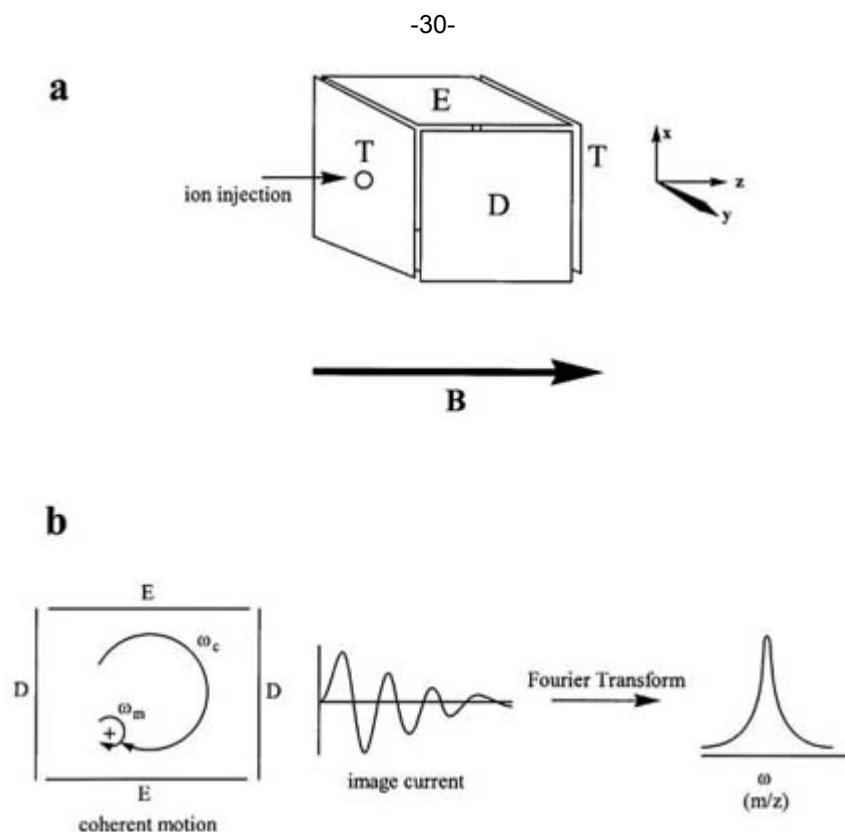


Figure B1.7.18. (a) Schematic diagram of the trapping cell in an ion cyclotron resonance mass spectrometer: excitation plates (E); detector plates (D); trapping plates (T). (b) The magnetron motion (ω_m) due to the crossing of the magnetic and electric trapping fields is superimposed on the circular cyclotron motion (ω_c) taken up by the ions in the magnetic field. Excitation of the cyclotron frequency results in an image current being detected by the detector electrodes which can be Fourier transformed into a secular frequency related to the m/z ratio of the trapped ion(s).

The circular orbits described above are perturbed by an electrostatic field applied to the two endcap trapping electrodes (figure B1.7.18(b)). In addition to the trapping motion, the crossed electric and magnetic fields superimpose a magnetron motion, ω_m , on the ions (figure B1.7.18(b)). An idealized trapping cell would produce a DC quadrupolar potential in three dimensions. The resulting ion motion is independent of the axial and radial position in the cell. In practice, however, the finite size of the trapping cell produces irregularities in the potential that affect ion motion.

(A) ION TRAPPING

The component of the DC quadrupolar potential in the z -axis direction is described by the following equation.

-31-

$$V(z) = \frac{V_T}{2} + \frac{kz^2}{2}$$

where k is a constant and V_T is the trapping potential applied to the endcap electrodes of the trapping cell. The derivative of this relationship yields a linear electric field along the z -axis.

$$E(z) = -kz.$$

From this relationship, an expression can be derived for the trapping frequency, ω_t .

$$\omega_t = \left(\frac{kze}{m} \right)^{1/2}$$

which again is a function of ion charge and mass. In theory, this trapping frequency is harmonic and independent of the ion's position in the trapping cell, but in practice, the finite size of the trapping cell produces irregularities in ω_t . The efficiency with which ions are trapped and stored in the FT-ICR cell diminishes as the pressure in the cell increases (as opposed to the quadrupole ion trap, which requires helium buffer gas for optimal trapping). For this reason, FT-ICR instruments are typically operated below 10^{-5} Torr (and usually closer to 10^{-8} Torr).

(B) ION DETECTION

In the other types of mass spectrometer discussed in this chapter, ions are detected by having them hit a detector such as an electron multiplier. In early ICR instruments, the same approach was taken, but FT-ICR uses a very different technique. If an RF potential is applied to the excitation plates of the trapping cell (figure B1.7.18(b)) equal to the cyclotron frequency of a particular ion m/z ratio, resonant excitation of the ion trajectories takes place (without changing the cyclotron frequency). The result is ion trajectories of higher

kinetic energy and larger radii inside the trapping cell. In addition, all of the ions with that particular m/z ratio take up orbits that are coherent (whereas they were all out of phase prior to resonant excitation). This coherent motion induces an image current on the detector plates of the trapping cell that has the same frequency as the cyclotron frequency (figure B1.7.18(b)). This image current is acquired over a period of time as the ion packet decays back to incoherent motion. The digitized time-dependent signal can be Fourier transformed to produce a frequency spectrum with one component, a peak at the cyclotron frequency of the ions. It is possible to resonantly excite the trajectories of all ions in the trapping cell by the application of a broad-band RF excitation pulse to the excitation electrodes. The resulting time-dependent image current, once Fourier transformed, yields a frequency spectrum with peaks due to each ion m/z in the trapping cell, and hence a mass spectrum. The intensities of the peaks are proportional to the concentrations of the ions in the cell.

B1.7.6.2 EXPERIMENTS USING FT-ICR

In many respects, the applications of FT-ICR are similar to those of the quadrupole ion trap, as they are both trapping instruments. The major difference is in the ion motion inside the trapping cell and the waveform detection. In recent

-32-

years there have been attempts to use waveform detection methods with quadrupole ion traps [46].

(A) COLLISION-INDUCED DISSOCIATION AND ION–MOLECULE REACTIONS

As with the quadrupole ion trap, ions with a particular m/z ratio can be selected and stored in the FT-ICR cell by the resonant ejection of all other ions. Once isolated, the ions can be stored for variable periods of time (even hours) and allowed to react with neutral reagents that are introduced into the trapping cell. In this manner, the products of bi-molecular reactions can be monitored and, if done as a function of trapping time, it is possible to derive rate constants for the reactions [47]. Collision-induced dissociation can also be performed in the FT-ICR cell by the isolation and subsequent excitation of the cyclotron frequency of the ions. The extra translational kinetic energy of the ion packet results in energetic collisions between the ions and background gas in the cell. Since the cell in FT-ICR is nominally held at very low pressures (10^{-8} Torr), CID experiments using the background gas tend not to be very efficient. One common procedure is to pulse a target gas (such as Ar) into the trapping cell and then record the CID mass spectrum once the gas has been pumped away. CID mass spectra obtained in this way are similar to those obtained on triple quadrupole and ion trap instruments.

(B) KINETIC STUDIES

Aside from the bimolecular reaction kinetics described above, it is possible to measure other types of kinetics with FT-ICR. Typically, for two species to come together in the gas phase to form a complex, the resulting complex will only be stable if a three body collision occurs. The third body is necessary to lower the internal energy of the complex below its dissociation threshold. Thus, complexes are generally made in high pressure ion sources. It is possible, however, for the complex to radiatively release excess internal energy. Dunbar and others [48, 49] have studied and modelled the kinetics of such ‘radiative association’ reactions in FT-ICR trapping cells because of the long time scales of the experiments (reaction progress is usually probed for many minutes). The rate constants for photon emission derived from the experimentally observed rate constants tend to be between 10 and 100 s^{-1} . It has also been found that ions can be dissociated by the absorption of blackbody radiation in the trapping cell (BIRD—blackbody infrared radiative dissociation) [50]. This technique, which is only feasible at the low pressures ($<10^{-8}$ Torr) and long trapping times inside the FT-ICR cell, allows the investigator to measure unimolecular decay rate constants of the order of 10^{-3} s^{-1} . Another approach to dissociation kinetics is time-resolved photodissociation [51]. Ions are photodissociated with laser

light in the visible and near UV and the product ion intensity is monitored as a function of time. Rate constants from 10^{-3} s^{-1} and higher can be measured with good precision using this technique.

(C) THERMOCHEMICAL STUDIES: THE BRACKETING METHOD

In an earlier section, measurements were described in which the equilibrium constant, K , for bimolecular reactions involving gas-phase ions and neutral molecules were determined. Another method for determining the proton or other affinity of a molecule is the bracketing method [52]. The principle of this approach is quite straightforward. Let us again take the case of a proton affinity determination as an example. In a reaction between a protonated base, B_1H^+ and a neutral molecule, B_2 , proton transfer from B_1 to B_2 will presumably occur only if the reaction is exothermic (in other words, if the PA of B_2 is greater than that of B_1). So, by choosing a range of bases B_1 covering a range of PA values and reacting them with a molecule of unknown PA, B_2 , the reactions leading to B_2H^+ can be monitored by the presence of this latter ion in the mass spectrum. The PA of B_2 can quickly be narrowed down provided the reference values are well established. The nature of this experiment requires that a bimolecular reaction takes place between the reference

-33-

base and unknown and thus these experiments are most commonly carried out in FT-ICR mass spectrometers, though quadrupole ion trap instruments have also been used.

REFERENCES

- [1] Wannier G H 1953 The threshold law for single ionization of atoms or ions by electrons *Phys. Rev.* **90** 817–25
- [2] Harrison A G 1992 *Chemical Ionization Mass Spectrometry* (Boca Raton, FL: Chemical Rubber Company)
- [3] Barber N, Bordoli R S, Elliot G J, Sedgwick R D and Tyler A N 1982 Fast atom bombardment mass spectrometry *Anal. Chem.* **54** 645A–57A
- [4] Karas M and Hillenkamp F 1988 Laser desorption ionization of proteins with molecular masses exceeding 10,000 Daltons *Anal. Chem.* **60** 2299–301
- [5] Gaskell S J 1997 Electrospray: principles and practice *J. Mass Spectrom.* **32** 677–88
- [6] DePaul S, Pullman D and Friedrich B 1993 A pocket model of seeded molecular beams *J. Phys. Chem.* **97** 2167–71
- [7] Miller D R 1988 Free jet sources *Atomic and Molecular Beam Methods* ed G Scoles (New York: Oxford University Press)
- [8] Thomson J J 1913 *Rays of Positive Electricity* (London: Longmans Green)
- [9] Cooks R G, Beynon J H, Capriolo R M and Lester G R 1973 *Metastable Ions* (Amsterdam: Elsevier)
- [10] Busch K L, Glish G L and McLuckey S A 1988 *Mass Spectrometry/Mass Spectrometry* (New York: VCH)
- [11] Cooks R G (ed) 1978 *Collision Spectroscopy* (New York: Plenum)
- [12] Hamdan M and Brenton A G 1991 High-resolution translational energy spectroscopy of molecular ions *Physics of Ion Impact Phenomena* ed D Mathur (Berlin: Springer)
- [13] Dawson P H 1976 *Quadrupole Mass Spectrometry and its Applications* (Amsterdam: Elsevier)
- [14] March R E and Hughes R J 1989 *Quadrupole Storage Mass Spectrometry* (New York: Wiley-Interscience)

- [15] March R E and Todd J F J 1995 *Practical Aspects of Ion Trap Mass Spectrometry* (Boca Raton, FL: Chemical Rubber Company)
- [16] Mathieu E 1986 *J. Math. Pure Appl.* **13** 137
- [17] Usypchuk L L, Harrison A G and Wang J 1992 Reactive collisions in quadrupole cells. Part I. Reaction of $[\text{CH}_3\text{NH}_2]^+$ with the isomeric butenes and pentenes *Org. Mass Spectrom.* **27** 777–82
- [18] Kebarle P 1988 Pulsed electron high pressure mass spectrometer *Techniques for the Study of Ion-Molecule Reactions* ed J M Farrar and W H Saunders (New York: Wiley–Interscience)
- [19] Szulejko J E and McMahon T B 1991 A pulsed electron beam, variable temperature, high pressure mass spectrometric re-evaluation of the proton affinity difference between 2-methylpropene and ammonia *Int. J. Mass Spectrom. Ion Proc.* **109** 279–94

-34-

- [20] Szulejko J E and McMahon T B 1993 Progress toward an absolute gas-phase proton affinity scale *J. Am. Chem. Soc.* **115** 7839–48
- [21] Berkowitz J, Ellison G B and Gutman D 1994 Three methods to measure RH bond energies *J. Phys. Chem.* **98** 2744–65
- [22] Davidson W R, Sunner J and Kebarle P 1979 Hydrogen bonding of water to onium ions. Hydration of substituted pyridinium ions and related systems *J. Am. Chem. Soc.* **101** 1675–80
- [23] Meot-Ner M 1984 Ionic hydrogen bond and ion solvation 2. Solvation of onium ions by 1–7 water molecules. Relations between monomolecular, specific and bulk hydration *J. Am. Chem. Soc.* **106** 1265–72
- [24] Smith D and Adams N G 1988 The selected ion flow tube (SIFT): studies of ion–neutral reactions *Advances in Atomic and Molecular Physics* ed D Bates and B Bederson (Boston, MA: Academic)
- [25] Ferguson E E, Fehsenfeld F C and Schmeltekopf A L 1969 *Adv. At. Mol. Phys.* **5** 1
- [26] Ferguson E E, Fehsenfeld F C and Albritton D L 1979 Ion chemistry of the earth's atmosphere *Gas Phase Ion Chemistry* vol 1, ed M T Bowers (New York: Academic)
- [27] Squires R R 1997 Atmospheric chemistry and the flowing afterglow technique *J. Mass Spectrom.* **32** 1271–72
- [28] Spanel P and Smith D 1997 SIFT studies of the reactions of H_3O^+ , NO^- and O_2^+ with a series of alcohols *Int. J. Mass Spectrom. Ion Proc.* **167/168** 375–88
- [29] Spanel P, Ji Y and Smith D 1997 SIFT study of the reactions of H_3O^+ , NO^- and O_2^+ with a series of aldehydes and ketones *Int. J. Mass Spectrom. Ion Proc.* **165/166** 25–37
- [30] Spanel P and Smith D 1998 SIFT studies of the reactions of H_3O^+ , NO^- and O_2^+ with a series of volatile carboxylic acids and esters *Int. J. Mass Spectrom. Ion Proc.* **172** 137–47
- [31] Ervin K M and Armentrout P B 1985 Translational energy dependence of $\text{Ar}^+ + \text{XY} \rightarrow \text{ArX}^+ + \text{Y}$ ($\text{XY} = \text{D}_2, \text{D}_2, \text{HD}$) from thermal to 30 eV c.m. *J. Chem. Phys.* **83** 166–89
- [32] Rodgers M T, Ervin K M and Armentrout P B 1997 Statistical modeling of CID thresholds *J. Chem. Phys.* **106** 4499–508
- [33] Asano K, Goeringer D and McLuckey S 1998 Dissociation kinetics in the quadrupole ion trap *Proc. 46th Conf. Am. Soc. Mass Spectrom.*
- [34] Brodbelt J, Liou C-C and Donovan T 1991 Selective adduct formation by dimethyl ether chemical ionization in a quadrupole ion trap mass spectrometer and a conventional ion source *Anal. Chem.* **63** 1205–9
- [35] Gronert S 1998 Estimation of effective ion temperatures in a quadrupole ion trap *J. Am. Soc. Mass Spectrom.* **9** 845–8

- [36] Guilhaus M 1995 Principles and instrumentation in time-of-flight mass spectrometry: physical and instrumental concepts *J. Mass Spectrom.* **30** 1519–32
- [37] Wiley W C and McLaren I H 1955 Time-of-flight mass spectrometer with improved resolution *Rev. Sci. Instrum.* **26** 1150–7
- [38] Nesselrodt D R, Potts A R and Baer T 1995 Stereochemical analysis of methyl-substituted cyclohexanes using 2+1 resonance enhanced multiphoton ionization *Anal. Chem.* **67** 4322–9
-

-35-

- [39] Muller-Dethlefs K and Schlag E W 1991 High resolution zero kinetic energy (ZEKE) photoelectron spectroscopy of molecular systems *Annu. Rev. Phys. Chem.* **42** 109–36
- [40] Baer T 1986 *Adv. Chem. Phys.* **64** 111
- [41] McIver R T 1970 A trapped ion analyzer cell for ion cyclotron resonance spectroscopy *Rev. Sci. Instrum.* **41** 555–8
- [42] Vartanian V H, Anderson J S and Laude D A 1995 Advances in trapped ion cells for Fourier transform ion cyclotron resonance mass spectrometry *Mass Spec. Rev.* **41** 1–19
- [43] Comisarow M B and Marshall A G 1996 Early development of Fourier transform ion cyclotron resonance (FT-ICR) spectroscopy *J. Mass Spectrom.* **31** 581–5
- [44] Amster I J 1996 Fourier transform mass spectrometry *J. Mass Spectrom.* **31** 1325–37
- [45] Freiser B S 1988 Fourier transform mass spectrometry *Techniques for the Study of Ion–Molecule Reactions* ed J M Farrar and W H Saunders (New York: Wiley–Interscience)
- [46] Soni M, Frankevich V, Nappi M, Santini R E, Amy J W and Cooks R G 1996 Broad-band Fourier transform quadrupole ion trap mass spectrometry *Anal. Chem.* **68** 3341–20
- [47] Grover R, Decouzon M, Maria P-C and Gal J-F 1996 Reliability of Fourier transform-ion cyclotron resonance determinations of rate constants for ion/molecule reactions *Eur. Mass Spectrom.* **2** 213–23
- [48] Cheng Y-W and Dunbar R C 1995 Radiative association kinetics of methyl-substituted benzene ions *J. Phys. Chem.* **99** 10 802–7
- [49] Fisher J J and McMahon T B 1990 Determination of rate constants for low pressure association reactions by Fourier transform-ion cyclotron resonance *Int. J. Mass Spectrom. Ion. Proc.* **100** 707–17
- [50] Dunbar R C and McMahon T B 1998 Activation of unimolecular reactions by ambient blackbody radiation *Science* **279** 194–7
- [51] Lin C Y and Dunbar R C 1994 Time-resolved photodissociation rates and kinetic modeling for unimolecular dissociations of iodotoluene ions *J. Phys. Chem.* **98** 1369–75
- [52] Born M, Ingemann S and Nobbering N M M 1994 Heats of formation of mono-halogen substituted carbenes. Stability and reactivity of CHX^- (X = F, Cl, Br and I) radical anions *J. Am. Chem. Soc.* **116** 7210–17
-

FURTHER READING

Cooks R G, Benynon J H, Capriolo R M and Lester G R 1973 *Metastable Ions* (Amsterdam: Elsevier)

This is the seminal book on metastable ions, their chemistry and experimental observation. It is a must for anyone starting out in gas-phase ion chemistry.

Busch K L, Glish G L and McLuckey S A 1988 *Mass Spectrometry/Mass Spectrometry* (New York: VCH)

This is one of the newer books covering tandem spectrometry and is a useful resource for the beginner and experienced mass

spectrometrist.

-36-

Cooks R G (ed) 1978 *Collision Spectroscopy* (New York: Plenum)

This volume deals with the various physical methods for studying collisions between projectile and target species. Theories of collisional scattering, energy transfer and reactive interactions are presented.

Dawson P H 1967 *Quadrupole Mass Spectrometry and its Applications* (Amsterdam: Elsevier)

This is the standard reference volume for the theory and application of quadrupole mass spectrometry.

March R E and Rodd J F J 1995 *Practical Aspects of Ion Trap Mass Spectrometry* (Boca Raton, FL: Chemical Rubber Company)

This is a three volume set covering the physics and chemistry for quadrupole ion traps. It is a must for anyone using traps as part of their research.

Farrar J M and Saunders W H (eds) 1988 *Techniques for the Study of Ion–Molecule Reactions* (New York: Wiley–Interscience)

This volume contains excellent discussions of the various methods for studying ion–molecule reactions in the gas phase, including high pressure mass spectrometry, ion cyclotron resonance spectroscopy (and FT-ICR) and selected ion flow tube mass spectrometry.

-1-

B1.8 Diffraction: x-ray, neutron and electron

Edward Prince

B1.8.1 INTRODUCTION

Diffraction is the deflection of beams of radiation due to interference of waves that interact with objects whose size is of the same order of magnitude as the wavelengths. Molecules and solids typically have interatomic distances in the neighbourhood of a few Ångströms ($1 \text{ \AA} = 10^{-10} \text{ m}$), comparable to the wavelengths of x-rays with energies of the order of 10 keV. Neutrons and electrons also have wave properties, with wavelengths given by the de Broglie relation, $\lambda = h/mv$, where h is Planck's constant, m is the mass of the particle and v is its velocity. Neutron diffraction applications use neutrons with wavelengths in the range from 1 to 10 Å; most electron diffraction applications use wavelengths of the order of 0.05 Å, although low energy electron diffraction (LEED), used for studies of surfaces, employs electrons with wavelengths in the neighbourhood of 1 Å. All three techniques have extensive applications in physics, chemistry, materials science, mineralogy and molecular biology. Although perhaps the most familiar application is determination of the structures of crystalline solids, there are also applications to structural studies of amorphous solids, liquids and gases. Diffraction also plays an important role in imaging techniques such as electron microscopy.

B1.8.2 PRINCIPLES OF DIFFRACTION

B1.8.2.1 THE ATOMIC SCATTERING FACTOR

We shall first discuss the diffraction of x-rays from isolated atoms, because this case illustrates principles that can be generalized to most practical applications. Consider [1] an atom consisting of a nucleus surrounded by a spherically symmetric cloud of electrons that can be represented by a density function $\rho(\mathbf{r})$ and a plane wave that can be described by a vector \mathbf{s}_i normal to the wave front with magnitude $1/\lambda$, where λ is the wavelength. According to Huygens's principle, each point within the electron cloud is the source of a spherical wavelet whose amplitude is proportional to $\rho(\mathbf{r})$. At a distance large compared with the dimensions of the atom this spherical wavelet will approximate a new plane wave and we are interested in the amplitude of the wave formed by the interference of the wavelets originating at all points within the electron cloud. Referring to [figure B1.8.1](#) the difference in pathlength for a wave propagating in the direction of \mathbf{s}_f ($|\mathbf{s}_f| = |\mathbf{s}_i|$) and originating at point \mathbf{r} , relative to the wave originating at the origin, is $\Delta l = \lambda \mathbf{r} \cdot (\mathbf{s}_f - \mathbf{s}_i)$, and the difference in phase is therefore $\Delta\phi = 2\pi i \mathbf{r} \cdot (\mathbf{s}_f - \mathbf{s}_i)$. The *atomic scattering factor*, $f(\mathbf{s}_f)$, is the amplitude of the resultant wave in the direction parallel to \mathbf{s}_f which is the vector sum of all contributions. This is given by

$$f(\mathbf{s}_f) = C \int \rho(\mathbf{r}) \exp[2\pi i \mathbf{r} \cdot (\mathbf{s}_f - \mathbf{s}_i)] d\tau \quad (\text{B1.8.1})$$

where the integral is over the volume of the atom and C is a proportionality constant. $f(\mathbf{s}_f)$ is therefore proportional to the Fourier transform of the electron density distribution, $\rho(\mathbf{r})$.

-2-

To evaluate this integral, first let $\mathbf{Q} = 2\pi(\mathbf{s}_f - \mathbf{s}_i)$, let $Q = |\mathbf{Q}|$, let $r = |\mathbf{r}|$ and let α be the angle between \mathbf{r} and \mathbf{Q} . Now $|\mathbf{s}_f - \mathbf{s}_i| = 2|\mathbf{s}_i| \sin\theta = 2 \sin\theta/\lambda$, so that

$$\mathbf{r} \cdot (\mathbf{s}_f - \mathbf{s}_i) = 2r|\mathbf{s}_i| \sin\theta \cos\alpha = 2r \sin\theta \cos\alpha/\lambda.$$

The area of a ring around \mathbf{Q} with width $d\alpha$ at radius r is $2\pi r^2 \sin\alpha d\alpha$, so

$$d\tau = 2\pi r^2 \sin\alpha d\alpha dr.$$

Because $\rho(\mathbf{r})$ is spherically symmetric, the number of electrons in this volume element is $\rho(r) d\tau$. Letting $x = Qr \cos\alpha$, $d\alpha = -dx/(Qr \sin\alpha)$. Then, making all substitutions,

$$f(\mathbf{Q}) = C \int_0^\infty \frac{2\pi r^2}{Qr} \rho(r) dr \int_{-Qr}^{Qr} \exp(ix) dx \quad (\text{B1.8.2a})$$

$$= 4\pi C \int_0^\infty r^2 \rho(r) \frac{\sin(Qr)}{Qr} dr. \quad (\text{B1.8.2b})$$

If $\theta = 0$, so that $Q = 0$, this reduces to

$$f(0) = 4\pi C \int_0^\infty r^2 \rho(r) dr \quad (\text{B1.8.3})$$

so that the integral is the total charge in the electron cloud. The constant C has the units of a length, and is

conventionally chosen so that $f(Q)$ is a multiple of the ‘classical electron radius’, 2.818×10^{-15} m.

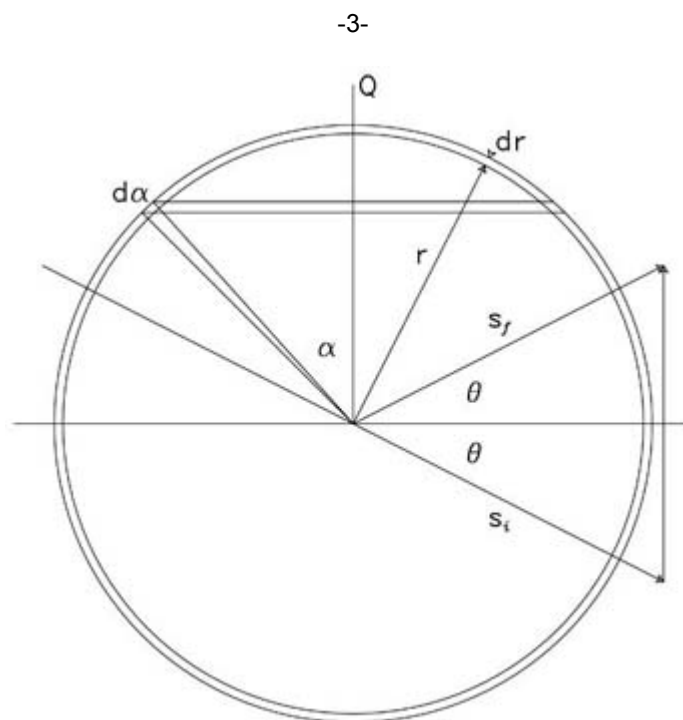


Figure B1.8.1. The atomic scattering factor from a spherically symmetric atom. The volume element is a ring subtending angle α with width $d\alpha$ at radius r and thickness dr .

The electron distribution, $\rho(r)$, has been computed by quantum mechanics for all neutral atoms and many ions and the values of $f(Q)$, as well as coefficients for a useful empirical approximation, are tabulated in the *International Tables for Crystallography* vol C [2]. In general, $f(Q)$ is a maximum equal to the nuclear charge, Z , for $Q = 0$ and decreases monotonically with increasing Q .

Because the neutron has a magnetic moment, it has a similar interaction with the clouds of unpaired d or f electrons in magnetic ions and this interaction is important in studies of magnetic materials. The magnetic analogue of the atomic scattering factor is also tabulated in the *International Tables* [3]. Neutrons also have direct interactions with atomic nuclei, whose mass is concentrated in a volume whose radius is of the order of 10^{-5} times the characteristic neutron wavelength. Thus $\rho(r)$ differs from zero only when $\sin(Qr)/Qr$ is effectively equal to one, so that $f(Q)$ is a constant independent of Q . Whereas the x-ray interaction depends on the total number of electrons in the cloud, and therefore on the nuclear charge, the neutron's interaction with a nucleus results from nuclear forces that vary in a haphazard manner from one isotope to another, lie within a rather narrow range and can even be negative, meaning that the Huygens wavelet from such a nucleus has a phase differing by π from the phase of one from a nucleus whose scattering factor is positive. The neutron scattering factors, or *scattering lengths*, conventionally denoted by b , have magnitudes in the range $1-10 \times 10^{-15}$ m [4].

The atomic scattering factor for electrons is somewhat more complicated. It is again a Fourier transform of a density of scattering matter, but, because the electron is a charged particle, it interacts with the nucleus as well as with the electron cloud. Thus $\rho(r)$ in equation (B1.8.2b) is replaced by $\varphi(r)$, the electrostatic potential of an electron situated at radius r from the nucleus. Under a range of conditions the electron scattering factor, $f_e(Q)$, can be represented in terms

of the x-ray atomic scattering factor by the so-called Mott–Bethe formula,

$$f_e(Q) = 2\pi \frac{me^2}{h^2\epsilon_0} [Z - f_x(Q)]/Q^2 \quad (\text{B1.8.4})$$

where m is the mass of the electron, e is its charge and ϵ_0 is the permittivity of free space.

B1.8.2.2 DIFFRACTION FROM CLUSTERS OF ATOMS

The derivation of [equation \(B1.8.1\)](#) makes no use of the assumption of spherical symmetry and it is, in fact, a very general result that the amplitude of a scattered wave is the Fourier transform of a density of scattering matter. Although there are examples of experimental observations of scattering from isolated atoms, one being the scattering of neutrons by a dilute solid solution of paramagnetic atoms in a diamagnetic matrix, diffraction from molecules, such as that of electrons in a gas, or from particles of contrasting density in a uniform medium are much more important. Examples of the latter are colloidal suspensions in a fluid and precipitates in an alloy. Note that the particles of contrasting density can also be voids: Babinet's principle requires that the amplitude of a scattered wave due to a negative difference be the complex conjugate of that due to a positive difference and, because the intensity of scattered radiation is proportional to the square of the modulus of the amplitude, the diffraction patterns are indistinguishable.

If individual scattering particles are far enough apart and their spatial distribution is such that the relative phases of their contributions to a scattered wave are random, the intensity distribution in the diffraction pattern will be the sum of contributions from all particles [5]. If the particles are identical (monodisperse) but have random orientations, or if they differ in size and shape (polydisperse), the resulting pattern will reflect an ensemble average over the sample. In either case there will be a spreading of the incident beam, so-called *small-angle scattering*. How small the angles are depends on the wavelength of the radiation and the size of the particles: long wavelengths give larger angles, but they also tend to be more strongly absorbed by the sample, so that there is a trade-off between resolution and intensity.

B1.8.2.3 DIFFRACTION FROM CRYSTALLINE SOLIDS

(A) BRAGG'S LAW

The diffraction of x-rays was first observed in 1912 by Laue and coworkers [6]. A plausible, though undocumented, story says that the classic experiment was inspired by a seminar given by P P Ewald, whose doctoral thesis was a purely theoretical study of the interaction of electromagnetic waves with an array of dipoles located at the nodes of a three-dimensional lattice. At the time it was hypothesized that crystals were composed of parallelepipedal building blocks, *unit cells*, fitted together in three dimensions and that x-rays were short-wavelength, electromagnetic radiation, but neither hypothesis had been confirmed experimentally. The Laue experiment confirmed both, but the application of x-ray diffraction to the determination of crystal structure was introduced by the Braggs.

W L Bragg [7] observed that if a crystal was composed of copies of identical unit cells, it could then be divided in many ways into slabs with parallel, plane faces whose distributions of scattering matter were identical and that if the pathlengths travelled by waves reflected from successive, parallel planes differed by integral multiples of the

wavelength there would be strong, constructive interference. Figure B1.8.2 shows a projection of parallel planes separated by a distance d and a plane wave with wavelength λ whose normal makes an angle θ with these reflecting planes. It is evident that the crests of waves reflected from successive planes will be in phase if $\lambda = 2d \sin\theta$, a relation that is known as *Bragg's law*. (It appears that W H Bragg played no role in the formulation of this relation, so it is correctly Bragg's law, not Bragg's law. In textbooks the relation is often stated in the form $n\lambda = 2d \sin\theta$, where n is the 'order' of the reflection, but in crystallography the order is conventionally incorporated in the definition of d .)

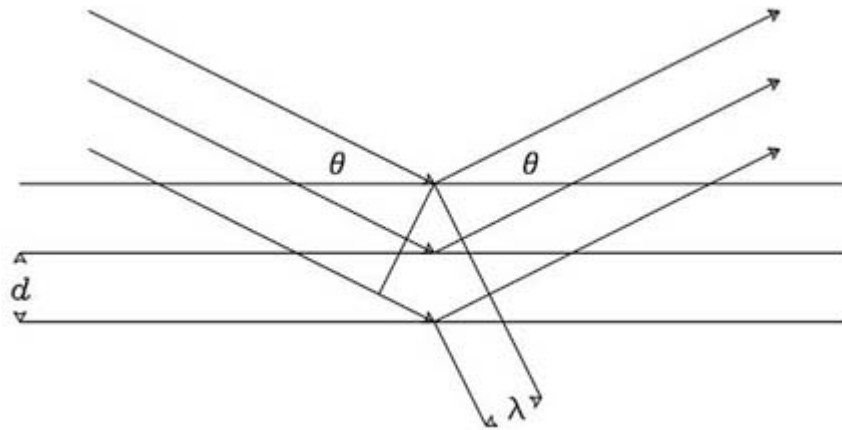


Figure B1.8.2. Bragg's law. When $\lambda = 2d \sin \theta$, there is strong, constructive interference.

(B) THE RECIPROCAL LATTICE

The vertices of the unit cells form an array of points in three-dimensional space, a *space lattice*. The edges of the parallelepiped can be defined by three non-coplanar vectors, \mathbf{a} , \mathbf{b} and \mathbf{c} , and then any lattice point can then be defined by a vector $\mathbf{r} = u\mathbf{a} + v\mathbf{b} + w\mathbf{c}$, where u , v and w are integers. Any point in the crystal can be specified by a vector $\mathbf{r} + \mathbf{x}$, where \mathbf{x} represents a vector within the unit cell. The periodicity of the crystal specifies that $\rho(\mathbf{r} + \mathbf{x}) = \rho(\mathbf{x})$. The families of parallel planes are specified by their *Miller indices*, conventionally denoted by h , k and l . The three points that define the plane closest to the origin are \mathbf{a}/h , \mathbf{b}/k and \mathbf{c}/l , with the understanding that if any of the indices is equal to zero, the plane is parallel to the corresponding vector. To find solutions to the Bragg equation it is necessary to determine the value of d . If \mathbf{a} , \mathbf{b} and \mathbf{c} are orthogonal, this is easy, but for most crystals they are not, and the computation is greatly simplified by use of the *reciprocal lattice*. The reciprocal lattice, which was introduced by J W Gibbs [8] and applied to the crystallographic problem by Ewald [9], is defined by three vectors, $\mathbf{a}^* = (\mathbf{b} \times \mathbf{c})/V$, $\mathbf{b}^* = (\mathbf{c} \times \mathbf{a})/V$ and $\mathbf{c}^* = (\mathbf{a} \times \mathbf{b})/V$, where $V = \mathbf{a} \cdot \mathbf{b} \times \mathbf{c}$ is the volume of the unit cell. It can easily be shown [10] that the vector $\mathbf{d}^* = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$ is perpendicular to the planes defined by the Miller indices h , k and l , and that $|\mathbf{d}^*| = 1/d$. Bragg's law then becomes $\sin\theta = \lambda|\mathbf{d}^*|/2$. Note that if $\mathbf{d}^* = s_f - s_i$, as shown in figure B1.8.3 this condition will be satisfied on the surface of a sphere (the *Ewald sphere*) passing through the origin of the reciprocal lattice whose centre is at the point $-\mathbf{s}_i$ in reciprocal space.

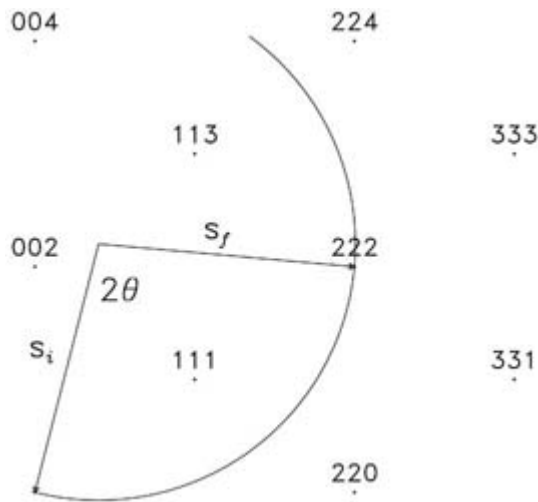


Figure B1.8.3. Ewald's reciprocal lattice construction for the solution of the Bragg equation. If $\mathbf{s}_f - \mathbf{s}_i$ is a vector of the reciprocal lattice, Bragg's law is satisfied for the corresponding planes. This occurs if a reciprocal lattice point lies on the surface of a sphere with radius $1/\lambda$ whose centre is at $-\mathbf{s}_i$.

(C) THE STRUCTURE AMPLITUDE

The amplitude and therefore the intensity, of the scattered radiation is determined by extending the Fourier transform of [equation \(B1.8.1\)](#) over the entire crystal and Bragg's law expresses the fact that this transform has values significantly different from zero only at the nodes of the reciprocal lattice. The amplitude varies, however, from node to node, depending on the transform of the contents of the unit cell. This leads to an expression for the *structure amplitude*, denoted by $F(hkl)$, of the form

$$F(hkl) = C \int_0^c dz \int_0^b dy \int_0^a \rho(x, y, z) \exp \left[2\pi i \left(h \frac{x}{a} + k \frac{y}{b} + l \frac{z}{c} \right) \right] dx \quad (\text{B1.8.5})$$

where $a = |\mathbf{a}|$, $b = |\mathbf{b}|$, $c = |\mathbf{c}|$ and x , y and z are the coordinates of a point in a (not necessarily orthogonal) Cartesian system defined by \mathbf{a} , \mathbf{b} and \mathbf{c} . Making use of the fact that the unit cell contents consist of atoms, each of which has its own atomic scattering factor, $f_j(d^*)$, this can be written

$$F(hkl) = C \sum_{j=1}^N f_j(d^*) \exp \left[2\pi i \left(h \frac{x}{a} + k \frac{y}{b} + l \frac{z}{c} \right) \right] \quad (\text{B1.8.6})$$

where $d^* = |\mathbf{d}^*|$ and the sum is over N atoms in the unit cell.

[Equation \(B1.8.6\)](#) assumes that all unit cells really *are* identical and that the atoms are fixed in their equilibrium positions. In real crystals at finite temperatures, however, atoms oscillate about their mean positions and also may be displaced from their average positions because of, for example, chemical inhomogeneity. The effect of this is, to a first approximation, to modify the atomic scattering factor by a convolution of $\rho(\mathbf{r})$ with a trivariate Gaussian density function, resulting in the multiplication of $f_j(d^*)$ by $\exp(-M)$, where

$$M = 8\pi^2 \overline{u_j^2} \sin^2 \theta / \lambda^2 \quad (\text{B1.8.7})$$

and $\overline{u_j^2}$ is the mean square displacement of the centre of atom j parallel to \mathbf{d}^* . The factor $\exp(-M)$ is the *atomic displacement factor*, or, in older literature, the *temperature factor* or, for early workers in the field, the *Debye–Waller factor*.

(D) DIFFRACTION OF NEUTRONS FROM NONMAGNETIC AND MAGNETIC CRYSTALS

Diffraction of neutrons [11] from nonmagnetic crystals is similar to that of x-rays, with the neutron atomic scattering factor substituted for the x-ray one. For magnetic crystals below their ordering temperatures, however, the neutron's magnetic moment interacts with an ordered array of electron spins, with the strength of the interaction being proportional to $\sin \alpha$, where α is the angle between \mathbf{d}^* and the electron spin axis, and the phases of the wavelets originating at different atoms depend on the relative orientations of their magnetic moments as well as on the path length. The electron spins may all point in the same direction, a *ferromagnet*, or those on different atoms may point in opposite directions, in equal numbers and in an ordered arrangement, an *antiferromagnet*, or they may be arranged in more complicated ways having a net magnetic moment, a so-called *ferrimagnet*. In the absence of an applied magnetic field the crystal tends to divide into domains in which the electron spins point along different, symmetry-equivalent directions and the diffracted intensities are averaged over the various possible values of the angle between the magnetic moment and \mathbf{d}^* . Furthermore, the magnetic diffraction and the nuclear diffraction do not interfere with one another, and the nuclear and magnetic intensities simply add together, although in many cases the magnetic unit cell is larger than the nuclear unit cell, which produces additional diffraction peaks.

If a magnetic field is applied to the crystal, the domains become aligned and the nuclear and magnetic wavelets *do* interfere with one another. Then the amplitude of the diffracted wave depends on the orientation of the neutron spin. In special cases, $\text{Co}_{0.92}\text{Fe}_{0.08}$ is an example, the interference may be totally destructive for one neutron spin state and the diffracted beam becomes polarized. If a crystal of one of these materials is used as a monochromator, the diffraction of this polarized beam is a particularly sensitive probe for the study of magnetic structures.

(E) DIFFRACTION OF ELECTRONS FROM CRYSTALS

Diffraction of electrons from single crystals [12] differs from the diffraction of x-rays or neutrons because the interaction of electrons with matter is much stronger. Conventional electron diffraction is performed as an adjunct to electron microscopy. In fact, the same instrument, a transmission electron microscope, commonly serves for both, with the configuration of the electron optics determining whether a diffraction pattern is magnified, or the diffracted beams are recombined to form an image. Because of the strong interaction, specimens must be thin, and accelerating

voltages must be large, of the order of 100 keV, so that the corresponding wavelength is much shorter than interatomic distances in a crystal, and the Ewald sphere is large compared with the spacing between points of the reciprocal lattice. As a result, when the direction of the incident beam is perpendicular to a reciprocal lattice plane, the small spread of the reciprocal lattice due to mosaic spread in the crystal produces a diffraction pattern that consists of spots with the structure of the reciprocal lattice plane.

Another mode of electron diffraction, low energy electron diffraction or LEED [13], uses incident beams of electrons with energies below about 100 eV, with corresponding wavelengths of the order of 1 Å. Because of the very strong interactions between the incident electrons and the atoms in the crystal, there is very little penetration of the electron waves into the crystal, so that the diffraction pattern is determined entirely by the

arrangement of atoms close to the surface. Thus, in contrast to high energy diffraction, where the pattern is formed by transmission through a thin, crystalline film, and the diffracted beams make angles of only a fraction of a degree with the incident beam, the pattern in LEED is formed by reflection from a surface, and the diffracted beam may be in any direction away from the surface. Furthermore, because there is no significant interference with scattered wavelets coming from below the surface, the reciprocal lattice can be considered to consist not of points but of lines perpendicular to the surface that will always intersect the Ewald sphere, so that even with a monochromatic incident beam there will always be a pattern of spots on a photographic plate or a fluorescent screen.

Although the structure of the surface that produces the diffraction pattern must be periodic in two dimensions, it need not be the same substance as the bulk material. Thus LEED is a particularly sensitive tool for studying the structures and properties of thin layers adsorbed epitaxially on the surfaces of crystals.

B1.8.2.4 DIFFRACTION FROM NONCRYSTALLINE SOLIDS

We have seen that the intensities of diffraction of x-rays or neutrons are proportional to the squared moduli of the Fourier transform of the scattering density of the diffracting object. This corresponds to the Fourier transform of a convolution, $P(\mathbf{s})$, of the form

$$P(\mathbf{s}) = \int \rho(\mathbf{r})\rho(\mathbf{r} + \mathbf{s}) d\mathbf{r}. \quad (\text{B1.8.8})$$

The integrand in this expression will have a large value at a point \mathbf{r} if $\rho(\mathbf{r})$ and $\rho(\mathbf{r}+\mathbf{s})$ are both large, and $P(\mathbf{s})$ will be large if this condition is satisfied systematically over all space. It is therefore a self- or autocorrelation function of $\rho(\mathbf{r})$. If $\rho(\mathbf{r})$ is periodic, as in a crystal, $P(\mathbf{s})$ will also be periodic, with a large peak when \mathbf{s} is a vector of the lattice and also will have a peak when \mathbf{s} is a vector between any two atomic positions. The function $P(\mathbf{s})$ is known as the *Patterson function*, after A L Patterson [14], who introduced its application to the problem of crystal structure determination.

(A) DIFFRACTION FROM GLASSES

There are two classes of solids that are not crystalline, that is, $\rho(\mathbf{r})$ is not periodic. The more familiar one is a glass, for which there are again two models, which may be called the random network and the random packing of hard spheres. An example of the first is silica glass or fused quartz. It consists of tetrahedral SiO_4 groups that are linked at their vertices by Si–O–Si bonds, but, unlike the various crystalline phases of SiO_2 , there is no systematic relation between

the orientations of neighbouring tetrahedra. In the random packing of spheres there is no regular arrangement of atoms even at short range and the coordination of any particular atom may have a wide variety of configurations. The two types of glass have similar diffraction properties, so we do not need to discuss them separately.

If the material is not periodic (but is isotropic), the integral in equation (B1.8.8) becomes spherically symmetric, and reduces for large values of s ($= |\mathbf{s}|$) to a constant equal to the average value of $\rho(\mathbf{r})^2$. In either the sphere-packing model or the random-network model, however, there is always a shortest interatomic distance and $\rho(\mathbf{r})$ falls to a small value between the atoms. The integrand will then have, on average, small values when s is equal to an atomic radius and large values when s is equal to a typical interatomic distance. The integrand and therefore $P(\mathbf{s})$, will have smaller ripples as s increases through additional coordination shells. Because the diffracted intensity is proportional to the Fourier transform of $P(\mathbf{s})$, it will also have broad

maxima and minima as $\sin \theta/\lambda$ increases.

(B) QUASICRYSTALS

The other type of noncrystalline solid was discovered in the 1980s in certain rapidly cooled alloy systems. D Shechtman and coworkers [15] observed electron diffraction patterns with sharp spots with fivefold rotational symmetry, a symmetry that had been, until that time, assumed to be impossible. It is easy to show that it is impossible to fill two- or three-dimensional space with identical objects that have rotational symmetries of orders other than two, three, four or six, and it had been assumed that the long-range periodicity necessary to produce a diffraction pattern with sharp spots could only exist in materials made by the stacking of identical unit cells. The materials that produced these diffraction patterns, but clearly could not be crystals, became known as *quasicrystals*.

Although details of quasicrystal structure remain uncertain, the circumstances under which diffraction patterns with ‘impossible’ symmetries can occur have become clear [16]. It is impossible to construct an object that has long-range periodicity using identical units with these symmetries, but it is not necessary for the object itself to have that symmetry. It is only necessary that its Patterson function be symmetric. The electron diffraction patterns observed by Shechtman actually have the symmetry of a regular icosahedron, and it is possible to build a structure with this symmetry using two rhombohedra, each having faces whose acute angle corners have an angle, α , equal to $2 \arctan[2/(1 + \sqrt{5})] = 63.435^\circ$. One of them has three acute angle corners meeting at a vertex, making a prolate rhombohedron, while the other has three obtuse angle corners meeting at a vertex, making an oblate rhombohedron. Large objects made from these two rhombohedra contain vectors parallel to all of the fivefold axes of the regular icosahedron, although different subsets of them appear in different, finite regions. More importantly, although there is no long range periodicity, departures from periodicity are bounded, which produces, as in a crystal, families of parallel planes with alternately higher and lower density. This in turn produces the observed, sharp diffraction spots.

B1.8.3 STRUCTURE DETERMINATION

We have thus far discussed the diffraction patterns produced by x-rays, neutrons and electrons incident on materials of various kinds. The experimentally interesting problem is, of course, the inverse one: given an observed diffraction pattern, what can we infer about the structure of the object that produced it? Diffraction patterns depend on the Fourier transform of a density distribution, but computing the inverse Fourier transform in order to determine the density distribution is difficult for two reasons. First, as can be seen from [equation \(B1.8.1\)](#), the Fourier transform is

-10-

defined for all values of s_f , but it can be measured only for values of $|s_f - s_i|$ less than $2/\lambda$. For practical reasons λ cannot be arbitrarily small, so that the Fourier transform can never be measured over its entire range. Second, the value of the Fourier transform is in general a complex number. Denoting $-\mathbf{h}$, $-\mathbf{k}$ and $-\mathbf{l}$ by \mathbf{h} , \mathbf{k} , and \mathbf{l} , respectively, [equation \(B1.8.6\)](#) shows that, for a crystal, $I(\mathbf{hkl}) = |F(\mathbf{hkl})|^2 = F(\mathbf{hkl})F(\mathbf{hkl})^*$, it will be the same independent of the phase of $F(\mathbf{hkl})$. As a result, the structural information that can be determined is either restricted to averaged properties that do not depend on phase information, or the phase must be determined by methods other than the simple measurement of diffraction intensities.

B1.8.3.1 SMALL-ANGLE SCATTERING

Materials have many properties that are important, scientifically and technologically, that do not depend on the details of long-range structure. For example, consider a solution of globular macromolecules in a solvent

of contrasting scattering density. If the solution is not too highly concentrated, so that intermolecular interactions can be neglected, the diffraction pattern will be the sum of the diffraction patterns of all individual molecules. Under these conditions all diffracted radiation makes a small angle with the incident beam. Although all molecules are identical, they can have all possible orientations relative to the incident beam, so the diffraction will be that from a spherically averaged distribution. The intensity of diffraction is proportional to the squared modulus of the Fourier transform of the density distribution, which is the Fourier transform of its Patterson function. An expression for the intensity, $I(\mathbf{Q})$, can be derived by substituting $\mathbf{P}(\mathbf{r})$ for $\rho(\mathbf{r})$ in [equation \(B1.8.2b\)](#), giving

(B1.8.9a)

where C is a scale factor dependent on the conditions of the experiment. This can be rewritten

(B1.8.9b)

The inverse of this Fourier sine transform is

(B1.8.10a)

It is conventional to express the structural information in terms of a *pair distance distribution function*, or PDDF [5], which is defined by $p(r) = r^2 P(r)$. Using this, [equation \(B1.8.10\)](#) becomes

(B1.8.10b)

-11-

[Equation \(B1.8.9a\)](#) and [Equation \(B1.8.10b\)](#) both involve integrals whose upper limits are infinite, but this does not present a serious problem, because $\mathbf{P}(\mathbf{r})$ is zero for r greater than the largest diameter of the molecule, and $I(\mathbf{Q})$ has a value significantly different from zero only for small angles. The functions $p(r)$ and $I(\mathbf{Q})$ both contain information about the sizes and shapes of the molecules. It is customary to plot the logarithm of $I(\mathbf{Q})$ as a function of Q and such a plot for a spherical molecule has broad maxima with sharp minima between them, while less symmetric molecules produce curves with smaller ripples or a smooth falloff with increasing angle. A useful property of the molecule is the radius of gyration, R_g , which is a measure of the distribution of scattering density. This may be determined from the relation

$$R_g^2 = \frac{\int_0^\infty r^2 p(r) dr}{\int_0^\infty p(r) dr}. \quad (\text{B1.8.11})$$

The volume of a uniform density molecule may be found from the relation

$$V = C'' \frac{\int_0^\infty p(r) dr}{\int_0^\infty Q^2 I(Q) dQ} \quad (\text{B1.8.12})$$

where $C'' = 8\pi^3$ if the measurements are on an absolute scale. In practice, measurements are relative to a standard sample of known structure.

Although this discussion has been in terms of molecules in solution, the same principles apply to other cases, such as precipitates in an alloy or composites of ceramic particles dispersed in a polymer. The density, $\rho(r)$, is

not relative to a vacuum, but is rather relative to a uniform medium. For x-rays this means electron densities, but for neutrons, because the atomic scattering factor is different from one isotope to another, an effect that is very large for the two stable isotopes of hydrogen, there can be wide variations in contrast depending on the isotopic compositions of the different components of the sample. This ‘contrast variation’ makes small-angle neutron scattering (SANS) a very versatile tool for the study of microstructure.

It has been shown that spherical particles with a distribution of sizes produce diffraction patterns that are indistinguishable from those produced by triaxial ellipsoids. It is therefore possible to assume a shape and determine a size distribution, or to assume a size distribution and determine a shape, but not both simultaneously.

B1.8.3.2 PAIR DISTRIBUTION FUNCTIONS

Another application in which useful information can be obtained in the absence of knowledge of the phase of the Fourier transform is the study of glasses and of crystals that contain short-range order but are disordered over long ranges. Here the objective is to determine a pair distribution function (PDF) [17], which is a generalization of the Patterson function that describes the probability of finding pairs of atoms separated by a vector $\mathbf{r}_j - \mathbf{r}_k$. For various

-12-

reasons these studies are most easily done with neutrons and most of them have been done with glasses. In a glass the long-range structure may be assumed to be isotropic. Setting $r = |\mathbf{r}_j - \mathbf{r}_k|$, the particle density, $\rho(\mathbf{r})$, at distance \mathbf{r} from another particle, can be represented to a good approximation by

$$\rho(\mathbf{r}) - \rho_0 = C \int_0^\infty Q [I_{\text{obs}}(Q) - I_{\text{inc}}] \sin(Qr) \, dQ \quad (\text{B1.8.13})$$

where ρ_0 is the mean overall density and I_{inc} is an isotropic incoherent scattering that is the only source of scattering at sufficiently large Q .

B1.8.3.3 CRYSTAL STRUCTURE DETERMINATION

(A) TRIAL AND ERROR

Laue’s original experiment established that x-rays were short-wavelength electromagnetic radiation and that crystals were composed of periodically repeated arrays of identical units, but it did not establish any scale for the wavelength or the sizes of the crystalline units. W L Bragg [18] observed that the positions of the spots in a diffraction photograph produced by zincblende, ZnS, could be explained by a model (see figure B1.8.4 in which the fundamental units were arranged on a face-centred cubic (f.c.c.) lattice. The same model explained the patterns of sodium chloride, potassium bromide and potassium iodide. (Interestingly, Bragg’s initial model for potassium chloride was based on what is now called a primitive cubic lattice. This was an artifact resulting from the near identity of the atomic scattering factors of potassium and chlorine.) By observing the relative intensities of the diffraction spots and applying elementary principles of group theory, Bragg proposed models for the arrangements of atoms that turned out to be correct.

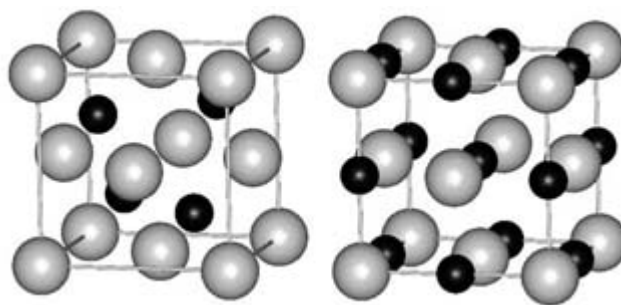


Figure B1.8.4. Two of the crystal structures first solved by W L Bragg. On the left is the structure of zincblende, ZnS. Each sulphur atom (large grey spheres) is surrounded by four zinc atoms (small black spheres) at the vertices of a regular tetrahedron, and each zinc atom is surrounded by four sulphur atoms. On the right is the structure of sodium chloride. Each chlorine atom (grey spheres) is surrounded by six sodium atoms (black spheres) at the vertices of a regular octahedron, and each sodium atom is surrounded by six chlorine atoms.

-13-

At about the same time R A Millikan [19] measured the charge on the electron. Dividing this into the electric charge required to electroplate a gram atomic weight of silver yielded a value, good to within 1%, for Avogadro's number, the number of formula units in a mole. Knowledge of the density and of Avogadro's number leads immediately to knowledge of the size of the unit cell, and thence to knowledge of the wavelength of the radiation producing the diffraction pattern. Bragg's models for the first crystal structures were deduced from Laue photographs, which use continuum radiation, but discovery of the characteristic spectral lines of the elements followed soon after, and H G J Moseley [20] used them to straighten out several anomalies in the periodic table of the elements and to predict the existence of several elements that had not previously been observed.

With the discovery of the x-ray line spectra it became possible to determine, at least relative to the slightly uncertain wavelengths, the sizes and also the symmetries, of the unit cells and the approximate sizes of the atoms. The theory of space groups, which had been worked out by mathematicians, principally A M Schönflies, in the 19th century, has always played a vital role in structural crystallography. With the structures of most of the solid elements and many of their binary compounds it was necessary only to calculate how many atoms would fit into the unit cell and to choose from a limited set of possible positions the ones that best accounted for the relative intensities of the diffraction spots. In sodium chloride, for example, the symmetry of the diffraction pattern shows that the unit cell is a cube and the fact that the indices, h , k and l , are either all odd or all even shows that the cube is face centred. There is room for only four atoms each of sodium and chlorine and it is observed that those reflections with the indices all even are much stronger than those with the indices all odd. This is consistent with a model that has sodium atoms at the corners and at the centres of the faces of the cube and chlorine atoms at the centres of the edges and at the body centre.

Potassium chloride actually has the same structure as sodium chloride, but, because the atomic scattering factors of potassium and chlorine are almost equal, the reflections with the indices all odd are extremely weak, and could easily have been missed in the early experiments. The zincblende form of zinc sulphide, by contrast, has the same pattern of all odd and all even indices, but the pattern of intensities is different. This pattern is consistent with a model that again has zinc atoms at the corners and the face centres, but the sulphur positions are displaced by a quarter of the body diagonal from the zinc positions.

In all of these structures the atomic positions are fixed by the space group symmetry and it is only necessary to determine which of a small set of choices of positions best fits the data. According to the theory of space groups, all structures composed of identical unit cells repeated in three dimensions must conform to one of 230 groups that are formed by combining one of 14 distinct *Bravais lattices* with other symmetry operations.

These include *rotation axes* of orders two, three, four and six and *mirror planes*. They also include *screw axes*, in which a rotation operation is combined with a translation parallel to the rotation axis in such a way that repeated application becomes a translation of the lattice, and *glide planes*, where a mirror reflection is combined with a translation parallel to the plane of half of a lattice translation. Each space group has a *general position* in which the three position coordinates, x , y and z , are independent, and most also have *special positions*, in which one or more coordinates are either fixed or constrained to be linear functions of other coordinates. The properties of the space groups are tabulated in the *International Tables for Crystallography* vol A [21].

The first crystal structure to be determined that had an adjustable position parameter was that of pyrite, FeS_2 . In this structure the iron atoms are at the corners and the face centres, but the sulphur atoms are further away than in zincblende along a different threefold symmetry axis for each of the four iron atoms, which makes the unit cell primitive.

-14-

Unfortunately for modern crystallographers, all of the crystal structures that could be solved by the choose-the-best-of-a-small-number-of-possibilities procedure had been solved by 1920. Bragg has been quoted as saying that the pyrite structure was ‘very complicated’, but he wrote, in about 1930, ‘It must be realized, however, that (cases having one or two parameters) are still extremely simple. The more typical crystal may have ten, twenty, or forty parameters, to all of which values must be assigned before the analysis of the structure is complete.’ This statement is read with amusement by a modern crystallographer, who routinely works with hundreds and frequently with thousands of parameters.

(B) PATTERSON METHODS

We have seen that the intensities of diffraction are proportional to the Fourier transform of the Patterson function, a self-convolution of the scattering matter and that, for a crystal, the Patterson function is periodic in three dimensions. Because the intensity is a positive, real number, the Patterson function is not dependent on phase and it can be computed directly from the data. The squared structure amplitude is

$$|F(hkl)|^2 = I(hkl)/Lp \quad (\text{B1.8.14})$$

where $I(hkl)$ is the integrated intensity of the hkl reflection, L is the so-called *Lorentz factor*, which depends on the experimental geometry and p is a polarization factor, which is equal to one for nuclear scattering of neutrons and depends on the scattering angle, 2θ , for x-rays. From this the Patterson function is

$$P(x, y, z) = \sum_{hkl} |F(hkl)|^2 \cos[2\pi(hx + ky + lz)] \quad (\text{B1.8.15})$$

where the sum is over all values of h , k and l . In practice $I(hkl)$ can be measured only over a finite range of h , k and l and the resulting truncation introduces ripples into the Patterson function.

The Patterson function has peaks corresponding to all interatomic vectors in the density function, the height of a peak being proportional to the product of the atomic scattering factors of the two atoms. Thus, although the Patterson function contains superpositions of the structure as if each atom is in turn placed at the origin and, therefore, has so many peaks that it is difficult to interpret except for very simple structures, there are several features that give important information about the underlying density function. If one or two of the atoms in the unit cell have much higher atomic numbers and, therefore, large values of the atomic scattering factors for

x-rays, the peaks in the Patterson function that correspond to vectors between them will stand out from the rest. Peaks corresponding to vectors between the heavy atoms and lighter ones will also be higher than those corresponding to vectors between lighter atoms, which may reveal features of the environment of the heavy atom. If neutron diffraction is used to study crystals that contain atoms with negative scattering factors, especially hydrogen, but also manganese and titanium, the Patterson function will have negative regions corresponding to vectors between the negative scatterers and other atoms.

If the space group contains screw axes or glide planes, the Patterson function can be particularly revealing. Suppose, for example, that parallel to the c axis of the crystal there is a 2_1 screw axis, one that combines a 180° rotation with

-15-

a translation of $c/2$. Then for an atom at position (x, y, z) there will be another at $(-x, -y, z + \frac{1}{2})$. The section of the Patterson function at $z = \frac{1}{2}$ will therefore contain a peak at position $(2x, 2y)$ for every atom in the unique part of the cell, the *asymmetric unit*. Because this property was first applied to structure determination by D Harker, these special sections of Patterson functions are known as *Harker sections* [22]. If there is a single heavy atom in the asymmetric unit, the Harker section can completely determine the position of the atom. This plays a critical role in the method of *isomorphous replacement*, which we discuss below.

(C) MORE TRIAL AND ERROR

With diffraction data alone, in the absence of phase information, it is always possible to put restrictions on the choice of space group and in many cases it is possible to determine the space group uniquely. Careful measurement of the positions of diffraction spots determines the dimensions of the unit cell and assigns it to one of seven symmetry systems, triclinic, monoclinic, orthorhombic, trigonal, tetragonal, hexagonal, and cubic. The 14 Bravais lattices divide into five basic types, designated primitive, single-face centred, all-face centred, body centred and rhombohedral, which can be distinguished by special patterns of observed and unobserved reflections. We have already discussed the all-face centred lattice, in which the indices are either all odd or all even. In a body centred cell the sum of the indices is always even, while in a primitive cell there are no restrictions.

If one or two of the indices are zeros, there may be additional restrictions. We have seen that a 2_1 screw axis parallel to the c axis of the unit cell produces pairs of atoms at (x, y, z) and $(-x, -y, z + \frac{1}{2})$. From [equation \(B1.8.6\)](#) we can write

$$F(00l) = C \sum_{j=1}^{N/2} f_j(d^*) \{ \exp[2\pi i l z_j] + \exp[2\pi i l (\frac{1}{2} + z_j)] \} \quad (\text{B1.8.16a})$$

which can be written

$$F(00l) = C \sum_{j=1}^{N/2} f_j(d^*) [1 + (-1)^l] \exp(2\pi i l z_j). \quad (\text{B1.8.16b})$$

All terms in the sum vanish if l is odd, so $(00l)$ reflections will be observed only if l is even. Similar restrictions apply to classes of reflections with two indices equal to zero for other types of screw axis and to classes with one index equal to zero for glide planes. These *systematic absences*, which are tabulated in the *International Tables for Crystallography* vol A, may be used to identify the space group, or at least limit the

choices.

The presence of a 2_1 screw axis and a glide plane perpendicular to it implies also the existence of a centre of symmetry, so that, for an atom at (x, y, z) , there is another one at $(-x, -y, -z)$. Equation (B1.8.6) can then be written

$$F(hkl) = C \sum_{j=1}^{N/2} f_j(d^*) [\exp[2\pi i(hx_j + ky_j + lz_j)] + \exp[-2\pi i(hx_j + ky_j + lz_j)]]. \quad (\text{B1.8.17})$$

-16-

The two exponential terms are complex conjugates of one another, so that all structure amplitudes must be real and their phases can therefore be only zero or π . (Nearly 40% of all known structures belong to monoclinic space group $P2_1/c$. The systematic absences of $(0k0)$ reflections when k is odd and of $(h0l)$ reflections when l is odd identify this space group and show that it is centrosymmetric.) Even in the absence of a definitive set of systematic absences it is still possible to infer the (probable) presence of a centre of symmetry. A J C Wilson [23] first observed that the probability distribution of the magnitudes of the structure amplitudes would be different if the amplitudes were constrained to be real from that if they could be complex. Wilson and co-workers established a procedure by which the frequencies of suitably scaled values of $|F|$ could be compared with the theoretical distributions for centrosymmetric and noncentrosymmetric structures. (Note that Wilson named the statistical distributions *centric* and *acentric*. These were not intended to be synonyms for centrosymmetric and noncentrosymmetric, but they have come to be used that way.)

The knowledge that a crystal structure is centrosymmetric reduces the phase problem to one of determining signs, but it is still a formidable one. An extended trial-and-error method uses all available information, including that derived from Patterson methods, numbers of special positions in the unit cell, known interatomic distances, likely group configurations etc, to guess a trial structure and compute from it a set of signs, which are then used to compute a density map, or, more likely, a difference map, in which the Fourier coefficients are the differences between the values of $F(hkl)$ computed from the trial structure and their observed values. Features of the difference map suggest modifications to the trial structure and a new set of signs is used to compute an updated map. With luck, this procedure will converge in a few iterations to a reasonable structure.

(D) DIRECT METHODS

As the number of atoms in the asymmetric unit increases, the solution of a structure by any of these phase-independent methods becomes more difficult, and by 1950 a PhD thesis could be based on a single crystal structure. At about that time, however, several groups observed that the fact that the electron density must be non-negative everywhere could be exploited to place restrictions on possible phases. The first use of this fact was by D Harker and J S Kasper [24], but their relations were special cases of more general relations introduced by J Karle and H Hauptman [25]. Denoting by h_i the set of indices h_i, k_i, l_i , the Karle–Hauptman condition states that all matrices of the form

$$\begin{pmatrix} F(\mathbf{0}) & F(\mathbf{h}_1) & F(\mathbf{h}_2) & \cdots & F(\mathbf{h}_n) \\ F^*(\mathbf{h}_1) & F(\mathbf{0}) & F(\mathbf{h}_2 - \mathbf{h}_1) & \cdots & F(\mathbf{h}_n - \mathbf{h}_1) \\ F^*(\mathbf{h}_2) & F^*(\mathbf{h}_2 - \mathbf{h}_1) & F(\mathbf{0}) & \cdots & F(\mathbf{h}_n - \mathbf{h}_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ F^*(\mathbf{h}_n) & F^*(\mathbf{h}_n - \mathbf{h}_1) & F^*(\mathbf{h}_n - \mathbf{h}_2) & \cdots & F(\mathbf{0}) \end{pmatrix}$$

must be positive definite. Defining $U(h_i)$ by $U(h_i) = F(h_i)/F(\mathbf{0})$ and taking a 3×3 matrix for an example, this

condition implies that the determinant

$$D(\mathbf{h}_1, \mathbf{h}_2) = \begin{vmatrix} 1 & U(\mathbf{h}_1) & U(\mathbf{h}_2) \\ U^*(\mathbf{h}_1) & 1 & U(\mathbf{h}_2 - \mathbf{h}_1) \\ U^*(\mathbf{h}_2) & U^*(\mathbf{h}_2 - \mathbf{h}_1) & 1 \end{vmatrix} \geq 0. \quad (\text{B1.8.18})$$

-17-

From this some tedious but straightforward algebra leads to

$$(\text{B1.8.19})$$

The two factors on the right are both positive, real numbers less than one. If the magnitudes of $U(\mathbf{h}_1)$ and $U(\mathbf{h}_2)$ are both close to one, therefore, the magnitude of the difference between the terms within the brackets on the left (complex numbers in general) must be small.

Karle and Hauptman showed that the fact that the crystal is composed of discrete atoms implies that large enough determinants of the type in [inequality \(B1.8.18\)](#) must vanish, leading to exact relations among sets of phases. For structures with moderately large numbers of atoms in the asymmetric unit ‘large enough’ may be very large indeed, and relations such as [inequality \(B1.8.19\)](#) may not represent much of a restriction on the phase of $U(\mathbf{h}_2 - \mathbf{h}_1)$. Further development of these principles by Hauptman, Karle, I L Karle, M M Woolfson and many others [26] showed that, although no one of these relations would put a significant restriction on a phase, reasonable assumptions about probability distributions would lead to statistical tests that assigned high probabilities to sufficient numbers of phases so that the correct structure could be identified in a density map. These developments, together with the revolution in computing power, have made the solution of structures with up to several hundred atoms in the asymmetric unit a matter of routine.

(E) ISOMORPHOUS REPLACEMENT

While direct methods have opened up structural chemistry with hundreds of atoms in the asymmetric unit, many of the most interesting studies are of biological macromolecules, particularly proteins, which may have thousands of atoms in the asymmetric unit. Furthermore, all biological molecules are chiral, which means that the space groups in which they crystallize can never possess centres of symmetry or mirror (or glide) planes. Although the phases of some sets of reflections may be restricted by symmetry, most structure amplitudes are complex and with large structures the statistical techniques do not supply sufficient information to be useful. The first successful method of determining phases in macromolecular structure studies was the method of *isomorphous replacement*, in which a crystal of a protein is treated chemically to incorporate a small number of heavy atoms into the crystal without disturbing very much the arrangement of the protein molecules. In favourable cases two or more heavy-atom derivatives can be prepared in which the arrangements of the heavy atoms are different. The contribution of the protein molecule to the structure amplitude is assumed to be the same in the derivatives as in the native protein and the interatomic vectors of the heavy atoms stand out sufficiently in a Patterson map to allow the heavy-atom positions to be determined.

Referring to [figure B1.8.5](#) the radii of the three circles are the magnitudes of the observed structure amplitudes of a reflection from the native protein, F_p , and of the same reflection from two heavy-atom derivatives, F_{d1} and F_{d2} . We assume that we have been able to determine the heavy-atom positions in the derivatives and F_{h1} and F_{h2} are the calculated heavy-atom contributions to the structure amplitudes of the derivatives. The centres of the derivative circles are at points $-F_{h1}$ and $-F_{h2}$ in the complex plane, and the three circles intersect at one point, which is therefore the complex value of F_p . The phases for as many reflections as possible can then be

used to compute a density map. The protein molecule is a chain of amino acid residues, whose sequence can be determined by biochemical means and each

-18-

residue consists of a backbone portion, whose length is essentially the same for all amino acids and a side chain. With lots of both skill and luck, and sophisticated computer hardware and software, a model of the chain can be fitted into the density map to obtain a trial structure that can be refined.

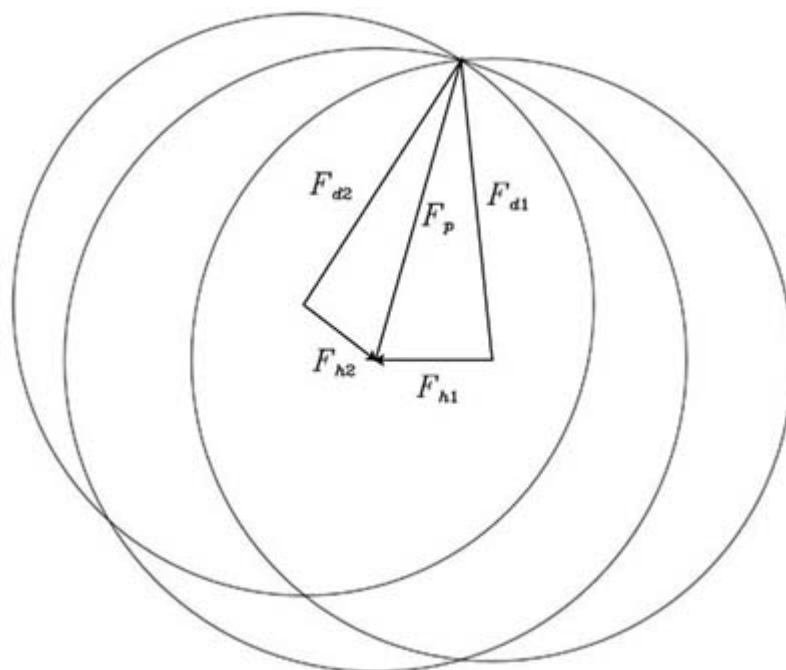


Figure B1.8.5. F_p , F_{d1} and F_{d2} are the measured structure amplitudes of a reflection from a native protein and from two heavy-atom derivatives. F_{h1} and F_{h2} are the heavy atom contributions. The point at which the three circles intersect is the complex value of F_p .

(F) MULTIPLE-WAVELENGTH ANOMALOUS DIFFRACTION

A technique that employs principles similar to those of isomorphous replacement is *multiple-wavelength anomalous diffraction* (MAD) [27]. The expression for the atomic scattering factor in [equation \(B1.8.2b\)](#) is strictly accurate only if the x-ray wavelength is well away from any characteristic absorption edge of the element, in which case the atomic scattering factor is real and $F(\overline{hkl}) = F(hkl)^*$. Since the diffracted intensity is proportional to $|F(hkl)|^2$, the diffraction process in effect introduces a centre of symmetry into all data, a fact that is known as Friedel's law. If the wavelength is near an absorption edge, however, the atomic scattering factor becomes complex and the phases of the contributions of an atom to $F(hkl)$ and $F(\overline{hkl})$ differ. The increasingly widespread availability of synchrotron radiation has made it possible to collect diffraction data at several wavelengths, including some near an absorption edge of one or more of the elements in the crystal. The differences between the intensities of hkl and $F(\overline{hkl})$ then serve in a role similar to that played by the differences between the intensities of native and derivative data in isomorphous replacement.

-19-

B1.8.4 EXPERIMENTAL TECHNIQUES

There are many experimental techniques for diffraction studies, depending on whether the materials producing the diffraction are crystalline or amorphous solids, liquids or gases. Crystalline materials are further subdivided according to whether the sample is a single crystal or a powder composed of many small crystals, frequently of more than one phase. All techniques include a source of radiation, a system for holding and manipulating the sample and a means of detecting the scattered radiation.

B1.8.4.1 SOURCES OF RADIATION

(A) X-RAYS

X-rays for diffraction are generated in two ways. The most common is to bombard a metallic anode in a vacuum tube with electrons emitted thermionically from a hot cathode, thereby exciting the characteristic radiation from the anode material, which is usually copper or molybdenum, although some other metals are used for special purposes. If the accelerating voltage in the tube is well above that required to eject a K shell electron from an atom of the anode material, most of the x-radiation emitted will be in the characteristic lines of the K series on top of a continuous, *Bremsstrahlung* spectrum. $K\beta$ and higher energy lines may be filtered out using a suitable metallic filter, or the characteristic line may be selected by reflection from a monochromator crystal.

The other type of x-ray source is an electron synchrotron, which produces an extremely intense, highly polarized and, in the direction perpendicular to the plane of polarization, highly collimated beam. The energy spectrum is continuous up to a maximum that depends on the energy of the accelerated electrons, so that x-rays for diffraction experiments must either be reflected from a monochromator crystal or used in the Laue mode. Whereas diffraction instruments using vacuum tubes as the source are available in many institutions worldwide, there are synchrotron x-ray facilities only in a few major research institutions. There are synchrotron facilities in the United States, the United Kingdom, France, Germany and Japan.

(B) NEUTRONS

Neutrons for diffraction experiments are also produced in two ways. Thermal neutrons from a nuclear reactor are reflected from a monochromator crystal and Bragg's law is satisfied for neutrons scattered from the sample by measuring the scattering angle, 2θ . In a spallation source short pulses of protons bombard a heavy metal target and high energy neutrons are produced by nuclear reactions. These neutrons interact with a moderator, giving a somewhat longer pulse of neutrons with a spectrum that extends down to thermal energies and therefore to wavelengths up to a few Ångströms. Diffraction from a sample a few metres away from the moderator is observed at a fixed angle and the relation between wavelength and velocity causes Bragg's law to be satisfied at some time after the initial pulse.

As with synchrotron x-rays, neutron diffraction facilities are available at only a few major research institutions. There are research reactors with diffraction facilities in many countries, but the major ones are in North America, Europe and Australia. There are fewer spallation sources, but there are major ones in the United States and the United Kingdom.

(C) ELECTRONS

As noted earlier, most electron diffraction studies are performed in a mode of operation of a transmission electron microscope. The electrons are emitted thermionically from a hot cathode and accelerated by the electric field of a conventional electron gun. Because of the very strong interactions between electrons and matter, significant diffracted intensities can also be observed from the molecules of a gas. Again, the source of electrons is a conventional electron gun.

B1.8.4.2 DETECTORS

Detectors for the three types of radiation are similar and may be classified in two categories, photographic and electronic. In addition to photographic films and plates, photographic detectors also include fluorescent screens and image plates, in which x-rays produce a latent image in a storage phosphor. In the dark the phosphor emits radiation very slowly, but exposure to light from a laser stimulates fluorescence, which then can be observed by a photomultiplier tube and converted to an electronic signal. Because neutrons interact weakly with most materials, image plates and fluorescent screens must contain one of the elements, such as gadolinium, that have isotopes with high absorption cross-sections. Photographic detection of neutrons usually uses a fluorescent screen to enhance the image.

There are many types of electronic detector. The original form of electronic detector was the Geiger counter, but it was replaced many years ago by the proportional counter, which allows selection of radiation of a particular type or energy. Proportional counters for x-rays are filled with a gas such as xenon, and those for neutrons are filled with a gas containing a neutron-absorbing isotope, usually ^3He . Recently these gases have been used to construct position-sensitive area detectors for both x-rays and neutrons.

B1.8.4.3 SINGLE-CRYSTAL DIFFRACTION

Many different geometrical arrangements are commonly used for measurements of diffraction of x-rays and neutrons from single crystals. All have a mechanism for setting the crystal so that Bragg's law is satisfied for some set of crystal planes and for placing a detector in the proper position to observe the reflection. In the original x-ray diffraction experiments of Laue and co-workers the x-rays had a broad spectral distribution, so that for any angular position of a crystal and any interplanar spacing there were x-rays with the proper wavelength to satisfy Bragg's law. Laue photographs reveal the internal symmetry of the crystal and are therefore used to determine the symmetry and orientation of the crystal. For crystal structure determination it is necessary to measure accurate intensities and it is usual to use a monochromatic beam of x-rays or neutrons.

For diffraction studies with monochromatic radiation, the crystal is commonly mounted on an Eulerian cradle, which can rotate the crystal so that the normal to any set of planes bisects the angle between the incident and reflected beams, which is set for reflection from planes with a particular value of the interplanar spacing, d .

If the detection system is an electronic, area detector, the crystal may be mounted with a convenient crystal direction parallel to an axis about which it may be rotated under the control of a computer that also records the diffracted intensities. Because the orientation of the crystal is known at the time an x-ray photon or neutron is detected at a particular point on the detector, the indices of the crystal planes causing the diffraction are uniquely determined. If

films are used, additional information is needed to index the pattern. The crystal is either oscillated through a narrow angular range, so that only a small number of planes can come into the reflecting position, or the film is moved, with a mask covering all but a small part of it, so that the exposed part of it is coordinated with the angular position of the crystal. There are two common moving-film methods, the Weissenberg method, in

which a cylindrical film is moved parallel to the rotation axis of the crystal and the precession method, in which the normal to a reciprocal lattice plane is moved along a circular cone while a flat film and a circular slit are both moved in such a way that the positions of the spots on the film correspond to the points of the reciprocal lattice plane.

One form of electron diffraction is similar to the precession method, except that the ‘single crystal’ is a grain of a polycrystalline foil. Figure B1.8.6 shows an electron diffraction pattern produced when the beam is directed down a fivefold symmetry axis of a quasicrystal. Because of the very short wavelength the cone angle is so small that it lies within the mosaic spread of the grain, and the resulting diffraction pattern, after magnification by the electron optics, closely resembles a precession pattern made with x-rays. In this technique the divergence of the electron beam is extremely small, and the diffraction spots correspond to lattice points in a plane of the reciprocal lattice that passes through the origin. The diffraction pattern therefore has a centre of symmetry. In convergent beam electron diffraction (CBED) [28] (see [figure B1.8.7](#) the divergence of the electron beam is still only a few tenths of a degree, but the resultant smearing of the Ewald sphere allows it to intersect layers of the reciprocal lattice adjacent to the one passing through the origin, so that a region of broadened diffraction spots is surrounded by one or more rings of additional spots corresponding to points in these adjacent planes. Because Friedel’s law does not apply in those planes, the pattern more closely reflects the true symmetry of the crystal.

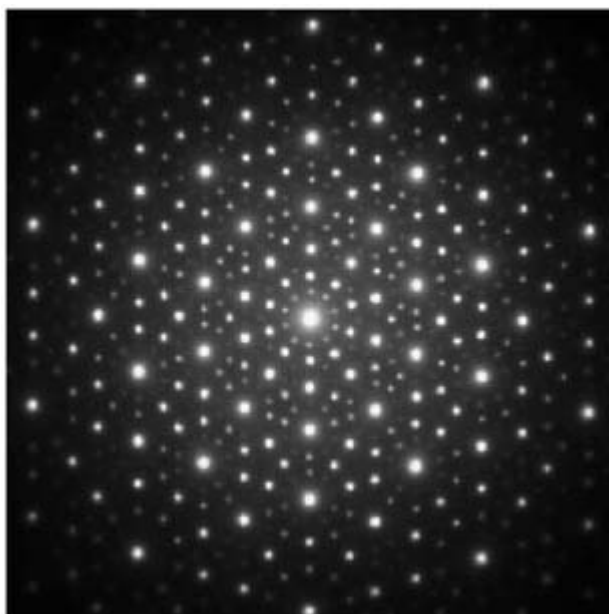


Figure B1.8.6. An electron diffraction pattern looking down the fivefold symmetry axis of a quasicrystal. Because Friedel’s law introduces a centre of symmetry, the symmetry of the pattern is tenfold. (Courtesy of L Bendersky.)

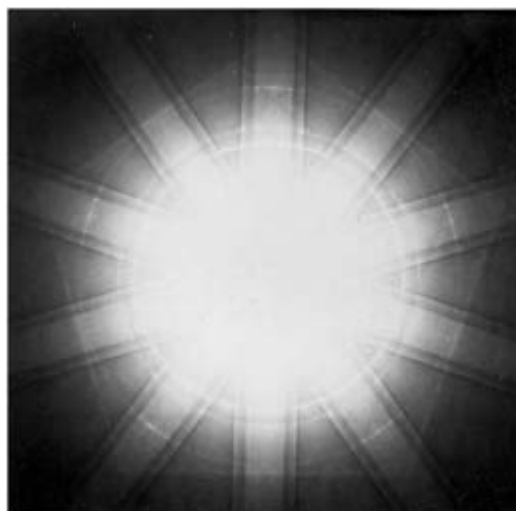


Figure B1.8.7. A convergent beam diffraction pattern of the fivefold axis of a quasicrystal, as in [figure B1.8.6](#). The diffraction rings show that the symmetry is fivefold, not tenfold. (Courtesy of L Bendersky.)

An experimental technique that is useful for structure studies of biological macromolecules and other crystals with large unit cells uses neither the broad, ‘white’, spectrum characteristic of Laue methods nor a sharp, monochromatic spectrum, but rather a spectral band with $\Delta\lambda/\lambda \approx 20\%$. Because of its relation to the Laue method, this technique is called *quasi-Laue*. It was believed for many years that the Laue method was not useful for structure studies because reflections of different orders would be superposed on the same point of a film or an image plate. It was realized recently, however, that, if there is a definite minimum wavelength in the spectral band, more than 80% of all reflections would contain only a single order. Quasi-Laue methods are now used with both neutrons and x-rays, particularly x-rays from synchrotron sources, which give an intense, white spectrum.

B1.8.4.4 POWDER DIFFRACTION

Many scientifically and technologically important substances cannot be prepared as single-crystals large enough to be studied by single crystal diffraction of x-rays and, especially, neutrons. If a sample composed of a very large number (of order 10^{10} or more) of very small ($10\ \mu\text{m}$ or smaller) crystals are irradiated by a monochromatic beam of x-rays or neutrons, there will be some crystals with the right orientation to reflect from all possible sets of crystal planes with interplanar spacings greater than $\lambda/2$. The resulting diffraction pattern contains intensity peaks that are characteristic for any crystalline compound. The pattern corresponds to a uniform distribution of reciprocal lattice points on the surface of a sphere and analysis of the pattern to determine the size and shape of the unit cell can be a difficult (but generally computationally tractable) problem. Nevertheless, in addition to structure determination, powder diffraction is an extremely powerful tool for phase identification and, because a mixture of crystalline phases will give the characteristic patterns of all phases present, quantitative phase analysis. Also if, because of mechanical deformation, for example, the powder sample is not spherically uniform, powder diffraction can reveal the nature of the preferred orientation. Because of the weak interaction and, therefore, high penetration of the neutron, neutron powder diffraction is particularly useful for preferred orientation (texture) studies of bulk materials.

X-ray powder diffraction studies are performed both with films and with counter diffractometers. The powder photograph was developed by P Debye and P Scherrer and, independently, by A W Hull. The Debye–Scherrer camera has a cylindrical specimen surrounded by a cylindrical film. In another commonly used powder

camera, developed by A Guinier, a convergent beam from a curved, crystal monochromator passes through a thin, flat sample and is focused on the film. The common x-ray powder diffractometer uses so-called Bragg–Brentano (although it was apparently developed by W Parrish) focusing. A divergent beam from the line focus of an x-ray tube is reflected from a flat sample and comes to an approximate focus at a receiving slit.

Powder diffraction studies with neutrons are performed both at nuclear reactors and at spallation sources. In both cases a cylindrical sample is observed by multiple detectors or, in some cases, by a curved, position-sensitive detector. In a powder diffractometer at a reactor, collimators and detectors at many different 2θ angles are scanned over small angular ranges to fill in the pattern. At a spallation source, pulses of neutrons of different wavelengths strike the sample at different times and detectors at different angles see the entire powder pattern, also at different times. These slightly displaced patterns are then ‘time focused’, either by electronic hardware or by software in the subsequent data analysis.

B1.8.5 FRONTIERS

Starting from the truly heroic solution of the structure of penicillin by D C Hodgkin (née Crowfoot) and coworkers [29], x-ray diffraction has been the means of molecular structure determination (and the basis of many Nobel prizes in addition to Hodgkin’s) of important compounds, including natural products, of which minute quantities were available for analysis, enabling chemical synthesis and further study. Knowledge of the molecular structure leads in turn to an understanding of reaction mechanisms and, in the case of biological molecules in particular, to an understanding of enzyme function and how drugs can be designed to promote desirable reactions and inhibit undesirable ones.

The development of neutron diffraction by C G Shull and coworkers [30] led to the determination of the existence, previously only a hypothesis, of antiferromagnetism and ferrimagnetism. More recently neutron diffraction, because of its sensitivity to light elements in the presence of heavy ones, played a crucial role in demonstrating the importance of oxygen content in high-temperature superconductors.

The development of synchrotron x-ray sources has resulted in a vast expansion of the capability of x-ray diffraction for determining macromolecular structure, but advances are still limited by the rarity and expense of synchrotron facilities. Correspondingly, the use of neutron diffraction has always been inhibited by the relatively low intensities available and the resulting need for large samples and long data collection times. With both synchrotrons and neutron sources observation time at existing facilities is chronically oversubscribed. Thus there is a need to develop both instruments and methodologies for maximum utilization of the sources.

REFERENCES

- [1] James R W 1965 *The Optical Principles of the Diffraction of X-Rays* (Cornell University Press) ch III
- [2] Maslen E N, Fox A G and O’Keefe M A 1999 X-ray scattering *International Tables for Crystallography* 2nd edn, vol C, ed A J C Wilson and E Prince (Dordrecht: Kluwer) section 6.1.1
- [3] Brown P J 1999 Magnetic form factors *International Tables for Crystallography* 2nd edn, vol C, ed A J C Wilson and E Prince (Dordrecht: Kluwer) section 4.4.5

- [4] Sears V F 1999 Scattering lengths for neutrons *International Tables for Crystallography* 2nd edn, vol C, ed A J C Wilson and E Prince (Dordrecht: Kluwer) section 4.4.4
- [5] Glatter O and May R 1999 Small angle techniques *International Tables for Crystallography* 2nd edn, vol C, ed A J C Wilson and E Prince (Dordrecht: Kluwer) chapter 2.6
- [6] Friedrich W, Knipping P and von Laue M 1912 Interferenz-Erscheinungen bei Röntgenstrahlen *Sitzungsberichte der Königlich Bayerischen Akademie der Wissenschaften zu München* pp 303–22
- [7] Bragg W L 1913 The diffraction of short electromagnetic waves by a crystal *Proc. Camb. Phil. Soc.* **17** 43–58
- [8] Gibbs J W 1928 *Elements of Vector Analysis (Collected Works of J. Willard Gibbs Vol II, Part 2)* (New York: Longmans Green)
- [9] Ewald P P 1921 Das reziproke Gitter in der Strukturtheorie *Z. Kristallogr.* **56** 129–56
- [10] Prince E 1994 *Mathematical Techniques in Crystallography and Materials Science* 2nd edn (Heidelberg: Springer)
- [11] Bacon G E 1962 *Neutron Diffraction* 2nd edn (Oxford: Clarendon)
- [12] Thomas G and Goringe M J 1981 *Transmission Electron Microscopy of Materials* (New York: Wiley)
- [13] van Hove M A, Weinberg W H and Chan C-H 1986 *Low Energy Electron Diffraction: Experiment, Theory, and Surface Structure Determination* (Berlin: Springer)
- [14] Patterson A L 1934 A Fourier series method for the determination of the components of interatomic distances in crystals *Phys. Rev.* **46** 372–6
- [15] Shechtman D, Blech I, Gratias D and Cahn J W 1984 Metallic phase with long range orientational order and no translational symmetry *Phys. Rev. Lett.* **53** 1951–3
- [16] Prince E 1987 Diffraction patterns from tilings with fivefold symmetry *Acta Crystallogr. A* **43** 393–400
- [17] Toby B H and Egami T 1992 Accuracy of pair distribution function analysis applied to crystalline and noncrystalline materials *Acta Crystallogr. A* **48** 336–46
- [18] Bragg W L 1913 The structure of some crystals as indicated by their diffraction of X-rays *Proc. R Soc. A* **89** 248–60
- [19] Millikan R A A new modification of the cloud method of determining the elementary electrical charge and the most probable value of that charge *Phil. Mag.* **19** 209–28

- [20] Moseley H G J 1913 The high-frequency spectra of the elements *Phil. Mag.* **26** 1024–34
- [21] Hahn Th (ed) 1992 *International Tables for Crystallography* vol A (Dordrecht: Kluwer)
- [22] Harker D 1936 The application of the three-dimensional Patterson method and the crystal structures of proustite, Ag_3AsS_3 , and pyrrargyrite, Ag_3SnS_3 *J. Chem. Phys.* **4** 381–90
- [23] Wilson A J C 1949 The probability distribution of X-ray intensities *Acta Crystallogr.* **2** 318–21
- [24] Harker D and Kasper J S 1948 Phases of Fourier coefficients directly from crystal diffraction data *Acta Crystallogr.* **1** 70–5
- [25] Karle J and Hauptman H 1950 The phases and magnitudes of the structure factors *Acta Crystallogr.* **3** 181–7
- [26] Woolfson M M 1987 Direct methods—from birth to maturity *Acta Crystallogr. A* **43** 593–612
- [27] Hendrickson W A 1991 Determination of macromolecular structures from anomalous diffraction of synchrotron radiation *Science* **254** 51–8
- [28] Tanaka M and Terauchi M 1985 *Convergent-Beam Electron Diffraction* JEOL, Tokyo

- [29] Crowfoot D, Bunn C W, Rogers-Low B W and Turner-Jones A 1949 The X-ray crystallographic investigation of the structure of penicillin *Chemistry of Penicillin* ed H T Clarke, J R Johnson and R Robinson (Princeton, NJ: Princeton University Press) pp 310–66
- [30] Shull C G, Strauser W A and Wollan E O 1951 Neutron diffraction by paramagnetic and antiferromagnetic substances *Phys. Rev.* **83** 333–45
-

FURTHER READING

Wilson A J C and Prince E (eds) 1999 *Mathematical, Physical and Chemical Tables (International Tables for Crystallography C)* 2nd edn (Dordrecht: Kluwer)

A comprehensive compilation of articles written by experts in their fields about all aspects of diffraction, with extensive lists of further references.

Zachariasen W H 1945 *Theory of X-ray Diffraction in Crystals* (New York: Dover)

The classic text on the diffraction of x-rays.

Bacon G E 1962 *Neutron Diffraction* 2nd ed (Oxford: Clarendon)

Likewise, the classic text on neutron diffraction.

Thomas G and Goringe M J 1981 *Transmission Electron Microscopy of Materials* (New York: Wiley)

A practical introduction to electron microscopy and diffraction.

-1-

B1.9 Scattering: light, neutrons, X-rays

Benjamin S Hsiao and Benjamin Chu

B1.9.1 INTRODUCTION

Scattering techniques using light, neutrons and x-rays are extremely useful to study the structure, size and shape of large molecules in solids, liquids and solutions. The principles of the scattering techniques, which involve the interaction of radiation with matter, are the same. However, the data treatment for scattering from light, neutrons or x-rays can be quite different because the intrinsic property of each radiation and its interactions with matter are different. One major difference in the data treatment arises from the states of matter. Although the general equations describing the interaction between radiation and matter are valid for all classes of materials, unique analytical treatments have been made to suit the different states of matter.

In this chapter, the general principles of the scattering phenomenon and specific data treatments for the material (isotropic and anisotropic) in both solid and solution states are presented. These treatments are useful for the analysis of scattering data by light, neutrons or x-rays from different material systems such as crystalline polymers, complex fluids (including colloidal suspensions and solutions of biological species), multicomponent systems (including microemulsions and nanocomposites) and oriented polymers. For detailed theoretical derivations, the reader should refer to the many excellent textbooks and review articles that deal with the subjects of scattering from light [[1](#), [2](#), [3](#), [4](#), [5](#), [6](#) and [7](#)], neutrons [[7](#), [8](#), [9](#), [10](#) and [11](#)] and x-rays [[7](#), [11](#)],

[12](#), [13](#), [14](#) and [15](#)]. We, however, will not discuss the detailed instrumentation for different scattering experiments as this topic has been well illustrated in some of the above references [[3](#), [4](#), [5](#), [6](#) and [7](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#) and [15](#)]. Also absent will be the analysis for thin films and interfaces which warrants a separate chapter by itself. The main focus of this chapter is to provide a comprehensive overview to the field of scattering from materials (with emphasis on polymers) including an appropriate comparison between the different techniques. Selected example studies will also be included at the end of each section to illustrate the applications of some advanced scattering techniques.

B1.9.2 INTERACTION OF RADIATION AND MATTER

In free space, electromagnetic radiation consists of simultaneous electric and magnetic fields, which vary periodically with position and time. These fields are perpendicular to each other and to the direction of wave propagation. The electromagnetic radiation consists of a wide range of wavelengths from 10^{-10} m (x-rays) to 10^4 m (low frequency radio waves). Visible light has wavelengths from 400 nm (violet) to 700 nm (red), which constitutes a very small fraction of the electromagnetic spectrum.

As the electromagnetic radiation interacts with matter, the resultant radiation may follow several pathways depending on the wavelength and the material characteristics. Scattering without loss of energy is termed elastic. Elastic scattering from the periodic structure of matter emitting radiation of wavelength with the same magnitude leads to diffraction phenomena. The radiation may be slowed down by the refraction phenomenon. The radiation may also be absorbed

-2-

by the material. The absorbed energy may be transferred to different modes of motion, dissipated as heat or re-emitted as radiation at a different frequency. Energetic photons from x-rays or ultra-violet (UV) radiation can produce dissociation of chemical bonds leading to chemical reactions ejecting photoelectrons (known as x-ray photoelectron spectroscopy, XPS, or electron spectroscopy for chemical analysis, ESCA). Fluorescence occurs when the transfer of the residual energy to electronic modes takes place. Raman scattering occurs when the energy is transferred to or from rotational or vibrational modes. The characterization techniques based on these different phenomena are described in other chapters of section B.

Light scattering arises from fluctuations in refractive index or polarizability and x-ray scattering arises from fluctuations in electron density. Both are dependent on interactions of radiation with extra-nuclear electrons ([figure B1.9.1](#)) and will be discussed together. If we consider an incident beam having an electric field $E = E_0 \cos(2\pi r/\lambda - \omega t)$, where E_0 is the magnitude, λ is the wavelength in vacuum, r is the distance of the observer from the scatterer and ω is the angular frequency. From electromagnetic radiation incident upon an atom (with polarizability, α), a dipole moment $m = \alpha E$ will be induced. The oscillating dipole will serve as a source of secondary radiation (this is scattering) with amplitude E_s [[16](#)],

$$E_s = \frac{1}{c^2 r} \frac{d^2 m}{dt^2} \cos \varphi \quad (\text{B1.9.1})$$

where c is the velocity of light and φ is the angle between the plane of the polarization and the dipole moment. Thus we obtain

$$E_s = \frac{-\alpha E_0 \omega^2}{c^2 r} \cos \varphi \cos(\omega t - \phi) \quad (\text{B1.9.2})$$

where ϕ is a phase angle which takes into account that the wave must travel a distance ($r = d$) to reach the observer ($\phi = 2\pi d/\lambda$). These equations presume that the electric field at the scattering position is not modified by the induced-dielectric environment (the Rayleigh–Gans approximation). Equation (B1.9.2) is thus termed Rayleigh scattering, which holds true for light scattering provided that the light frequency is small when compared with the resonance frequency of the electrons. For x-ray scattering, the frequency of electromagnetic radiation is higher than the resonance frequency of the electrons. In this case, Thomson scattering prevails and the scattering amplitude becomes

$$E_s = \frac{-e^2 \pi E_0}{m_0 c^2 r} \cos \varphi \cos(\omega t - \phi) \quad (\text{B1.9.3})$$

where e is the electron charge and m_0 is the electron mass. Thus, all electrons scatter x-rays equally and the x-ray scattering ability of an atom depends on the number of electrons, which is proportional to the atomic number, Z , in the atom. It should be noted that Rayleigh scattering is dependent on frequency and polarizability, but Thomson scattering is not. As a result, more polarizable molecules (larger, conjugated, more aromatic) are better Rayleigh scatterers than others. Neutron scattering depends upon nuclear properties being related to fluctuations in the neutron scattering cross

-3-

section σ between the scatterer and the surroundings. Hence, hydrogen can be a strong neutron scatterer in an isotope environment, but it is a weak electron scatterer.

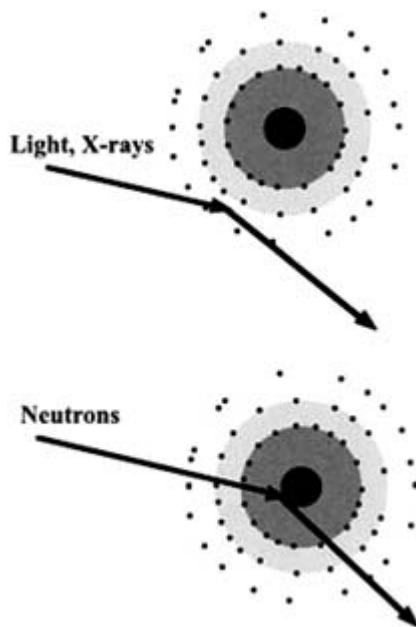


Figure B1.9.1. Diagrams showing that x-ray and light scattering involve extra-nuclear electrons, while neutron scattering depends on the nature of the atomic nucleus.

The generalized scattering equation can be expressed by a complex exponential form

$$(E_s)_j = E_0 K_j \exp[i(\omega t - \phi_j)] \quad (\text{B1.9.4})$$

where the subscript j refers to the scattering from the j th element, $(E_s)_j$ represents the amplitude of the scattering of the j th scatterer and K_j is proportional to the scattering power of the j th scatterer. For a collection of scattering elements, the total field strength (amplitude) of the scattered waves is

$$E_s = \sum_j (E_s)_j = E_0 \sum_j K_j \exp[i(\omega t - \phi_j)]. \quad (\text{B1.9.5})$$

All scattering phenomena (light, x-rays and neutrons) can be interpreted in terms of this equation (B1.9.5). These techniques differ mainly in the structural entities that contribute to the K_j term. For light, the refractive index or polarizability is the principal contributor; for x-rays, the electron density is the contributor; and for neutrons, the nature of the scattering nucleus is the contributor. Equation (B1.9.5) thus represents a starting point for the discussion of the interference problem presented below.

-4-

B1.9.3 LIGHT SCATTERING

The intensity of light scattering, I_s , for an isolated atom or molecule is proportional to the mean squared amplitude

$$I_s = K \langle E_s^2 \rangle \quad (\text{B1.9.6})$$

where the constant K is equal to $c/4\pi$ for electromagnetic radiation and $\langle \rangle$ represents an average operation. Combining equations (B1.9.2) and (B1.9.6), we have

$$I_s = K \frac{\alpha^2 E_0^2 \omega^4}{c^4 r^2} \langle \cos^2 \varphi \cos^2(\omega t - \phi) \rangle. \quad (\text{B1.9.7})$$

As the average is over all values of time,

$$\langle \cos^2(\omega t - \phi_j) \rangle = \langle \cos^2 x \rangle = \left(\int_0^{2\pi} \cos^2 x \, dx \right) \left(\int_0^{2\pi} dx \right)^{-1} = \frac{1}{2} \quad (\text{B1.9.8})$$

so equation (B1.9.7) can be simplified to

$$I_s = K \frac{\alpha^2 E_0^2 \omega^4}{2c^4 r^2} \cos^2 \varphi. \quad (\text{B1.9.9})$$

The incident intensity, I_0 , is given by

$$I_0 = K \langle E^2 \rangle = K E_0^2 \langle \cos^2(\omega t - \phi) \rangle = \frac{1}{2} K E_0^2. \quad (\text{B1.9.10})$$

The ratio of the scattered to the incident intensity is given by

$$\frac{I_s}{I_0} = \frac{\alpha^2 \omega^4}{c^4 r^2} \cos^2 \varphi. \quad (\text{B1.9.11})$$

-5-

Equation (B1.9.11) is valid only for plane polarized light. For unpolarized incident light, the beam can be resolved into two polarized components at right angles to each other. The scattered intensity can thus be expressed as (figure B1.9.2)

$$I_s = I_{s1} + I_{s2}. \quad (\text{B1.9.12})$$

As a result, equation (B1.9.11) becomes

$$\frac{I_s}{I_0} = \frac{\alpha^2 \omega^4}{2c^4 r^2} (\cos^2 \varphi_1 + \cos^2 \varphi_2) = \frac{\alpha^2 \omega^4}{2c^4 r^2} (1 + \cos^2 2\theta) \quad (\text{B1.9.13})$$

where we let $\varphi_1 = 0$ and $2\theta = \varphi_2$. Note that in light scattering, one often defines $\theta = \varphi_2$ with θ being the scattering angle. Herein, we define 2θ as the scattering angle to be consistent with x-ray scattering described in B1.9.4. Since the frequency term may be converted to wavelength (in vacuum)

$$\frac{\omega^4}{c^4} = \frac{16\pi^4 \nu^4}{c^4} = \frac{16\pi^4}{\lambda^4} \quad (\text{B1.9.14})$$

where ν is the frequency, equation (B1.9.13) becomes

$$\frac{I_s}{I_0} = \frac{8\pi^4 \alpha^2}{\lambda^4 r^2} (1 + \cos^2 2\theta). \quad (\text{B1.9.15})$$

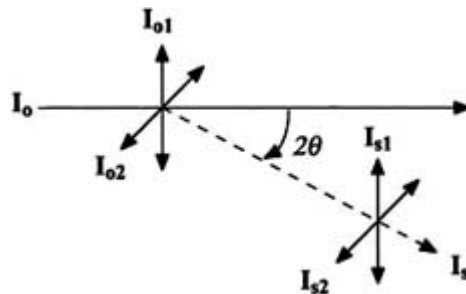


Figure B1.9.2. Resolution of a plane unpolarized incident beam into polarized scattering components.

-6-

This equation has the following implications.

- (1) At $2\theta = 0^\circ$, the scattering comprises both components of polarization of the incident beam; at $2\theta = 90^\circ$, the scattering comprises only one half of the incident beam. Consequently, the scattered light at 90° will be plane polarized.
- (2) Since $I_s \propto 1/\lambda^4$, this indicates that the shorter wavelengths (such as blue) scatter more than longer ones (such as red). Therefore, the light from a clear sky being blue is due to scattering by gas molecules in the atmosphere.

B1.9.3.1 SCATTERING FROM A COLLECTION OF OBJECTS

For random locations of N scattering objects in volume V , the scattered intensity can be found by summing the scattering from each object:

$$\frac{I_s}{I_0} = \frac{8\pi^4 N \alpha^2}{\lambda^4 r^2} (1 + \cos^2 2\theta). \quad (\text{B1.9.16})$$

A modified form of equation (B1.9.16) is usually used to express the scattering power of a system in terms of the 'Rayleigh ratio' defined as

$$R = \frac{(I_s/I_0)(r^2)}{V(1 + \cos^2 2\theta)} = \frac{8\pi^4 \alpha^2}{\lambda^4} \left(\frac{N}{V} \right). \quad (\text{B1.9.17})$$

In this case, the scattering serves as a means for counting the number of molecules (or particles, or objects) per unit volume (N/V). It is seen that the polarizability, α , will be greater for larger molecules, which will scatter more. If we take the Clausius–Mosotti equation [16]:

$$(n^2 - 1)/(n^2 + 2) = \frac{4}{3}\pi \left(\frac{N}{V} \right) \alpha \quad (\text{B1.9.18})$$

and consider $n \approx 1$ (n is the refractive index), then

$$\alpha = \frac{n - 1}{2\pi} \left(\frac{V}{N} \right). \quad (\text{B1.9.19})$$

If the scattering particles are in a dielectric solvent medium with solvent refractive index n_0 , we can define the excess

polarizability ($\alpha_{\text{ex}} = \alpha(\text{solution}) - \alpha(\text{solvent})$) as

$$(\text{B1.9.20})$$

$$\alpha_{\text{ex}} = \frac{n^2 - n_0^2}{4\pi} \left(\frac{V}{N} \right).$$

If the weight concentration C is used, $C = MN/(N_a V)$, where M is the molecular weight of the particle, N_a is Avogadro's number and $n \approx n_0$, then the above relationship becomes

$$\alpha_{\text{ex}} = \frac{(n + n_0)(n - n_0)}{4\pi C} \frac{CV}{N} = \frac{n_0}{2\pi} \left(\frac{\partial n}{\partial C} \right) \left(\frac{M}{N_a} \right) \quad (\text{B1.9.21})$$

where $\partial n/\partial C \approx (n - n_0)/C$ at constant temperature. The excess Rayleigh ratio $R_{\text{ex}} = R(\text{solution}) - R(\text{solvent})$ has the form

$$R_{\text{ex}} = \frac{2\pi^2 n_0^2}{\lambda^4 N_a} \left(\frac{\partial n}{\partial C} \right)^2 CM = HCM \quad (\text{B1.9.22})$$

where H is the optical constant for unpolarized incident light. For polarized light, the constant H can be twice as large due to the factor $(1 + \cos^2 2\theta)$. It is noted that this factor of 2 depends on the definition of R , i.e., whether the $1 + \cos^2 2\theta$ term is absorbed by R .

$$R_{\text{ex,p}} = \frac{4\pi^2 n_0^2}{\lambda^4 N_a} \left(\frac{\partial n}{\partial C} \right)^2 CM = HCM \quad (\text{B1.9.23})$$

where $R_{\text{ex,p}}$ is the excess Rayleigh ratio for polarized light. The above treatment is valid for molecules that are small compared to the wavelength of the incident beam.

B1.9.3.2 SCATTERING FROM A SOLUTION OF LARGE MOLECULES

For molecules having dimensions comparable with the wavelength, phase differences will occur between waves scattered from different regions of the molecule. These phase differences result in an angular dependence of the scattered intensity. The reduction may be expressed in terms of a particle interference factor $P(2\theta)$ such that

$$P(2\theta) = \frac{R_{\text{ex}}(\text{experimental})}{R_{\text{ex}}(\text{no interference})} \quad (\text{B1.9.24})$$

-8-

where 2θ is the scattering angle. We again caution the reader that the conventional symbol for the scattering angle by light is θ . Herein we use the symbol 2θ to be consistent with x-ray and neutron scattering described later. The interference factor also follows the expression

$$P(2\theta) = \left\langle \left[\sum_j \alpha_j \cos(\omega t - \phi_j) \right]^2 \right\rangle \left(\left[\sum_j \alpha_j \cos(\omega t) \right]^2 \right)^{-1} \quad (\text{B1.9.25})$$

where the summation is over all parts of a molecule. The nature of $P(2\theta)$ may be qualitatively deduced from figure B1.9.3. For scattering in the forward direction (θ_1) the path difference between the rays from elements A and B of the molecule ($d_B - d_A$) is less than that at the backward scattering angle ($2\theta_2$) to observer O_2 , ($d'_B - d'_A$). So, a greater phase difference occurs at $2\theta_2$. If the dimensions of the molecule (or particle) are less than the wavelength, destructive interference occurs and $P(2\theta)$ will decrease. If the molecular (or particle) dimensions are much greater than the probing wavelength, both destructive and constructive interference can occur leading to maxima and minima in $P(2\theta)$. For $2\theta = 0^\circ$, no path difference exists and $P(2\theta) = 1$. Thus, the scattering technique can be used to estimate the size of the molecule (or particle).

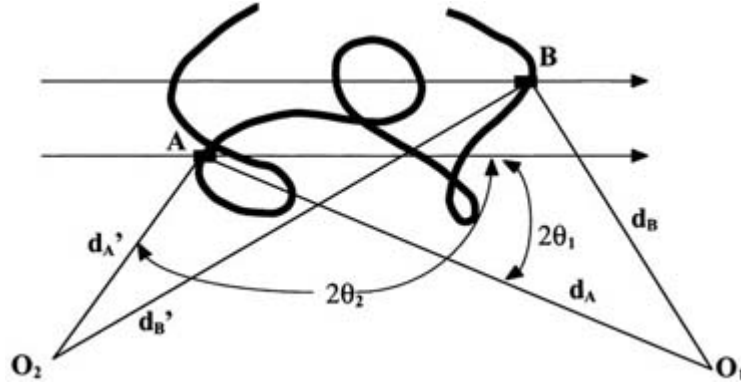


Figure B1.9.3. Variation of $P(2\theta)$ as a function of scattering angle.

Equation (B1.9.5) gives the total amplitudes of scattering from a collection of objects and is a good starting point for the derivation of interference phenomena associated with molecular size.

$$E_s = E_0 \sum_j K_j \exp[i\omega(t - d_j/c)] \quad (\text{B1.9.26})$$

where d_j is the distance between the scattering element and the observation point P (shown in [figure B1.9.4](#) . In [figure B1.9.4](#) , the following relationships can be approximated as

$$d_j = \mathbf{r}_j \cdot \mathbf{s}_0 + D - \mathbf{r}_j \cdot \mathbf{s}_1 = D + \mathbf{r}_j \cdot (\mathbf{s}_0 - \mathbf{s}_1) = D + (\mathbf{r}_j \cdot \mathbf{s}) \quad (\text{B1.9.27})$$

-9-

where D represents the distance between the observation point P and the origin O , \mathbf{s}_0 is the unit vector in the incident beam direction, \mathbf{s}_1 is the unit vector in the scattered beam direction and \mathbf{r}_j is the vector to the j th scattering element. Equation (B1.9.26) thus becomes

$$E_s = E_0 \sum_j K_j e^{i\omega t} e^{-i\omega(\mathbf{r}_j \cdot \mathbf{s})/c} e^{-i\omega D/c} = F e^{-ikD} \quad (\text{B1.9.28})$$

where $F = \sum E_0 K_j e^{i\omega t} e^{-ik(\mathbf{r}_j \cdot \mathbf{s})}$, which is defined as the structure or form factor of the object, and $k = \omega/c = 2\pi/\lambda$. For a system with a large number of scattering elements, the summation in equation (B1.9.28) may be replaced by an integral

$$\begin{aligned}
F &= E_0 e^{i\omega t} \sum_j K_j \exp[-ik(\mathbf{r}_j \cdot \mathbf{s})] \\
&= E_0 e^{i\omega t} \int_r \rho(r) \exp[-ik(\mathbf{r}_j \cdot \mathbf{s})] d^3r.
\end{aligned}
\tag{B1.9.29}$$

The term $E_0 e^{i\omega t}$ is related to the incident beam, which is often omitted in the theoretical derivation (as follows). The second term in (B1.9.29) represents the amplitude scattered by a three dimensional element with a volume element d^3r and $\rho(r)$ being the density profile. From now on, we will simplify the symbol for the dot product of two vectors $\mathbf{r}_j \cdot \mathbf{s}$ as $r_j s$ and other similar products. If we consider the spherical polar coordinates (figure B1.9.5), this volume element becomes

$$d^3r = \int_{\phi=0}^{2\pi} \int_{\varphi=0}^{\pi} \int_{r=0}^{\infty} r^2 \sin \varphi \, dr \, d\varphi \, d\phi \tag{B1.9.30}$$

so that more generally,

$$F = \int_{\phi=0}^{2\pi} \int_{\varphi=0}^{\pi} \int_{r=0}^{\infty} \rho(r, \varphi, \phi) e^{-ik(r_j s)} r^2 \sin \varphi \, d\phi \, d\varphi \, dr. \tag{B1.9.31}$$

For spherically symmetric systems, we can derive the following expression from (B1.9.31) [17]:

$$F = 4\pi \int_{r=0}^{\infty} \rho(r) \frac{\sin qr}{qr} r^2 \, dr \tag{B1.9.32}$$

-10-

where $q = 4\pi(\sin \theta)/\lambda$, with 2θ being the scattering angle. The interference factor described previously (equation (B1.9.25)) may be expressed in terms of the intramolecular particle scattering factor as

$$P(2\theta) = \left(\int_{r=0}^{\infty} \rho(r) \frac{\sin qr}{qr} r^2 \, dr \right) \left(\int_{r=0}^{\infty} \rho(r) r^2 \, dr \right)^{-1}. \tag{B1.9.33}$$

If we define the radius of gyration, R_g , by

$$R_g^2 = \left(\int_0^{\infty} \rho(r) r^2 \, dr \right) \left(\int_0^{\infty} \rho(r) \, dr \right)^{-1} \tag{B1.9.34}$$

then equation (B1.9.33) can be expressed as [18]

$$P(2\theta) = 1 - \frac{q^2 R_g^2}{3} + \dots = 1 - \frac{16\pi^2 R_g^2}{3\lambda^2} \sin^2(2\theta/2) + \dots. \tag{B1.9.35}$$

To measure the molecular weight of the molecule, we can modify equation (B1.9.23) to take into account the intramolecular interference in the dilute solution range,

$$\frac{HC}{R_{\text{ex}}} \cong \frac{1}{MP(2\theta)} + 2A_2C \quad (\text{B1.9.36})$$

where A_2 is the second virial coefficient. By combining (B1.9.35) with (B1.9.36), we obtain

$$\frac{HC}{R_{\text{ex}}} \cong \frac{1}{M} \left(1 + \frac{q^2 R_g^2}{3} \right) + 2A_2C. \quad (\text{B1.9.37})$$

This is the basic equation for monodisperse particles in light scattering experiments. We can derive three relationships by extrapolation.

-11-

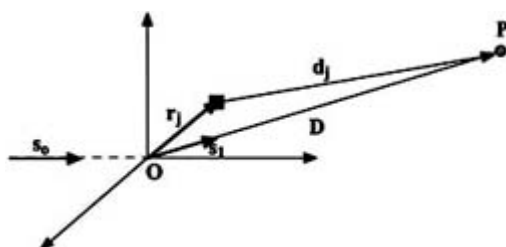


Figure B1.9.4. Geometrical relations between vectors associated with incident and scattered light.

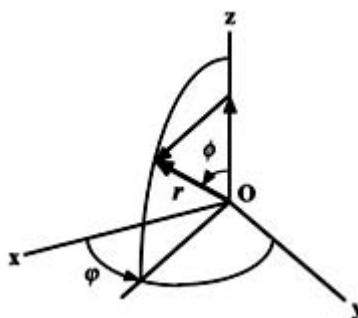


Figure B1.9.5. Geometrical relations between the Cartesian coordinates in real space, the spherical polar coordinates and the cylindrical polar coordinates.

$$\lim_{C \rightarrow 0} \frac{HC}{R_{\text{ex}}} \cong \frac{1}{M} \left(1 + \frac{q^2 R_g^2}{3} \right) \quad (\text{B1.9.38})$$

$$\lim_{q \rightarrow 0} \frac{HC}{R_{\text{ex}}} \cong \frac{1}{M} + 2A_2C \quad (\text{B1.9.39})$$

$$\lim_{q \rightarrow 0, C \rightarrow 0} \frac{HC}{R_{\text{ex}}} \cong \frac{1}{M}. \quad (\text{B1.9.40})$$

A graphical method, proposed by Zimm (thus termed the Zimm plot), can be used to perform this double extrapolation to determine the molecular weight, the radius of gyration and the second virial coefficient. An example of a Zimm plot is shown in [figure B1.9.6](#) where the light scattering data from a solution of poly

(tetrafluoroethylene) (PTFE) ($M_w = (2.9 \pm 0.2) \times 10^5 \text{ g mol}^{-1}$; $A_2 = -(6.7 \pm 1.3) \times 10^{-5} \text{ mol cm}^3 \text{ g}^{-2}$ and $R_g = 17.8 \pm 2.4 \text{ nm}$) in oligomers of poly(chlorotrifluoroethylene) (as solvents) at 340°C is shown [19]. The dashed lines represent the extrapolated values at $C = 0$ and $2\theta = 0$.

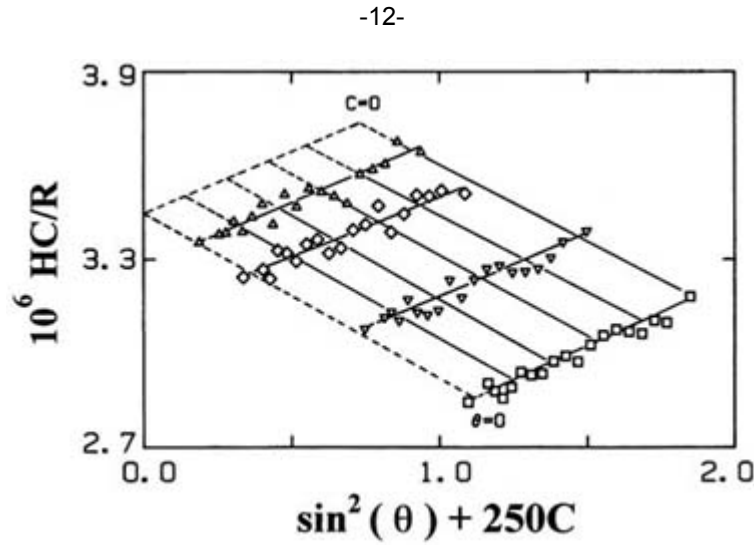


Figure B1.9.6. A typical Zimm plot; data obtained from a solution of poly(tetrafluoroethylene) (PTFE) ($M_w = (2.9 \pm 0.2) \times 10^5 \text{ g mol}^{-1}$; $A_2 = -(6.7 \pm 1.3) \times 10^{-5} \text{ mol cm}^3 \text{ g}^{-2}$ and $R_g = 17.8 \pm 2.4 \text{ nm}$) in oligomers of poly(chlorotrifluoroethylene) (as solvents) at 340°C . (Reprinted with permission from Chu *et al* [19].)

B1.9.3.3 CALCULATE SCATTERED INTENSITY

There are two ways to calculate the scattered intensity. One is to first calculate the magnitude of the structure factor F by summing the amplitude of scattering elements in the system and then multiply it by F^* (the conjugate of F). This method is best for the calculation of the scattered intensity from discrete particles such as spheres or rods. Two examples are illustrated as follows.

- (1) The scattering from an *isolated sphere* may be calculated from equation (B1.9.32). This derivation assumes that the sphere is uniform, with its density profile $\rho(r) = \rho_0$ if $r < r_0$ and $\rho(r) = 0$ if $r > r_0$ (surrounded by a non-scattering material). With this assumption, equation (B1.9.32) becomes

$$F_{\text{sp}} = 4\pi\rho_0 \int_{r=0}^{r_0} \frac{\sin qr}{qr} r^2 dr. \quad (\text{B1.9.41})$$

By changing the variable, $x = qr$, we obtain

$$F_{\text{sp}} = \frac{4\pi\rho_0}{q^3} \int_{x=0}^{x=U} x \sin(x) dx. \quad (\text{B1.9.42})$$

We can integrate the above equation by parts and derive

$$F_{\text{sp}} = \frac{4\pi\rho_0}{q^3}[\sin U - U \cos U] = V_{\text{sp}}\rho_0\Phi(U) \quad (\text{B1.9.43})$$

where V_{sp} is the sphere volume and

$$\Phi(U) = \text{the sphere scattering function} = \frac{3}{U^3}[\sin U - U \cos U] \quad (\text{B1.9.44})$$

with the parameter $U = qr_0$. The intensity then is proportional to F^2 ($R = K_1 FF^*$, where K_1 is a calibration constant for light scattering).

(2)

This treatment may be extended to *spheres core-shell structure*. If the core density is ρ_1 to r_1 , the shell density is ρ_2 in the range r_1 to r_2 , the density of the surrounding medium is ρ_0 , then the structure factor becomes

$$F_{\text{cs}} = V_1(\rho_1 - \rho_0)\Phi(U_1) + V_2(\rho_2 - \rho_0)\Phi(U_2)$$

The second method to calculate the scattered intensity (or R : the Rayleigh ratio) is to square the sum in I

$$R(q) = K_1 \sum_i \sum_j \rho_i \rho_j \exp[i(qr_{ij})].$$

For a continuous system with assorted scattering elements, the sum can be replaced by integration and be expressed as

$$R(q) = K_1 V \langle \eta^2 \rangle \int \gamma(r) \exp[i(qr)] d^3r$$

where V is the scattering volume, η is the heterogeneity of the fluctuation parameter defined as $\eta = \rho - \langle \rho \rangle$ and the *correlation function* defined as

$$\gamma(r_{ij}) = \langle \eta_i \eta_j \rangle.$$

Equation (B1.9.47) applies to the general scattering expression of any system. With spherical symmetry, the scattered intensity becomes [20]

$$R(q) = 4\pi K_1 V \langle \eta^2 \rangle \int_0^\infty \gamma(r) r^2 \frac{\sin(qr)}{qr} dr. \quad (\text{B1.9.49})$$

In the case of anisotropic systems with a cylindrical symmetry (such as rods or fibres), the scattered intensity can be expressed as (derivation to be made later in [section B1.9.4](#)):

$$R(q) = 4\pi K_1 V \langle \eta^2 \rangle \int_0^\infty \gamma(r) J_0(q_r r) r dr \int_0^\infty \gamma(z) \cos(q_z z) dz \quad (\text{B1.9.50})$$

where the subscript r represents the operation along the radial direction, the subscript z represents the operation along the cylinder axis, J_0 is the zero-order Bessel function and q_r and q_z are scattering vectors along the r (equator) and z (meridian) directions, respectively.

B1.9.3.4 ESTIMATE OBJECT SIZE

One of the most important functions in the application of light scattering is the ability to estimate the object dimensions. As we have discussed earlier for dilute solutions containing large molecules, [equation \(B1.9.38\)](#) can be used to calculate the ‘radius of gyration’, R_g , which is defined as the mean square distance from the centre of gravity [12]. The combined use of [equation \(B1.9.38\)](#), [equation \(B1.9.39\)](#) and [equation \(B1.9.40\)](#) (the Zimm plot) will yield information on R_g , A_2 and molecular weight.

The above approximation, however, is valid only for dilute solutions and with assemblies of molecules of similar structure. In the event that concentration is high where intermolecular interactions are very strong, or the system contains a less defined morphology, a different data analysis approach must be taken. One such approach was derived by Debye *et al* [21]. They have shown that for a random two-phase system with sharp boundaries, the correlation function may carry an exponential form.

$$\gamma(r) = e^{-r/a_c} \quad (\text{B1.9.51})$$

where a_c is a correlation length describing the dimension of heterogeneities. Substitution of (B1.9.51) into [\(B1.9.47\)](#) gives rise to the expression

-15-

$$R(q) = \frac{8\pi K_1 \langle \eta^2 \rangle a_c^3}{[1 + q^2 a_c^2]^2}. \quad (\text{B1.9.52})$$

Based on this equation, one can make a ‘Debye–Bueche’ plot by plotting $[R(q)]^{-1/2}$ versus q^2 and determine the slope and the intercept of the curve. The correlation length thus can be calculated as [21]

$$a_c = \left(\frac{\text{slope}}{\text{intercept}} \right)^{1/2}. \quad (\text{B1.9.53})$$

B1.9.4 X-RAY SCATTERING

X-ray scattering arises from fluctuations in electron density. The general expression of the absolute scattered intensity $I_{\text{abs}}(q)$ (simplified as $I(q)$ from now on) from the three-dimensional objects immersed in a different

density medium, similar to (B1.9.47), can be expressed as:

$$I(q) = K_x V \langle \eta^2 \rangle \int \gamma(r) e^{i(qr)} d^3r \quad (\text{B1.9.54})$$

where V is the volume of the scatterer, $\langle \eta^2 \rangle$ is the square of the electron density fluctuations, $\gamma(r)$ is the correlation function and K_x is a calibration constant depending on the incident beam intensity and the optical apparatus geometry (e.g. polarization factor) given by

$$K_x = I_0 \left(\frac{e^2}{m_0 c^2} \right)^2 \frac{1}{D_s^2} \frac{1 + \cos^2 2\theta}{2} \quad (\text{B1.9.55})$$

where e , m_0 are the charge and mass of an electron, c is the velocity of light, D_s is the sample to detector distance and 2θ is the scattering angle.

If the scattering system is isotropic, equation (B1.9.54) can be expressed in spherical polar coordinates (the derivation is similar to equation (B1.9.32)):

$$I(q) = 4\pi K_x V \langle \eta^2 \rangle \int_0^\infty \gamma(r) r^2 \frac{\sin(qr)}{qr} dr. \quad (\text{B1.9.56})$$

-16-

This expression is very similar to (B1.9.49). If the scattering system is anisotropic, equation (B1.9.54) can then be expressed in cylindrical polar coordinates (see figure B1.9.5 :

$$I(q) = I(q_r, \psi, q_z) = K_x V_s \langle \eta^2 \rangle \iiint \gamma(r, \varphi, z) e^{i(rq_r \cos(\varphi - \psi) + zq_z)} r dr d\varphi dz \quad (\text{B1.9.57})$$

where r represents the distance along the radial direction and z represents the distance along the cylindrical axis (in real space), q_r and q_z are correspondent scattering momenta along the r and z directions in reciprocal space, φ is the polar angle in real space and ψ is the polar angle in reciprocal space. Equation (B1.9.57) is a general expression for cylinders without any assumptions. If we consider the scatterers having the geometry of a cylinder, it is reasonable to assume $\gamma(r, \varphi, z) = \gamma(r, \varphi) \gamma(z)$, which indicates that the two correlation functions along the radial (equatorial) direction and the cylindrical (meridional) direction are independent. In addition, the term $\gamma(\varphi) = 1$ can be applied, which represents the symmetry of a cylinder. Equation (B1.9.57) thus can be rewritten as

$$I(q) = K_x V \langle \eta^2 \rangle \int_0^\infty \gamma(r) \left(\int_0^{2\pi} e^{irq_r \cos(\varphi - \psi)} d\varphi \right) r dr \int_{-\infty}^\infty \gamma(z) e^{izq_z} dz. \quad (\text{B1.9.58})$$

If we define $\delta = \varphi - \psi$ and $rq_r = u$, then

$$(\text{B1.9.59})$$

$$\int_0^{2\pi} e^{irq_r \cos(\varphi-\psi)} d\varphi = \int_0^{2\pi} e^{iu \cos \delta} d\delta = 2\pi J_0(u)$$

where $J_0(u)$ is the zeroth order Bessel function of the first kind. Also, we have

$$\begin{aligned} \int_{-\infty}^{\infty} \gamma(z) e^{izq_z} dz &= \int_{-\infty}^{\infty} \gamma(z) (\cos(zq_z) + i \sin(zq_z)) dz \\ &= \int_{-\infty}^{\infty} \gamma(z) \cos(zq_z) dz \end{aligned} \quad (\text{B1.9.60})$$

because both $\gamma(z)$ and $\cos(zq_z)$ are even functions (i.e., $\gamma(z) = \gamma(-z)$ and $\cos(zq_z) = \cos(-zq_z)$) and $\sin(zq_z)$ is an odd function. The integral of the latter ($\gamma(z)\sin(zq_z)$, an odd function) is an even function integrated from $-\infty$ to ∞ , which becomes zero. Combining equations (B1.9.59) and (B1.9.60) into equation (B1.9.58), we obtain

$$I(q) = K_x V \langle \eta^2 \rangle \int_0^{\infty} \gamma(r) 2\pi J_0(q_r r) r dr \int_0^{\infty} 2\gamma(z) \cos(q_z z) dz. \quad (\text{B1.9.61})$$

-17-

This equation is the same as (B1.9.50) for light scattering.

B1.9.4.1 PARTICLE SCATTERING

The general expression for particle scattering can best be described by the correlation function $\gamma(r)$. Using the definition in (B1.9.48), we have

$$\gamma(r) = (\Delta\rho)^2 \gamma_0(r) \quad (\text{B1.9.62})$$

where $\Delta\rho$ is the electron density difference between the particle and the surrounding medium and is assumed to be constant, $\gamma(0) = 1$ and $\gamma(r) = 1$ for $r \geq D$ (the diameter of the particle). The function $\gamma_0(r)$ can be expressed by a distribution function $G(l)$ with l being the *intersection length* or the *chord length* of the particle (figure B1.9.7) [12, 13]

$$\gamma_0(r) = \frac{1}{\bar{l}} \int_r^D (l-r) G(l) dl \quad (\text{B1.9.63})$$

with

$$\bar{l} = \int_0^D l G(l) dl. \quad (\text{B1.9.64})$$

By differentiation, we can derive

$$\frac{d\gamma_0(r)}{dr} = -\frac{1}{\bar{l}} \int_r^D G(l) dl \quad (\text{B1.9.65})$$

$$\frac{d^2\gamma_0(r)}{dr^2} = \frac{1}{\bar{l}} G(l).$$

The distance distribution function $p(r)$ has a clear geometrical definition. It is defined as

$$p(r) = \gamma(r)r^2. \quad (\text{B1.9.66})$$

For homogeneous particles, it represents the number of distances within the particle. For inhomogeneous particles, it has to take into account the different electron density of the volume elements. Thus it represents the number of pairs of difference in electrons separated by the distance r . A qualitative description of shape and internal structure of the

-18-

particle can be obtained directly from $p(r)$. In addition, several structural parameters can be determined quantitatively [22]. We can describe several analytical forms of the distance distribution function for different shapes of homogeneous particles as follows.

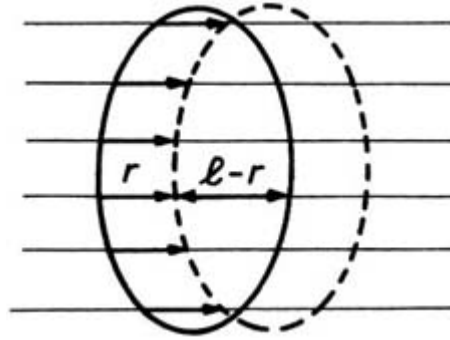


Figure B1.9.7. Diagram to illustrate the relationship between chord length (ℓ) and r in a particle.

- (1) Globular particles

$$p(r) = 12x^2(2 - 3x + x^3) \quad (\text{B1.9.67})$$

where $x = r/D$.

- (2) Rodlike particles

$$p(r) = \frac{1}{2\pi} \rho_c^2 A^2 (L - r) \quad (\text{B1.9.68})$$

where ρ_c is the particle electron density, A is the cross-section of the rod and L is the length of the rod particle.

- (3) Flat particles, i.e. particles elongated in two dimensions (discs, flat prisms)

$$p(r) = \frac{16}{\pi} x(\arccos(x) - x\sqrt{1-x^2}) \quad (\text{B1.9.69})$$

where $x=r/D$.

-19-

The radius of gyration of the whole particle, R_g , can be obtained from the distance distribution function $p(r)$ as

$$R_g^2 = \left(\int_0^\infty p(r)r^2 dr \right) \left(\int_0^\infty p(r) dr \right)^{-1}. \quad (\text{B1.9.70})$$

This value can also be obtained from the innermost part of the scattering curve.

$$I(q) = I(0) \exp\left(-\frac{R_g^2 q^2}{3}\right) \quad (\text{B1.9.71})$$

where $I(0)$ is the scattered intensity at zero scattering angle. A plot of $\log I(q)$ versus q^2 , which is known as the Guinier plot, should show a linear descent with a negative slope ($= -R_g^2/3$) related to the radius of gyration.

In practical data analysis, we are interested in extracting information about the size, shape and distribution of the scattering particles. The most widely used approach for this purpose is the indirect Fourier transformation method, pioneered by O Glatter [23, 24, and 25]. This approach can be briefly illustrated as follows. In dilute solutions with spherical particles, where the interparticle scattering is negligible, the distance distribution function $p(r)$ can be directly calculated from the scattering data through Fourier transformation (combining equations (B1.9.56) and (B1.9.66))

$$I(q) = 4\pi K_x V \langle \eta^2 \rangle \int_0^\infty p(r) \frac{\sin(qr)}{qr} dr. \quad (\text{B1.9.72})$$

This process is shown in figure [figure B1.9.8](#) where T_1 represents the Fourier transformation. If there are additional instrumentation effects desmearing the data (such as the slit geometry, wavelength distribution etc), appropriate inverse mathematical transformations (T_2, T_3, T_4) can be used to calculate $p(r)$. If the particle has a certain shape, the $\sin(x)/x$ term in (B1.9.72) must be replaced by the form factor according to the shape assumed. If concentrations increase, the interparticle scattering (the so-called structure factor) should be considered, which will be discussed later in [section B1.9.5](#). A unique example to illustrate the usefulness of the distance distribution function is the study of a DNA-dependent RNA polymerase core enzyme [26, 27] ([figure B1.9.9](#)). The RNA polymerase enzyme is known to have four subunits but in two possible configurations (model 1 and model 2). Model 1 has a configuration with the two larger subunits having a centre-to-centre distance of 5 nm, and model 2 has a more open configuration having a centre-to-centre distance of 7 nm. From the experimental data (open circle), it is clear that model 1 gives a better fit to the data.

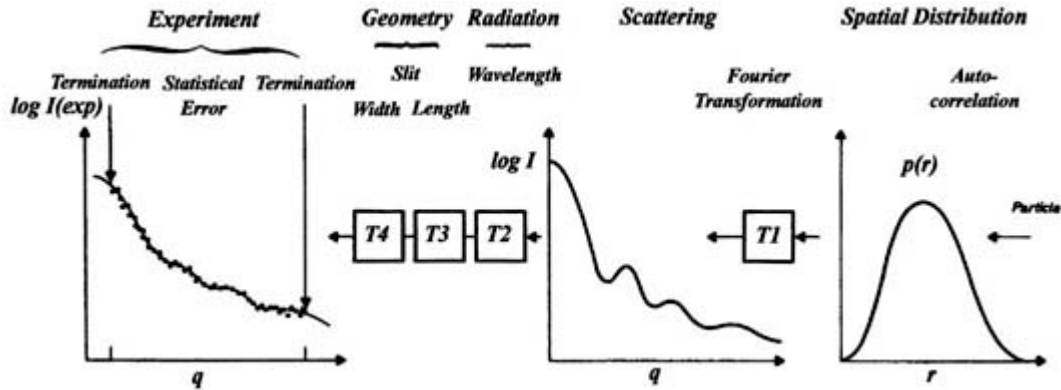


Figure B1.9.8. Schematic diagram of the relationship between a particle distribution and the measured experimental scattering data. This figure is duplicated from [14], with permission from Academic Press.

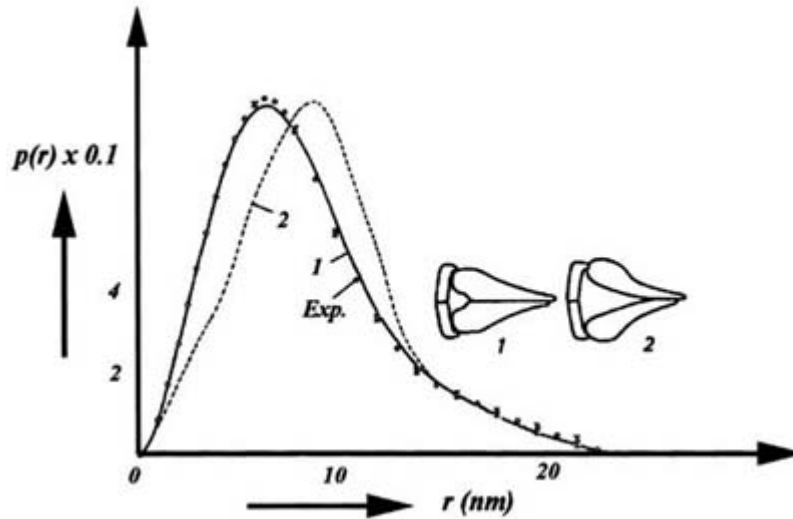


Figure B1.9.9. Comparison of the distance distribution function $p(r)$ of a RNA-polymerase core enzyme from the experimental data (open circle) and the simulation data (using two different models). This figure is duplicated from [27], with permission from Elsevier Science.

B1.9.4.2 NONPARTICULATE SCATTERING

If we consider the scattering from a general two-phase system (figure B1.9.10) distinguished by indices 1 and 2) containing constant electron density in each phase, we can define an average electron density $\bar{\rho}$ and a mean square density fluctuation as:

$$\bar{\rho} = \varphi_1 \rho_1 + \varphi_2 \rho_2 \tag{B1.9.73}$$

$$\overline{\eta^2} = (\rho_1 - \rho_2)^2 \varphi_1 \varphi_2 \tag{B1.9.74}$$

where φ and ρ are the fractions of the total volume V . Equation (B1.9.74) is directly related to the invariant Q of the system

$$Q = 2\pi^2 V \overline{\eta^2} = 2\pi^2 V (\rho_1 - \rho_2)^2 \varphi_1 \varphi_2. \quad (\text{B1.9.75})$$

In this case, the correlation function $\gamma(r)$ becomes

$$\gamma(r) = (\rho_1 - \rho_2)^2 \varphi_1 \varphi_2 \gamma_0(r) \quad (\text{B1.9.76})$$

where $\gamma_0(r)$ is the normalized correlation function, which is related to the geometry of the particle.

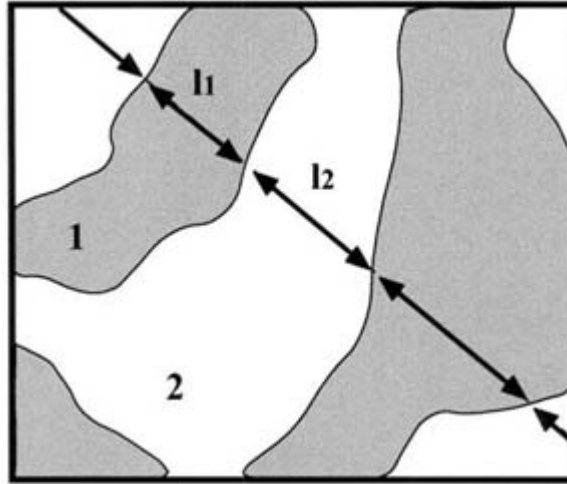


Figure B1.9.10. Non-particulate random two-phase system.

There are several geometrical variables that one can extract from the correlation function approach. First, a correlation

volume v_c can be defined, which is related to the extrapolated intensity at zero angle $I(0)$.

$$I(0) = V (\rho_1 - \rho_2)^2 \varphi_1 \varphi_2 v_c. \quad (\text{B1.9.77})$$

With the use of invariant Q , we obtain

$$v_c = \frac{2\pi^2}{Q} I(0). \quad (\text{B1.9.78})$$

If we consider the chord length distribution, we can express the alternating chords l_1 and l_2 as:

$$l_1 = 4 \frac{V}{S} \varphi_1 \quad l_2 = 4 \frac{V}{S} \varphi_2 \quad (\text{B1.9.79})$$

where S is the total surface area. The mean chord length thus becomes

$$\bar{l} = l_1\varphi_2 = l_2\varphi_1 = 4\frac{V}{S}\varphi_1\varphi_2. \quad (\text{B1.9.80})$$

As we will discuss in the next section, the scattered intensity $I(q)$ at very large q values will be proportional to the q^{-4} term. This is the well known Porod approximation, which has the relationship

$$\lim_{q \rightarrow \infty} I(q) = V\varphi_1\varphi_2(\rho_1 - \rho_2)^2 \frac{8\pi}{\bar{l}} \frac{1}{q^4}. \quad (\text{B1.9.81})$$

It is sometimes more convenient to normalize the absolute scattered intensity $I(q)$ and use the following expression:

$$\lim_{q \rightarrow \infty} I(q)q^4/Q = \frac{1}{\pi\varphi_1\varphi_2} \frac{S}{V}. \quad (\text{B1.9.82})$$

Thus the ratio of S/V can be determined by the limiting value of $I(q)q^4$. We will discuss the Porod approximation next.

B1.9.4.3 POROD APPROXIMATION

According to the Porod law [28], the intensity in the tail of a scattering curve from an *isotropic* two-phase structure having sharp phase boundaries can be given by [equation \(B1.9.81\)](#). In fact, this equation can also be derived from the general expression of scattering ([B1.9.56](#)). The derivation is as follows. If we assume $qr = u$ and use the Taylor expansion at large q , we can rewrite ([B1.9.56](#)) as

$$\begin{aligned} I(q) &= \frac{4\pi K_x V \langle \eta^2 \rangle}{q^3} \int_0^\infty \gamma(r) u \sin(u) du \\ &= \frac{4\pi K_x V \langle \eta^2 \rangle}{q^3} \left(\int_0^\infty \gamma(0) u \sin(u) du + \int_0^\infty \frac{\gamma'(0)}{q} u^2 \sin(u) du \right. \\ &\quad \left. + \int_0^\infty \frac{\gamma''(0)}{2q^2} u^3 \sin(u) du + \dots \right) \end{aligned} \quad (\text{B1.9.83})$$

where γ' and γ'' are the first and second derivatives of γ . The following expressions can be derived to simplify the above equation.

$$\int_0^\infty x \sin(x) dx = 0, \int_0^\infty x^2 \sin(x) dx = -2, \int_0^\infty x^3 \sin(x) dx = 0. \quad (\text{B1.9.84})$$

Then

$$I(q) = -\frac{8\pi K_x V \langle \eta^2 \rangle}{q^4} \gamma'(0) + O(q^{-2n}, d^{2n-3} \gamma(0) / dr^{2n-3}) \quad (n = 3, 4, 5 \dots). \quad (\text{B1.9.85})$$

The second term in equation (B1.9.85) rapidly approaches zero in the large q region, thus

$$\lim_{q \rightarrow \infty} I(q) = -\frac{8\pi K_x V \langle \eta^2 \rangle}{q^4} \gamma'(0). \quad (\text{B1.9.86})$$

This equation is the same as [equation \(B1.9.81\)](#), which is termed the Porod law. Thus, the scattered intensity $I(q)$ at very large q values will be proportional to the q^{-4} term, this relationship is valid only for sharp interfaces.

The least recognized forms of the Porod approximation are for the *anisotropic* system. If we consider the cylindrical scattering expression of [equation \(B1.9.61\)](#), there are two principal axes (z and r directions) to be discussed

-24-

$$I(q) = K_x V \langle \eta^2 \rangle \int_0^\infty \gamma(r) 2\pi J_0(q_r r) r dr \int_0^\infty 2\gamma(z) \cos(q_z z) dz. \quad (\text{B1.9.61})$$

where q_r and q_z are the scattering vectors along the r or z directions, respectively.

- (1) For the component of the scattered intensity along the equatorial direction (i.e., perpendicular to the cylinder direction, $q_z = 0$), [equation \(B1.9.61\)](#) can be simplified as

$$I(q_r) = K'_x V \langle \eta^2 \rangle \int_0^\infty \gamma(r) 2\pi J_0(q_r r) r dr \quad (\text{B1.9.87})$$

where $K'_x (=K_x \int_0^\infty 2\gamma(z) dz)$ is a new constant, because the integral of $\gamma(z)$ is a constant. In this equation, the correlation function $\gamma(r)$ can again be expanded by the Taylor series in the region of small r , which becomes

$$I(q_r) = \frac{2\pi K'_x V \langle \eta^2 \rangle}{q_r^2} \left(\int_0^\infty \gamma(0) u J_0(u) du + \int_0^\infty \frac{\gamma'(0)}{q_r} u^2 J_0(u) du + \int_0^\infty \frac{\gamma''(0)}{2q_r^2} u^3 J_0(u) du + \dots \right). \quad (\text{B1.9.88})$$

The following relationships can be used to simplify the above equation

$$\int_0^\infty x J_0(x) dx = 0, \int_0^\infty x^2 J_0(x) dx = -1, \int_0^\infty x^3 J_0(x) dx = 0. \quad (\text{B1.9.89})$$

Thus, we obtain

$$(\text{B1.9.90})$$

$$I(q_r) = -\frac{2\pi K'_x V \langle \eta^2 \rangle}{q_r^3} \gamma'(0) + O(q_r^{-2n-1}, d^{2n-1} \gamma(0)/dr^{2n-1}) \quad (n = 2, 3, 4 \dots).$$

This expression holds true only in the large q_r region in reciprocal space (or small r in real space). Since the second term in equation (B1.9.88) rapidly approaches zero at large q_r , we have

$$\lim_{q_r \rightarrow \infty} I(q_r) = -\frac{2\pi K'_x V \langle \eta^2 \rangle}{q_r^3} \gamma'(0). \quad (\text{B1.9.91})$$

-25-

This equation is the Porod law for the large-angle tail of the scattering curve along the equatorial direction, which indicates that the equatorial scattered intensity $I(q_r)$ is proportional to q_r^{-3} in the Porod region of an anisotropic system. Cohen and Thomas have derived the following relationships for the two-dimensional two-phase system (with sharp interfaces) such as fibres [29].

$$\langle \eta^2 \rangle = v_1(1 - v_1)(\rho_1 - \rho_2)^2 \quad (\text{B1.9.92})$$

$$\gamma'(0) = -\frac{L_I}{A} \frac{1}{\pi v_1(1 - v_1)} \quad (\text{B1.9.93})$$

where v represents the area fraction of a phase, L_I is the length of the interface between the two phases and A is the total cross sectional area.

- (2) For the component of the scattered intensity along the meridional direction (i.e., parallel to the cylinder direction, $q_r = 0$), equation (B1.9.61) can be rewritten as

$$I(q_z) = K''_x V \langle \eta^2 \rangle \int_0^\infty 2\gamma(z) \cos(q_z z) dz \quad (\text{B1.9.94})$$

where $K''_x (= K_x \int_0^\infty 2\pi \gamma(r) r dr)$ is a constant, because the integral of $\gamma(r)r$ is a constant. Again, we can expand the term $\gamma(z)$ using the Taylor series in the small z region to derive the Porod law. Equation (B1.9.94) thus becomes ($q_z z = v$).

$$I(q_z) = \frac{2K''_x V \langle \eta^2 \rangle}{q_z} \left(\int_0^\infty \gamma(0) \cos(v) dv + \int_0^\infty \frac{\gamma'(0)}{q_z} v \cos(v) dv + \int_0^\infty \frac{\gamma''(0)}{2q_z^2} v^2 \cos(v) dv + \dots \right). \quad (\text{B1.9.95})$$

The following expressions can be derived to simplify (B1.9.95)

$$\int_0^\infty \cos(x) dx = 0, \int_0^\infty x \cos(x) dx = -1, \int_0^\infty x^2 \cos(x) dx = 0. \quad (\text{B1.9.96})$$

Thus,

$$I(q_z) = -\frac{2K_x'' V_s \langle \eta^2 \rangle}{q_z^2} \gamma'(0) + O(q_z^{-2n}, d^{2n-1} \gamma(0) / dr^{2n-1}) \quad (n = 2, 3, 4 \dots). \quad (\text{B1.9.97})$$

Again, the second term in (B1.9.97) rapidly approaches zero at large q_z , thus we obtain

$$\lim_{q_z \rightarrow \infty} I(q_z) = -\frac{2K_x'' V \langle \eta^2 \rangle}{q_z^2} \gamma'(0). \quad (\text{B1.9.98})$$

If the scatterers are elongated along the fibre direction and the two phases and their interfaces have the same consistency in both radial and z directions, a different expression of the Porod law can be derived.

$$\lim_{q_z \rightarrow \infty} I(q_z) = -\frac{2K_x'' V (\rho_1 - \rho_2)^2 C}{\pi q_z^2 A} \quad (\text{B1.9.99})$$

This is the Porod law for the large angle tail of the scattering curve in the meridional direction. In this case, the scattered intensity is proportional to q_z^{-2} at large scattering angles.

B1.9.4.4 SCATTERING FROM SEMICRYSTALLINE POLYMERS

Semicrystalline polymers are ideal objects to be studied by small-angle x-ray scattering (SAXS), because electron density variations of the semicrystalline morphology (with alternating crystalline and amorphous structures) have a correlation length of several hundred Ångströms, which falls in the resolution range of SAXS (1–100 nm). In addition, the semicrystalline structures can usually be described by assuming electron density variations to occur in one coordinate only. In this case, the scattered intensity $I(q)$ can be described by a one-dimensional correlation function $\gamma_1(r)$.

The scattered intensity measured from the isotropic three-dimensional object can be transformed to the one-dimensional intensity function $I_1(q)$ by means of the Lorentz correction [15]

$$I_1(q) = I(q) 4\pi q^2 \quad (\text{B1.9.100})$$

where the term $4\pi q^2$ represents the scattering volume correction in space. In this case, the correlation and interface distribution functions become

$$\gamma_1(r) = \left(\int_0^\infty I_1(q) \cos(qr) dq \right) / Q \quad (\text{B1.9.101})$$

$$q_1(r) = \partial^2(\gamma_1(r))/\partial r^2 = \left(- \int_0^\infty I_1(q)q^2 \cos(qr) dq \right) / Q \quad (\text{B1.9.102})$$

where $Q (= \int_0^\infty I_1(q) dq = 4\pi \int_0^\infty I(q)q^2 dq)$ is the invariant. The above two equations are valid only for lamellar structures.

Lamellar morphology variables in semicrystalline polymers can be estimated from the correlation and interface distribution functions using a two-phase model. The analysis of a correlation function by the two-phase model has been demonstrated in detail before [30, 31]. The thicknesses of the two constituent phases (crystal and amorphous) can be extracted by several approaches described by Strobl and Schneider [32]. For example, one approach is based on the following relationship:

$$x_1 x_2 = \frac{B}{L} \quad (\text{B1.9.103})$$

where x_1 and x_2 are the linear fractions of the two phases within the lamellar morphology, B is the value of the abscissa when the ordinate is first equal to zero in $\gamma_1(r)$ and L represents the long period determined as the first maximum of $\gamma_1(r)$ (figure B1.9.11).

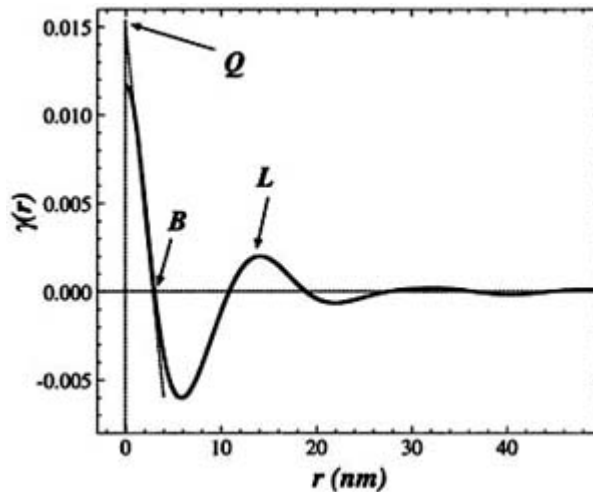


Figure B1.9.11. The analysis of correlation function using a lamellar model.

The analysis of the interface distribution function $g_1(r)$ is relatively straightforward [33]. The profile of $g_1(r)$ can be directly calculated from the Fourier transformation of the interference function or by taking the second derivative of the correlation function (B1.9.102). In the physical sense, the interface distribution function represents the probability of finding an interface along the density profile. A positive value indicates an even number of interfaces within a real space distance with respect to the origin. A negative value indicates an odd number of interfaces within the corresponding distance. With lamellar morphology, odd numbers of interfaces correspond to integral numbers of long periods. The shape of the probability distribution with distance for a given interface manifests as the shape of the corresponding peak on the interface distribution function. These distributions can be deconvoluted to reveal more detailed morphological parameters [34]. The schematic diagram of the relationships between the one-dimensional electron density profile, $\rho(r)$, correlation function,

$\gamma_1(r)$, and interface distribution function, $g_1(r)$, is shown in figure B1.9.12. In general, we find the values of the long period calculated from different methods, such as a conventional analysis by using Bragg's law, the correlation function and the interface distribution function, to be quite different. However, their trends as functions of time and temperature are usually similar. The ordering of these long periods indicates the heterogeneity of the lamellar distributions in the morphology [35].

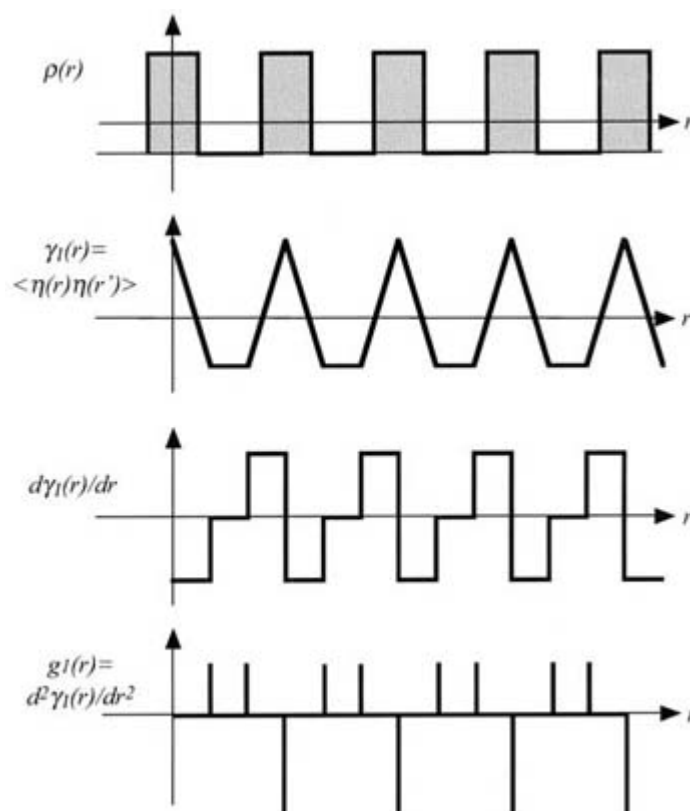


Figure B1.9.12. The schematic diagram of the relationships between the one-dimensional electron density profile, $\rho(r)$, correlation function $\gamma_1(r)$ and interface distribution function $g_1(r)$.

In oriented systems (fibres or stretched films), the scattered image often appears as a two-bar or a four-point pattern with the scattering maximum at or near the meridian (fibre axis). The one-dimensional scattered intensity along the meridian must be calculated by the projection method using the following formalism

$$I_1(q_z) = \int_0^\infty I(q_r, q_z) q_r dq_r. \quad (\text{B1.9.104})$$

This intensity can be used to calculate the correlation function (B1.9.101) and the interface distribution function (B1.9.102) and to yield the lamellar crystal and amorphous layer thicknesses along the fibre.

Recently, a unique approach for using the correlation function method has been demonstrated to extract morphological variables in crystalline polymers from time-resolved synchrotron SAXS data. The principle of the calculation is based on two alternative expressions of Porod's law using the form of interference function [33, 36]. This approach enables a continuous estimate of the Porod constant, corrections for liquid scattering

and finite interface between the two phases, from the time-resolved data. Many detailed morphological variables such as lamellar long period, thicknesses of crystal and amorphous phases, interface thickness and scattering invariant can be estimated. An example analysis of isothermal crystallization in poly (ethyleneterephthalate) (PET) at 230°C measured by synchrotron SAXS is illustrated here. Time-resolved synchrotron SAXS profiles after the removal of background scattering (air and windows), calculated correlation function profiles and morphological variables extracted by using the two-phase crystal lamellar model are shown in figure B1.9.13. Two distinguishable stages are seen in this figure (the first stage was collected at 5 seconds per scan; the later stage was collected at 30 seconds per scan). It is seen that the long period L_c^M and crystal lamellar thickness l_c decrease with time. This behaviour can be explained by the space filling of thinner secondary crystal lamellae after the initial formation of thicker primary lamellae during isothermal crystallization [36].

-30-

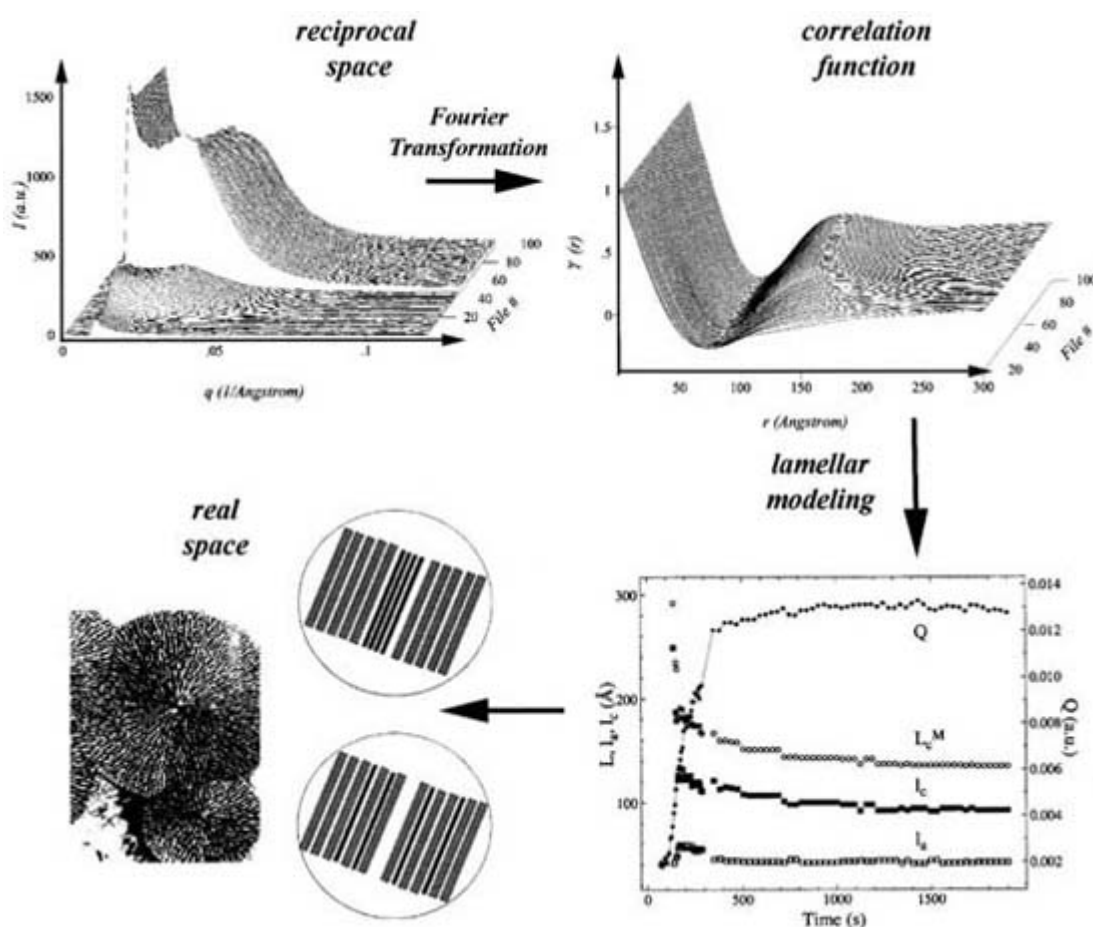


Figure B1.9.13. Time-resolved SAXS profiles during isothermal crystallization (230 °C) of PET (the first 48 scans were collected with 5 seconds scan time, the last 52 scans were collected with 30 seconds scan time); calculated correlation functions $\gamma(r)$ (normalized by the invariant Q) and lamellar morphological variables extracted from the correlation functions (invariant Q , long period L_c^M , crystal lamellar thickness l_c and interlamellar amorphous thickness l_a).

-31-

B1.9.5 NEUTRON SCATTERING

Neutron scattering depends upon nuclear properties, which are related to fluctuations in the neutron scattering cross section σ between the scatterer and the surroundings. The scattered amplitude from a collection of scatterers can thus be written as (similar to (B1.9.29)):

$$F(q) = \sum b_j \exp(iqr_j) \quad (\text{B1.9.105})$$

where b_j is referred to as the ‘scattering length’ of the object j ; its square is related to the scattering cross section ($\sigma_j = 4\pi b_j^2$) of the j th object. The value of b_j depends on the property of the nucleus and is generally different for different isotopes of the same element. As neutron scattering by nature is a nuclear event and the wavelength used in neutron scattering is much larger than the nuclear dimensions, the intranuclear interference of waves due to neutron scattering need not be considered such that b_j is normally independent of scattering angle.

In neutron scattering, the scattered intensity is often expressed in terms of the differential scattering cross section:

$$\frac{\partial \sigma}{\partial \Omega} = \sum_{i,j}^N \langle b_i b_j \exp(iq(r_j - r_i)) \rangle \quad (\text{B1.9.106})$$

where $d\Omega$ is the small solid angle into which the scattered neutrons are accepted. We can mathematically divide the above equation into two parts:

$$\frac{\partial \sigma}{\partial \Omega} = \sum_{i,j}^N \langle b_i b_j \exp(iq(r_j - r_i)) \rangle \quad (\text{B1.9.107})$$

If we define

$$\overline{\Delta b^2} = \langle b^2 \rangle - \langle b \rangle^2 \quad (\text{B1.9.108})$$

we can obtain the following relationship:

$$\frac{\partial \sigma}{\partial \Omega} = N \overline{\Delta b^2} + \langle b \rangle^2 \sum_{i,j}^N \langle \exp(iq(r_j - r_i)) \rangle \quad (\text{B1.9.109})$$

where $\langle b \rangle^2 = \overline{b^2}$. The first term represents the incoherent scattering, which depends only on the fluctuations of the scattering length b occupying the different positions. The second term is the coherent scattering, which

depends on the positions of all the scattering centres and is responsible for the angular dependence of the scattered intensity. In this chapter, we shall only focus on the phenomenon of coherent scattering. Using the concept of correlation function (as in equation (B1.9.54)), we obtain

$$I(q) = \left(\frac{\partial \sigma}{\partial \Omega} \right)_{\text{coh}} = K_n b^2 \int \gamma(r) e^{-iqr} d^3r \quad (\text{B1.9.110})$$

where K_n is a calibration constant depending on the incident flux and apparatus geometry. The expressions of scattered intensity for isotropic and anisotropic systems can be obtained similarly to equations (B1.9.49) and (B1.9.50).

B1.9.5.1 SCATTERING FROM MULTICOMPONENT SYSTEMS

Equation (B1.9.110) is for a system containing only one type of scattering length. Let us consider the system containing more than one species, such as a two-species mixture with N_1 molecules of scattering length b_1 and N_2 of scattering length b_2 . The coherent scattered intensity $I(q)$ becomes

$$I(q) = K_n \left\{ b_1^2 \sum_{i_1}^{N_1} \sum_{j_1}^{N_1} \langle \exp(-iqr_{ij}) \rangle + 2b_1 b_2 \sum_{i_1}^{N_1} \sum_{j_2}^{N_2} \langle \exp(-iqr_{ij}) \rangle + b_2^2 \sum_{i_2}^{N_2} \sum_{j_2}^{N_2} \langle \exp(-iqr_{ij}) \rangle \right\} \quad (\text{B1.9.111})$$

or

$$I(q) = K_n \left\{ b_1^2 \int \gamma_{11}(r) e^{-iqr} d^3r + 2b_1 b_2 \int \gamma_{12}(r) e^{-iqr} d^3r + b_2^2 \int \gamma_{22}(r) e^{-iqr} d^3r \right\} \quad (\text{B1.9.112})$$

-33-

where γ_{ij} is the correlation function between the local density of constituents i and j . We can define three partial structural factors, S_{11} , S_{12} and S_{22}

$$S_{ij}(r) = K_n \int \gamma_{ij}(r) e^{-iqr} d^3r. \quad (\text{B1.9.113})$$

Thus

$$I(q) = b_1^2 S_{11}(q) + 2b_1 b_2 S_{12}(q) + b_2^2 S_{22}(q). \quad (\text{B1.9.114})$$

If the two constituting molecular volumes are identical in a two component system, we can obtain [37, 38]

$$S_{11}(q) = S_{22}(q) = -S_{12}(q) \quad (\text{B1.9.115})$$

or

$$I(q) = (b_1 - b_2)^2 S_{11}(q) = (b_1 - b_2)^2 S_{22}(q) = -(b_1 - b_2)^2 S_{12}(q). \quad (\text{B1.9.116})$$

Let us assume that we have $p + 1$ different species; equation (B1.9.111) can be generalized as

$$I(q) = \sum_{i=1}^p b_i^2 S_{ii}(q) + 2 \sum_{i<j}^p b_i b_j S_{ij}(q). \quad (\text{B1.9.117})$$

B1.9.5.2 PROPERTIES OF $S(Q)$

As we have introduced the structure factor $S(q)$ (B1.9.113), it is useful to separate this factor into two categories of interferences for a system containing N scattering particles [9]:

$$S(q) = N[P(q) + NQ(q)]. \quad (\text{B1.9.118})$$

The first term, $P(q)$, represents the interferences within particles and its contribution is proportional to the number of particle, N . The second term, $Q(q)$, involves interparticle interferences and is proportional to the number of pairs of particles, N^2 .

Let us consider the scattered intensity from a binary incompressible mixture of two species (containing N_1 molecules of particle 1 and N_2 molecules of particle 2) as in (B1.9.112); we can rewrite the relationship as

$$I(q) = b_1^2 N_1 [P_1(q) + N_1 Q_{11}(q)] + b_2^2 N_2 [P_2(q) + N_2 Q_{22}(q)] + 2b_1 b_2 N_1 N_2 Q_{12}(q) \quad (\text{B1.9.119})$$

where $P_i(q)$ is the intramolecular interference of species i and $Q_{ij}(q)$ is the intermolecular interferences between species i and j .

Two of the most important functions in the application of neutron scattering are the use of deuterium labelling for the study of molecular conformation in the bulk state and the use of deuterium solvent in polymer solutions. In the following, we will consider several different applications of the general formula to deuteration.

- (1) Let us first consider two identical polymers, one deuterated and the other not, in a melt or a glassy state. The two polymers (degree of polymerization d) differ from each other only by scattering lengths b_H and b_D . If the total number of molecules is N , x is the volume fraction of the deuterated species ($x = N_D / N$, with $N_D + N_H = N$). According to equation (B1.9.116), we obtain

$$I(q) = (b_D - b_H)^2 S_{DD}(q) \quad (\text{B1.9.120})$$

or S_{HH} or $-S_{HD}$. The coherent scattering factor can further be expressed in terms of $P(q)$ and $Q(q)$ as (see equation (B1.9.117))

$$\begin{aligned} S_{DD} &= xNd^2P(q) + x^2N^2d^2Q(q) = (1-x)Nd^2P(q) + (1-x)^2d^2N^2Q(q) \\ &= -x(1-x)N^2d^2Q(q). \end{aligned} \quad (\text{B1.9.121})$$

Thus

$$NQ(q) = -P(q). \quad (\text{B1.9.122})$$

By combining equations (B1.9.120) and (B1.9.122), we have

$$I(q) = (b_D - b_H)^2 x(1-x)Nd^2P(q). \quad (\text{B1.9.123})$$

-35-

- (2) Next, let us consider the case of a system made up of two polymers with different degrees of polymerization d_D for the deuterated species and d_H for the other. The generalized expression of [B1.9.122](#) becomes:

$$-N_D d_D^2 P_D(q) = N_D^2 d_D^2 Q_D(q) + N_D N_H d_D d_H Q_{DH}(q). \quad (\text{B1.9.124})$$

- (3) Let us take two polymers (one deuterated and one hydrogenated) and dissolve them in a solvent (or another polymer) having a scattering length b_0 . The coherent scattered intensity can be derived from [\(B1.9.117\)](#), which gives

$$\begin{aligned} I(q) &= (b_D - b_0)^2 S_{DD}(q) + (b_H - b_0)^2 S_{HH}(q) \\ &\quad + 2(b_D - b_0)(b_H - b_0)S_{HD}(q) \end{aligned} \quad (\text{B1.9.125})$$

where

$$\begin{aligned} S_{DD} &= xNd^2P(q) + x^2N^2d^2Q(q) \\ S_{HH} &= (1-x)Nd^2P(q) + (1-x)^2N^2d^2Q(q) \\ S_{HD} &= -x(1-x)Nd^2Q(q). \end{aligned} \quad (\text{B1.9.126})$$

Thus, we have

$$I(q) = (b_D - b_H)^2 x(1-x)Nd^2P(q) + (xb_D + (1-x)b_H - b_0)^2 Nd^2[P(q) + dQ(q)]. \quad (\text{B1.9.127})$$

The second term of the above equation gives an important adjustment to contrast variation between the

solvent and the polymer. If one adjusts the scattering length b_0 of the solvent by using a mixture of deuterated and hydrogenated solvent such that

$$xb_D + (1 - x)b_H - b_0 = 0 \quad (\text{B1.9.128})$$

then we can obtain (B1.9.123). This experiment thus yields directly the form factor $P(q)$ of the polymer molecules in solution even at high polymer concentrations.

B1.9.5.4 ANALYSIS OF MOLECULAR PARAMETERS

There are many different data analysis schemes to estimate the structure and molecular parameters of polymers from the neutron scattering data. Herein, we will present several common methods for characterizing the scattering profiles, depending only on the applicable q range. These methods, which were derived based on different assumptions, have

-36-

different limitations. We caution the reader to check the limitations of each method before its application.

- (1) If we deal with a solution at very low concentrations, we can ignore the interactions between the particles and express the scattered intensity as

$$\log I(q) = \log I(0) - \frac{1}{3}q^2 R_g^2 + \dots \quad (\text{B1.9.129})$$

where R_g is the radius of gyration. This equation is similar to (B1.9.35). If one plots $\log[I(q)]$ as a function of q^2 , the initial part is a straight line with a negative slope proportional to R_g^2 , which is called the Guinier plot. This approach is only suitable for scattering in the low qR_g range and in dilute concentrations. A similar expression proposed by Zimm has a slightly different form:

$$\frac{Nd^2}{S(q)} = (1 + \frac{1}{3}q^2 R_g^2 + \dots) \quad (\text{B1.9.130})$$

where N is the number of scattering objects and d is the degree of polymerization. This equation is similar to equation (B1.9.35). Thus if one plots $1/S(q)$ as function q^2 , the initial slope is $R_g^2/3$.

The above radius of gyration is for an isotropic system. If the system is anisotropic, the mean square radius of gyration is equal to

$$\langle R_g^2 \rangle = \langle R_x^2 \rangle + \langle R_y^2 \rangle + \langle R_z^2 \rangle \quad (\text{B1.9.131})$$

where R_x , R_y and R_z are the components of the radius of gyration along the x , y , z axes. For the isotropic system

$$\langle R_x^2 \rangle = \langle R_y^2 \rangle + \langle R_z^2 \rangle = \frac{1}{3} \langle R_g^2 \rangle. \quad (\text{B1.9.132})$$

- (2) In the intermediate and high q range, the analysis becomes quite different. The qualitative interpretation for the scattering profile at the high q range may make the use of scaling argument proposed by de Gennes [39]. If we neglect the intermolecular interactions, we can write

$$S(q) = Nd^2 P(qR_g) = V\varphi d P(qR_g) \quad (\text{B1.9.133})$$

-37-

where V is the volume of the sample, φ is the volume fraction occupied by the scattering units in polymer with a degree of polymerization d and qR_g is a dimensionless quantity which is associated with a characteristic dimension. Typically, the term $P(qR_g)$ can be approximated by $(qR_g)^{-\beta}$. Since the relationships between R_g and z are known as follows:

a Gaussian chain	$R_g \approx d^{0.5}$	
a chain with excluded volume (Flory)	$R_g \approx d^{0.6}$	
a rod	$R_g \approx d^{1.0}$	(B1.9.134)
general expression	$R_g \approx d^a$.	

This leads to the following equation:

$$S(q) = V\varphi d (qd^a)^{-\beta} = V\varphi q^{-\beta} d^{1-a\beta}. \quad (\text{B1.9.135})$$

In order to have $S(q)$ independent of d , the power of d must be zero giving $a = 1/\beta$. This gives rise to the following relationships:

a Gaussian chain	$S(q) = q^{-2}$	
a chain with excluded volume (Flory)	$S(q) = q^{-1.66}$	(B1.9.136)
a rod	$S(q) = q^{-1}$	

- (3) The quantitative analysis of the scattering profile in the high q range can be made by using the approach of Debye *et al* as in equation (B1.9.52). As we assume that the correlation function $\gamma(r)$ has a simple exponential form $\gamma(r) = \exp(-r/a_c)$, where a_c is the correlation length, the scattered intensity can be expressed as

$$I(q) = \frac{8\pi a_c^3 b_v^2 \varphi (1 - \varphi)}{(1 + q^2 a_c^2)^2} \quad (\text{B1.9.137})$$

where φ is the volume fraction of the component (with scattering length b_1 and volume of the monomer of the polymer v_1), $(1 - \varphi)$ is the volume fraction of the solvent (with scattering length b_0 and volume of the monomer of the other polymer v_0) and $b_v = b_1/v_1 - b_0/v_0$. Thus the correlation length a_c can be calculated by plotting $I(q)^{-1/2}$ versus q^2 , using equation (B1.9.53) (as light scattering).

A more general case of continuously varying density was treated by Ornstein and Zernicke for scattering of

-38-

opalescence from a two-phase system [40]. They argued that

$$\gamma(r) = \text{constant} \frac{\exp(-r/\xi)}{r} \quad (\text{B1.9.138})$$

where ξ is a characteristic length. This leads to

$$S(q) = \frac{B}{1 + q^2 \xi^2} \quad (\text{B1.9.139})$$

where B is a constant.

B1.9.5.5 CALCULATE THERMODYNAMIC PARAMETER

In polymer solutions or blends, one of the most important thermodynamic parameters that can be calculated from the (neutron) scattering data is the enthalpic interaction parameter χ between the components. Based on the Flory–Huggins theory [41, 42], the scattering intensity from a polymer in a solution can be expressed as

$$\frac{1}{s(q)} = \frac{1}{\varphi z P(q)} + \frac{1}{\varphi_0} - 2\chi \quad (\text{B1.9.140})$$

where $s(q) = S(q)/N_T$ corresponds to the scattered intensity from a volume equal to that of a molecule of solvent, N_T is the ratio of the solution volume to the volume of one solvent molecule ($N_T = V/v_s$), φ is the volume fraction occupied by the polymer and φ_0 is the volume fraction occupied by the solvent, v_s is the solvent specific volume. For a binary polymer blend, equation (B1.9.140) can be generalized as:

$$\frac{1}{s(q)} = \frac{1}{\varphi_1 z_1 P_1(q)} + \frac{1}{\varphi_2 z_2 P_2(q)} - 2\chi \quad (\text{B1.9.141})$$

where the subscripts 1 and 2 represent a solution of polymer 1 in polymer 2. The generalized Flory–Huggins model in the de Gennes formalism with the random phase approximation has the form:

$$\frac{V}{I(q)} = \frac{1}{b_v^2} \left\{ \frac{1}{(C_1 M_1 / N_A) P_1(q) (v_1^s)^2} + \frac{1}{(C_2 M_2 / N_A) P_2(q) (v_2^s)^2} - \frac{2\chi}{v_0} \right\} \quad (\text{B1.9.142})$$

where b_v is the contrast factor as described before, N_A is Avogadro's number, C is the concentration, M is the molecular weight, v_s is the specific volume ($C_1 v_1^s + C_2 v_2^s = 1$) and v_0 is an arbitrary reference volume.

In the case of low concentration and low q expansion, equation (B1.9.142) can be expressed by replacing $P_i(q)$ with the Guinier approximation (B1.9.130)

$$\frac{V u_0 b_v^2}{I(q)} = \left\{ \frac{1}{\varphi_1 z_1} + \frac{1}{\varphi_2 z_2} - 2\chi \right\} + \frac{1}{3} \left\{ \frac{R_{g1}^2}{\varphi_1 z_1} + \frac{R_{g2}^2}{\varphi_2 z_2} \right\} q^2. \quad (\text{B1.9.143})$$

Thus the slope of the $I(q)^{-1}$ versus q^2 plot is related to the values of two radii of gyration.

B1.6.1 CONCLUDING REMARKS

In this chapter, we have reviewed the general scattering principles from matter by light, neutrons and x-rays and the data treatments for the different states of matter. The interaction between radiation and matter has the same formalism for all three cases of scattering, but the difference arises from the intrinsic property of each radiation. The major difference in data treatments results from the different states of matter. Although we have provided a broad overview of the different treatments, the content is by no means complete. Our objective in this chapter is to provide the reader a general background for the applications of scattering techniques to materials using light, neutrons and x-rays.

REFERENCES

- [1] van de Hulst H C 1957 *Light Scattering by Small Particles* (New York: Wiley)
- [2] Kerker M 1969 *The Scattering of Light and Other Electromagnetic Radiation* (New York: Academic)
- [3] Berne B J and Pecora R 1976 *Dynamic Light Scattering* (New York: Wiley-Interscience)
- [4] Chen S H, Chu B and Nossal R (eds) 1981 *Scattering Techniques Applied to Supramolecular and Nonequilibrium Systems* (New York: Plenum)
- [5] Chu B 1991 *Laser Light Scattering, Basic Principles and Practice* 2nd edn (New York: Academic) (See also the first edition (published in 1974) that contains more mathematical derivations.)
- [6] Schmitz K S 1990 *An Introduction to Dynamic Light Scattering by Macromolecules* (New York: Academic)
- [7] Linden P and Zemb Th (eds) 1991 *Neutron, X-ray and Light Scattering: Introduction to an Investigation Tool for Colloidal and Polymeric Systems* (Amsterdam: North-Holland)
- [8] Lovesey S W 1984 *Theory of Neutron Scattering from Condensed Matter* vol 1 (Oxford: Oxford University Press)
- [9] Higgins J S and Benoît H C 1994 *Polymers and Neutron Scattering* (Oxford: Oxford Science)
- [10] Wignall G D 1987 *Encyclopedia of Polymer Science and Engineering* vol 12 (New York: Wiley) p 112
- [11] Wignall G D, Crist B, Russell T P and Thomas E L (eds) 1987 *Mater. Res. Soc. Symp. Proc.* (Pittsburgh, PA: Materials Research Society) vol 79

- [12] Guinier A and Fournet G 1955 *Small Angle Scattering of X-rays* (New York: Wiley)

- [13] Brumberger H (ed) 1967 *Small-Angle X-ray Scattering* (New York: Gordon and Breach)
- [14] Glatter O and Kratky O 1982 *Small Angle X-ray Scattering* (New York: Academic)
- [15] Baltá-Calleja F J and Vonk G G 1989 *X-ray Scattering of Synthetic Polymers* (New York: Elsevier)
- [16] Ditchburn R W 1953 *Light* (New York: Wiley-Interscience)
- [17] Debye P 1915 *Ann. Phys., Lpz.* **46**809
- [18] Pusey P N and Tough R A 1985 *Dynamic Light Scattering ed R Pecora (New York: Plenum) ch 4*
- [19] Chu B, Wu C and Zuo J 1987 *Macromolecules* **20** 700
- [20] Debye P and Bueche A M 1949 *J. Appl. Phys.* **20** 518
- [21] Debye P, Anderson H R and Brumberger H 1957 *J. Appl. Phys.* **28** 679
- [22] Glatter O 1979 *J. Appl. Crystallogr.* **12** 166
- [23] Glatter O 1977 *J. Appl. Crystallogr.* **10** 415
- [24] Glatter O 1981 *J. Appl. Crystallogr.* **14** 101
- [25] Strey R, Glatter O, Schubert K V and Kaler E W 1996 *J. Chem. Phys.* **105** 1175
- [26] Meisenberger O, Pilz I and Heumann H 1980 *FEBS Lett.* **112** 39
- [27] Meisenberger O, Heumann H and Pilz I 1980 *FEBS Lett.* **122** 117
- [28] Porod G 1952 *Kolloid Z. Z. Polym.* **125** 51
Porod G 1952 *Kolloid Z. Z. Polym.* **125** 108
- [29] Cohen Y and Thomas E 1988 *Macromolecules* **21** 433
Cohen Y and Thomas E 1988 *Macromolecules* **21** 436
- [30] Vonk C G and Kortleve G 1967 *Colloid Polym. Sci.* **220** 19
- [31] Vonk G G 1973 *J. Appl. Crystallogr.* **6** 81
- [32] Strobl G R and Schneider M 1980 *J. Polym. Sci. B* **18** 1343
- [33] Ruland W 1977 *Colloid. Polym. Sci.* **255** 417
- [34] Stribeck N and Ruland W 1978 *J. Appl. Crystallogr.* **11** 535
- [35] Santa Cruz C, Stribeck N, Zachmann H G and Baltá-Calleja F J 1991 *Macromolecules* **24** 5980
- [36] Hsiao B S and Verma R K 1998 *J. Synchrotron Radiat.* **5** 23
- [37] Daoud M, Cotton J P, Farnoux B, Jannink G, Sarma S and Benoît H H 1975 *Macromolecules* **8** 804
- [38] Akcasu A, Summerfield G C, Jahshan S N, Han C C, Kim C Y and Yu H 1980 *J. Polym. Sci. Polym. Phys.* **18** 863
- [39] de Gennes P G 1979 *Scaling Concepts in Polymer Physics* (Ithaca, NY: Cornell University Press)
- [40] Ornstein L S and Zernike F 1914 *Proc. Acad. Sci. Amsterdam* **17** 793
- [41] Flory P J 1942 *J. Chem. Phys.* **10** 51
- [42] Huggins M L 1942 *J. Phys. Chem.* **46** 151

B1.10 Coincidence techniques

Michael A Coplan

B1.10 INTRODUCTION

Time is a fundamental variable in almost all experimental measurements. For some experiments, time is

measured directly, as in the determination of radioactive half-lives and the lifetimes of excited states of atoms and molecules. Velocity measurements require the measurement of the time required for an object to travel a fixed distance. Time measurements are also used to identify events that bear some correlation to each other and as a means for reducing background noise in experiments that would otherwise be difficult or impossible to perform. Examples of time correlation measurements are so-called coincidence measurements where two or more separate events can be associated with a common originating event by virtue of their time correlation. Positron annihilation in which a positron and an electron combine to yield two gamma rays, is an example of this kind of coincidence measurement. Electron impact ionization of atoms in which an incident electron strikes an atom, ejects an electron and is simultaneously scattered is another example. In such an experiment the ejected and scattered electrons originate from the same event and their arrival at two separate detectors is correlated in time. In one of the very first coincidence experiments Bothe and Geiger [1] used the time correlation between the recoil electron and inelastically scattered x-ray photon, as recorded on a moving film strip, to identify Compton scattering of an incident x-ray and establish energy conservation on the microscopic level. An example of the use of time correlation to enhance signal-to-noise ratios can be found in experiments where there is a background signal that is uncorrelated with the signal of interest. The effect of penetrating high-energy charged particles from cosmic rays can be eliminated from a gamma ray detector by construction of an anti-coincidence shield. Because a signal from the shield will also be accompanied by a signal at the detector, these spurious signals can be eliminated.

Experiments in almost all subjects from biophysics to high-energy particle physics and cosmic ray physics use time correlation methods. There are a few general principles that govern all time correlation measurements and these will be discussed in sufficient detail to be useful in not only constructing experiments where time is a parameter, but also in evaluating the results of time measurements and optimizing the operating parameters. In a general way, all physical measurements either implicitly or explicitly have time as a variable. Recognizing this is essential in the design of experiments and analysis of the results. The rapid pace of improvements and innovation in electronic devices and computers have provided the experimenter with electronic solutions to experimental problems that in the past could only be solved with custom hardware.

B1.10.2 STATISTICS

B1.10.2.1 CORRELATED AND RANDOM EVENTS

Correlated events are related in time and this time relation can be measured either with respect to an external clock or to the events themselves. Random or uncorrelated events bear no fixed time relation to each other but, on the other

-2-

hand, their very randomness allows them to be quantified. Consider the passing of cars on a busy street. It is possible to calculate the probability, $P_{\bar{n}}(n)$ that n cars pass within a given time interval in terms of the average number of cars, \bar{n} , in that interval where $P_{\bar{n}}(n)$, the Poisson distribution is given by [2].

$$P_{\bar{n}}(n) = \frac{\bar{n}^n}{n!} e^{-\bar{n}}.$$

For example, if one finds that 100 cars pass a fixed position on a highway in an hour, then the average number of cars per minute is 100/60. The probability that two cars pass in a minute is given by $\frac{1}{2}0.6^2e^{-0.6} = 0.10$. The probability that three times the average number of cars pass per unit time is 0.02. $P_{\bar{n}}(n)$ and the integral of $P_{\bar{n}}(n)$ for \bar{n} are shown in figure B1.10.1. It is worthwhile to note that only a single parameter, \bar{n} , the average value, is sufficient to define the function; moreover, the function is not symmetric about \bar{n} .

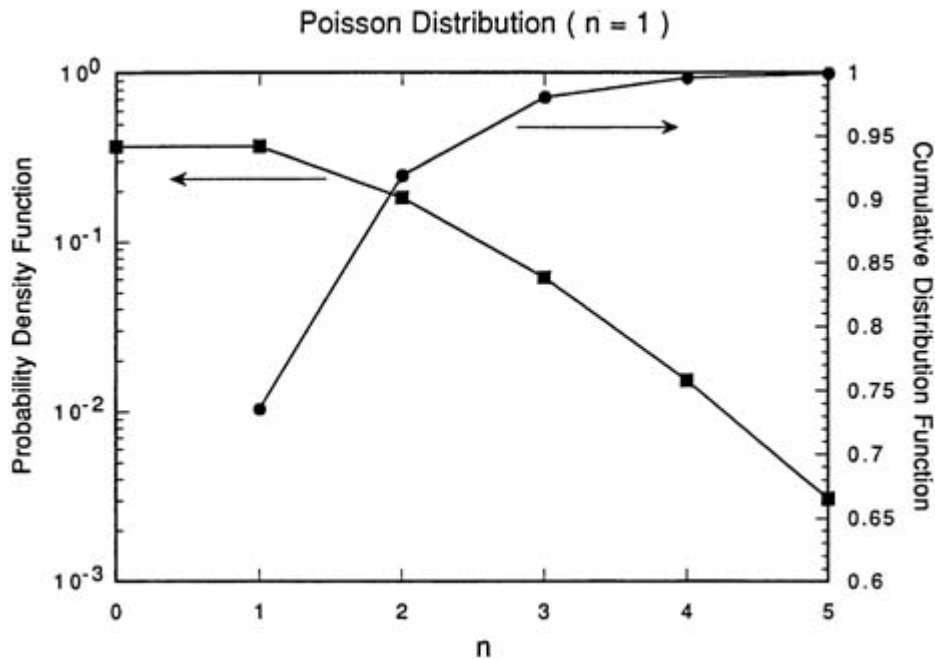


Figure B1.10.1. Poisson distribution for \bar{n} (left vertical axis). Cumulative Poisson distribution for $\bar{n} = 1$ (right vertical axis). The cumulative distribution is the sum of the values of the distribution from 0 to n , where $n = 1, 2, 3, 4, 5$ on this graph.

This example of passing cars has implications for counting experiments. An arrangement for particle counting is shown in [figure B1.10.2](#). It consists of the source of particles, a detector, preamplifier, amplifier/discriminator, counter, and a storage device for recording the results. The detector converts the energy of the particle to an electrical signal that is amplified by a low-noise preamplifier to a level sufficient to be amplified and shaped by the amplifier. The discriminator converts the signal from the amplifier to a standard electrical pulse of fixed height and width, provided that the amplitude of the signal from the amplifier exceeds a set threshold. The counter records the number of pulses from the discriminator for a set period of time. The factors that affect the measurement are counting rate, signal durations, and processing times.

-3-

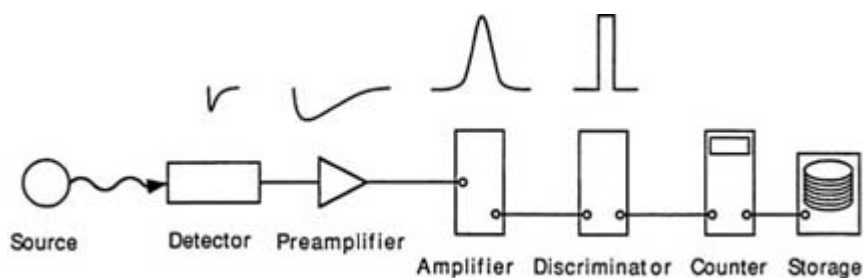


Figure B1.10.2. Schematic diagram of a counting experiment. The detector intercepts signals from the source. The output of the detector is amplified by a preamplifier and then shaped and amplified further by an amplifier. The discriminator has variable lower and upper level thresholds. If a signal from the amplifier exceeds the lower threshold while remaining below the upper threshold, a pulse is produced that can be registered by a preprogrammed counter. The contents of the counter can be periodically transferred to an on-line storage device for further processing and analysis. The pulse shapes produced by each of the devices are shown schematically above them.

In counting experiments, the instantaneous rate at which particles arrive at the detector can be significantly

different from the average rate. In order to assess the rate at which the system can accept data, it is necessary to know how the signals from the detector, preamplifier, amplifier and discriminator each vary with time. For example, if signals are arriving at an average rate of 1 kHz at the detector, the average time between the start of each signal is 10^{-3} s. If we consider a time interval of 10^{-3} s, the average number of signals arriving in that interval is one; if as many as three pulses can be registered in the 10^{-3} s then 99% of all pulses will be counted. To accommodate the case of three pulses in 10^{-3} s it is necessary to recognize that the pulses will be randomly distributed in the 10^{-3} s interval. To register the three randomly distributed pulses in the 10^{-3} s interval requires approximately another factor of three in time resolution, with the result that the system must have the capability of registering events at a 10 kHz uniform rate in order to be able to register with 99% efficiency randomly arriving events arriving at a 1 kHz average rate. For 99.9% efficiency, the required system bandwidth increases to 100 kHz. Provided that the discriminator and counter are capable of handling the rates, it is then necessary to be sure that the duration of all of the electronic signals are consistent with the required time resolution. For 1 kHz, 10 kHz and 100 kHz bandwidths this means signal durations of 0.1, 0.01, and 0.001 ms, respectively. An excellent discussion of the application of statistics to physics experiments is given by Melissinos [3].

The processing times of the electronic units must also be taken into consideration. There are propagation delays associated with the active devices in the electronics circuits and delays associated with the actual registering of events by the counter. Processing delays are specified by the manufacturers of counters in terms of maximum count rate: a 10 MHz counter may only count at a 10 MHz rate if the input signals arriving at a uniform rate. For randomly arriving signals with a 10 MHz average rate, a system with a 100 MHz bandwidth is required to record 99% of the incoming events.

B1.10.2.2 STATISTICAL UNCERTAINTIES

In counting experiments the result that is sought is a rate or the number of events registered per unit time. For

-4-

convenience we divide the total time of the measurement, T , into n equal time intervals, each of length T/n . If there are N_i counts registered in interval i , then the mean or average rate is given by

$$\frac{1}{n} \sum_{i=1}^n \frac{N_i}{T/n} = \frac{N_N}{T}$$

where N_N is the total number of counts registered during the course of the experiment. To assess the uncertainty in the overall measured rate, we assume that the individual rate measurements are statistically distributed about the average value, in other words, they arise from statistical fluctuations in the arrival times of the events and not from uncertainties in the measuring instruments. Here the assumption is that all events are registered with 100% efficiency and that there is negligible uncertainty in the time intervals over which the events are registered.

When the rate measurement is statistically distributed about the mean, the distribution of events can be described by the Poisson distribution, $P_{\bar{n}}(n)$, given by

$$P_{\bar{n}}(n) = \frac{\bar{n}^n}{n!} e^{-\bar{n}}$$

where n is the number of events per unit time and the average number of events per unit time is \bar{n} . The uncertainty in the rate is given by the standard deviation and is equal to the square root of the average rate, $\sqrt{\bar{n}}$. Most significant is the relative uncertainty, $\sqrt{N}/t_T/N/t_T$. The relative uncertainty decreases as the

number of counts per counting interval increases. For a fixed experimental arrangement the number of counts can be increased by increasing the time of the measurement. As can be seen from the formula, the relative uncertainty can be made arbitrarily small by increasing the measurement time; however, improvement in relative uncertainty is proportional to the reciprocal of the square root of the measurement time. To reduce the relative uncertainty by a factor of two, it is necessary to increase the measurement time by a factor of four. One soon reaches a point where the fractional improvement in relative uncertainty requires a prohibitively long measurement time.

In practice, the length of time for an experiment depends on the stability and reliability of the components. For some experiments, the solar neutrino flux and the rate of decay of the proton being extreme examples, the count rate is so small that observation times of months or even years are required to yield rates of sufficiently small relative uncertainty to be significant. For high count rate experiments, the limitation is the speed with which the electronics can process and record the incoming information.

In this section we have examined the issue of time with respect to the processing and recording of signals and also with regard to statistical uncertainty. These are considerations that are the basis for the optimization of more complex experiments where the time correlation between sets of events or among several different events are sought.

B1.10.3 TIME-OF-FLIGHT EXPERIMENTS

Time-of-flight experiments are used to measure particle velocities and particle mass per charge. The typical experiment

requires *start* and *stop* signals from detectors located at the beginning and end of the flight path, see figure B1.10.3. The *time-of-flight* is then the time difference between the signals from the stop and start detectors. The *start* signal is often generated by the opening of a shutter and *stop* by the arrival of the particle at a *stop* detector. Alternatively, the *start* signal may be generated by the particle passing through a *start* detector that registers the passing of the particle without altering its motion in a significant way. The result of accumulating thousands of time-of-flight signals is shown in [figure B1.10.4](#) where the number of events with a given time-of-flight is plotted against time-of-flight specified as channel number. The data for the figure were acquired for a gas sample of hydrogen mixed with air. The different flight times reflect the fact that the ions all have the same kinetic energy.

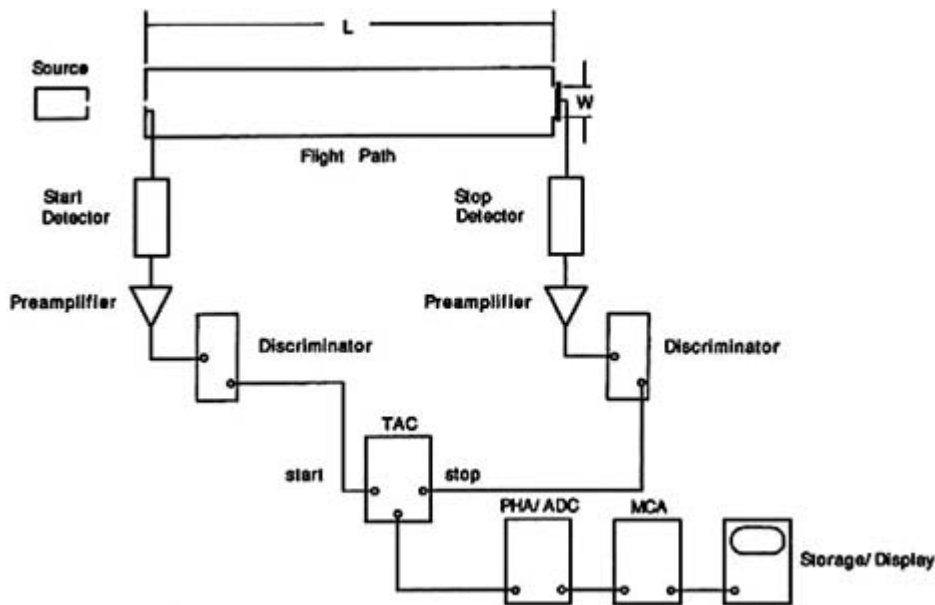


Figure B1.10.3. Time-of-flight experiment. Detectors at the beginning and end of the flight path sense the passage of a particle through the entrance and exit apertures. The width of the exit aperture, W , determines the amount of transverse velocity particles can have and still be detected at the end of the flight path. Transverse velocity contributes to the dispersion of flight times for identical particles with the same kinetic energies. Detector signals are amplified by a preamplifier; threshold discriminators produce standard pulses whenever the incoming signals exceed an established threshold level. Signals from the start and stop discriminators initiate the operation of the TAC (time-to-amplitude converter). At the end of the TAC cycle, a pulse is produced with amplitude proportional to the time between the start and stop signals. The TAC output pulse amplitude is converted to a binary number by a pulse height analyser (PHA) or analogue-to-digital converter (ADC). The binary number serves as an address for the multichannel analyser (MCA) that adds one to the number stored at the specified address. In this way, a collection of flight times, a time spectrum, is built up in the memory of the MCA. The contents of the MCA can be periodically transferred to a storage device for analysis and display.

B1.10.3.1 SOURCES OF UNCERTAINTIES

The measurement of velocity is given by $L(t_2 - t_1)$, where L is the effective length of the flight path and $t_2 - t_1$ is the difference in time between the arrival of the particle at the stop detector (t_2) and the start detector (t_1). The uncertainty

-6-

in the velocity is

$$\frac{1}{\Delta t} dL - \frac{L}{\Delta t^2} d\Delta t$$

where $\Delta t = t_2 - t_1$.

The relative uncertainty is

$$\sqrt{\left(\frac{dL}{L}\right)^2 + \left(\frac{d\Delta t}{\Delta t}\right)^2}$$

In a conventional time-of-flight spectrometer, the transverse velocities of the particles and the angular acceptance of the flight path and *stop* detector determine the dispersion in L . The dispersion in flight times is given by the time resolution of the *start* and *stop* detectors and the associated electronics. When a shutter is used there is also a time uncertainty associated with its opening and closing. A matter that cannot be overlooked in time-of-flight measurements is the rate at which measurements can be made. In the discussion the implicit assumption has been that only one particle is in the flight path at a time. If there is more than one, giving rise to multiple *start* and *stop* signals, it is not possible to associate a unique *start* and *stop* signal with each particle. It cannot be assumed that the first start signal and the first stop signals have been generated by the same particle, because the faster particles can overtake the slower ones in the flight path. When shutters are used, the opening time of the shutter must be sufficiently short to allow no more than one particle to enter the flight path at a time. These considerations give rise to constraints on the rate which measurements can be made. If the longest time-of-flight is T_{\max} , the particles must not be allowed to enter the flight path at a rate that exceeds $0.1/T_{\max}$. Detector response time and the processing time of the electronics should be taken into consideration when calculating T_{\max} . Time-of-flight experiments are inherently inefficient because the rate at which the shutter is opened is set by the time-of-flight of the slowest particle while the duration of the shutter opening is set by the total flux of particles incident on the shutter. The maximum uncertainty in the measured time-of-flight is, on the other hand, determined by T_{\min} , the time for the fastest particle to traverse the flight path.

Many of the electronics in the time-of-flight system are similar to those in the counting experiment, with the exception of the time-to-amplitude converter (TAC) and analogue-to-digital converter (ADC). The TAC has *start* and *stop* inputs and an output. The output is a pulse of amplitude (height) proportional to the time difference between the *stop* and *start* pulses. Different scale settings allow adjustment of both the pulse height and time range. Typical units also include *true start* and *busy* outputs that allow monitoring of the input *start* rate and the interval during which the unit is busy processing input signal and therefore unavailable for accepting new signals.

The output of the TAC is normally connected to the input of a pulse height analyser (PHA), or ADC and a multichannel analyser (MCA). The PHA/ADC assigns a binary number to the height of the input pulse; 0 for a zero amplitude pulse and 2^n for the maximum amplitude pulse, where n is an integer that can be selected according to the application. The binary number is used as the address for a multichannel analyser with 2^n address locations in the MCA. For each address from the PHA/ADC, a one is added to the contents of the address location. If, for example, a TAC has a range of 0 to 4 V output amplitude for *stop/start* time differences of 200 ns and the PHA/ADC assigns binary 0 to the 0 V amplitude signal and binary 255 to the 4 V amplitude signal, each of the 256 channels will correspond to 0.78 ns. With time, a histogram of flight times is built up in the memory of the MCA. [Figure B1.10.4](#) is an illustration of a time-of-flight spectrum for a sample of air that has been ionized and accelerated to 2000 eV.

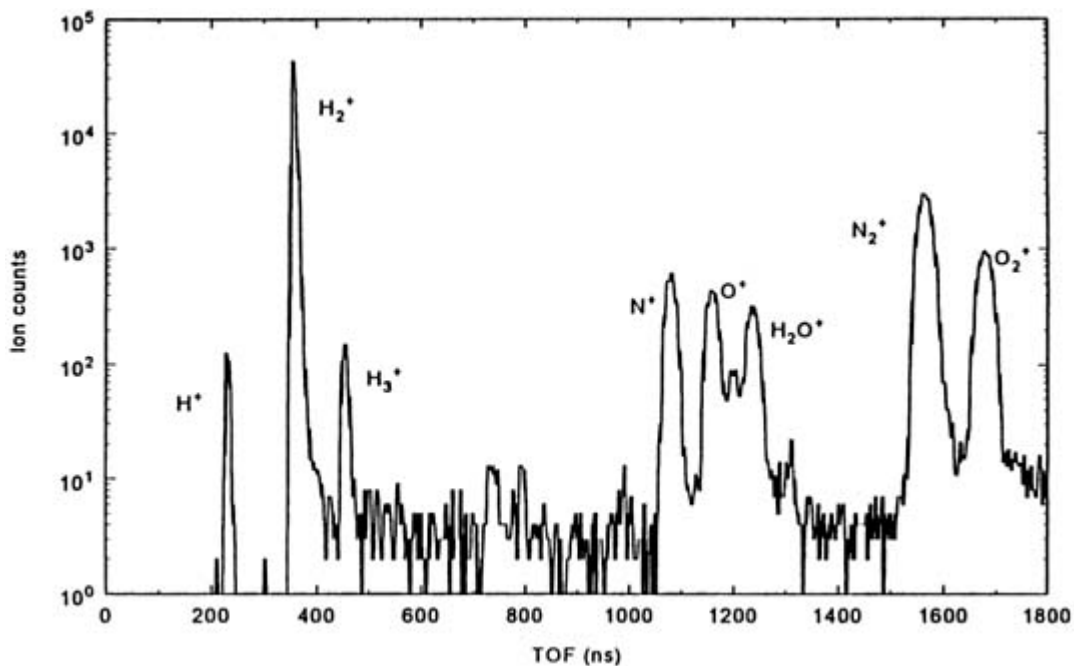


Figure B1.10.4. Time-of-flight histogram for ions resulting from the ionization of a sample of air with added hydrogen. The ions have all been accelerated to the same energy (2 keV) so that their time of flight is directly proportional to the reciprocal of the square root of their mass.

An alternative to the TAC and PHA/ADC is the time-to-digital converter (TDC), a unit that combines the functions of the TAC and PHA/ADC. There are *start* and *stop* inputs and an output that provides a binary number directly proportional to the time difference between the stop and start signals. The TAC can be directly connected to a MCA or PC with the appropriate digital interface.

B1.10.3.2 ELECTRONICS LIMITATIONS

By storing the binary outputs of a PHA/ADC or a TDC in the form of a histogram in the memory of a MCA, computer analyses can be performed on the data that take full account of the limitations of the individual components of the measuring system. For the preamplifiers and discriminators, time resolution is the principal consideration. For the TAC, resolution and processing time can be critical. The PHA/ADC or a TDC provide the link between the analogue circuits of the preamplifiers and discriminators and the digital input of the MCA or PC. The conversion from analogue-to-digital by a PHA/ADC or TDC is not perfectly linear and the deviations from linearity are expressed in terms of differential and integral nonlinearities. Differential nonlinearity is the variation of input signal amplitude over the width of a single time channel. Integral nonlinearity is the maximum deviation of the measured time channel number from a least squares straight line fit to a plot of signal amplitude as a function of channel number.

B1.10.4 LIFETIME MEASUREMENTS

B1.10.4.1 GENERAL CONSIDERATIONS

Lifetime measurements have elements in common with both counting and time-of-flight experiments [4, 5]. In a lifetime experiment there is an initiating event that produces the system that subsequently decays with the emission of radiation, particles or both. Decay is statistical in character; taking as an example nuclear decay,

at any time t , each nucleus in a sample of n nuclei has the same probability of decay in the interval dt . Those nuclei that remain then have the same probability of decaying in the next interval dt . The rate of decay is given by $dN/dt = -kN$. The constant k , with units $1/\text{time}$, is called the *lifetime* and depends on the system and the nature of the decay. Integration of the first-order differential equation gives the exponential decay law, $N(t) = N_0 e^{-t/\tau}$, where N_0 is the number of systems (atoms, molecules, nuclei) initially created and $N(t)$ is the number that remain after a time t . The constant $\tau = 1/k$ can be obtained by measuring the time for the sample to decay to $1/e$ of the initial size.

A more direct method for lifetime measurements is the delayed coincidence technique [6] in which the time between an initiation event and the emission of a decay product is measured. A schematic diagram of an apparatus used for the measurement of atomic lifetimes is shown in figure B1.10.5. The slope of the graph of the natural log of the number of decay events as a function of time delay gives the lifetime directly. The precision with which the slope can be determined increases with the number of measurements. With 10^5 separate time determinations, times to 10τ can be sampled providing a range sufficient for a determination of τ to a few per cent. Enough time must be allowed between each individual measurement to allow for long decay times. This requires that the experimental conditions be adjusted so that on average one decay event is recorded for every 100 initiation events. The delayed coincidence method can routinely measure lifetimes from a few nanoseconds to microseconds. The lower limit is set by the excitation source and the time resolution of the detector and electronics. Lifetimes as short as 10 ps can be measured with picosecond pulsed laser excitation, microchannel plate photomultipliers, a GHz preamplifier and a fast timing discriminator. Instrument stability and available time set the upper limit for lifetime measurements.

-9-

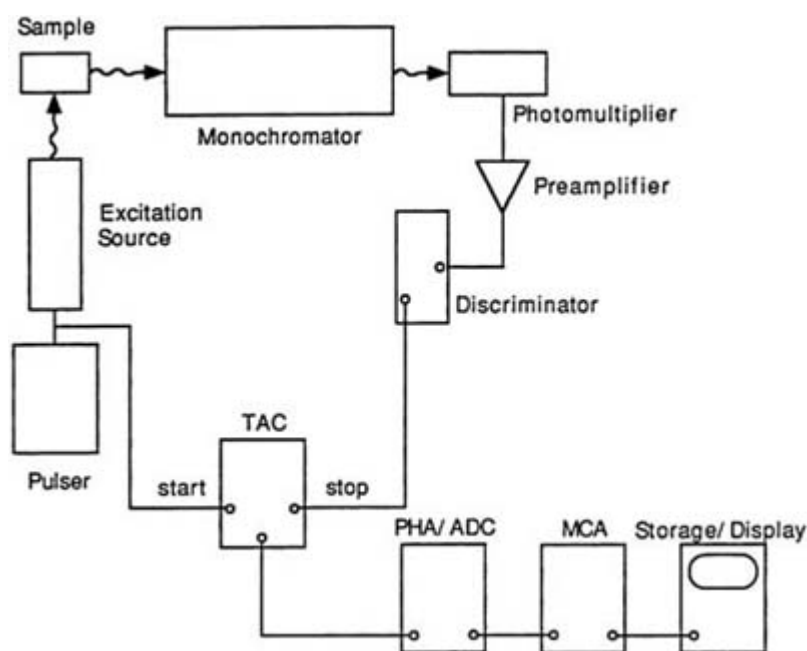


Figure B1.10.5. Lifetime experiment. A pulser triggers an excitation source that produces excited species in the sample. At the same time the pulser provides the *start* signal for a time-to-amplitude converter (TAC). Radiation from the decay of an excited species is detected with a photomultiplier at the output of a monochromator. The signal from the photomultiplier is amplified and sent to a discriminator. The output of the discriminator is the *stop* signal for the TAC. The TAC output pulse amplitude is converted to a binary number by a pulse height analyser (PHA) or analogue-to-digital converter (ADC). The binary number serves as an address for the multichannel analyser (MCA) that adds one to the number stored at the specified address. The pulser rate must accommodate decay times at least a factor of 10 longer than the lifetime of the specie under study. The excitation source strength and sample density are adjusted to have at most one detected decay event per pulse.

The delayed coincidence method has been applied to the fluorescent decay of laser excited states of biological molecules. By dispersing the emitted radiation from the decaying molecules with a polychromator and using an array of photodetectors for each wavelength region it is possible to measure fluorescent lifetimes as a function of the wavelength of the emitted radiation. The information is used to infer the conformation of the excited molecules [7].

B1.10.4.2 MULTIPLE HIT TIME-TO-DIGITAL CONVERSION

Both lifetime and time-of-flight measurements have low duty cycles. In the case of the lifetime measurements sufficient time between initiation events must be allowed to accommodate the detection of long decays. Moreover, the signal rate has to be adjusted to allow for the detection of at most one decay event for every initiation event. In the case of time-of-flight measurements enough time must be allowed between measurements to allow for the slowest particle to traverse the flight path. As with the case of lifetime measurements, each initiation event must give rise to no more than one particle in the flight path.

-10-

In order to circumvent the signal limitations of lifetime and time-of-flight measurements multiple hit time-to-digital converters can be used. Typically, these instruments have from eight to sixteen channels and can be used in the ‘common start’ or ‘common stop’ mode. When used in the common start mode a single start signal arms all of the channels with successive ‘stop’ signals directed to each of the channel inputs in sequence. In this way the signal rate can be increased by a factor equal to the number of channels in the TDC. A counter and demultiplexer/data selector [8] perform the function of routing the stop signals to the different channel inputs in sequence. A schematic diagram of the arrangement is shown in figure B1.10.6.

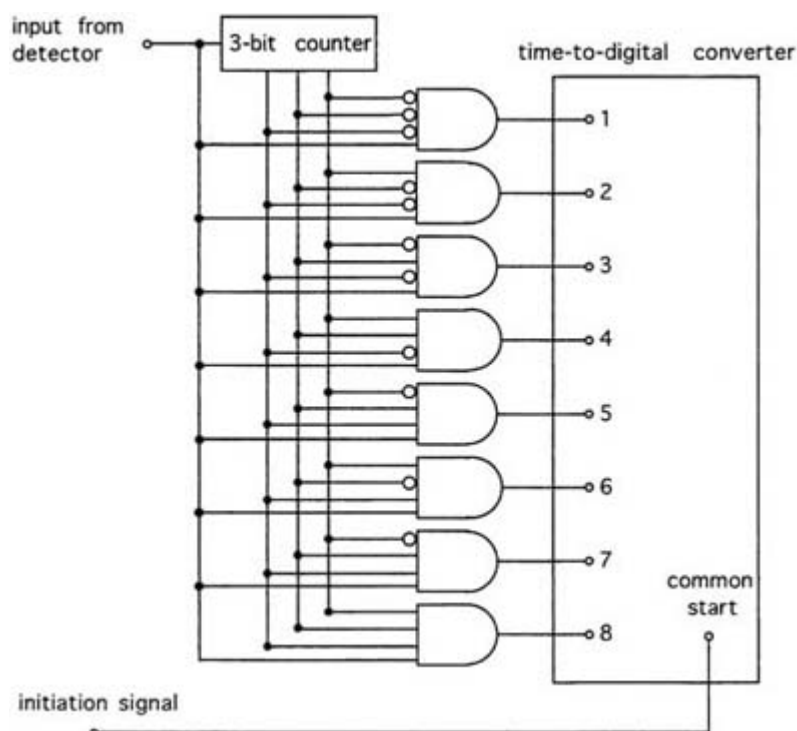


Figure B1.10.6. Multiple hit time-to-digital conversion scheme. Signals from the detector are routed to eight different TDC channels by a demultiplexer/data selector according to the digital output of the 3-bit counter that tracks the number of detector output signals following the initiation signal. The initiation signal is used as the common start. Not shown are the delays and signal conditioning components that are necessary to ensure the correct timing between the output of the counter and the arrival of the pulses on the common data line. Control logic to provide for the counter to reset after eight detector signals is also required.

B1.10.5 COINCIDENCE EXPERIMENTS

Coincidence experiments explicitly require knowledge of the time correlation between two events. Consider the example of electron impact ionization of an atom, [figure B1.10.7](#). A single incident electron strikes a target atom or molecule and ejects an electron from it. The incident electron is deflected by the collision and is identified as the scattered electron. Since the scattered and ejected electrons arise from the same event, there is a time correlation

-11-

between their arrival times at the detectors.

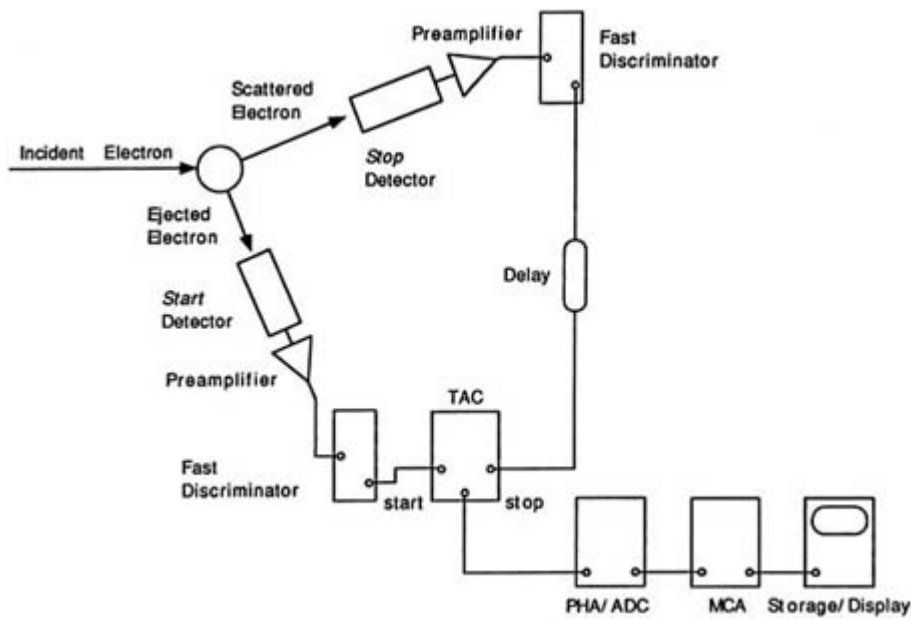


Figure B1.10.7. Electron impact ionization coincidence experiment. The experiment consists of a source of incident electrons, a target gas sample and two electron detectors, one for the scattered electron, the other for the ejected electron. The detectors are connected through preamplifiers to the inputs (start and stop) of a time-to-amplitude converter (TAC). The output of the TAC goes to a pulse-height-analyser (PHA) and then to a multichannel analyser (MCA) or computer.

Coincidence experiments have been common in nuclear physics since the 1930s. The widely used coincidence circuit of Rossi [9] allowed experimenters to determine, within the resolution time of the electronics of the day, whether two events were coincident in time. The early circuits were capable of submicrosecond resolution, but lacked the flexibility of today's equipment. The most important distinction between modern coincidence methods and those of the earlier days is the availability of semiconductor memories that allow one to now record precisely the time relations between all particles detected in an experiment. We shall see the importance of this in the evaluation of the statistical uncertainty of the results.

In a two detector coincidence experiment, of which figure B1.10.7 is an example, pulses from the two detectors are amplified and then sent to discriminators, the outputs of which are standard rectangular pulses of constant amplitude and duration. The outputs from the two discriminators are then sent to the *start* and *stop* inputs of a TAC or TDC. Even though a single event is responsible for ejected and scattered electrons, the two electrons will not arrive at the detectors at identical times because of differences in path lengths and electron velocities. Electronic propagation delays and cable delays also contribute to the *start* and *stop* signals not arriving at the inputs of the TAC or TDC at identical times; sometimes the *start* signal arrives first, sometimes

the *stop* signal is first. If the signal to the *start* input arrives after the signal to the *stop*, that pair of events will not result in a TAC/TDC output. The result can be a 50% reduction in the number of recorded coincidences. To overcome this limitation a time delay is inserted in the *stop* line between the discriminator and the TAC/TDC. This delay can be a length of coaxial cable (typical delays are of the order of 1 ns/ft) or an electronic circuit. The purpose is to ensure that the stop signal always arrives at the *stop* input of the TAC/TDC after the *start* signal. A perfect coincidence will be recorded at a time difference approximately equal to the delay time.

-12-

When a time window twice the duration of the delay time is used, perfect coincidence is at the centre of the time window and it is possible to make an accurate assessment of the background by considering the region to either side of the perfect coincidence region. An example of a time spectrum is shown in figure B1.10.8.

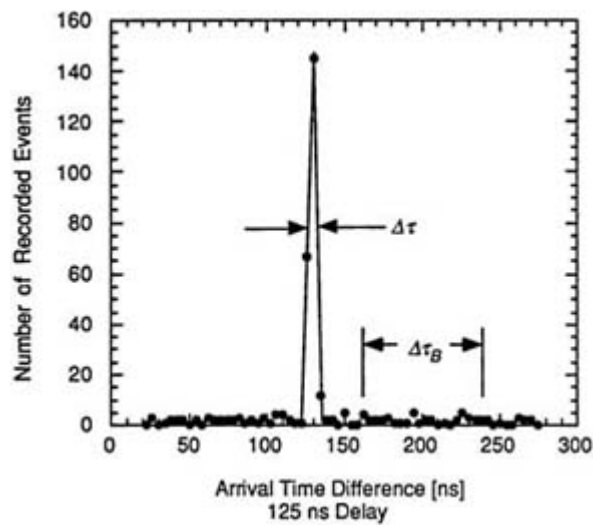


Figure B1.10.8. Time spectrum from a double coincidence experiment. Through the use of a delay in the lines of one of the detectors, signals that occur at the same instant in both detectors are shifted to the middle of the time spectrum. Note the uniform background upon which the true coincidence signal is superimposed. In order to decrease the statistical uncertainty in the determination of the true coincidence rate, the background is sampled over a time $\Delta\tau_B$ that is much larger than the width of the true coincidence signal, $\Delta\tau$.

B1.10.5.1 SIGNAL AND BACKGROUND

Referring to [figure B1.10.7](#) consider electrons from the event under study as well as from other events all arriving at the two detectors. The electrons from the event under study are correlated in time and result in a peak in the time spectrum centred approximately at the delay time. There is also a background level due to events that bear no fixed time relation to each other. If the average rate of the background events in each detector is R_1 and R_2 , then the rate that two such events will be recorded within time $\Delta\tau$ is given by R_B , where

$$R_B = R_1 R_2 \Delta\tau.$$

Let the rate of the event under study be R_A . It will be proportional to the cross section for the process under study, σ_A , the incident electron current, I_0 , the target density, n , the length of the target viewed by the detectors, ℓ , the solid angles subtended by the detectors, $\Delta\omega_1$ and $\Delta\omega_2$ the efficiency of the detectors, ε_1 and ε_2 .

$$R_A = \sigma_A I_0 n \ell \Delta\omega_1 \Delta\omega_2 \varepsilon_1 \varepsilon_2.$$

The product $I_0 n^{\ell}$, depends on the properties of the region from which the two electrons originate and is called the

-13-

source function, S . The properties of the detectors are described by the product, $\Delta\omega_1 \cdot \Delta\omega_2 \varepsilon_1 \varepsilon_2$.

For the background, each of the rates, R_1 and R_2 , will be proportional to the source function, the cross sections for single electron production and the properties of the individual detectors,

Combining the two expressions for R_1 and R_2

Comparing the expressions for the background rate and the signal rate one sees that the background increases as the square of the source function while the signal rate is proportional to the source function. The signal-to-background rate, R_{AB} , is then

It is important to note that the signal is always accompanied by background. We now consider the signal and background after accumulating counts over a time T . For this it is informative to refer to [figure B1.10.8](#) the time spectrum. The total number of counts within an arrival time difference $\Delta\tau$ is N_T and this number is the sum of the signal counts, $N_A = R_A T$, and the background counts, $N_B = R_B T$,

The determination of the background counts must come from an independent measurement, typically in a region of the time spectrum outside of the signal region, yet representative of the background within the signal region. The uncertainty in the determination of the signal counts is given by the square root of the uncertainties in the total counts and the background counts

The essential quantity is the relative uncertainty in the signal counts, $\delta N_A / N_A$. This is given by

-14-

Expressing R_A in terms of R_{AB} results in the following formula

$$\delta N_A / N_A = \sqrt{\frac{(R_{AB} + 2) \Delta\tau}{\frac{\sigma_A^2}{\sigma_1 \sigma_2} \Delta\omega_1 \Delta\omega_2 \varepsilon_1 \varepsilon_2 T}}$$

There are a number of observations to be drawn from the above formula: the relative uncertainty can be reduced to an arbitrarily small value by increasing T , but because the relative uncertainty is proportional to $1/\sqrt{T}$, a reduction in relative uncertainty by a factor of two requires a factor of four increase in collection time. The relative uncertainty can also be reduced by reducing $\Delta\tau$. Here, it is understood that $\Delta\tau$ is the smallest time window that just includes all of the signal. $\Delta\tau$ can be decreased by using the fastest possible detectors, preamplifiers and discriminators and minimizing time dispersion in the section of the experiment ahead of the detectors.

The signal and background rates are not independent, but are coupled through the source function, S , as a consequence the relative uncertainty in the signal decreases with the signal-to-background rate, $R_{AB} = 1$, a somewhat unanticipated result. Dividing $\delta N_A / N_A$ by its value at $R_{AB} = 1$ gives a reduced relative uncertainty $[\delta N_A / N_A]_R$ equal to $\sqrt{(R_{AB} + 2)/3}$. A plot of $[\delta N_A / N_A]_R$ as a function of R_{AB} is shown in [figure B1.9](#).

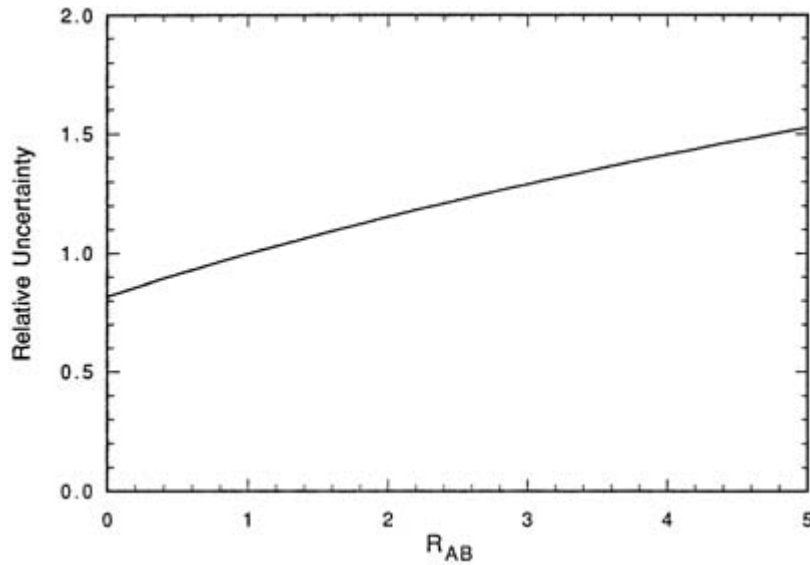


Figure B1.10.9. Plot of the reduced relative uncertainty of a double coincidence experiment as a function of the signal-to-background ratio. Note that the relative uncertainty decreases as the signal-to-background rate decreases.

To illustrate this result, consider a case where there is one signal count and one background count in the time window $\Delta\tau$. The signal-to-background ratio is 1 and the relative uncertainty in the signal is $\sqrt{2+1}/1 = 1.7$. By increasing the source strength by a factor of 10 the signal will be increased by a factor of 10 and the background by a factor of 100. The signal-to-background ratio is now 0.1, but the relative uncertainty in the signal is $\sqrt{110+100}/10 = 1.45$, a clear improvement over the larger signal-to-background case.

-15-

Another method for reducing the relative uncertainty is to increase the precision of the measurement of background. The above formulae are based on an independent measurement of background over a time window, $\Delta\tau$, that is equivalent to the time window within which the signal appears. If a larger time window for the background is used, the uncertainty in background determination can be correspondingly reduced. Let a time window of width $\Delta\tau_B$ be used for the determination of background, where $\Delta\tau / \Delta\tau_B = \rho < 1$. If the rate at which counts are accumulated in time window $\Delta\tau_B$ is R_{BB} , the background counts to be subtracted from the total counts in time window $\Delta\tau$ becomes $\rho R_{BB} T = \rho N_{BB}$. The uncertainty in the number of background counts to be subtracted from the total of signal plus background is $\sqrt{\rho^2 N_{BB}}$. The relative uncertainty of the signal is then

$$\sqrt{\frac{1 + 2\rho^2 N_{BB}/N_A}{N_A}} = \sqrt{\frac{1 + 2\rho N_B/N_A}{N_A}}$$

The expression for the reduced relative uncertainty is then

$$[\delta N_A / N_{AB}]_R = \sqrt{\frac{(R_{AB} + 2\rho)}{3}}$$

For coincidence experiments where the detectors record time correlated events, maximum efficiency and sensitivity is attained when the detectors accept particles originating from the same source volume, in other words, the length, ℓ , that defines the common volume of the source region seen by the detectors should be the same as the lengths and corresponding volumes for each of the detectors separately. If the two volumes are different, only the common volume seen by the detectors is used in the calculation of coincidence rate. Events that take place outside the common volume, but within the volume accepted by one of the detectors will only contribute to the background. This is shown schematically in [figure B1.10.10](#). The situation is potentially complex if the source region is not uniform over the common volume viewed by the detectors and if the efficiencies of the detectors varies over the volume they subtend. Full knowledge of the geometric acceptances of the detectors and the source volume is necessary to accurately evaluate the size and strength of the source. Such an analysis is typically done numerically using empirical values of the view angles of the detectors and the spatial variation of target density and incident beam intensity.

-16-

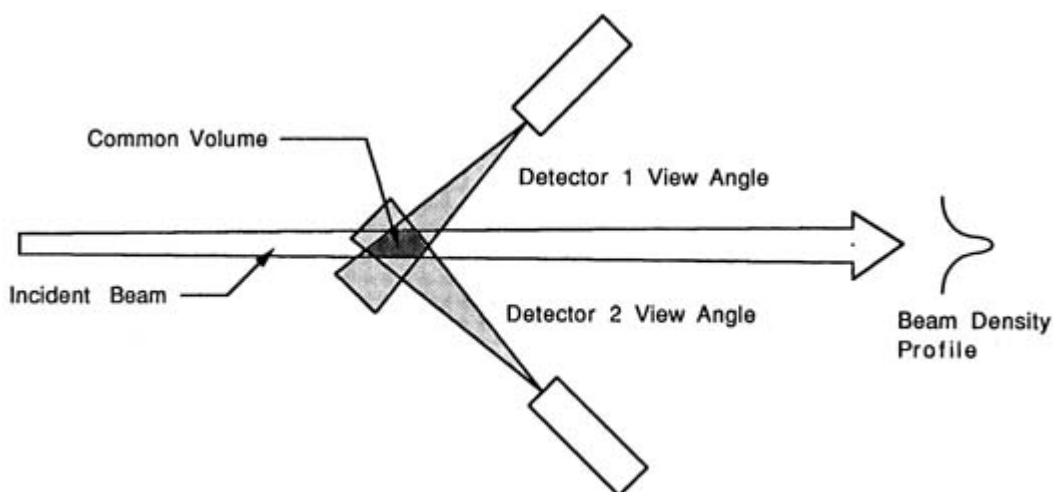


Figure B1.10.10. Schematic diagram of the effect of detector view angles on coincidence rate. The view angles of two detectors are shown along with the common view angle. Maximum signal collection efficiency is achieved when the individual view angles have the maximum overlap and when the overlap coincides with the maximum density of the incident beam.

B1.10.5.3 MULTIPLE PARAMETER MEASUREMENTS

It is often the case that a time correlation measurement alone is not sufficient to identify a particular event of interest. For this reason, coincidence measurements are often accompanied by the measurement of another parameter such as energy, spin, polarization, wavelength, etc. For the case of the electron impact ionization experiment of [figure B1.10.7](#) electrostatic energy analysers can be placed ahead of the detectors so that only electrons of a preselected energy arrive at the detectors. In this case the coincidence rate must take into account the energy bandpass of the analysers and the fact that in this example the energies of the scattered and ejected electrons are not independent, but are coupled through the relation that the energy of the incident electron equals the sum of the energies of the scattered and ejected electrons and the binding energy of the ejected electron

$$E_0 = E_1 + E_2 + BE,$$

where E_0 is the energy of the incident electron, E_1 is the energy of the scattered electron, E_2 is the energy of the ejected electron and BE is the binding energy of the ejected electron. The coincidence rate now has an additional term, ΔE , that is the overlap of the energy bandwidths of the two detectors, ΔE_1 and ΔE_2 , subject to

the energy conservation constraint. An often used approximation to ΔE is $\Delta E = \sqrt{\Delta E_1^2 + \Delta E_2^2}$. The background rates in the detectors depend only on the individual energy bandwidths, and are independent of any additional constraints. Maximum efficiency is achieved when the energy bandwidths are only just wide enough to accept the coincident events of interest.

B1.10.5.4 MULTIPLE DETECTORS

The arrangement for a single coincidence measurement can be expanded to a multiple detector arrangement. If, for example, it is necessary to measure coincidence rates over an angular range, detectors can be placed at the angles of

-17-

interest and coincidence rates measured between all pairs of detectors. This is a much more efficient way of collecting data than having two detectors that are moved to different angles after each coincidence determination. Depending on the signal and background rates, detector outputs can be multiplexed and a detector identification circuit used to identify those detectors responsible for each coincidence signal. If detectors are multiplexed, it is well to remember that the overall count rate is the sum of the rates for all of the detectors. This can be an important consideration when rates become comparable to the reciprocal of system dead time.

B1.10.5.5 PREPROCESSING

It is often the case that signal rates are limited by the processing electronics. Consider a coincidence time spectrum of 100 channels covering 100 ns with a coincidence time window of 10 ns. Assume a signal rate of 1 Hz and a background rate of 1 Hz within the 10 ns window. This background rate implies an uncorrelated event rate of 10 kHz in each of the detectors. To register 99% of the incoming events, the dead time of the system can be no larger than 10 μ s. The dead time limitation can be substantially reduced with a preprocessing circuit that only accepts events falling within 100 ns of each other (the width of the time spectrum). One way to accomplish this is with a circuit that incorporates delays and gates to only pass signals that fall within a 100 ns time window. With this circuit the number of background events that need to be processed each second is reduced from 10,000 to 10, and dead times as long as 10 ms can be accommodated while maintaining a collection efficiency of 99%. In a way this is an extension of the multiparameter processing in which the parameter is the time difference between processed events. A preprocessing circuit is discussed by Goruganthu et al [10].

B1.10.5.6 TRIPLE COINCIDENCE MEASUREMENTS

A logical extension of the coincidence measurements described above is the triple coincidence measurement shown schematically in [figure B1.10.11](#). Taking as an example electron impact double ionization, the two ejected electrons and the scattered electron are correlated in time. On a three-dimensional graph with vertical axes representing the number of detected events and the horizontal axes representing the time differences between events at detectors 1 and 2 and 1 and 3, a time correlated signal is represented as a three-dimensional peak with a fixed base width in the horizontal plane. A triple coincidence time spectrum is shown in [figure B1.10.12](#). Unlike the situation for double coincidence measurements the background has four sources, random rates in each of the three detectors, correlated events in detectors 1 and 2 with an uncorrelated signal in 3, correlated events in detectors 1 and 3 with an uncorrelated signal in 2 and correlated events in detectors 2 and 3 with an uncorrelated signal in detector 1. The first source gives a background that is uniform over the full horizontal plane. The three other sources produce two walls that are parallel to the two time axes and a third wall that lies at 45° to the time axes. These can be seen in [figure B1.10.12](#).

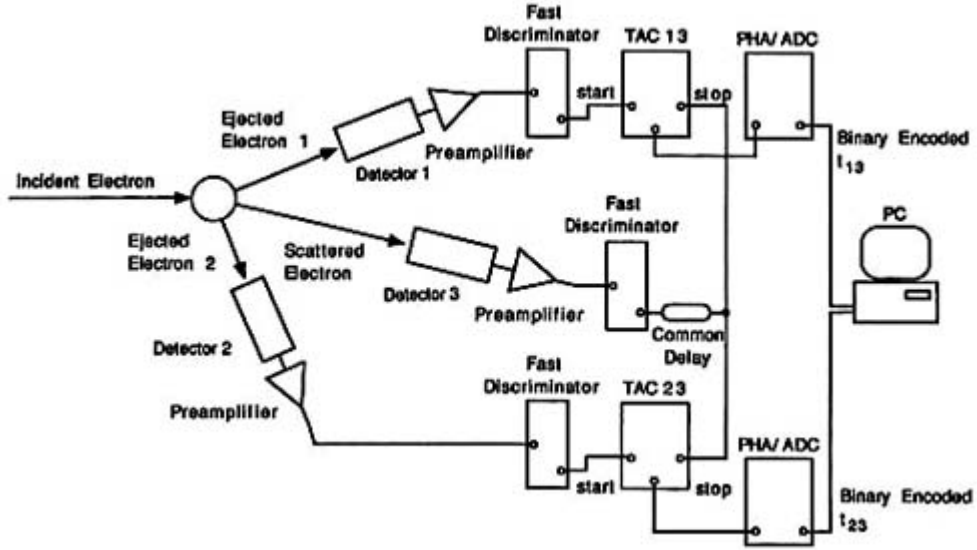


Figure B1.10.11. Electron impact double ionization triple coincidence experiment. Shown are the source of electrons, target gas, three electron detectors, one for the scattered electron and one for each of the ejected electrons. Two time differences, t_{13} and t_{23} , are recorded for each triple coincidence. t_{13} is the difference in arrival times of ejected electron 1 and the scattered electron; t_{23} is the difference in arrival times of ejected electron 2 and the scattered electron. Two sets of time-to-amplitude converters (TACs) and pulse height analysers/analogue-to-digital converters (PHA/ADC) convert the times to binary encoded numbers that are stored in the memory of a computer. The data can be displayed in the form of a two-dimensional histogram (see figure B1.10.12).

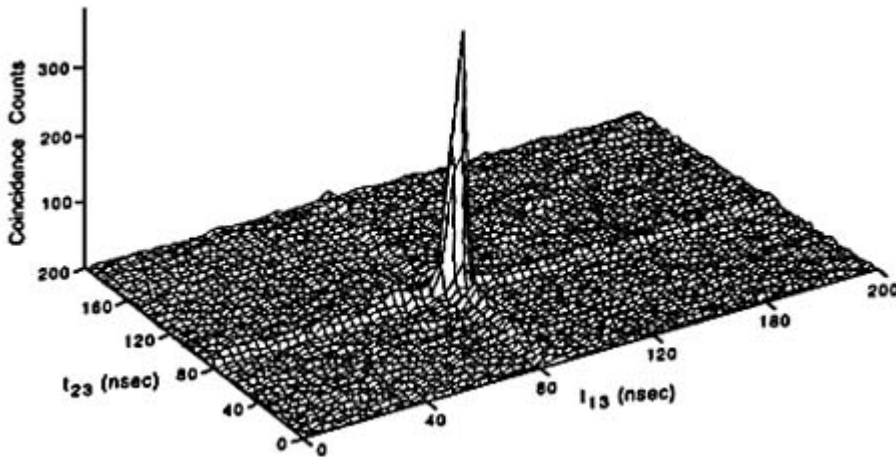


Figure B1.10.12. Schematic diagram of a two-dimensional histogram resulting from the triple coincidence experiment shown in figure B1.10.10. True triple coincidences are superimposed on a uniform background and three walls corresponding to two electron correlated events with a randomly occurring third electron.

For any region of the horizontal plane with dimensions $\Delta\tau_{12}$ $\Delta\tau_{13}$, the uniform background rate is

$$R_{123} = R_1 R_2 R_3 \Delta\tau_{12} \Delta\tau_{13} = [S\sigma_1 \Delta\omega_1 \varepsilon_1][S\sigma_2 \Delta\omega_2 \varepsilon_2][S\sigma_3 \Delta\omega_3 \varepsilon_3] \Delta\tau_{12} \Delta\tau_{13}$$

where R_1 , R_2 and R_3 are the random background rates in detectors 1, 2 and 3. The background due to two time correlated events and a single random event is $R_{12} + R_{13} + R_{23}$, where

$$R_{12} = [S\sigma_{12}\Delta\omega_1\Delta\omega_2\varepsilon_1\varepsilon_2][S\sigma_3\Delta\omega_3\varepsilon_3]\Delta\tau_{13}$$

$$R_{13} = [S\sigma_{13}\Delta\omega_1\Delta\omega_3\varepsilon_1\varepsilon_3][S\sigma_2\Delta\omega_2\varepsilon_2]\Delta\tau_{12}$$

$$R_{23} = [S\sigma_{23}\Delta\omega_2\Delta\omega_3\varepsilon_2\varepsilon_3][S\sigma_1\Delta\omega_1\varepsilon_1]\Delta\tau_{23}.$$

The signal rate R_A , is

$$R_A = S\sigma_A\Delta\omega_1\Delta\omega_2\Delta\omega_3\varepsilon_1\varepsilon_2\varepsilon_3$$

where the symbols have the same meaning as in the treatment of double coincidences. If the signal falls within a two-dimensional time window $\Delta\tau_{12}\Delta\tau_{13}$, then the signal-to-background rate ratio is $R_A/[R_{123} + R_{12} + R_{13} + R_{23}] = \chi$, and the statistical uncertainty in the number of signal counts accumulated in time T is

$$\delta N_A/N_A = \sqrt{\frac{1 + 2/\chi}{K_A S T}}$$

where $K_A = \sigma_A\Delta\omega_1\Delta\omega_2\Delta\omega_3\varepsilon_1\varepsilon_2\varepsilon_3$. In contrast to simpler double coincidence experiments, S , the source function is not directly proportional to $1/\rho$. The full expression for $\delta N_A/N_A$ is

$$\delta N_A/N_A = \sqrt{\frac{1 + 2/\chi}{\frac{K_A}{2} \left[-\frac{K_1}{K_2} + \sqrt{\left(\frac{K_1}{K_2}\right)^2 + 4\left(\frac{K_A}{K_2}\right)\frac{1}{\chi}} \right] T}}$$

where

$$K_1 = \sigma_1\Delta\omega_1\varepsilon_1\sigma_2\Delta\omega_2\varepsilon_2\sigma_3\Delta\omega_3\varepsilon_3\Delta\tau_{12}\Delta\tau_{13}$$

and

$$K_2 = \sigma_{12}\Delta\omega_1\Delta\omega_2\varepsilon_1\varepsilon_2\sigma_3\Delta\omega_3\varepsilon_3\Delta\tau_{13} + \sigma_{13}\Delta\omega_1\Delta\omega_3\varepsilon_1\varepsilon_3\sigma_2\Delta\omega_2\varepsilon_2\Delta\tau_{12} + \sigma_{23}\Delta\omega_2\Delta\omega_3\varepsilon_2\varepsilon_3\sigma_1\Delta\omega_1\varepsilon_1\Delta\tau_{23}.$$

-20-

As is the case for the double coincidence arrangement $\delta N_A/N_A$ is inversely proportional to the square root of the acquisition time, however $\delta N_A/N_A$ does not approach a minimum value in the limit of $\chi = 0$, but rather has a minimum in the region between $\chi = 0$ and $\chi = \infty$, the precise value of χ depending on the experimental conditions and the magnitudes of the cross sections for the different electron producing events. A detailed treatment of this topic with examples is given by Dupré et al [11].

B1.10.6 ANTI-COINCIDENCE

In high-energy physics experiments there can be many interfering events superimposed on the events of interest. An example is the detection of gamma rays in the presence of high-energy electrons and protons. The

electrons, protons and gamma rays all produce very similar signals in the solid state detectors that are used, and it is not possible to distinguish the gamma rays from the charged particles. A technique that is frequently used is to surround the gamma ray detectors with a plastic scintillation shield that produces a flash of light when traversed by a charged particle, but is transparent to the gamma rays. Light from the shield is coupled to photomultiplier detectors via light pipes. Signals that occur simultaneously in the photomultipliers and solid state detector are due to high-energy charged particles entering the instrument and are excluded from any analysis. An example of such an anti-coincidence circuit can be found in the energetic gamma ray experiment telescope (EGRET) on the gamma ray observatory (GRO) space craft that was launched in 1991 and continues to provide information on the energies and sources of gamma rays in the 20 MeV to 30 GeV energy range. Figure B1.10.13 is a schematic diagram of the EGRET experiment showing the anticoincidence shield.

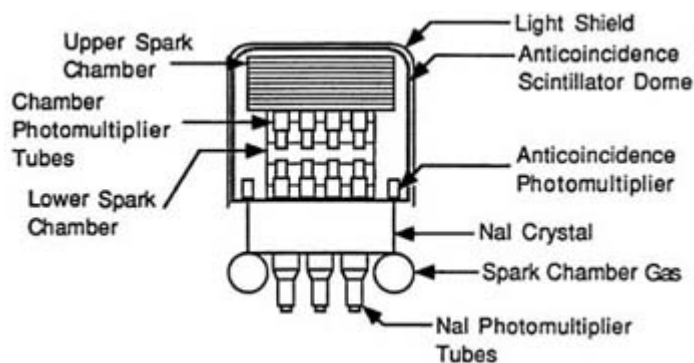


Figure B1.10.13. EGRET experiment showing spark chambers and photomultiplier tubes for the detection of gamma rays. The gamma ray detectors are surrounded by an anticoincidence scintillator dome that produces radiation when traversed by a high-energy charged particle but is transparent to gamma rays. Light pipes transmit radiation from the dome to photomultiplier tubes. A signal in a anticoincidence photomultiplier tube causes any corresponding signal in the gamma ray detectors to be ignored. The NaI crystal detector and photomultiplier tubes at the bottom of the unit provide high-resolution energy analysis of gamma rays passing through the spark chambers.

REFERENCES

- [1] Bothe W and Geiger H 1925 Über das Wesen des Comptoneffekts; ein experimenteller Beitrag zur Theorie der Strahlung *Z. Phys.* **32** 639
- [2] Bevington P R 1969 *Data Reduction and Error Analysis for the Physical Sciences* (New York: McGraw Hill) pp 36–43
- [3] Melissinos A C 1966 *Experiments in Modern Physics* (New York: Academic) ch 10
- [4] Demas J N 1983 *Excited State Lifetime Measurements* (New York: Academic)
- [5] Lakowicz J R 1983 *Principles of Fluorescence Spectroscopy* (New York: Plenum) ch 3
- [6] Klose J Z 1966 Atomic lifetimes in neon I *Phys. Rev.* **141** 181
- [7] Knutson J R 1988 Time-resolved laser spectroscopy in biochemistry *SPIE* **909** 51–60
- [8] Millman J and Grabel A 1987 *Microelectronics* 2nd edn (New York: McGraw Hill) pp 279–84
- [9] Rossi B 1930 *Nature* **125** 636
- [10] Goruganthu R R, Coplan M A, Moore J H and Tossell J A 1988 (e,2e) momentum spectroscopic study of the interaction of $-\text{CH}_3$ and $-\text{CF}_3$ groups with the carbon-carbon triple bond *J. Chem. Phys.* **89** 25
- [11] Dupré C, Lahmam-Bennani A and Duguet A 1991 About some experimental aspects of double and triple coincidence techniques to study electron impact double ionizing processes *Meas. Sci. Technol.* **2** 327

B1.11 NMR of liquids

Oliver W Howarth

B1.11.1 INTRODUCTION

Nuclear magnetic resonance (NMR) was discovered by Bloch, Purcell and Pound in 1945, as a development of the use of nuclear spins in low-temperature physics. Its initial use was for the accurate measurement of nuclear magnetic moments. However, increases in instrumental precision led to the detection of chemical shifts (B1.11.5) and then of spin–spin couplings (B1.11.6). This stimulated use by chemists. There have been spectacular improvements in sensitivity, resolution and computer control since then, so that NMR equipment is now essential in any laboratory for synthetic chemistry. Within moments or hours, it can determine the structure and, if desired, conformation of most medium-sized molecules in the solution phase. For this reason, a large pharmaceutical company will typically generate several hundred NMR spectra in one working day. NMR is also widely used in biochemistry for the much more challenging problem of determining the structures of smaller proteins and other biomolecules. The rates and extents of molecular motions can also be measured, through measuring the rates of energy transfer to, from and between nuclei (relaxation, B1.13). Outside of chemistry, it is used in a different mode for medical and other imaging, and for the detection of flow and diffusion in liquids (B1.14). It can also be used for clinical and *in vivo* studies, as the energies involved present no physiological dangers.

B1.11.2 NUCLEAR SPINS

NMR depends on manipulating the collective motions of nuclear spins, held in a magnetic field. As with every rotatable body in nature, every nucleus has a spin quantum number I . If $I = 0$ (e.g. ^{12}C , ^{16}O) then the nucleus is magnetically inactive, and hence ‘invisible’. If, as with most nuclei, $I > 0$, then the nucleus must possess a magnetic moment, because of its charge in combination with its angular momentum $\hbar\sqrt{I(I+1)}$. This makes it detectable by NMR. The most easily detected nuclei are those with $I = \frac{1}{2}$ and with large magnetic moments, e.g. ^1H , ^{13}C , ^{19}F , ^{31}P . These have two allowed states in a magnetic (induction) field B_0 : $m_{1/2}$ and $m_{-1/2}$, with angular momentum components $\pm\hbar/2$. Thus these nuclear magnets lie at angles $\pm \cos^{-1}((\frac{1}{2})/(\frac{3}{4}))^{1/2} = 54.7^\circ$ to B_0 . They also precess rapidly, like all gyroscopes, at a rate ν_L . This is called the Larmor frequency, after its discoverer. The $2I + 1$ permitted angles for other values of I differ from the above angles, but they can never be zero or 180° , because of the uncertainty principle.

The energy difference ΔE corresponding to the permitted transitions, $\Delta m = \pm 1$, is given by

$$\Delta E = \gamma \hbar B_0 / 2\pi = h\nu_L.$$

Therefore, in NMR, one observes collective nuclear spin motions at the Larmor frequency. Thus the frequency of NMR detection is proportional to B_0 . Nuclear magnetic moments are commonly measured either by their magnetogyric ratio γ , or simply by their Larmor frequency ν_L in a field where ^1H resonates at 100 MHz (symbol Ξ).

B1.11.2.1 PRACTICABLE ISOTOPES

Almost all the stable elements have at least one isotope that can be observed by NMR. However, in some cases the available sensitivity may be inadequate, especially if the compounds are not very soluble. This may be because the relevant isotope has a low natural abundance, e.g. ^{17}O , or because its resonances are extremely broad, e.g. ^{33}S . Tables of such nuclear properties are readily available [1], and isotopic enrichment may be available as an expensive possibility. Broad resonances are common for nuclei with $I > 0$, because these nuclei are necessarily non-spherical, and thus have electric quadrupole moments that interact strongly with the electric field gradients present at the nuclei in most molecules. Linewidths may also be greatly increased by chemical exchange processes at appropriate rates (B2.7), by the presence of significant concentrations of paramagnetic species and by very slow molecular tumbling, as in polymers and colloids. For nuclei with $I = 1/2$, a major underlying cause of such broadening is the magnetic dipolar interaction of the nucleus under study with nearby spins. This and other interactions also lead to very large linewidths in the NMR spectra of solids and of near-solid samples, such as gels, pastes or the bead-attached molecules used in combinatorial chemistry. However, their effect may be reduced by specialized techniques (B1.13).

Fortunately, the worst broadening interactions are also removed naturally in most liquids and solutions, or at least greatly reduced in their effect, by the tumbling motions of the molecules, for many of the broadening interactions vary as $(3 \cos^2\theta - 1)$ where θ is the angle between the H–H vector and B_0 , and so they average to zero when θ covers the sphere isotropically. As a result, the NMR linewidths for the lighter spin- $\frac{1}{2}$ nuclei in smallish molecules will commonly be less than 1 Hz. Resolution of this order is necessary for the adequate resolution of the shifts and couplings described below. It requires expensive magnets and physically homogeneous samples. The presence of irregular interfaces degrades resolution by locally distorting the magnetic field, although the resulting spectra may still have enough resolution for some purposes, such as *in vivo* NMR.

It is occasionally desirable to retain a small proportion of molecular orientation, in order to quantitate the dipolar interactions present, whilst minimizing their contribution to the linewidth. Partial orientation may be achieved by using a nematic solvent. In large, magnetically anisotropic molecules it may occur naturally at the highest magnetic fields.

Figure B1.11.1 shows the range of radiofrequencies where resonances may be expected, between 650 and 140 MHz, when $B_0 = 14.1$ T, i.e. when the ^1H resonance frequency is 600 MHz. There is one bar per stable isotope. Its width is the reported chemical shift range (B1.11.5) for that isotope, and its height corresponds to the log of the sensitivity at the natural abundance of the isotope, covering about six orders of magnitude. The radioactive nucleus ^3H is also included, as it is detectable at safe concentrations and useful for chemical labelling. It is evident that very few ranges overlap. This, along with differences in linewidth, means that a spectrometer set to detect one nucleus is highly unlikely to detect any other in the same experiment.

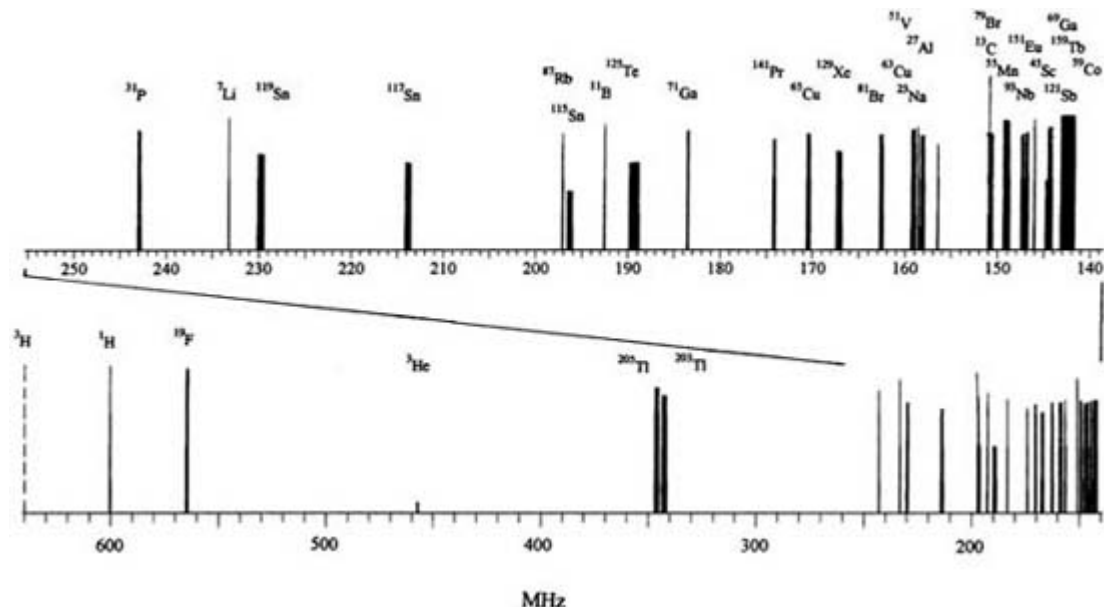


Figure B1.11.1. Resonance frequencies for different nuclei in a field of 14.1 T. Widths indicate the quoted range of shifts for each nucleus, and heights indicate relative sensitivities at the natural isotopic abundance, on a log scale covering approximately six orders of magnitude. Nuclei resonating below 140 MHz are not shown.

B1.11.2.2 PRACTICABLE SAMPLES

Once the above restrictions on isotope, solubility, chemical lability and paramagnetism are met, then a very wide range of samples can be investigated. Gases can be studied, especially at higher pressures. Solutions for ^1H or ^{13}C NMR are normally made in deuteriated solvents. This minimizes interference from the solvent resonances and also permits the field to be locked to the frequency as outlined below. However, it is possible to operate with an unavoidably non-deuteriated solvent, such as water in a biomedical sample, by using a range of solvent-suppression techniques. An external field–frequency lock may be necessary in these cases.

NMR spectra from a chosen nucleus will generally show all resonances from all such nuclei present in solution. Therefore, if at all possible, samples should be chemically pure, to reduce crowding in the spectra, and extraneous compounds such as buffers should ideally not contain the nuclei under study. When chromatographic separations are frequently and routinely essential, then on-line equipment that combines liquid chromatography with NMR is available [2]. Some separation into subspectra is also possible within a fixed sample, using specialized equipment, when the components have different diffusion rates. The technique is called diffusion ordered spectroscopy, or DOSY [3]. If the spectral crowding arises simply from the complexity of the molecule under study, as in proteins, then one can resort to selective isotopic labelling: for example, via gene manipulation. A wide range of experiments is available that are selective for chosen pairs of isotopes, and hence yield greatly simplified spectra.

The available sensitivity depends strongly on the equipment as well as the sample. ^1H is the nucleus of choice for most experiments. 1 mg of a sample of a medium-sized molecule is adequate for almost all types of ^1H -only spectra, and with specialized equipment one can work with nanogram quantities. At this lower level, the problem is not so much sensitivity as purity of sample and solvent. ^{13}C NMR at the natural isotopic abundance of 1.1% typically requires 30 times this quantity of material, particularly if non-protonated carbon atoms are to be studied. Most other nuclei necessitate larger amounts of sample, although ^{31}P [4] and ^{19}F are useful exceptions. *In vivo* spectroscopy generally calls for custom-built probes suited for the organ or organism under study. Human *in vivo* spectra usually require a magnet having an especially wide bore, along with a moveable detection coil. They can reveal abnormal metabolism and internal tissue damage.

NMR can be carried out over a wide range of temperatures, although there is a time and often a resolution penalty in using temperatures other than ambient. An effective lower limit of $\sim -150\text{ }^\circ\text{C}$ is set by the lack of solvents that are liquid below this. Temperatures above $\sim 130\text{ }^\circ\text{C}$ require special thermal protection devices, although measurements have even been made on molten silicates.

B1.11.3 THE NMR EXPERIMENT

B1.11.3.1 EQUIPMENT AND RESONANCE

Figure B1.11.2 represents the essential components of a modern high-resolution NMR spectrometer, suitable for studies of dissolved samples. The magnet has a superconducting coil in a bath of liquid He, jacketed by liquid N_2 . The resulting, persistent field B_0 ranges from 5.9 to 21.1 T, corresponding to ^1H NMR at 250 to 900 MHz. This field can be adjusted to a homogeneity of better than 1 ppb over the volume of a typical sample. The sample is commonly introduced as 0.5 ml of solution held in a 5 mm OD precision glass tube. Microcells are available for very small samples, and special tubes are also available for, for example, pressurized or flow-through samples. The sample can be spun slowly in order to average the field inhomogeneity normal to the tube's axis, but this is not always necessary, for the reduction in linewidth may only be a fraction of a hertz. The field aligns the spins as described above, and a Boltzmann equilibrium then develops between the various m_l Zeeman states. A typical population difference between the two ^1H levels is 1 part in 10^5 . Thus 10^{-5} of the ^1H spins in the sample are not paired. These are collectively called the bulk nuclear magnetization, and in many ways they behave in combination like a classical, magnetized gyroscope. The time required for the establishment of the Boltzmann equilibrium is approximately five times the spin-lattice relaxation time, T_1 (B1.14). Because the population difference between the two ^1H levels is proportional to the field, the sensitivity of NMR also rises with the field.

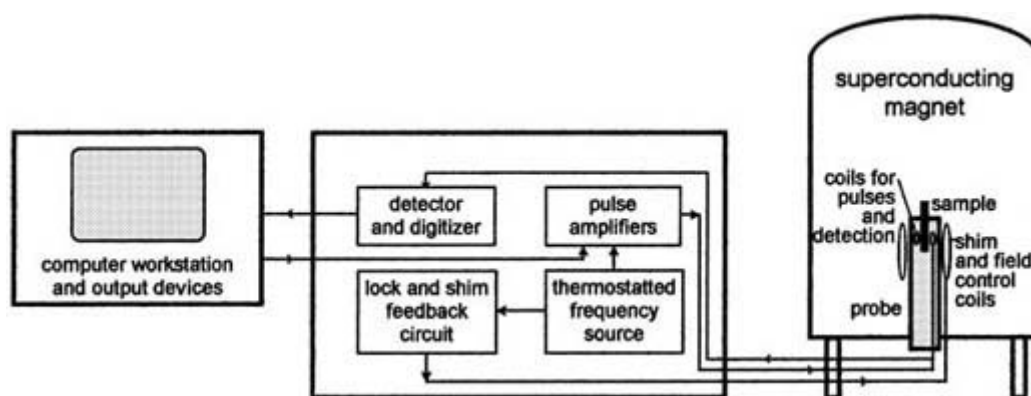


Figure B1.11.2. Simplified representation of an NMR spectrometer with pulsed RF and superconducting magnet. The main magnetic field B_0 is vertical and centred on the sample.

The bulk magnetization is stimulated into precessional motion around B_0 by a radiofrequency (RF) pulse at ν_L , applied through a solenoid-like coil whose axis is perpendicular to B_0 . This motion amounts to a nuclear magnetic resonance. Typically, 10^{-5} s of RF power tilts the magnetization through 90° and thus constitutes a 90° pulse. The coil is then computer-switched from being a transmitter to becoming a receiver, with the free precessional motion of the magnetization generating a small RF voltage in the same coil, at or near the frequency ν_L . The decay time of this oscillating voltage is of the order of T_1 , unless speeded by contributions from exchange or field inhomogeneity. After appropriate amplification and heterodyne detection it is reduced to an audiofrequency and digitized. These raw data are commonly called the free induction decay, or FID,

although the term applies more properly to the original bulk precession.

One further consequence of the use of a pulse, as against the older method of applying a single RF frequency, is that its underlying frequency is broadened over a range of the order of the reciprocal of the pulse length. Indeed, the RF power is in many cases virtually uniform or 'white' across a sufficient range to stimulate all the nuclei of a given isotope at once, even though they are spread out in frequency by differences of chemical shift. Thus, the repeat time for the NMR measurement is approximately T_1 , typically a few seconds, rather than the considerably longer time necessary for sweeping the frequency or the field. Most FIDs are built up by gathering data after repeated pulses, each usually less than 90° in order to permit more rapid repetition. This helps to average away noise and other imperfections, relative to the signal, and it also permits more elaborate experiments involving sequences of multiple pulses and variable interpulse delays.

However, it also necessitates a strictly constant ratio of field to frequency, over the duration of the experiment. Although the master frequency source can be held very constant by a thermostatted source, the field is always vulnerable to local movements of metal, and to any non-persistence of the magnet current. Therefore the field is locked to the frequency through a feedback loop that uses continuous, background monitoring of the ^2H solvent resonance. The probe containing the sample and coil will also normally have at least one further frequency channel, for decoupling experiments (B1.11.6). The lock signal is also simultaneously employed to maximize the field homogeneity across the sample, either manually or automatically, via low-current field correction 'shim' coils. A feedback loop maximizes the height of the lock signal and, because the peak area must nevertheless remain constant, thereby minimizes the peak's width.

-6-

The digitized FID can now be handled by standard computer technology. Several 'spectrum massage' techniques are available for reducing imperfections and noise, and for improving resolution somewhat. The FID is then converted into a spectrum by a discrete Fourier transformation. Essentially, digital sine and cosine waves are generated for each frequency of interest, and the FID is multiplied by each of these in turn. If the product of one particular multiplication does not average to zero across the FID, then that frequency, with that phase, is also present in the FID. The resulting plot of intensity versus frequency is the spectrum. It is normally 'phased' by appropriate combination of the sine and cosine components, so as to contain only positive-going peaks. This permits the measurement of peak areas by digital integration, as well as giving the clearest separation of peaks.

The information within the spectrum can then be presented in many possible ways. In a few cases, it is possible to identify the sample by a fully automatic analysis: for example, by using comparisons with an extensive database. However, most analyses require the knowledge outlined in the following sections.

Many other pulsed NMR experiments are possible, and some are listed in the final sections. Most can be carried out using the standard equipment described above, but some require additions such as highly controllable, pulsed field gradients, shaped RF pulses for (for example) single-frequency irradiations, and the combined use of pulses at several different frequencies.

B1.11.4 QUANTITATION

The simplest use of an NMR spectrum, as with many other branches of spectroscopy, is for quantitative analysis. Furthermore, in NMR all nuclei of a given type have the same transition probability, so that their resonances may be readily compared. The area underneath each isolated peak in an NMR spectrum is proportional to the number of nuclei giving rise to that peak alone. It may be measured to $\sim 1\%$ accuracy by digital integration of the NMR spectrum, followed by comparison with the area of a peak from an added standard.

The absolute measurement of areas is not usually useful, because the sensitivity of the spectrometer depends on factors such as temperature, pulse length, amplifier settings and the exact tuning of the coil used to detect resonance. Peak intensities are also less useful, because linewidths vary, and because the resonance from a given chemical type of atom will often be split into a pattern called a multiplet. However, the relative overall areas of the peaks or multiplets still obey the simple rule given above, if appropriate conditions are met. Most samples have several chemically distinct types of (for example) hydrogen atoms within the molecules under study, so that a simple inspection of the number of peaks/multiplets and of their relative areas can help to identify the molecules, even in cases where no useful information is available from shifts or couplings.

This is illustrated in [figure B1.11.3](#) the integrated ^1H NMR spectrum of commercial paracetamol in deuteriodimethylsulfoxide solvent. The paracetamol itself gives five of the integrated peaks or multiplets. Two other integrated peaks at 3.4 and 1.3 ppm, plus the smaller peaks, arise from added substances. The five paracetamol peaks have area ratios (left to right) of 1:1:2:2:3. These tally with the paracetamol molecule (see diagram). The single H atoms are OH and NH respectively, the double ones are the two distinct pairs of hydrogens on the aromatic ring and the triple ones are the methyl group. Few other molecules will give these ratios, irrespective of peak position.

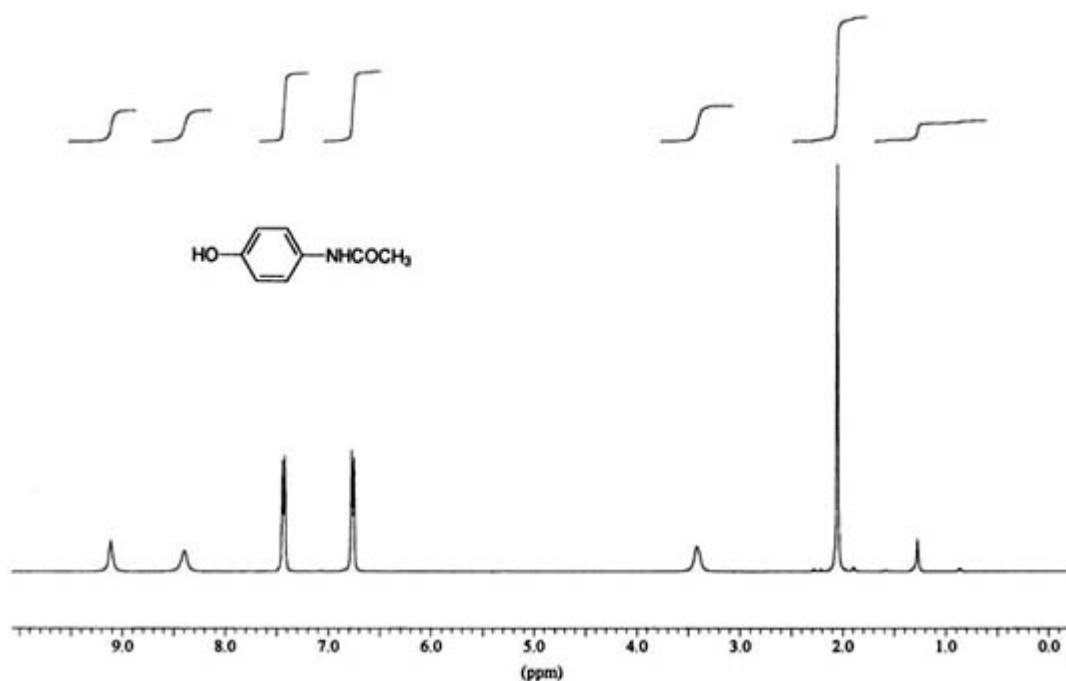


Figure B1.11.3. 400 MHz ^1H NMR spectrum of paracetamol (structure shown) with added integrals for each singlet or multiplet arising from the paracetamol molecule.

The other peaks demonstrate the power of NMR to identify and quantitate all the components of a sample. This is very important for the pharmaceutical industry. Most of the peaks, including a small one accidentally underlying the methyl resonance of paracetamol, arise from stearic acid, which is commonly added to paracetamol tablets to aid absorption. The integrals show that it is present in a molar proportion of about 2%. The broader peak at 3.4 ppm is from water, present because no attempt was made to dry the sample. Such peaks may be identified either by adding further amounts of the suspected substance, or by the more fundamental methods to be outlined below. If the sample were less concentrated, then it would also be possible to detect the residual hydrogen atoms in the solvent, i.e. from its deuteriodimethylsulfoxide- d^5 impurity, which resonates in this case at 2.5 ppm.

It is evident from the figure that impurities can complicate the use of NMR integrals for quantitation. Further complications arise if the relevant spins are not at Boltzmann equilibrium before the FID is acquired. This may occur either because the pulses are repeated too rapidly, or because some other energy input is present, such as decoupling. Both of these problems can be eliminated by careful timing of the energy inputs, if strictly accurate integrals are required.

Their effects are illustrated in [figure B1.11.4](#) which is a ^1H -decoupled ^{13}C NMR spectrum of the same sample of paracetamol, obtained without such precautions. The main peak integrals are displayed both as steps and also as numbers below the peaks. The peaks from stearic acid are scarcely visible, but the dms- d_6 solvent multiplet at 30 ppm is prominent, because ^{13}C was also present at natural abundance in the solvent. The paracetamol integrals should ideally be in the ratios 1:1:1:2:2:1, corresponding to the carbonyl carbon, the four chemically distinct ring carbons and the methyl carbon to the right. However, the first three peaks correspond to carbons that have no attached H, and their integrals are reduced by factors of between 2 and 3. The methyl peak is also slightly reduced.

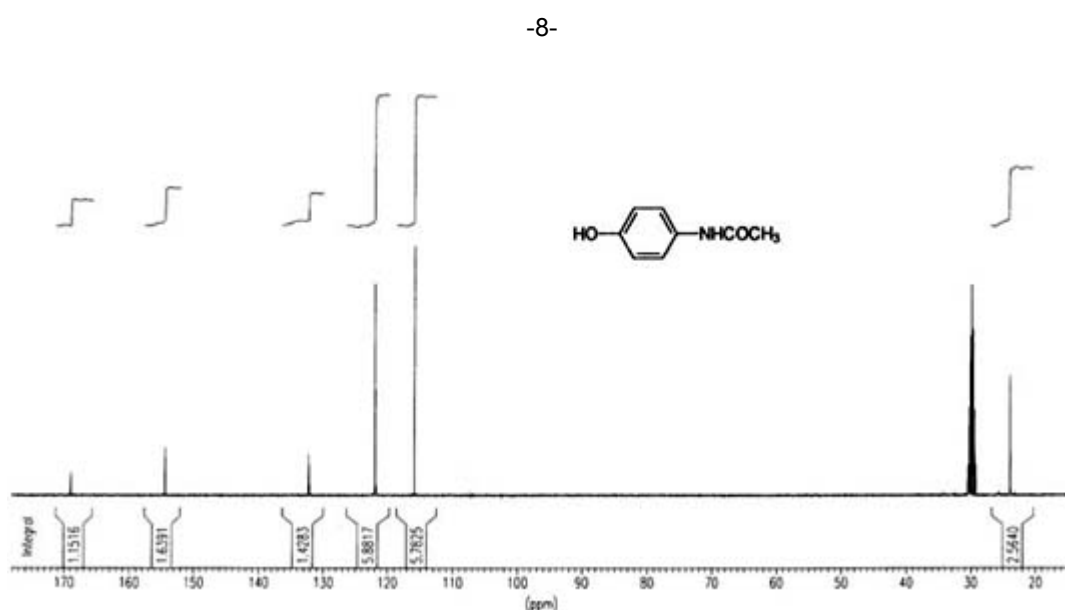


Figure B1.11.4. Hydrogen-decoupled 100.6 MHz ^{13}C NMR spectrum of paracetamol. Both graphical and numerical peak integrals are shown.

These reductions arose because the spectrum was obtained by the accumulation of a series of FIDs, generated by pulses spaced by less than the $5 \times T_1$ interval mentioned above, so that a full Boltzmann population difference was not maintained throughout. The shortfall was particularly acute for the unprotonated carbons, for these have long relaxation times, but it was also significant for the methyl group because of its rapid rotational motion. These losses of intensity are called 'saturation'. If they are intentionally introduced by selective irradiation just prior to the pulse, then they become the analogues of hole-bleaching in laser spectroscopy. They are useful for the selective reduction of large, unwanted peaks, as in solvent suppression. Also, because they are a property of the relevant nuclei rather than of their eventual chemical environment they can be used as transient labels in kinetic studies.

The integrals in figure B1.11.4 are, however, also distorted by a quite different mechanism. The spectrum was obtained in the standard way, with irradiation at all the relevant ^1H frequencies so as to remove any couplings from ^1H . This indirect input of energy partly feeds through to the ^{13}C spins via dipolar coupling, to produce intensity gains at all peaks, but particularly at protonated carbons, of up to $\times 3$, and is an example of the nuclear Overhauser enhancement. The phenomenon is quite general wherever T_1 is dominated by dipolar interactions, and when any one set of spins is saturated, whether or not decoupling also takes place. It was discovered by A W Overhauser in the context of the dipolar interactions of electrons with nuclei [5]. The

NOEs between ^1H nuclei are often exploited to demonstrate their spatial proximity, as described in the final section. It is possible to obtain decoupled ^{13}C NMR spectra without the complications of the NOE, by confining the decoupling irradiation to the period of the FID alone, and then waiting for $10 \times T_1$ before repeating the process. However, the $\times 3$ enhancement factor is useful and also constant for almost all protonated carbons, so that such precautions are often superfluous. Both carbon and hydrogen integrals are particularly valuable for identifying molecular symmetry.

-9-

Figure B1.11.5 is an example of how relative integrals can determine structure even if the peak positions are not adequately understood. The decavanadate anion has the structure shown, where oxygens lie at each vertex and vanadiums at the centre of each octahedron. An aqueous solution of decavanadate was mixed with about 8 mol% of molybdate, and the three peaks from the remaining decavanadate were then computer-subtracted from the ^{51}V NMR spectrum of the resulting equilibrium mixture [6]. The remaining six peaks arise from a single product and, although their linewidths vary widely, their integrals are close to being in a 2:2:2:1:1:1 ratio. This not only suggests that just one V has been replaced by Mo, but also identifies the site of substitution as being one of the four MO_6 octahedra not lying in the plane made by vanadiums 1, 2 and 3. No other site or extent of substitution would give these integral ratios.

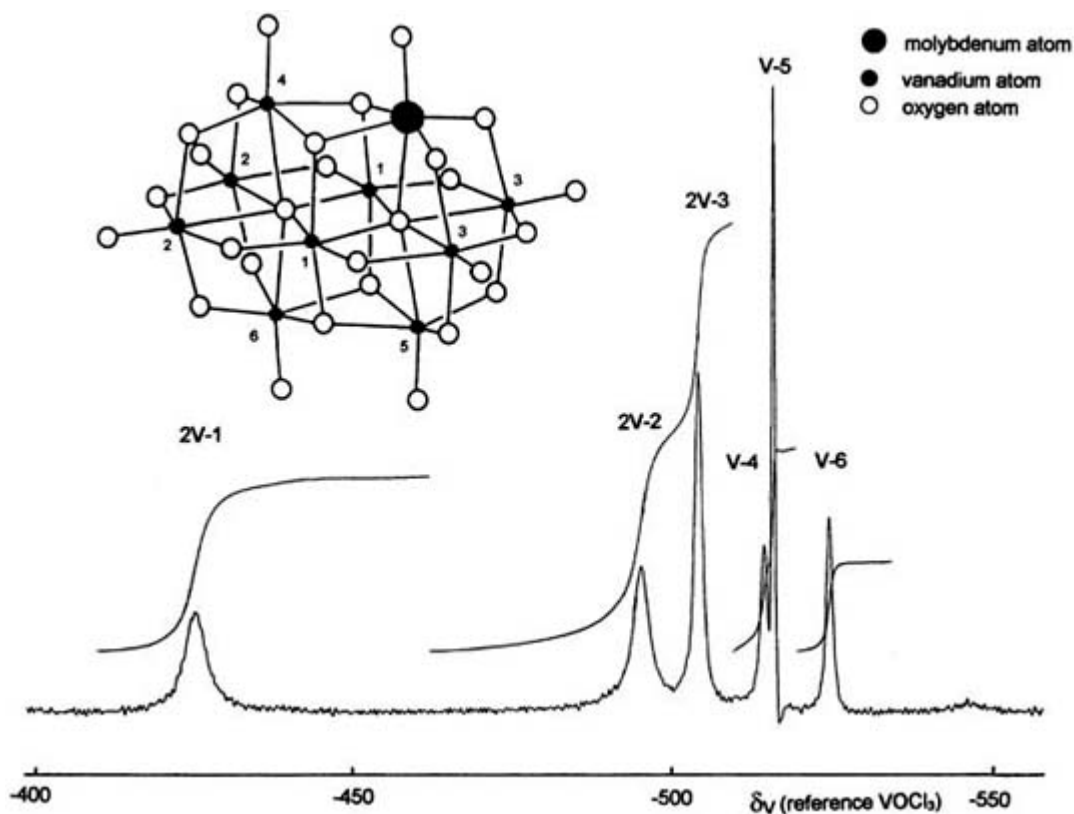


Figure B1.11.5. 105 MHz ^{51}V NMR subtraction spectrum of the $[\text{MoV}_9\text{O}_{28}]^{5-}$ anion (structure shown). The integrals are sufficient to define the position of the Mo atom.

When T_1 is very short, which is almost always true with nuclei having $I > 1/2$, the dipolar contribution to relaxation will be negligible and, hence, there will be no contributions to the integral from either NOE or saturation. However, resonances more than about 1 kHz wide may lose intensity simply because part of the FID will be lost before it can be digitized, and resonances more than 10 kHz wide may be lost altogether. It is also hard to correct for minor baseline distortions when the peaks themselves are very broad.

B1.11.5 CHEMICAL SHIFTS

Strictly speaking, the horizontal axis of any NMR spectrum is in hertz, as this axis arises from the different frequencies that make up the FID. However, whilst this will be important in the section that follows (on coupling), it is generally more useful to convert these frequency units into fractions of the total spectrometer frequency for the nucleus under study. The usual units are parts per million (ppm). Because frequency is strictly proportional to field in NMR, the same ppm units also describe the fractional shielding or deshielding of the main magnetic field, by the electron clouds around or near the nuclei under study. The likely range of such shieldings is illustrated in [figure B1.11.1](#) [figure B1.11.3](#) and [figure B1.11.4](#).

B1.11.5.1 CALIBRATION

The ppm scale is always calibrated relative to the appropriate resonance of an agreed standard compound, because it is not possible to detect the NMR of bare nuclei, even though absolute shieldings can be calculated with fair accuracy for smaller atoms. Added tetramethylsilane serves as the standard for ^1H , ^2H , ^{13}C and ^{29}Si NMR, because it is a comparatively inert substance, relatively immune to changes of solvent, and its resonances fall conveniently towards one edge of most spectra. Some other standards may be unavoidably less inert. They are then contained in capillary tubes or in the annulus between an inner and an outer tube, where allowance must be made for the jump in bulk magnetic susceptibility between the two liquids. For dilute solutions where high accuracy is not needed, a resonance from the deuteriated solvent may be adequate as a secondary reference. The units of chemical shift are δ . This unit automatically implies ppm from the standard, with the same sign as the frequency, and so measures deshielding. Hence $\delta = 10^6(\nu - \nu_{\text{ref}})/\nu_{\text{ref}}$

B1.11.5.2 LOCAL CONTRIBUTIONS TO THE CHEMICAL SHIFT

The shielding at a given nucleus arises from the virtually instantaneous response of the nearby electrons to the magnetic field. It therefore fluctuates rapidly as the molecule rotates, vibrates and interacts with solvent molecules. The changes of shift with rotation can be large, particularly when double bonds are present. For example, the ^{13}C shift of a carbonyl group has an anisotropy comparable to the full spectrum width in [figure B1.11.4](#). Fortunately, these variations are averaged in liquids, although they are important in the NMR of solids. This averaging process may be visualized by imagining the FID emitted by a single spin. If the emission frequency keeps jumping about by small amounts, then this FID will be made up from a series of short segments of sine waves. However, if the variations in frequency are small compared to the reciprocal of the segment lengths, then the composite wave will be close to a pure sine wave. Its frequency will be the weighted average of its hidden components and the rapid, hidden jumps will not add to the linewidth. This is described as ‘fast exchange on the NMR timescale’ and it is discussed more fully in B.2.7. The same principles apply to rapid chemical changes, such as aqueous protonation equilibria. In contrast, somewhat slower exchange processes increase linewidths, as seen for the OH and NH resonances in [figure B1.11.3](#). In this sample, the individual molecules link transiently via hydrogen bonding, and are thus in exchange between different H-bonded oligomers.

The two primary causes of shielding by electrons are diamagnetism and temperature-independent paramagnetism (TIP). Diamagnetism arises from the slight unpairing of electron orbits under the influence of the magnetic field. This always occurs so as to oppose the field and was first analysed by Lamb [7]. A simplified version of his formula,

appropriate for a single orbital in isolated atoms, shows that the diamagnetic shielding contribution to δ is

proportional to $-\rho_e \langle r^2 \rangle$ where ρ_e is the electron density and $\langle r^2 \rangle$ is the average squared radius of the electron orbit [8]. Thus diamagnetic shielding is lowered and, hence, δ is increased, by the attachment of an electronegative group to the atom under study, for this decreases both ρ_e and $\langle r^2 \rangle$. Similarly, in conjugated systems, δ will often approximately map the distribution of electronic charge between the atoms.

The TIP contribution has the opposite effect on δ . It should not be confused with normal paramagnetism, for it does not require unpaired electrons. Instead, it arises from the physical distortion of the electron orbitals by the magnetic field. A highly simplified derivation of the effect of TIP upon δ shows it to be proportional to $+\langle r^{-3} \rangle / \Delta E$, with r as above. ΔE is a composite weighting term representing the energy gaps between the occupied and the unoccupied orbitals. A small ΔE implies orbitals that are accessible and hence vulnerable to magnetic distortion. ΔE is particularly large for H, making the TIP contribution small in this case, but for all other atoms TIP dominates δ . The effect of an electronegative group on the TIP of an atom is to shrink the atom slightly and, thus, to increase δ , i.e. it influences chemical shift in the same direction as described above in the case of diamagnetism, but for a different reason.

These effects are clearly seen for saturated molecules in [table B1.11.1](#), which correlates the ^1H and ^{13}C NMR chemical shifts of a series of compounds $\text{X}-\text{CH}_3$ with the approximate Pauling electronegativities of X. Both δ_{H} and δ_{C} increase with X as predicted, although the sequence is slightly disordered by additional, relativistic effects at the C atoms, when X is a heavy atom such as I. The correlations of shift with changes in the local electric charge and, hence, with orbital radius, are also seen for an aromatic system in [figure B1.11.6](#). Here the hydrogen and carbon chemical shifts in phenol, quoted relative to benzene, correlate fairly well both with each other and also with the charge distribution deduced by other means.

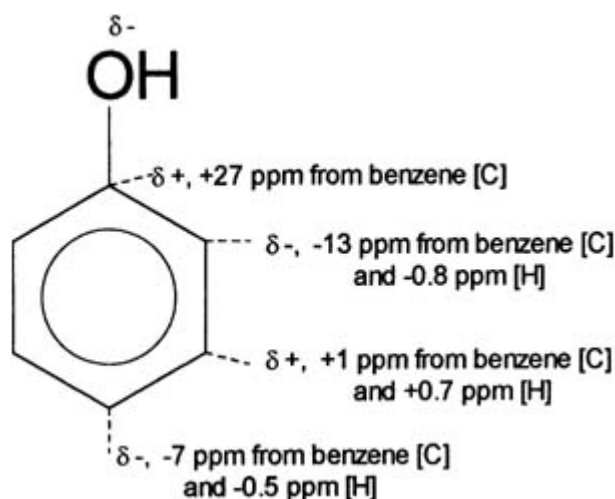


Figure B1.11.6. ^1H and ^{13}C chemical shifts in phenol, relative to benzene in each case. Note that δ (H or C) approximately follows δ (the partial charge at C).

Table B1.11.1. Effect of an electronegative substituent upon methyl shifts in $\text{X}-^{13}\text{C}^1\text{H}_3$.

X	δ_{H}	δ_{C}	Electronegativity of X
---	---------------------	---------------------	------------------------

Si(CH ₃) ₃	0.0	0.0	1.8
H	0.13	-2.3	2.1
CH ₃	0.88	5.7	2.5
I	2.16	-20.7 ^a	2.5
NH ₂	2.36	28.3	2.8
Br	2.68	10.0 ^a	2.8
Cl	3.05	25.1	3.0
OH	3.38	49.3	3.5
F	4.26	75.4	4.0

^a Heavy-atom relativistic effects influence these shifts.

The effects of TIP also appear in [figure B1.11.3](#) and [figure B1.11.4](#). In the ¹³C NMR spectrum, all the resonances of the sp² carbons lie above 100 ppm (a useful general rule of thumb) because ΔE is smaller for multiple bonds. The highest shifts are for the carbonyl C at 169 ppm and the ring C attached to oxygen at 155 ppm, because of the high electronegativity of O. In the ¹H spectrum, H atoms attached to sp² carbons also generally lie above 5 ppm and below 5 ppm if attached to sp³ carbons. However, the NH and OH resonances have much less predictable shifts, largely governed by the average strength of the associated hydrogen bonds. When these are exceptionally strong, as in nucleic acids, δ_{H} can be as high as 30 ppm, whereas most normal H shifts lie in the range 0–10 ppm.

[Table B1.11.2](#) gives a different example of the effects of TIP. In many fluorine compounds, ΔE is largely determined by the energy difference between the bonding and antibonding orbitals in the bond to F, and hence by the strength of this bond. The ¹⁹F shifts correlate nicely in this way. For example, C–F bonds are notoriously strong and so give low values of δ_{F} , whereas the remarkable reactivity of F₂ depends on the weakness of the F–F bond, which also gives F₂ a much higher chemical shift. The weakness of the bond to F is even more apparent in the chemical shifts of the explosive compounds XeF₆ and FOOF. In some compounds the spectra also reflect the presence of distinct types of fluorine within the molecule. For example, ClF₃ has a T-shaped structure with the Cl–F bond stronger for the stem of the T, giving this single F atom a lower shift.

-13-

Table B1.11.2. Fluorine-19 chemical shifts.

Organic compounds	δ_{F}	Inorganic compounds	δ_{F}	Metal fluorides	δ_{F}
-------------------	---------------------	---------------------	---------------------	-----------------	---------------------

CH ₃ F	-272	HF	-221	MoF ₆	-278
CH ₃ CH ₂ F	-213	LiF	-210	SbF ₅	-108
C ₆ F ₆	-163	[BF ₄] ⁻	-163	WF ₆	+166
CH ₂ F ₂	-143	BF ₃	-131	ReF ₇	+345
F ₂ C=CF ₂	-135	ClF ₃	-4 (1), +116 (2)		
H ₂ C=CF ₂	-81	IF ₇	+170		
CF ₃ R	-60 to -70	IF ₅	+174 (1), +222 (4)		
CF ₂ Cl ₂	-8	ClF ₅	+247 (1), +412 (4)		
CFCl ₃	0 (reference)	XeF ₂	+258		
CF ₂ Br ₂	+7	XeF ₄	+438		
CFBr ₃	+7	XeF ₆	+550		
		F ₂	+421.5		
		ClF	+448.4		
		FOOF	+865		

The shifts for other nuclei are not usually so simple to interpret, because more orbitals are involved, and because these may have differing effects on $\langle r^{-3} \rangle$ and on ΔE . Nevertheless, many useful correlations have been recorded [9]. A general rule is that the range of shifts increases markedly with increasing atomic number, mainly because of decreases in ΔE and increases in the number of outer-shell electrons. The shift range is particularly large when other orbital factors also lower and vary ΔE , as in cobalt[III] complexes. However, the chemical shifts for isotopes having the same atomic number, such as ¹H, ²H and ³H, are almost identical. Each nucleus serves merely to report the behaviour of the same electron orbitals, except for very small effects of isotopic mass on these orbitals.

B1.11.5.3 CHEMICAL SHIFTS ARISING FROM MORE DISTANT MOIETIES

Hydrogen shifts fall within a narrow range and, hence, are proportionally more susceptible to the influence of more distant atoms and groups. As an example, aromatic rings have a large diamagnetic anisotropy, because the effective radius of the electron orbit around the ring is large. This affects the chemical shifts of nearby atoms by up to 5 ppm, depending on their position relative to the ring. It also explains the relatively high shifts of the hydrogens directly attached to the ring. The diamagnetism of the ring is equivalent to a solenoid opposing B_0 . Its lines of force necessarily loop backwards outside the ring and therefore deshield atoms in or near the ring plane. The effect is illustrated in figure B1.11.7 for a paracyclophane. The methylene shift values fall both above and below the value of 1.38, expected in the absence of the aromatic moiety, according to the average positioning of the methylene hydrogens relative to the ring plane. Similar but smaller shifts are

observed with simple double and triple bonds, because these also generate a significant circulation of electrons. Polycyclic aromatic systems generate even larger shifts, in regions near to more than one ring.

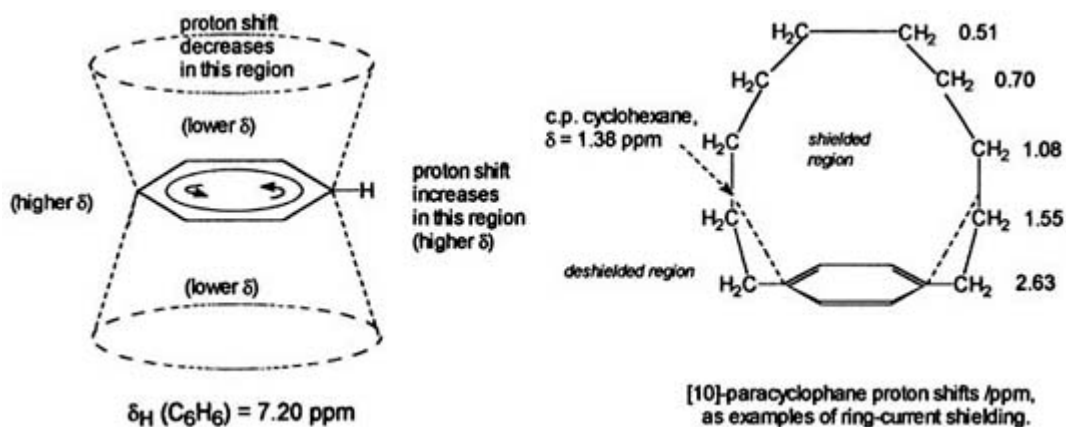


Figure B1.11.7. ^1H chemical shifts in [10]-paracyclophane. They have values on either side of the 1.38 ppm found for large polymethylene rings and, thus, map the local shielding and deshielding near the aromatic moiety, as depicted in the upper part of the figure.

Shifts are also affected by steric compression of any kind on the atom under study. The effect on a C atom can reduce δ_C by up to 10 ppm. For example, the ^{13}C chemical shifts of the methyl carbons in but-2-ene are 5 ppm lower in the *cis* isomer than in the *trans*, because in the *cis* case the methyl carbons are only about 3 Å apart. Steric shifts are particularly important in the ^{13}C NMR spectra of polymers, for they make the peak positions dependent on the local stereochemistry and, hence, on the tacticity of the polymer [10].

B1.11.5.4 SHIFT REAGENTS

Nearby paramagnetic molecules can also have a similar effect on shifts. Their paramagnetic centre will often possess a substantially anisotropic *g*-factor. If the paramagnetic molecule then becomes even loosely and temporarily attached to the molecule under study, this will almost always lead to significant shift changes, for the same reasons as with nearby multiple bonds. Furthermore, if the molecule under study and its paramagnetic attachment are both chiral, the small variations in the strength and direction of the attachment can effect a separation of the shifts of the two chiral forms

and thus detect any enantiomeric excess. The paramagnetic species thus acts as a chiral shift reagent. In practice, the effect of the paramagnetic additive will be complicated, because there may be further shifts arising from direct leakage of free electron density between the paired molecules. There will also be broadening due to the large magnetic dipole of the unpaired electrons. These complications can be reduced by a judicious choice of shift reagent. It may also be possible to use a diamagnetic shift reagent such as 2,2'-trifluoromethyl-9-anthrylethanol and thus avoid the complications of paramagnetic broadening altogether, whilst hopefully retaining the induced chiral separation of shifts [11].

B1.11.5.5 THE PREDICTION OF CHEMICAL SHIFTS

Enormous numbers of chemical shifts have been recorded, particularly for ^1H and ^{13}C . Many algorithms for the prediction of shifts have been extracted from these, so that the spectra of most organic compounds can be predicted at a useful level of accuracy, using data tables available in several convenient texts [12, 13, 14 and 15]. Alternatively, computer programs are available that store data from 10^4 – 10^5 spectra and then use direct

structure-comparison methods to predict the ^1H and ^{13}C chemical shifts of new species.

Shifts can also be predicted from basic theory, using higher levels of computation, if the molecular structure is precisely known [16]. The best calculations, on relatively small molecules, vary from observation by little more than the variations in shift caused by changes in solvent. In all cases, it is harder to predict the shifts of less common nuclei, because of the generally greater number of electrons in the atom, and also because fewer shift examples are available.

B1.11.6 THE DETECTION OF NEIGHBOURING ATOMS–COUPLINGS

The ^1H NMR spectrum in figure B1.11.2 shows two resonances in the region around 7 ppm, from the two types of ring hydrogens. Each appears as a pair of peaks rather than as a single peak. This splitting into a pair, or ‘doublet’, has no connection with chemical shift, because if the same sample had been studied at twice the magnetic field, then the splitting would have appeared to halve, on the chemical shift scale. However, the same splitting does remain strictly constant on the concealed, frequency scale mentioned previously, because it derives not from B_0 , but instead from a fixed energy of interaction between the neighbouring hydrogen atoms on the ring. Each pair of peaks in the present example is called a doublet: a more general term for such peak clusters is a multiplet. The chemical shift of any symmetrical multiplet is the position of its centre, on the ppm scale, whether or not any actual peak arises at that point. The area of an entire multiplet, even if it is not fully symmetric, obeys the principles of section B1.11.4 above. It follows that the individual peaks in a multiplet may be considerably reduced in intensity compared with unsplit peaks, especially if many splittings are present.

The splittings are called J couplings, scalar couplings or spin–spin splittings. They are also closely related to hyperfine splittings in ESR. Their great importance in NMR is that they reveal interactions between atoms that are chemically nearby and linked by a bonding pathway. The J value of a simple coupling is the frequency difference in hertz between the two peaks produced by the coupling. It is often given by the symbol nJ , where n is the number of bonds that separate the coupled atoms. Thus, in combination with chemical shifts and integrals, they will often allow a chemical structure to be determined from a single ^1H NMR spectrum, even of a previously unknown molecule of some complexity. Their presence also permits a wide range of elegant experiments where spins are manipulated by carefully timed pulses. However, they also complicate spectra considerably if more than a few interacting atoms are present,

-16-

particularly at low applied fields, when the chemical shift separations, converted back to hertz, are not large in comparison with the couplings. Fortunately some of the aforesaid elegant experiments can extract information on the connectivities of atoms, even when the multiplets are too complex or overlapped to permit a peak-by-peak analysis.

B1.11.6.1 COUPLING MECHANISMS

The simplest mechanism for spin–spin couplings is described by the Fermi contact model. Consider two nuclei linked by a single electron-pair bond, such as ^1H and ^{19}F in the hydrogen fluoride molecule. The magnetic moment of ^1H can be either ‘up’ or ‘down’ relative to B_0 , as described earlier, with either possibility being almost equally probable at normal temperatures. If the ^1H is ‘up’ then it will have a slight preferential attraction for the ‘down’ electron in the bond pair, i.e. the one whose magnetic moment lies antiparallel to it, for magnets tend to pair in this way. The effect will be to unpair the two electrons very slightly and, thus, to make the ^{19}F nucleus more likely to be close to the ‘up’ electron and thus slightly favoured in energy, if it is itself ‘down’. The net result is that the ^1H and ^{19}F nuclei gain slightly in energy when they are mutually antiparallel. The favourable and unfavourable arrangements are summarized as follows:

energetically favourable: $H\uparrow e\downarrow\uparrow e\downarrow F$ $H\downarrow e\uparrow\downarrow e\uparrow F$ energy gain $\frac{1}{4}J$
energetically unfavourable: $H\uparrow e\downarrow\uparrow e\uparrow F$ $H\downarrow e\uparrow\downarrow e\downarrow F$ energy loss $\frac{1}{4}J$.

These energy differences then generate splittings as outlined in [figure B1.11.8](#). If the energies of the two spins, here generalized to I and S, in the magnetic field are corrected by the above $\pm \frac{1}{4}J$ quantities, then the transitions will move from their original frequencies of ν_I and ν_S to $\nu_I \pm \frac{1}{2}J$ and $\nu_S \pm \frac{1}{2}J$. Thus, both the I and the S resonances will each be symmetrically split by J Hz. In the case of $^1H^{19}F$, $^1J_{HF} = +530$ Hz. The positive sign applies when the antiparallel nuclear spin configuration is found, and the '1' superscript refers to the number of bonds separating the two spins.

[Figure B1.11.8](#) does not apply accurately in the not uncommon cases when I and S have Larmor frequencies whose separation $\Delta\nu$ is not large compared with J_{IS} . In these cases the two spin states where I and S are antiparallel have very similar energies, so that the coupling interaction mixes their spin states, analogously to the non-crossing rule in UV-visible spectroscopy. As $\Delta\nu$ falls, the peak intensities in the multiplet alter so as to boost the component(s) nearest to the other multiplet, at the expense of the other components, whilst keeping the overall integral of the multiplet constant. This 'roofing' or 'tenting' is an example of a second-order effect on couplings. It is actually useful up to a point, in that it helps the analysis of couplings in a complex molecule, by indicating how they pair off. Some examples are evident in subsequent figures. [Figure B1.11.10](#) for example, shows mutual roofing of the H-6a and H-6b multiplets, and also of H-3 with H-2. Roofing is also seen, more weakly, in [Figure B1.11.9](#), where every multiplet is slightly tilted towards the nearest multiplet with which it shares a coupling. Indeed, it is unusual not to see second-order distortions in 1H NMR spectra, even when these are obtained at very high field.

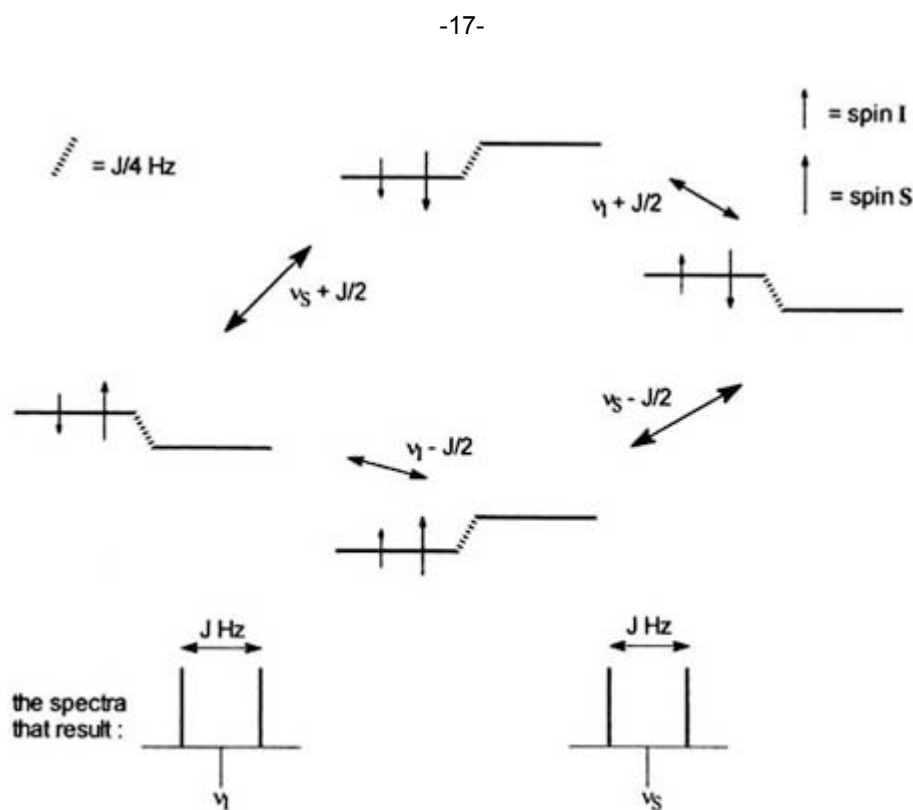


Figure B1.11.8. Combined energy states for two spins I and S (e.g. 1H and ^{13}C) with an exaggerated representation of how their mutual alignment *via* coupling affects their combined energy.

As $\Delta\nu$ gets even smaller, the outer components of the multiplet become invisibly small. Further splittings may also appear in complex multiplets. In the extreme case that $\Delta\nu = 0$, the splittings disappear altogether, even though the physical coupling interaction is still operative. This is why the methyl resonance in [figure B1.11.3](#) appears as a single peak. It is occasionally necessary to extract accurate coupling constants and chemical shifts from complex, second-order multiplets, and computer programs are available for this.

B1.11.6.2 FACTORS THAT DETERMINE COUPLING CONSTANTS

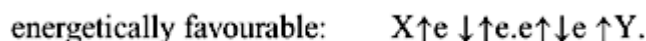
Because J arises from the magnetic interactions of nuclei, the simplest factor affecting it is the product $\gamma_I\gamma_S$ of the two nuclear magnetogyric ratios involved. For example, $^1J_{DF}$ in $^2H^{19}F$ is 82 Hz, i.e. $^1J_{HF} \times \gamma_D/\gamma_H$. This totally predictable factor is sometimes discounted by quoting the reduced coupling constant $K_{IS} \equiv 4\pi^2 J_{IS}/h\gamma_I\gamma_S$.

A second determining factor in the Fermi contact mechanism is the requirement that the wavefunction of the bonding orbital has a significant density at each nucleus, in order for the nuclear and the electron magnets to interact. One consequence of this is that K correlates with nuclear volume and therefore rises sharply for heavier nuclei. Thus the 1K constants in the XH_4 series with $X = ^{13}C, ^{29}Si, ^{73}Ge, ^{119}Sn$ and ^{207}Pb are respectively +41.3, +84.9, +232, +430 and +938 $N A^{-2} m^{-3}$. Here the average value of $(K/\text{nuclear mass number}) = 3.5 \pm 1$.

-18-

The presence of a spin-spin splitting therefore means that the interacting atomic orbitals must each possess significant s character, because all orbitals other than s have zero density at the nucleus. The $^1J_{CH}$ values for CH_4 , $H_2C=CH_2$ and $HC\equiv CH$ illustrate the dependence of J upon s character. They are respectively +124.9, +156.2 and +249 Hz, and so they correlate unusually precisely with the s characters of the carbon orbitals bonding to H, these being respectively 25, 33.3 and 50%. A coupling also shows that the bond that carries the coupling cannot be purely ionic. A significant J value demonstrates covalency, or at least orbital overlap of some kind.

The next large influence on nJ is the number of bonds n in the bonding pathway linking the two spins and their stereochemistry. When $n > 1$, then there will be at least two electron pairs in the pathway, and J will thus be attenuated by the extent to which one electron pair can influence the next. In simple cases, the insertion of a further electron pair will reverse the sign of J , because parallel electrons in different, interacting orbitals attract each other slightly, following Hund's rules. Thus, for a two-bond pathway, a favoured spin configuration will now be



Some other general rules have been extracted; they come particularly from $^nJ_{HH}$ data, but also have a wider validity in many cases.

- $n = 2$ 2J is often low, because of competing pathways involving different orbitals. A typical value for chemically distinct H atoms in a methylene group is -14 Hz. However, it can rise to zero, or even to a positive value, in $C = CH_2$ moieties.
- $n = 3$ 3J is almost always positive and its magnitude often exceeds that of 2J . It always depends in a predictable way on the dihedral angle ϕ between the outer two of the three bonds in the coupling pathway. Karplus first showed theoretically that 3J varies to a good approximation as $A \cos^2\phi + B \cos\phi$, where A and B are constants, and also that $A \gg B$ [17]. His equation has received wide-ranging

experimental verification. For a typical HCCH pathway, with not particularly electronegative substituents, $A = 13$ Hz and $B = -1$ Hz. This predicts ${}^3J = 14.0$ Hz for $\phi = 180^\circ$ and 2.75 Hz for $\phi = 60^\circ$. Thus, a typical HCCH₃ coupling, where the relevant dihedral angles are 60° (twice) and 180° (once) will average to $(2 \times 2.75 + 14)/3 = 6.5$ Hz, as the C–C bond rotates. Almost all other 3J couplings follow a similar pattern, with 3J being close to zero if $\phi = 90^\circ$, even though the values of A vary widely according to the atoms involved. The pattern may be pictured as a direct, hyperconjugative interaction between the outer bonding orbitals, including their rear lobes. It is only fully effective when these orbitals lie in the same plane. The Karplus equation offers a valuable way of estimating bond angles from couplings, especially in unsymmetrical molecular fragments such as CH–CH₂, where the presence of two 3J couplings eliminates any ambiguities in the values of ϕ .

- $n = 4$ 4J couplings are often too small to resolve. However, they are important in cases where the relevant orbitals are aligned appropriately. If the molecular fragment HC–C=CH has the first CH bond approximately parallel to the C=C π orbital, the resulting hyperconjugative interaction will give rise to an ‘allylic’ coupling of about -2 Hz. A similar coupling arises in a saturated HC–C–CH fragment if the bonds lie approximately in a W configuration. In such cases the rear lobes of the CH bonding orbitals touch, thus offering an extra electronic pathway for coupling. Indeed, all such contacts give rise to couplings, even if the formal bonding pathway is long.

-19-

Although the Fermi contact mechanism dominates most couplings, there are smaller contributions where a nuclear dipole physically distorts an orbital, not necessarily of s type [18]. There are many useful compilations of J and K values, especially for HH couplings (see [9], ch 4, 7–21 and [12, 13, 14 and 15]).

B1.11.6.3 MULTIPLE COUPLINGS

In principle, every nucleus in a molecule, with spin quantum number I , splits every other resonance in the molecule into $2I + 1$ equal peaks, i.e. one for each of its allowed values of m_I . This could make the NMR spectra of most molecules very complex indeed. Fortunately, many simplifications exist.

- As described above, most couplings fall off sharply when the number of separating bonds increases.
- Nuclei with a low isotopic abundance will contribute correspondingly little to the overall spectra of the molecule. Thus, splittings from the 1.1% naturally abundant ${}^{13}\text{C}$ isotope are only detectable at low level in both ${}^1\text{H}$ and ${}^{13}\text{C}$ NMR spectra, where they are called ${}^{13}\text{C}$ sidebands.
- Nuclei with $I > 1/2$ often relax so rapidly that the couplings from them are no longer detected, for reasons analogous to shift averaging by chemical exchange. For example, couplings from Cl, Br and I atoms are invisible, even though these halogens all have significant magnetic moments.
- Chemical exchange will similarly average couplings to zero, if it takes place much faster than the value of the coupling in hertz.
- Selected nuclei can also have their couplings removed deliberately (decoupled) by selective irradiation during the acquisition of the FID.

Despite these simplifications, a typical ${}^1\text{H}$ or ${}^{19}\text{F}$ NMR spectrum will normally show many couplings. [Figure B1.11.9](#) is the ${}^1\text{H}$ NMR spectrum of propan-1-ol in a dilute solution where the exchange of OH hydrogens between molecules is slow. The underlying frequency scale is included with the spectrum, in order to emphasize how the couplings are quantified. Conveniently, the shift order matches the chemical order of the atoms. The resonance frequencies of each of the 18 resolved peaks can be quantitatively explained by the four chemical shifts at the centre of each multiplet, plus just three values of 3J , two of these being in fact almost the same. If the hydrogen types are labelled in chemical order from the methyl hydrogens H₃-3 to the

hydroxyl hydrogen then the three coupling types visibly present are ${}^3J_{23} = 2a$, ${}^3J_{12} = 2b$ and ${}^3J_{1-OH} = 2c$ Hz, with $a \approx b$ because of the strong chemical similarity of all the HCCH bonding pathways. Consider first the methyl resonance H_3-3 , and its interaction with the two equivalent hydrogens $H-2$ and $H-2'$. Each of the three equivalent hydrogens of H_3-3 will first be split into a 1:1 doublet by $H-2$, with peak positions $\delta_3 \pm a$, i.e. a Hz on either side of the true chemical shift δ_3 , reconverted here into frequency units. $H-2'$ will further split each component, giving three peaks with positions $\delta_3 \pm a \pm a$. When all the possible \pm combinations are considered, they amount to peaks at δ_3 (twice) plus $\delta_3 \pm 2a$ (once each) and are conveniently called a (1:2:1) triplet. The OH multiplet is similarly a (1:2:1) triplet with peaks at δ_{OH} (twice) plus $\delta_{OH} \pm 2c$ (once each).

-20-

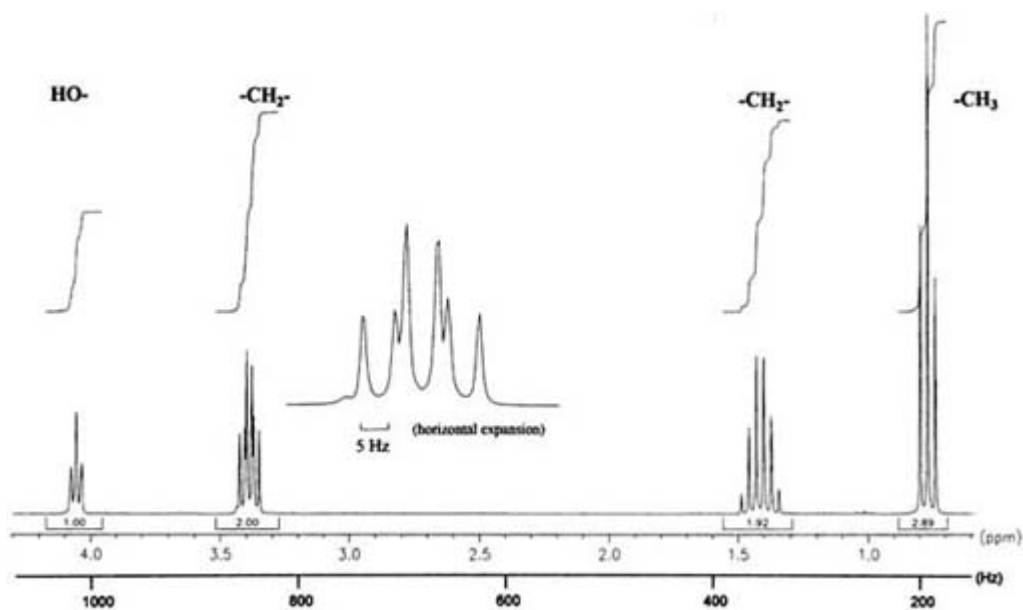


Figure B1.11.9. Integrated 250 MHz ${}^1\text{H}$ NMR spectrum of dilute propan-1-ol in dimethylsulfoxide solvent. Here, the shift order parallels the chemical order. An expansion of the H_2-1 multiplet is included, as is the implicit frequency scale, also referenced here to TMS = 0.

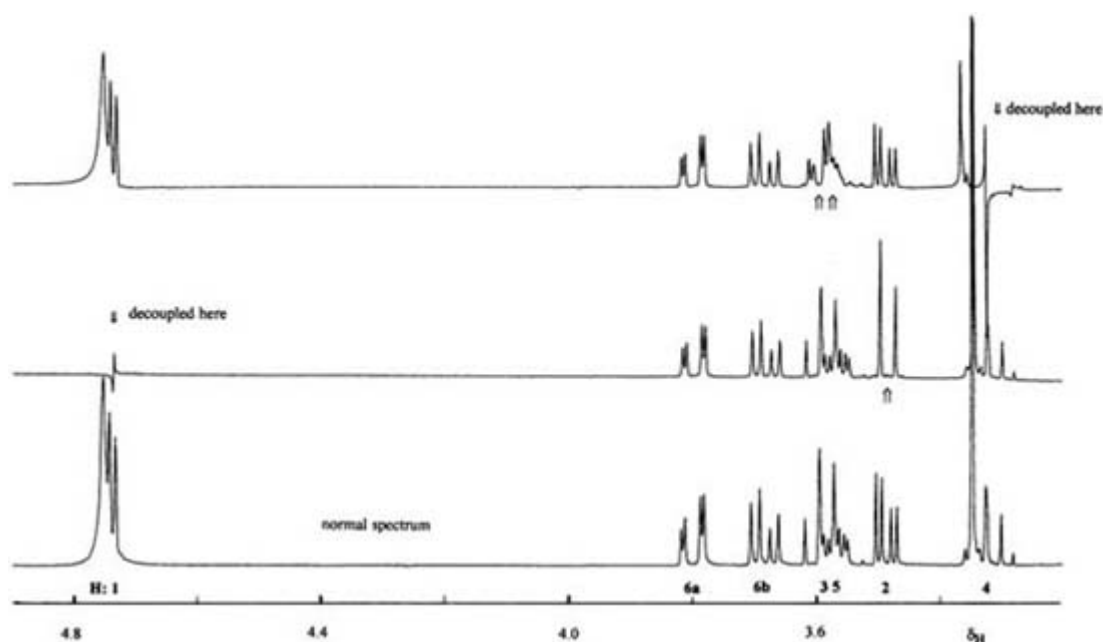


Figure B1.11.10. 400 MHz ${}^1\text{H}$ NMR spectrum of methyl- α -glucopyranose (structure as in figure B1.11.12)

together with the results of decoupling at H-1 (centre trace) and at H-4 (upper trace).

-21-

Each H-1 hydrogen contributes to a more complex multiplet, with peaks at all combinations of the frequencies $\delta_1 \pm b \pm b \pm c$, thus making a triplet of doublets whose peak positions are $\delta_1 \pm c$ (twice each) and $\delta_1 \pm 2b \pm c$ (once each). Finally, each H-2 hydrogen contributes to the sextet at $\delta_2 = 1.42$ ppm, whose individual components appear at $\delta_2 \pm a \pm a \pm a \pm b \pm b$. If we take $b = a$, then the possible combinations of these frequencies amount to the following peak positions, with their relative heights given in brackets: $\delta_2 \pm 5a$ (1), $\delta_2 \pm 3a$ (5), $\delta_2 \pm a$ (10). Note that the relative intensities follow Pascal's triangle and also that the separation of the outermost line of a multiplet must always equal the sum of the underlying couplings, when only spin-1/2 nuclei are involved.

The same principles apply to couplings from spins with $I > 1/2$, where these are not seriously affected by relaxation. [Figure B1.11.4](#) illustrates a common case. The solvent resonance at 30 ppm is a 1:3:6:7:6:3:1 multiplet arising from the $^{13}\text{C}^2\text{H}_3$ carbons in the solvent deuterioacetone. Each of the three deuterium nuclei splits the carbon resonance into a 1:1:1 triplet, corresponding to the three possible Zeeman states of an $I = 1$ nucleus. Thus the overall peak positions are the possible combinations of $\delta_{\text{C}} + (a \text{ or } 0 \text{ or } -a) + (a \text{ or } 0 \text{ or } -a) + (a \text{ or } 0 \text{ or } -a)$, where the coupling constant a is ~ 20 Hz.

With relatively simple spectra, it is usually possible to extract the individual coupling constants by inspection, and to pair them by size in order to discover what atoms they connect. However, the spectra of larger molecules present more of a challenge. The multiplets may overlap or be obscured by the presence of several unequal but similarly sized couplings. Also, if any chiral centres are present, then the two hydrogens in a methylene group may no longer have the same chemical shift, and in this case they will also show a mutual 2J coupling. Fortunately, several powerful aids exist to meet this challenge: decoupling and a range of multidimensional spectra.

B1.11.6.4 DECOUPLING

Several examples have already been given of resonances that merge because the underlying molecular or spin states interchange more rapidly than their frequency separation. Similar interchanges can also be imposed so as to remove a coupling in a controllable way. One irradiates a selected resonance so as to give its magnetization no preferred direction, within the timescale set by the coupling to be removed. This can be achieved selectively for any resonance removed from others by typically 20 Hz, although multiplets nearly as close as this will unavoidably be perturbed in a noticeable and predictable way by the irradiation process.

[Figure B1.11.10](#) offers an example. It shows the 400 MHz ^1H NMR spectrum of α -1-methylglucopyranose, below two further spectra where ^1H -decoupling has been applied at H-1 and H-4 respectively. The main results of the decouplings are arrowed. Irradiation at the chemical shift position of H-1 removes the smaller of the two couplings to H-2. This proves the saccharide to be in its α form, i.e. with $\phi \approx 60^\circ$ rather than 180° , according to the Karplus relationship given previously. Note that both the H-1 resonance and the overlapping solvent peak are almost totally suppressed by the saturation, caused by the decoupling irradiation. The H-4 resonance is a near-triplet, created by the two large and nearly equal couplings to H-3 and H-5. In both these cases, $\phi \approx 180^\circ$. The spectrum in this shift region is complicated by the methyl singlet and by some minor peaks from impurities. However, these do not affect the decoupling process, beyond being severely distorted by it. Genuine effects of decoupling are seen at the H-3 and the H-5 resonances only.

Single-frequency decoupling is easy and rapidly carried out. However, it may be limited by the closeness of different multiplets. Also, it will not normally be possible to apply more than one frequency of decoupling irradiation at a time. Fortunately, these disadvantages do not apply to the equivalent multidimensional methods.

It is also usually possible to remove all the couplings from a particular isotope, e.g. ^1H , provided that one only wishes to observe the spectrum from another isotope, e.g. ^{13}C . Either the decoupling frequency is noise-modulated to cover the relevant range of chemical shifts, or else the same decoupling is achieved more efficiently, and with less heating of the sample, by using a carefully designed, continuous sequence of composite pulses. [Figure B1.11.4](#) is a ^1H -decoupled ^{13}C spectrum: this is sometimes abbreviated to $^{13}\text{C}\{^1\text{H}\}$. In the absence of decoupling, the resonances would each be split by at least four short- and long-range couplings from ^1H atoms, and the signal-to-noise ratio would drop accordingly. The largest couplings would arise from the directly attached hydrogens. Thus, one might use a ^{13}C NMR spectrum without decoupling to distinguish CH_3 (broadened 1:3:3:1 quartets) from CH_2 (1:2:1 triplets) from CH (1:1 doublets) from C (singlets). However, more efficient methods are available for this, such as the DEPT and INEPT pulse sequences outlined below. Two-dimensional methods are also available if one needs to detect the connectivities revealed by the various ^{13}C - ^1H couplings.

B1.11.6.5 POLARIZATION TRANSFER

Couplings can also be exploited in a quite different way, that lies behind a wide range of valuable NMR techniques. If just one component peak in a multiplet X can be given a non-equilibrium intensity, e.g. by being selectively inverted, then this necessarily leads to large changes of intensity in the peaks of any other multiplet that is linked via couplings to X. [Figure B1.11.11](#) attempts to explain this surprising phenomenon. Part A shows the four possible combined spin states of a $^{13}\text{C}^1\text{H}$ molecular fragment, taken as an example. These are the same states as in [Figure B1.11.8](#), but attention is now drawn to the populations of the four spin states, each reduced by subtracting the 25% population that would exist at very low field, or alternatively at infinite temperature. The figures above each level are these relative differences, in convenient units. The intensity of any one transition, i.e. of the relevant peak in the doublet, is proportional to the difference of these differences, and is therefore proportionally relative to unity for any ^{13}C transition at Boltzmann equilibrium, and 4 for any ^1H transition.

The only alteration in part B is that the right hand ^1H transition has been altered so as to interchange the relevant populations. In practice, this might be achieved either by the use of a highly selective soft pulse, or by a more elaborate sequence of two pulses, spaced in time so as to allow the Larmor precessions of the two doublet components to dephase by 180° . The result is to produce a ^{13}C doublet whose relative peak intensities are 5: -3, in place of the original 1:1. Further, similar manipulations of the ^{13}C spins can follow, so as to re-invert the -3 component. This results in a doublet, or alternatively a singlet after decoupling, that has four times the intensity it had originally. The underlying physical process is called polarization transfer. More generally, any nucleus with magnetogyric ratio γ_A , coupled to one with γ_B , will have its intensity altered by γ_B/γ_A . This gain is valuable if $\gamma_B \gg \gamma_A$. The technique is then called INEPT (insensitive nucleus enhancement by polarization transfer) [19]. It has an added advantage. The relaxation rate of a high- γ nucleus is usually much less than that of a low- γ nucleus, because it has a bigger magnetic moment to interact with its surroundings. In the INEPT experiment, the spin populations of the low- γ nucleus are driven by the high- γ nucleus, rather than by natural relaxation. Hence the repeat time for accumulating the FID is shortened to that appropriate for the high- γ nucleus.

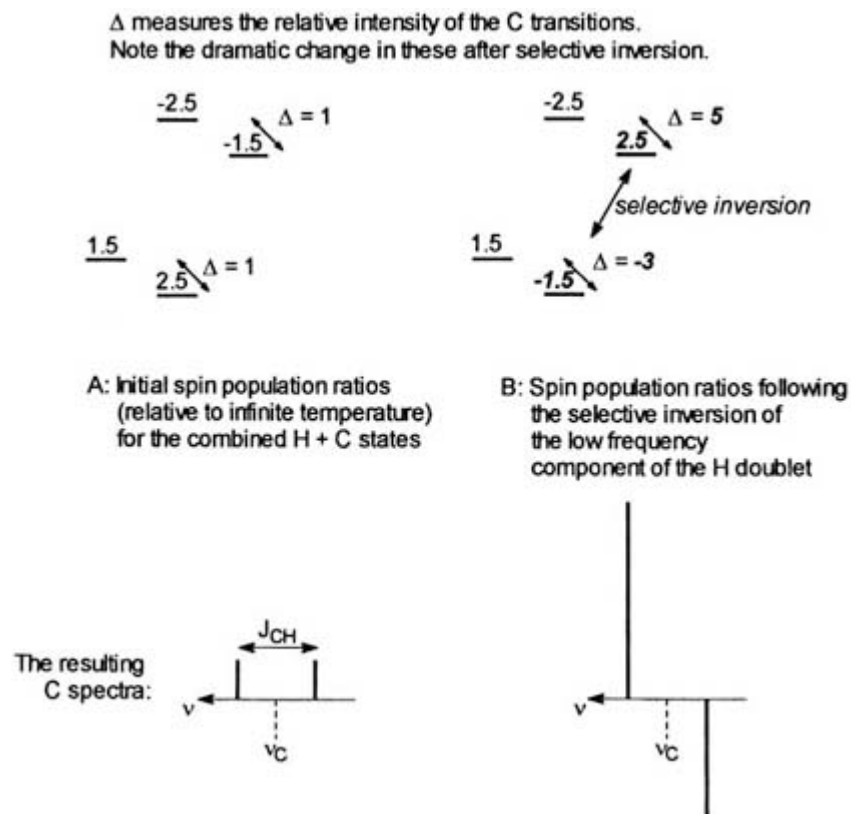


Figure B1.11.11. Polarization transfer from ^1H to ^{13}C (see the text). The inversion of one H transition also profoundly alters the C populations.

The same general methodology can also be applied to edit (for example) a decoupled ^{13}C NMR spectrum into four subspectra, for the CH_3 , CH_2 , CH and C moieties separately. A common variant method called DEPT (distortionless enhancement by polarization transfer) uses non-standard pulse angles, and is a rapid and reliable way for assigning spectra of medium complexity [20].

More generally, note that the application of almost any multiple pulse sequence, where at least two pulses are separated by a time comparable to the reciprocal of the coupling constants present, will lead to exchanges of intensity between multiplets. These exchanges are the physical method by which coupled spins are correlated in 2D NMR methods such as correlation spectroscopy (COSY) [21].

B1.11.7 TWO-DIMENSIONAL METHODS

The remarkable stability and controllability of NMR spectrometers permits not only the precise accumulation of FIDs over several hours, but also the acquisition of long series of spectra differing only in some stepped variable such as an interpulse delay. A peak at any one chemical shift will typically vary in intensity as this series is traversed. All the sinusoidal components of this variation with time can then be extracted, by Fourier transformation of the variations. For example, suppose that the normal 1D NMR acquisition sequence (relaxation delay, 90° pulse, collect FID) $_n$ is replaced by the 2D sequence (relaxation delay, 90° pulse, delay τ - 90° pulse, collect FID) $_n$ and that τ is increased linearly from a low value to create the second dimension. The polarization transfer process outlined in the previous section will then cause the peaks of one multiplet to be modulated in intensity, at the frequencies of any other multiplet with which it shares a coupling.

The resulting data set constitutes a rectangular or square array of data points, having a time axis in both dimensions. This is converted via Fourier transformation in both dimensions, giving the corresponding array of points in a 2D spectrum, with each axis being in δ units for convenience. These are most conveniently plotted as a contour map. In the above example the experiment is COSY-90, '90' referring to the second pulse, and the map should be at least approximately symmetrical about its diagonal. Any off-diagonal or 'cross' peaks then indicate the presence of couplings. Thus, a cross-peak with coordinates (δ_A, δ_B) indicates a coupling that connects the multiplets A and B. The spectrum is normally simplified by eliminating superfluous, mirror-image peaks, either with phase-cycled pulses and appropriate subtractions or by the use of carefully controlled, linear pulsed field gradients. No special equipment is needed in a modern spectrometer, although the data sets are typically 1 Mbyte or larger. The time requirement is only about 16 times that for a 1D spectrum, in favourable cases, and may be less if pulsed field gradients are used.

B1.11.7.1 HOMONUCLEAR COSY SPECTRA

Figure B1.11.12 shows the 2D COSY-45 contour plot of the same α -1-methylglucopyranose compound as in a previous figure. The corresponding 1D spectrum, plotted directly above, is in fact the projection of the 2D spectrum onto the horizontal shift axis. The hydrogen assignments are added underneath the multiplets. The analysis of such a spectrum begins by selecting one peak, identifiable either by its distinctive chemical shift or by mapping its coupling pattern onto the expected pattern of molecular connectivity. Here, the doublet at 4.73 ppm uniquely has the high shift expected when a CH group bears two O substituents. The coupling pattern can now be identified by noting the horizontal and vertical alignments of all the strongest diagonal and off-diagonal resonances. These alignments must be exact, within the limits of the digitization, as the COSY process does not cause any shifts. The H-1 to H-2 correlation exemplifies this precision, in that the cross-peak whose centre has coordinates at δ (3.47, 4.73) is precisely aligned with the centres of the H-2 and H-1 multiplets respectively. The same principle allows one to see that the cross-peak at δ (3.47, 3.59) arises from the (H-2 to H-3) coupling, whereas the more complex multiplet to the right of it must come from the overlap of the (H-3 to H-4) and the (H-5 to H-4) cross-peaks, respectively at δ (3.59, 3.33) and δ (3.56, 3.33). In this way it is possible to distinguish the H-3 and H-5 multiplets, respectively at 3.59 and 3.56 ppm, even though they overlap in the 1D spectrum. This illustrates the power of multidimensional NMR to separate overlapping resonances. In a similar way, the 2D spectrum makes it clear that the OCH_3 resonance at 3.35 ppm has no coupling connection with any other hydrogen in the saccharide, so that its shift overlap with H-4 is purely accidental.

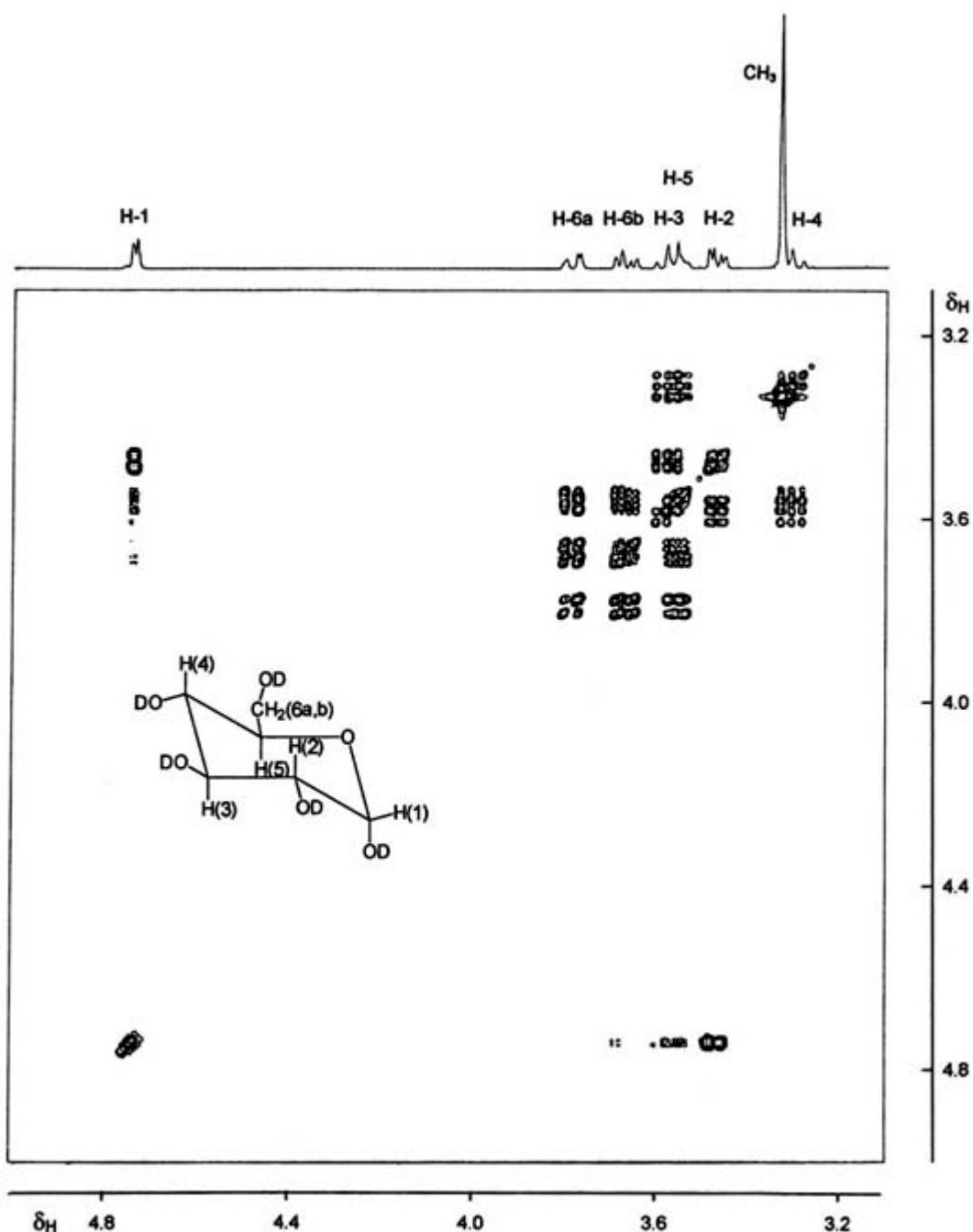


Figure B1.11.12. ^1H - ^1H COSY-45 2D NMR spectrum of methyl- α -glucopyranose (structure shown). The coupling links and the approximate couplings can be deduced by inspection.

Several other features of the figure merit attention. The two hydroxymethyl resonances, H-6a at 3.79 ppm and H-6b at 3.68 ppm, are separated in shift because of the chiral centres present, especially the nearby one at C-5. (One should note that a chiral centre will always break the symmetry of a molecule, just as an otherwise symmetrical coffee mug loses its symmetry whilst grasped by a right hand.) H-5 is thus distinctive in showing three coupling connections, and could be assigned by this alone, in the absence of other information. Also, the (H-6a to H-6b) cross-peak gives the general impression of a tilt parallel to the diagonal, whereas some of the other cross-peaks show the opposite tilt. This appearance of a tilt in the more complex cross-peaks is the deliberate consequence of completing the COSY pulse sequence with a 45° rather than a 90° pulse. It actually arises from selective changes of intensity in the component peaks of the off-diagonal multiplet. The 'parallel'

pattern arises from negative values of the coupling constant responsible for the cross-peak, i.e. the ‘active’ coupling. It therefore usually shows that the active coupling is of 2J or 4J type, whereas an ‘antiparallel’ pattern usually arises from a 3J active coupling.

Another advantage of a COSY spectrum is that it can yield cross-peaks even when the active coupling is not fully resolved. This can be useful for assigning methyl singlets, for example in steroids, and also for exploiting couplings comparable with the troublesomely large linewidths in protein spectra, for example, or the $^{11}\text{B}\{^1\text{H}\}$ spectra of boranes. However, it does also mean that longer-range couplings may appear unexpectedly, such as the weak (H-1 to H-3), (H-1 to H-6b) and (H-6a to H-4) couplings in [figure B1.11.12](#). Their appearance can yield stereochemical information such as the existence of bonding pathways having appropriate conformations, and they are usually recognizable by their comparatively weak intensities.

Many variations of the basic homonuclear COSY experiment have been devised to extend its range. A brief guide to some classes of experiment follows, along with a few of the common acronyms.

- (i) Experiments using pulsed field gradients [23]. These can be very rapid if concentration is not a limiting factor, but they require added equipment and software. Acronym: a prefix of ‘g’.
- (ii) Multiple quantum methods. These employ more complex pulse sequences, with the aim of suppressing strong but uninteresting resonances such as methyl singlets. Acronyms include the letters MQ, DQ, etc. They are also useful in isotope-selective experiments, such as INADEQUATE [24]. For example, the ^{13}C – ^{13}C INADEQUATE experiment amounts to a homonuclear ^{13}C COSY experiment, with suppression of the strong singlets from uncoupled carbons.
- (iii) Extended or total correlation methods. These reveal linked couplings, involving up to every atom in a group, where at least one reasonably large coupling exists for any one member atom. They are valuable for identifying molecular moieties such as individual amino acids in a peptide, for they can remove ambiguities arising from peak overlaps. Acronym: TOCSY (total correlation spectroscopy).
- (iv) Correlations exploiting nuclear Overhauser enhancements (NOEs) in place of couplings, such as NOESY. These are discussed in [section B1.11.8](#).
- (v) Correlations that emphasize smaller couplings. These are simply achieved by the addition of appropriate, fixed delays in the pulse sequence.

Homonuclear techniques such as J -resolved spectroscopy also exist for rotating all multiplets through 90° , to resolve overlaps and also give a 1D spectrum from which all homonuclear couplings have been removed [26].

-27-

The theory of these and other multidimensional NMR methods requires more than can be visualized using magnetization vectors. The spin populations must be expressed as the diagonal elements of a density matrix, which not only has off-diagonal elements indicating ordinary, single-quantum transitions such as those described earlier, but also other off-diagonal elements corresponding to multiple-quantum transitions or ‘coherences’. The pulses and time developments must be treated as operators. The full theory also allows experiments to be designed so as to minimize artefacts [27, 28 and 29].

B1.11.7.2 HETERONUCLEAR CORRELATION SPECTRA

Similar experiments exist to correlate the resonances of different types of nucleus, e.g. ^{13}C with ^1H , provided that some suitable couplings are present, such as $^1J_{\text{CH}}$. It is necessary to apply pulses at both the relevant frequencies and it is also desirable to be able to detect either nucleus, to resolve different peak clusters. Detection through the nucleus with the higher frequency is usually called reverse-mode detection and generally gives better sensitivity. The spectrum will have the two different chemical shift scales along its axes

and therefore will not be symmetrical.

A ^1H (detected)- ^{13}C shift correlation spectrum (common acronym HMQC, for heteronuclear multiple quantum coherence, but sometimes also called COSY) is a rapid way to assign peaks from protonated carbons, once the hydrogen peaks are identified. With changes in pulse timings, this can also become the HMBC (heteronuclear multiple bond connectivity) experiment, where the correlations are made via the smaller $^2J_{\text{CH}}$ and $^3J_{\text{CH}}$ couplings. This helps to assign quaternary carbons and also to identify coupling and, hence, chemical links, where H-H couplings are not available. Similar experiments exist for almost any useful pairing of nuclei: those to ^{15}N are particularly useful in the spectra of suitably labelled peptides. [Figure B1.11.13](#) shows a simple ^{13}C - ^1H shift correlation spectrum of the same saccharide as in the previous two figures, made by exploiting the $^1J_{\text{CH}}$ couplings. The detection was via ^{13}C and so the spectrum has good resolution in the ^{13}C dimension, here plotted as the horizontal axis, but rather basic resolution in the vertical ^1H dimension. The ^1H resolution is nonetheless sufficient to show that C-6 is a single type of carbon attached to two distinct types of hydrogen. Inspection of the ^1H axis shows that the approximate centre of each 2D multiplet matches the shifts in the previous spectra. This affords an unambiguous assignment of the carbons.

Even more complex sequences can be applied for the simultaneous correlation of, for example, ^{15}N shifts with ^1H - ^1H COSY spectra. Because the correlation of three different resonances is highly specific, the chance of an accidental overlap of peaks is greatly reduced, so that very complex molecules can be assigned. Furthermore, the method once again links atoms where no simple H-H couplings are available, such as across a peptide bond. Such 3D NMR methods generate enormous data sets and, hence, very long accumulation times, but they allow the investigation of a wide range of labelled biomolecules.

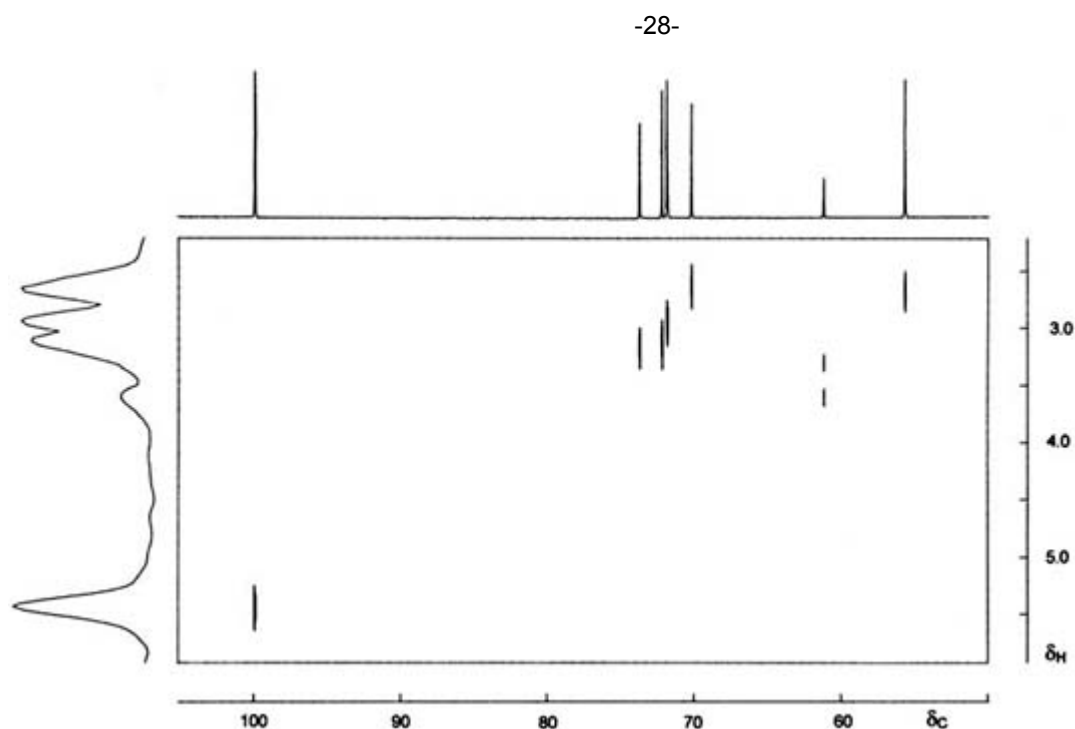


Figure B1.11.13. ^{13}C - ^1H shift correlation via $^1J_{\text{CH}}$. This spectrum of methyl- α -glucopyranose (structure as in [figure B1.11.12](#)) permits unambiguous ^{13}C assignments once the ^1H assignments have been determined as in [figure B1.11.12](#).

B1.11.8 SPATIAL CORRELATIONS

J couplings are not the only means in NMR for showing that two atoms lie close together. Pairs of atoms, particularly ^1H , also affect each other through their dipolar couplings. Even though the dipolar splittings are averaged away when a molecule rotates isotropically, the underlying magnetic interactions are still the major contributors to the ^1H T_1 . Each pairwise interaction of a particular H_A with any other H_B , separated by a distance r_{AB} in a rigid molecule, will contribute to $1/T_{1A}$ in proportion to $(r_{AB})^{-6}$.

Although it is possible to detect this interaction by the careful measurement of all the T_1 s in the molecule, it may be detected far more readily and selectively, *via* the mutual NOE. When a single multiplet in a ^1H NMR spectrum is selectively saturated, the resulting input of energy leaks to other nuclei, through all the dipolar couplings that are present in the molecule [29]. This alters the intensities and likewise the integrals of their resonances. The A spin of an isolated pair of spins, A and B, in a molecule tumbling fairly rapidly, will have its intensity increased by the multiplicative factor $1 + \gamma_B/2\gamma_A$. This amounts to a gain of 50% when A and B are both ^1H , although it will normally be less in practice, because of the competing interactions of other spins and other relaxation mechanisms. If B is ^1H and A is ^{13}C , the corresponding enhancement is 299%. In this case there need be no competition, for all the ^1H resonances can be saturated simultaneously, using the techniques of broadband decoupling. Also, dipolar coupling usually dominates the relaxation of all carbons bearing hydrogens. The threefold intensity gain is a valuable bonus in $^{13}\text{C}\{^1\text{H}\}$ NMR.

-29-

B1.11.8.1 NOE-DIFFERENCE SPECTRA

Even if the intensity changes are only of the order of 1%, they may nevertheless be reliably detected using difference spectra [30]. Figure B1.11.14 shows an NOE-difference ^1H NMR spectrum of slightly impure aspirin, over a normal spectrum of the same sample. The difference spectrum was obtained by gently saturating the methyl singlet at 2.3 ppm, for a period of 2 s just before collecting the spectrum, and then by precisely subtracting the corresponding spectrum acquired without this pre-irradiation. As the pre-irradiation selectively whitewashes the methyl peak, the result of the subtraction is to create a downwards-going resultant peak. The other peaks subtract away to zero, unless the energy leakage from the pre-irradiation has altered their intensity. In the figure, the four aromatic H resonances to the left subtract to zero, apart from minor errors arising from slight instabilities in the spectrometer. In contrast, the broad OH resonance at 5 ppm in the upper spectrum proves that this hydrogen lies close to the CH_3 group. Thus, aspirin must largely possess the conformation shown on the right. If the left-hand conformation had been significant, then the aromatic H at 7.2 ppm would have received a significant enhancement.

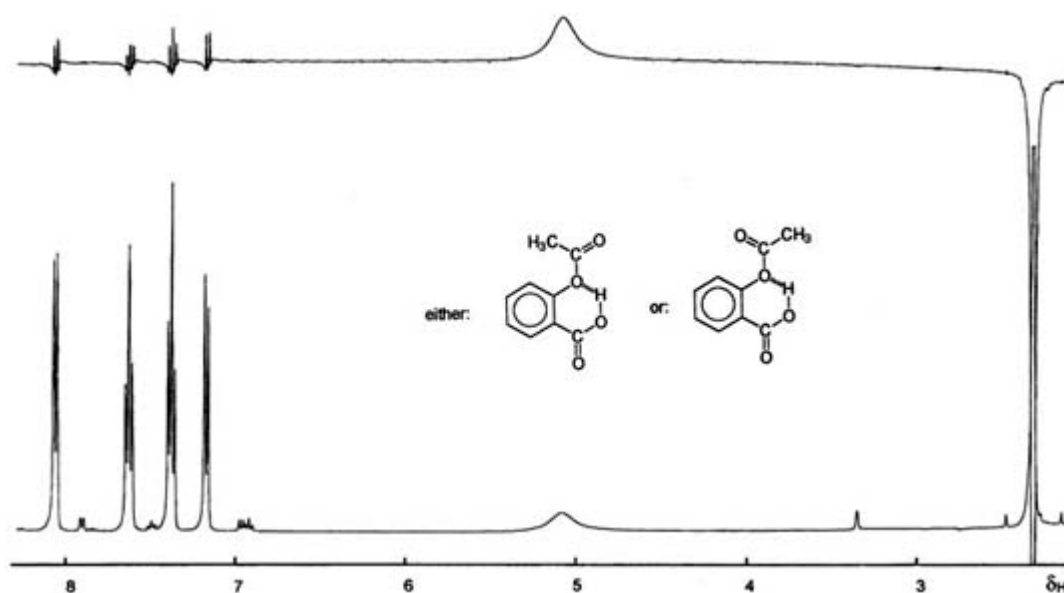


Figure B1.11.14. ^1H NOE-difference spectrum (see the text) of aspirin, with pre-saturation at the methyl resonance, proving that the right-hand conformation is dominant.

NOE-difference spectroscopy is particularly valuable for distinguishing stereoisomers, for it relies solely on internuclear distances, and thus avoids any problems of ambiguity or absence associated with couplings. With smallish molecules, it is best carried out in the above 1D manner, because ~ 2 s are necessary for the transmission of the NOE. The transmission process becomes more efficient with large molecules and is almost optimal for proteins. However, problems can occur with molecules of intermediate size [31]. A 2D version of the NOE-difference experiment exists, called NOESY.

-30-

B1.11.8.2 PROTEIN STRUCTURES

If multidimensional spectra of both COSY and NOESY types can be obtained for a protein, or any comparable structured macromolecule, and if a reasonably complete assignment of the resonances is achieved, then the NOESY data can be used to determine its structure [31, 32 and 33]. Typically, several hundred approximate H–H distances will be found via the NOESY spectrum of a globular protein having mass around 20 000 Da. These can then be used as constraints in a molecular modelling calculation. The resulting structures can compare in quality with those from x-ray crystallography, but do not require the preparation of crystals. Related spectra can also elucidate the internal flexibility of proteins, their folding pathways and their modes of interaction with other molecules. Such information is vital to the pharmaceutical industry in the search for new drugs. It also underpins much biochemistry.

REFERENCES

- [1] Harris R K 1996 Nuclear spin properties and notation *Encyclopedia of NMR* vol 5, ed D M Grant and R K Harris (Chichester: Wiley) pp 3301–14
- [2] Dorn H C 1984 ^1H NMR—a new detector for liquid chromatography *Anal. Chem.* **56** 747A–58A
- [3] Morris K F and Johnson C S Jr 1993 Resolution of discrete and continuous molecular size distributions by means of diffusion-ordered 2D NMR spectroscopy *J. Am. Chem. Soc.* **115** 4291–9
- [4] Quin L D and Verkade J G (eds) 1994 *Phosphorus-31 NMR Spectral Properties in Compound Characterization and Structural Analysis* (New York: VCH)
- [5] Overhauser A W 1953 Polarization of nuclei in metals *Phys. Rev.* **92** 411–15
- [6] Howarth O W, Pettersson L and Andersson I 1989 Monomolybdonavanadate and *cis*- and *trans*-dimolybdo-octavanadate *J. Chem. Soc. Dalton Trans.* 1915–23
- [7] Lamb W 1941 Internal diamagnetic fields *Phys. Rev.* **60** 817–19
- [8] Lynden-Bell R M and Harris R K 1969 *Nuclear Magnetic Resonance Spectroscopy* (London: Nelson) pp 81–3
- [9] Mason J (ed) 1987 *Multinuclear NMR* (New York: Plenum) ch 7–21
- [10] Tonelli A S 1996 *Polymer Spectroscopy* ed A Fawcett (Chichester: Wiley) ch 2
- [11] Martin M M, Martin G J and Delpuech J-J (eds) 1980 Use of chemicals as NMR auxiliary reagents *Practical NMR Spectroscopy* (London: Heyden) ch 10
- [12] Williams D H and Fleming I 1995 *Spectroscopic Methods in Organic Chemistry* (London: McGraw-Hill) ch 3
- [13] Breitmaier E and Voelter W 1986 *Carbon-13 NMR Spectroscopy: High Resolution Methods and Applications in Organic Chemistry* (New York: VCH)
- [14] Pretsch E, Clerc T, Seibl J and Simon W 1983 *Tables of Spectral Data for Structural Determination of Organic Compounds* Engl. edn. (Berlin: Springer)

- [15] Silverstein R M, Bassler G C and Morrill T C 1981 *Spectrometric Identification of Organic Compounds* (New York: Wiley) ch 4 and 5
- [16] Jameson C J and Mason J 1987 The chemical shift *Multinuclear NMR* ed J Mason (New York: Plenum) ch 3
- [17] Karplus M 1959 Contact electron spin coupling of nuclear magnetic moments *J. Chem. Phys.* **30** 11–15
-

-31-

- [18] Venanzi T J 1982 Nuclear magnetic resonance coupling constants and electronic structure in molecules *J. Chem. Educ.* **59** 144–8
- [19] Morris G A and Freeman R 1979 Enhancement of nuclear magnetic resonance signals by polarization transfer *J. Am. Chem. Soc.* **101** 760–2
- [20] Doddrell D M, Pegg D T and Bendall M R 1982 Distortionless enhancement of NMR signals by polarization transfer *J. Magn. Reson.* **48** 323–7
- [21] Bax A and Freeman R 1981 Investigation of complex networks of spin–spin coupling by two-dimensional NMR *J. Magn. Reson.* **44** 542–61
- [22] Hurd R E 1990 Gradient enhanced spectroscopy *J. Magn. Reson.* **87** 422–8
- [23] Bax A, Freeman R and Kempell S P 1980 Natural abundance ^{13}C – ^{13}C coupling observed via double quantum coherence *J. Am. Chem. Soc.* **102** 4849–51
- [24] Braunschweiler L and Ernst R R 1983 Coherence transfer by isotropic mixing: application to proton correlation spectroscopy *J. Magn. Reson.* **53** 521–8
- [25] Aue W P, Kharran J and Ernst R R 1976 Homonuclear broadband decoupling and two-dimensional J-resolved NMR spectroscopy *J. Chem. Phys.* **64** 4226–7
- [26] Ernst R R, Bodenhausen G and Wokaun A 1987 *Principles of Nuclear Magnetic Resonance in One and Two Dimensions* (Oxford: Clarendon)
- [27] Bax A 1982 *2-Dimensional Nuclear Magnetic Resonance in Liquids* (Delft: Delft University Press)
- [28] Freeman R 1997 *Spin Choreography* (Oxford: Spektrum) ch 3
- [29] Neuhaus D and Williamson M 1989 *The Nuclear Overhauser Effect in Structural and Conformational Analysis* (New York: VCH)
- [30] Bax A and Davis D G 1985 Practical aspects of two-dimensional transverse NOE spectroscopy *J. Magn. Reson.* **63** 207–13
- [31] Wüthrich K 1996 Biological macromolecules: structural determination in solution *Encyclopedia of NMR* vol 2, ed D M Grant and R K Harris (Chichester: Wiley) pp 932–9
- [32] Markley J R and Opella S J (eds) 1997 *Biological NMR Spectroscopy* (Oxford: Oxford University Press)
- [33] Oschkinat H, Müller T and Dieckmann T 1994 Protein structure determination with three- and four-dimensional spectroscopy *Angew. Chem. Int. Ed. Engl.* **33** 277–93
-

FURTHER READING

Abraham R J, Fisher J and Loftus P 1988 *Introduction to NMR Spectroscopy* (Chichester: Wiley)

A first text that concentrates on ^1H and ^{13}C NMR. Level suitable for undergraduates, although complete beginners might need more help.

Williams D H and Fleming I 1995 *Spectroscopic Methods in Organic Chemistry* (London: McGraw-Hill)

Includes basic interpretation, 30 valuable tables of data, and a concise introduction to multidimensional NMR spectra from an interpretational point of view.

Mason J (ed) 1987 *Multinuclear NMR* (New York: Plenum)

The most recent comprehensive text concentrating on the entire Periodic Table. Individual elements are also covered from time to time in monographs and reviews, e.g. in *Progress in NMR Spectroscopy*.

Braun S, Kalinowski H-O and Berger S 1998 *150 and More Basic NMR Experiments* (Weinheim: VCH)

A fairly straightforward 'how to do it' book for the practising spectroscopist.

Sanders J K M and Hunter B J 1993 *Modern NMR Spectroscopy: a Guide for Chemists* (Oxford: Oxford University Press)

An informative second-level text with a practical flavour.

Martin G E and Zekster A S 1988 *Two-dimensional NMR Methods for Establishing Molecular Connectivity* (Weinheim: VCH)

Contains a wealth of practical experience.

Hoch J C and Stern A S 1996 *NMR Data Processing* (New York: Wiley-Liss)

Careful treatment of what happens to the data after they have been collected.

Ernst R R, Bodenhausen G and Wokaun A 1987 *Principles of Nuclear Magnetic Resonance in One and Two Dimensions* (Oxford: Clarendon)

Authoritative at a high theoretical level.

Freeman R 1996 *Spin Choreography* (Oxford: Spektrum)

Exceptionally elegant exposition of a wide range of advanced but much-used NMR concepts.

Grant D M and Harris R K (eds) 1996 *Encyclopedia of NMR* (Chichester: Wiley)

Very comprehensive eight-volume coverage.

B1.12 NMR of solids

R Dupree and M E Smith

B1.12.1 INTRODUCTION

Solid-state NMR has long been used by physicists to study a wide range of problems such as superconductivity, magnetism, the electronic properties of metals and semiconductors, ionic motion etc. The early experiments mostly used 'wide line' NMR where high resolution was not required but with the development of the technique, particularly the improvements in resolution and sensitivity brought about by magic angle spinning (B1.12.4.3), and decoupling and cross polarization (B1.12.4.4), solid-state NMR has become much more widely used throughout the physical and, most recently, biological sciences. Although organic polymers were the first major widespread application of high-resolution solid-state NMR, it has found application to many other types of materials, from inorganics such as aluminosilicate microporous materials, minerals and glasses to biomembranes. Solid-state NMR has become increasingly multinuclear and the utility of the technique is evidenced by the steady and continued increase in papers that use the technique to characterize materials. There is no doubt that the solid-state NMR spectrometer has become a central piece of equipment in the modern materials physics and chemistry laboratories.

The principal difference from liquid-state NMR is that the interactions which are averaged by molecular motion on the NMR timescale in liquids lead, because of their anisotropic nature, to much wider lines in solids. Extra information is, in principle, available but is often masked by the lower resolution. Thus, many of the techniques developed for liquid-state NMR are not currently feasible in the solid state. Furthermore, the increased linewidth and the methods used to achieve high resolution put more demands on the spectrometer. Nevertheless, the field of solid-state NMR is advancing rapidly, with a steady stream of new experiments forthcoming.

This chapter summarizes the interactions that affect the spectrum, describes the type of equipment needed and the performance that is required for specific experiments. As well as describing the basic experiments used in solid-state NMR, and the more advanced techniques used for distance measurement and correlation, some emphasis is given to nuclei with spin $I > \frac{1}{2}$ since the study of these is most different from liquid-state NMR.

B1.12.2 FUNDAMENTALS

B1.12.2.1 INTERACTION WITH EXTERNAL MAGNETIC FIELDS

NMR is accurately described as a coherent radiofrequency (RF) spectroscopy of the nuclear magnetic energy levels. The physical basis of the technique is the lifting of the degeneracy of the different m_z nuclear spin states through interaction with an external magnetic field, creating a set of energy levels. The total energy separation between these levels is determined by a whole range of interactions. The nuclear spin Hamiltonian has parts corresponding to the experimental conditions, which are termed external, and parts that result from the sample itself, which are called internal. The internal part provides information about the physical and electronic structure of the sample.

-2-

The total interaction energy of the nucleus may be expressed as a sum of the individual Hamiltonians given in equation B1.12.1, (listed in table B1.12.1) and are discussed in detail in several excellent books [1, 2, 3 and 4].

$$H^{\text{TOT}} = H^Z + H^{\text{RF}} + H^{\text{D}} + H^{\text{CS}} + H^{\text{K}} + H^{\text{J}} + H^{\text{P}} + H^{\text{Q(1)}} + H^{\text{Q(2)}} + \dots \quad (\text{B1.12.1})$$

The large static applied magnetic field (B_0) produces the Zeeman interaction ($= -\hbar\omega_0 I_z$, where I_z is the z -component of I (the nuclear spin) with eigenvalues m_z ($-I \leq m_z \leq I$), figure B1.12.1(a) with the nuclear magnetic dipole moment μ ($= \gamma\hbar I$, where γ is the gyromagnetic ratio of the nucleus). The B_0 field is taken to define the z -axis in the laboratory frame and gives an interaction energy of

$$H^Z = \mu B_0 = -\gamma\hbar B_0 m_z = -\hbar\omega_0 m_z \quad (\text{B1.12.2})$$

where ω_0 is the Larmor frequency in angular frequency. In this chapter only the high field limit is considered whereby the nuclear spin states are well described by the Zeeman energy levels, and all the other interactions can be regarded as perturbations of these spin states. Any nucleus that possesses a magnetic moment is technically accessible to study by NMR, thus only argon and cerium of the stable elements are excluded as they possess only even-even isotopes.

Table B1.12.1 Summary of main interactions important to NMR.

H^m	Interaction	T_{ij}	Typical size A_i (Hz)	Comments
H^Z	Zeeman	Unitary B_0	10^7-10^9	Interaction with main magnetic field
H^{RF}	RF	Unitary B_1	10^3-10^5	Interaction with RF field
H^D	Dipolar	D	I 10^3-10^4	Through space spin-spin interaction, axially symmetric traceless tensor
H^{CS}	Chemical shielding	σ	B_0 10^2-10^5	Alteration of the magnetic field by the electrons
H^J	Indirect spin	J	I $1-10^3$	Spin-spin interaction mediated via the bonding electrons through the contact interaction
H^P	Paramagnetic		S 10^2-10^5	Interaction with isolated unpaired electrons
H^K	Knight shift	K	S 10^2-10^5	Interaction with conduction electrons via the contact interaction
H^Q	Quadrupolar	Eq	I 10^3-10^7	Interaction of nuclear quadrupolar moment with the electric field gradient (q) and I twice to produce effectively an I^2 interaction

-3-

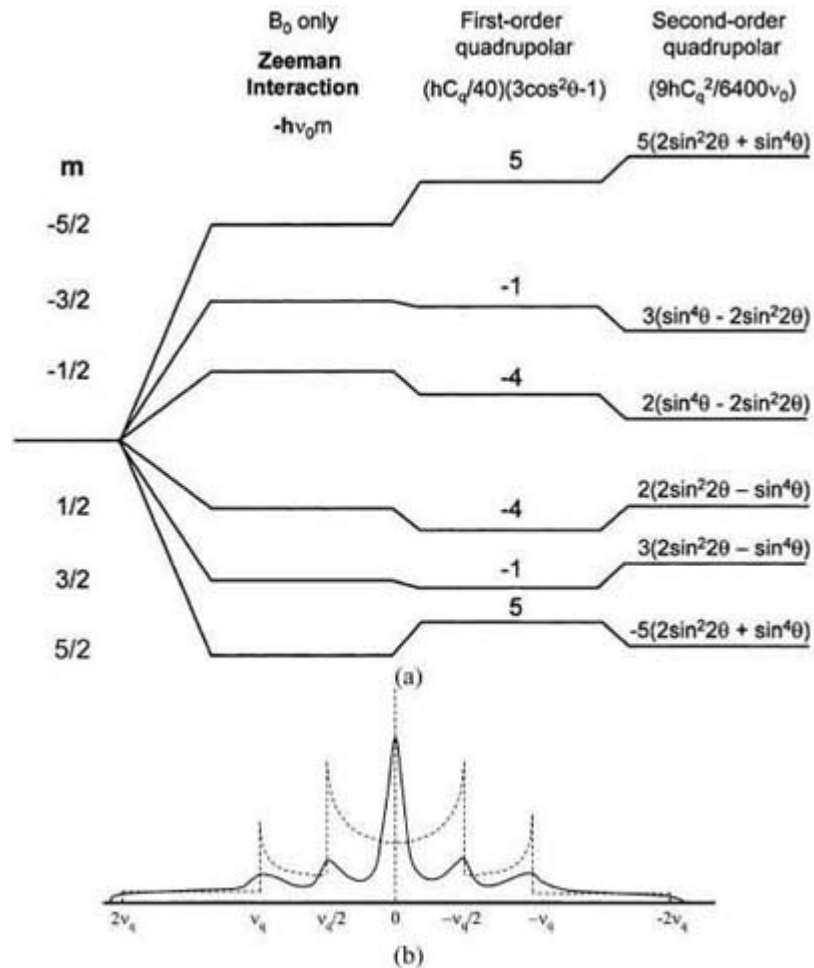


Figure continued on next page

-4-

Figure B1.12.1. (a) Energy level diagram for an $I = \frac{1}{2}$ nucleus showing the effects of the Zeeman interaction and first- and second-order quadrupolar effect. The resulting spectra show static powder spectra for (b) first-order perturbation for all transitions and (c) second-order broadening of the central transition. (d) The MAS spectrum for the central transition.

-5-

For a spin- $\frac{1}{2}$ nucleus the two m_z values $\pm\frac{1}{2}$ have energies of $\pm\gamma\hbar B_0/2$ giving an energy separation of $\gamma\hbar B_0$. Thermal equilibrium produces a Boltzmann distribution between these energy levels and produces the bulk nuclear magnetization of the sample through the excess population which for a sample containing a total of N spins is $\sim N\gamma\hbar B_0/2kT$. For example for ^{29}Si in an applied magnetic field of 8.45 T the excess in the populations at room temperature is only 1 in 10^5 so that only a small number of the total number of spins contribute to the signal, and this is possibly the greatest weakness of NMR. As the magnetization is directly proportional to the applied magnetic field, sensitivity provides one of the drives to higher applied magnetic fields. The dependence on γ explains why some nuclei are more favoured as the ease with which a nucleus can be observed depends upon the receptivity (R_x)

$$R_x = \gamma^3 C_x I_x (I_x + 1) \quad (\text{B1.12.3})$$

where C_x is the natural abundance of the isotope being considered [3].

In spectroscopy involving electromagnetic (em) radiation both spontaneous and stimulated events can occur. NMR is a relatively low-frequency em spectroscopy so that spontaneous events are very unlikely ($\sim 10^{-22} \text{ s}^{-1}$) and stimulated events therefore dominate. This means that NMR is a coherent spectroscopy so that the excess in population can be made to work in a concerted way. The NMR experiment involves measurement of the energy level separation by application of a time-varying magnetic field \mathbf{B}_1 orthogonal to \mathbf{B}_0 . \mathbf{B}_1 excites transitions (through I_+ and I_- , the conventional spin raising and lowering operators) when its frequency (ω) is close to ω_0 , typically in the RF region of 10 MHz to 1 GHz. The Hamiltonian for this interaction is

$$H^{\text{RF}} = (-\gamma\hbar B_1/2)(I_+ e^{-i\omega t} + I_- e^{+i\omega t}). \quad (\text{B1.12.4})$$

B1.12.2.2 LOCAL INTERACTIONS

In diamagnetic insulating solids spin- $\frac{1}{2}$ nuclei experience a range of interactions that include magnetic dipolar (H^{D}) interaction through space with nearby nuclear magnetic moments, chemical shielding (H^{CS}), modification of the magnetic field at the nucleus due to the surrounding electrons and indirect spin-spin coupling (H^{J})—interaction of magnetic moments mediated by intermediate electron spins. In materials that contain paramagnetic centres the unpaired electrons can interact strongly with the nuclei (H^{P}) and possibly cause very large shifts and severe broadening of the NMR signal. The fluctuating magnetic fields produced by the electron spins can produce very efficient relaxation. Hence, for solids where the nuclei are slowly relaxing and will dissolve paramagnetic ions, small amounts ($\sim 0.1 \text{ mol}\%$) are added to aid relaxation. In materials

containing conduction electrons these can also interact strongly with the nuclear spin via a contact interaction H^K that produces relaxation and a change in resonance position termed the Knight shift, both of which provide important information on the nature of the density of states at the Fermi surface. Nuclei with spin $I > \frac{1}{2}$ are also affected by the electric quadrupole interaction (H^Q), an interaction between the nuclear electric quadrupole moment and the gradient in the electric field at the nucleus. Although this is an electrical interaction it depends on the magnetic quantum number (m_z) so affects the NMR spectrum [2]. The background of the quadrupole interaction is given in the classic paper by Cohen and Reif [5].

All interactions associated with NMR can be expressed as tensors and may be represented by a general expression

-6-

$$H^m = k^m I_i T_{ij}^m A_j^m \quad (\text{B1.12.5})$$

where H^m is one of the component Hamiltonians in equation B1.12.1. For each interaction there is a constant k , a 3×3 second-rank tensor T_{ij} , and then another vector quantity that the spin (I) interacts with: this is either a field or a spin. Normally three numbers are needed to describe a 3×3 tensor relating two vectors and these numbers are usually the isotropic value, the anisotropy and the asymmetry [6]. Their exact definition can vary even though there are conventions that are normally used, so that any paper should be examined carefully to see how the quantities are being defined. Note that the chemical shielding is fundamentally described by the tensor σ although in experiment data it is the chemical shift δ that is normally reported. The shielding and the shift are related by $\delta = 1 - \sigma$ [3, 6].

In a typical isotropic powder the random distribution of particle orientations means the principal axes systems (where the tensor only has diagonal elements) will be randomly distributed in space. In the presence of a large magnetic field, this random distribution gives rise to broadening of the NMR spectrum since the exact resonance frequency of each crystallite will depend on its orientation relative to the main magnetic field.

Fortunately, to first order, all these interactions have similar angular dependences of $(3 \cos^2 \theta - 1 + \eta \sin^2 \theta \cos^2 \varphi)$ where η is the asymmetry parameter of the interaction tensor ($\eta = 0$ for axial symmetry). Lineshapes can provide very important information constraining the local symmetry of the interaction that can often be related to some local structural symmetry.

Of the NMR-active nuclei around three-quarters have $I \geq 1$ so that the quadrupole interaction can affect their spectra. The quadrupole interaction can be significant relative to the Zeeman splitting. The splitting of the energy levels by the quadrupole interaction alone gives rise to pure nuclear quadrupole resonance (NQR) spectroscopy. This chapter will only deal with the case when the quadrupole interaction can be regarded as simply a perturbation of the Zeeman levels.

The electric field gradient is again a tensor interaction that, in its principal axis system (PAS), is described by the three components $V_{x'x'}$, $V_{y'y'}$ and $V_{z'z'}$, where ' indicates that the axes are not necessarily coincident with the laboratory axes defined by the magnetic field. Although the tensor is completely defined by these components it is conventional to recast these into 'the electric field gradient' $eq = V_{z'z'}$, the largest component, and the asymmetry parameter $\eta_Q = |V_{y'y'} - V_{x'x'}|/V_{z'z'}$. The electric field gradient is set up by the charge distribution outside the ion (e.g. Al^{3+}) but the initially spherical charge distribution of inner shells of electrons of an ion will themselves become polarized by the presence of the electrical field gradient to lower their energy in the electric field. This polarization produces an electric field gradient at the nucleus itself of $eq_n = eq(1 - \gamma_\infty)$ where $(1 - \gamma_\infty)$ is the Sternheimer antishielding factor which is a measure of the magnification of eq due to this polarization of the core electron shells [7]. Full energy band structure calculations of electric field gradients show how important the contribution of the electrons on the ion itself are compared to the lattice. The quadrupole Hamiltonian (considering axial symmetry for simplicity) in the laboratory frame, with

θ the angle between the z' -axis of the quadrupole PAS and B_0 , is

$$H^Q = (hC_Q/(8I(2I-1)))[(3\cos^2\theta-1)(3I_z-a)+3\sin\theta\cos\theta(I_+I_+I_-) + (I_+I_-I_z)+3\sin^2\theta/2(I_+^2+I_-^2)] \quad (\text{B1.12.6})$$

where $C_Q = (e^2qQ/h)(1-\gamma_\infty)$ and $a = I(I+1)$. In the limit $H^Z \gg H^Q$ a standard perturbation expansion using the eigenstates of H^Z is applicable. The first-order term splits the spectrum into $2I$ components ([figure B1.12.1 \(a\)](#)) of intensity $|\langle m-1|I_-|m\rangle|^2$ ($\propto a - m(m-1)$) at frequency

-7-

$$\nu_m^{(1)} = (3C_Q/4I(2I-1))(3\cos^2\theta-1)(m_z - 1/2). \quad (\text{B1.12.7})$$

This perturbation can cause the non-central transitions (i.e. $m_z \neq \frac{1}{2}$) to be shifted ([figure B1.12.1\(b\)](#)) sufficiently far from the Larmor frequency such that these transitions become difficult to observe with conventional pulse techniques. This is particularly important for spin-1 nuclei (of which the most important ones are ^2H and ^{14}N) as there is no central transition ($m_z = \frac{1}{2}$) and all transitions are broadened to first order.

Fortunately, for non-integer quadrupolar nuclei for the central transition $\nu_m^{(1)} = 0$ and the dominant perturbation is second order only (equation B1.12.8) which gives a characteristic lineshape ([figure B1.12.1\(c\)](#)) for axial symmetry):

$$\nu_m^{(2)} = (-9C_Q^2/64\nu_0I^2(2I-1)^2)(a-3/4)(1-\cos^2\theta)(9\cos^2\theta-1). \quad (\text{B1.12.8})$$

This angular dependence is different from the first-order perturbations so that the conventional technique of removing linebroadening in solids, MAS (see below), cannot completely remove this interaction at the same time as removing the first-order broadening. Hence, the resolution of MAS spectra from quadrupolar nuclei is usually worse than for spin- $\frac{1}{2}$ nuclei and often characteristic lineshapes are observed. If this is the case, it is usually possible to deduce the NMR interactions C_Q , η and δ_{iso} providing valuable information about the sample.

B1.12.2.3 BASIC EXPERIMENTAL PRINCIPLES OF FT NMR

The essence of NMR spectroscopy is to measure the separation of the magnetic energy levels of a nucleus. The original method employed was to scan either the frequency of the exciting oscillator or to scan the applied magnetic field until resonant absorption occurred. However, compared to simultaneous excitation of a wide range of frequencies by a short RF pulse, the scanned approach is a very time-inefficient way of recording the spectrum. Hence, with the advent of computers that could be dedicated to spectrometers and efficient Fourier transform (FT) algorithms, pulsed FT NMR became the normal mode of operation. Operating at constant field and frequency also produced big advantages in terms of reproducibility of results and stability of the applied magnetic field. In an FT NMR experiment a pulse of RF close to resonance, of duration T_p , is applied to a coil. If the pulse is applied exactly on resonance (i.e. the frequency of the applied em radiation exactly matches that required for a transition) it produces a resultant magnetic field orthogonal to B_0 in the rotating frame which causes a coherent oscillation of the magnetization. The magnetization is consequently tipped by an angle $\theta_p (= \gamma B_1 T_p)$ away from the direction defined by B_0 . After a $\pi/2$ -pulse all the magnetization is in a plane transverse to the direction to B_0 and, hence, is termed transverse magnetization. B_0 exerts a torque on the transverse magnetization which will consequently Larmor precess about B_0 . The rotating magnetization is then providing an alternating flux linkage with the NMR coil that, through Faraday's law of electromagnetic induction, will produce a voltage in the NMR coil. The transverse magnetization

decays through relaxation processes (see chapter B1.13). The observed signal is termed the free induction decay (FID). Adding coherently n FIDs together improves S/N by $n^{1/2}$ compared to a single FID.

In the linear approximation there is a direct Fourier relationship between the FID and the spectrum and, in the great majority of experiments, the spectrum is produced by Fourier transformation of the FID. It is a tacit assumption that everything behaves in a linear fashion with, for example, uniform excitation (or effective RF field) across the spectrum. For many cases this situation is closely approximated but distortions may occur for some of the broad lines that may be encountered in solids. The power spectrum $P(\nu)$ of a pulse applied at ν_0 is given by a sinc^2 function [8]

-8-

$$P(\nu) = [\sin^2 \pi T_p (\nu_0 - \nu)] / (\nu_0 - \nu)^2. \quad (\text{B1.12.9})$$

The spectral frequency range covered by the central lobe of this sinc^2 function increases as the pulselength decreases. For a spectrum to be undistorted it should really be confined to the middle portion of this central lobe (figure B1.12.2). There are a number of examples in the literature of solid-state NMR where the resonances are in fact broader than the central lobe so that the ‘spectrum’ reported is only effectively providing information about the RF-irradiation envelope, not the shape of the signal from the sample itself. The sinc^2 function describes the best possible case, with often a much stronger frequency dependence of power output delivered at the probe-head. (It should be noted here that other excitation schemes are possible such as adiabatic passage [9] and stochastic excitation [10] but these are only infrequently applied.) The excitation/recording of the NMR signal is further complicated as the pulse is then fed into the probe circuit which itself has a frequency response. As a result, a broad line will not only experience non-uniform irradiation but also the intensity detected per spin at different frequency offsets will depend on this probe response, which depends on the quality factor (Q). The quality factor is a measure of the sharpness of the resonance of the probe circuit and one definition is the resonance frequency/halfwidth of the resonance response of the circuit (also $= \omega_0 L/R$ where L is the inductance and R is the probe resistance). Hence, the width of the frequency response at 100 MHz is about 1 MHz. Hence, direct FT-pulse observation of broad spectral lines becomes impractical with pulse techniques for linewidths greater than ~ 200 kHz. For a great majority of NMR studies carried out on nuclei such as ^1H , ^{13}C and ^{29}Si this does not really impose any limitation on their observation. Broader spectral lines can be reproduced by pulse techniques, provided that corrections are made for the RF-irradiation and probe responses but this requires careful calibration. Such corrections have been most extensively used for examining satellite transition spectra from quadrupolar nuclei [11].

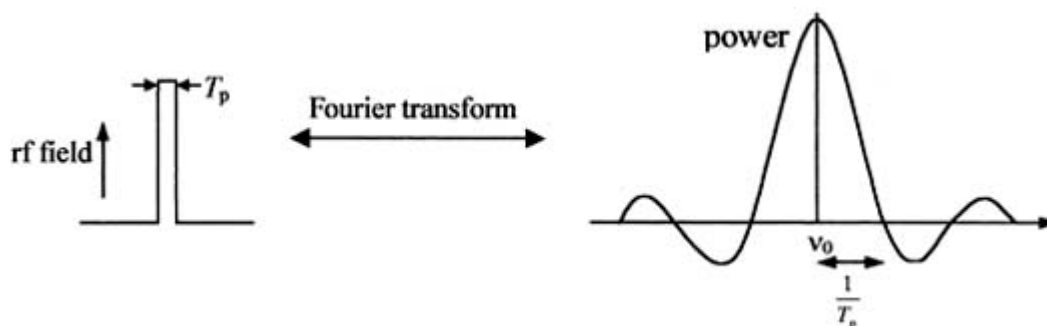


Figure B1.12.2. Power distribution for an RF pulse of duration T_p applied at frequency ν_0 .

Another problem in many NMR spectrometers is that the start of the FID is corrupted due to various instrumental deadtimes that lead to intensity problems in the spectrum. The spectrometer deadtime is made up of a number of sources that can be apportioned to either the probe or the electronics. The loss of the initial part of the FID is manifest in a spectrum as a rolling baseline and the preferential loss of broad components of

the spectrum. In the best cases the deadtime is $\leq 2 \mu\text{s}$, but even this can still lead to severe distortion of broad spectral lines. Baseline correction can be achieved by use of either simple spline fits using spectrometer software (including back-prediction) or the use of analytical functions which effectively amount to full-intensity reconstruction. Many spectrometer software packages now contain correction routines, but all such procedures should be used with extreme caution.

-9-

B1.12.3 INSTRUMENTATION

B1.12.3.1 OVERVIEW OF A PULSE FT NMR SPECTROMETER

The basic components of a pulse FT NMR spectrometer are shown schematically in figure B1.12.3. It can be seen that, in concept, a NMR spectrometer is quite simple. There is a high-field magnet, which these days is nearly always a superconducting solenoid magnet that provides the basic Zeeman states on which to carry out the NMR experiment. The probe circuit containing the sample in the NMR coil is placed in the magnetic field. The probe is connected to the transmitter that is gated to form the pulses that produce the excitation. The probe is also connected to the receiver and it requires some careful design to ensure that the receiver that is sensitive to μV does not see any of the large-excitation voltages produced by the transmitter. The relatively simple concept of the experiment belies the extensive research and development effort that has gone into developing the components of NMR spectrometers. Although all components are important, emphasis is placed on three central parts of the spectrometer: namely the magnet, the probe and signal detection.

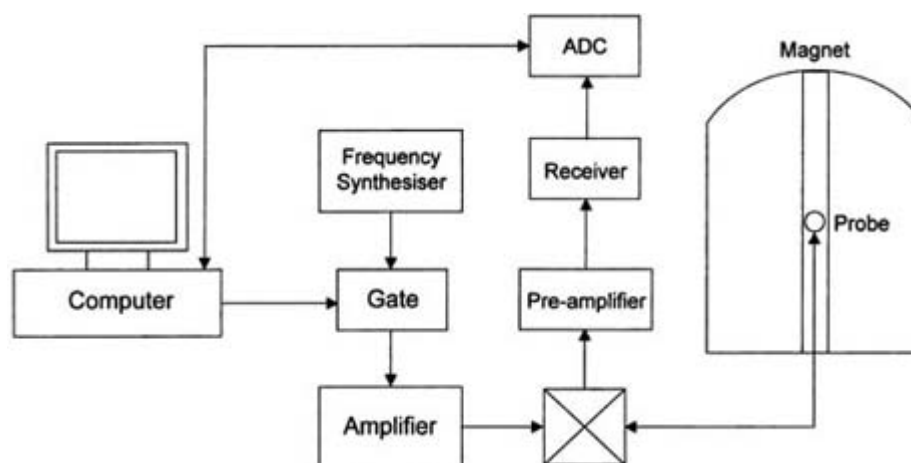


Figure B1.12.3. Schematic representation showing the components of a pulse FT NMR spectrometer.

B1.12.3.2 MAGNETS

Much effort goes into producing ever higher magnetic fields, and the highest currently commercially available for solid-state NMR is 18.8 T. Standard instruments are now considered to be 4.7–9.4 T. The drive for higher fields is based on the increased chemical shift dispersion (in hertz) and the increase in sensitivity via both the Boltzmann factor and higher frequency of operation. For solid-state NMR of half-integer spin quadrupolar nuclei there is the additional advantage that the second-order quadrupolar broadening of the central transition decreases inversely proportionally with B_0 . Superconducting solenoids dominate based on Nb_3Sn or NbTi multifilament wire kept in liquid helium. However, fields and current densities now used are close to the critical limits of these materials demanding improved materials technology [12]. The principle of operation is very simple: a high current is passed through a long coil of wire, with typically 40–100 A of current circulating around several kilometres of wire. This means that the magnet stores significant amounts of energy in its field ($=1/2LI^2$; L is the solenoid's inductance and I is the current flowing) of up to 10 MJ.

A superconducting magnet consists of a cryostat, main coil, superconducting shim set and a means for attaching the current supply to the main coil (figure B1.12.4). The cryostat consists of two vessels for the liquid cryogens, an inner one for helium and the outer one for nitrogen. Then, to insulate these, there are several vacuum jackets with a radiation shield. The aim is to reduce heat leakage to the inner chamber to conserve helium. Superconducting magnets in NMR are usually operated in persistent mode, which means that, after a current is introduced, the start and end of the main coil are effectively connected so that the current has a continuous path within the superconductor and the power supply can then be disconnected. To achieve this the circuits within the cryostat have a superconducting switch. The coil circuits are also designed to cope with a sudden, irreversible loss of superconductivity, termed a quench. There are resistors present (called dump resistors) to disperse the heating effect and prevent damage to the main coil when a quench occurs.

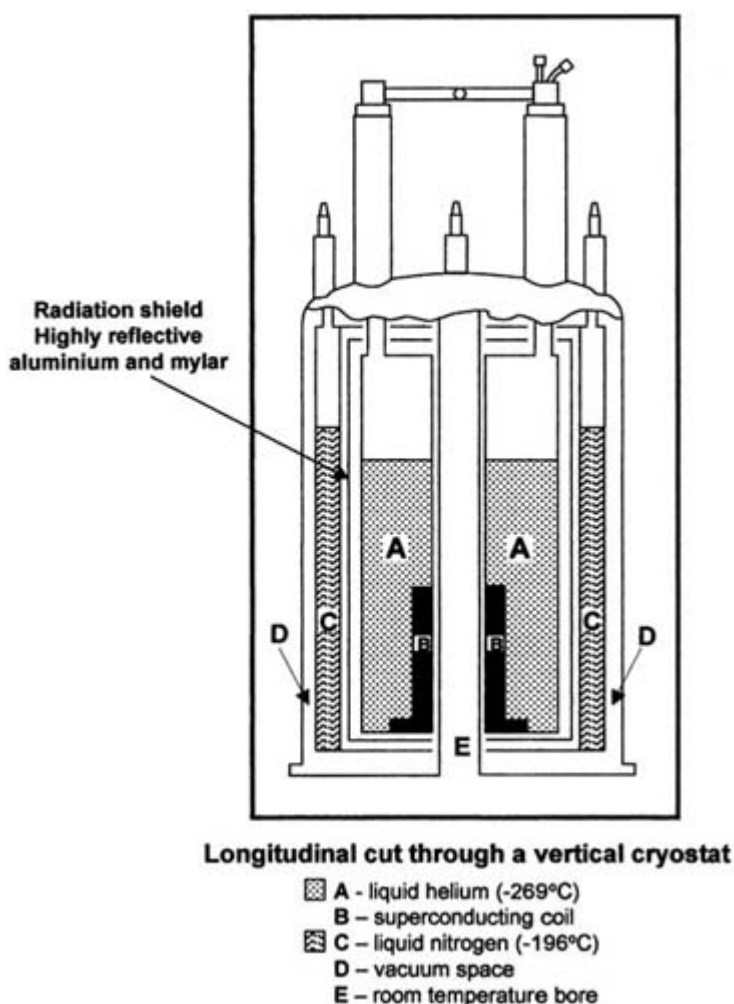


Figure B1.12.4. Construction of a high-field superconducting solenoid magnet.

Higher magnetic fields exist than those used in NMR but the NMR experiment imposes constraints in addition to just magnitude. An NMR spectroscopy experiment also demands homogeneity and stability of the magnetic field. Long-term stability is aided by persistent-mode operation and the drift should be $<2 \times 10^{-7}$ a day. Homogeneity requirements for solid-state NMR experiments are typically up to 2×10^{-9} over a volume of $\sim 1 \text{ cm}^3$. The main coil alone is unable to produce this level of homogeneity so there is also a set of smaller

superconducting coils called cryoshims. The number of these cryoshims depends on the design and also the purpose of the magnet (e.g. solid-state NMR, high-resolution NMR, imaging) but typically varies between three and eight. Although for many wide-line solid-state NMR experiments the homogeneity produced by the cryoshims is sufficient, most commercial spectrometers also have a room-temperature shim set which further improves the homogeneity of the magnetic field. A final consideration for the magnet is the accessible room-temperature bore of the magnet. A standard liquids magnet has a bore of 52 mm diameter. However most solid-state NMR spectroscopists prefer 89 mm as this gives much more room for the probe, allowing the use of larger and more robust electrical components for handling high powers and for accommodating some of the more specialist probe designs (e.g. double angle rotation, dynamic angle spinning (see [section B1.12.4.5](#)) etc).

B1.12.3.3 PROBES

The heart of an NMR spectrometer is the probe, which is essentially a tuned resonant circuit with the sample contained within the main inductance (the NMR coil) of that circuit. Usually a parallel tuned circuit is used with a resonant frequency of $\omega_0 = (LC)^{-1/2}$. The resonant frequency is obviously the most important probe parameter but the input impedance, which should be 50Ω , and Q are also extremely important. Several designs of coil exist each having advantages in specific applications. The traditional coil design, particularly applicable at lower frequencies and for solids, is the conventional solenoid. For external access of large samples, Helmholtz or saddle coils are used, such as are in widespread use for high-resolution liquid-state NMR studies. For large samples, again with external access, birdcage coils are finding increasing uses especially in magnetic resonance imaging. There are competing requirements for the probe which can be characterized in terms of Q ; pulselength ($\propto Q^{-1/2}$), deadtime ($\propto Q$), maximum voltage ($\propto Q$), bandwidth ($\propto Q^{-1}$) and sensitivity ($\propto Q^{1/2}$). A probe cannot be designed that will lead to all these being optimized simultaneously because of their differing Q -dependence so that compromise and focus on the most important aspects for a specific application is required. The probe needs to be constructed from robust electronic components, as they often have to withstand high voltages (many kilovolts).

Often linear circuit analysis is applied but in probes designed for solids, and therefore high-power operation, nonlinear effects can occur. Furthermore, in doubly and triply tuned probes used for decoupling, cross-polarization, and for some of the more sophisticated pulse sequences such as REDOR, TEDOR etc even small voltages generated at the second resonant frequency are unacceptable. They can swamp the NMR signal, given that the coil is part of more than one resonant circuit. Detailed consideration of the design criterion for double-tuned CP-MAS probes has been given [13]. These days, a number of sequences demand triply resonant circuits with two channels tunable over the lower frequency ranges and the third tunable to the high- γ nuclei (e.g. ^1H , ^{19}F).

B1.12.3.4 SIGNAL DETECTION

In addition to the deadtime mentioned earlier, magnetoacoustic ringing (up to 200 μs) can be very significant without careful probe design. Samples that exhibit piezoelectric behaviour can produce very long response times of up to 10 ms. The microvolt NMR signal generated in the coil needs amplification prior to detection and digitization. The first stage is about 30 dB amplification in the preamplifier with the most important characteristic being the noise figure (NF), essentially a measure of the noise added to the signal by the amplifier. Careful consideration is necessary for the production of a low noise figure and rapid recovery from saturation and, again, some compromise is required, with recovery of $<2 \mu\text{s}$. The preamp. should also have good linearity.

The amplified signal is passed to a double-balanced mixer configured as a phase-sensitive detector where the two inputs are the NMR signal (ω_0) and the frequency of the synthesizer (ω_{ref}) with the output proportional to $\cos(\omega_0 - \omega_{\text{ref}})t + \theta) + \cos((\omega_0 + \omega_{\text{ref}})t + \theta)$. The sum frequency is much larger than the total bandwidth of the spectrometer so it is lost, leaving only the difference frequency. Phase-sensitive detection is equivalent to examining the NMR signal in the rotating reference frame and if the frequencies ω_0 and ω_{ref} are equal a constant output is obtained (neglecting relaxation effects). Most modern spectrometers employ two phase-sensitive detectors which have reference signals that differ in phase by 90° , termed quadrature phase-sensitive detection [14]. This scheme can distinguish whether a signal is above or below the reference frequency and allows the transmitter frequency to be placed at the centre of the range of interest, improving pulse power efficiency and signal-to-noise. (This is not possible with a single detector, which can only provide the magnitude of the offset.) Imbalance in the two channels and non- 90° angles between them can give rise to quadrature images and should be minimized in the spectrometer set-up. Phase cycling includes application of different phase pulses and normally four phases 90° apart are available. However, much more sophisticated phase cycles are now required and variable phases can be generated by using a digital synthesizer. In particular, there is increasing demand for much smaller phase shifts than 90° .

The signal is then digitized ready for storage in the computer memory. Digitizers are characterized by the number of bits (usually 12 or 16, determining the dynamic range), the rate of digitization (determining the dwell time) and the size of memory capable of storing the data points. Until very recently the ability to record narrow spectral objects over a broad range of frequencies has been limited, usually by the on-board computer memory, but commercial spectrometers have now addressed this problem.

B1.12.4 EXPERIMENTAL TECHNIQUES

B1.12.4.1 CLASSIFICATION OF NUCLEI

The sensitivity in an NMR experiment is directly proportional to the number of spins, making quantification of the amount of a particular element present straightforward, at least for spin $I = \frac{1}{2}$ nuclei. Furthermore, the large variation in the gyromagnetic ratio, γ , means that, except for pathological cases, the resonant frequencies are sufficiently different from one element to another that there is no possibility of confusing different elements. The ease of obtaining a spectrum, and to a certain extent the type of experiment undertaken, depends upon γ , the nuclear spin and the natural abundance of the isotope concerned. The large variation in γ means that the sensitivity, which is proportional to γ^3 (see [section B1.12.2.1](#)), varies by $>10^4$ from (say) ^1H and ^{19}F which have the largest γ to ^{109}Ag which has one of the smallest. For spin $I = \frac{1}{2}$ nuclei the ability to obtain a useful signal is also dependent on the spin-lattice relaxation time, T_1 , as well as the sensitivity. As the principal cause of relaxation in spin- $\frac{1}{2}$ systems is usually dipolar and thus proportional to (at least) γ^2 , relaxation times can be very long for low-gamma nuclei making experiments still more difficult.

Nuclei with spin $I > \frac{1}{2}$, about three-quarters of the periodic table, have a quadrupolar moment and as a consequence are affected by any electric field gradient present (see [section B1.12.2.2](#)). The lines can be very broad even when techniques such as MAS (see [section B1.12.4.3](#)) are used, thus it is convenient to divide nuclei into groups depending upon ease of observation and likely width of the line. This is done in [table B1.12.2](#) for the most commonly studied elements. These have been divided into six categories: (a) spin $I = \frac{1}{2}$ nuclei which are readily observable, (b) spin $I = \frac{1}{2}$ nuclei which are observable with difficulty because either the isotopic abundance is low, in which case isotopic enrichment is sometimes used (e.g. ^{15}N), or because the γ is very small (e.g. ^{183}W), (c) spin $I = 1$ nuclei where the quadrupolar interaction can lead to very broad lines (d) non-integer spin $I > \frac{1}{2}$ nuclei which are readily observable, (e) non-integer spin $I > \frac{1}{2}$ nuclei which are readily observable only in relatively symmetric environments because the quadrupolar interaction can

strongly broaden the line and (f) those spin $I > \frac{1}{2}$ nuclei whose quadrupole moment is sufficiently large that they can only be observed in symmetric environments where the electric field gradient is small. Of course the boundaries between the categories are not 'rigid'; as technology improves and magnetic fields increase there is movement from (f) to (e) and (e) to (d). In each case the sensitivity at natural abundance relative to ^{29}Si (for which resonance is readily observed) is given. More complete tables of nuclear properties relevant to NMR are given in several books on NMR [1, 2].

-14-

Table B1.12.2 Classification of nuclei according to spin and ease of observation.

Isotope	Sensitivity at natural abundance relative to ^{29}Si	Isotope	Sensitivity at natural abundance relative to ^{29}Si
$I = \frac{1}{2}$			
(a) Readily observable			
^1H	2700	^{113}Cd	3.6
^{13}C	0.48	^{119}Sn	12
^{19}F	2200	^{125}Te	6.0
^{29}Si	1.00	^{195}Pt	9.1
^{31}P	180	^{199}Hg	2.6
^{77}Se	1.4	^{205}Tl	35
^{89}Y	0.32	^{207}Pb	5.6
(b) Observable with difficulty or requiring isotopic enrichment			
$^{15}\text{N}^*$	1.0×10^{-2}	^{109}Ag	0.13
$^{57}\text{Fe}^*$	1.0×10^{-3}	^{183}W	2.8×10^{-2}
^{109}Rh	8.5×10^{-2}		
(c) Integer $I = 1$			
$^2\text{D}^*$	3.9×10^{-3}	^{14}N	2.7
Non-integer $I = \frac{1}{2}$			
(d) Readily observable			
^7Li	730	^{23}Na	250
^9Be	37	^{27}Al	560
^{11}B	360	^{51}V	1030
$^{17}\text{O}^*$	2.9×10^{-2}	^{133}Cs	130
(e) Readily observable only in relatively symmetric environments			
^{25}Mg	0.73	^{59}Co	750
$^{33}\text{S}^*$	4.7×10^{-2}	^{65}Cu	96
^{37}Cl	1.7	^{67}Zn	0.32
^{39}K	1.3	^{71}Ga	150
$^{43}\text{Ca}^*$	2.3×10^{-2}	^{81}Br	133
^{45}Sc	820	^{87}Rb	133
^{55}Mn	470	^{93}Nb	1300
(f) Observable in very symmetric environment			
^{49}Ti	0.44	^{105}Pd	0.68
^{53}Cr	0.23	^{115}In	910

⁶¹ Ni	0.11	¹²¹ Sb	250
⁷³ Ge	0.30	¹²⁷ I	260
⁷⁵ As	69	¹³⁵ Ba	0.89
⁸⁷ Sr	0.51	¹³⁹ La	160
⁹¹ Zr	2.9	²⁰⁹ Bi	380
⁹⁵ Mo	1.4		

* Isotopically enriched.

-15-

B1.12.4.2 STATIC BROAD-LINE EXPERIMENTS

Static NMR powder patterns offer one way of characterizing a material, and if spectral features can be observed and the line simulated then accurate determination of the NMR interaction parameters is possible. The simplest experiment to carry out is one-pulse acquisition and despite the effects outlined above that corrupt the start of the time domain signal, the singularities are relatively narrow spectral features and their position can be recorded. The magnitude of the interaction can then be estimated. A common way to overcome deadtime problems is to form a signal with an effective time zero point outside the deadtime, i.e. an echo. There is a huge multiplicity of methods for forming such echoes. Most echo methods are two-pulse sequences, with the classic spin echo consisting of $90^\circ - \tau - 180^\circ$ which refocuses at τ after the second pulse. The echo decay shape is a good replica of the original FID and its observation can be used to obtain more reliable and quantitative information about solids than from the one-pulse experiment.

To accurately determine broad spectral lineshapes from echoes hard RF pulses are preferred for uniform excitation. Often an echo sequence with phase cycling first proposed by Kunwar *et al* [15] has been used, which combines phase cycling to remove quadrature effects and to cancel direct magnetization (the remaining FID) and ringing effects. The rotation produced by the second pulse in the two-pulse echo experiment is not critical. In practice, the best choice is to make the second pulse twice the length of the first, with the actual length a trade-off between sensitivity and uniformity of the irradiation. In recording echoes there is an important practical consideration in that the point of applying the echo is to move the effective $t = 0$ position for the FID to being outside the region where the signal is corrupted. However, in order that phasing problems do not re-emerge, a data sampling rate should be used that is sufficient to allow the effective $t = 0$ point to be accurately defined. If T_2 allows the whole echo (both before and after the maximum, i.e. $t = 0$) to be accurately recorded without an unacceptably large loss of intensity, there is then no need to accurately define the new $t = 0$ position. Fourier transformation of the whole echo (which effectively amounts to integration between $\pm\infty$) followed by magnitude calculation removes phasing errors producing a pure absorption lineshape with the signal-to-noise $\sqrt{2}$ larger than that obtained by transforming from the echo maximum.

Even if echoes are used, there are still difficulties in recording complete broad spectral lines with pulsed excitation. Several approaches have been adopted to overcome these difficulties based on the philosophy that although the line is broad it can be recorded using a series of narrow-banded experiments. One of these approaches is to carry out a spin-echo experiment using relatively weak RF pulses, recording only the intensity of the on-resonance magnetization and repeating the experiment at many frequencies to map out the lineshape. This approach has been successfully used in a series of studies. An example is ⁹¹Zr from the polymorphs of ZrO₂ (figure B1.12.5) where mapping out the lineshape clearly shows differences in NMR interaction [16]. This approach works but is extremely tedious because each frequency step requires accurate retuning of the probe. An alternative is to sweep the main magnetic field. There are several examples of sweeping the main magnetic field for solids dating from the earliest days of NMR but only a limited number reported using superconducting magnets, with a recent example being for ²⁷Al in α -Al₂O₃ [17]. It is now

possible to have a single NMR spectrometer that is capable of both conventional high-resolution spectroscopy and also field sweep operation. As with the stepped-frequency experiment, relatively soft pulses are applied, and although strictly the on-resonance part of the magnetization should be used, it has been shown experimentally that using the spin-echo intensity directly accurately reproduces the lineshape (figure B1.12.6) [18]. Recording the full distortion-free lineshape including the satellite transitions then allows accurate determination of the asymmetry parameter.

-16-

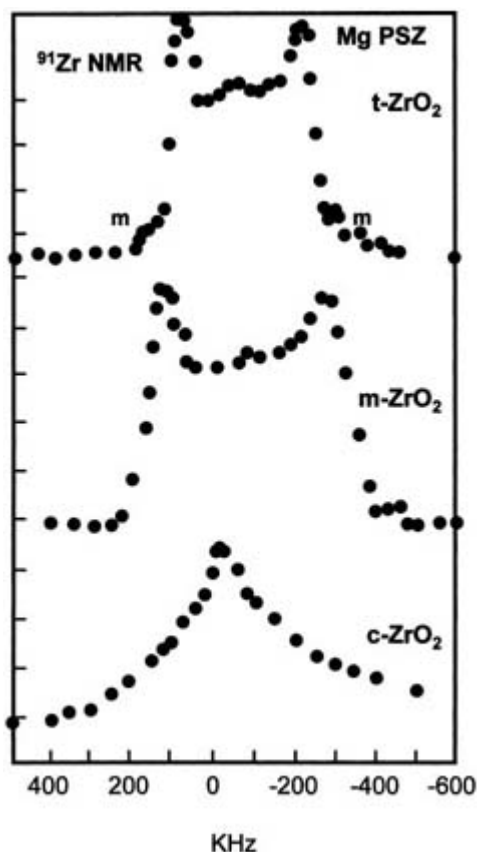


Figure B1.12.5. ^{91}Zr static NMR lineshapes from ZrO_2 polymorphs using frequency-stepped spin echoes.

For quadrupolar nuclei, the dependence of the pulse response on ν_Q/ν_1 has led to the development of quadrupolar nutation, which is a two-dimensional (2D) NMR experiment. The principle of 2D experiments is that a series of FIDs are acquired as a function of a second time parameter (e.g. here the pulse length applied). A double Fourier transformation can then be carried out to give a 2D data set (F_1 , F_2). For quadrupolar nuclei while the pulse is on the experiment is effectively being carried out at low field with the spin states determined by the quadrupolar interaction. In the limits $\nu_Q \ll \nu_1$ and $\nu_Q \gg \nu_1$ the pulse response lies at ν_1 and $(I + \frac{1}{2})\nu_1$ respectively so is not very discriminatory. However, for $\nu_Q \sim \nu_1$ the pulse response is complex and allows C_Q and η to be determined by comparison with theoretical simulation. Nutation NMR of quadrupolar nuclei has largely been limited by the range of RF fields that puts ν_Q in the intermediate region. This approach has been extended by Kentgens by irradiating off-resonance, producing a larger effective nutation field, and matches ν_1 to ν_Q [19].

-17-

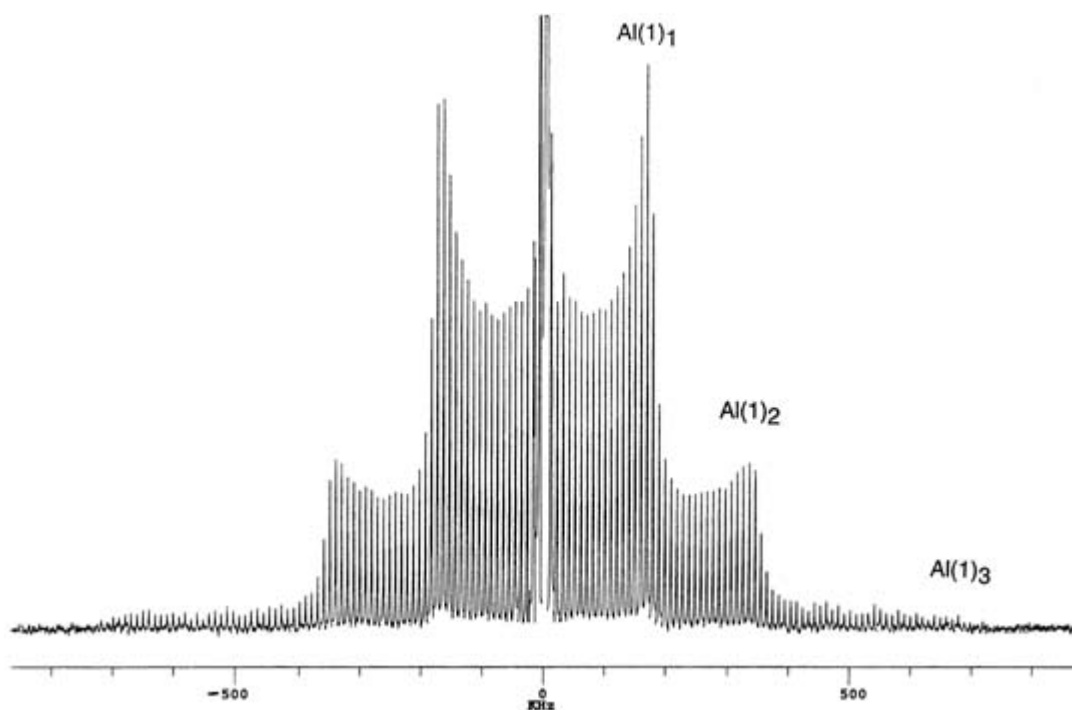


Figure B1.12.6. Field-swept ^{27}Al NMR spectrum from $\alpha\text{-Al}_2\text{O}_3$.

B1.12.4.3 MAGIC ANGLE SPINNING

All of the anisotropic interactions described in [section B1.12.2](#) are second-rank tensors so that, in polycrystalline samples, lines in solids can be very broad. In liquids, the rapid random molecular tumbling brought about by thermal motion averages the angular dependences to zero, leaving only the isotropic part of the interaction. The essence of the spinning technique is to impose a time dependence on these interactions externally so as to reduce the anisotropic part, which to first order has the same Legendre polynomial, $P_2(\cos \theta)$, dependence for all the interactions, in a similar but not identical manner as in liquids. This time dependence is imposed by rapidly rotating the whole sample container (termed the rotor) at the so-called ‘magic angle’ where $3 \cos^2 \theta = 1$, i.e. $\theta = 54^\circ 44'$. ‘Rapid’ means that the spinning speed should be faster than the homogeneous linewidth. (A spectral line is considered homogeneous when all nuclei can be considered as contributing intensity to all parts of the line, so that the intrinsic linewidth associated with each spin is the same as the total linewidth.) Essentially, this is determined by the dipolar coupling: in proton-rich systems the proton–proton coupling can be >40 kHz, beyond the range of current technology where the maximum commercially available spinning speed is ~ 35 kHz. Fortunately, in most other systems the line is inhomogeneously broadened, i.e. is made up of distinct contributions from individual spins in differently oriented crystallites which merge to give the composite line, the intrinsic width associated with each spin being considerably narrower than the total linewidth. To cause effective narrowing of these lines the spinning rate needs only to exceed the intrinsic linewidth and typically a few kilohertz is sufficient to narrow lines where the chemical shift is the dominant linebroadening mechanism. In [figure B1.12.7](#) both the static and MAS ^{29}Si spectrum of a sample of sodium disilicate ($\text{Na}_2\text{Si}_2\text{O}_7$) crystallized from a glass is shown as an example. Whilst the static spectrum clearly indicates an axial chemical shift powder pattern, it gives no evidence of more than one silicon site. The MAS spectrum clearly shows four resolved lines from the different polymorphs present in the material whose widths are ~ 100 times less than the chemical shift anisotropy.

Note that if the spinning speed is less than the static width then a series of spinning sidebands are produced,

separated from the isotropic line by the spinning speed (these are also visible in figure B1.12.7). For multisite systems, spinning sidebands can make interpretation of the spectrum more difficult and it may be necessary to do experiments at more than one spinning speed in order to determine the isotropic peak and to eliminate overlap between sidebands and main peaks. However, the presence of sidebands can be useful since from the amplitude of the spinning sideband envelope one can deduce the complete chemical tensor, giving additional information about the local site environment [20].

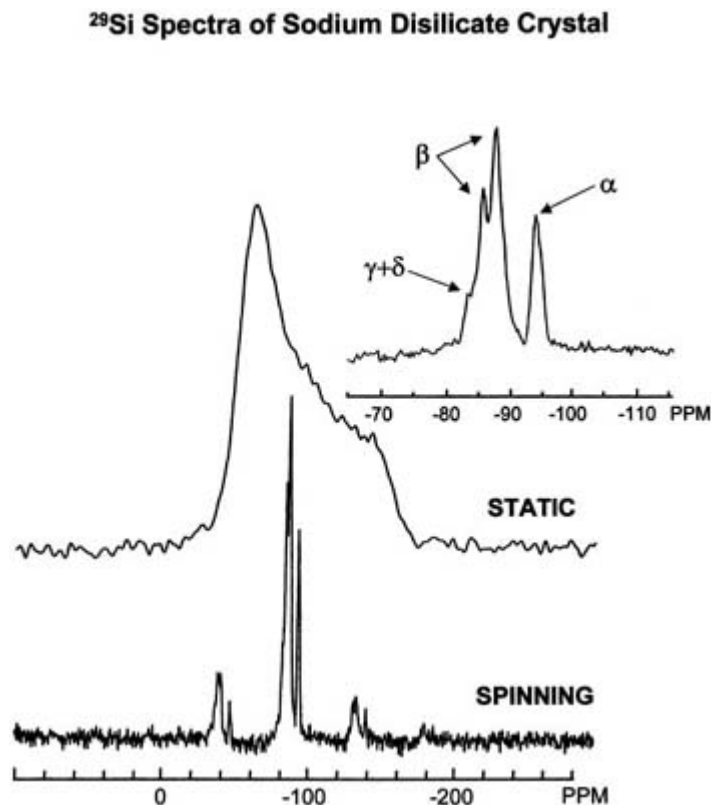


Figure B1.12.7. Static and MAS ²⁹Si NMR spectra of crystalline Na₂Si₂O₅.

The interpretation of MAS experiments on nuclei with spin $I > \frac{1}{2}$ in non-cubic environments is more complex than for $I = \frac{1}{2}$ nuclei since the effect of the quadrupolar interaction is to spread the $\pm m \leftrightarrow \pm(m - 1)$ transition over a frequency range $(2m - 1)\nu_Q$. This usually means that for non-integer nuclei only the $\frac{1}{2} \leftrightarrow -\frac{1}{2}$ transition is observed since, to first order in the quadrupolar interaction, it is unaffected. However, usually second-order effects are important and the angular dependence of the $\frac{1}{2} \leftrightarrow -\frac{1}{2}$ transition has both $P_2(\cos \theta)$ and $P_4(\cos \theta)$ terms, only the first of which is cancelled by MAS. As a result, the line is narrowed by only a factor of ~ 3.6 , and it is necessary to spin faster than the residual linewidth $\Delta\nu_Q$ where

$$\Delta\nu_Q = f(I) \frac{C_Q^2(6 + \eta)^2}{224\nu_0} \quad (\text{B1.12.10})$$

to narrow the line. The resulting lineshape depends upon both C_Q and η and is shown in [figure B1.12.8](#) for several different asymmetry parameters. The centre of gravity of the line does not occur at the isotropic

chemical shift (see [figure B1.12.1\(d\)](#)); there is a quadrupolar shift

$$\nu_{m,\text{cg}}^{(2)} = \frac{3}{40} f(I) \frac{C_Q^2}{\nu_0} \left(1 + \frac{\eta^2}{3} \right). \quad (\text{B1.12.11})$$

If the lineshape can be clearly distinguished, then C_Q , η and δ_{iso} can be readily determined even for overlapping lines, although it should be noted that most computer packages used to simulate quadrupolar lineshapes assume ‘infinite’ spinning speed. Significant differences between the experimental lineshape and simulation can occur if one is far from this limit and caution is required. A further problem for MAS of quadrupolar nuclei is that the angle must be set very accurately (which can be difficult and time consuming with some commercial probes) to obtain the true lineshape of broad lines. In many samples (e.g. glasses and other disordered systems) featureless lines are observed and the centre of gravity must then be used to estimate δ_{iso} and the electric field gradient since

$$\delta_{\text{iso}} = \delta_{\text{cg}} - \frac{3}{40} f(I) \frac{C_Q^2}{\nu_0^2} \left(1 + \frac{\eta^2}{3} \right) \quad (\text{B1.12.12})$$

and a plot of ν_{cg} against $1/\nu_0^2$ will give δ_{iso} and $C_Q^2(1 + \eta^2/3)$. $f(I)$ is the spin-dependent factor $[I(I+1) - 3/4]/I^2(2I-1)^2$ and is given in table B1.12.3.

Table B1.12.3 Spin-dependent factor $f(I)$ for the isotropic second-order quadrupole shift.

<i>I</i>				
$\frac{3}{2}$	$\frac{5}{2}$	$\frac{9}{2}$	$\frac{2}{25}$	$\frac{1}{54}$

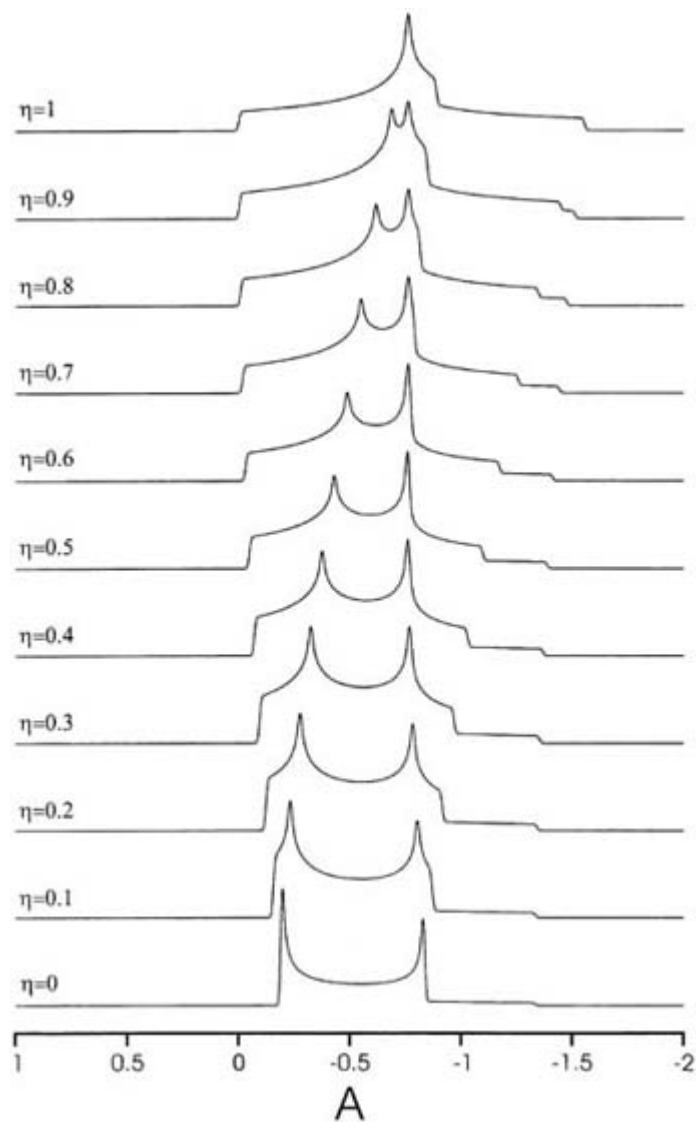


Figure B1.12.8. MAS NMR lineshapes from the central transition lineshape for non-integer quadrupole lineshapes with various η_Q ($A = (I(I+1) - 3/4)v_Q^2/v_0$).

The second-order quadrupolar broadening of the $\frac{1}{2} \leftrightarrow -\frac{1}{2}$ transition can be further reduced by spinning at an angle other than 54.7° (VAS), the width being a minimum between $60\text{--}70^\circ$. The reduction is only ~ 2 however, and dipolar and shift anisotropy broadening will be reintroduced, thus VAS has only found limited application.

B1.12.4.4 DECOUPLING AND CROSS-POLARIZATION

In proton-containing systems such as organic materials the dipole–dipole interaction is usually sufficiently large that current spinning speeds are not sufficient to narrow the line. However, in a heteronuclear (I,S) spin system if a large RF magnetic field is applied at the resonance frequency of the spin, I , one wishes to decouple, the magnetization will precess around the effective field in the rotating frame and the average IS dipolar coupling will tend to zero. This technique is most commonly applied to remove the $^1\text{H}\text{--}^{13}\text{C}$ dipolar coupling, but can be used for any system. Decoupling is usually combined with cross-polarization (CP) and the pulse sequence is shown in [figure B1.12.9](#). A 90° pulse is applied to the I spin system and the phase of the RF is then shifted by 90° to spin lock the magnetization in the rotating frame. The S spin system RF is now

turned on with an amplitude such that $\gamma_I B_{1I} = \gamma_S B_{1S}$ i.e. in the rotating frame both spin systems are precessing at the same rate (the Hartmann–Hahn condition) and thus magnetization can be transferred via the flip-flop term in the dipolar Hamiltonian. The length of time that the S RF is on is called the contact time and must be adjusted for optimum signal. Since magnetization transfer is via the dipole interaction, in general the signal from S spin nuclei closest to the I spins will appear first followed by the signal from S spins further away. Thus one use of CP is in spectral editing. However, the main use is in signal enhancement since the S spin magnetization will be increased by (γ_I/γ_S) which for ^1H - ^{13}C is ~ 4 and for ^1H - ^{15}N is ~ 9 . In addition, the S spin relaxation time T_1 is usually much longer than that of the I spin system (because γ is smaller); as a consequence in CP one can repeat the experiment at a rate determined by T_{1I} rather than T_{1S} producing a further significant gain in signal to noise. For spin $I, S = \frac{1}{2}$ systems the equation describing the signal is given by

$$M_0(\text{CP}) = M_0 \left(\frac{\gamma_I}{\gamma_S} \right) \left[\exp \left(- \frac{t}{T_{1\rho I}} \right) - \exp \left(- \left\{ \frac{1}{T_{1\rho S}} + \frac{1}{T_{IS}} \right\} t \right) \right] \quad (\text{B1.12.13})$$

where $T_{1\rho I}$ and $T_{1\rho S}$ are the relaxation times in the rotating frame of the I and S spin systems respectively and T_{IS} , which is dependent upon the strength of the dipole coupling between the I and S spin systems, is the characteristic time for the magnetization of the S spin system to increase via contact with the I spins. Generally, $T_{1\rho S}$ is very much longer than T_{IS} and can be ignored and although $T_{1\rho I}$ is shorter it too is usually much longer than T_{IS} . In this case maximum signal will occur for a contact time $t \sim T_{IS} \ln(T_{IS}/T_{1\rho I})$. If one wishes to do quantitative measurements using CP it is necessary to plot signal amplitude versus contact time and then fit to equation B1.12.13. A typical plot is shown for glycine in [figure B1.12.10](#). It can be seen from the inset that T_{IS} for the CH_2 peak at ~ 40 ppm is much shorter ($20 \mu\text{s}$) than that for the COOH peak at 176 ppm ($570 \mu\text{s}$). Carbon spectra can be very complex. Various pulse sequences have been used to simplify the spectra. The most common is dipolar dephasing, in which there is a delay after the contact time (during which no decoupling takes place) before the signal is acquired. The signal from strongly coupled carbon (e.g. CH_2) will decay rapidly during this time leaving only weakly coupled carbons visible in the spectrum.

-22-

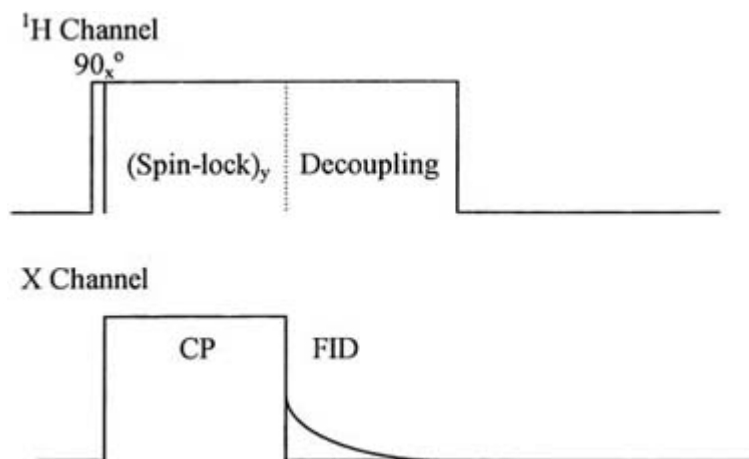


Figure B1.12.9. Pulse sequence used for CP between two spins (I, S).

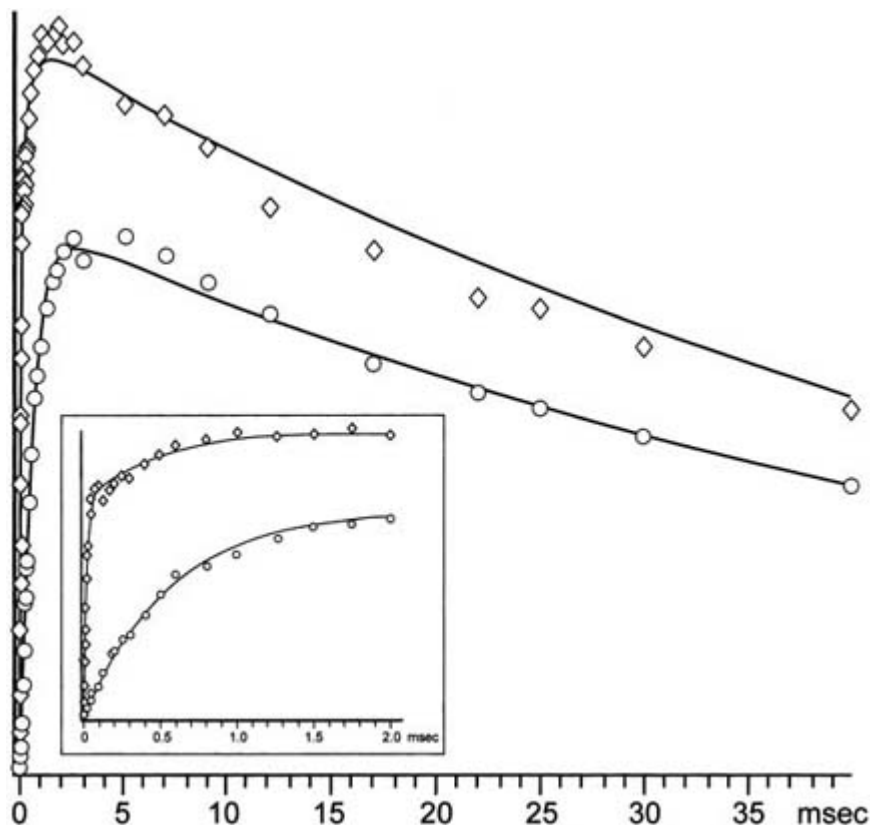


Figure B1.12.10. ^{13}C CP data from the two carbons in glycine as a function of contact time. The signal for short contact times is shown in the inset where the effect of the different $T_{1\rho}$ values can be clearly seen.

-23-

Whilst MAS is effective (at least for spin $I = \frac{1}{2}$) in narrowing inhomogeneous lines, for abundant spin systems with a large magnetic moment such as protons (or fluorine) the homonuclear dipole coupling can be >30 kHz, too strong to be removed by MAS at currently available spinning speeds. Thus there have been many multiple pulse sequences designed to remove homonuclear coupling. When combined with MAS they are called CRAMPS (combined rotation and multiple pulse spectroscopy). The MREV-8 [6] sequence is one of the more commonly used and robust sequences; however, they all require a special probe because of the high powers needed, together with a very well tuned and set up spectrometer. Furthermore, to be effective the period of rotation needs to be much greater than the cycle time of the sequence, thus their use is somewhat restricted.

B1.12.4.5 HIGH-RESOLUTION SPECTRA FROM QUADRUPOLEAR NUCLEI

Although MAS is very widely applied to non-integer spin quadrupolar nuclei to probe atomic-scale structure in solids, such as distinguishing AlO_4 and AlO_6 environments [21], simple MAS about a single axis cannot produce a completely averaged isotropic spectrum. As the second-order quadrupole interaction contains both second-rank ($\propto 3 \cos^2 \theta - 1$ ($P_2(\cos \theta)$)) and fourth-rank ($\propto 35 \cos^4 \theta - 30 \cos^2 \theta + 3$ ($P_4(\cos \theta)$)) terms it can be seen from figure B1.12.11 that spinning at 54.7° can only partially average $35 \cos^4 \theta - 30 \cos^2 \theta + 3$. If a characteristic well defined lineshape can be resolved the NMR interaction parameters can be deduced. Even if overlapping lines are observed, provided a sufficient number of features can be discerned, fitting, especially constrained by field variation, will allow the interactions to be accurately deduced. However, there are still many cases where better resolution from such nuclei would be extremely helpful, and over the last decade or so there have been many ingenious approaches to achieve this. Here, four of the main approaches will be briefly examined, namely satellite transition MAS spectroscopy, double-angle rotation (DOR), dynamic angle spinning (DAS) and multiple quantum (MQ) MAS. The latter three produce more complete averaging of the interactions by imposing more complex time dependences on

the interactions. DOR and DAS do this directly on the spatial coordinates, whereas MQ also manipulates the spin part of the Hamiltonian. For each method the background to producing better-resolution solid-state NMR spectra will be given, and the pros and cons of each detailed. Extensive referencing to general reviews [22, 23 and 24] and reviews of specific techniques is given where more details can be found. Then, by way of illustration, comparison will be made of these methods applied to ^{27}Al NMR of kyanite.

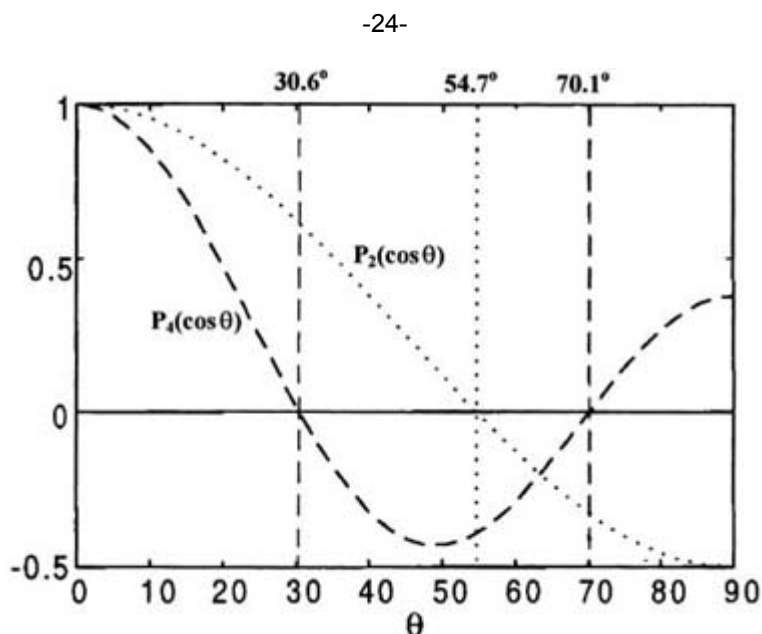


Figure B1.12.11. Angular variation of the second- and fourth-rank Legendre polynomials.

(A) SATELLITE TRANSITION MAS

Physical background. MAS will narrow the inhomogeneously broadened satellite transitions to give a series of sharp sidebands whose intensity envelopes closely follow the static powder pattern so that the quadrupole interaction can be deduced. The work of Samoson [25] gave real impetus to satellite transition spectroscopy by showing that both the second-order quadrupolar linewidths and isotropic shifts are functions of I and m_z . Some combinations of I and m_z produce smaller second-order quadrupolar effects on the satellite lines than for the central transition, thus offering better resolution and more accurate determination of $\delta_{\text{iso, cs}}$. The two cases where there are distinct advantages of this approach over using the central transition are $(\pm\frac{3}{2}, \pm\frac{1}{2})$ for $I = \frac{5}{2}$ and $(\pm\frac{5}{2}, \pm\frac{3}{2})$ for $I = \frac{9}{2}$. ^{27}Al has been the focus of much of the satellite transition work and has been used for a range of compounds including ordered crystalline, atomically disordered and amorphous solids. The work of the group of Jaeger has greatly extended the practical implementation and application of this technique and a comprehensive review is given in [26].

Advantages. The experiment can be carried out with a conventional fast-spinning MAS probe so that it is straightforward to implement. For recording the satellite transition lineshapes it offers better signal-to-noise and is less susceptible to deadtime effects than static measurements. As the effects differ for each m_z value, a single satellite transition experiment is effectively the same as carrying out multiple field experiments on the central transition.

Disadvantages. The magic angle must be extremely stable and accurately set. The spinning speed must show good stability over the duration of the experiment. The probe needs to be accurately tuned and careful correction for irradiation and detection variations with frequency, and baseline effects are necessary. The gain

in resolution only applies to $I = \frac{5}{2}$ and $\frac{9}{2}$.

-25-

(B) DOUBLE ANGLE SPINNING

Physical background. DOR offers the most direct approach to averaging $P_2(\cos \theta)$ and $P_4(\cos \theta)$ terms simultaneously, by making the spinning axis a continually varying function of time [26]. To achieve this, DOR uses a spinning rotor (termed the inner rotor) which moves bodily by enclosing it in a spinning outer rotor, thereby forcing the axis of the inner rotor to describe a complicated but continuous trajectory as a function of time. The effect of this double rotation is to introduce modulation of the second-order quadrupolar frequency of the central transition in the laboratory frame of the form

$$H^{Q(2)}(t) = \frac{hC_Q^2}{8I(2I-1)\nu_0} [A_0(I, m, \eta_Q) + A_1(I, m, \eta_Q)P_2(\cos \beta_1)P_2(\cos \beta_2)F_1(\theta, \phi) + A_2(I, m, \eta_Q)P_4(\cos \beta_1)P_4(\cos \beta_2)F_2(\theta, \phi) + \text{terms } \propto \cos(\omega_1 t + \gamma_2)] \quad (\text{B1.12.14})$$

where β_1 and β_2 are the angles between the outer rotor and the magnetic field and the angle between the axes of the two rotors, respectively, and ϕ and θ describe the orientation of the principal axis system of a crystallite in the inner rotor. ω_1 is the angular frequency of the outer rotor and γ_2 an angle describing the position of the outer rotor in the laboratory frame. More details of the functions A and F can be found by comparison with [equation B1.12.10](#), [equation B1.12.11](#), [equation B1.12.12](#) and [equation B1.12.13](#) in [24]. β_1 and β_2 can be chosen so that $P_2(\cos \beta_1) = 0$ (54.74°) and $P_4(\cos \beta_2) = 0$ (30.56° or 70.15°).

Simulation of the complete DOR spectrum (centreband plus the spinning sidebands) will yield the NMR interaction parameters. However, it is most usual to perform the experiment to give improved resolution and simply quote the measured peak position, which appears at the sum of the isotropic chemical and second-order quadrupole shifts. DOR experiments at more than one applied magnetic field will allow these different isotropic contributions to be separated and hence provide an estimate of the quadrupole interaction. This approach is similar to that using the field variation of the centre of gravity of the MAS centreband ([equation B1.12.12](#)) but has the advantage that the narrower, more symmetric line makes determination of the correct position more precise. For experiments carried out at two magnetic fields where the Larmor frequencies are ν_{01} and ν_{02} for the measured DOR peak positions (in ppm) at the two magnetic fields at $\delta_{\text{dor}1,2}$ then

$$\delta_{\text{iso,cs}} = \frac{\nu_{01}^2 \delta_{\text{dor}1} - \nu_{02}^2 \delta_{\text{dor}2}}{\nu_{01}^2 - \nu_{02}^2} \quad (\text{B1.12.15})$$

and

$$\left[\frac{3(a - \frac{3}{4})}{40I^2(2I-1)^2} \right] C_Q^2 \left(1 + \frac{\eta_Q^2}{3} \right) = \frac{\nu_{01}^2 \nu_{02}^2 (\delta_{\text{dor}1} - \delta_{\text{dor}2})}{\nu_{01}^2 - \nu_{02}^2}. \quad (\text{B1.12.16})$$

Advantages. DOR works well if the quadrupolar interaction is dominant and the sample is highly crystalline, with some extremely impressive gains in resolution. Provided that the correct RF-excitation conditions are employed the spectral information is directly quantitative.

-26-

Disadvantages. The technique requires investment in a specialized, complex probe-head that requires considerable experience to use effectively. The relatively slow rotation speed of the large outer rotor can lead to difficulties in averaging strong homogeneous interactions and produces many closely spaced spectral sidebands. In disordered solids where there is a distribution of isotropic chemical shifts, quite broad sidebands can result that may coalesce at the slow rotation rates used. Currently, the maximum actual spinning speed that can be routinely obtained in the latest system with active computer control of the gas pressures is ~1500 Hz. By the use of synchronous triggering [28] this effectively amounts to a spinning speed of 3 kHz. Undoubtedly the technology associated with the technique will continue to improve leading to increased spinning speeds and thus expanding the application of the technique. Also, the RF coil encloses the whole system, and the filling factor is consequently small leading to relatively low sensitivity. The large coil size also means that the RF generated is quite low and that double tuning for CP is difficult, although such an experiment has been performed.

(C) DYNAMIC ANGLE SPINNING

Physical background. DAS is a 2D NMR experiment where the evolution of the magnetization is divided into two periods and the sample is spun about a different axis during each period [27, 29]. During the first evolution time t_1 the sample is spun at an angle of θ_1 . The magnetization is then stored along the z -axis and the angle of the spinning axis is changed to θ_2 . After the rotor is stabilized (~30 ms) the magnetization is brought into the xy -plane again and a signal is acquired. The second-order quadrupole frequency of an individual crystallite depends on the angle of the spinning axis. So during t_1 the quadrupole frequency will be ν_1 , and ν_2 during t_2 . If ν_2 is of opposite sign to ν_1 the magnetization from the crystallite will be at its starting position again at some time during t_2 . One can choose both angles in such a way that the signals from each individual crystallite will be at the starting position at exactly the same time. In other words, an echo will form and the effect of the second-order quadrupolar broadening is removed at the point of echo formation. To achieve this cancellation to form an echo then

$$P_2(\cos \theta_1) = -k P_2(\cos \theta_2) \quad \text{and} \quad P_4(\cos \theta_1) = -k P_4(\cos \theta_2) \quad (\text{B1.12.17})$$

must both be satisfied where k is the scaling factor. There is a continuous set of solutions for θ_1 and θ_2 , the so-called DAS complementary angles, and each set has a different scaling factor. For these solutions, the second-order quadrupole powder pattern at θ_1 is exactly the scaled mirror image of the pattern at θ_2 and an echo will form at $t_2 = kt_1$. For the combination $\theta_1 = 30.56^\circ$, $\theta_2 = 70.12^\circ$ the $P_4(\cos \theta)$ terms are zero and the scaling factor $k = 1.87$. A practically favoured combination is $\theta_1 = 37.38^\circ$, $\theta_2 = 79.19^\circ$ as the scaling factor $k = 1$ and the spectra are exact mirror images so that an echo will form at $t_1 = t_2$. There are several ways in which the DAS spectra can be acquired; one can acquire the entire echo, which means that the resulting 2D spectrum will be sheared. Some additional processing is then required to obtain an isotropic spectrum in $F1$. The acquisition could also start at the position of the echo so that an isotropic spectrum in $F1$ is obtained directly. A third possibility is to carry the experiment out as a pseudo 1D experiment where only the top of the echo is acquired as a function of t_1 . In this case the isotropic spectrum is acquired directly but there is no saving in the duration of the experiment.

Advantages. Compared to DOR, a small rotor can be used allowing relatively fast spinning speeds; high RF powers can be attained and if the coil is moved with the rotor a good filling factor can be obtained. In the isotropic dimension high-resolution spectra are produced and the second dimension retains the anisotropic information.

Disadvantages. Again, a specialized probe-head is necessary to allow the rotor axis to be changed, usually using a stepper motor and a pulley system. Angle switching needs to be as fast as possible and reproducible.

For an RF coil that changes orientation with the rotor the RF field will vary with the angle and the RF leads need to be flexible and resistant to metal fatigue. A major limitation of the DAS technique is that it cannot be used on compounds with a short t_1 because of the time needed to reorient the spinning axis. If magnetization is lost through t_1 and/or spin diffusion effects the signal can become very weak. This factor has meant that DAS has been most useful on ^{17}O where the dipole–dipole interaction is weak but has not had the impact on NMR of nuclei such as ^{27}Al that it might otherwise have had.

(D) MQMAS

Physical background. Relatively recently (1995) a new experiment emerged that has already had a great impact on solid-state NMR spectroscopy of quadrupolar nuclei [30]. The 2D multiple quantum magic angle spinning (2D MQMAS) experiment greatly enhances resolution of the spectra of half-integer spin quadrupolar nuclei. Basically, this experiment correlates the $(m, -m)$ multiple quantum transition to the $(\frac{1}{2}, -\frac{1}{2})$ transition. The resolution enhancement stems from the fact that the quadrupole frequencies for both transitions are correlated. At specific times the anisotropic parts of the quadrupole interaction are refocused and an echo forms. The frequency of an $(m, -m)$ transition is given by

$$\nu_p = C_0(p)\nu_0^Q - \frac{7}{18}C_4(p)\nu_4^Q(\theta, \phi) \quad (\text{B1.12.18})$$

where $p = 2m$ is the order of the coherence, $p = 1$ for the $(\frac{1}{2}, -\frac{1}{2})$ transition, $p = 3$ for the $(\frac{3}{2}, -\frac{3}{2})$ transition, etc. The coefficients are defined as

$$\begin{aligned} C_0(p) &= p \left(I(I+1) - \frac{3}{4}p^2 \right) \\ C_4(p) &= p \left(18I(I+1) - \frac{34}{4}p^2 - 5 \right). \end{aligned} \quad (\text{B1.12.19})$$

The isotropic part ν_0^Q of the quadrupole frequency is

$$\nu_0^Q = -(\nu_q^2(3 + \eta_q^2))/90\nu_0 \quad [\text{Hz}] \quad (\text{B1.12.20})$$

and the anisotropic part $\nu_4^Q(\theta, \phi)$ is given by

$$\begin{aligned} \nu_4^Q(\theta, \phi) &= (\nu_q^2/112\nu_0) \times [(7/18)(3 - \eta_q \cos 2\phi)^2 \sin^4 \theta \\ &\quad + (2\eta_q \cos 2\phi - 4 - (2/9)\eta_q^2) \sin^2 \theta + (2/45)\eta_q^2 + (4/5)]. \end{aligned} \quad (\text{B1.12.21})$$

Numerous schemes exist that are used to obtain 2D MQMAS spectra [24]. The simplest form of the experiment is when the MQ transition is excited by a single, high-power RF pulse, after which the MQ coherence is allowed to evolve for a time t_1 (figure B1.12.12). After the evolution time, a second pulse is applied which converts the MQ coherence into a $p = -1$ coherence which is observed during t_2 . The signal is then acquired immediately after the second pulse and the echo will form at a time $t_2 = |QA|t_1$. Both pulses are non-selective and will excite all coherences to a varying degree. Selection of the coherences of interest is achieved by cycling the pulse sequence through the appropriate phases. Phase cycles can be easily worked out

by noting that an RF phase shift of ϕ degrees is seen as a phase shift of $p\phi$ degrees for a p -quantum coherence [31]. After a 2D Fourier transformation the resonances will show up as ridges lying along the quadrupole anisotropy (QA) axis. The isotropic spectrum can be obtained by projection of the entire 2D spectrum on a line through the origin ($\nu_1 = \nu_2 = 0$) perpendicular to the QA axis. figure B1.12.12 shows some of the many different pulse sequences and their coherence pathways that can be used for the 2D MQMAS experiment.

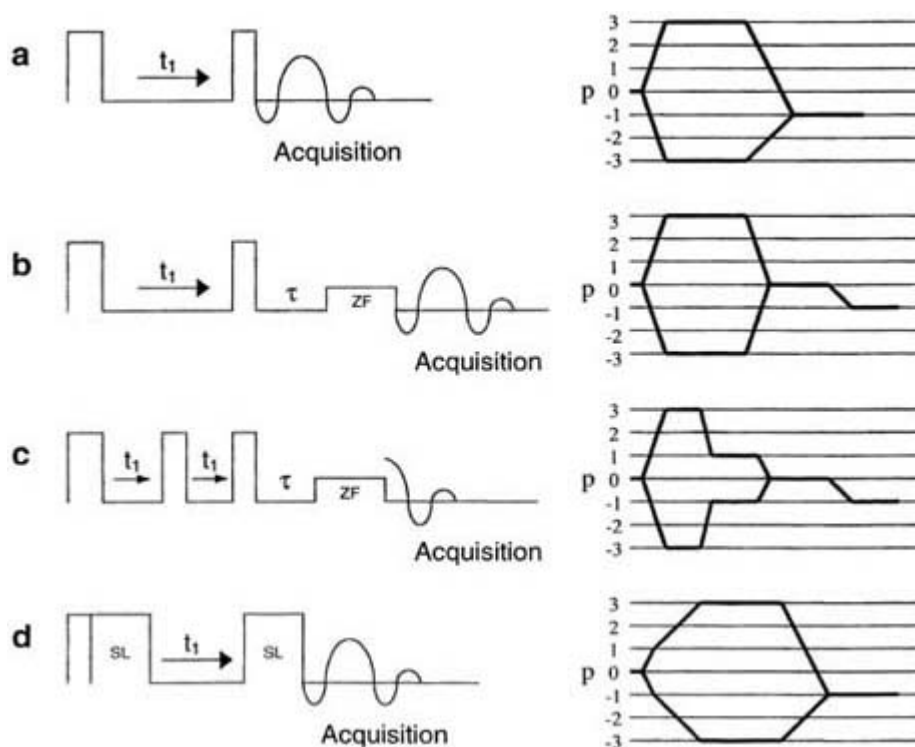


Figure B1.12.12. Pulse sequences used in multiple quantum MAS experiments and their coherence pathways for (a) two-pulse, (b) z-filter, (c) split- t_1 with z-filter and (d) RIACT (II).

The isotropic shift and the quadrupole-induced shift (QIS) can easily be obtained from the data. The QIS is different in both dimensions, however, and is given by

$$\delta_{\text{qis}}^p = \frac{C_0(p)v_0^Q}{p\nu_0} \times 10^6 \quad [\text{in ppm}]. \quad (\text{B1.12.22})$$

-29-

δ_{cg}^p , the position of the centre of gravity with $p = 1$ for the single quantum dimension and $p = \Delta m$ for the multiple quantum dimension, is given by

$$\delta_{\text{cg}}^p = \delta_{\text{iso}} + \delta_{\text{qis}}^p. \quad (\text{B1.12.23})$$

Hence the isotropic shift can be easily retrieved using

$$\delta_{\text{iso}} = \frac{C_0(p)\delta_{\text{cg}}^{p=1} - pC_0(1)\delta_{\text{cg}}^p}{C_0(p) - pC_0(1)} \quad [\text{in ppm}] \quad (\text{B1.12.24})$$

and the isotropic quadrupolar shift by

$$\nu_0^Q = \frac{(\delta_{\text{cg}}^{p=1} - \delta_{\text{cg}}^p)}{C_0(1) - C_0(p)/p} \nu_0 \quad \text{[in Hz].} \quad (\text{B1.12.25})$$

Shearing of the data is performed to obtain isotropic spectra in the $F1$ dimension and to facilitate easy extraction of the 1D slices for different peaks. Shearing is a projection of points that lie on a line with a slope equal to the anisotropy axis onto a line that is parallel to the $F2$ axis [24]. Shearing essentially achieves the same as the split- t_1 experiment or delayed acquisition of the echo. Although sheared spectra may look more attractive, they do not add any extra information and they are certainly not necessary for the extraction of QIS and δ_{iso} values.

Advantages. The experiment can be readily carried out with a conventional probe-head, although the fastest spinning and highest RF powers available are useful. The pulse sequences are relatively easy to set up (compared to DAS and DOR) and the results are usually quite straightforward to interpret in terms of the number of sites and determination of the interactions.

Disadvantages. A researcher new to the subject will be confronted with a large number of schemes for collecting, processing and presenting the data which can be very confusing. The relationship between the measured peak positions and the NMR interaction parameters crucially depends on the processing and referencing conventions adopted. There is one clear distinction between the two main approaches: either the MQ evolution is regarded as having taken place only in the evolution time (t_1), or the period up to the echo is also regarded as being part of the evolution time which is then $(1 + QA)t_1$. A detailed critique of these two approaches and the consequences of adopting each has recently been given by Man [32]. Shearing data introduces an extra processing step, which may introduce artefacts. The key point in determining the quadrupole parameters is the accuracy of measuring the position of the centre of gravity of the resonance. Both the excitation efficiency as well as significant intensity in the spinning sidebands can adversely affect the accuracy of determining the centre of gravity. MQ spectra show a strong dependence on excitation efficiency that is strongly dependent on the value of C_Q relative to the RF field strength. This means that often the spectra are non-quantitative and sites with a large C_Q can be lost completely from the spectrum.

(E) APPLICATION OF HIGH-RESOLUTION METHODS TO ^{27}Al IN KYANITE

Kyanite, a polymorph of Al_2SiO_5 , has four distinct AlO_6 sites. The crystal structure and site geometries (e.g. Al–O bondlengths) are well characterized. The quadrupole interactions vary from 3.6 to 10.1 MHz, providing a good test of the different approaches to achieving high-resolution NMR spectra from quadrupolar nuclei. Both single-crystal studies [33] and NQR [34] have accurately determined the quadrupole interaction parameters. MAS studies of the central transition have been reported from 7 to 18.8 T [35, 36, 37 and 38]. An example of a 17.55 T spectrum is shown in figure B1.12.13(a) along with a simulation of the individual components and their sum. It can be clearly seen that at even this high field there is extensive overlap of the four components. However, many distinct spectral features exist that constrain the simulation and these can be followed as a function of applied magnetic field. As the field is increased the second-order quadrupole effects decrease. Comparing simulations at 4.7 and 9.4 T (figure B1.12.13(b)) the field variation simply scales the width of the spectrum, being a factor of two narrower (in hertz) at the higher field. By extending the simulation to higher fields, 18 T can be seen to be a poor choice as there is considerable overlap of the resonances with virtually no shift dispersion between the sites. 135.3 T would provide a completely resolved spectrum under simple MAS.

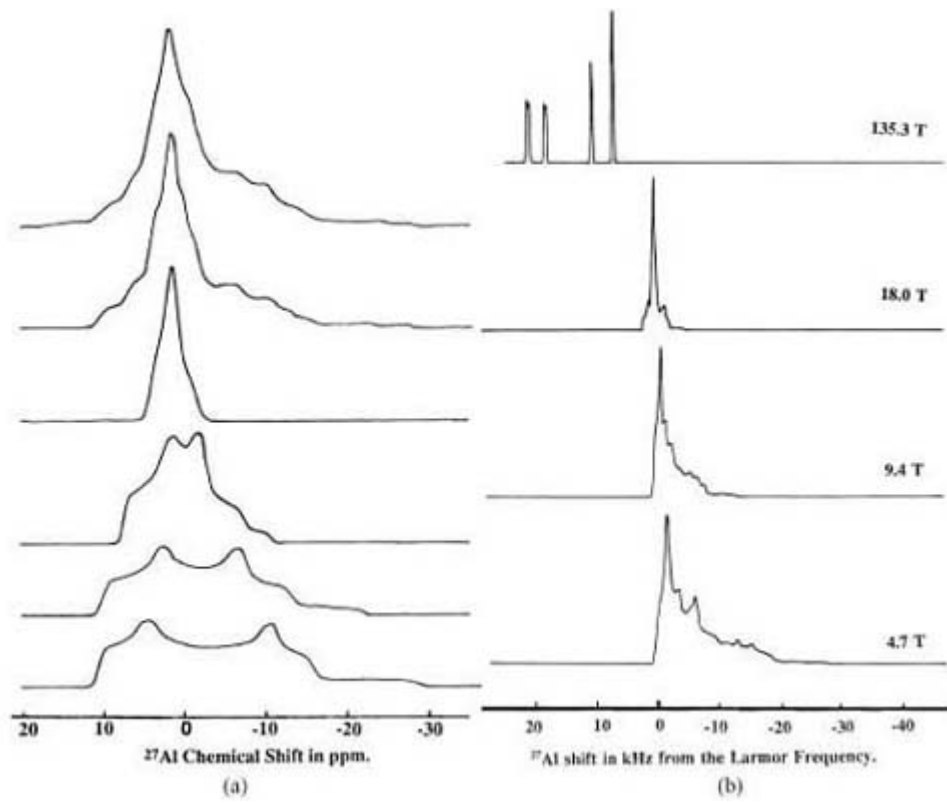


Figure continued on next page

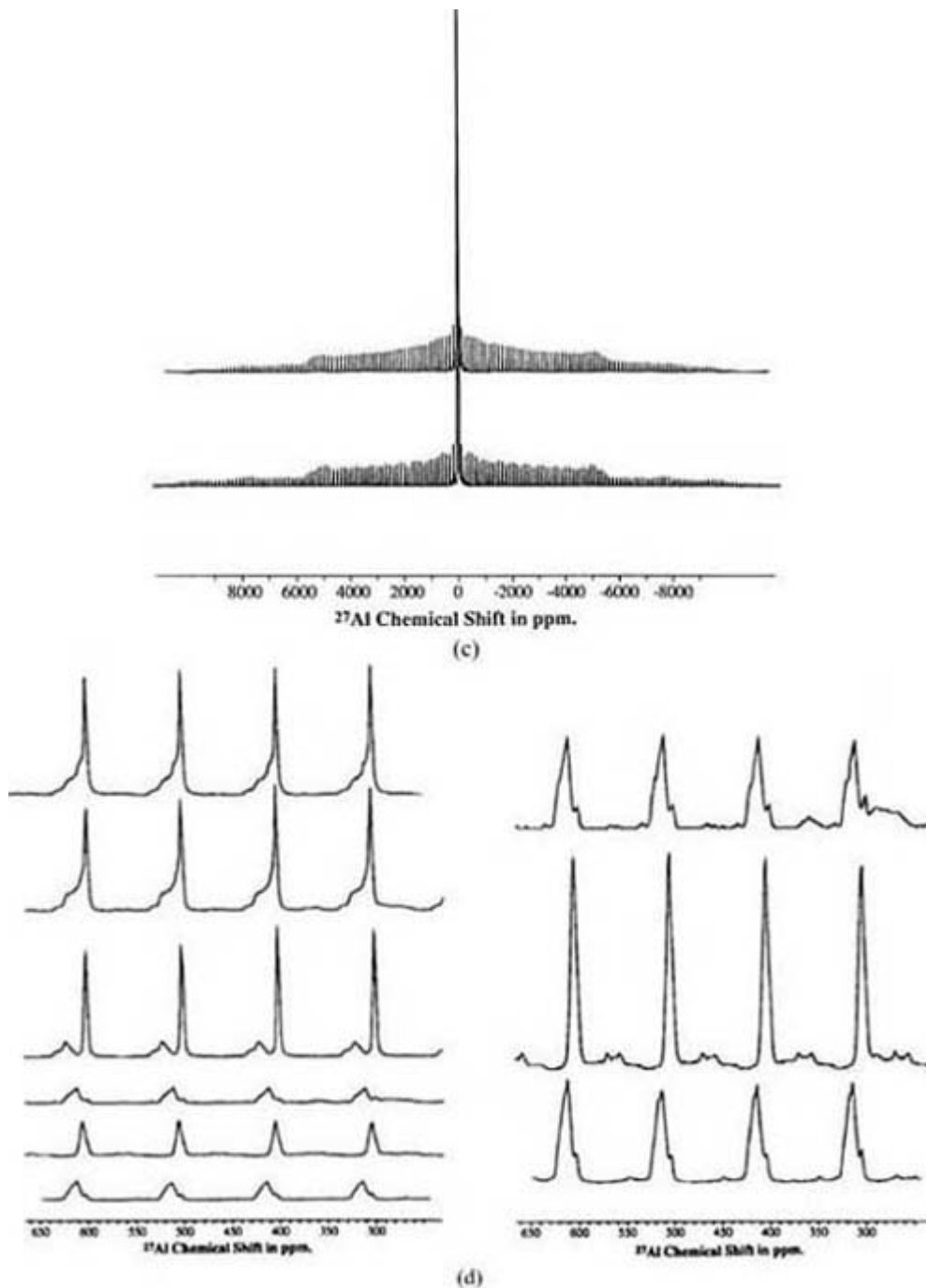


Figure B1.12.13. ^{27}Al MAS NMR spectra from kyanite (a) at 17.55 T along with the complete simulation and the individual components, (b) simulation of centreband lineshapes of kyanite as a function of applied magnetic field, and the satellite transitions showing (c) the complete spinning sideband manifold and (d) an expansion of individual sidebands and their simulation.

Satellite transition MAS NMR provides an alternative method for determining the interactions. The intensity envelope of the spinning sidebands are dominated by site A2 (using the crystal structure nomenclature) which has the smallest C_Q , resulting in the intensity for the transitions of this site being spread over the smallest range ($\propto C_Q$), and will have the narrowest sidebands ($\propto C_Q^2/\nu_0$) of all the sites (figure B1.12.13(c)) [37]. The simulation of this envelope provides additional constraints on the quadrupole interaction parameters for this site. Expanding the sidebands (shown for the range 650 to 250 ppm in figure B1.12.13(d)) reveals distinct second-order lineshapes with each of the four sites providing contributions from the $(\pm\frac{1}{2}, \pm\frac{3}{2})$ and $(\pm\frac{3}{2}, \pm\frac{5}{2})$

transitions. The improved resolution provided by the $(\pm\frac{1}{2}, \pm\frac{3}{2})$ transition compared to the central transition is clear. For the A2 site both of the contributing transitions are clearly seen while for the other three only the $(\pm\frac{1}{2}, \pm\frac{3}{2})$ contribution is really observable until the vertical scale is increased by six, which then shows the outer transition for all sites, especially A3. The simulation of all three transitions provides an internal check on the interaction parameters for each site.

DOR provides significantly higher resolution than MAS [37, 39]. At 11.7 T a series of relatively narrow resonances and accompanying sidebands are observed under DOR (figure B1.12.14(a)). The relatively slow spinning speed of the outer rotor results in numerous sidebands and the isotropic line is identified by collecting spectra at several different spinning speeds. If the isotropic position is then collected as a function of B_0 the quadrupole interaction parameters can then be deduced. MQ MAS provides an alternative approach for producing high resolution [38, 40], with the whole 2D data set shown at 9.4 T along with the isotropic projections at 11.7 and 18.8 T (figure B1.12.14(b)) [38]. All three isotropic triple quantum (3Q) projections show only three resolved lines as the NMR parameters from the two sites (A1, A4) with the largest C_Q means that their resonances are superimposed at all fields. This is confirmed by the 9.4 T 3Q data where an RF field of 280 kHz was employed making the data more quantitative: the three resonances with isotropic shifts of 43.0, 21.1 and 8.0 ppm had intensities of 2.1:1.1:1.0 respectively. At 11.7 and 18.8 T the MQ MAS NMR spectra collected are not quantitative since the RF fields to excite the 3Q transitions were not strong enough. For $I = \frac{5}{2}$ the isotropic shifts are $-\frac{10}{17}$ of the value compared to direct MAS at the same field, so MQ data effectively produces results at a 'negative' applied magnetic field thereby more strongly constraining the NMR interaction parameters deduced from isotropic shift against B_0^{-2} plots.

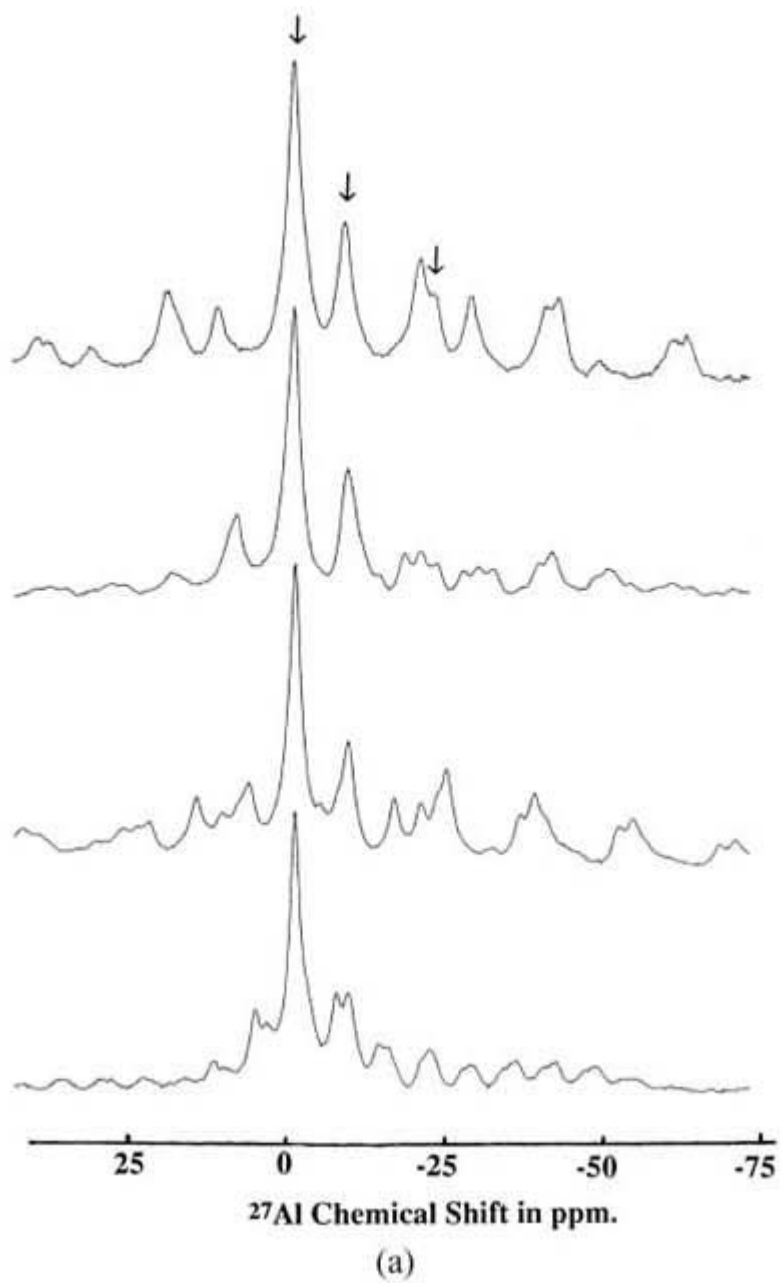


Figure continued on next page

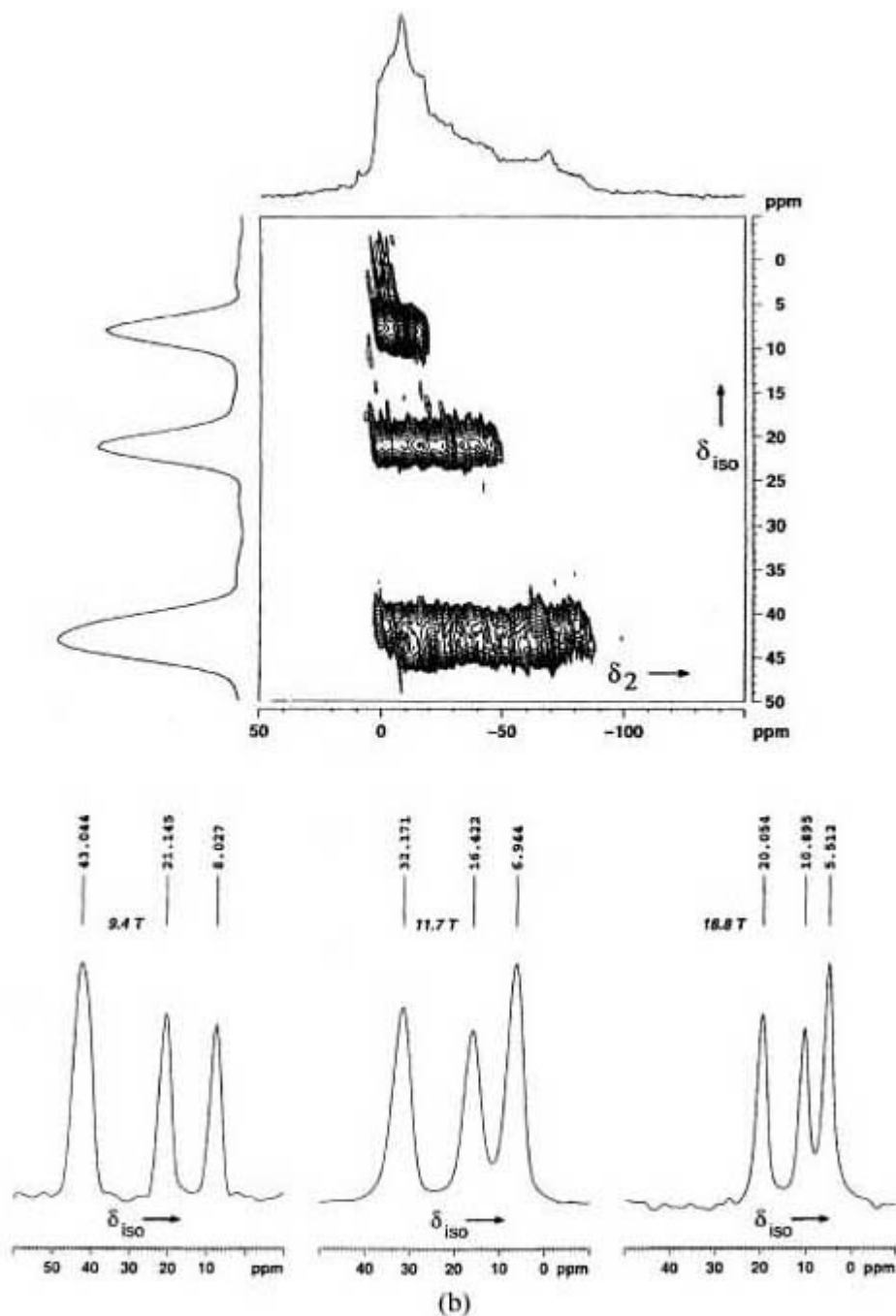


Figure B1.12.14. (a) ^{27}Al DOR NMR spectrum of kyanite at 11.7 T at various spinning speeds and (b) the ^{27}Al MQ MAS NMR spectrum of kyanite at 9.4 T (top) along with the isotropic projections at 11.7 and 18.8 T. (The MQ MAS NMR data are taken from [38] with the permission of Academic Press.)

When ^{27}Al MAS NMR spectra were first collected at moderate B_0 with relatively slow MAS rates (<4 kHz) there was much confusion about the quantitative integrity of such spectra. In fact, provided the correct excitations are employed (i.e. $\nu_1 \gg \nu_r$ and $2\pi\nu_1 T_p \ll 1$) all that is necessary to make MAS NMR spectra quantitative is to know what fraction of the different transitions are contributing to the centreband. In practice this usually amounts to estimating the fraction of the $(\frac{1}{2}, -\frac{1}{2})$ transition that contributes to the centreband, which depends on $\nu_Q^2/\nu_0\nu_r$ [41]. At 17.55 T the intensity distribution between the sites in the centreband was 23%:27%:26%:24% (A1:A2:A3:A4). Once the correction factors are taken into account, within experimental error all sites are equally populated. DAS is not reported for kyanite because of its shortcomings for fast-

relaxing nuclei such as ^{27}Al . The data on kyanite demonstrates how accurately the interaction parameters and intensities can be extracted for quadrupole nuclei by using a combination of advanced techniques.

B1.12.4.6 DISTANCE MEASUREMENTS, DIPOLAR SEQUENCES AND CORRELATION EXPERIMENTS

The dipolar coupling between two nuclei I, S is $(\mu_0/4\pi)(\hbar/2\pi)\gamma_I\gamma_S r_{IS}^{-3}$, thus in systems containing pairs of nuclei measurement of the dipolar coupling will give the distance between them. The simplest pulse sequence for doing this, SEDOR (spin echo double resonance) [2], is shown in [figure B1.12.15\(a\)](#). A normal ‘spin echo’ sequence is applied to spin I which refocuses the heteronuclear dipole coupling and the chemical shift anisotropy producing an echo at 2τ . However, if a 180° pulse is applied to the S spin system it will invert the sign of any IS dipole coupling and reduce the echo intensity. By varying the 180° pulse position a set of difference signals can be used to determine the dipolar coupling and hence the internuclear distance for an isolated spin pair system. The SEDOR sequence is only useful for static samples and the range of distances accessible is restricted by T_2 , however, it can give useful qualitative information in systems where the spins are not isolated pairs.

MAS averages the heteronuclear dipole interaction giving a much increased resolution and increased T_2 . The REDOR (rotational echo double resonance) [42] sequence enables the heteronuclear dipole coupling to be measured for a spinning sample. There are several versions of this experiment; in one ([figure B1.12.15\(b\)](#)) a rotor-synchronized echo is applied to I whilst a series of rotor-synchronized 180° pulses are applied to S . As with SEDOR, the attenuated signal is subtracted from the normal echo signal to determine the REDOR fraction. In suitably labelled systems dipolar couplings as small as 25 Hz have been measured corresponding to a distance of 6.7 Å [43].

TEDOR (transferred echo double resonance) [44, 45] is another experiment for measuring internuclear distances whilst spinning. In this ([figure B1.12.15\(c\)](#)) the S spin is observed. There is first a REDOR sequence on the I spin, the magnetization is then transferred by applying 90° pulses on both spins and this is followed by a REDOR sequence on the S spins. Both REDOR and TEDOR require accurate setting of the 180° pulse and good long-term spectrometer stability to be effective. An advantage of the TEDOR sequence is that there is no background signal from uncoupled spins; however, the theoretical maximum efficiency is 50% giving a reduced signal. Whilst both REDOR and TEDOR can work well for spin $I, S = \frac{1}{2}$ pairs, for quadrupolar nuclei the situation is more complex [24]. The echo sequence must be applied to the quadrupolar nucleus if only one of the pair has spin $I > \frac{1}{2}$ since accurate inversion is rarely possible for spins $> \frac{1}{2}$. A sequence designed specifically for a quadrupolar nucleus is TRAPDOR ([figure B1.12.15\(d\)](#)) in which the dephasing spin is always quadrupolar. One cannot obtain accurate values of the dipolar coupling with this sequence but it can be used to give qualitative information about spatial proximity. There are many other pulse sequences designed to give distance information via the dipole–dipole interaction, most of limited applicability. One, designed specifically for homonuclei, that works well in labelled compounds is DRAMA [46] and a comparison of the more popular homonuclear sequences is given in [47].

Figure B1.12.15. Some double-resonance pulse sequences for providing distance information in solids: (a) SEDOR, (b) REDOR, (c) TEDOR and (d) TRAPDOR. In all sequences the narrow pulses are 90° and the wide pulses 180° . For sequences that employ MAS the number of rotor cycles (N_c) is shown along the bottom.

In liquid-state NMR two-dimensional correlation of spectra has provided much useful information. However, in solids the linewidths (T_2), even under MAS, are usually such that experiments like COSY and INADEQUATE which rely on through-bond J coupling cannot be used (although there are some notable exceptions). The most commonly used heteronuclear correlation (HETCOR) experiments in solids all rely on dipolar coupling and can thus be complicated by less local effects. Nevertheless, they are able to correlate specific sites in the MAS spectrum of one nucleus with sites of the second nucleus which are nearby. The simplest versions of the experiment are just a two-dimensional extension of CP in which the pulse that generates magnetization is separated from the matching pulses by a time which is incremented to give the second dimension. An example of the usefulness of the technique is shown in figure B1.12.16 where the ^1H - ^{31}P correlation of two inorganic hydrated phosphates, (a) brushite and (b) a bone, are shown [48]. In both cases two phosphorus sites that completely overlap in the 1D MAS spectrum are clearly visible in the 2D spectrum.

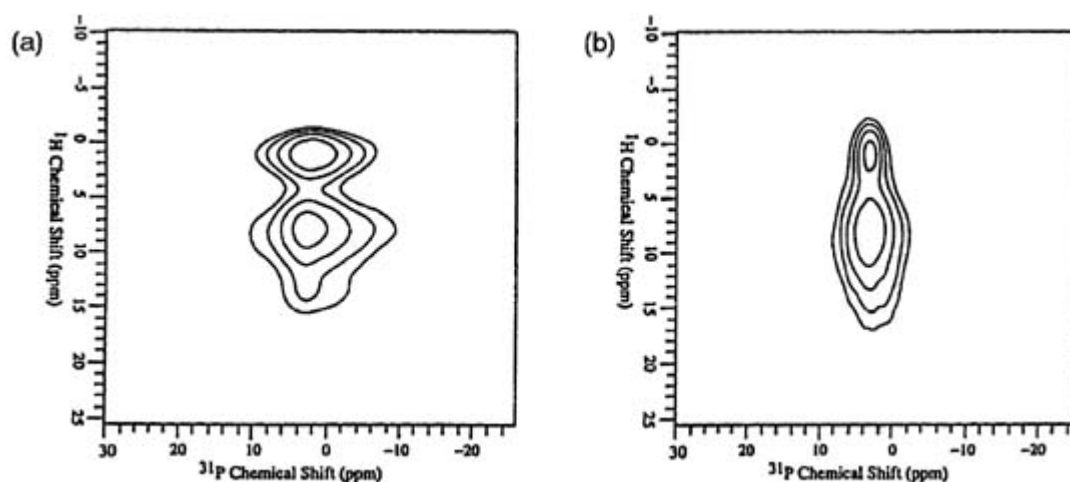


Figure B1.12.16. ^1H - ^{31}P HETCOR NMR spectra from (a) brushite and (b) bone. (Adapted from [48] with the permission of Academic Press.)

REFERENCES

- [1] Abragam A 1983 *Principles of Nuclear Magnetism* (Oxford: Oxford University Press)
- [2] Slichter C P 1990 *Principles of Magnetic Resonance* (Berlin: Springer)
- [3] Harris R K 1984 *NMR Spectroscopy* (London: Pitman)
- [4] Spiess H W and Schmidt-Rohr K *Multidimensional NMR of Polymers* (New York: Academic)
- [5] Cohen M H and Reif F 1954 *Solid State Phys.* **5** 736
- [6] Gerstein B C and Dybowski C 1985 *Transient Techniques in NMR of Solids* (New York: Academic)
- [7] Sternheimer R M 1958 *Phys. Rev.* **37** 736
- [8] Fukushima E and Roeder S B W 1981 *Experimental Pulse NMR* (Reading, MA: Addison-Wesley)
- [9] Kentgens A P M 1991 *J. Magn. Reson.* **95** 619
- [10] Liao M Y, Chew B G M and Zax D B 1995 *Chem. Phys. Lett.* **242** 89

- [11] Kunath-Fandrei G 1998 *PhD Thesis* University of Jena (Aachen: Shaker)
- [12] Laukien D D and Tschopp W H 1993 *Concepts in Magnetic Resonance* **6** 255

- [13] Doty F D, Connick T J, Ni X Z and Clingan M N 1988 *J. Magn. Reson.* **77** 536
- [14] Derome A E 1987 *Modern NMR Techniques for Chemistry Research* (Oxford: Pergamon)
- [15] Kunwar A C, Turner G L and Oldfield E 1986 *J. Magn. Reson.* **69** 124
- [16] Bastow T J and Smith M E 1994 *Solid State NMR* **3** 17
- [17] Wu X, Juban E A and Butler L G 1994 *Chem. Phys. Lett.* **221** 65
- [18] Poplett I J F and Smith M E 1998 *Solid State NMR* **11** 211
- [19] Kentgens A P M 1998 *Prog. NMR Spectrosc.* **32** 141
- [20] Herzfeld J and Berger A E 1980 *J. Chem. Phys.* **73** 6021
- [21] Smith M E 1993 *Appl. Magn. Reson.* **4** 1
- [22] Freude D and Haase J 1993 *NMR Basic Principles and Progress* vol 29, ed P Diehl *et al* (Berlin: Springer) p 1
- [23] Kentgens A P M 1997 *Geoderma* **80** 271
- [24] Smith M E and van Eck E R H 1999 *Prog. NMR Spectrosc.* **34** 159
- [25] Samoson A 1985 *Chem. Phys. Lett.* **119** 29
- [26] Jaeger C 1994 *NMR Basic Principles and Progress* ed B Blumich and R Kosfeld (Berlin: Springer) vol 31, p 135
- [27] Zwanziger J and Chmelka B F 1994 *NMR Basic Principles and Progress* ed B Blumich and R Kosfeld (Berlin: Springer) vol 31, p 202
- [28] Samoson A and Lippmaa E 1989 *J. Magn. Reson.* **89** 410
- [29] Mueller K T, Sun B Q, Chingas G C, Zwanziger J W, Terao T and Pines A 1990 *J. Magn. Reson.* **86** 470
- [30] Medek A, Harwood J S and Frydman L 1995 *J. Am. Chem. Soc.* **117** 12 779
- [31] Ernst R R, Bodenhausen G and Wokaun A 1987 *Principles of NMR in One and Two Dimensions* (Oxford: Clarendon)
- [32] Man P P 1998 *Phys. Rev. B* **58** 2764
- [33] Hafner S S and Raymond M 1967 *Am. Mineral.* **52** 1632
- [34] Lee D and Bray P J 1991 *J. Magn. Reson.* **94** 51
- [35] Lippmaa E, Samoson A and Magi M 1986 *J. Am. Chem. Soc.* **108** 1730
- [36] Alemany L B, Massiot D, Sherriff B L, Smith M E and Taulelle F 1991 *Chem. Phys. Chem.* **117** 301
- [37] Smith M E, Jaeger C, Schoenhofer R and Steuernagel S 1994 *Chem. Phys. Lett.* **219** 75
- [38] Alemany L B, Steuernagel S, Amoureux J-P, Callender R L and Barron A R 1999 *Chem. Phys. Lett.* **14** 1
- [39] Xu Z and Sherriff B L 1993 *Appl. Magn. Reson* **4** 203
- [40] Baltisberger J H, Xu Z, Stebbins J F, Wang S H and Pines A 1996 *J. Am. Chem. Soc.* **118** 7209
- [41] Massiot D, Bessada C, Coutures J P and Taulelle F 1990 *J. Magn. Reson.* **90** 231
- [42] Gullion T and Schaefer J 1989 *Adv. Magn. Reson.* **13** 57
- [43] Merritt M E, Goetz J, Whitney D, Chang C P P, Heux L, Halary J L and Schaefer J 1998 *Macromolecules* **31** 1214
- [44] Van Eck E R H and Veeman W S 1993 *Solid State NMR* **2** 307
- [45] Hing A W, Vega S and Schaefer J 1993 *J. Magn. Res. A* **103** 151
- [46] Tycko R and Dabbagh G 1990 *Chem. Phys. Lett.* **173** 461
- [47] Baldus M, Geurts D G and Meier B H 1998 *Solid State NMR* **11** 157
- [48] Santos R A, Wind R A and Bronniman C E 1990 *J. Magn. Reson. B* **105** 183

FURTHER READING

Stejskal E O and Memory J D 1994 *High Resolution NMR in the Solid State* (Oxford: Oxford University Press)

Introductory text, fairly mathematical, concentrates on spin $I = \frac{1}{2}$ systems, good references.

Slichter C P 1989 *Principles of Magnetic Resonance* 3rd edn (Berlin: Springer)

Comprehensive coverage from a physics viewpoint.

Engelhardt G and Michel D 1987 *High Resolution Solid State NMR of Silicates and Zeolites* (New York: Wiley)

Good coverage of NMR in these solids up to 1987.

Blümich B *et al* (eds) 1994 *Solid State NMR* vol 1 (Berlin: Springer)

NMR basic principles and progress; specialised monograph giving detailed descriptions of specific areas of solid state NMR.

Schmidt-Rohr K and Spiess H W 1994 *Multidimensional Solid State NMR and Polymers* (New York: Academic)

A comprehensive text which discusses advanced experiments with particular reference to polymers.

Fitzgerald J J (ed) 1999 *Solid State NMR Spectroscopy of Inorganic Materials* (Washington, DC: American Chemical Society)

Gives a range of current examples of solid state NMR applied to inorganic materials.

Grant D M and Harris R K (eds) 1996 *Encyclopaedia of NMR* (New York: Wiley)

Contains articles on all aspects of NMR.

-1-

B1.13 NMR relaxation rates

Jozef Kowalewski

B1.13.1 INTRODUCTION

Nuclear magnetic resonance (NMR) spectroscopy deals with the interactions among nuclear magnetic moments (nuclear spins) and between the spins and their environment in the presence of a static magnetic field B_0 , probed by radiofrequency (RF) fields. The couplings involving the spins are relatively weak which results, in particular in low-viscosity liquids, in narrow lines and spectra with a rich structure. The NMR experiments are commonly carried out in the time domain: the spins are manipulated by sequences of RF pulses and delays, creating various types of non-equilibrium spin state, and the NMR signal corresponding to magnetization components perpendicular to B_0 is detected as a function of time (the free induction decay, FID). In terms of the kinetics, the weakness of the interactions results in slow decays (typically milliseconds to seconds) of non-equilibrium states. The recovery processes taking the ensembles of nuclear spins back to equilibrium and some related phenomena, are called *nuclear spin relaxation*. The relaxation behaviour of nuclear spin systems is both an important source of information about the molecular structure and dynamics and a factor to consider in optimizing the design of experiments.

The concept of 'relaxation time' was introduced to the vocabulary of NMR in 1946 by Bloch in his famous equations of motion for nuclear magnetization vector \mathbf{M} [1]:

(B1.13.1)

$$\frac{dM_z}{dt} = \frac{M_0 - M_z}{T_1}$$

$$\frac{dM_{x,y}}{dt} = -\frac{M_{x,y}}{T_2}$$

The phenomenological Bloch equations assume the magnetization component along B_0 (the longitudinal magnetization M_z) to relax exponentially to its equilibrium value, M_0 . The time constant for the process is called the spin–lattice or longitudinal relaxation time, and is denoted T_1 . The magnetization components perpendicular to B_0 (the transverse $M_{x,y}$ magnetization, $M_{x,y}$) are also assumed to relax in an exponential manner to their equilibrium value of zero. The time constant for this process is called the spin–spin or transverse relaxation time and is denoted T_2 . The inverse of a relaxation time is called a relaxation rate.

The first microscopic theory for the phenomenon of nuclear spin relaxation was presented by Bloembergen, Purcell and Pound (BPP) in 1948 [2]. They related the spin–lattice relaxation rate to the transition probabilities between the nuclear spin energy levels. The BPP paper constitutes the foundation on which most of the subsequent theory has been built, but contains some faults which were corrected by Solomon in 1955 [3]. Solomon noted also that a correct description of even a very simple system containing two interacting spins, requires introducing the concept of cross-relaxation, or magnetization exchange between the spins. The subsequent development has been rich and the goal of this entry is to provide a flavour of relaxation theory (section b1.13.2), experimental techniques (section b1.13.3) and applications (section b1.13.4).

-2-

The further reading list for this entry contains five monographs, a review volume and two extensive reviews from the early 1990s. The monographs cover the basic NMR theory and the theoretical aspects of NMR relaxation. The review volume covers many important aspects of modern NMR experiments in general and relaxation measurements in particular. The two reviews contain more than a thousand references to application papers, mainly from the eighties. The number of literature references provided in this entry is limited and, in particular in the theory and experiments sections, priority is given to reviews rather than to original articles.

B1.13.2 RELAXATION THEORY

We begin this section by looking at the Solomon equations, which are the simplest formulation of the essential aspects of relaxation as studied by NMR spectroscopy of today. A more general Redfield theory is introduced in the next section, followed by the discussion of the connections between the relaxation and molecular motions and of physical mechanisms behind the nuclear relaxation.

B1.13.2.1 THE SOLOMON THEORY FOR A TWO-SPIN SYSTEM

Let us consider a liquid consisting of molecules containing two nuclei with the spin quantum number 1/2. We denote the two spins I and S and assume that they are distinguishable, i.e. either belong to different nuclear species or have different chemical shifts. The system of such two spins is characterized by four energy levels and by a set of transition probabilities between the levels, cf. Figure B1.13.1. We assume at this stage that the two spins interact with each other only by the dipole–dipole (DD) interaction. The DD interaction depends on the orientation of the internuclear axis with respect to B_0 and is thus changed (and averaged to zero on a sufficiently long time scale) by molecular tumbling. Taking this motional variation into consideration, it is quite straightforward to use time-dependent perturbation theory to derive transition probabilities between the pairs of levels in Figure B1.13.1. Briefly, the transition probabilities are related to spectral density functions

(*vide infra*), which measure the intensity of the local magnetic fields fluctuating at the frequencies corresponding to the energy level differences in [Figure B1.13.1](#). Solomon [3] showed that the relaxation of the longitudinal magnetization components, proportional to the expectation values of I_z and S_z operators, was related to the populations of the four levels and could be described by a set of two coupled equations:

$$\begin{aligned} \frac{d\langle I_z \rangle}{dt} &= -(W_0 + 2W_{1I} + W_2)(\langle I_z \rangle - I_z^0) - (W_2 - W_0)(\langle S_z \rangle - S_z^0) \\ \frac{d\langle S_z \rangle}{dt} &= -(W_2 - W_0)(\langle I_z \rangle - I_z^0) - (W_0 + 2W_{1S} + W_2)(\langle S_z \rangle - S_z^0) \end{aligned} \quad (\text{B1.13.2})$$

or

$$\begin{aligned} \frac{d\langle I_z \rangle}{dt} &= -\rho_I(\langle I_z \rangle - I_z^0) - \sigma_{IS}(\langle S_z \rangle - S_z^0) \\ \frac{d\langle S_z \rangle}{dt} &= -\sigma_{IS}(\langle I_z \rangle - I_z^0) - \rho_S(\langle S_z \rangle - S_z^0). \end{aligned} \quad (\text{B1.13.3})$$

-3-

I_z^0 and S_z^0 are the equilibrium magnetization for the two spins and ρ^I and ρ^S are the corresponding decay rates (spin–lattice relaxation rates). The symbol σ_{IS} denotes the cross-relaxation rate. The general solutions of [equation B1.13.2](#) or [equation B1.13.3](#) for $\langle I_z \rangle$ or $\langle S_z \rangle$ are sums of two exponentials, i.e. the longitudinal magnetizations in a two-spin system do not follow the Bloch equations. Solomon demonstrated also that the simple exponential relaxation behaviour of the longitudinal magnetization was recovered under certain limiting conditions.

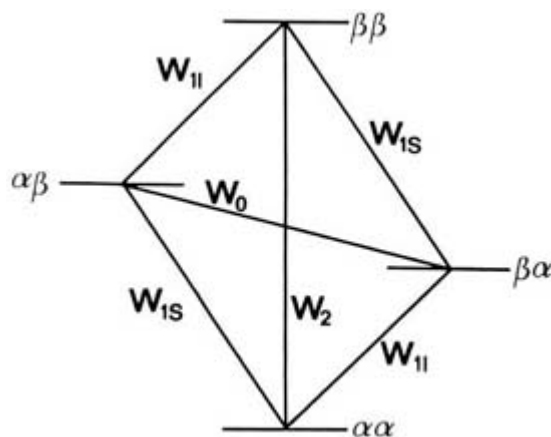


Figure B1.13.1. Energy levels and transition probabilities for an IS spin system. (Reproduced by permission of Academic Press from Kowalewski J 1990 *Annu. Rep. NMR Spectrosc.* **22** 308–414.)

(i) The two spins are identical (e.g. the two proton spins in a water molecule). We then have $W_{1I}=W_{1S}=W^1$ and the spin–lattice relaxation rate is given by:

$$T_1^{-1} = 2(W_1 + W_2). \quad (\text{B1.13.4})$$

(ii) One of the spins, say S , is characterized by another, faster, relaxation mechanism. We can then say that the

S spin remains in thermal equilibrium on the time scale of the I -spin relaxation. This situation occurs in paramagnetic systems, where S is an electron spin. The spin–lattice relaxation rate for the I spin is then given by:

$$\rho_I = T_{1I}^{-1} = W_0 + 2W_1 + W_2. \quad (\text{B1.13.5})$$

(iii) One of the spins, say S again, is saturated by an intense RF field at its resonance frequency. These are the conditions applying for e.g. carbon-13 (treated as the I spin) under broad-band decoupling of protons (S spins). The relaxation rate is also in this case given by equation B1.13.5. In addition, we then observe the phenomenon referred to as the (hetero)nuclear Overhauser enhancement (NOE), i.e. the steady-state solution for $\langle I_z \rangle$ is modified to

$$\langle I_z \rangle_{\text{steady state}} = I_z^0 + \frac{\sigma_{IS}}{\rho_I} S_z^0. \quad (\text{B1.13.6})$$

-4-

B1.13.2.2 THE REDFIELD THEORY

A more general formulation of relaxation theory, suitable for systems with scalar spin–spin couplings (J couplings), is known as the Wangness, Bloch and Redfield (WBR) theory or the Redfield theory [4]. In analogy with the Solomon theory, the Redfield theory is also based on the second-order perturbation theory, which in certain situations (unusual for nuclear spin systems in liquids) can be a limitation. Rather than dealing with the concepts of magnetizations or energy level populations in the Solomon formulation, the Redfield theory is given in terms of density operator (for a general review of the density operator formalism, see ‘Further reading’). Briefly, the density operator describes the average behaviour of an ensemble of quantum mechanical systems. It is usually expressed by expansion in a suitable operator basis set. For the discussion of the IS system above, the appropriate 16-dimensional basis set can e.g. consist of the unit operator, E , the operators corresponding to the Cartesian components of the two spins, $I_x, I_y, I_z, S_x, S_y, S_z$ and the products of the components of I and the components of S . These 16 operators span the Liouville space for our two-spin system. If we concentrate on the longitudinal relaxation (the relaxation connected to the distribution of populations), the Redfield theory predicts the relaxation to follow a set of three coupled differential equations:

$$\frac{d}{dt} \begin{pmatrix} \langle I_z \rangle \\ \langle S_z \rangle \\ \langle 2I_z S_z \rangle \end{pmatrix} = - \begin{pmatrix} \rho_I & \sigma_{IS} & \delta_{I,IS} \\ \sigma_{IS} & \rho_S & \delta_{S,IS} \\ \delta_{I,IS} & \delta_{S,IS} & \rho_{IS} \end{pmatrix} \begin{pmatrix} \langle I_z \rangle - I_z^0 \\ \langle S_z \rangle - S_z^0 \\ \langle 2I_z S_z \rangle \end{pmatrix}. \quad (\text{B1.13.7})$$

The difference compared to [equation B1.13.2](#) or [equation B1.13.3](#) is the occurrence of the expectation value of the $2I_z S_z$ operator (the two-spin order), characterized by its own decay rate ρ_{IS} and coupled to the one-spin longitudinal operators by the terms $\delta_{I,IS}$ and $\delta_{S,IS}$. We shall come back to the physical origin of these terms below.

The matrix on the rhs of equation B1.13.7 is called the relaxation matrix. To be exact, it represents one block of a larger, block-diagonal relaxation matrix, defined in the Liouville space for the two-spin system. The remaining part of the large matrix (sometimes also called the relaxation supermatrix) describes the relaxation of coherences, which can be seen as generalizations of the transverse components of the magnetization vector. In systems without degeneracies, each of the coherences decays exponentially, with its own T_2 . The Redfield theory can be used to obtain expressions for the relaxation matrix elements for arbitrary spin systems and for any type of relaxation mechanism. In analogy with the Solomon theory, also these more general relaxation

rates are expressed in terms of various spectral density functions.

B1.13.2.3 MOLECULAR MOTIONS AND SPIN RELAXATION

Nuclear spin relaxation is caused by fluctuating interactions involving nuclear spins. We write the corresponding Hamiltonians (which act as perturbations to the static or time-averaged Hamiltonian, determining the energy level structure) in terms of a scalar contraction of spherical tensors:

$$H_1(t) = \sum_{q=-j}^j (-1)^q F^{(q)}(t) A^{(-q)}. \quad (\text{B1.13.8})$$

-5-

j is the rank of the tensor describing the relevant interactions, which can be 0, 1 or 2. $A^{(-q)}$ are spin operators and $F^{(q)}(t)$ represent classical functions related to the lattice, i.e. to the classically described environment of the spins. The functions $F^{(q)}(t)$ are, because of random molecular motions, stochastic functions of time. A fluctuating (stochastic) interaction can cause transitions between the energy levels of a spin system (and thus transfer the energy between the spin systems and its environment) if the power spectrum of the fluctuations contains Fourier components at frequencies corresponding to the relevant energy differences. In this sense, the transitions contributing to the spin relaxation and originating from randomly fluctuating interactions, are not really fundamentally different from the transitions caused by coherent interactions with electromagnetic radiation. According to theory of stochastic processes (the Wiener–Khinchin theorem) [5], the power available at a certain frequency, or the spectral density function $J(\omega)$ at that frequency, is obtained as a Fourier transform of a time correlation function (tcf) characterizing the stochastic process. A tcf $G(\tau)$ for a stochastic function of time $F^{(q)}(t)$ is defined:

$$G(\tau) = \langle F^{(-q)}(t) F^{(q)}(t + \tau) \rangle \quad (\text{B1.13.9})$$

with the corresponding spectral density:

$$J(\omega) = \int_{-\infty}^{+\infty} G(\tau) \exp(i\omega\tau) d\tau \quad (\text{B1.13.10})$$

where the symbol $\langle \rangle$ denotes ensemble average. The function in the lhs of equation B1.13.9 is called the auto-correlation function. The prefix ‘auto’ refers to the fact that we deal with an ensemble average of the product of a stochastic function taken at one point in time and the same function at another point in time, the difference between the two time points being τ . In certain situations in relaxation theory, we also need cross-correlation functions, where we average the product of two different stochastic functions, corresponding to different Hamiltonians of [equation B1.13.8](#).

We now come back to the important example of two spin 1/2 nuclei with the dipole–dipole interaction discussed above. In simple physical terms, we can say that one of the spins senses a fluctuating local magnetic field originating from the other one. In terms of the Hamiltonian of [equation B1.13.8](#), the stochastic function of time $F^{(q)}(t)$ is proportional to $Y_{2q}(\theta, \phi)/r_{\text{IS}}^3$, where Y_{2q} is an $l = 2$ spherical harmonic and r_{IS} is the internuclear distance. If the two nuclei are directly bonded, r_{IS} can be considered constant and the random variation of $F^{(q)}(t)$ originates solely from the reorientation of the molecule-fixed internuclear axis with respect to the laboratory frame. The auto-tcf for Y_{2q} can be derived, assuming that the stochastic time dependence can be described by the isotropic rotational diffusion equation. We obtain then:

$$G(\tau) = G(0) \exp(-\tau/\tau_c) \quad (\text{B1.13.11})$$

with the corresponding spectral density:

$$J(\omega) = G(0) \frac{2\tau_c}{1 + \omega^2\tau_c^2}. \quad (\text{B1.13.12})$$

-6-

We call τ_c the correlation time: it is equal to $1/6 D_R$, where D_R is the rotational diffusion coefficient. The correlation time increases with increasing molecular size and with increasing solvent viscosity. [equation B1.13.11](#) and [equation B1.13.12](#) describe the rotational Brownian motion of a rigid sphere in a continuous and isotropic medium. With the Lorentzian spectral densities of [equation B1.13.12](#), it is simple to calculate the relevant transition probabilities. In this way, we can use e.g. [equation B1.13.5](#) to obtain T^{1-1} for a carbon-13 bonded to a proton as well as the corresponding T^{2-} , as a function of the correlation time for a given magnetic field, cf. Figure B1.13.2. The two rates are equal in the region $\omega\tau_c \ll 1$, referred to as the ‘extreme narrowing’. In the extreme narrowing regime, the relaxation rates are independent of the magnetic field and are simply given by a product of the square of an interaction strength constant (the dipole coupling constant, DCC) and the correlation time.

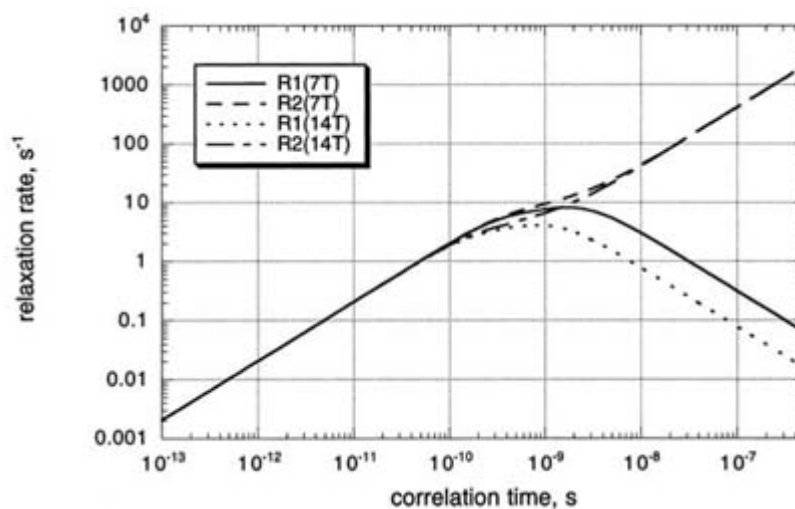


Figure B1.13.2. Spin–lattice and spin–spin relaxation rates (R_1 and R_2 , respectively) for a carbon-13 spin directly bonded to a proton as a function of correlation time at the magnetic fields of 7 and 14 T.

In order to obtain a more realistic description of reorientational motion of internuclear axes in real molecules in solution, many improvements of the tcf of [equation B1.13.11](#) have been proposed [6]. Some of these models are characterized in [table B1.13.1](#). The entry ‘number of terms’ refers to the number of exponential functions in the relevant tcf or, correspondingly, the number of Lorentzian terms in the spectral density function.

-7-

Table B1.13.1 Selected dynamic models used to calculate spectral densities.

Dynamic model	Parameters influencing the spectral densities	Number of terms	Comment	Ref.–
Isotropic rotational diffusion	Rotational diffusion coefficient, $D_R = 1/6 \tau_c$	1	Useful as a first approximation	Further reading
Rotational diffusion, symmetric top	Two rotational diffusion coefficients, D_{\parallel} and D_{\perp} , the angle θ between the symmetry axis and the internuclear axis	3	Rigid molecule, requires the knowledge of geometry	[7]
Rotational diffusion, asymmetric top	Three rotational diffusion coefficients, two polar angles θ and τ	5	Rigid molecule, rather complicated	[8]
Isotropic rotational diffusion with one internal degree of freedom	Rotational diffusion coefficient, D_3 , internal motion rate parameter, angle between the internal rotation axis and the internuclear axis	3	Useful for e.g. methyl groups	[9]
'Model-free'	Global and local correlation times, generalized order parameter, S	2	Widely used for non-rigid molecules	[10]
Anisotropic model-free	D_{\parallel} and D_{\perp} , θ , local correlation time, generalized order parameter, S	4	Allows the interpretation of data for non-spherical, non-rigid molecules	[11]

B1.13.2.4 RELAXATION MECHANISMS

The DD interaction discussed above is just one—admittedly a very important one—among many possible sources of nuclear spin relaxation, collected in [table B1.13.2](#) (not meant to be fully comprehensive). The interactions listed there are often used synonymously with 'relaxation mechanisms' and more detailed descriptions of various mechanisms can be found in 'Further reading'. We can note in [table B1.13.2](#) that one particular type of motion—reorientation of an axis—can cause random variations of many $j = 2$ interactions: DD for different intramolecular axes, CSA, quadrupolar interaction. In fact, all the interactions with the same tensor rank can give rise to the interference or cross-terms with each other. In an important paper by Szymanski *et al* [20], the authors point out that the concept of a 'relaxation mechanism' is more appropriate to use referring to a pair of interactions, rather than to a single one. The role of interference terms has been reviewed by Werbelow [21]. They often contribute to relaxation matrices through coherence or polarization transfer and through higher forms of order or coherence. The relevant spectral densities are of cross-correlation (rather than auto-correlation) type. For example, the off-diagonal δ -terms in the relaxation matrix of [equation B1.13.7](#), connecting the one- and two-spin longitudinal order, are caused by cross-correlation between the DD and CSA interactions, which explains why they are called 'cross-correlated relaxation rates'.

Table B1.13.2 Interactions giving rise to nuclear spin relaxation.

Interaction	Tensor rank	Process causing the random changes	Comment	Ref.
-------------	-------------	------------------------------------	---------	------

Intramolecular dipole-dipole (DD)	2	Reorientation of the inter-nuclear axis	Very common for $I = 1/2$	Further reading
Intermolecular DD	2	Distance variation by translational diffusion	Less common	[12]
Chemical shift anisotropy (CSA)	2	Reorientation of the CSA principal axis	Increases with the square of the magnetic field	[13]
Intramolecular quadrupolar	2	Reorientation of the electric field gradient principal axis	Dominant for $I \geq 1$ (covalently bonded)	[14]
Intermolecular quadrupolar	2	Fluctuation of the electric field gradient, moving multipoles	Common for $I \geq 1$ in free ions in solution	[15]
Antisymmetric CSA	1	Reorientation of a pseudo-vector	Very uncommon	[13]
Spin-rotation	1	Reorientation and time dependence of angular momentum	Small molecules only	[16]
Scalar coupling	0	Relaxation of the coupled spin or exchange	Can be important for T_2	Further reading
Hyperfine interaction (dipolar and scalar)	2,0	Electron relaxation, may be complicated	Paramagnetic systems and impurities	[17–19]

Before leaving the subsection on relaxation mechanisms, I wish to mention the connections between relaxation and chemical exchange (exchange of magnetic environments of spins by a chemical process). The chemical exchange and relaxation determine together the NMR lineshapes, the exchange affects the measured relaxation rates and it can also act as a source of random variation of spin interactions. The relaxation effects of chemical exchange have been reviewed by Woessner [22].

B1.13.3 EXPERIMENTAL METHODS

Relaxation experiments were among the earliest applications of time-domain high-resolution NMR spectroscopy, invented more than 30 years ago by Ernst and Anderson [23]. The progress of the experimental methodology has been enormous and only some basic ideas of the experiment design will be presented here. This section is divided into three subsections. The first one deals with Bloch equation-type experiments, measuring T_1 and T_2 when such quantities can be defined, i.e. when the relaxation is monoexponential. As a slightly oversimplified rule of thumb, we can say that this happens in the case of isolated spins. The two subsections to follow cover multiple-spin effects.

B1.13.3.1 SPIN-LATTICE AND SPIN-SPIN RELAXATION RATES

Measurements of spin-lattice relaxation time, T_1 , are the simplest relaxation experiments. A straight-forward method to measure T_1 is the inversion-recovery experiment, the principle of which is illustrated in [figure B1.13.3](#). The equilibrium magnetization M_0 or $M_z(\infty)$ (cf. [figure B1.13.3\(A\)](#)), directed along B_0 (the z -axis), is first inverted by a 180° pulse (a π -pulse), a RF pulse with the duration τ_{180° and the RF magnetic field B_1 (in the the direction perpendicular to B_0) chosen so that the magnetization is nutated by 180° around the B_1 direction. The magnetization immediately after the 180° pulse is directed along the $-z$ direction (cf. [figure B1.13.3\(B\)](#)) and starts to relax following [equation B1.13.1](#). After a variable delay τ , when the $M_z(\tau)$ has reached the stage depicted in [figure B1.13.3\(C\)](#), a 90° pulse is applied. This pulse nutates the magnetization along the z -axis to the x,y -plane (cf. [figure B1.13.3\(D\)](#)), where it can be detected in the form of a FID. If

required, the experiment can be repeated a number of times to improve the signal-to-noise ratio, waiting for about $5T^1$ (recycle delay) between scans to allow for return to equilibrium. The subsequent Fourier transformation (FT) of the FID gives a spectrum. The experiment is repeated for different values of the delay τ and the measured line intensities are fitted to an exponential expression $S(\tau) = A + B \exp(-\tau / T^1)$. The inversion-recovery experiments are often performed for multiline spectra of low-natural abundance nuclei, such as ^{13}C or ^{15}N , under the conditions of broadband saturation (decoupling) of the abundant proton spins. The proton (S -spin) decoupling renders the relaxation of the I spin of the dipolarly coupled IS -spin system monoexponential; we may say that the decoupling results in 'pseudo-isolated' I spins. An example of a ^{13}C inversion-recovery experiment for a trisaccharide, melezitose, is shown in [figure B1.13.4](#).

-10-

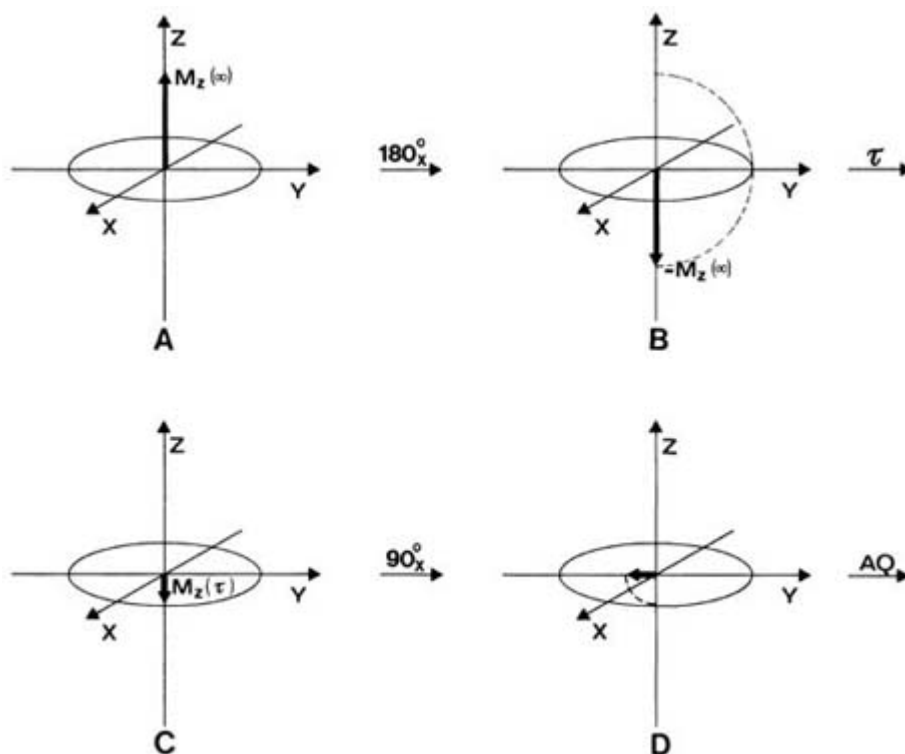


Figure B1.13.3. The inversion-recovery experiment. (Reproduced by permission of VCH from Banci L, Bertini I and Luchinat C 1991 *Nuclear and Electron Relaxation* (Weinheim: VCH).)

-11-

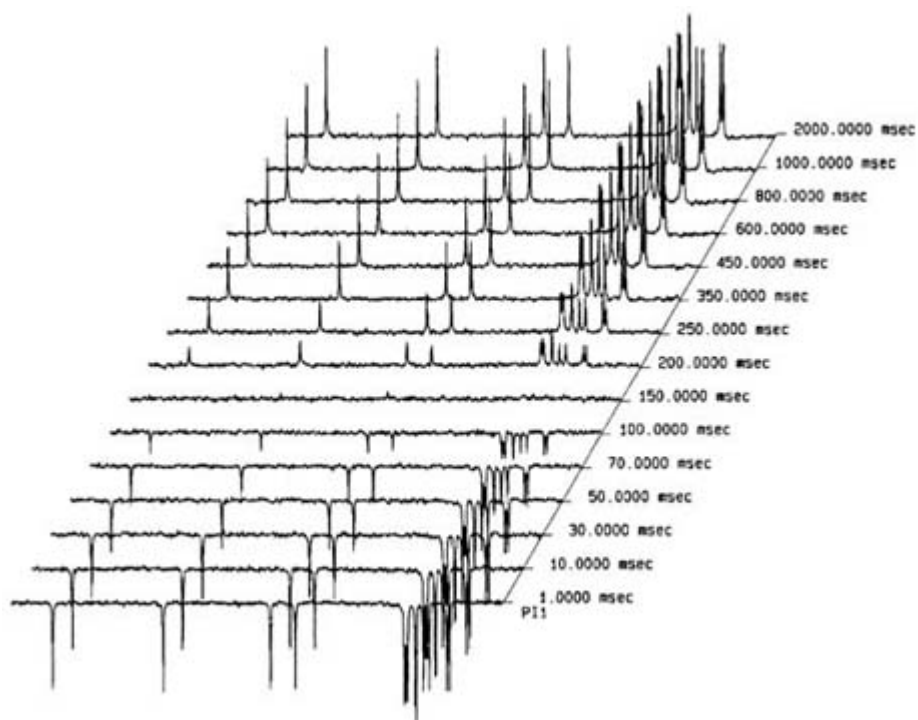


Figure B1.13.4. The inversion-recovery determination of the carbon-13 spin–lattice relaxation rates in melezitose. (Reproduced by permission of Elsevier from Kowalewski J and Mäler L 1997 *Methods for Structure Elucidation by High-Resolution NMR* ed Gy Batta, K E Kövér and Cs Szántay (Amsterdam: Elsevier) pp 325–47.)

NMR spectroscopy is always struggling for increased sensitivity and resolution, as well as more efficient use of the instrument time. To this end, numerous improvements of the simple inversion-recovery method have been proposed over the years. An early and important modification is the so-called fast inversion recovery (FIR) [25], where the recycle delay is made shorter than $5T_1$ and the experiment is carried out under the steady-state rather than equilibrium conditions. A still more time-saving variety, the super-fast inversion recovery (SUFIR) has also been proposed [26].

Several other improvements of the inversion-recovery scheme employ advanced tools of modern NMR spectroscopy: polarization transfer and two-dimensional spectroscopy (see further reading). The basic design of selected pulse sequences is compared with the simple inversion-recovery scheme in figure B1.13.5 taken from Kowalewski and Mäler [24], where references to original papers can be found. The figure B1.13.5(a), where thick rectangular boxes denote the 180° I -spin pulses and thin boxes the corresponding 90° pulses, is a representation of the inversion-recovery sequence with the continuous saturation of the protons. In figure B1.13.5(b), the inverting I -spin pulse is replaced by a series of pulses, separated by constant delays and applied at both the proton and the I -spin resonance frequencies, which creates a more strongly polarized initial I -spin state (the polarization transfer technique). In figure B1.13.5(c), a two-dimensional (2D) NMR technique is employed. This type of approach is particularly useful when the sample contains many heteronuclear IS spin pairs, with different I s and different S s characterized by slightly different resonance frequencies (chemical shifts), resulting in crowded spectra. In a generic 2D experiment, the NMR signal is sampled as a function of two time variables: t_2 is the running time during which the FID is acquired (different

points in the FID have different t_2). In addition, the pulse sequence contains an evolution time t_1 , which is systematically varied in the course of the experiment. The double Fourier transformation of the data matrix $S(t_1, t_2)$ yields a two-dimensional spectrum. In the example of figure B1.13.5(c), the polarization is transferred first from protons to the I spins, in the same way as in figure B1.13.5(b). This is followed by the evolution

time, during which the information on the various I -spin resonance frequencies is encoded. The next period is the analogue of the delay τ of the simple inversion-recovery experiment. The final part of the sequence contains an inverse polarization transfer, from I spins to protons, followed by the proton detection. The resulting 2D spectrum, for a given delay τ , has the proton chemical shifts on one axis and the shifts of the J -coupled I -spin on the other one. We can thus call the experiment the proton– I -spin correlation experiment. This greatly improves the spectral resolution. Spectra with several different τ delays are acquired and the I spin T_1 is determined by fitting the intensity decay for a given peak in the 2D spectrum.

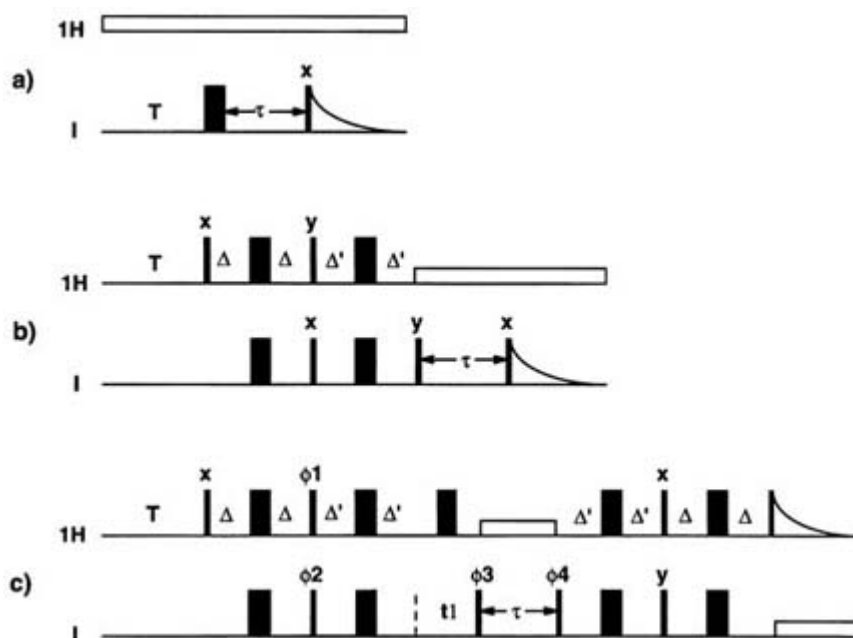


Figure B1.13.5. Some basic pulse sequences for T_1 measurements for carbon-13 and nitrogen-15. (Reproduced by permission of Elsevier from Kowalewski J and Mäler L 1997 *Methods for Structure Elucidation by High-Resolution NMR* ed Gy Batta, K E Kövér and Cs Szántay (Amsterdam: Elsevier) pp 325–47.)

The discussion above is concerned with T_1 experiments under high resolution conditions at high magnetic field. In studies of complex liquids (polymer solutions and melts, liquid crystals), one is often interested to obtain information on rather slow (micro- or nanosecond time scale) motions measuring T_1 at low magnetic fields using the field-cycling technique [27]. The same technique is also invaluable in studies of paramagnetic solutions. Briefly, the non-equilibrium state is created by keeping the sample at a certain, moderately high polarizing field B_0 and then rapidly switching B_0 to another value, at which we wish to measure T_1 . After the variable delay τ , the field is switched again and the signal is detected.

The spin–spin relaxation time, T_2 , defined in the Bloch equations, is simply related to the width $\Delta\nu_{1/2}$ of the Lorentzian line at the half-height: $\Delta\nu_{1/2} = 1/\pi T^2$. Thus, it is in principle possible to determine T_2 by measuring the linewidth. This simple approach is certainly useful for rapidly relaxing quadrupolar nuclei and has also been demonstrated to work for $I=1/2$ nuclei, provided the magnetic field homogeneity is ascertained. The more usual practice, however, is to suppress the inhomogeneous broadening caused by the spread of the magnetic field values (and thus resonance frequencies) over the sample volume, by the spin-echo technique. Such experiments are in general more difficult to perform than the spin–lattice relaxation time measurements. The most common echo sequence, the $90^\circ\text{--}\tau\text{--}180^\circ\text{--}\tau\text{--}\text{echo}$, was originally proposed by Carr and Purcell [28] and modified by Meiboom and Gill [29]. After the initials of the four authors, the modified sequence is widely known as the CPMG method. The details of the behaviour of spins under the spin-echo sequence can be found

in modern NMR monographs (see further reading) and will not be repeated here. We note, however, that complications can arise in the presence of scalar spin–spin couplings.

An alternative procedure for determining the transverse relaxation time is the so-called $T_{1\rho}$ experiment. The basic idea of this experiment is as follows. The initial 90° I -spin pulse is applied with B_1 in the x -direction, which turns the magnetization from the z -direction to the y -direction. Immediately after the initial pulse, the B_1 RF field is switched to the y -direction so that \mathbf{M} and B_1 become collinear. The notation $T_{1\rho}$ (T_1 in the rotating frame) alludes to the fact that the decay of \mathbf{M} along B_1 is similar to the relaxation of longitudinal magnetization along B_0 . The $T_{1\rho}$ in liquids is practically identical to T_2 . The measurements of T_2 or $T_{1\rho}$ can, in analogy with T_1 studies, also utilize the modern tools increasing the sensitivity and resolution, such as polarization transfer and 2D techniques.

B1.13.3.2 CROSS-RELAXATION AND NUCLEAR OVERHAUSER ENHANCEMENT

Besides measuring T_1 and T_2 for nuclei such as ^{13}C or ^{15}N , relaxation studies for these nuclei also include measurements of the NOE factor, cf. [equation B1.13.6](#). Knowing the $T_{1I}^{-1}(\rho_I)$ and the steady-state NOE (measured by comparing the signal intensities in the presence and in the absence of the saturating field), we can derive the cross-relaxation rate, σ_{IS} , which provides an additional combination of spectral densities, useful for e.g. molecular dynamics studies.

The most important cross-relaxation rate measurements are, however, performed in homonuclear networks of chemically shifted and dipolarly coupled proton spins. The subject has been discussed in two books [[30](#), [31](#)]. There is large variety of experimental procedures [[32](#)], of which I shall only mention a few. A simple method to measure the homonuclear NOE is the NOE-difference experiment, in which one measures spectral intensities using low power irradiation at selected narrow regions of the spectrum, before applying the observe pulse. This corresponds to different individual protons acting as the saturated S spins. The difference spectrum is obtained by subtracting the spectrum obtained under identical conditions, but with the irradiation frequency applied in a region without any peaks to saturate. The NOE-difference experiments are most often applied in a semi-quantitative way in studies of small organic molecules.

For large molecules, such as proteins, the main method in use is a 2D technique, called NOESY (nuclear Overhauser effect spectroscopy). The basic experiment [[33](#), [34](#)] consists of three 90° pulses. The first pulse converts the longitudinal magnetizations for all protons, present at equilibrium, into transverse magnetizations which evolve during the subsequent evolution time t_1 . In this way, the transverse magnetization components for different protons become labelled by their resonance frequencies. The second 90° pulse rotates the magnetizations to the $-z$ -direction.

-14-

The interval between the second and third pulse is called the mixing time, during which the spins evolve according to the multiple-spin version of [equation B1.13.2](#) and [equation B1.13.3](#) and the NOE builds up. The final pulse converts the longitudinal magnetizations, present at the end of the mixing time, into detectable transverse components. The detection of the FID is followed by a recycle delay, during which the equilibrium is recovered and by the next experiment, e.g. with another t_1 . After acquiring the 2D data matrix $S(t^1, t^2)$ (for a given mixing time) and the double Fourier transformation, one obtains a 2D spectrum shown schematically in [figure B1.13.6](#). The individual cross-relaxation rates for pairs of spins can be obtained by following the build-up of the cross-peak intensities as a function of the mixing time for short mixing times (the so-called initial rate approximation). For longer mixing times and large molecules, the cross-peaks show up in a large number of positions, because of multiple transfers called spin-diffusion. The analysis then becomes more complicated, but can be handled based on a generalization of the Solomon equation to many spins (the complete relaxation matrix treatment) [[35](#)].

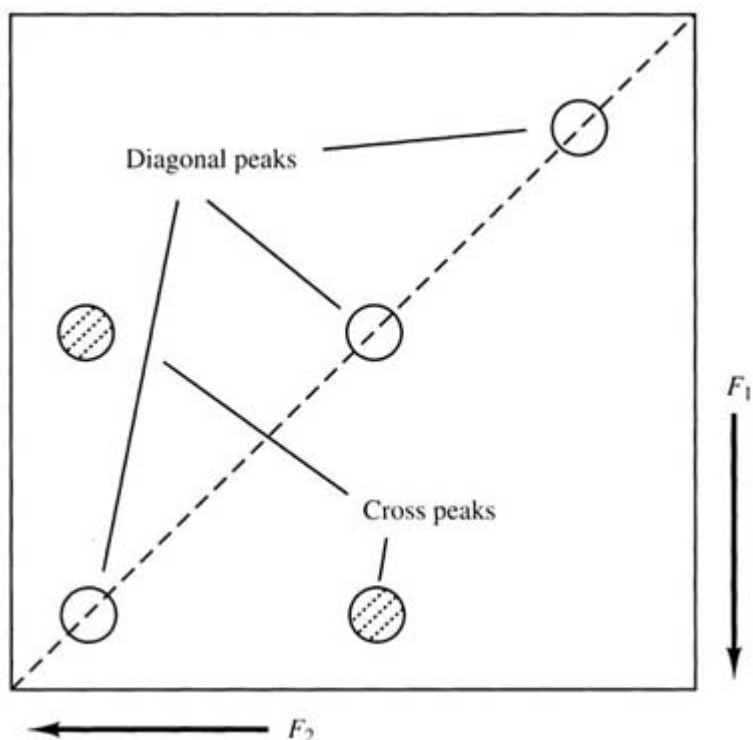


Figure B1.13.6. The basic elements of a NOESY spectrum. (Reproduced by permission of Wiley from Williamson M P 1996 *Encyclopedia of Nuclear Magnetic Resonance* ed D M Grant and R K Harris (Chichester: Wiley) pp 3262–71).

As seen in [equation B1.13.2](#) and [equation B1.13.3](#), the cross-relaxation rate σ_{IS} is given by $W^2 - W^0$, the difference between two transition probabilities. Assuming the simple isotropic rotational diffusion model, each of the transition probabilities is proportional to a Lorentzian spectral density (cf. [equation B1.13.12](#)), taken at the frequency of the corresponding transition. For the homonuclear case, W_2 corresponds to a transition at high frequency ($\omega_1 + \omega_S \approx 2\omega_1$), while W_0 is proportional to a Lorentzian at ($\omega_1 - \omega_S \approx 0$). When the product $\omega_1\tau_c$ is small, W_2 is larger than W_0 and the cross-relaxation rate is positive. When the product $\omega_1\tau_c$ is large, the Lorentzian function evaluated at $2\omega_1$ is much smaller than at zero-frequency and σ_{IS} changes sign. The corresponding NOESY peak intensities in a two-spin

system are shown as a function of the mixing time in figure B1.13.7. Clearly, the intensities of the cross-peaks for small molecules ($\omega^2\tau_c^2 \ll 1$) have one sign, while the opposite sign pertains for large molecules ($\omega^2\tau_c^2 \gg 1$). At a certain critical correlation time, we obtain no NOESY cross-peaks. In such a situation and as complement to the NOESY experiments, one can perform an experiment called ROESY (rotating frame Overhauser effect spectroscopy) [36]. The relation between NOESY and ROESY is similar to that between T_1 and $T_{1\rho}$.

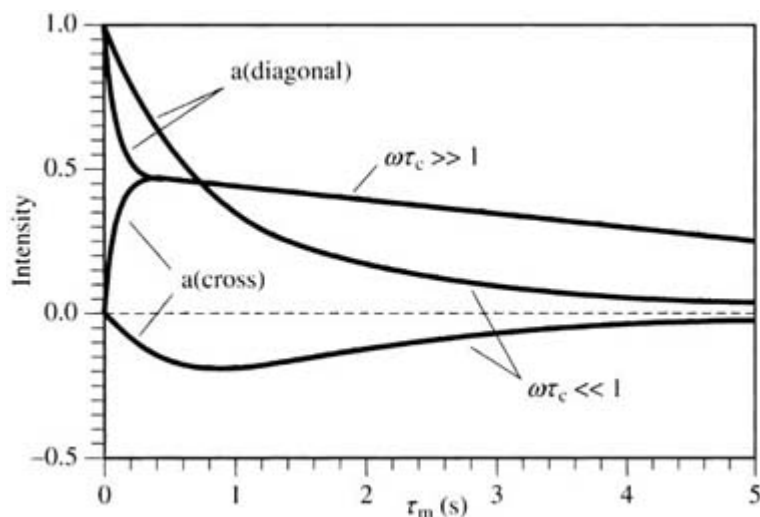


Figure B1.13.7. Simulated NOESY peak intensities in a homonuclear two-spin system as a function of the mixing time for two different motional regimes. (Reproduced by permission of Wiley from Neuhaus D 1996 *Encyclopedia of Nuclear Magnetic Resonance* ed D M Grant and R K Harris (Chichester: Wiley) pp 3290–301.)

B1.13.3.3 CROSS-CORRELATED RELAXATION

Studies of cross-correlated relaxation have received increasing attention during the last decades. The general strategies for creating and detecting different types of spin-ordered state and for measuring the transfer rates have been discussed by Canet [37]. I shall concentrate here on measurements of the DD-CSA interference terms in two-spin systems, the $\delta_{I,IS}$ and $\delta_{S,IS}$ terms in equation B1.13.7. Let us consider a system where the two spins have their resonances sufficiently far apart that we can construct pulses selective enough to manipulate one of them at a time (this is automatically fulfilled for a heteronuclear case; in the homonuclear this requires specially shaped low power RF pulses). One way to measure the longitudinal cross-correlated relaxation rates is to invert one of the spins by a 180° pulse and to detect the build-up of the two-spin order. The two-spin order $\langle 2I_{zSz} \rangle$ is not detectable directly but, if one of the spins is exposed to a 90° pulse, the two-spin order becomes converted into a detectable signal in the form of an antiphase doublet, cf. Figure B1.13.8 (the corresponding one-spin order subjected to a 90° pulse gives rise to an in-phase doublet). To separate the two types of order in a clean way, one can use an RF pulse trick called the double-quantum filter. There are many ways to optimize the above method as well as other schemes to measure the $\langle I_z \rangle$ to $\langle 2I_{zSz} \rangle$ transfer rates. One such scheme uses a set of three selective NOESY experiments, where three 90° pulses strike the spins in the sequence *III*, *SSS* and *IIS* [38]. Another scheme uses an extended sequence of 180° pulses, followed by a detecting pulse [39].

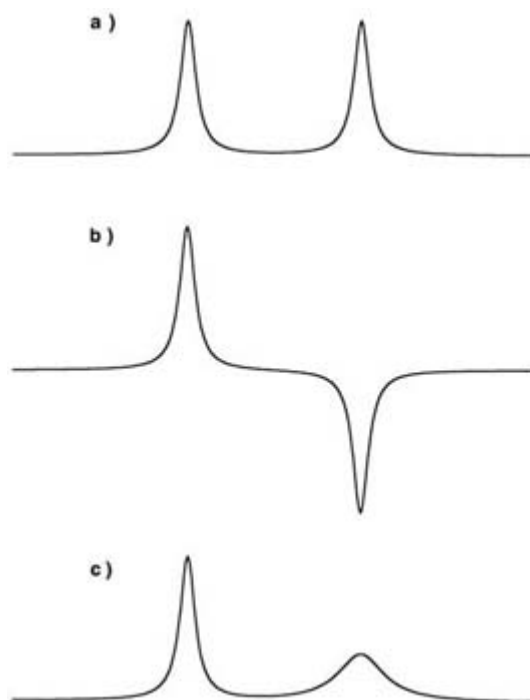


Figure B1.13.8. Schematic illustration of (a) an antiphase doublet, (b) an in-phase doublet and (c) a differentially broadened doublet. The splitting between the two lines is in each case equal to J , the indirect spin–spin coupling constant.

The cross-correlation effects between the DD and CSA interactions also influence the transverse relaxation and lead to the phenomenon known as differential line broadening in a doublet [40], cf. Figure B1.13.8. There is a recent experiment, designed for protein studies, that I wish to mention at the end of this section. It has been proposed by Pervushin *et al* [41], is called TROSY (transverse relaxation optimized spectroscopy) and employs the differential line-broadening in a sophisticated way. One works with the ^{15}N –proton J -coupled spin system. When the system is subjected to a 2D nitrogen–proton correlation experiment, the spin–spin coupling gives rise to four lines (a doublet in each dimension). The differential line broadening results in one of the four lines being substantially narrower than the other ones. Pervushin *et al* [41] demonstrated that the broader lines can be suppressed, resulting in greatly improved resolution in the spectra of large proteins.

B1.13.4 APPLICATIONS

In this section, I present a few illustrative examples of applications of NMR relaxation studies within different branches of chemistry. The three subsections cover one ‘story’ each, in order of increasing molecular size and complexity of the questions asked.

B1.13.4.1 SMALL MOLECULES: THE DUAL SPIN PROBE TECHNIQUE

Small molecules in low viscosity solutions have, typically, rotational correlation times of a few tens of picoseconds, which means that the extreme narrowing conditions usually prevail. As a consequence, the interpretation of certain relaxation parameters, such as carbon-13 T_1 and NOE for proton-bearing carbons, is very simple. Basically, the DCC for a directly bonded CH pair can be assumed to be known and the experiments yield a value of the correlation time, τ_c . One interesting application of the measurement of τ_c is to follow its variation with the site in the molecule (motional anisotropy), with temperature (the correlation

time increases often with decreasing temperature, following an Arrhenius-type equation) or with the composition of solution. The latter two types of measurement can provide information on intermolecular interactions.

Another application of the knowledge of τ_c is to employ it for the interpretation of another relaxation measurement in the same system, an approach referred to as the dual spin probe technique. A rather old, but illustrative, example is the case of *tris*(pentane-2,4-dionato)aluminium(III), $\text{Al}(\text{acac})^3$. Dechter *et al* [42] reported measurements of carbon-13 spin–lattice relaxation for the methine carbon and of the aluminium-27 linewidth in $\text{Al}(\text{acac})^3$ in toluene solution. ^{27}Al is a quadrupolar nucleus ($I=5/2$) and the linewidth gives directly the T_2 , which depends on the rotational correlation time (which can be assumed the same as for the methine CH axis) and the strength of the quadrupolar interaction (the quadrupolar coupling constant, QCC). Thus, the combination of the carbon-13 and aluminium-27 measurements yields the QCC. The QCC in this particular case was also determined by NMR in a solid sample. The two measurements agree very well with each other (in fact, there is a small error in the paper [42]: the QCC from the linewidth in solution should be a factor of 2π larger than what is stated in the article). More recently, Champmartin and Rubini [43] studied carbon-13 and oxygen-17 (another $I=5/2$ quadrupolar nucleus) relaxation in pentane-2,4-dione (Hacac) and $\text{Al}(\text{acac})^3$ in solution. The carbon measurements were performed as a function of the magnetic field. The methine carbon relaxation showed, for both compounds, no field dependence, while the carbonyl carbon T_1 increased linearly with $\frac{2}{\omega_0}$. This indicates the CSA mechanism and allows an estimate of its interaction strength, the anisotropy of the shielding tensor. Also this quantity could be compared with solid state measurements on $\text{Al}(\text{acac})^3$ and, again, the agreement was good. From the oxygen-17 linewidth, the authors obtained also the oxygen-17 QCC. The chemically interesting piece of information is the observation that the QCC changes only slightly between the free acid and the trivalent metal complex.

B1.13.4.2 OLIGOSACCHARIDES: HOW FLEXIBLE ARE THEY?

Oligosaccharides are a class of small and medium-sized organic molecules, subject to intense NMR work. I present here the story of two disaccharides, sucrose and $\alpha\text{-D-Manp-(1}\rightarrow\text{3)-}\beta\text{-D-Glcp-OMe}$ and a trisaccharide melezitose. McCain and Markley [44] published a carbon-13 T_1 and NOE investigation of sucrose in aqueous solution as a function of temperature and magnetic field. At low temperatures, a certain field dependence of the parameters could be observed, indicating that the extreme narrowing conditions might not be fulfilled under these circumstances. Taking the system out of extreme narrowing (which corresponds to correlation times of few hundred picoseconds) renders the relaxation rates field dependent and allows the investigators to ask more profound questions concerning the interaction strength and dynamics. Kovacs *et al* [45] followed up the McCain–Markley study by performing similar experiments on sucrose in a 7:3 molar $\text{D}^2\text{O/DMSO-d}^6$ solvent mixture. The solvent mixture has about four times higher viscosity than water and is a cryo-solvent. Thus, it was possible to obtain the motion of the solute molecule far from the extreme narrowing region and to make a quantitative determination of the effective DCC, in a way which is related to the Lipari–Szabo method [10]. Mäler *et al* [46] applied a similar experimental approach (but extended also to include the T_2 measurements) and the Lipari–Szabo analysis to study the molecular dynamics of melezitose in the same

mixed solvent. Somewhat different dynamic behaviour of different sugar residues and the exocyclic hydroxymethyl groups was reported.

The question of the flexibility (or rigidity) of the glycosidic linkage connecting the different sugar units in oligosaccharides is a hot issue in carbohydrate chemistry. The findings of Mäler *et al* [46] could be interpreted as indicating a certain flexibility of the glycosidic linkages in melezitose. The question was posed more directly by Poppe and van Halbeek [47], who measured the intra- and interresidue proton–proton NOEs and rotating frame NOEs in sucrose in aqueous solution. They interpreted the results as proving the non-rigidity of

the glycosidic linkage in sucrose. Mäler *et al* [48] investigated the disaccharide α -D-Manp-(1 \rightarrow 3)- β -D-Glcp-OMe by a combination of a carbon-13 spin-lattice relaxation study with measurements of the intra- and interresidue proton-proton cross-relaxation rates in a water-DMSO mixture. They interpreted their data in terms of the dynamics of the CH axes, the intra-ring HH axis and the trans-glycosidic HH axis being described by a single set of Lipari-Szabo parameters. This indicated that the inter-residue HH axis did not sense any additional mobility as compared to the other axes. We shall probably encounter a continuation of this story in the future.

B1.13.4.3 HUMAN UBIQUITIN: A CASE HISTORY FOR A PROTEIN

Human ubiquitin is a small (76 amino acid) and well characterized protein. I choose to illustrate the possibilities offered by NMR relaxation studies of proteins [49] through this example. The 2D NMR studies for the proton resonance assignment and a partial structure determination through NOESY measurements were reported independently by Di Stefano and Wand [50] and by Weber *et al* [51]. Schneider *et al* [52] studied a uniformly nitrogen-15 labelled species of human ubiquitin and reported nitrogen-15 T_1 (at two magnetic fields) and NOE values for a large majority of amide nitrogen sites in the molecule. The data were interpreted in terms of the Lipari-Szabo model, determining the generalized order parameter for the amide NH axes. It was thus possible to identify the more flexible and more rigid regions in the protein backbone. Tjandra *et al* [53] extended this work, both in terms of experiments (adding T_2 measurements) and in terms of interpretation (allowing for the anisotropy of the global reorientation, by means of the anisotropic Lipari-Szabo model [10]). This provided a more quantitative interpretation of the molecular dynamics. During the last two years, Bax and coworkers have, in addition, determined the CSA for the nitrogen-15 and proton for most of the amide sites, as well as for the α -carbons, through measurements of the cross-correlated relaxation rates in the nitrogen-15-amide proton or the C_α - H_α spin pairs [54, 55 and 56]. The CSA values could, in turn, be correlated with the secondary structure, hydrogen bond length etc. It is not likely that the ubiquitin story is finished either.

ACKNOWLEDGMENTS

I am indebted to Dr Dan Bergman and Mr Tomas Nilsson for valuable comments on the manuscript.

REFERENCES

- [1] Bloch F 1946 Nuclear induction *Phys. Rev.* **70** 460–74
 - [2] Bloembergen N, Purcell E M and Pound R V 1948 Relaxation effects in nuclear magnetic resonance absorption *Phys. Rev.* **73** 679–712
-

- [3] Solomon I 1955 Relaxation processes in a system of two spins *Phys. Rev.* **99** 559–65
- [4] Redfield A G 1996 Relaxation theory: density matrix formulation *Encyclopedia of Nuclear Magnetic Resonance* ed D M Grant and R K Harris (Chichester: Wiley) pp 4085–92
- [5] Van Kampen N G 1981 *Stochastic Processes in Physics and Chemistry* (Amsterdam: North-Holland)

- [6] Woessner D E 1996 Brownian motion and correlation times *Encyclopedia of Nuclear Magnetic Resonance* ed D M Grant and R K Harris (Chichester: Wiley) pp 1068–84
- [7] Woessner D E 1962 Nuclear spin relaxation in ellipsoids undergoing rotational Brownian motion *J. Chem. Phys.* **37** 647–54
- [8] Huntress W T 1970 The study of anisotropic rotation of molecules in liquids by NMR quadrupolar relaxation *Adv. Magn. Reson.* **4** 1–37
- [9] Woessner D E 1962 Spin relaxation processes in a two-proton system undergoing anisotropic reorientation *J. Chem. Phys.* **36** 1–4
- [10] Lipari G and Szabo A 1982 Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules 1. Theory and range of validity *J. Am. Chem. Soc.* **104** 4546–59
- [11] Barbato G, Ikura M, Kay L E, Pastor R W and Bax A 1992 Backbone dynamics of calmodulin studied by ¹⁵N relaxation using inverse detected two-dimensional NMR spectroscopy: the central helix is flexible *Biochemistry* **31** 5269–78
- [12] Hwang L-P and Freed J H 1975 Dynamic effects of pair correlation functions on spin relaxation by translational diffusion in liquids *J. Chem. Phys.* **63** 4017–25
- [13] Anet F A L and O'Leary D J 1992 The shielding tensor. Part II. Understanding its strange effects on relaxation *Concepts Magn. Reson.* **4** 35–52
- [14] Werbelow L G 1996 Relaxation theory for quadrupolar nuclei *Encyclopedia of Nuclear Magnetic Resonance* ed D M Grant and R K Harris (Chichester: Wiley) pp 4092–101
- [15] Roberts J E and Schnitker J 1993 Ionic quadrupolar relaxation in aqueous solution—dynamics of the hydration sphere *J. Phys. Chem.* **97** 5410–17
- [16] McClung R E D 1996 Spin-rotation relaxation theory *Encyclopedia of Nuclear Magnetic Resonance* ed D M Grant and R K Harris (Chichester: Wiley) pp 4530–5
- [17] Banci L, Bertini I and Luchinat C 1991 *Nuclear and Electron Relaxation* (Weinheim: VCH)
- [18] Bertini I, Luchinat C and Aime S 1996 *NMR of Paramagnetic Substances* (Amsterdam: Elsevier)
- [19] Kowalewski J 1996 Paramagnetic relaxation in solution *Encyclopedia of Nuclear Magnetic Resonance* ed D M Grant and R K Harris (Chichester: Wiley) pp 3456–62
- [20] Szymanski S, Gryff-Keller A M and Binsch G A 1986 Liouville space formulation of Wangness–Bloch–Redfield theory of nuclear spin relaxation suitable for machine computation. I. Fundamental aspects *J. Magn. Reson.* **68** 399–432

- [21] Werbelow L G 1996 Relaxation processes: cross correlation and interference terms *Encyclopedia of Nuclear Magnetic Resonance* ed D M Grant and R K Harris (Chichester: Wiley) pp 4072–8
- [22] Woessner D E 1996 Relaxation effects of chemical exchange *Encyclopedia of Nuclear Magnetic Resonance* ed D M Grant and R K Harris (Chichester: Wiley) pp 4018–28
- [23] Ernst R R and Anderson W A 1986 Application of Fourier transform spectroscopy to magnetic resonance *Rev. Sci. Instrum.* **37** 93–102
- [24] Kowalewski J and Mäler L 1997 Measurements of relaxation rates for low natural abundance $I=1/2$ nuclei *Methods for Structure Elucidation by High-Resolution NMR* ed Gy Batta, K E Kövér and Cs

Szántay (Amsterdam: Elsevier) pp 325–47

- [25] Canet D, Levy G C and Peat I R 1975 Time saving in ^{13}C spin-lattice relaxation measurements by inversion-recovery *J. Magn. Reson.* **18** 199–204
 - [26] Canet D, Mutzenhardt P and Robert J B 1997 The super fast inversion recovery (SUFIR) experiment *Methods for Structure Elucidation by High-Resolution NMR* ed Gy Batta, K E Kövér and Cs Szántay (Amsterdam: Elsevier) pp 317–23
 - [27] Koenig S H and Brown R D 1990 Field-cycling relaxometry of protein solutions and tissue—implications for MRI *Prog. Nucl. Magn. Reson. Spectrosc.* **22** 487–567
 - [28] Carr H Y and Purcell E M 1954 Effects of diffusion on free precession in nuclear magnetic resonance experiments *Phys. Rev.* **94** 630–8
 - [29] Meiboom S and Gill D 1958 Modified spin-echo method for measuring nuclear relaxation times *Rev. Sci. Instrum.* **29** 688–91
 - [30] Noggle J H and Schirmer R E 1971 *The Nuclear Overhauser Effect* (New York: Academic)
 - [31] Neuhaus D and Williamson M P 1989 *The Nuclear Overhauser Effect in Structural and Conformational Analysis* (New York: VCH)
 - [32] Neuhaus D 1998 Nuclear Overhauser effect *Encyclopedia of Nuclear Magnetic Resonance* ed D M Grant and R K Harris (Chichester: Wiley) pp 3290–301
 - [33] Jeener J, Meier B H, Bachmann P and Ernst R R 1979 Investigation of exchange processes by two-dimensional NMR spectroscopy *J. Chem. Phys.* **71** 4546–53
 - [34] Williamson M P 1996 NOESY *Encyclopedia of Nuclear Magnetic Resonance* ed D M Grant and R K Harris (Chichester: Wiley) pp 3262–71
 - [35] Borgias B A, Gochin M, Kerwood D J and James T L 1990 Relaxation matrix analysis of 2D NMR data *Prog. NMR Spectrosc.* **22** 83–100
 - [36] Bax A and Grzesiek S 1996 ROESY *Encyclopedia of Nuclear Magnetic Resonance* ed D M Grant and R K Harris (Chichester: Wiley) pp 4157–67
 - [37] Canet D 1989 Construction, evolution and detection of magnetization modes designed for treating longitudinal relaxation of weakly coupled spin 1/2 systems with magnetic equivalence *Prog. NMR Spectrosc.* **21** 237–91
-

- [38] Di Bari L, Kowalewski J and Bodenhausen G 1990 Magnetization transfer modes in scalar-coupled spin systems investigated by selective 2-dimensional nuclear magnetic resonance exchange experiments *J. Chem. Phys.* **93** 7698–705
- [39] Levitt M H and Di Bari L 1994 The homogeneous master equation and the manipulation of relaxation networks *Bull. Magn. Reson.* **16** 94–114
- [40] Farrar T C and Stringfellow T C 1996 Relaxation of transverse magnetization for coupled spins *Encyclopedia of Nuclear Magnetic Resonance* ed D M Grant and R K Harris (Chichester: Wiley) pp 4101–7
- [41] Pervushin K, Riek R, Wider G and Wüthrich K 1997 Attenuated T^2 relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very

large biological macromolecules in solution *Proc. Natl Acad. Sci. USA* **94** 12 366–71

- [42] Dechter J J, Henriksson U, Kowalewski J and Nilsson A-C 1982 Metal nucleus quadrupole coupling constants in aluminum, gallium and indium acetylacetonates *J. Magn. Reson.* **48** 503–11
- [43] Champmartin D and Rubini P 1996 Determination of the O-17 quadrupolar coupling constant and of the C-13 chemical shielding tensor anisotropy of the CO groups of pentane-2,4-dione and beta-diketonate complexes in solution. NMR relaxation study *Inorg. Chem.* **35** 179–83
- [44] McCain D C and Markley J L 1986 Rotational spectral density functions for aqueous sucrose: experimental determination using ¹³C NMR *J. Am. Chem. Soc.* **108** 4259–64
- [45] Kovacs H, Bagley S and Kowalewski J 1989 Motional properties of two disaccharides in solutions as studied by carbon-13 relaxation and NOE outside of the extreme narrowing region *J. Magn. Reson.* **85** 530–41
- [46] Mäler L, Lang J, Widmalm G and Kowalewski J 1995 Multiple-field carbon-13 NMR relaxation investigation on melezitose *Magn. Reson. Chem.* **33** 541–8
- [47] Poppe L and van Halbeek H 1992 The rigidity of sucrose—just an illusion? *J. Am. Chem. Soc.* **114** 1092–4
- [48] Mäler L, Widmalm G and Kowalewski J 1996 Dynamical behavior of carbohydrates as studied by carbon-13 and proton nuclear spin relaxation *J. Phys. Chem.* **100** 17 103–10
- [49] Dayie K T, Wagner G and Lefèevre J F 1996 Theory and practice of nuclear spin relaxation in proteins *Annu. Rev. Phys. Chem.* **47** 243–82
- [50] Di Stefano D L and Wand A J 1987 Two-dimensional ¹H NMR study of human ubiquitin: a main chain directed assignment and structure analysis *Biochemistry* **26** 7272–81
- [51] Weber P L, Brown S C and Mueller L 1987 Sequential ¹H NMR assignment and secondary structure identification of human ubiquitin *Biochemistry* **26** 7282–90
- [52] Schneider D M, Dellwo M J and Wand A J 1992 Fast internal main-chain dynamics of human ubiquitin *Biochemistry* **31** 3645–52
- [53] Tjandra N, Feller S E, Pastor R W and Bax A 1995 Rotational diffusion anisotropy of human ubiquitin from N-15 NMR relaxation *J. Am. Chem. Soc.* **117** 12 562–6
- [54] Tjandra N, Szabo A and Bax A 1996 Protein backbone dynamics and N-15 chemical shift anisotropy from quantitative measurement of relaxation interference effects *J. Am. Chem. Soc.* **118** 6986–91

-22-

- [55] Tjandra N and Bax A 1997 Solution NMR measurement of amide proton chemical shift anisotropy in N-15-enriched proteins. Correlation with hydrogen bond length *J. Am. Chem. Soc.* **119** 8076–82
- [56] Tjandra N and Bax A 1997 Large variations in C-13(alpha) chemical shift anisotropy in proteins correlate with secondary structure *J. Am. Chem. Soc.* **119** 9576–7

FURTHER READING

Hennel J W and Klinowski J 1993 *Fundamentals of Nuclear Magnetic Resonance* (Harlow: Longman)

A good introductory textbook, includes a nice and detailed presentation of relaxation theory at the level of Solomon equations.

Goldman M 1988 *Quantum Description of High-Resolution NMR in Liquids* (Oxford: Clarendon)

A good introductory treatment of the density operator formalism and two-dimensional NMR spectroscopy, nice presentation of Redfield relaxation theory.

Canet D 1996 *Nuclear Magnetic Resonance Concepts and Methods* (Chichester: Wiley)

A good NMR textbook, with a nice chapter on relaxation.

Ernst R R, Bodenhausen G and Wokaun A 1987 *Principles of Nuclear Magnetic Resonance in One and Two Dimensions* (Oxford: Clarendon Press)

An extensive presentation of the modern NMR theory, standard text for 2D NMR.

Abraham A 1961 *Principles of Nuclear Magnetism* (Oxford: Clarendon)

An extensive presentation of the fundamentals of NMR. A very good chapter on relaxation, including general theory and mechanisms. A real classic 'still going strong'.

Batta Gy, Kövér K E and Szántay Cs (eds) 1997 *Methods for Structure Elucidation by High-Resolution NMR* (Amsterdam: Elsevier)

A nice collection of review articles dealing with various experimental aspects of modern NMR spectroscopy; several papers on relaxation.

Kowalewski J 1990 Nuclear spin relaxation in diamagnetic fluids Part 1 *Annu. Rep. NMR Spectrosc.* **22** 308–414; 1991 Part 2: *Annu. Rep. NMR Spectrosc.* **23** 289–374

An extensive review of the relaxation literature of the 1980s.

B1.14 NMR imaging (diffusion and flow)

Ute Goerke, Peter J McDonald and Rainer Kimmich

B1.14.1 INTRODUCTION

Nuclear magnetic resonance imaging and magnetic resonance determinations of flow and diffusion are increasingly coming to the fore as powerful means for characterizing dynamic processes in diverse areas of materials science covering physics, chemistry, biology and engineering. In this chapter, the basic concepts of the methods are introduced and a few selected examples presented to show the power of the techniques. Although flow and diffusion through bulk samples can be measured, they are primarily treated here as parameters to be mapped in an imaging experiment. To that end, imaging is dealt with first, followed by flow and diffusion, along with other contrast parameters such as spin-relaxation times and chemical shift.

The great diversity of applications of magnetic resonance imaging (MRI) has resulted in a plethora of techniques which at first sight can seem bemusing. However, at heart they are built on a series of common

building blocks which the reader will progressively come to recognize. The discussion of imaging is focused very much on just three of these—slice selection, phase encoding and frequency encoding, which are brought together in perhaps the most common imaging experiment of all, the spin warp sequence [1, 2 and 3]. This sequence is depicted in figure B1.14.1. Its building blocks are discussed in the following sections. Blocks for contrast enhancement and parameter selectivity can be added to the sequence.

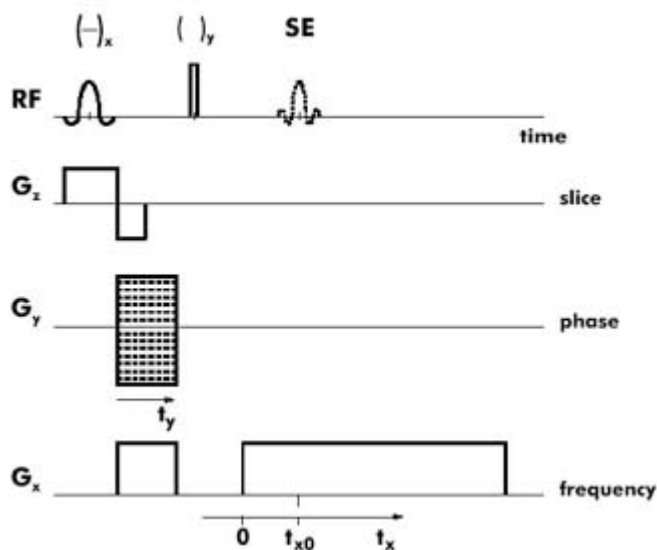


Figure B1.14.1. Spin warp spin-echo imaging pulse sequence. A spin echo is refocused by a non-selective 180° pulse. A slice is selected perpendicular to the z-direction. To frequency-encode the x-coordinate the echo SE is acquired in the presence of the readout gradient. Phase-encoding of the y-dimension is achieved by incrementing the gradient pulse G_y .

B1.14.2 FUNDAMENTALS OF SPATIAL ENCODING

The classical description of magnetic resonance suffices for understanding the most important concepts of magnetic resonance imaging. The description is based upon the Bloch equation, which, in the absence of relaxation, may be written as

$$\frac{d\mathbf{M}(\mathbf{r}, t)}{dt} = \gamma \mathbf{M}(\mathbf{r}, t) \times \mathbf{B}(\mathbf{r}, t).$$

The equation describes the manner in which the nuclear magnetization, \mathbf{M} , at position \mathbf{r} and time t precesses about the magnetic flux density, \mathbf{B} , in which it is found. The constant γ is the magnetogyric ratio of the nuclides under study. The precessional frequency, ω , is given by the Larmor equation,

$$\omega = \gamma B. \tag{B1.14.1}$$

Magnetic resonance imaging, flow and diffusion all rely upon manipulating spatially varying magnetic fields in such a manner as to encode spatial information within the accumulated precession of the magnetization. In

most MRI implementations, this is achieved with three orthogonal, constant, pulsed magnetic field gradients which are produced using purpose built current windings carrying switched currents. The gradient fields are superimposed on the normal static, applied field B_0 . Although the current windings produce gradient fields with components in all directions, for sufficiently high applied flux densities and small enough gradients, it is sufficient to consider only the components in the direction of the applied field, conventionally assigned the z -direction. Accordingly, the gradients are referred to as $G_x = \frac{\partial B_z}{\partial x}$, $G_y = \frac{\partial B_z}{\partial y}$ and $G_z = \frac{\partial B_z}{\partial z}$. The local polarizing field at position \mathbf{r} is then given by

$$\mathbf{B}(\mathbf{r}) = (B_0 + \mathbf{G} \cdot \mathbf{r})\mathbf{u}_z.$$

The Bloch equation is simplified, and the experiment more readily understood, by transformation into a frame of reference rotating at the frequency $\omega_0 = \gamma B_0$ about the z -axis whereupon:

$$\frac{dM'(\mathbf{r}, t)}{dt} = \gamma M'(\mathbf{r}, t) \times (\mathbf{G} \cdot \mathbf{r})\mathbf{u}_z.$$

The transverse magnetization may be described in this frame by a complex variable, m , the real and imaginary parts of which represent the real and imaginary components of observable magnetization respectively:

$$m(\mathbf{r}, t) = M'_x(\mathbf{r}, t) + iM'_y(\mathbf{r}, t).$$

-3-

The components of the Bloch equation are hence reduced to

$$\begin{aligned} \frac{dM'_z(\mathbf{r}, t)}{dt} &= 0 \\ \frac{dm(\mathbf{r}, t)}{dt} &= -i\gamma(\mathbf{G} \cdot \mathbf{r})m(\mathbf{r}, t). \end{aligned}$$

The z -component of the magnetization is constant. The evolution of the transverse magnetization is given by

$$m(\mathbf{r}, t) = m(\mathbf{r}, 0) \exp[-i\Omega(\mathbf{r})t] \tag{B1.14.2}$$

where $\Omega(\mathbf{r}) = \gamma \mathbf{G} \cdot \mathbf{r}$ is the offset frequency relative to the resonance frequency ω_0 . Spin position is encoded directly in the offset frequency Ω . Measurement of the offset frequency forms the basis of the frequency encoding of spatial information, the first building block of MRI. It is discussed further in subsequent sections. In a given time $t = \tau$ the transverse magnetization accumulates a spatially varying phase shift, $\omega(\mathbf{r})\tau$. Measurement of the phase shift forms the basis of another building block, the second encoding technique, phase encoding, which is also further discussed below. However, before proceeding with further discussion of either of these two, we turn our attention to the third key component of an imaging experiment—slice selection.

B1.14.2.1 SLICE SELECTION

A slice is selected in NMR imaging by applying a radio frequency (RF) excitation pulse to the sample in the presence of a magnetic field gradient. This is in contrast to spatial encoding, where the magnetization following excitation is allowed to freely evolve in the presence of a gradient. A simple appreciation of how

slice selection works is afforded by comparing the spread of resonance frequencies of the nuclei in the sample with the frequency bandwidth of the RF pulse. The resonance frequencies of nuclear spins in a sample placed centrally in a magnetic field, B_0 , with a superimposed constant gradient \mathbf{G} vary linearly across the sample between $\omega_0 \pm \gamma G d_s / 2$ where d_s is the dimension of the sample in the gradient direction. To a first approximation, the radio-frequency excitation pulse (of duration t_w and carrier frequency ω_0) contains frequency components in the range $\omega_0 \pm \pi / t_w$. If the pulse bandwidth is significantly less than the spread of frequencies within the sample, the condition for a so called *soft pulse*, then the pulse excites nuclei only in a central slice of the sample, perpendicular to the gradient, where the frequencies match. A slice at a position other than the centre of the sample can be chosen by offsetting the excitation carrier frequency, ω_0 .

A more detailed description of the action of an arbitrary pulse on a sample in a gradient can be obtained from a solution of the Bloch equations, either numerically or using advanced analytic techniques. In general this is complicated since the effective field in the rotating frame is composed not only of the spatially varying gradient field in the z -direction but also the transverse excitation field which, in general, varies with time. What follows is therefore an approximate treatment which nonetheless provides a surprisingly accurate description of many of the more commonly used slice selection pulses [4].

Suppose that a gradient, $\mathbf{G} = G_z \mathbf{u}_z$, is applied in the z -direction. In a local frame of reference rotating about the combined polarizing and gradient fields at the frequency $\omega = \omega_0 + \gamma(\mathbf{G} \cdot \mathbf{r})$, an excitation pulse $\mathbf{B}_1(t)$ applied at the central resonance

-4-

frequency of ω_0 and centred on $t = 0$ is seen to rotate at the offset frequency $\gamma(\mathbf{G} \cdot \mathbf{r})$ and therefore to have components given by

$$B_1(t) \left\{ \cos \left[\gamma(\mathbf{G} \cdot \mathbf{r}) \left(t + \frac{t_w}{2} \right) \right] \mathbf{u}'_x + \sin \left[\gamma(\mathbf{G} \cdot \mathbf{r}) \left(t + \frac{t_w}{2} \right) \right] \mathbf{u}'_y \right\}.$$

The excitation field is the only field seen by the magnetization in the rotating frame. The magnetization precesses about it. Starting from equilibrium ($\mathbf{M}_0 = M_0 \mathbf{u}_z$), transverse components are created and develop according to

$$\frac{d\mathbf{m}(\mathbf{r}, t)}{dt} = i\gamma M'_z(\mathbf{r}, t) B_1(t) \exp \left[-i\gamma(\mathbf{G} \cdot \mathbf{r}) \left(t + \frac{t_w}{2} \right) \right].$$

The simplifying approximation of a linear response is now made, by which it is assumed that rotations about different axes may be decoupled. This is only strictly valid for small rotations, but is surprisingly good for larger rotations too. This means that $M'_z(\mathbf{r}, t) \approx M_0(\mathbf{r})$ constant. Accordingly, at the end of the pulse the transverse magnetization is given by

$$m\left(\mathbf{r}, t = \frac{t_w}{2}\right) \approx i\gamma M_0(\mathbf{r}) \exp\left[-i\gamma(\mathbf{G} \cdot \mathbf{r})\frac{t_w}{2}\right] \int_{-\frac{t_w}{2}}^{+\frac{t_w}{2}} B_1(t) \exp[-i\gamma(\mathbf{G} \cdot \mathbf{r})t] dt. \quad (\text{B1.14.3})$$

The integral describes the spatial amplitude modulation of the excited magnetization. It represents the excitation or slice profile, $g(z)$, of the pulse in real space. As B_1 drops to zero for t outside the pulse, the integration limits can be extended to infinity whereupon it is seen that the excitation profile is the Fourier transform of the pulse shape envelope:

$$g(z) = \gamma \int_{-\infty}^{+\infty} B_1(t) \exp(-i\gamma G_z t z) dt.$$

For a soft pulse with a rectangular envelope

$$B_1(t) = \begin{cases} B_1 & \text{for } -\frac{t_w}{2} \leq t \leq \frac{t_w}{2} \\ 0 & \text{otherwise} \end{cases}$$

-5-

and a carrier frequency resonant at the position $z=z_0$ the excitation profile is sinc shaped. In normalized form, it is:

$$\frac{g(z - z_0)}{g(0)} = \text{sinc}\left[\frac{1}{2\pi}(z - z_0)\gamma G_z t_w\right]$$

where

$$\text{sinc}(x) = \frac{\sin(\pi x)}{\pi x}.$$

A sinc-shape has side lobes which impair the excitation of a distinct slice. Other pulse envelopes are therefore more commonly used. Ideally, one would like a rectangular excitation profile which results from a sinc-shaped pulse with an infinite number of side lobes. In practice, a finite pulse duration is required and therefore the pulse has to be truncated, which causes oscillations in the excitation profile. Another frequently used pulse envelope is a Gaussian function:

$$B_1(t) = B_1(0) e^{-at^2}$$

which produces a Gaussian slice profile:

$$\frac{g(z - z_0)}{g(0)} = e^{-\Omega_{z_0}^2/(4a)}$$

where $\Omega_{z_0} = \gamma G_z(z - z_0)$ and a is a pulse-width parameter. The profile is centred on z_0 at which position the transmitter frequency is resonant. The full width at half maximum of the slice, $\Delta_{1/2}$, is determined by the constant:

$$\Delta z_{1/2} = \frac{4}{\gamma G_z} \sqrt{a \ln 2}.$$

For a given half width at half maximum in the time domain, $\Delta t_{1/2} = 2\sqrt{\frac{\ln 2}{a}}$, the slice width $\Delta z_{1/2}$ decreases with increasing gradient strength G_z .

Closer examination of [equation B1.14.3](#) reveals that, after the slice selection pulse, the spin isochromats at different positions in the gradient direction are not in phase. Rather they are rotated by $i \exp(-i\gamma G_z z \frac{t_w}{2})$ and therefore have a phase $(\gamma G_z z t_w - \pi)/2$. Destructive interference of isochromats at different locations leads to cancellation of the total transverse magnetization.

The dephased isochromats may be refocused by applying one or more RF and trimming gradient pulses after the excitation. In the pulse sequence shown in [figure B1.14.1](#) the magnetization is refocused by a negative trimming z -gradient pulse of the same amplitude, but half the duration as the original slice selection gradient pulse. The trimming gradient causes a precession of the magnetization at every location which exactly compensates that occurring during the initial excitation (save for the constant factor π). A similar effect may be achieved using a (non-selective) 180° pulse and a trimming gradient pulse of the same sign.

B1.14.2.2 FREQUENCY ENCODING

Once a slice has been selected and excited, it is necessary to encode the ensuing NMR signal with the coordinates of nuclei within the slice. For each coordinate (x and y) this is achieved by one of two very closely related means, frequency encoding or phase encoding [1]. In this section we consider the former and in the next, the latter. In the section after that we show how the two are combined in the most common imaging experiment.

As before, we note that the resonance frequency of a nucleus at position \mathbf{r} is directly proportional to the combined applied static and gradient fields at that location. In a gradient $\mathbf{G} = G_x \mathbf{u}_x$, orthogonal to the slice selection gradient, the nuclei precess (in the usual frame rotating at ω_0) at a frequency $\omega = \gamma G_{xx}$. The observed signal therefore contains a component at this frequency with an amplitude proportional to the local spin density. The total signal is of the form

$$S(t_x) = \int_{-\infty}^{+\infty} dx m(x) e^{-i\gamma G_x x t_x}$$

from which it is seen that the spin density in the x -direction is recovered by a Fourier transform of the signal with respect to time

$$m(x) \propto \int dt_x S(t_x) \exp(i\gamma G_x x t_x).$$

In practice, it is generally preferable to create and record the signal in the form of an echo well clear of pulse

ring-down and other artifacts associated with defining the zero of time. This can be done either by first applying a negative gradient lobe followed by a positive gradient—a gradient echo—or by including a 180° inversion pulse between two positive gradients—a spin echo. [Figure B1.14.1](#) demonstrates the spin echo. A trimming x -gradient of duration t_y is placed before the 180° pulse which inverts the phase of the magnetization, so that, with reference to [figure B1.14.1](#).

$$m(\mathbf{r}, t_x) = m(\mathbf{r}, 0) e^{i\gamma G_x x t_y} e^{-i\gamma G_x x t_x}.$$

The maximum signal appears at the echo centre when the exponent disappears for all \mathbf{r} at time $t_x = t_{x0}$. If the magnitude of the read gradient and the corresponding trimming gradient is the same, then $t_{x0} = t_y$. The magnetization profile is obtained by Fourier transforming the echo.

B1.14.2.3 PHASE ENCODING

If a gradient pulse is applied for a fixed evolution time t_y the magnetization is dephased by an amount dependent on the gradient field. The signal phase immediately after the gradient varies linearly in the direction of the gradient. For a y -gradient $\mathbf{G} = G_y \mathbf{u}_y$ of duration t_y as shown in [figure B1.14.1](#) we have:

$$m(\mathbf{r}, G_y) = m(\mathbf{r}, 0) e^{-i\gamma G_y y t_y}$$

-7-

For phase encoding the phase twist is most commonly varied by incrementing G_y in a series of subsequent transients as this results in a constant transverse relaxation attenuation of the signal at the measurement position. The signal intensity as a function of G_y is

$$S(G_y) = \int_{-\infty}^{+\infty} dy m(y) e^{-i\gamma G_y y t_y}.$$

The magnetization profile in the y -direction is recovered by Fourier transformation with respect to G_y .

B1.14.2.4 2D SPIN-ECHO FT IMAGING AND K-SPACE

We now bring all the elements of the imaging experiment together within the typical spin-warp imaging sequence [10] previously depicted in [figure B1.14.1](#). A soft 90° pulse combined with slice-selection gradient G_z excites a slice in the x/y -plane. The spin isochromats are refocused by the negative z -gradient lobe. A subsequent spin echo, SE , is formed by a hard 180° pulse inserted after the slice selection pulse. The G_x -gradient either side of the 180° pulse first dephases and then rephases the magnetization so that an echo forms at twice the time separation, τ , between the two pulses. The x -dimension is encoded by acquiring the echo in the presence of this gradient, often known as a readout gradient. The y -dimension is phase encoded using the gradient G_y which is incremented in subsequently measured transients. The signal intensity S of the echo is the superposition of all the transverse magnetization originating from the excited slice. The acquired data set is described by:

$$S(t_x, G_y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} dx dy m(x, y) e^{-i\gamma G_x x (t_x - t_{x0})} e^{i\gamma G_y y t_y}.$$

The magnetization density is recovered by a two-dimensional Fourier transform of the data with respect to t_x

and G_y .

From a more general point of view, components k_j , $j=x,y,z$ of a wave vector \mathbf{k} which describes the influence of all gradient pulses may be defined as follows: $k_j(t) = \int_0^t \gamma G_j(t') dt'$. For the 2D imaging pulse sequence discussed here, $k_x = \gamma G_x(t_x - t_{x0})$ and $k_y = -\gamma G_y t_y$ (the negative signs result from the inversion pulse). Spatial encoding is the sampling of \mathbf{k} -space and the acquired data set is then:

$$S(k_x, k_y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} dx dy m(x, y) e^{-ik_x x} e^{-ik_y y}.$$

The image, i.e., the spatially resolved distribution of the magnetization $m(x,y)$, is reconstructed by two-dimensional Fourier transformation with respect to k_x and k_y . In the absence of other interactions and encodings discussed below, it represents the spin density distribution of the sample.

There is of course no requirement to confine the slice selection to the z -gradient. The gradients may be used in any combination and an image plane selected in any orientation without recourse to rotating the sample.

-8-

Another frequently used imaging method is gradient-recalled spin-echo imaging (figure B1.14.2). In this method, the 180° pulse of the spin warp experiment is omitted and the first lobe of the G_x gradient is instead inverted. Otherwise the experiment is the same. As the refocusing 180° pulse is omitted, the echo time T_E can be adjusted to be shorter than in the spin-echo version. Therefore gradient-recalled echo and variants of this technique are used when samples with shorter T_2 are to be imaged. On the other hand, this method is more susceptible to off-resonance effects, e.g., due to chemical shift or magnetic field inhomogeneities.

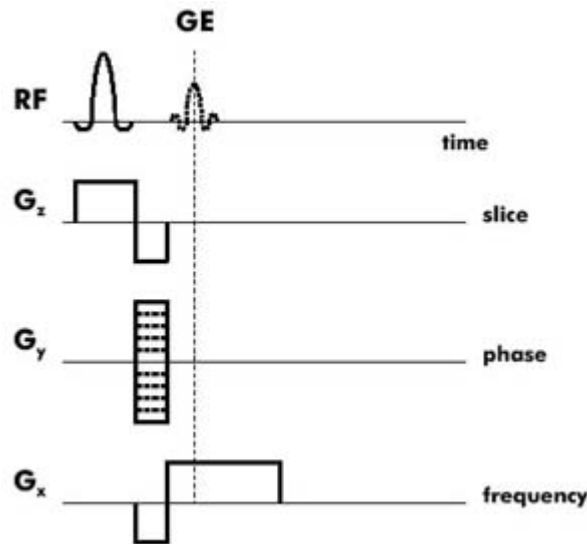


Figure B1.14.2. Gradient-recalled echo pulse sequence. The echo is generated by deliberately dephasing and refocusing transverse magnetization with the readout gradient. A slice is selected in the z -direction and x - and y -dimension are frequency and phase encoded, respectively.

The imaging methods just described require n transients each containing l data points to be acquired so as to construct a two-dimensional image of a matrix of $n \times l$ pixels. Since it is necessary to wait a time of the order of the spin-lattice relaxation time, T_1 , for the spin system to recover between the collection of transients, the total imaging time is in excess of nT_1 for a single average. This may mean imaging times of the order of minutes. The \mathbf{k} -space notation and description of imaging makes it easy to conceive of single transient

imaging experiments in which, by judicious switching of the gradients so as to form multiple echoes, the whole of k -space can be sampled in a single transient.

Techniques of this kind go by the generic title of echo-planar imaging methods [5, 6, 7 and 8] and in the case of full three-dimensional imaging, echo-volumar imaging. A common echo-planar imaging pulse sequence—that for blipped echo-planar imaging—is shown in figure B1.14.3. Slice selection is as before. The alternate positive and negative pulses of x -gradient form repeated gradient echoes as k -space is repeatedly traversed in the positive and negative x -directions. These echoes are used for frequency encoding. The initial negative pulse of y -gradient followed by the much smaller pulses of positive y -gradient ensure that each traverse in the x -direction is for a different value of k_y , starting from an extreme negative value.

-9-

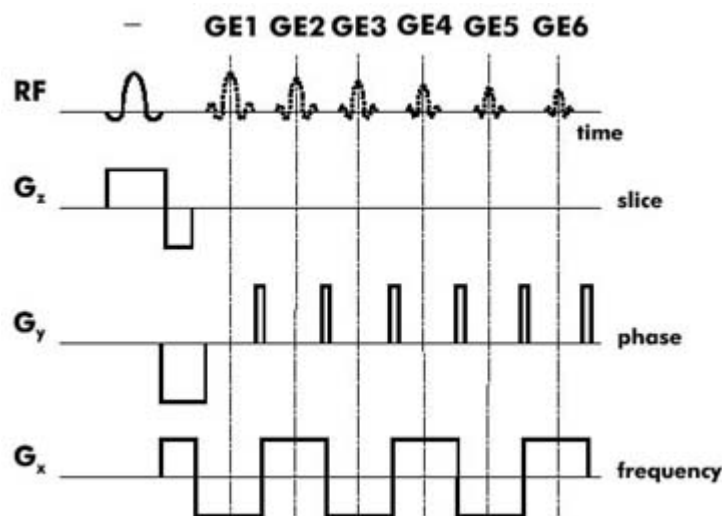


Figure B1.14.3. Echo-planar imaging (EPI) pulse sequence. In analogy to the pure gradient-echo recalled pulse sequence a series of echoes GE1–GE6 is refocused by alternately switching between positive and negative readout gradients. During the readout gradient switching a small phase-encoding gradient pulse (blip) is applied. The spatial phase encoding is hence stepped through the acquired echo train.

In cases where it is not possible to rapidly switch the gradients, frequency-encoded profiles may be acquired in different directions by rotating the gradient and/or sample orientation between transients. The image is reconstructed using filtered back-projection algorithms [9, 10]. The two-dimensional raster of k -space for the spin warp experiment shown in figure B1.14.1 is shown in figure B1.14.4(a) and that for the blipped echo-planar method in figure B1.14.4(b). The raster for back-projection methods is shown in figure B1.14.4(c).

-10-

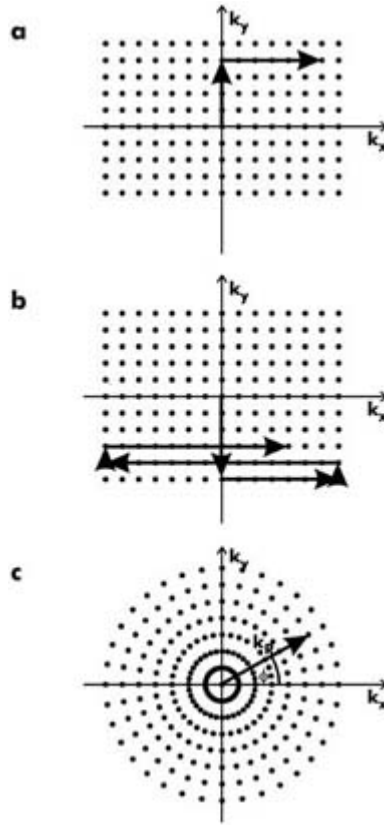


Figure B1.14.4. k -space representations of the (a) spin-echo imaging pulse sequence in [figure B1.14.1](#), (b) echo-planar imaging sequence in [figure B1.14.3](#) and (c) back-projection imaging. In (a) and (b) the components of the wave vectors k_x and k_y represent frequency and phase encoding, respectively. For back-projection (c) solely a readout gradient for frequency encoding is used. The direction of this gradient is changed stepwise from 0 to 180° in the x/y -plane (laboratory frame) by appropriate superposition of the x - and y -gradient. The related wave vector \mathbf{k}_R then rotates around the origin sampling the k_x/k_y -plane. As the sampled points are not equidistant in k -space, an image reconstruction algorithm different from the two-dimensional Fourier transform has to be used [9, 10].

B1.14.2.5 RESOLUTION

The achievable spatial resolution is limited by several effects. The first is the maximum gradient strength and encoding times available. Bearing in mind [equation B1.14.1](#) and $t_x^{\max} = 1/\Delta f_{\max}$ the pixel size resulting from the Fourier transform of the frequency encoded echo is given by

$$\Delta x = \frac{2\pi}{\gamma G_x t_x^{\max}} \quad (\text{B1.14.4})$$

where t_x^{\max} is the total data acquisition time. The maximum useful encoding time is limited by the transverse relaxation time of the nuclei, T_2 . Therefore the best resolution which can be achieved is of the order of [14]

$$\Delta x^{\text{best}} = \frac{2\pi}{\gamma G_x T_2}. \quad (\text{B1.14.5})$$

A similar result holds for phase encoding where the gradient strength appearing in the expression for the resolution is the maximum gradient strength available. According to these results the resolution can be improved by raising the gradient strength. However, for any spectrometer system, there is a maximum gradient strength which can be switched within a given rise time due to technical limitations. Although modern gradient coil sets are actively shielded to avoid eddy currents in the magnet, in reality systems with pulsed gradients in excess of 1 T m^{-1} are rare. Since the T_2 of many commonly imaged, more mobile, samples is of the order of 10 ms, resolution limits in ^1H imaging are generally in the range one to ten micrometres. At this resolution, considerable signal averaging is generally required in order to obtain a sufficient signal to noise ratio and the imaging time may extend to hours. Moreover, a slice thickness of typically $500 \mu\text{m}$, which is significantly greater than the lateral resolution, is frequently used to improve signal to noise. In solids, T_2 is generally very much shorter than in soft matter and high resolution imaging is not possible without recourse either to sophisticated line narrowing techniques [11], to magic-echo refocusing variants [2] or to very high gradient methods such as stray field imaging [13].

Motion, and in particular diffusion, causes a further limit to resolution [14, 15]. First, there is a physical limitation caused by spins diffusing into adjacent voxels during the acquisition of a transient. For water containing samples at room temperature the optimal resolution on these grounds is about $5 \mu\text{m}$. However, as will be seen in subsequent sections, diffusion of nuclei in a magnetic field gradient causes an additional attenuation of the signal in the time domain. In the presence of a steady gradient, it is $\exp(-\gamma^2 G^2 D t^3/3)$. Hence, the linewidth for spins diffusing in a gradient is of the order of

$$\Delta f_D = 0.6(\gamma^2 G^2 D/3)^{\frac{1}{3}}$$

where D is the diffusion coefficient, so that the best resolution becomes in analogy to the derivation of the equation B1.14.4 and equation B1.14.5

$$\Delta x^{\text{best}} \approx 2.6 \left(\frac{D}{\gamma G} \right)^{\frac{1}{3}} .$$

In practice, internal gradients inherent to the sample resulting from magnetic susceptibility changes at internal interfaces can dominate the applied gradients and lead to strong diffusive broadening just where image resolution is most required. Again the resolution limits tend to be on the ten micrometre scale.

B1.14.3 CONTRASTS IN MR IMAGING

Almost without exception, magnetic resonance ‘images’ are more than a simple reflection of nuclear spin density throughout the sample. They crudely visualize one or more magnetic resonance measurement parameters with which the NMR signal intensity is weighted. A common example of this kind is spin-relaxation-weighted image contrast. Mapping refers to encoding of the value of an NMR measurement parameter within each image pixel and thereby the creation of a map of this parameter. An example is a velocity map. The power of MRI compared to other imaging modalities is the large range of dynamic and microscopic structural contrast parameters which the method can encode, ‘visualise’ and map.

B1.14.3.1 RELAXATION

Transverse relaxation weighting is perhaps the most common form of contrast imposed on a magnetic

resonance image. It provides a ready means of differentiating between more mobile components of the sample such as low viscous liquids which are generally characterized by long T_2 values of the order of seconds and less mobile components such as elastomers, fats and adsorbed liquids with shorter T_2 values of the order of tens of milliseconds [6]. Transverse relaxation contrast is, in fact, a natural consequence of the spin warp imaging technique described in the previous section. As already seen, data are recorded in the form of a spatially encoded spin echo. Only those nuclides in the sample with a T_2 of the order of, or greater than, the echo time, $T_E=2\tau\approx 2t_y$, contribute significantly to the echo signal. Consequently, the image reflects the distribution of nuclides for which $T_2 \geq T_E$. Often, a crude distinction between two components in a sample, one more mobile than the other, is made on the basis of a single T_2 -weighted image in which the echo time is chosen intermediate between their respective T_2 values. For quantitative T_2 -mapping, images are recorded at a variety of echo times and subsequently analysed by fitting single- or multi-modal relaxation decays to the image intensity on a pixel by pixel basis. The fit parameters are then used to generate a true T_2 -map in its own right.

An example of the application of T_2 -weighted imaging is afforded by the imaging of the dynamics of chemical waves in the Belousov–Zhabotinsky reaction shown in [figure B1.14.5](#) [16]. In these images, bright bands correspond to an excess of Mn^{3+} ions with a long T_2 and dark bands to an excess of Mn^{2+} ions with a short T_2 .

-13-

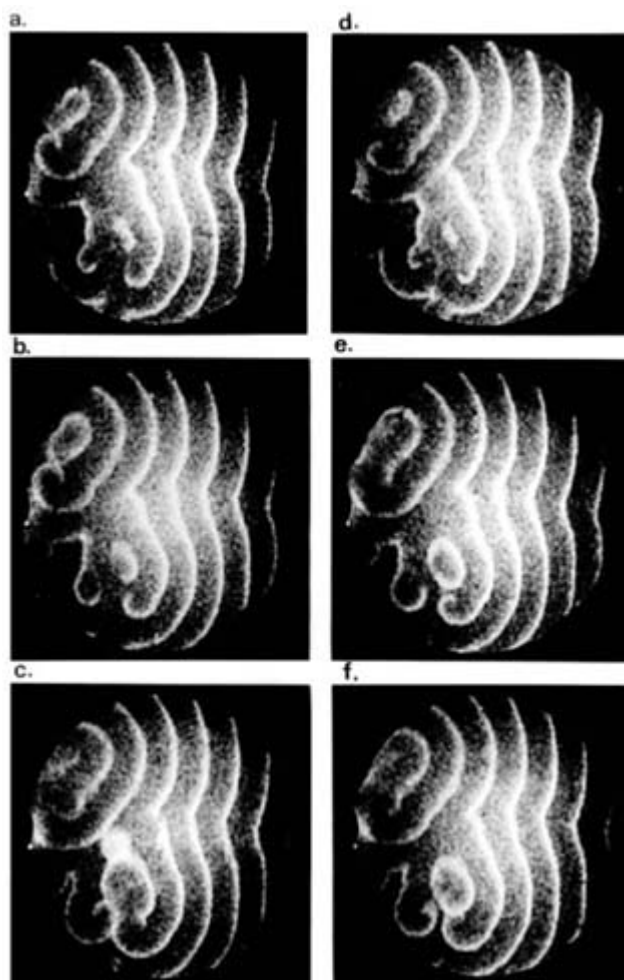


Figure B1.14.5. T_2 -weighted images of the propagation of chemical waves in an Mn^{2+} catalysed Belousov–Zhabotinsky reaction. The images were acquired in 40 s intervals (a) to (f) using a standard spin echo pulse sequence. The slice thickness is 2 mm. The diameter of the imaged pill box is 39 mm. The bright bands

correspond to an excess of Mn^{3+} ions with long T_2 , dark bands to an excess of Mn^{2+} with a short T_2 . (From [16]).

Another powerful contrast parameter is spin–lattice, or T_1 , relaxation. Spin–lattice relaxation contrast can again be used to differentiate different states of mobility within a sample. It can be encoded in several ways. The simplest is via the repetition time, T_R , between the different measurements used to collect the image data. If the repetition time is sufficiently long such that $T_R \gg T_1$ for all nuclei in the sample, then all nuclei will recover to thermal equilibrium between measurements and will contribute equally to the image intensity. However, if the repetition time is reduced, then those nuclei for which $T_R < T_1$ will not recover between measurements and so will not contribute to the subsequent measurement. A steady state rapidly builds up in which only those nuclei with $T_1 \ll T_R$ contribute in any significant manner. As with T_2 -contrast, single images recorded with a carefully selected T_R may be used to select crudely a short T_1 component of a sample.

-14-

The mathematical description of the echo intensity as a function of T_2 and T_1 for a repeated spin-echo measurement has been calculated on the basis that the signal before one measurement cycle is exactly that at the end of the previous cycle. Under steady state conditions of repeated cycles, this must therefore equal the signal at the end of the measurement cycle itself. For a spin-echo pulse sequence such as that depicted in [figure B1.14.1](#) the echo magnetization is given by [17]

$$M_{\text{echo}} = M_0 \frac{[1 - 2 \exp(-\frac{T_R - T_E/2}{T_1}) + \exp(-\frac{T_R}{T_1})] \exp(-\frac{T_E}{T_2})}{1 + \cos \theta \exp(-\frac{T_R}{T_1})} \sin \theta$$

where M_0 is the equilibrium magnetization and θ is the tip angle of the radio-frequency excitation pulse and where it is assumed that there is total dephasing of the magnetization between cycles. Other expressions applicable to other situations are to be found in the literature. In practice, some kind of relaxation-weighting of image contrast is always present and can hardly be avoided.

Other methods to encode T_1 contrast include saturation recovery [18, 19] and T_1 nulling techniques [20]. In the latter, a 180° pulse is applied some time T_0 before the image data acquisition. This pulse inverts the magnetization. In the interval T_0 the magnetization recovers according to $[1 - 2 \exp(-t/T_1)]$ so that at the time of image data excitation it is $[1 - 2 \exp(-T_0/T_1)]$. Nuclei for which $T_0 = 0.693 T_1$ have zero magnetization at this time and so do not contribute to the signal intensity. This method may be used to suppress a large background component in an image, such as that due to bulk water. With saturation recovery, a train of radio-frequency pulses is applied to the sample some time T_{SR} prior to the data acquisition sequence. The train of pulses, which are often applied with ever decreasing spacing between them, serves to saturate the equilibrium magnetization. Only those nuclides for which $T_{SR} < T_1$ recover to equilibrium prior to the image acquisition proper and so only these nuclei contribute to the image intensity. Multiple images recorded as a function of T_{SR} may be used to build a true T_1 -map by fitting a relaxation recovery curve to the data on a pixel by pixel basis. An example of brine in a sandstone core is depicted in [Figure B1.14.6](#)[21]. Both M_0 - and T_1 -maps, which were obtained from fitting with a stretched exponential, clearly show layers in the stone. The spin–lattice relaxation presumably correlates with spatially varying pore sizes and surface relaxivity.

-15-

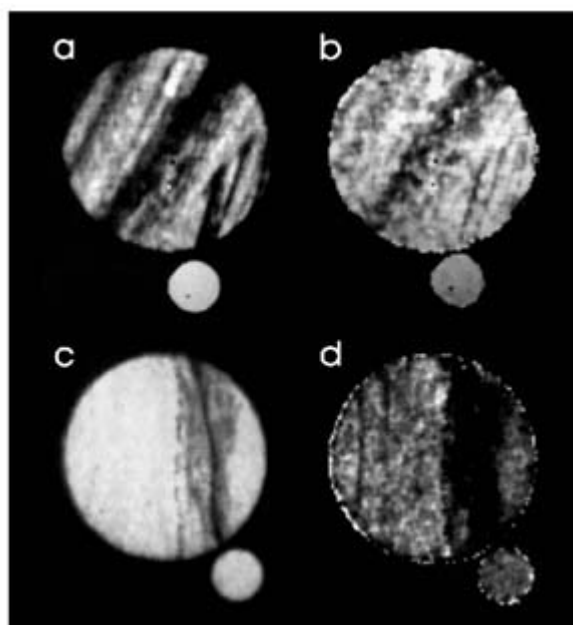


Figure B1.14.6. T_1 -maps of a sandstone reservoir core which was soaked in brine. (a), (b) and (c), (d) represent two different positions in the core. For T_1 -contrast a saturation pulse train was applied before a standard spin-echo imaging pulse sequence. A full T_1 -relaxation recovery curve for each voxel was obtained by incrementing the delay between pulse train and imaging sequence. M_0 - ((a) and (c)) and T_1 -maps ((b) and (d)) were calculated from stretched exponentials which are fitted to the magnetization recovery curves. The maps show the layered structure of the sample. Presumably T_1 -relaxation varies spatially due to inhomogeneous size distribution as well as surface relaxivity of the pores. (From [21].)

B1.14.3.2 CHEMICALLY RESOLVED IMAGING

In many instances, it is important that some form of chemical selectivity be applied in magnetic resonance imaging so as to distinguish nuclei in one or more specific molecular environment(s). There are many ways of doing this and we discuss here just three. The first option is to ensure that one of the excitation RF pulses is a narrow bandwidth, frequency selective pulse applied in the absence of any gradient [22]. Such a pulse can be made specific to one particular value of the chemical shift and thereby affects only nuclei with that chemical shift. In practice this can be a reasonable method for the specific selection of fat or oil or water in a mixed hydrocarbon/water system.

A higher level of sophistication involves obtaining a full chemical shift spectrum within each image pixel. The chemical information can be encoded either before or after the image encoding. The important requirement is that the spin system is allowed to evolve without gradients and that data are recorded as a function of this chemical shift evolution time. The data can then be Fourier transformed with respect to the evolution time as well as the standard imaging variables so as to yield the spatially resolved chemical shift spectrum. In this respect chemical shift imaging is like a four-dimensional imaging experiment [23]. A post-encoding sequence suitable for this purpose is shown in [figure B1.14.7](#). The chemical shift encoding part—a spin echo in the absence of gradients—comes after the image encoding part—a two-dimensional phase encoding experiment. The 180° pulse refocuses all chemical-shift-induced dephasing occurring during the spatial encoding.

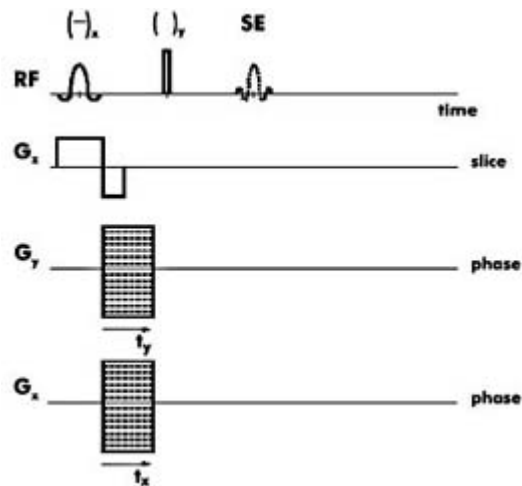
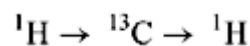


Figure B1.14.7. Chemical shift imaging sequence [23]. Both x - and y -dimensions are phase encoded. Since line-broadening due to acquiring the echo in the presence of a magnetic field gradient is avoided, chemical shift information is retained in the echo.

The third alternative is a more robust, sensitive and specialized form of the first, in that only hydrogen nuclei indirectly spin–spin coupled to ^{13}C in a specific molecular configuration are imaged. In achieving selectivity, the technique exploits the much wider chemical shift dispersion of ^{13}C compared to ^1H . The method involves cyclic transfer from selected ^1H nuclei to indirectly spin–spin coupled ^{13}C nuclei and back according to the sequence



Called CYCLCROP (cyclic cross polarization) [24], the method works by first exciting all ^1H magnetization. Cross polarization pulses are then applied at the specific Larmor frequencies of the ^1H – ^{13}C pair of interest so as to transfer coherence from ^1H to ^{13}C . The transfer pulses must satisfy the Hartmann–Hahn condition

$$\gamma_{\text{C}} B_{1\text{C}} = \gamma_{\text{H}} B_{1\text{H}}$$

$B_{1\text{H}}$ and $B_{1\text{C}}$ are the excitation magnetic field strengths and must be applied for a time of the order of $1/J$ where J is the spin–spin coupling constant. Following magnetization transfer, the ^{13}C magnetization is stored along the z -axis and the remaining ^1H magnetization from other molecular groupings is destroyed by a series of pulses and homospoil gradients. The stored magnetization specific to the coupled ^1H – ^{13}C pair is then returned to the ^1H by a second pair of cross polarization pulses. CYCLCROP chemical selective excitation may replace the initial excitation in a standard

2DFT imaging experiment with the slice selection moved to the 180° pulse in order to yield a ^{13}C edited image. A number of pulse schemes for the cross coupling are known, each with various advantages in terms of low radio-frequency power deposition, tolerance of pulse artifact, breadth of the spectral bandwidth etc. In figure B1.14.8. CYCLCROP has been used to map ^{13}C labelled sucrose in the stem of a castor bean seedling [25]. Its arrival and accumulation are visualized in a series of subsequently acquired ^{13}C -selective images.

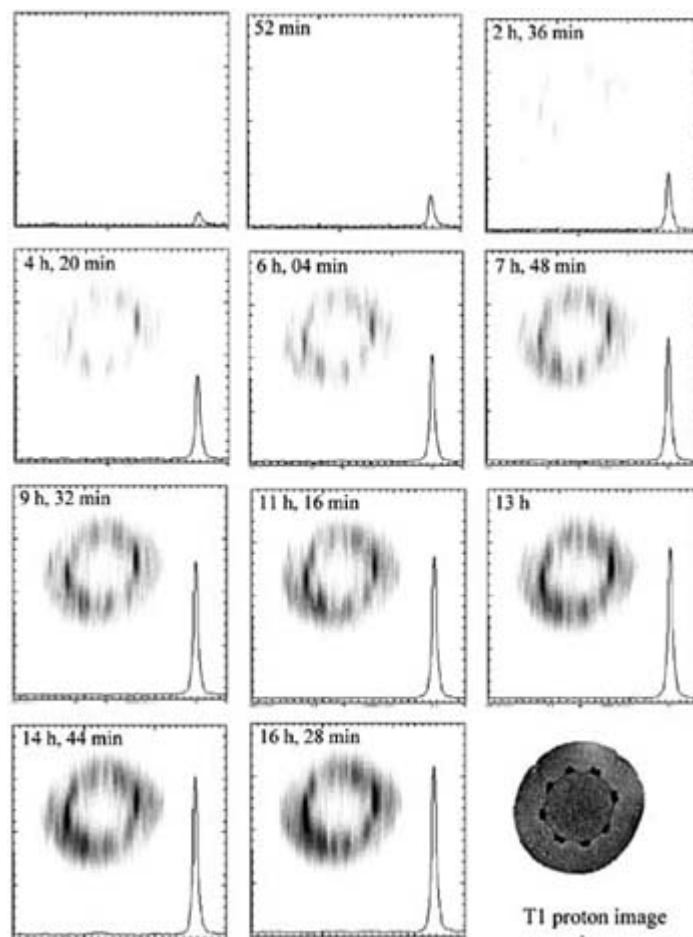


Figure B1.14.8. Time course study of the arrival and accumulation of ^{13}C labelled sucrose in the stem of a castor bean seedling. The labelled tracer was chemically, selectively edited using CYCLCROP (cyclic cross polarization). The first image in the upper left corner was taken before the incubation of the seedling with enriched hexoses. The time given in each image represents the time elapsed between the start of the incubation and the acquisition. The spectrum in the lower right corner of each image shows the total intensity of ^{13}C nuclei. At later times, enriched sucrose is visible in the periphery of the stem. Especially high intensities are detected in the vascular bundles. The last image represents a micrograph of the stem structure showing the position of the vascular bundles (dark features). (From [25]).

B1.14.4 FLOW AND DIFFUSION

NMR is an important technique for the study of flow and diffusion, since the measurement may be made highly sensitive to motion without in any way influencing the motion under study. In analogy to many non-NMR-methods, mass transport can be visualized by imaging the distribution of magnetic tracers as a function of time. Tracers may include paramagnetic contrast agents which, in particular, reduce the transverse relaxation time of neighbouring nuclei and therefore appear as T_2 -contrast in an image. The ^{13}C cross polarization method with enriched compounds may also be used as a tracer experiment. More sophisticated tracer methods include so called ‘tagging’ experiments in which the excitation of nuclei is spatially selective. The spatial evolution of the selected nuclei is followed. This example is discussed in section B1.14.4.5.

Generally, however, the application of tracer methods remains a rarity compared to methods which directly exploit the motion sensitivity of the NMR signal. The detection of motion is based on the sensitivity of the

signal phase to translational movements of nuclei in the presence of magnetic field gradients [26, 27]. Using the large magnetic field gradient in the stray field of superconducting magnets, displacements as small as 10 nm in slowly diffusing polymer melts can be detected. At the other extreme, velocities of the order of $m\ s^{-1}$, such as occur in blood-filled arteries, can be measured.

B1.14.4.1 COHERENT AND STATIONARY FLOW

The displacement of a spin can be encoded in a manner very similar to that used for the phase encoding of spatial information [28, 29 and 30]. Consider a spin j with position $r(t)$ moving in a magnetic field gradient G . The accumulated phase, ϕ_j , of the spin at time t is given by

$$\phi_j(t) = -\gamma \int_0^t G(r') \cdot r_j(t') dt' \tag{B1.14.6}$$

In order to encode displacement as opposed to average position, the gradient is applied in such a manner as to ensure that $\int_0^t G(r') dt' = 0$. Generally, this means applying gradient pulses in bipolar pairs or applying unimodal gradient pairs either side of a 180° RF inversion pulse, with the advantage that the necessary careful balancing of the gradient amplitudes is more straightforward.

We first examine how this works for the case of coherent flow. A typical pulse sequence is shown in [figure B1.14.9](#). This sequence creates a spin echo using two unipolar gradient pulses on either side of a 180° pulse. The duration of each gradient pulse of strength G_v is δ . The centres of the gradient pulses are separated by Δ .

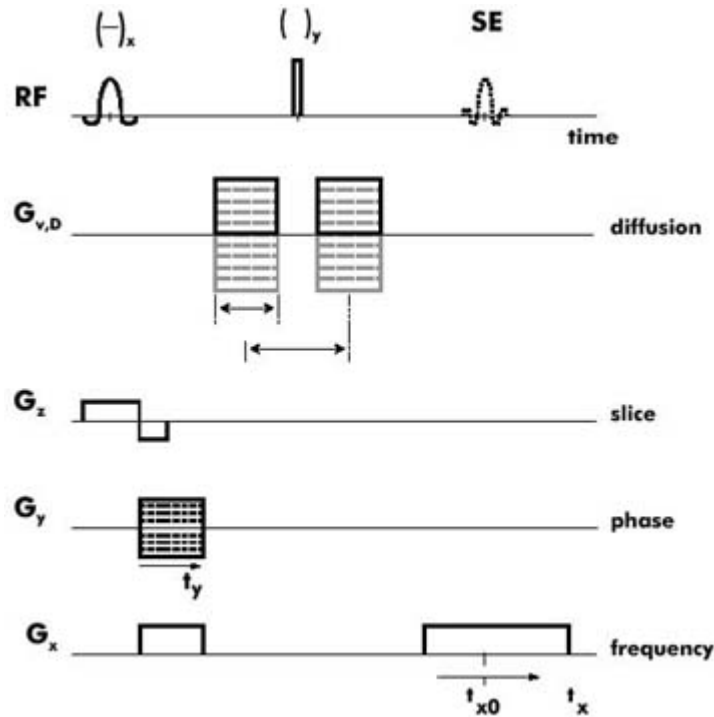


Figure B1.14.9. Imaging pulse sequence including flow and/or diffusion encoding. Gradient pulses $G_{v,D}$ before and after the inversion pulse are supplemented in any of the spatial dimensions of the standard spin-echo imaging sequence. Motion weighting is achieved by switching a strong gradient pulse pair $G_{v,D}$ (see solid black line). The steady-state distribution of flow (coherent motion) as well as diffusion (spatially

incoherent motion) in a sample is encoded by incrementing $G_{v,D}$ (see dashed grey lines). The measured data set then consists of two spatial and a motion-encoded dimension. Velocity and/or diffusion maps can be rendered by three-dimensional Fourier transformation.

Under steady-state flow conditions (coherent motion), a Taylor series can be applied to describe the time-dependent position of the fluid molecules:

$$\mathbf{r}(t) = \mathbf{r}_0 + \mathbf{v}t + \frac{1}{2}\mathbf{a}t^2 + \dots$$

If terms of higher order than linear in t are neglected, the transverse magnetization evolves in the presence of the first bipolar gradient pulse according to (equation B1.14.2 and equation B1.14.6):

$$m(t) = m(0) e^{-i\gamma\mathbf{G}_v \cdot \mathbf{r}_0 t - i\frac{1}{2}\gamma\mathbf{G}_v \cdot \mathbf{v}t^2} \quad 0 \leq t \leq \delta.$$

The phase of the transverse magnetization is inverted by the 180° pulse and the magnetization after the second gradient pulse and therefore at the echo centre is:

-20-

$$\begin{aligned} m(\delta, \Delta) &= m(0) e^{+\gamma\mathbf{G}_v \cdot \mathbf{r}_0 \delta + \gamma\frac{1}{2}\mathbf{G}_v \cdot \mathbf{v}\delta^2} e^{-\gamma\mathbf{G}_v \cdot \mathbf{r}_0[\delta+\Delta] - \gamma\mathbf{G}_v \cdot \mathbf{v}[(\delta+\Delta)^2 - \Delta^2]} \\ &= m(0) e^{-i\gamma\mathbf{G}_v \cdot \mathbf{v}\delta\Delta}. \end{aligned}$$

The echo phase does not depend on the initial position of the nuclei, only on their displacement, $\mathbf{v}\Delta$, occurring in the interval between the gradient pulses. Analysis of the phase of the echo yields a measure of flow velocity in a bulk sample. Spatial resolution is easily obtained by the incorporation of additional imaging gradients. One way of doing this is illustrated in figure B1.14.9. The first part of the experiment is the same flow encoding experiment as just discussed. The velocity-encoded echo is the excitation for the subsequent 2DFT experiment which is as previously discussed. Where both stationary and moving spins are present, these superimpose in the image. A variety of methods exist for separating the two, including cycling the phase of the velocity encoding gradients or making measurements at two or more strengths of the velocity encoding gradient. In the latter, a wave number $\mathbf{k}_v = \gamma\mathbf{G}_v \cdot \mathbf{v}$ can be defined adding an additional dimension to spatial \mathbf{k} -space. Fourier transformation of this dimension directly produces the velocity spectrum of each voxel. As an example, velocity maps of flow through an extruder are shown in figure B1.14.10 [31].

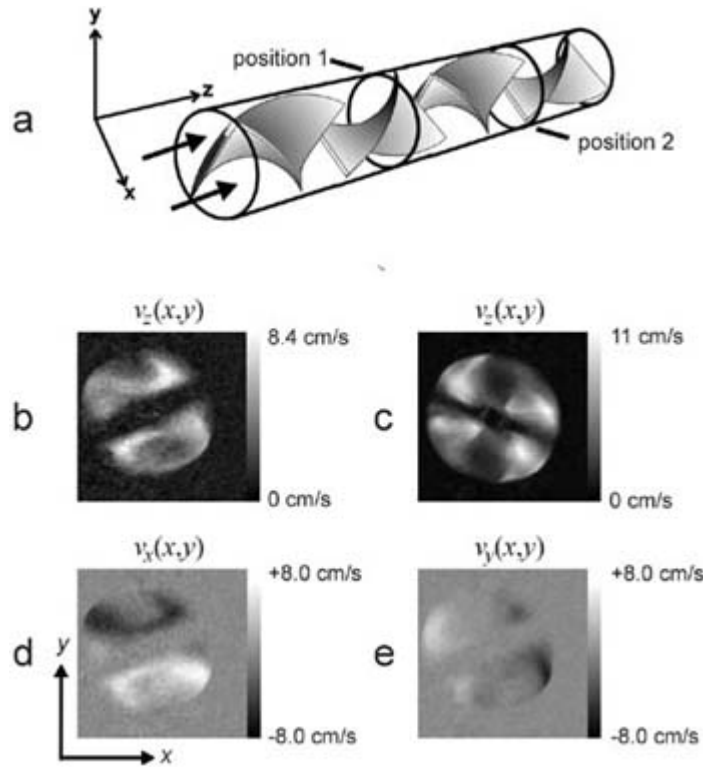


Figure B1.14.10. Flow through an KENICS mixer. (a) A schematic drawing of the KENICS mixer in which the slices selected for the experiment are marked. The arrows indicate the flow direction. Maps of the z -component of the velocity at position 1 and position 2 are displayed in (b) and (c), respectively. (d) and (e) Maps of the x - and the y -velocity component at position 1. The FOV (field of view) is 10 mm. (From [31].)

B1.14.4.2 PULSATING MOTION

Flow which fluctuates with time, such as pulsating flow in arteries, is more difficult to experimentally quantify than steady-state motion because phase encoding of spatial coordinate(s) and/or velocity requires the acquisition of a series of transients. Then a different velocity is detected in each transient. Hence the phase-twist caused by the motion in the presence of magnetic field gradients varies from transient to transient. However if the motion is periodic, e.g., $\mathbf{v}(\mathbf{r}, t) = \mathbf{v}_0 \sin\{\omega_v t + \phi_0\}$ with a spatially varying amplitude $\mathbf{v}_0 = \mathbf{v}_0(\mathbf{r})$, a pulsation frequency $\omega_v = \omega_v(\mathbf{r})$ and an arbitrary phase ϕ_0 , the phase modulation of the acquired data set is described as follows:

$$S(k_x, k_y, k_v) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} dx dy dv m(x, y, v) e^{-ik_x x - ik_y y + ik_v v} \sin(\omega_v l_p T_R + \phi_0)$$

where $t = l_p T_R$ is the time at which the l_p th transient is acquired. Since k_y and k_v are 'incremented' for phase encoding in subsequent transients, they are linked to the phase twist caused by $v(t)$ via the parameter l_p . The reconstruction of the dimensions y and v by Fourier transformation is therefore affected, making an interpretation difficult.

Nevertheless, averaging provides information about the motional parameters such as velocity amplitude and

pulsation frequency. If the repetition time, the delay between subsequent transients, is not equal to a multiple of the pulsation period the motion appears to be temporally uncorrelated. In this case, a temporal average over all velocity values is obtained by accumulating a sufficient number of transients. This causes a broad phase distribution resulting in signal attenuation similar to the one caused by diffusion, a spatially incoherent phenomenon. The images provide quantitative information about the distribution of motion and velocity amplitude. The temporal characteristics of the pulsation are detected by omitting phase encoding. Using a gradient pulse pair of constant magnitude for motion weighting the signal phase is then solely a function of the velocity. Since the y determined from the intensity modulation due to the motion in a series of transients.

Figure B1.14.11 shows the application of averaging techniques for the characterization of pulsating motions which start on about the fourth incubation day in quail eggs [32]. Dark areas in the incoherent motion-weighted images in figure B1.14.11 and represent strong motion, white no uncorrelated motion. They are localized at the suspected position of the embryo above the egg yolk which is the black region in the middle of the egg. The signal attenuation strongly depends on the probed velocity component indicating the anisotropic nature of the motion. In figure B1.14.12(a) and (b) a time series of profiles through the region of motion with spatially incoherent motion weighting were acquired before and after the start of pulsations. At the later incubation stage the modulation of the signal intensity due to temporally periodic motion is clearly visible. Fourier transformation (figure B1.14.12(d)) reveals a pulsation frequency of about 0.4 Hz.

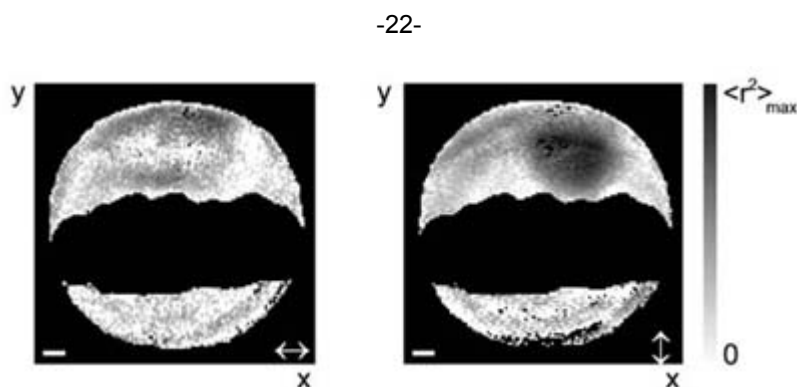


Figure B1.14.11. Amplitude-weighted images of (temporally) uncorrelated motions in a quail egg at an incubation period at about 140 h. A standard gradient echo sequence supplemented with strong bipolar gradient pulses for motion weighting was used. A high number of transients ($N_t = 490$) was acquired for each phase-encoding step to adequately average out temporal fluctuation of the motion. The intensity in the images shown corresponds to the signal ratio with and without motion weighting. Light grey shades hence represent no signal attenuation, darker shades strong signal attenuation due to uncorrelated motion. Pixels with signal below the noise level are set black as is the case in the egg yolk (black region in the middle of the egg) due to comparatively short T_2 . The white double arrows indicate the probed velocity component. Both images show signal attenuation due to strong motion in the region above the egg yolk where the embryo presumably is located. Furthermore, the attenuation of the signal appears to be much stronger for the y -velocity component than for the x -component indicating strongly anisotropic motion. The white bar represents 2 mm. (From [32].)

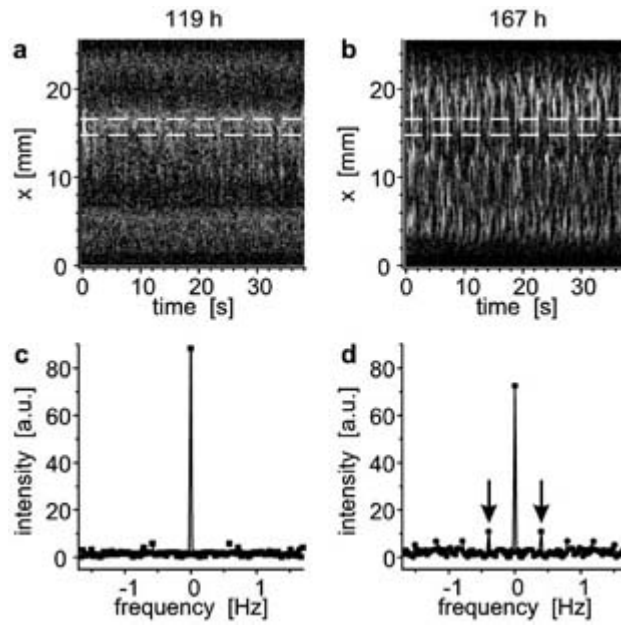


Figure B1.14.12. Study of the temporal fluctuation of motion in a quail egg at the incubation times 119 h and 167 h. Spatial phase encoding of the y -dimension was omitted to increase the rapidity of the imaging experiment. Profiles were obtained by 1D Fourier transformation of echoes which were acquired in the presence of a readout gradient for frequency encoding of the x -coordinate. A strong gradient pulse pair for (spatially) incoherent motion weighting was applied during the evolution period of the magnetization. A series of subsequent single-scan profiles were measured at the two different incubation times 119 h (a) and 167 h (c). Temporal fluctuations of the signal intensity which are not visible in (a) reveal themselves at the later incubation stage. The intensity modulation which was caused by temporally fluctuating motion was analysed by Fourier transformation. The spectra which were calculated from the integral intensities of the pixels between the two dashed lines in (a) and (c) are displayed in (b) and (d) for the two incubation times. The line at 0 Hz is due to the constant baseline offset. The arrows in (d) mark the peak representing a frequency of 0.4 Hz of the pulsating motion. (From [32].)

B1.14.4.3 DIFFUSION AND PSEUDO-DIFFUSION

If magnitude and/or direction of velocity vary on a length scale below spatial resolution, the detected motion is incoherent. (Self-) diffusion certainly falls into this category, but randomly oriented flow is also spatially incoherent. A well known example is the blood flow through brain capillaries, which are smaller than a voxel [33]. Incoherent flow is often referred to as pseudo-diffusion. An apparent diffusion coefficient which can be significantly bigger than the self-diffusion coefficient is then defined. Pulse sequences to measure coherent flow (figure B1.14.9) can also be used for (spatially) incoherent motion although the theory has to be reconsidered at this point [34, 35 and 36].

The observed magnetization in the echo is the superposition of all the spins j at different positions r_j :

$$\begin{aligned} m(T_E) &= \langle m(0) e^{i\gamma G_D \cdot (r_j(\Delta) - r_j(0))\delta} \rangle_j \\ &= \langle m(0) e^{i\varphi_j(t)} \rangle_j. \end{aligned}$$

It is expanded into:

$$m(T_E) = m(0) \sum_j P(\varphi_j) e^{i\varphi_j(t)}$$

where $P(\varphi_j)$ is the probability of phase φ_j . The underlying motional process influences this phase distribution. To demonstrate the principle, the simple case of normal isotropic diffusion will be discussed [27]. The solution of Fick's diffusion equation together with the central limit theorem implies that, for a constant gradient, a Gaussian phase distribution function with a mean squared phase twist $\overline{\varphi^2(t)}$ applies at any instant

$$P(\varphi) = \frac{1}{(2\pi\overline{\varphi^2})^{1/2}} \exp(-\varphi^2/\overline{\varphi^2})$$

leading to the transverse magnetization

$$m(t) = m(0) e^{-\overline{\varphi^2(t)}/2}. \quad (\text{B1.14.7})$$

The relationship between mean squared phase shift and mean squared displacement can be modelled in a simple way as follows: This motion is mediated by small, random jumps in position occurring with a mean interval τ_j . If the jump size in the gradient direction is ϵ , then after n jumps at time $t=n\tau_j$, the displacement of a spin is

$$E(n\tau_j) = \sum_{i=1}^n \epsilon a_i$$

where a_i is a randomly either +1 or -1. Hence, from the relation $\varphi = \gamma G_D \delta E(n\tau_j)$, the phase shift distribution imposed by the diffusion measurement gradients G_D of duration δ is

$$\overline{\varphi^2} = \gamma^2 \delta^2 G_D^2 \epsilon^2 \overline{\left(\sum_{i=1}^n a_i \right)^2}.$$

-25-

The summation averages to n . Using the definition of the diffusion coefficient, $D = \epsilon^2 / (2\tau_j)$, and the diffusion time, $\Delta = n\tau_j$, [equation B1.14.7](#) gives

$$m(t) = m(0) \exp(-\gamma^2 \delta^2 G_D^2 D \Delta).$$

This is the factor by which the echo magnetization is attenuated as a result of diffusion. More elaborate calculations, which account for phase displacements due to diffusion occurring during the application of the gradient pulses yield

$$m(t) = m(0) \exp[-\gamma^2 \delta^2 G_D^2 D (\Delta - \delta/3)].$$

This expression can be used for pulsed field gradient spin-echo experiments and also for spin-echo experiments in which the gradient is applied continuously.

A measure of the echo attenuation within each pixel of an image created using the pulse sequence of [figure B1.14.9](#) perhaps by repeating the experiment with different values of G_D and/or δ , gives data from which a true diffusion map can be constructed [37, 38].

In principle, it is possible to measure both flow and diffusion in a single experiment: the echo is both phase shifted and attenuated. It is also possible to account for the presence of background gradients arising from the sample itself which can be significant. The exact form of the echo in these circumstances has been calculated and the results are to be found in the literature [39, 40, 41 and 42]. Furthermore, in systems in which T_2 is relatively short compared to T_1 , the stimulated echo comprising three 90° pulses can be used instead of a 90° – t – 180° spin-echo sequence [43]. In this case, the fact that the velocity encoding time, Δ , can be made significantly longer outweighs the fact that only half of the magnetization is refocused after the third pulse.

B1.14.4.4 RESTRICTED DIFFUSION PULSED FIELD GRADIENT MICROSCOPY

Diffusometry and spatially resolved magnetic resonance are usefully combined in an alternate technique to imaging which is increasingly coming to the fore. It has been dubbed both q -space microscopy [4] and restricted diffusion pulsed field gradient (PFG) microscopy. The method probes the microstructure of a sample on the micrometre scale—significantly smaller than conventional MRI permits—by measuring the effects of restricted diffusion of a translationally mobile species. The technique yields parameters characteristic of the average structure of the whole sample (the experiment done without spatial resolution) or of the sample within an image pixel (sub-millimetre scale) if the experiment is done in combination with conventional imaging. The method works because of a Fourier relationship which exists between the observed echo attenuation in a pulsed field gradient diffusometry experiment and the propagator, which describes the molecular motion $P(\mathbf{r}; \mathbf{r}', t)$. $P(\mathbf{r}; \mathbf{r}', t)$ is the conditional probability of finding a diffusing spin at location \mathbf{r}' at time t given its initial location, \mathbf{r} . This propagator depends intimately on the microstructure of the sample. Following on from the above analysis, the echo attenuation is

$$E(G_D, \delta, \Delta) = \int_{\mathbf{r}} \rho(\mathbf{r}) \int_{\mathbf{r}'} P(\mathbf{r}; \mathbf{r}', \Delta) \exp[i\gamma\delta G_D \cdot (\mathbf{r} - \mathbf{r}')] d\mathbf{r}' d\mathbf{r}$$

-26-

where $\rho(\mathbf{r})$ is the spin density defined by the microstructure. In the long timescale limit, the diffusing spins sample the whole of the available space and $P(\mathbf{r}; \mathbf{r}', t)$ becomes independent of \mathbf{r} such that

$$P(\mathbf{r}; \mathbf{r}', \infty) = \rho(\mathbf{r}') = \rho(\mathbf{r})$$

then

$$E(G_D, \delta, \infty) = \left| \int_{\mathbf{r}} \rho(\mathbf{r}) \exp[i\gamma\delta(G_D \cdot \mathbf{r})] d\mathbf{r} \right|^2 = |S(\mathbf{q})|^2$$

where $\mathbf{q} = (2\pi)^{-1} \gamma \delta G_D$ and $S(\mathbf{q})$ is a structure factor, defined by the above expression with direct analogies in optics and neutron scattering. Measurement of echo attenuation and hence $S(\mathbf{q})$ and calculations of microstructure have been reported for both model and real systems including porous media and emulsions.

As an example figure B1.14.13 shows the droplet size distribution of oil drops in the cream layer of a decane-in-water emulsion as determined by PFG [45]. Each curve represents the distribution at a different height in the cream with large drops at the top of the cream. The inset shows the PFG echo decay trains as a function of

height and the curves to which the data were fitted using a Stokes-velocity-based model of the creaming process.

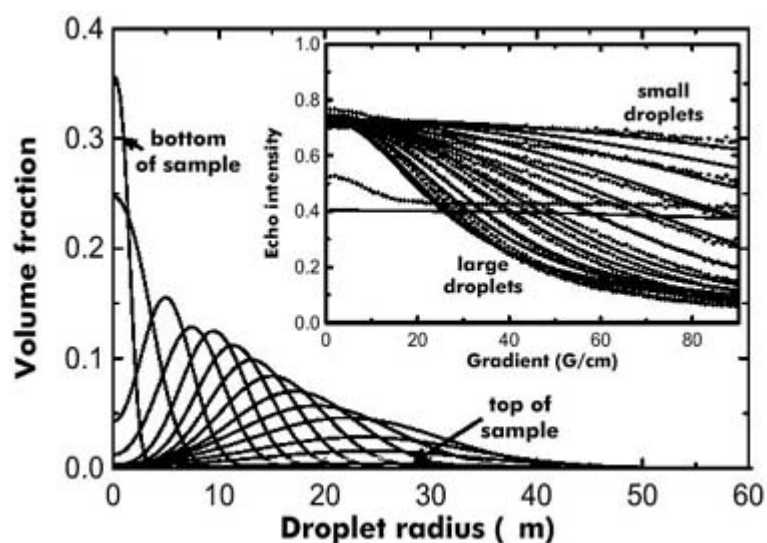


Figure B1.14.13. Derivation of the droplet size distribution in a cream layer of a decane/water emulsion from PGSE data. The inset shows the signal attenuation as a function of the gradient strength for diffusion weighting recorded at each position (top trace = bottom of cream). A Stokes-based velocity model (solid lines) was fitted to the experimental data (solid circles). The curious horizontal trace in the centre of the plot is due to partial volume filling at the water/cream interface. The droplet size distribution of the emulsion was calculated as a function of height from these NMR data. The most intense narrowest distribution occurs at the base of the cream and the curves proceed logically up through the cream in steps of 0.041 cm. It is concluded from these data that the biggest droplets are found at the top and the smallest at the bottom of the cream. (From [45].)

B1.14.4.5 PLANAR TAGGING

A more qualitative means of visualising flows is by ‘multi-plane tagging’ [46, 47]. The principle of tagging experiments is to excite magnetic resonance only in stripes across the image plane and to observe the spatial evolution of these stripes with time (figure B1.14.14). The excitation can be achieved using a comb of radio-frequency pulses each of narrow flip angle applied to a sample in a magnetic field gradient (figure B1.14.15 (a)) [48]. The Fourier transform of this excitation shows distinct maxima occurring at frequency intervals given by the reciprocal of the pulse spacing. The overall excitation bandwidth—which must be sufficient to cover the frequency spread of all nuclei in the sample—is determined by the individual pulse widths, and the sharpness of the maxima by the number of pulses. This frequency response is illustrated in figure B1.14.15(b) from which it is clear how the excitation is achieved. The action of the comb can be understood more qualitatively as follows. Each pulse tips all spins by a small angle. Between the pulses the spins precess by an amount dependent on their position in the gradient. Only at positions where the precessional phases between pulses is equal to $2n\pi$ do the pulses have a cumulative effect in tipping the magnetization through a large angle.

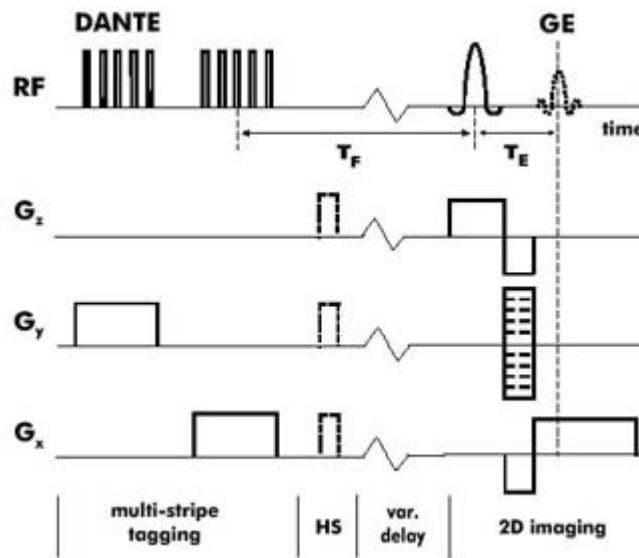


Figure B1.14.14. Pulse sequence for multi-plane tagging. A magnetization saturation grid is prepared in the multi-plane tagging section. After a certain time of flight T_F this grid is imaged using a standard imaging pulse sequence. The motion of tagged spins is visualized by displacements of the grid lines.

-28-

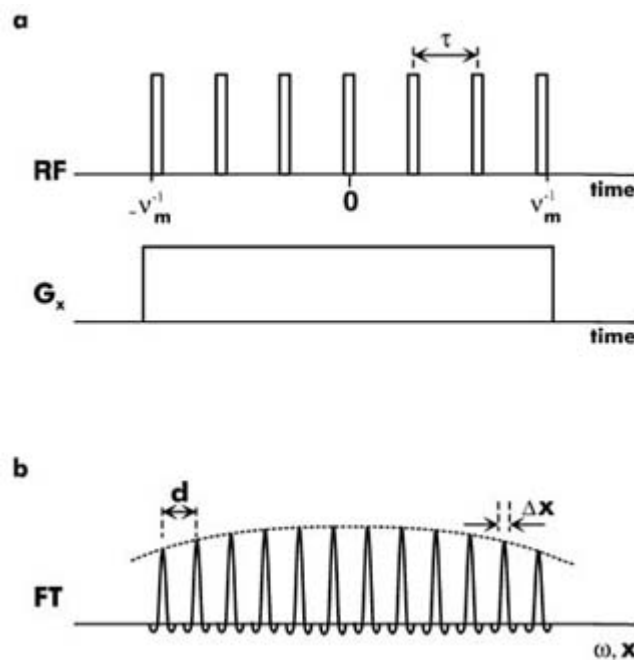


Figure B1.14.15. Preparation of a magnetization grid by means of DANTE pulse trains. The pulse sequence for one-dimensional tagging is shown in (a). As depicted in (b) the spectrum of a pulse train consists of a comb of peaks. To tag spins in certain regions of the sample the DANTE pulse train is applied while a magnetic field gradient G_x is switched. The magnetization is then saturated in planes which are located at equidistant positions corresponding to the spectral peaks of the DANTE pulse train and which are orthogonal to the spatial coordinate $x = \omega / (\gamma G_x)$. A magnetization saturation grid is prepared by subsequently using the sequence (a) with magnetic field gradients in the two directions of the imaging plane coordinates. Saturated spins (bearing transverse magnetization) are visualized as black grid lines in an image. These lines of thickness $\Delta x = 2\pi v_m / (\gamma G_x)$ are separated by $d = 2\pi / (\gamma G_x \tau)$ where τ is the delay from one RF pulse to the next.

Following excitation of this kind, a standard spin warp imaging protocol can be used to create the image. By varying the delay between excitation and image acquisition, flow is visualized by different degrees of distortions of the saturated magnetization grid. An example of a rotating cylinder filled with water–oil mixture is shown in [figure B1.14.16](#) [49].

-29-

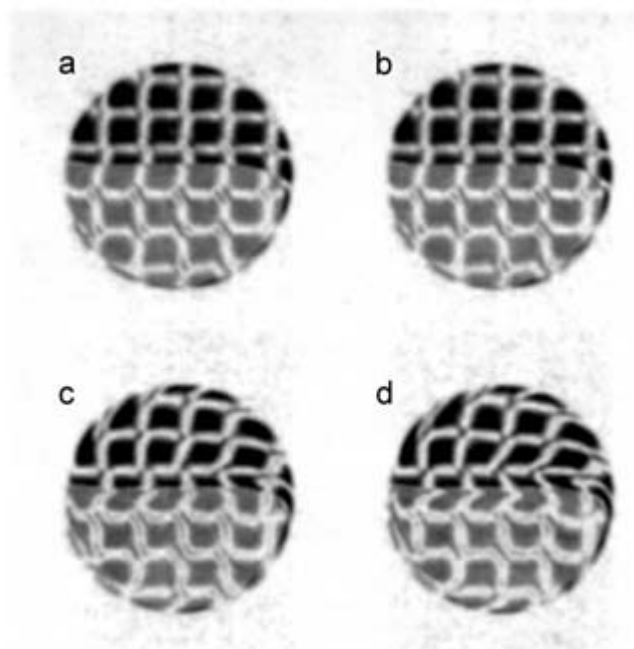


Figure B1.14.16. Multi-plane tagging experiment for a 1:1 water (bottom) and oil (top) mixture in the cylinder, which rotates clockwise. The rotation rate was 0.5 rev s^{-1} , and the tagging delay times T_F are (a) 1 ms, (b) 25 ms, (c) 50 ms and (d) 100 ms. The interface between the fluids is clearly shown. The misalignment in the horizontal direction at the interface is caused by chemical shift between water and oil. Flow was mainly detected in a thin layer near the cylinder and a layer along the water–oil interface in the centre that flows to the right. (From [49].)

REFERENCES

- [1] Edelstein W A, Hutchison J M S, Johnson G and Redpath T W 1980 Spin warp NMR imaging and applications to human whole-body imaging *±B* **25** 751–6
 - [2] Johnson G, Hutchison J M S, Redpath T W and Eastwood L M 1983 Improvements in performance time for simultaneous three-dimensional NMR imaging *J. Magn. Reson.* **54** 374–84
 - [3] Morris P G 1986 *Nuclear Magnetic Resonance Imaging in Medicine and Biology* (Oxford: Clarendon)
 - [4] Bailes D R and Bryant D J 1984 NMR Imaging *Contemp. Phys.* **25** 441–75
 - [5] Mansfield P 1977 Multi-planar image formation using NMR spin echoes *J. Phys. C: Solid State Phys.* **10** L55–L58
 - [6] Mansfield P and Morris P G 1982 NMR imaging in biomedicine (New York: Academic)
 - [7] Stehling M K, Turner R and Mansfield P 1991 Echo-planar imaging: magnetic-resonance-imaging in a fraction of a second *Science* **254** 43–50
-

- [8] Mansfield P and Pykett I L 1978 Biological and medical imaging by NMR *J. Magn. Reson.* **29** 355–73
- [9] Lauterbur P C 1973 Image formation by induced local interactions: examples employing nuclear magnetic resonance *Nature* **242** 190–1
- [10] Lauterbur P C 1974 Magnetic resonance zeugmatography *Pure Appl. Chem.* **40** 149–57
- [11] Smith M E and Strange J H 1996 NMR techniques in material physics: a review *Meas. Sci. Technol.* **7** 449–75
- [12] Hafner S, Demco D E and Kimmich R 1996 Magic echoes and NMR imaging of solids *Solid State Nucl. Magn. Reson.* **6** 275–93
- [13] McDonald P J 1997 Stray field magnetic resonance imaging *Prog. Nucl. Magn. Reson. Spectrosc.* **30** 69–99
- [14] Callaghan P T 1993 *Principles of Nuclear Magnetic Resonance Microscopy* (Oxford: Clarendon)
- [15] Ahn C B and Cho Z H 1989 A generalized formulation of diffusion effects in μm resolution nuclear magnetic-resonance imaging *Med. Phys.* **16** 22–8
- [16] Tzalmona A, Armstrong R L, Menzinger M, Cross A and Lemaire C 1990 Detection of chemical waves by magnetic resonance imaging *Chem. Phys. Lett.* **174** 199–202
- [17] Wehrli F W, MacFall J R, Shutts D, Breyer R and Herfkens R J 1984 Mechanisms of contrast in NMR imaging *J. Comput. Assist. Tomogr.* **8** 369–80
- [18] Osment P A, Packer K J, Taylor M J, Attard J J, Carpenter T A, Hall L D, Harrod N J and Doran S J 1990 NMR imaging of fluids in porous solids *Phil. Trans. R. Soc A* **333** 441–52
- [19] Attard J J, Carpenter T A, Hall L D, Davies S, Taylor M J and Packer K J 1991 Spatially resolved T_1 relaxation measurements in reservoir cores *Magn. Reson. Imaging* **9** 815–19
- [20] Balcom B J, Fischer A E, Carpenter T A and Hall L D 1993 Diffusion in aqueous gels—mutual diffusion-coefficients measured by one-dimensional nuclear-magnetic-resonance imaging *J. Am. Chem. Soc.* **115** 3300–305
- [21] Attard J J, Doran S J, Herrod N J, Carpenter T A and Hall L D 1992 Quantitative NMR spin-lattice-relaxation imaging of brine in sandstone reservoir cores *J. Magn. Reson.* **96** 514–25
- [22] Haase A and Frahm J 1985 Multiple chemical-shift-selective NMR imaging using stimulated echoes *J. Magn. Reson.* **64** 94–102
- [23] Maudsley A A, Hilal S K, Perman W H and Simon H E 1983 Spatially resolved high-resolution spectroscopy by 4-dimensional NMR *J. Magn. Reson.* **51** 147–52
- [24] Kunze C and Kimmich R 1994 Proton-detected ^{13}C imaging using cyclic-J cross polarization *Magn. Reson. Imaging* **12** 805–10
- [25] Heidenreich M, Spyros A, Köckenberger W, Chandrakumar N, Bowtell R and Kimmich R 1998 CYCLCROP mapping of ^{13}C labelled compounds: perspectives in polymer sciences and plant physiology *Spatially Resolved Magnetic Resonance, Proc. 4th Int. Conf. on Magnetic Resonance Microscopy and Macroscopy* ed P Blümli, B Blümich, R E Botto and E Fukushima (Weinheim: Wiley-VCH) pp 21–52
- [26] Hahn E L 1950 Spin Echoes *Phys. Rev.* **80** 580–94
- [27] Carr H Y and Purcell E M 1954 Effects of diffusion on free precession in nuclear magnetic resonance experiments *Phys. Rev.* **94** 630–8
-

- [28] Moran P R 1982 A flow velocity zeugmatographic interlace for NMR imaging in humans *Magn. Reson. Imaging* **1** 197–203
- [29] Redpath T W, Norris D G, Jones R A and Hutchison J M S 1984 A new method of NMR flow imaging \pm B **29** 891–5
- [30] Bryant D J, Payne J A, Firmin D N and Longmore D B 1984 Measurement of flow with NMR imaging using a gradient pulse and phase difference technique *J. Comput. Assist. Tomogr.* **8** 588–93
- [31] Rombach K, Laukemper-Ostendorf and Blümmler P 1998 Applications of NMR flow imaging in materials science *Spatially Resolved Magnetic Resonance, Proc. 4th Int. Conf. on Magnetic Resonance Microscopy and Macroscopy* ed P Blümmler, B Blümich, R E Botto and E Fukushima (Weinheim: Wiley–VCH) pp 517–29
- [32] Goerke U and Kimmich R 1998 NMR-imaging techniques for quantitative characterization of periodic motions: 'incoherent averaging' and 'spectral side band analysis' *Spatially Resolved Magnetic Resonance, Proc. 4th Int. Conf. on Magnetic Resonance Microscopy and Macroscopy* ed P Blümmler, B Blümich, R E Botto and E Fukushima (Weinheim: Wiley–VCH) pp 499–506
- [33] Le Bihan D, Breton E, Lallemand D, Aubin M-L, Vignaud J and Laval-Jeantet M 1988 Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging 1988 *Radiology* **168** 497–505
- [34] Stepisnik J 1981 Analysis of NMR self-diffusion measurements by a density-matrix calculation *Physica B/C* **104** 350–64
- [35] Stepisnik J 1985 Measuring and imaging of flow by NMR *Prog. Nucl. Magn. Reson. Spectrosc.* **17** 187–209
- [36] Kärger J, Pfeifer H and Heink W 1988 Principles and application of self-diffusion measurements by nuclear magnetic resonance *Adv. Magn. Res.* **12** 1–89
- [37] Taylor D G and Bushell M C 1985 The spatial-mapping of translational diffusion-coefficients by the NMR imaging technique \pm B **30** 345–9
- [38] Basser P J, Mattiello J and LeBihan D 1994 MR diffusion tensor spectroscopy and imaging *Biophys. J.* **66** 259–67
- [39] Williams W D, Seymour E F W and Cotts R M 1978 A pulsed-gradient multiple-spin-echo NMR technique for measuring diffusion in the presence of background magnetic field gradients 1978 *J. Magn. Reson.* **31** 271–82
- [40] Karlicek R F Jr and Lowe I J 1980 A modified pulsed gradient technique for measuring diffusion in the presence of large background gradients *J. Magn. Reson.* **37** 75–91
- [41] Lucas A J, Gibbs S J, Jones E W G, Peyron M, Derbyshire J A and Hall L D 1993 Diffusion imaging in the presence of static magnetic-field gradients *J. Magn. Reson. A* **104** 273–82
- [42] Lian J, Williams D S and Lowe I J 1994 Magnetic resonance imaging of diffusion in the presence of background gradients and imaging of background gradients *J. Magn. Reson. A* **106** 65–74
- [43] Tanner J E 1970 Use of the stimulated echo in NMR diffusion studies *J. Chem. Phys.* **52** 2523–6
- [44] Callaghan P T, Eccles C D and Xia Y 1988 NMR microscopy of dynamic displacements— k -space and q -space imaging *J. Phys. E: Sci. Instrum.* **21** 820–2
- [45] McDonald P J, Ciampi E, Keddie J L, Heidenreich M and Kimmich R, Magnetic resonance determination of the spatial dependence of the droplet size distribution in the cream layer of oil-in-water emulsions: evidence for the effects of depletion flocculation *Phys. Rev. E*, submitted

- [46] Axel L and Dougherty L 1989 Heart wall motion—improved method of spatial modulation of magnetization for MR imaging *Radiology* **172** 349–50
- [47] Zerhouni E A, Parrish D M, Rodgers W J, Yang A and Shapiro E P 1988 Human-heart-tagging with MR imaging—

a method for noninvasive assessment of myocardial motion *Radiology* **169** 59–63

- [48] Morris G A and Freeman R 1978 Selective excitation in Fourier transform nuclear magnetic resonance *J. Magn. Reson.* **29** 433–62
- [49] Jeong E K, Altobelli S A and Fukushima E 1994 NMR imaging studies of stratified flows in a horizontal rotating cylinder *Phys. Fluids* **6** 2901–6

FURTHER READING

Blümich B and Kuhn W (eds) 1992 *Magnetic Resonance Microscopy Methods and Applications in Materials Science, Agriculture and Biomedicine* (Weinheim: Wiley–VCH)

Blümler P, Blümich B, Botto R E and Fukushima E (eds) 1998 *Spatially Resolved Magnetic Resonance*, Proc. 4th Int. Conf. on *Magnetic Resonance Microscopy and Macroscopy* (Weinheim: Wiley–VCH)

These two references give an excellent overview over the most recent examples in this research field.

Callaghan P T 1993 *Principles of Nuclear Magnetic Resonance Microscopy* (Oxford: Clarendon)

Kimmich R 1997 *NMR—Tomography, Diffusometry, Relaxometry* (Berlin: Springer)

Two standard text books which are recommended to researchers and graduate students who seek deeper insight into methodology and theory.

-1-

B1.15 EPR methods

Stefan Weber

B1.15.1 INTRODUCTION

Systems containing unpaired electron spins, such as free radicals, biradicals, triplet states, most transition metal and rare-earth ions and some point defects in solids form the playground for electron paramagnetic resonance, EPR, also called electron spin resonance, ESR, or electron magnetic resonance, EMR. The fundamentals of EPR spectroscopy are very similar to the more familiar nuclear magnetic resonance (NMR) technique. Both deal with interactions of electromagnetic radiation with magnetic moments, which in the case of EPR arise from electrons rather than nuclei. With few exceptions, unpaired electrons lead to a non-vanishing spin of a particle that can be used as a spectroscopic probe. In EPR spectroscopy such molecules are studied by observing the magnetic fields at which they come into resonance with monochromatic electromagnetic radiation. Since species with unpaired electron spins are relatively rare compared to the multitude of species with magnetic nuclei, EPR is less widely applicable than NMR or even optical spectroscopy which has clear advantages with its ability to detect diamagnetic as well as paramagnetic states. What appears to be a drawback, however, can turn into an invaluable advantage, for instance, when selectively studying paramagnetic ions or molecules buried in a large protein environment. With its inherent specificity for those reactants, intermediates or products that carry unpaired electron spins, together with its high spectral resolution, EPR has excelled over many other techniques in, for example, unravelling the primary events of photosynthesis. Similarly, many key intermediates in this process have been identified by

EPR. By appending a paramagnetic fragment—a so-called ‘spin-label’—to a molecule of biological importance, in effect one has acquired a probe to supply data on the interactions and dynamics of biological molecules. Very many systems of biomedical interest have had their structure and function elucidated by application of modern EPR techniques. Also EPR has allowed chemists to probe into the details of reaction mechanisms by using the technique of spin trapping to identify reactive radical intermediates. As one last example of the many successes of EPR the identification of paramagnetic species in insulators and semiconductors is worth mentioning.

More than 50 years after its invention by the Russian physicist Zavoisky (for a review of the EPR history see [1]), advanced EPR techniques presently applied in the above mentioned areas of physics, chemistry and biology include time-resolved continuous wave (CW) and pulsed EPR (Fourier transform (FT) EPR and electron-spin echo (ESE) detected EPR) at various microwave (MW) frequencies and multiple-resonance EPR methods such as electron–nuclear double resonance (ENDOR) and electron–nuclear–nuclear TRIPLE resonance in the case of electron and nuclear transitions and electron–electron double resonance (ELDOR) in the case of different electron spin transitions. High-field/high-frequency EPR and ENDOR have left the developmental stage, and a wide range of significant applications continues to emerge. The range of multi-frequency EPR spectroscopy is now extending from radiofrequencies (RFs) in near-zero fields up to several hundred gigahertz in superconducting magnets or Bitter magnets.

In this article only the most important and frequently applied EPR methods will be introduced. For more extensive treatments of CW and pulsed EPR the reader is referred to some excellent review articles that will be specified in the respective sections of this article. A good starting point for further reading is provided by a number of outstanding textbooks which have been written on the various aspects of EPR in general [2, 3, 4, 5, 6, 7 and 8]. Interested readers

-2-

might also appreciate the numerous essays on various magnetic resonance topics that are published on a bimonthly basis in the educational journal ‘Concepts in Magnetic Resonance’.

B1.15.2 EPR BACKGROUND

B1.15.2.1 SPINS AND MAGNETIC MOMENTS

Historically, the recognition of electron spin can be traced back to the famous Stern–Gerlach experiment in the early 1920s. Stern and Gerlach observed that a beam of silver atoms was split into two components deflected in different directions when passing through an inhomogeneous magnetic field. The observation could only be explained with the concept of a half-integral angular momentum ascribed to an intrinsic spin of the electron. EPR spectroscopy relies on the behaviour of the electron angular momentum and its associated magnetic moment in an applied magnetic field.

If the angular momentum of a free electron is represented by a spin vector $\mathbf{S}=(S_x, S_y, S_z)$, the magnetic moment μ_S is related to \mathbf{S} by

$$(B1.15.1)$$

where g_e is a dimensionless number called the electron g -factor and $\beta_e = |e|/(2m_e) = 9.2740154 \times 10^{-24} \text{ J T}^{-1}$ is the Bohr magneton; e is the electronic charge, $h/(2\pi)$ is Planck’s constant and m_e is the electron mass. The negative sign in equation (b1.15.1) indicates that, because of the negative charge of the electron, the magnetic moment vector is antiparallel to the spin (since $g_e < 0$). In the quantum theory μ_S and \mathbf{S} are treated as (vector)

operators. Suppose that the angular momentum operator \mathbf{S} is defined in units of \hbar , then \mathbf{S}^2 has the eigenvalues $S(S+1)$, where S is either integer or half integer. The magnitude of the angular momentum itself is given by the square root of the eigenvalue of \mathbf{S}^2 , which is $\hbar\sqrt{S(S+1)}$. Any component of \mathbf{S} (for example S_z) commutes with \mathbf{S}^2 , so that simultaneously eigenvalues of both \mathbf{S}^2 and S_z may be specified, which are $S(S+1)$ and M_S , respectively. M_S has $(2S+1)$ allowed values running in integral steps from $-S$ to $+S$.

Classically, the interaction energy of a magnetic moment μ_S in an applied magnetic field \mathbf{B} is

$$(B1.15.2)$$

For a quantum mechanical system μ_S is replaced by the appropriate operator, equation (b1.15.1) to obtain the Hamiltonian for a free electron in a magnetic field,

$$(B1.15.3)$$

If the magnetic field is B_0 in the z -direction, $\mathbf{B} = (0,0,B_0)$, the scalar product simplifies and the Hamiltonian becomes

$$(B1.15.4)$$

-3-

The eigenvalues of this Hamiltonian are simple, being only multiples $g_e\beta_e B_0$ of the eigenvalues of S_z . Therefore, the allowed energies are $E_{M_S} = g_e\beta_e M_S B_0$. For a simple system of one unpaired electron, $S = \frac{1}{2}$ and $M_S = \pm\frac{1}{2}$, which results in two energy states which are degenerate in zero field and whose energy separation increases linearly with B_0 . This is summarized in figure B1.15.1 where the two states are also labelled with their eigenfunctions $|+\frac{1}{2}\rangle \equiv |\alpha\rangle$ and $|-\frac{1}{2}\rangle \equiv |\beta\rangle$ to indicate the $M_S = +\frac{1}{2}$ and $M_S = -\frac{1}{2}$ eigenstates for $S = \frac{1}{2}$, respectively. The lowest state has $M_S = -\frac{1}{2}$ (since $g_e < 0$), so that the projection of \mathbf{S} along the z -axis, S_z , is antiparallel to the field, but in accordance with physical expectation the z -component of the magnetic moment is parallel to the field (see equation (b1.15.1)). The splitting of the electron spin energy levels by a magnetic field is referred to as the Zeeman effect.

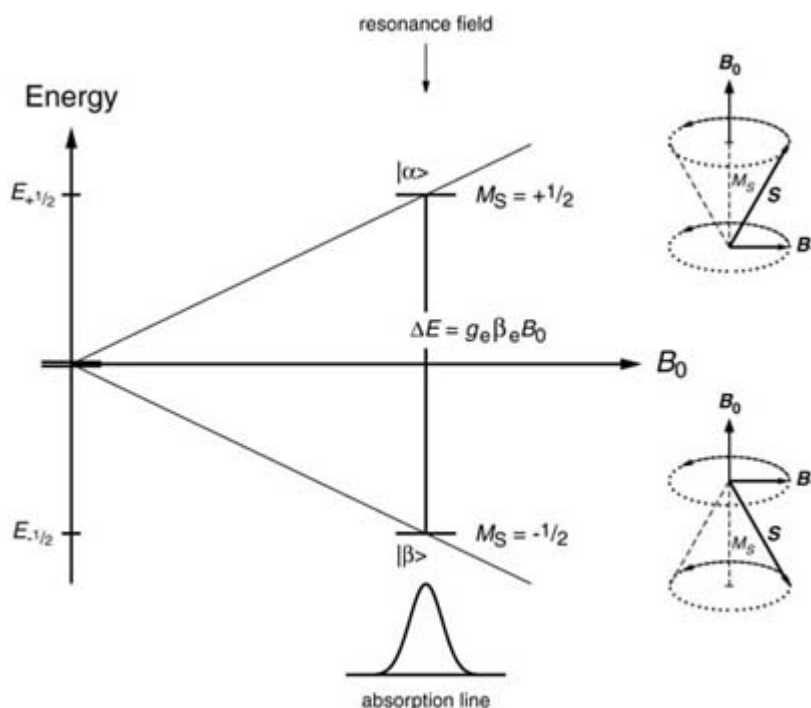


Figure B1.15.1. Energy levels for an electron ($S = \frac{1}{2}$) as a function of the applied magnetic field B_0 . $E_{1/2} = +\frac{1}{2} g_e \beta_e B_0$ and $E_{-1/2} = -\frac{1}{2} g_e \beta_e B_0$ represent the energies of the $M_S = +\frac{1}{2}$ and $M_S = -\frac{1}{2}$ states, respectively.

Under the influence of the external magnetic field \mathbf{B} , the spin \mathbf{S} and the magnetic moment μ_S perform a precessional motion about the axis pointing along the direction of \mathbf{B} . In the absence of additional magnetic fields, the angle between \mathbf{S} and \mathbf{B} does not change and the motion of the spin about the \mathbf{B} axis may be illustrated as cones, as shown on the right-hand side of figure B1.15.1 for the two possible orientations of \mathbf{S} . The z -component of the magnetic moment is sharply defined, while the x - and y -components are not, since they oscillate in the xy -plane. The precession frequency of μ_S and \mathbf{S} about \mathbf{B} is called the Larmor frequency, $\omega_0 = g_e \beta_e B_0 / \hbar$.

-4-

The transition between the two eigenstates can be induced by the application of microwave (MW) radiation with frequency ω and its magnetic field vector linearly polarized in the xy -plane. The oscillating \mathbf{B}_1 -field of this radiation can be formally decomposed into two counter-rotating circularly polarized components, one of which is rotating in the same direction as the Larmor precession of the spins. The other component of the MW field does not interact significantly with the electron spins and can be neglected. If ω is different from ω_0 the precessing magnetic moment will not seriously be affected by \mathbf{B}_1 , for its component in the xy -plane will pass in and out of phase with \mathbf{B}_1 and there will be no resultant interaction. Transitions occur only near the resonance condition of

$$\omega = \omega_0 = g_e \beta_e B_0 / \hbar. \quad (\text{B1.15.5})$$

On exact resonance, μ_S and \mathbf{B}_1 can remain in phase so the precessing magnetic moment experiences a constant field \mathbf{B}_1 in the xy -plane (see figure B1.15.1). It will respond to this by precessing about it with frequency $\omega_1 = g_e \beta_e B_1 / \hbar$.

In the language of quantum mechanics, the time-dependent \mathbf{B}_1 -field provides a perturbation with a nonvanishing matrix element joining the stationary states $|\alpha\rangle$ and $|\beta\rangle$. If the rotating field is written in terms of an amplitude B_1 a perturbing term in the Hamiltonian is obtained

$$\mathcal{H}' = g_e \beta_e B_1 (S_x \cos(\omega t) + S_y \sin(\omega t)). \quad (\text{B1.15.6})$$

The operators S_x and S_y have matrix elements between states $|\alpha\rangle$ and $|\beta\rangle$ or, in general, between states $|M_S\rangle$ and $|M_S \pm 1\rangle$ and consequently induce transitions between levels adjacent in energy.

If $B_1 \ll B_0$ first-order perturbation theory can be employed to calculate the transition rate for EPR (at resonance)

$$W_{\beta \leftarrow \alpha} = \pi \frac{\omega_1^2}{2} \rho(\omega). \quad (\text{B1.15.7})$$

In equation (b1.15.7) $\rho(\omega)$ is the frequency distribution of the MW radiation. This result obtained with explicit evaluation of the transition matrix elements occurring for simple EPR is just a special case of a much more general result, Fermi's golden rule, which is the basis for the calculation of transition rates in general:

$$(\text{B1.15.8})$$

$$W_{M_S \leftarrow M_S+1} = 2\pi |\langle M_S | \mathcal{H}' | M_S + 1 \rangle|^2 \rho(\omega).$$

Using the selection rule for allowed transitions the relative intensity for the transition from the state $|M_S\rangle$ to $|M_S+1\rangle$ is given by

$$W_{M_S \leftarrow M_S+1} \propto (S(S+1) - M_S(M_S+1)). \quad (\text{B1.15.9})$$

-5-

The transition between levels coupled by the oscillating magnetic field \mathbf{B}_1 corresponds to the absorption of the energy required to reorient the electron magnetic moment in a magnetic field. EPR measurements are a study of the transitions between electronic Zeeman levels with $\Delta M_S = \pm 1$ (the selection rule for EPR).

The g -factor for a free electron, $g_e = 2.002\,319\,304\,386(20)$, is one of the most accurately known physical constants. For a magnetic field of 1 T ($= 10^4$ G) the resonance frequency $\omega/(2\pi)$ is 28.024 945 GHz, approximately three orders of magnitude larger than is required for any nuclear resonance (because $\beta_e/\beta_N \approx 1836$). This corresponds to a wavelength of about 10 mm and is in the microwave (MW) region of the electromagnetic spectrum.

B1.15.2.2 THERMAL EQUILIBRIUM, MAGNETIC RELAXATION AND LORENTZIAN LINESHAPE

Application of an oscillating magnetic field at the resonance frequency induces transitions in both directions between the two levels of the spin system. The rate of the induced transitions depends on the MW power which is proportional to the square of $\omega_1 = \gamma_e B_1$ (the amplitude of the oscillating magnetic field) (see [equation \(b1.15.7\)](#)) and also depends on the number of spins in each level. Since the probabilities of upward ($|\beta\rangle \rightarrow |\alpha\rangle$) and downward ($|\alpha\rangle \rightarrow |\beta\rangle$) transitions are equal, resonance absorption can only be detected when there is a population difference between the two spin levels. This is the case at thermal equilibrium where there is a slight excess of spins in the energetically lower $|\beta\rangle$ -state. The relative population of the two-level system in thermal equilibrium is given by the Boltzmann distribution

$$\frac{N_\alpha}{N_\beta} = \exp\left(-\frac{\Delta E}{k_B T}\right) = \exp\left(-\frac{g_e \beta_e B_0}{k_B T}\right) \quad (\text{B1.15.10})$$

where N_α and N_β are the populations of the upper and lower spin states, respectively, ΔE is the energy difference separating the states, k_B is the Boltzmann constant and T is the temperature in Kelvin. The total number of spins is, of course, $N = N_\alpha + N_\beta$.

Computation of the fractional excess of the lower level,

$$\frac{N_\beta - N_\alpha}{N} = \frac{1 - \exp(-g_e \beta_e B_0 / (k_B T))}{1 + \exp(-g_e \beta_e B_0 / (k_B T))} \quad (\text{B1.15.11})$$

yields, for electrons in a magnetic field of 0.3 T at 300 K, a value of 7.6×10^{-4} , while for protons under the same conditions the value is only 1.2×10^{-6} . Thus, at thermal equilibrium, in EPR experiments one can virtually always take any nuclear spin state belonging to the same electron spin state to be equally populated. Because of the slightly larger number of spins occupying the lower energy level, there will be a net absorption of energy which results in an exponential decay of the initial population difference of the spin states. Eventually the levels would be equally populated (the spin system is then said to be saturated) if there were no

radiationless processes that restored the thermal equilibrium distribution of the population by dissipating the energy absorbed by the spin system to other degrees of freedom. These nonradiative transitions between the two states $|\alpha\rangle$ and $|\beta\rangle$ are called spin–lattice relaxation. Spin–lattice relaxation is possible because the spin system is coupled to fluctuating magnetic fields driven by the thermal motions of the surroundings which are at thermal equilibrium. These fluctuations can stimulate spin flips and,

-6-

therefore, this process leads to unequal probabilities of spontaneous transitions $|\alpha\rangle \rightarrow |\beta\rangle$ and $|\beta\rangle \rightarrow |\alpha\rangle$ and unequal populations at thermal equilibrium. In a magnetic resonance experiment one always has a competition between spin–lattice relaxation and the radiation field whose nature is to equalize the population of the levels. Qualitatively, T_1 is the time for the population difference to decay to 1/e of its equilibrium value after the perturbation (which in the case of magnetic resonance is the radiation field) is removed.

A second type of relaxation mechanism, the spin–spin relaxation, will cause a decay of the phase coherence of the spin motion introduced by the coherent excitation of the spins by the MW radiation. The mechanism involves slight perturbations of the Larmor frequency by stochastically fluctuating magnetic dipoles, for example those arising from nearby magnetic nuclei. Due to the randomization of spin directions and the concomitant loss of phase coherence, the spin system approaches a state of maximum entropy. The spin–spin relaxation disturbing the phase coherence is characterized by T_2 .

A result of the relaxation processes is a shortened lifetime of the spin states giving rise to a broadening of the EPR line, which for most magnetic resonance lines dominated by homogeneous linewidth can be written as

$$f(\omega) = \frac{A\gamma M_0 T_2}{1 + \gamma^2 B_1^2 T_1 T_2 + T_2^2 (\omega_0 - \omega)^2}. \quad (\text{B1.15.12})$$

In equation (b1.15.12), M_0 is the z-component of the bulk magnetization vector, $\mathbf{M} = (1/V) \sum_i^N \boldsymbol{\mu}_i$ (unit $\text{J T}^{-1} \text{m}^{-3}$), for an ensemble of N spin magnetic moments at thermal equilibrium (in the absence of any resonant radiation), or in other words the net magnetic moment per unit volume, $\gamma = g_e \beta_e / \hbar$ is the gyromagnetic ratio and A is a proportionality constant to include instrumental factors. The lineshape function $f(\omega)$ has a maximum at $\omega = \omega_0$ and it decreases for high power levels (i.e. for large B_1) and when the spin–lattice relaxation is not fast enough to maintain the population difference. This decrease is called saturation. If the saturation factor s is defined by

$$s = \frac{1}{1 + \gamma^2 B_1^2 T_1 T_2} \quad (\text{B1.15.13})$$

then $f(\omega)$ has the form

$$f(\omega) = \frac{As\gamma M_0 T_2}{1 + sT_2^2 (\omega_0 - \omega)^2}. \quad (\text{B1.15.14})$$

Well below saturation $s \approx 1$, and so the lineshape function becomes

$$f(\omega) = \frac{A\gamma M_0 T_2}{1 + T_2^2 (\omega_0 - \omega)^2}. \quad (\text{B1.15.15})$$

This is the famous Lorentzian function which is very often found for spectra of radicals in solution. In order to determine the relaxation times T_1 and T_2 , a series of EPR spectra is recorded with the MW power varying from a condition of negligible saturation ($B_1^2 \gamma^2 T_1 T_2 \ll 1; s \approx 1$) to one of pronounced saturation ($B_1^2 \gamma^2 T_1 T_2 > 1; s < 1$). T_2 is then calculated from the linewidth below saturation by means of the expression

$$T_2 = \frac{2}{\Delta\omega_{1/2}^0} \quad (\text{B1.15.16})$$

where $\Delta\omega_{1/2}^0$ is the half width at half height of the magnetic resonance absorption line in the limit $B_1 \rightarrow 0$ ($s \rightarrow 1$). For T_1 one obtains

$$T_1 = \frac{\Delta\omega_{1/2}^0}{2} \left(\frac{1/s - 1}{\omega_1^2} \right). \quad (\text{B1.15.17})$$

One of the principal experimental advantages of this method of determining relaxation times is that it may be carried out with standard EPR spectrometers using CW-detected EPR lines [9, 10]. A discussion of more direct measurements of T_1 and T_2 using time-resolved EPR techniques is deferred to a later point (see sections B1.15.4 and B1.15.6.3(b)).

B1.15.2.3 SPIN HAMILTONIAN

To characterize and interpret EPR spectra one needs to obtain transition frequencies and transition probabilities between the $(2S + 1)$ spin states. All interactions of the spins of electrons and nuclei with the applied magnetic field and with each other that lead to energy differences between states with different angular momenta have to be considered. The interactions are expressed in terms of operators representing the spins, with various coupling coefficients for the different interactions. The contributions of all these interactions make up the spin Hamiltonian that will be given in energy units throughout this text. Since, in principle, the spin Hamiltonian has no effect on the spatial part of the electronic wavefunction, the energy of the spin system in a certain state characterized by the quantum numbers M_S and M_I can be derived from the time-independent Schrödinger equation. The EPR spectrum is then interpreted as the allowed transitions between the eigenvalues of the spin Hamiltonian.

(A) ELECTRON ZEEMAN INTERACTION

The first contribution considered here is the electron Zeeman interaction, i.e. the coupling of the magnetic dipole moment of the electron spin to the external magnetic field. For symmetry reasons the electron Zeeman interaction is isotropic for a free electron spin and is characterized by the Zeeman splitting constant g_e . The g -value of an unpaired electron in an atomic or molecular environment is very often different from g_e and may also be anisotropic, i.e. dependent on the orientation of the system relative to the magnetic field \mathbf{B} . The deviation from the spin-only value of the g -factor and the anisotropy result from the contribution of the orbital angular momentum to the total angular momentum of the electron. This phenomenon is called spin-orbit coupling. It leads to an anisotropic electron Zeeman interaction (EZI) which is usually formulated as

$$\mathcal{H}_{\text{EZI}} = \beta_e \mathbf{B} g \mathbf{S} \quad (\text{B1.15.18})$$

where the field and angular momentum vectors are coupled through a symmetric matrix \mathbf{g} of dimension 3×3 . In organic radicals orbital momenta are almost completely quenched by chemical bonding (with the exception of cases where the energies of the two orbitals are nearly degenerate), leading to only small deviations $\Delta g = |g_{ii} - g_e|$, $i=X, Y, Z$ of the principal values of \mathbf{g} from the free-electron value g_e (typically in a range from 10^{-5} to 5×10^{-2}). g -values very different from g_e are expected for first-row transition metal ions and for rare-earth ions where spin-orbit coupling is more complete. The \mathbf{g} -matrix of organic radicals reflects certain features of the electronic wavefunction of the paramagnetic species. The spatial distribution of the orbital carrying the unpaired spin can be influenced by interactions with other molecules, e.g. via hydrogen bonding. Therefore, a determination of the g -values and the orientation of the main axes of \mathbf{g} with respect to the molecular axis frame can give highly specific information on the interaction of the molecule with its surrounding.

(B) ELECTRON SPIN-SPIN INTERACTION

An atom or a molecule with the total spin of the electrons $S = 1$ is said to be in a triplet state. The multiplicity of such a state is $(2S+1)=3$. Triplet systems occur in both excited and ground state molecules, in some compounds containing transition metal ions, in radical pair systems, and in some defects in solids.

For a system with $S = 1$, there are three sublevels characterized by $M_S = \pm 1$ and $M_S = 0$. In contrast to systems with $S = \frac{1}{2}$, these sublevels may not be degenerate in the absence of an external magnetic field (see [figure B1.15.2](#)). The lifting of degeneracy of the spin states at zero field is called zero-field splitting (ZFS) and it is common for systems with $S \geq 1$. For triplet states of organic molecules ($S = 1$) the ZFS arises from the dipolar interaction of the two magnetic moments of the electron spins with each other. The interaction is described by an additional term

$$\mathcal{H}_{\text{ZFS}} = \mathbf{S} \mathbf{D} \mathbf{S} \quad (\text{B1.15.19})$$

that must be included in a spin Hamiltonian when $S \geq 1$. The spin-spin coupling (or ZFS) tensor \mathbf{D} is a symmetric and traceless ($\sum_{i=X,Y,Z} D_{ii} = 0$) second-rank tensor. Therefore, \mathbf{D} can be written in its principal axis frame with only two parameters

$$\mathcal{H}_{\text{ZFS}} = D(S_z^2 - \frac{1}{3}\mathbf{S}^2) + E(S_x^2 - S_y^2) \quad (\text{B1.15.20})$$

where $D \approx \frac{2}{3} D_{ZZ}$ is the axial and $E = \frac{1}{2} (D_{XX} - D_{YY})$ is the rhombic zero-field parameter. One may define an asymmetry parameter $\eta_D = E/D$ of the \mathbf{D} -tensor. The case of $\eta_D = 0$ (or $E = 0$) corresponds to an axially symmetric ZFS tensor ($D_{XX} = D_{YY}$) and two of the states will remain degenerate at zero magnetic field. The ZFS parameters can in general be determined from the EPR spectrum (for $\mathcal{H}_{\text{ZFS}} < \mathcal{H}_{\text{EZI}}$). In liquids the ZFS is averaged out to zero.

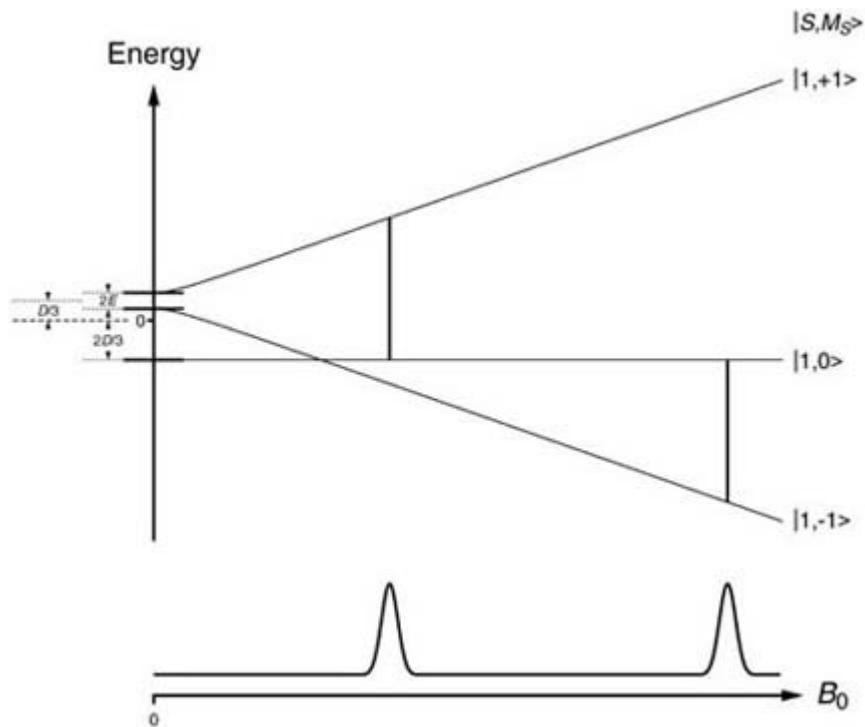


Figure B1.15.2. The state energies and corresponding eigenfunctions (high-field labels) as a function of the applied magnetic field B_0 for a system of spin $S = 1$ and $\mathbf{B} \parallel z$, shown for $D > 0$ and $E \neq 0$. The two primary transitions ($\Delta M_S = \pm 1$) are indicated for a constant frequency spectrum. Note that, because $E \neq 0$, the state energies vary nonlinearly with B_0 at low B_0 .

(C) ELECTRON SPIN EXCHANGE INTERACTION

For systems with two unpaired electrons, such as radical pairs, biradicals or triplets, there are four spin states that can be represented by the product functions $|M_{S1}\rangle \otimes |M_{S2}\rangle$, i.e. $|\alpha_1\alpha_2\rangle$, $|\alpha_1\beta_2\rangle$, $|\beta_1\alpha_2\rangle$ and $|\beta_1\beta_2\rangle$, where the subscripts indicate electron spin 1 and electron spin 2, respectively. In a paramagnetic centre of moderate size, however, it is more advantageous to combine these configurations into combination states because, in addition to the dipolar coupling between the spin magnetic moments, there is also an electrostatic interaction between the electron spins, the so-called exchange interaction, which gives rise to an energy separation between the singlet state, $|S\rangle \approx \frac{1}{\sqrt{2}}(|\alpha_1\beta_2\rangle - |\beta_1\alpha_2\rangle)$, and the triplet states ($|T_+\rangle = |\alpha_1\alpha_2\rangle$, $|T_0\rangle \approx \frac{1}{\sqrt{2}}(|\alpha_1\beta_2\rangle + |\beta_1\alpha_2\rangle)$, $|T_-\rangle = |\beta_1\beta_2\rangle$). The magnitude of the (isotropic) exchange interaction can be derived from the overlap of the wavefunctions and is described by the Hamiltonian

$$\mathcal{H}_{EX} = -2JS_1 \cdot S_2 \quad (\text{B1.15.21})$$

-10-

which denotes a scalar coupling between the spins $\mathbf{S}_1 = (S_{1x}, S_{1y}, S_{1z})$ and $\mathbf{S}_2 = (S_{2x}, S_{2y}, S_{2z})$. The energy separation between the $|S\rangle$ and $|T_0\rangle$ wavefunctions is determined by the exchange coupling constant J . For $J > 0$ the singlet state is higher in energy than the $|T_0\rangle$ -state. The observed properties of the system depend on the magnitude of J . If it is zero the two spins behave completely independently and one would have a true biradical. At the other extreme, when J is large the singlet lies far above the triplet and the magnetic resonance properties are solely determined by the interaction within the triplet manifold.

(D) HYPERFINE INTERACTION

The interaction of the electron spin's magnetic dipole moment with the magnetic dipole moments of nearby nuclear spins provides another contribution to the state energies and the number of energy levels, between which transitions may occur. This gives rise to the hyperfine structure in the EPR spectrum. The so-called hyperfine interaction (HFI) is described by the Hamiltonian

$$\mathcal{H}_{\text{HFI}} = \mathbf{SAI} \quad (\text{B1.15.22})$$

where \mathbf{A} is the HFI matrix and $\mathbf{I} = (I_x, I_y, I_z)$ is the vector representation of the nuclear spin. The HFI consists of two parts and therefore, equation (b1.15.22) can be separated into the sum of two terms

$$\mathcal{H}_{\text{HFI}} = \mathbf{SA}^{\text{dip}}\mathbf{I} + a\mathbf{S} \cdot \mathbf{I} \quad (\text{B1.15.23})$$

where the first term describes the anisotropic dipolar coupling through space between the electron spin and the nuclear spin. \mathbf{A}^{dip} is the symmetric and traceless dipolar HFI matrix. In the so-called point-dipole approximation, where both spins are assumed to be located, this part is given by

$$\mathcal{H} = -g\beta_e g_n \beta_n \left[\frac{\mathbf{I} \cdot \mathbf{S}}{r^3} - \frac{3(\mathbf{I}\mathbf{r})(\mathbf{S}\mathbf{r})}{r^5} \right]. \quad (\text{B1.15.24})$$

In equation (b1.15.24), \mathbf{r} is the vector connecting the electron spin with the nuclear spin, r is the length of this vector and g_n and β_n are the g -factor and the Bohr magneton of the nucleus, respectively. The dipolar coupling is purely anisotropic, arising from the spin density of the unpaired electron in an orbital of non-spherical symmetry (i.e. in p, d or f-orbitals) with a vanishing electron density at the nucleus. Since \mathbf{A}^{dip} is traceless the dipolar interactions are averaged out in isotropic fluid solution and only the orientation-independent isotropic coupling represented by the second term in equation (b1.15.23) gives rise to the observed hyperfine coupling in the spectrum. This isotropic contribution is called the (Fermi) contact interaction arising from electrons in s orbitals (spherical symmetry) with a finite probability ($|\psi(0)|^2$) of finding the electron at the nucleus. The general expression for the isotropic hyperfine coupling constant is

$$a = \frac{2\mu_0}{3} g\beta_e g_n \beta_n |\Psi(0)|^2. \quad (\text{B1.15.25})$$

Hence, a measurement of hyperfine coupling constants provides information on spin densities at certain positions in the molecule and thus renders a map of the electronic wavefunction.

The simplest system exhibiting a nuclear hyperfine interaction is the hydrogen atom with a coupling constant of 1420 MHz. If different isotopes of the same element exhibit hyperfine couplings, their ratio is determined by the ratio of the nuclear g -values. Small deviations from this ratio may occur for the Fermi contact interaction, since the electron spin probes the inner structure of the nucleus if it is in an s orbital. However, this so-called hyperfine anomaly is usually smaller than 1%.

(E) NUCLEAR ZEEMAN AND NUCLEAR QUADRUPOLE INTERACTION

While all contributions to the spin Hamiltonian so far involve the electron spin and cause first-order energy shifts or splittings in the EPR spectrum, there are also terms that involve only nuclear spins. Aside from their importance for the calculation of ENDOR spectra, these terms may influence the EPR spectrum significantly in situations where the high-field approximation breaks down and second-order effects become important. The first of these interactions is the coupling of the nuclear spin to the external magnetic field, called the

nuclear Zeeman interaction (NZI). Neglecting chemical shift anisotropies that are usually small and not resolved in ENDOR spectra it can be considered isotropic and written as

$$\mathcal{H}_{\text{NZI}} = -g_n \beta_n \mathbf{B} \cdot \mathbf{I}. \quad (\text{B1.15.26})$$

The negative sign in equation (b1.15.26) implies that, unlike the case for electron spins, states with larger magnetic quantum number have smaller energy for $g_n > 0$. In contrast to the g -value in EPR experiments, g_n is an inherent property of the nucleus. NMR resonances are not easily detected in paramagnetic systems because of sensitivity problems and increased linewidths caused by the presence of unpaired electron spins.

Since atomic nuclei are not perfectly spherical their spin leads to an electric quadrupole moment if $I \geq 1$ which interacts with the gradient of the electric field due to all surrounding electrons. The Hamiltonian of the nuclear quadrupole interactions can be written as tensorial coupling of the nuclear spin with itself

$$\mathcal{H}_{\text{NQI}} = \mathbf{I} \mathbf{P} \mathbf{I} \quad (\text{B1.15.27})$$

where \mathbf{P} is the quadrupole coupling tensor. Comparison with [equation \(b1.15.19\)](#) shows that the NQI can be formally treated in a way analogous to that for the ZFS. In liquids the NQI is averaged out to zero.

(F) THE COMPLETE SPIN HAMILTONIAN

The complete spin Hamiltonian for a description of EPR and ENDOR experiments is given by

$$\mathcal{H} = \mathcal{H}_{\text{EZI}} + \mathcal{H}_{\text{ZFS}} + \mathcal{H}_{\text{EX}} + \mathcal{H}_{\text{HFI}} + \mathcal{H}_{\text{NZI}} + \mathcal{H}_{\text{NQI}}. \quad (\text{B1.15.28})$$

The approximate magnitudes of the terms in equation (b1.15.28) are shown in an overview in [figure B1.15.3](#) (see also [2, 3, 11]).

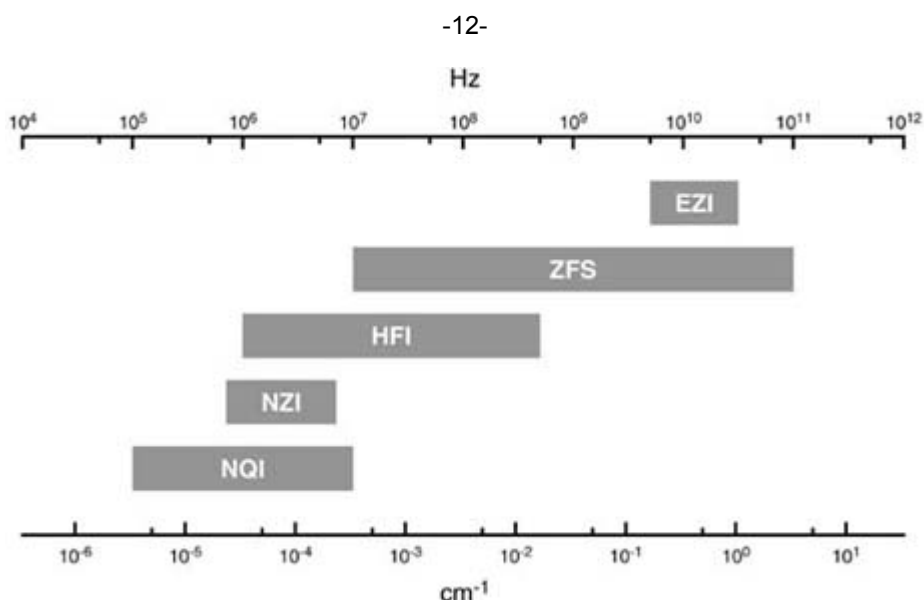


Figure B1.15.3. Typical magnitudes of interactions of electron and nuclear spins in the solid state (logarithmic scale).

B1.15.3 EPR INSTRUMENTATION

A typical CW-detected EPR spectrum is recorded by holding the frequency constant and varying the external magnetic field. In doing so one varies the separation between the energy levels to match it to the quantum of the radiation. Even though magnetic resonance could also be achieved by sweeping the frequency at a fixed magnetic field, high-frequency sources with a broad frequency range and simultaneously low enough noise characteristics have not yet been devised to be practical for frequency-swept EPR.

EPR absorption has been detected from zero magnetic field up to fields as high as 30 T corresponding to a MW frequency of 10^{12} Hz. There are various considerations that influence the choice of the radiation frequency. Higher frequencies, which require higher magnetic fields, give inherently greater sensitivity by virtue of a more favourable Boltzmann factor (see equation (b1.15.11)). However, several factors place limits on the frequency employed, so that frequencies in the MW region of the electromagnetic spectrum remain favoured. One limitation is the sample size; at frequencies around 40 GHz the dimensions of a typical resonant cavity are of the order of a few millimetres, thus restricting the sample volume to about 0.02 cm^3 . The requirement to reduce the sample size roughly compensates for the sensitivity enhancement in going to higher fields and frequencies in EPR. However, the sensitivity advantage persists if only small quantities of the sample are available. This is often the case for biological samples, particularly when single-crystal studies are intended. Second, high frequencies require high magnetic fields that are homogeneous over the sample volume. Sufficiently homogeneous magnetic fields above 2.5 T are difficult to produce with electromagnets. Superconducting magnets are commercially available for magnetic fields up to 22 T, but they are expensive and provide only small room-temperature bores, thus limiting the space available for the resonator. Third, the small size of MW components for high frequencies makes their fabrication technically difficult and costly. These and other factors have resulted in a choice of the frequency region around 10 GHz (usually denominated the X-band region) as the resonance frequency of most commercially available spectrometers. The most common frequency

-13-

bands for high-field/high-frequency EPR are Q-band (35 GHz) and W-band (95 GHz), where the wavelengths are 8 mm and 3 mm, respectively. In order to carry out EPR experiments in larger objects such as intact animals it appears necessary to use lower frequencies because of the large dielectric losses of aqueous samples. At L-band frequencies (1–2 GHz), with appropriate configurations of EPR resonators, whole animals the size of mice can be studied by insertion into the resonator. Table B1.15.1 lists typical frequencies and wavelengths, together with the resonant fields required for resonance of a free electron.

Table B1.15.1 Some frequencies and resonance fields (for $g = 2$) used in EPR.

Typical EPR frequency ν (GHz)	Vacuum wavelength λ (m)	Typical EPR field B_0 (T)	Band designation Frequency (GHz)
1.5	2×10^{-1}	0.054	L 0.39–1.55
3.0	1×10^{-1}	0.107	S 1.55–3.9
6.0	5×10^{-2}	0.214	C 3.9–6.2
9.5	3.2×10^{-2}	0.339	X 6.2–10.9
24	1.2×10^{-2}	0.856	K 10.9–36
36	8.3×10^{-3}	1.285	Q 36–46
50	6.0×10^{-3}	1.784	V 46–56
95	3.2×10^{-3}	3.390	W 56–100
140	2.1×10^{-3}	4.996	D
250	1.2×10^{-3}	8.921	—
360	8.3×10^{-4}	12.846	—
604	5.0×10^{-4}	21.552	—

The components of a typical EPR spectrometer operating at X-band frequencies [3, 4, 6] are shown in [figure B1.15.4](#).

Up to the present the MW radiation has usually been provided by reflex klystrons, which essentially consist of a vacuum tube and a pair of electrodes to produce an electron beam that is velocity modulated by a radio-frequency (RF) electric field. The net effect is the formation of groups or bunches of electrons. Using a reflector the bunched electron beam is turned around and will debunch, giving up energy to the cavity, provided the RF frequency and the beam and reflector voltages are properly adjusted. This will set up one of several stable klystron modes at the cavity frequency f_0 ; the mode corresponding to the highest output of power is usually the one utilized. A typical klystron has a mechanical tuning range allowing the klystron cavity frequency to be tuned over a range of 5–50% around f_0 . The adjustment of the reflector voltage allows one to vary the centre frequency of a given mode over a very limited range of 0.2–0.8%. Often a low-amplitude sine-wave reflector voltage modulation is employed as an integral part of an automatic frequency control (AFC). It is desirable that the klystron frequency be very stable; hence, fluctuations of the klystron temperature or of applied voltages must be minimized and mechanical vibrations suppressed. Klystrons are employed as generators of nearly monochromatic output radiation in the frequency range from 1 to 100 GHz. In commercial EPR spectrometers the klystron normally provides less than 1 W of continuous output power. Increasingly, however, solid-state devices, such as Gunn-effect oscillators and IMPATTs, are superseding klystron tubes.

-14-

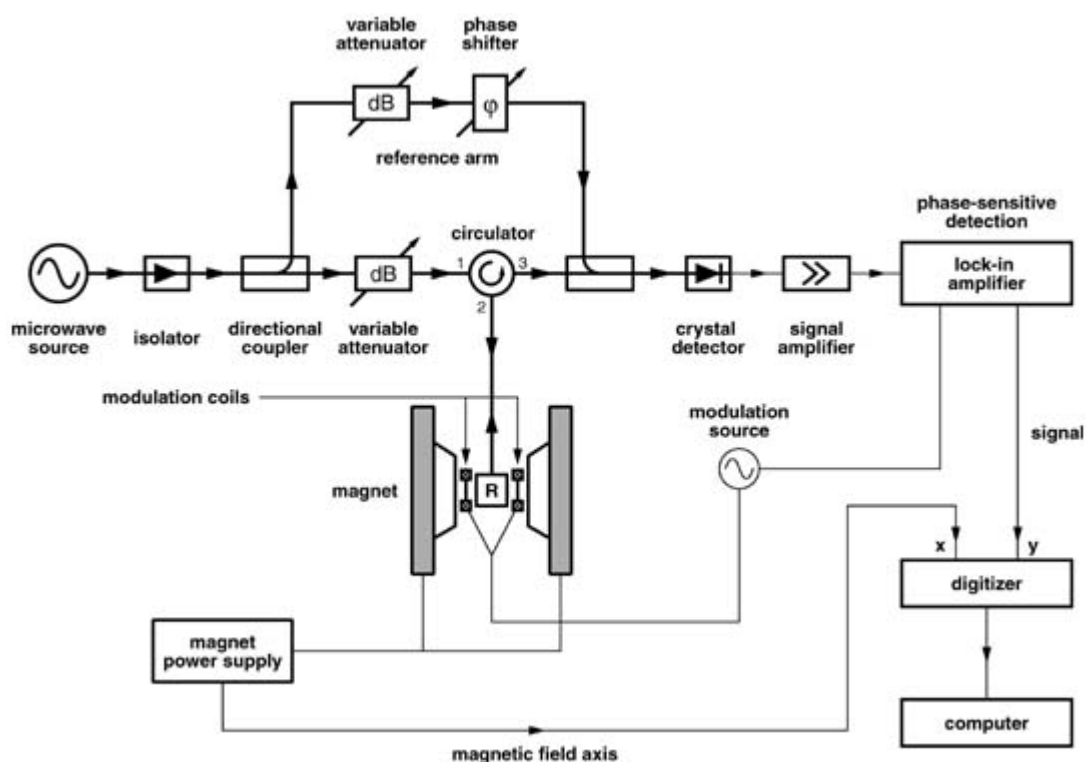


Figure B1.15.4. Block diagram of a typical EPR spectrometer operating at X-band frequencies.

Waveguides are commonly used to transmit microwaves from the source to the resonator and subsequently to the receiver. For not-too-high-frequency radiation (≤ 10 GHz) low-loss MW transmission can also be achieved using strip-lines and coaxial cables. At the output of a klystron an isolator is often used to prevent back-reflected microwaves to perturb the on-resonant klystron mode. An isolator is a microwave-ferrite device that permits the transmission of microwaves in one direction and strongly attenuates their propagation in the other direction. The principle of this device involves the Faraday effect, that is, the rotation of the polarization

planes of the microwaves.

The amount of MW power reaching the sample in the resonator is controlled by a variable attenuator. Like the isolator, the circulator is a non-reciprocal device that serves to direct the MW power to the resonator (port 1 → port 2) and simultaneously allows the signal reflected at resonance to go from the resonator directly to the receiver (port 2 → port 3).

Although achievement of resonance in an EPR experiment does not require the use of a resonant cavity, it is an integral part of almost all EPR spectrometers. A resonator dramatically increases the sensitivity of the spectrometer and greatly simplifies sample access. A resonant cavity is the MW analogue of a RF-tuned *RLC* circuit and many expressions derived for the latter may also be applied to MW resonators. A typical resonant cavity for microwaves is a box or a cylinder fabricated from high-conductivity metal and with dimensions comparable to the wavelength of the radiation. Each particular cavity size and shape can sustain oscillations in a number of different standing wave patterns, called resonator modes. Visual images and mathematical expressions of the distributions of the electric and magnetic field vectors within the cavity can be derived from Maxwell's equations with suitable boundary conditions.

-15-

The locations of the maxima of the \mathbf{B}_1 -field and the \mathbf{E} -field are different depending on the mode chosen for the EPR experiment. It is desirable to design the cavity in such a way that the \mathbf{B}_1 field is perpendicular to the external field \mathbf{B} , as required by the nature of the resonance condition. Ideally, the sample is located at a position of maximum \mathbf{B}_1 , because below saturation the signal-to-noise ratio is proportional to B_1 . Simultaneously, the sample should be placed at a position where the \mathbf{E} -field is a minimum in order to minimize dielectric power losses which have a detrimental effect on the signal-to-noise ratio.

The sharpness of the frequency response of a resonant system is commonly described by a factor of merit, called the quality factor, $Q = \nu / \Delta\nu$. It may be obtained from a measurement of the full width at half maximum $\Delta\nu$, of the resonator frequency response curve obtained from a frequency sweep covering the resonance. The sensitivity of a system (proportional to the inverse of the minimum detectable number of paramagnetic centres in an EPR cavity) critically depends on the quality factor

$$S \propto Q\eta \quad (\text{B1.15.29})$$

where $\eta = \int_0^t \mathbf{G}(t') dt' = \mathbf{0}_{B_1}^2 dV$ is the filling factor. The cavity types most commonly employed in EPR are the rectangular-parallelepiped cavity and the cylindrical cavity. The rectangular cavity is typically operated in a transverse electric mode, TE_{102} , which permits the insertion of large samples with low dielectric constants. It is especially useful for liquid samples in flat cells, which may extend through the entire height of the cavity. In the cylindrical cavity a TE_{011} mode is frequently used because of its fairly high Q -factor and the very strong B_1 along the sample axis.

Microwaves from the waveguide are coupled into the resonator by means of a small coupling hole in the cavity wall, called the iris. An adjustable dielectric screw (usually machined from Teflon) with a metal tip adjacent to the iris permits optimal impedance matching of the cavity to the waveguide for a variety of samples with different dielectric properties. With an appropriate iris setting the energy transmission into the cavity is a maximum and simultaneously reflections are minimized. The optimal adjustment of the iris screw depends on the nature of the sample and is found empirically.

Other frequently used resonators are dielectric cavities and loop-gap resonators (also called split-ring resonators) [12]. A dielectric cavity contains a diamagnetic material that serves as a dielectric to raise the effective filling factor by concentrating the \mathbf{B}_1 field over the volume of the sample. Hollow cylinders machined from fused quartz or sapphire that host the sample along the cylindrical axis are commonly used.

Loop-gap resonators consist of one or a series of cylindrical loops interrupted by at least one or several gaps. Loops and gaps act as inductive and capacitive elements, respectively. With a suitable choice of loop and gap dimensions, resonators operating at different resonance frequencies over a wide range of the MW spectrum can be constructed. Loop-gap resonators typically have low Q -factors. Their broad-bandwidth frequency response, $\Delta\nu$, makes them particularly useful in EPR experiments where high time resolution, $\tau_{\text{res}}=1/(2\pi\Delta\nu)$, because fast signal changes are required. Excellent filling factors, η , may be obtained with loop-gap devices; the high η makes up for the typically low Q to yield high sensitivity (see equation (b1.15.29)), valuable for small sample sizes and in pulsed EPR experiments. Coupling of microwaves into these cavities is most conveniently accomplished by a coupling loop that acts as an antenna. Typically, the distance between the antenna and the loop-gap resonator is varied in order to obtain optimal impedance matching.

When the applied magnetic field is swept to bring the sample into resonance, MW power is absorbed by the sample. This changes the matching of the cavity to the waveguide and some power is now reflected and passes via the circulator to the detector. This reflected radiation is thus the EPR signal.

-16-

The most commonly used detector in EPR is a semiconducting silicon crystal in contact with a tungsten wire, which acts as an MW rectifier. At microwatt powers, crystal detectors are typically non-linear and render a rectified current that is proportional to the MW power (i.e. proportional to B_1^2). In the milliwatt region, the rectified crystal current becomes proportional to the square root of the MW power (i.e. proportional to B_1), and the crystal behaves as a linear detector. In EPR spectroscopy it is preferred to operate the crystal rectifier in its linear regime. However, since the EPR signal is typically rather small, the diode needs to be biased to operate it at higher MW power levels. This can be done by slightly mismatching the cavity to the waveguide in order to increase the MW power back-reflected from the cavity, or by adding microwaves at a constant power level guided through the reference arm (often called the bypass arm) of the spectrometer. The reference arm takes microwaves from the waveguide ahead of the circulator and returns them with adjusted phase and power behind the circulator. When properly adjusted, the reference arm can also be used to detect the in-phase (χ') and out-of-phase (χ'') components of the EPR signal with respect to the phase of the microwaves.

When sweeping the magnetic field through resonance, a crystal detector renders a slowly varying DC signal which is not readily processed and which is superimposed by low-frequency noise contributions. To overcome this, a phase-sensitive detection technique utilizing small-amplitude magnetic field modulation is employed in most EPR spectrometers. Modulation of the magnetic field is achieved by placing a pair of Helmholtz coils on each side of the cavity along the axis of the external magnetic field. An alternating current is fed through them and a small oscillating magnetic field is induced which is superimposed on the external magnetic field. The effect of the modulation is depicted in [figure B1.15.5](#). Provided the amplitude of the modulation field is small compared to the linewidth of the absorption signal, $\Delta B_{1/2}$, the change in MW power at the detector will contain an oscillatory component at the modulation frequency whose amplitude will be proportional to the slope of the EPR line. A lock-in detector compares the modulated EPR signal from the crystal with a reference and only passes the components of the signal that have the proper frequency and phase.

The reference voltage comes from the same frequency generator that produces the field modulation voltage and this causes the EPR signal to pass through while most noise at frequencies other than the modulation frequency is suppressed. As a result of phase-sensitive detection using lock-in amplification one typically obtains the first derivative of the absorption line EPR signal. The application of field modulation, however, can cause severe lineshape distortion: to limit modulation-induced line broadening to below 1% of the undistorted linewidth, $\Delta B_{1/2}$ requires small modulation amplitudes ($B_{\text{mod}} < 0.15 \Delta B_{1/2}$ for Lorentzian lineshapes and $B_{\text{mod}} < 0.3 \Delta B_{1/2}$ for Gaussian lineshapes).

After the signal emerges from the lock-in amplifier it still contains a considerable amount of noise. Most of the noise contributions to the signal can be eliminated by passing the signal through a low-pass filter. The filter time constant is a measure of the cutoff frequency of the filter. If accurate linewidth and g -factor

measurements are intended, one must be careful to employ a sufficiently short response time because lineshape distortions may occur as a result of too intense filtering.

-17-

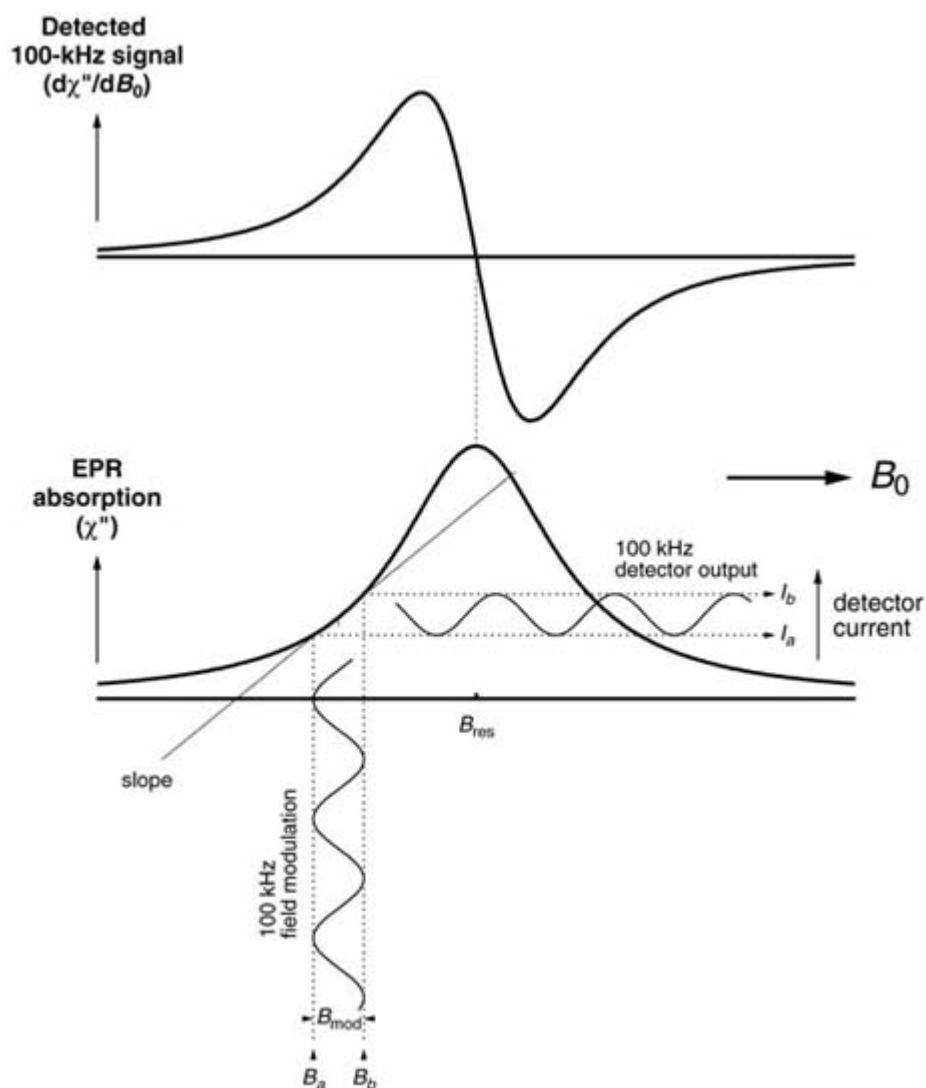


Figure B1.15.5. Effect of small-amplitude 100 kHz field modulation on the detector output current. The static magnetic field is modulated between the limits B_a and B_b . The corresponding detector current varies between the limits I_a and I_b . The upper diagram shows the recorded 100 kHz signal as a function of B_0 . After [3].

Figure B1.15.6 presents a liquid-phase EPR spectrum of an organic radical measured using a conventional EPR spectrometer like the one depicted in Figure B1.15.4. As is usual, the lines are presented as first derivatives $d\chi''/dB_0$ of the power absorbed by the spins. The spectrum shows a pronounced pattern of hyperfine lines arising from two different groups of protons (see also Figure B1.15.9). The number, spacing and intensity of the lines provides information on the molecular and electronic structure of the molecule carrying the unpaired electron spin. The individual lines have a Lorentzian lineshape with a homogeneous linewidth determined by T_2 . The most common case for inhomogeneously broadened lines giving rise to a Gaussian lineshape is unresolved hyperfine interactions arising from a large number of nonequivalent nuclei and anisotropies of the hyperfine coupling which will persist when recording EPR spectra of radicals in solids.

-18-

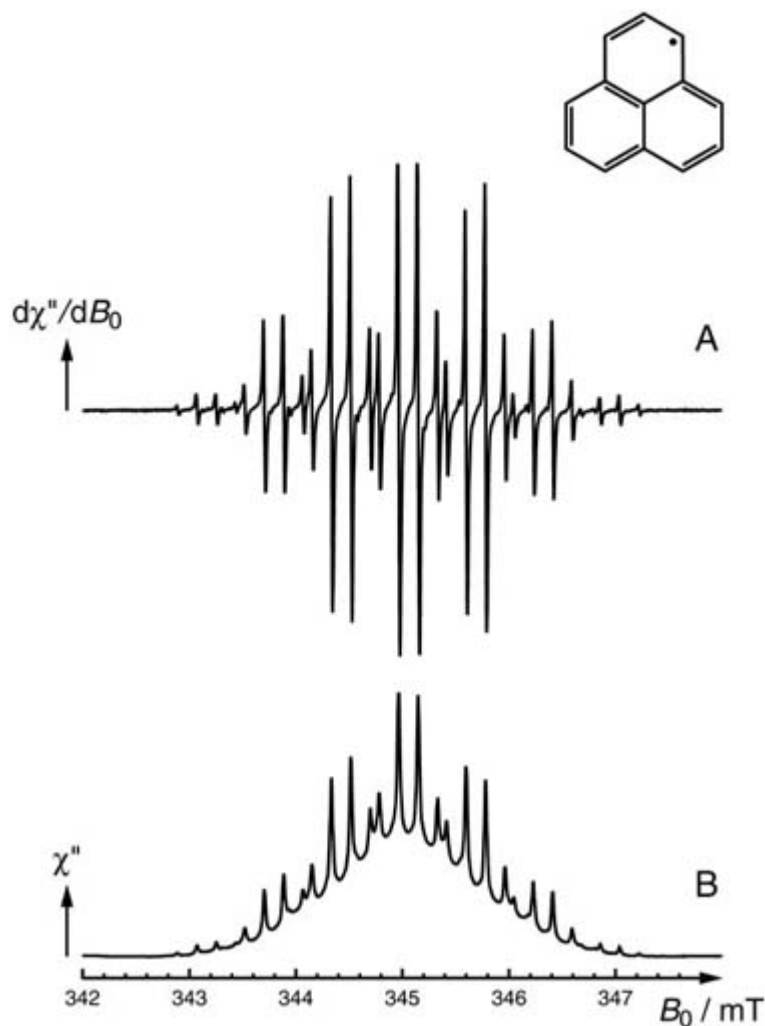


Figure B1.15.6. The EPR spectrum of the perinaphtheryl radical in mineral oil taken at room temperature. (A) First derivative of the EPR absorption χ'' with respect to the external magnetic field, B_0 . (B) Integrated EPR spectrum.

EPR has been successfully applied to radicals in the solid, liquid and gaseous phase. Goniometer techniques have been adopted to measure anisotropic magnetic interactions in oriented (e.g. single-crystal) and partially oriented (e.g. film) samples as a function of the sample orientation with respect to the external field. Variable temperature studies can provide a great deal of information about a spin system and its interactions with its environment. Therefore, low-temperature as well as high-temperature EPR experiments can be conducted by either heating or cooling the entire cavity in a temperature-controlled cryostat or by heating or cooling the sample in a jacket inserted into the cavity. Specialized cavity designs have also been worked out to perform EPR studies under specific conditions (e.g. high pressures). Sample irradiation is facilitated through shielded openings in the cavity.

The practical goal of EPR is to measure a stationary or time-dependent EPR signal of the species under scrutiny and subsequently to determine magnetic interactions that govern the shape and dynamics of the EPR response of the spin system. The information obtained from a thorough analysis of the EPR signal, however, may comprise not only the parameters enlisted in the previous chapter but also a wide range of other physical parameters, for example reaction rates or orientation order parameters.

B1.15.4 TIME-RESOLVED CW EPR METHODS

Although EPR in general has the potential to follow the concentration changes of short-lived paramagnetic intermediates, standard CW EPR using field modulation for narrow-band phase-sensitive detection is geared for high sensitivity and correspondingly has only a mediocre time resolution. Nevertheless, transient free radicals in the course of (photo-)chemical processes can be studied by measuring the EPR line intensity of a spectral feature as a function of time at a fixed value of the external magnetic field. Typically, the optimum time response of a commercial spectrometer which uses a CW fixed-frequency lock-in detection is in the order of 20 μs . By use of field modulation frequencies higher than the 100 kHz usually employed in commercial instruments, the time resolution can be increased by about an order of magnitude, which makes this method well suited for the study of transient free radicals on a microsecond timescale.

B1.15.4.1 TRANSIENT EPR SPECTROSCOPY

The time resolution of CW EPR can be considerably improved by removing the magnetic field modulation completely. Rather, a suitably fast data acquisition system is employed to directly detect the transient EPR signal as a function of time at a fixed magnetic field. In transient EPR spectroscopy (TREPR) [13, 14] paramagnetic species (e.g. free radicals, radical pairs, triplets or higher multiplet states) are generated on a nanosecond timescale by a short laser flash or radiolysis pulse and the arising time-dependent signals are detected in the presence of a weak MW magnetic field. For this purpose the standard EPR spectrometer shown in [figure B1.15.4](#) needs to be modified. The components for the field modulation may be removed, and the lock-in amplifier and digitizer are replaced by a fast transient recorder or a digital oscilloscope triggered by a photodiode in the light path of the laser. The response time of such a spectrometer is potentially controlled by the bandwidth of each individual unit. Provided that the MW components are adequately broadbanded and the laser flash is sufficiently short (typically a few nanoseconds), the time resolution can be pushed to the 10^{-8} s range (which is not far from the physical limit given by the inverse MW frequency) if a resonator with a low Q -value (and hence wide bandwidth) is used. Fortunately, the accompanying sensitivity loss (see [equation \(b1.15.29\)](#)) can be compensated to a large extent by using resonators with high filling factors [12]. Furthermore, excellent sensitivity is obtained in studies of photoprocesses where the light-generated paramagnetic species are typically produced in a state of high electron spin polarization (or, in other words, removed from the thermal equilibrium population of the electron spin states). Nevertheless, the EPR time profiles at a fixed magnetic field position are repeatedly measured in order to improve the signal-to-noise ratio by a factor \sqrt{N} , where N is the number of time traces averaged.

Following the pioneering work by Kim and Weissman [15], it has been demonstrated that TREPR works for a broad range of resonance frequencies from 4 GHz (S-band) up to 95 GHz (W-band). As an example the time-dependent EPR signal of the photo-generated triplet state of pentacene in a *para*-terphenyl single crystal obtained by TREPR at X-band is shown in [figure B1.15.7](#). Note that EPR signals taken in direct detection appear in the absorption (or dispersion)

mode, not in the usual derivative form associated with field modulation and phase-sensitive detection. Therefore, positive signals indicate absorptive (A) and negative signals emissive (E) EPR transitions.

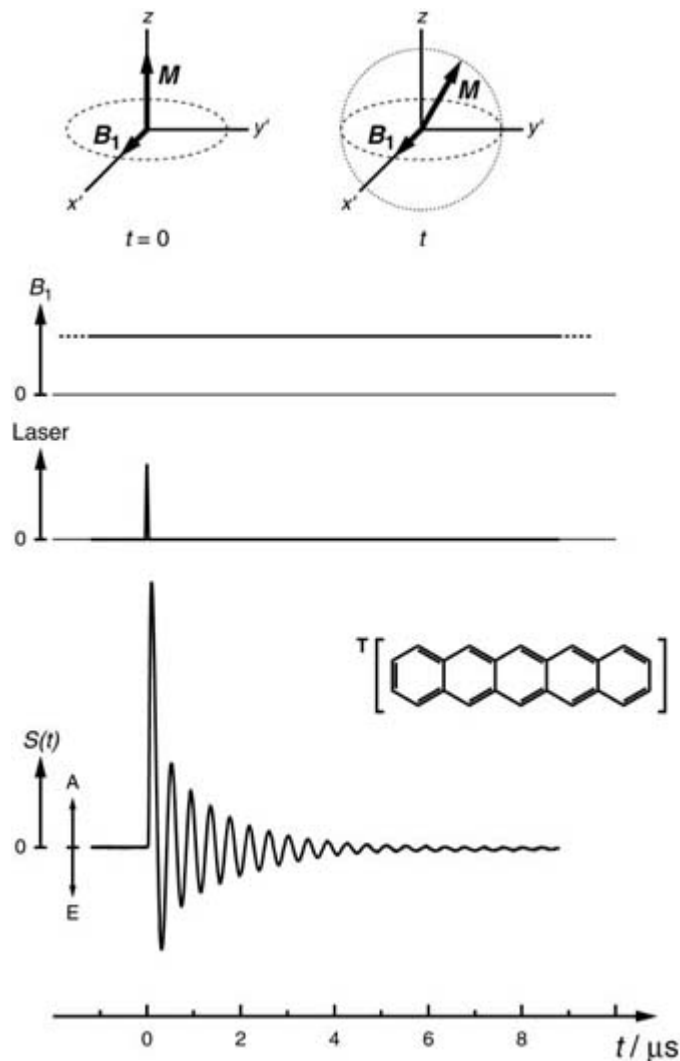


Figure B1.15.7. Transient EPR. Bottom: time-resolved EPR signal of the laser-flash-induced triplet state of pentacene in *p*-terphenyl. $B_1=0.085$ mT. Top: initially, the transient magnetization \mathbf{M} is aligned along $\mathbf{B} \parallel z$. In the presence of a MW magnetic field \mathbf{B}_1 the magnetization precesses about $\mathbf{B}_1 \parallel x'$ (rotating frame representation).

The following discussion of the time dependence of the EPR response in a TREPR experiment is based on the assumption that the transient paramagnetic species is long lived with respect to the spin relaxation parameters. When ω_1 is large compared to the inverse relaxation times, i.e. $\omega_1 = \gamma B_1 \gg T_1^{-1}, T_2^{-1}$, or, in other words, for high MW powers, the signal exhibits oscillations with a frequency proportional to the MW magnetic field B_1 . These so-called transient nutations are observed if resonance between an electron spin transition and a coherent radiation field is suddenly achieved. The phenomenon can be understood when viewing the motion of the magnetization vector, \mathbf{M} , in a

reference frame, (x', y', z) , rotating at the frequency ω of the MW field around the static external field ($\mathbf{B}_0 \parallel z$). Hence, in the rotating frame B_1 is a stationary field defined along the x' -axis as indicated in [figure B1.15.7](#). Paramagnetic species are created during the laser flash at $t = 0$ with their spins aligned with the applied magnetic field, which causes an initial magnetization $M_z(0)$ in this direction. This is acted upon by the radiation field at or near resonance, which rotates it to produce an orthogonal component, $M_{y'}(t)$, to which the observed signal is proportional. The signal initially increases as $M_{y'}(t)$ grows under the rotation of \mathbf{M} about \mathbf{B}_1 (at exact resonance) or $\mathbf{B}_0 + \mathbf{B}_1$ (near resonance), but, while the system approaches a new state via oscillations,

M continually decreases under the influence of spin–spin relaxation which destroys the initial phase coherence of the spin motion within the y/z -plane. In solid-state TREPR, where large inhomogeneous EPR linewidths due to anisotropic magnetic interactions persist, the long-time behaviour of the spectrometer output, $S(t)$, is given by

$$S(t) \propto \omega_1 J_0(\omega_1 t) e^{-t/(2T_2)} \quad (\text{B1.15.30})$$

where the oscillation of the transient magnetization is described by a Bessel function $J_0(\omega_1 t)$ of zeroth order, damped by the spin–spin relaxation time T_2 . At low MW powers ($\omega_1^2 T_1 T_2 \ll 1$) an exponential decay of the EPR signal is observed, governed by spin–lattice relaxation

$$S(t) \propto \omega_1 e^{-t/T_1}. \quad (\text{B1.15.31})$$

The rise time of the signals—independent of the chosen MW power—is proportional to the inverse inhomogeneous EPR linewidth. As can be seen from equations (b1.15.30) and (b1.15.31) a measurement of the ω_1 -dependence of the transient EPR signals provides a straightforward method to determine not only the relaxation parameters of the spin system but also the strength of the MW magnetic field B_1 at the sample. Spectral information can be obtained from a series of TREPR signals taken at equidistant magnetic field points covering the total spectral width. This yields a two-dimensional variation of the signal intensity with respect to both the magnetic field and the time axis. Transient spectra can be extracted from such a plot at any fixed time after the laser pulse as slices parallel to the magnetic field axis.

B1.15.4.2 MW-SWITCHED TIME INTEGRATION METHOD (MISTI)

An alternative method to obtain accurate values of the spin–lattice relaxation time T_1 is provided by the TREPR technique with gated MW irradiation, also called the MW-switched time integration method (MISTI) [13, 14]. The principle is quite simple. The MW field is switched on with a variable delay τ after the laser flash. The amplitude of the transient signal plotted as a function of τ renders the decay of the spin-polarized initial magnetization towards its equilibrium value. This method is preferred over the TREPR technique at low MW power (see equation (b1.15.31)) since the spin system is allowed to relax in the absence of any resonant MW field in a true spin–lattice relaxation process. The experiment is carried out by adding a PIN diode MW switch between the MW source and the circulator (see figure B1.15.4, and set between a pair of isolators. Since only low levels of MW power are switched (typically less than 1 W), as opposed to those in ESE and FT EPR, the detector need not be protected against high incident power levels.

As a summary it may be of interest to point out why TREPR spectroscopy and related methods remain important in the EPR regime, even though pulsed EPR methods are becoming more and more widespread. (1) For the case of an inhomogeneously broadened EPR line the time resolution of TREPR compares favourably with pulsed techniques.

(2) The low MW power levels commonly employed in TREPR spectroscopy do not require any precautions to avoid detector overload and, therefore, the full time development of the transient magnetization is obtained undiminished by any MW detection deadtime. (3) Standard CW EPR equipment can be used for TREPR requiring only moderate efforts to adapt the MW detection part of the spectrometer for the observation of the transient response to a pulsed light excitation with high time resolution. (4) TREPR spectroscopy proved to be a suitable technique for observing a variety of spin coherence phenomena, such as transient nutations [16], quantum beats [17] and nuclear modulations [18], that have been useful to interpret EPR data on light-induced spin-correlated radical pairs.

B1.15.5 MULTIPLE RESONANCE TECHNIQUES

In the previous chapters experiments have been discussed in which one frequency is applied to excite and detect an EPR transition. In multiple resonance experiments two or more radiation fields are used to induce different transitions simultaneously [19, 20, 21, 22 and 23]. These experiments represent elaborations of standard CW and pulsed EPR spectroscopy, and are often carried out to complement conventional EPR studies, or to refine the information which can in principle be obtained from them.

B1.15.5.1 ELECTRON–NUCLEAR DOUBLE RESONANCE SPECTROSCOPY (ENDOR)

It was noted earlier that EPR may at times be used to characterize the electronic structure of radicals through a measurement of hyperfine interactions arising from nuclei that are coupled to the unpaired electron spin. In very large radicals with low symmetry, where the presence of many magnetic nuclei results in a complex hyperfine pattern, however, the spectral resolution of conventional EPR is very often not sufficient to resolve or assign all hyperfine couplings. It was as early as 1956 that George Feher demonstrated that by electron–nuclear double resonance (ENDOR) the spectral resolution can be greatly improved [24]. In ENDOR spectroscopy the electron spin transitions are still used as means of detection because the sensitivity of the electron resonance measurement is far greater than that of the nuclear resonance. In brief, an EPR transition is saturated, which leads to a collapse of the observed EPR signal as the corresponding state populations equalize. If one now simultaneously irradiates the spin system with an RF field in order to induce transitions between the nuclear sublevels, the condition of saturation in the EPR transition is lifted as the nuclear sublevel populations shift, and there is a partial recovery of the EPR signal.

ENDOR transitions can be easily understood in terms of a simple system consisting of a single unpaired electron spin ($S=\frac{1}{2}$) coupled to a single nuclear spin ($I=\frac{1}{2}$). The interactions responsible for the various splittings are summarized in the following static Hamiltonian:

$$\mathcal{H} = \mathcal{H}_{EZI} + \mathcal{H}_{NZI} + \mathcal{H}_{HFI} = \beta_e \mathbf{B} g \mathbf{S} - g_n \beta_n \mathbf{B} \cdot \mathbf{I} + \mathbf{S} \mathbf{A} \mathbf{I}. \quad (\text{B1.15.32})$$

The coupling constants of the hyperfine and the electron Zeeman interactions are scalar as long as radicals in isotropic solution are considered, leading to the Hamiltonian

$$\mathcal{H} = g_e \beta_e \mathbf{B} \cdot \mathbf{S} - g_n \beta_n \mathbf{B} \cdot \mathbf{I} + a \mathbf{S} \cdot \mathbf{I}. \quad (\text{B1.15.33})$$

-23-

In the high-field approximation with $\mathbf{B} \parallel z$, the energy eigenvalues classified by the magnetic spin quantum numbers, M_S and M_I , are given by

$$E_{M_S, M_I} = g_e \beta_e B_0 M_S - g_n \beta_n M_I B_0 + a M_S M_I \quad (\text{B1.15.34})$$

where g_n and a may be positive or negative, thus leading to a different ordering of the levels. The energy level diagram for the case $a < 0$ and $|a|/2 < g_n \beta_n B_0$ is shown in figure B1.15.8. Adopting the notation $\hbar \omega_e = g_e \beta_e B_0$ and $\hbar \omega_N = g_n \beta_n B_0$, two EPR transitions are obtained,

$$\omega_{\text{EPR}} = \omega_e \pm a/(2\hbar) \quad (\text{B1.15.35})$$

which obey the selection rule $\Delta M_S = \pm 1$ and $\Delta M_I = 0$. The two ENDOR transitions are

$$\omega_{\text{ENDOR}}^{\pm} = |\omega_n \pm a/(2\hbar)| \quad (\text{B1.15.36})$$

which satisfy the selection rule $\Delta M_S = 0$ and $\Delta M_I = \pm 1$. The absolute value is used in equation (b1.15.36) to take into account the two cases $|a|/(2\hbar) < |\omega_n|$ and $|a|/(2\hbar) > |\omega_n|$. The corresponding ENDOR spectra are shown schematically in figures b1.15.8(B) and (C). Irrespective of the EPR line monitored, two ENDOR lines, separated by $|a|/\hbar$ and centred at $|\omega_n|$, are observed. For $|a|/(2\hbar) > |\omega_n|$ the two ENDOR transitions are given by $|a|/(2\hbar) \pm |\omega_n|$: again, two lines are observed; however, separated by $2|\omega_n|$ and centred at $|a|/(2\hbar)$.

-24-

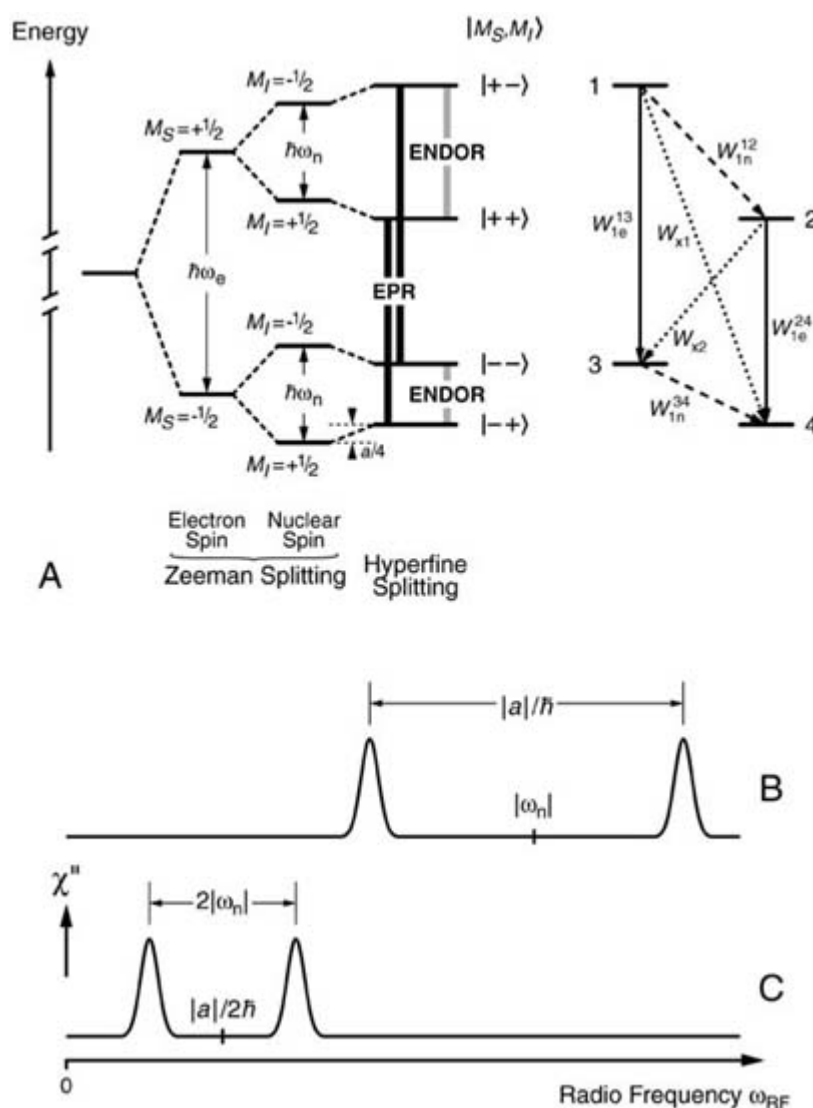


Figure B1.15.8. (A) Left side: energy levels for an electron spin coupled to one nuclear spin in a magnetic field, $S = I = \frac{1}{2}$, $g_n > 0$, $a < 0$, and $|a|/(2\hbar) < \omega_n$. Right side: schematic representation of the four energy levels with $|\pm\pm\rangle \equiv |M_S = \pm\frac{1}{2}, M_I = \pm\frac{1}{2}\rangle$. $|+-\rangle \equiv 1$, $|++\rangle \equiv 2$, $|--\rangle \equiv 3$ and $| -+\rangle \equiv 4$. The possible relaxation paths are characterized by the respective relaxation rates W . The energy levels are separated horizontally to distinguish between the two electron spin transitions. Bottom: ENDOR spectra shown when $|a|/(2\hbar) < |\omega_n|$ (B) and when $|\omega_n| < |a|/(2\hbar)$ (C).

For the simple system discussed above the advantages of performing double resonance do not become so

apparent: two lines are observed using either method, EPR or ENDOR. The situation dramatically changes when there are i groups of nuclei present, each group consisting of n_i magnetically equivalent nuclei with nuclear spin quantum number I_i , each one coupling to the unpaired electron spin with the hyperfine constant a_i . While the EPR spectrum will consist

-25-

of $\pi_i(2n_i I_i + 1)$ lines, for each group of equivalent nuclei, no matter how many nuclei there are or what their spin quantum number is, there will still be only two ENDOR lines separated by $|a_i|$ or $2\omega_{Ni}$. Hence, with increasing number of groups of nuclei the number of ENDOR lines increases only in an additive way. Since the ENDOR spectral lines are comparable in width to EPR lines, the reduced number of lines in the ENDOR spectrum results in a much greater effective resolution. Therefore, accurate values of the hyperfine couplings may be obtained from an ENDOR experiment even under conditions where the hyperfine pattern is not resolved in the EPR spectrum. In addition, ENDOR spectra become easier to interpret when there are nuclei with different magnetic moments involved. Their ENDOR lines normally appear in different frequency ranges and, from their Larmor frequencies, these nuclei can be immediately identified. ENDOR is also a well justified method when anisotropic hyperfine and nuclear quadrupole (for nuclei with $I \geq 1$) couplings in solids are to be measured. As an example, the ENDOR spectrum of the perinaphtheryl radical in liquid solution is depicted in figure B1.15.9 (see also figure B1.15.6 for a comparison with the CW EPR spectrum).

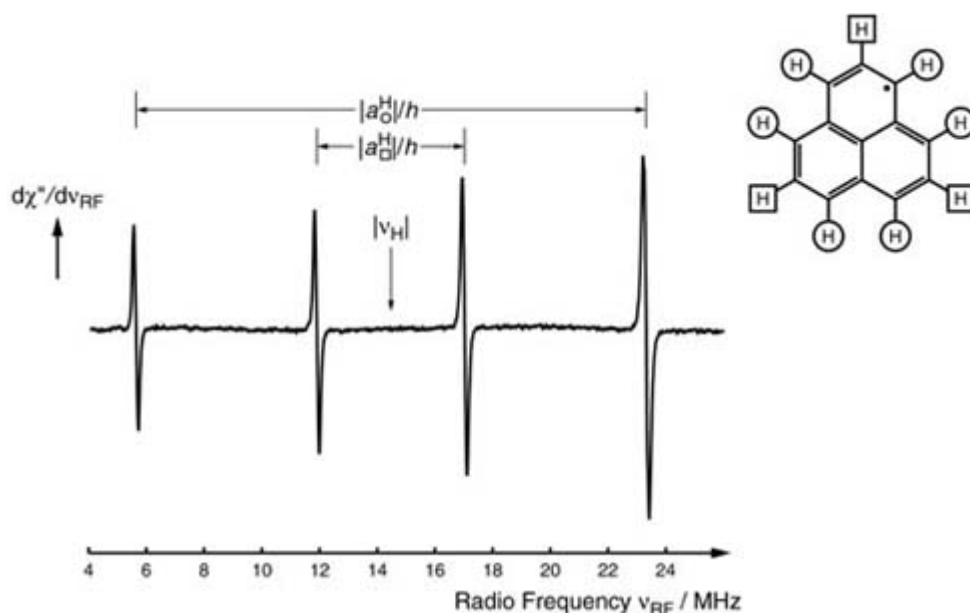


Figure B1.15.9. The ENDOR spectrum of the perinaphtheryl radical in mineral oil taken at room temperature. $|a_{\bigcirc}^H|/h = 17.68$ MHz and $|a_{\square}^H|/h = 5.12$ MHz are the hyperfine coupling constants for the protons in the position \bigcirc and \square , respectively. $|a_{\bigcirc}^H|/(g\beta_e) = 0.631$ mT and $|a_{\square}^H|/(g\beta_e) = 0.183$ mT.

The ENDOR experiment is performed at a constant external magnetic field by applying MW and RF fields in a continuous fashion. This technique is called CW ENDOR spectroscopy. The design of an ENDOR spectrometer differs only slightly from a basic CW or pulsed EPR spectrometer. A coil located around the sample tube within the resonant EPR cavity is used as an element in a RF transmitter circuit and is the source of a RF field. The basic elements of the RF circuit include a low-power signal source or sweeper and a high-power amplifier to produce an RF output signal that can be scanned over a wide frequency range (e.g. for proton ENDOR at X-band from approximately 4 to 30 MHz) at a power level up to 1 kW. To carry out an ENDOR experiment, the magnetic field B_0 is set at the resonance of one of the observed EPR transitions.

Then, the MW power is increased in order to partially saturate the EPR transition. The degree of saturation is provided by the saturation factor s defined earlier (see [equation \(b1.15.13\)](#)).

-26-

Finally, a strong RF oscillating field of varying frequency is applied to induce and saturate a transition within the nuclear sublevels. When the resonance condition for nuclear transitions is fulfilled, the saturated EPR transition can be desaturated by the ENDOR transition provided both transitions have energy levels in common. This desaturation of the EPR transition is detected as a change in the EPR absorption at characteristic frequencies ω_{ENDOR} , and constitutes the ENDOR response.

Phenomenologically, the ENDOR experiment can be described as the creation of alternative relaxation paths for the electron spins, which are excited with microwaves. In the four-level diagram of the $S=I=\frac{1}{2}$ system described earlier (see [figure B1.15.8](#)) relaxation can occur via several mechanisms: W_{1e} and W_{1n} describe the relaxation rates of the electron spins and nuclear spins, respectively. W_{x1} and W_{x2} are cross-relaxation rates in which electron and nuclear spin flips occur simultaneously. Excitation, for example, of the EPR transition $|-\rangle \leftrightarrow |+\rangle$ (i.e. $4 \leftrightarrow 2$) will equalize the population of both levels, 4 and 2, if the direct relaxation (characterized by the relaxation rate W_{1e}^{24}) cannot compete with the transitions induced by the resonant microwaves.

Simultaneous application of an RF field at a frequency corresponding to the $|+\rangle \leftrightarrow |-\rangle$ (i.e. $2 \leftrightarrow 1$) transition then opens a relaxation path via W_{1e}^{13} and W_{1n}^{34} or, more directly, via W_{x1} . The extent to which these relaxation bypasses can compete with the direct W_{1e}^{24} route controls the degree of desaturation of the EPR line, and, therefore, determines the ENDOR signal intensity, which, consequently, does not generally reflect the number of contributing nuclei (in contrast to EPR and NMR). The signal intensity observed depends very critically on the balance between the various relaxation rates and the magnitude of the MW and RF fields, B_1 and B_2 , respectively. Additional parameters to be varied in order to optimize the ENDOR signal-to-noise ratio are the radical concentration, the solvent viscosity and the temperature. The amplitude of ENDOR signals is furthermore influenced by the enhancement effect which occurs because the nucleus does not only experience the time-dependent magnetic field B_2 at the RF, but also an additional magnetic field component (the hyperfine field) due to the magnetic moment of the electron. Therefore, the effective field at the nucleus can be described as

$$B_2^{\text{eff}} = \kappa B_2 \quad (\text{B1.15.37})$$

where κ is the hyperfine enhancement factor. For isotropic HFI $\kappa = |1 - M_{\text{Sa}} / (\hbar \omega_N)|$. The hyperfine enhancement is one reason for the different intensities of the individual lines of an ENDOR line pair; at the same RF power the high-frequency line is usually more intense than the low-frequency one. Another reason for asymmetrical ENDOR line patterns is the effectiveness of the cross-relaxation paths: W_{x1} is in general different from W_{x2} , thus leading to an asymmetrical relaxation network and, as a consequence, to unequal signal intensities. In spectra of single crystals, powders and noncrystalline solids, however, the enhancement factor is governed by different hyperfine tensor components. This often leads to unexpected intensity patterns within ENDOR line pairs.

Despite the increased resolution of ENDOR compared to EPR, some restrictions concerning the information contents of ENDOR spectra persist: (1) unlike EPR, the relative signal intensities of ENDOR line pairs belonging to different groups of nuclei do not give any indication of the relative number of nuclei belonging to the individual groups; (2) the ENDOR spectrum does not give the sign of the hyperfine couplings, that is, one does not know which ENDOR transition belongs to which electron spin state. Both problems are addressed in triple resonance, which can be seen as an extension to ENDOR spectroscopy. Therefore, triple resonance experiments are very often carried out in order to supplement ENDOR data.

B1.15.5.2 ELECTRON–NUCLEAR–NUCLEAR TRIPLE RESONANCE

A refinement of the ENDOR experiment is electron–nuclear–nuclear triple resonance, now commonly denoted TRIPLE. In TRIPLE experiments one monitors the effect of a simultaneous excitation of two nuclear spin transitions on the level of the EPR absorption. Two versions, known as special TRIPLE (ST) and general TRIPLE (GT), are routinely performed on commercially available spectrometers.

(A) SPECIAL TRIPLE (ST)

The special TRIPLE technique [25, 26] is used for hyperfine couplings $|a|/(2h) < |\omega_N|$. In a typical experiment, RF is generated at two frequencies: one is fixed at the free nuclear frequency ω_N appropriate to the sort of nuclei under scrutiny and the second is swept. These two frequencies are multiplied to obtain the sum and the difference frequencies, $\omega_N \pm \omega_{RF2}$, which are used to irradiate the sample. The experiment can be understood using the energy level and relaxation scheme of figure B1.15.8. Both ENDOR transitions, ω_{ENDOR}^{\pm} (i.e. transitions $3 \leftrightarrow 4$ and $1 \leftrightarrow 2$), associated with the same nucleus are simultaneously excited. In cases of vanishing cross-relaxation the second saturating RF field enhances the efficiency of the relaxation bypass, thus increasing the signal intensity, particularly in cases where W_N is the rate-limiting step (because $W_n \ll W_c$). A second advantage of ST resonance over ENDOR is that when both RF fields are sufficiently strong to completely saturate nuclear transitions the EPR desaturation becomes independent of W_N . Consequently, the line intensities are no longer determined by the relaxation behaviour of the various nuclei, but rather reflect the number of nuclei involved in the transition. Finally, ST also has the advantage of higher resolution because the effective saturation of nuclear transitions results in smaller observed linewidths compared to ENDOR.

(B) GENERAL TRIPLE (GT)

In a general TRIPLE (GT) experiment one particular ENDOR transition is pumped with the first RF while the second RF is scanned over the whole range of nuclear resonances [27]. Therefore, nuclear transitions of different sets of nuclei of the same kind or of different kinds are saturated simultaneously and the effect on the ENDOR transitions for all the hyperfine couplings in the system is measured. Clearly, ST is included within GT. From the characteristic intensity changes of the high-frequency and low-frequency signals compared with those of the ENDOR signals the relative signs of the hyperfine coupling constants can easily be determined.

B1.15.5.3 ELECTRON–ELECTRON DOUBLE RESONANCE (ELDOR)

ELDOR is the acronym for electron–electron double resonance. In an ELDOR experiment [28] one observes a reduction in the EPR signal intensity of one hyperfine transition that results from the saturation of another EPR transition within the spin system. ELDOR measurements are still relatively rare but the experiment is firmly established in the EPR repertoire.

With help of the four-level diagram of the $S=I=\frac{1}{2}$ system (see figure B1.15.8) two common ways for recording ELDOR spectra will be illustrated. In frequency-swept ELDOR the magnetic field is set at a value that satisfies the resonance condition for one of the two EPR transitions, e.g. $4 \leftrightarrow 2$, at the fixed observe klystron frequency, ω_{obs} . The pump klystron is then turned on and its frequency, ω_{pump} , is swept. When the pump frequency passes through the value

that satisfies the resonance condition of the $3 \leftrightarrow 1$ transition, there is a decrease in the signal at the frequency ω_{obs} which constitutes the ELDOR signal. In field-swept ELDOR the pumping and observing MW frequencies are held fixed at predetermined values and the magnetic field is swept through the region of resonance. ELDOR experiments are technically more difficult than ENDOR: simultaneous EPR in one magnetic field for two different transitions requires irradiation simultaneously at two MW frequencies. That is, one requires a resonator tunable to two MW frequencies separated by a multiple of the hyperfine coupling. The development of loop-gap and split-ring resonators has, because of their wide bandwidth and the feasibility of high filling factors, made ELDOR a truly practical technique.

To analyse ELDOR responses, the reduction of the observed EPR transition at ω_{obs} is expressed quantitatively in terms of the ELDOR reduction factor

$$R = \frac{I(\omega_{\text{pump}} = 0) - I(\omega_{\text{pump}})}{I(\omega_{\text{pump}} = 0)} \quad (\text{B1.15.38})$$

where $I(\omega_{\text{pump}}=0)$ is the observed EPR intensity with pump power off, and $I(\omega_{\text{pump}})$ is the intensity with pump power on. The ELDOR technique is very sensitive to the various relaxation mechanisms involved. For the $S=I=\frac{1}{2}$ system R may be expressed in terms of the six relaxation rates between the four energy levels that are indicated in [figure B1.15.8](#). With the assumption $W_e^{13}=W_e^{24}=W_e$ and $W_N^{12}=W_N^{34}=W_N$ the ELDOR reduction factor is given by

$$R = \frac{W_n^2 - W_{x1} W_{x2}}{W_c(2W_n + W_{x1} + W_{x2}) + (W_n + W_{x1})(W_n + W_{x2})} \quad (\text{B1.15.39})$$

which shows that the ELDOR response will be a reduction if $W_N^2 > W_{x1} W_{x2}$. If modulation of dipolar hyperfine couplings is the dominant relaxation mechanism, this condition can be fulfilled for dilute radical concentrations at low temperatures. At high concentrations or sufficiently high temperatures, Heisenberg spin exchange or chemical exchange, which tends to equalize the population of all spin levels, is the dominant ELDOR mechanism.

ELDOR has been employed to study a number of systems such as inorganic compounds, organic compounds, biologically important compounds and glasses. The potential of ELDOR for studying slow molecular motions has been recognized by Freed and coworkers [29, 30].

B1.15.6 PULSED EPR SPECTROSCOPY

By far the greatest advantage of pulsed EPR [31, 32] lies in its ability to manipulate the spin system nearly at will and, thus, to measure properties that are not readily available from the CW EPR spectra. Nevertheless, EPR has long remained a domain of CW methods. In contrast to the rapid development of pulsed NMR spectroscopy, the utilization of the time domain in EPR took a much longer time, even though the underlying principles are essentially the same. There are several reasons for this slow development of pulsed EPR. (1) The large energies involved in electron spin interactions (see [figure B1.15.3](#)) can give rise to spectral widths of the order of 10–25% of the carrier frequency

(at X-band) as opposed to the ppm scale which applies to NMR. Consequently, with the exception of some organic radicals in solution and a few defect centres in single crystals, it is technically impossible to excite the

entire EPR spectrum by a pulse of electromagnetic radiation. (2) CW EPR records derivatives of absorption spectra by using magnetic field modulation in a range between 10 kHz and 100 kHz, a method that takes advantage of narrowband detection at the modulation frequency (see [figure B1.15.5](#)) and of better resolution of the derivative as compared to the absorption lineshape (see [figure B1.15.6](#)). Calculating the derivative from the absorption lineshape obtained with pulsed methods results in a decrease in the signal-to-noise ratio. For these reasons, CW EPR is typically more sensitive than pulsed EPR at a given resolution. (3) Finally, the fact that electron spin relaxation times are orders of magnitudes shorter than the nuclear spin relaxation times encountered in NMR makes the technology required to perform pulsed EPR experiments much more demanding.

In recent years, however, enormous progress has been made and with the availability of the appropriate MW equipment pulsed EPR has now emerged from its former shadowy existence. Fully developed pulse EPR instrumentation is nowadays commercially available [31, 33].

The practical goal for pulsed EPR is to devise and apply pulse sequences in order to isolate pieces of information about a spin system and to measure that information as precisely as possible. To achieve this goal it is necessary to understand how the basic instrumentation works and what happens to the spins during the measurement.

B1.15.6.1 PULSES AND THEIR EFFECTS

For an understanding of pulsed excitation of spin ensembles it is of fundamental importance to realize that radiation pulses actually contain ranges of frequencies: A burst of monochromatic microwaves at frequency ω_{MW} and of pulse duration t_p translates into a frequency spectrum of the pulse that has field components at all frequencies. The amplitude of the field drops off as one moves away from the carrier frequency ω_{MW} according to $B_1(\omega) \propto B_1 \sin(\omega_{\text{MW}t_p})/(\omega_{\text{MW}t_p})$. The excitation bandwidth of a specific pulse depends only on the pulse duration, t_p .

The effect of an MW pulse on the macroscopic magnetization can be described most easily using a coordinate system (x', y', z) which rotates with the frequency ω_{MW} about the z -axis defined by the applied field \mathbf{B} . Initially, the net magnetic moment vector \mathbf{M} is in its equilibrium position oriented parallel to the direction of the strong external field. In the rotating frame, \mathbf{B}_1 is a stationary field, which is assumed to be oriented parallel to the x' -axis of the rotating coordinate system. The result of applying a short intense MW pulse is to rotate the magnetization \mathbf{M} about the axis defined by \mathbf{B}_1 , i.e. the x' -axis, through the flip angle $\theta = \gamma_e B_1 t_p = \omega_1 t_p$, expressed in radians. When the duration of the MW radiation at a given MW power level is just long enough to flip \mathbf{M} into the $x'y'$ -plane, the pulse is defined as a $\pi/2$ -pulse. Immediately after the cessation of the pulse, \mathbf{M} has been rotated with its magnitude unaltered (if relaxation phenomena are negligible during the excitation pulse) to an orientation perpendicular to \mathbf{B}_1 at angle θ with respect to $B_0 \parallel z$. After this perturbation the system is then allowed to return to its equilibrium, or, after an appropriate delay, additional pulses with specific flip angles and phases are applied to further manipulate the spin system. With suitable apparatus, that is, the detection system aligned in the direction of the y' -axis of the rotating axis system, the temporal behaviour of the y' -component of the magnetization can be followed. The normalized FT of the function $M_{y'}(t)$ provides the lineshape which is analogous to that obtained from CW EPR experiments under nonsaturating conditions.

B1.15.6.2 INSTRUMENTATION

The design of a pulsed EPR spectrometer depends heavily on the required pulse length and pulse power which in turn are mainly dictated by the relaxation times of the paramagnetic species to be studied, but also by the type of experiment performed. When pulses of the order of a few nanoseconds are required (either to compete

with the relaxation times or to excite a broad spectral range) not only is high MW power needed to fulfill the condition $\gamma_e B_1 t_p = \pi/2$, but also the whole design of such a high power spectrometer becomes much more complex and the construction is more expensive. In most of today's pulsed EPR spectrometers the MW pulses are formed on a low-power level by fast switching diodes after the CW source. These low-power pulses are then fed into a pulsed high-power MW amplifier (typically a travelling-wave tube amplifier) capable of giving the requisite high power up to a few kW. The amplified pulses are then directed into a resonant cavity to excite the EPR transitions of the system. The MW power in the resonator grows as $[1 - \exp(-\omega_{MW}t/Q)]$ and decays as $\exp(-\omega_{MW}t/Q)$ in response to a square, resonant pulse. An important consideration in a pulsed EPR spectrometer is the detection deadtime, or how soon after a pulse the signal $M_{y'}(t)$ can be measured. Typically, the deadtime is taken to be the time when ringing from the resonator equals thermal noise. The choice of an appropriate resonator for pulsed EPR experiments is therefore always influenced by the conflicting demands for a short deadtime and good sensitivity: a low quality factor Q for bandwidth coverage and fast instrumental response should be combined with concentrated MW fields for short pulses and high filling factor, the latter in partial compensation for the loss in sensitivity in favour of fast time resolution. The spin system responds to the exciting MW pulses by producing a signal at a later time when the incident pulses are off. A low-noise amplifier amplifies the signal to a level well above the noise floor of the detector. Standard Schottky barrier diodes can be used as detectors up to a bandwidth of 5 MHz. For broader bandwidths, multiplying mixers can be employed to downconvert the signal from the sample to a video signal centred at zero frequency. (The mixer output is the sum or the difference of two input frequencies and the signal amplitudes are proportional to the input amplitudes.) A quadrature mixer has two inputs, one for the signal and one for a reference from the master MW oscillator. Its two outputs are in quadrature with each other: one is out of phase by 90° and the other in phase with the pulse phase. Quadrature detection means that, unlike the case in other spectrometers, the reference arm needs no phase shifter since the phase of the recorded signal can be adjusted digitally by taking a linear combination of the two quadrature components. Typically, the low-noise MW amplifier and the mixer detector are very sensitive to high-power reflections from the cavity and, therefore, have to be protected during the excitation pulse. A gated PIN-diode switch in front of the amplifier strongly attenuates the input of the detection system during the excitation pulses and, therefore, avoids saturation or permanent damage.

B1.15.6.3 PULSE EPR METHODS

(A) FOURIER TRANSFORM EPR (FT EPR)

All operating principles are the same as in FT NMR. A single short and intense MW pulse (typically a $\pi/2$ -pulse along x') is applied to flip the magnetizations into the $x'y'$ -plane of the rotating frame (see [figure B1.15.10\(A\)](#)). The induced signal proportional to $M_{y'}$, will decay due to transverse relaxation or sample inhomogeneities. This process is called free induction decay (FID). The complete spectrum is obtained without the need of a field sweep via FT of the FID. Under most conditions, the FT EPR spectrum measured using a single excitation pulse corresponds exactly to the CW EPR spectrum. Today's state-of-the-art pulsed EPR spectrometers feature $\pi/2$ -pulse lengths of $t_p \approx 5$ ns or less, corresponding to an excitation bandwidth of roughly 200 MHz. Therefore, FT EPR is applicable to not too wide spectral patterns consisting of narrow lines (as typical for free radicals in solution) with long enough T_2 so that the

FID does not die away before the deadtime has elapsed. In the case of inhomogeneously broadened EPR lines (as typical for free radicals in solids) the dephasing of the magnetizations of the individual spin packets (which all possess slightly different resonance frequencies) will be complete within the detection deadtime and, therefore, the FID signal will usually be undetectable.

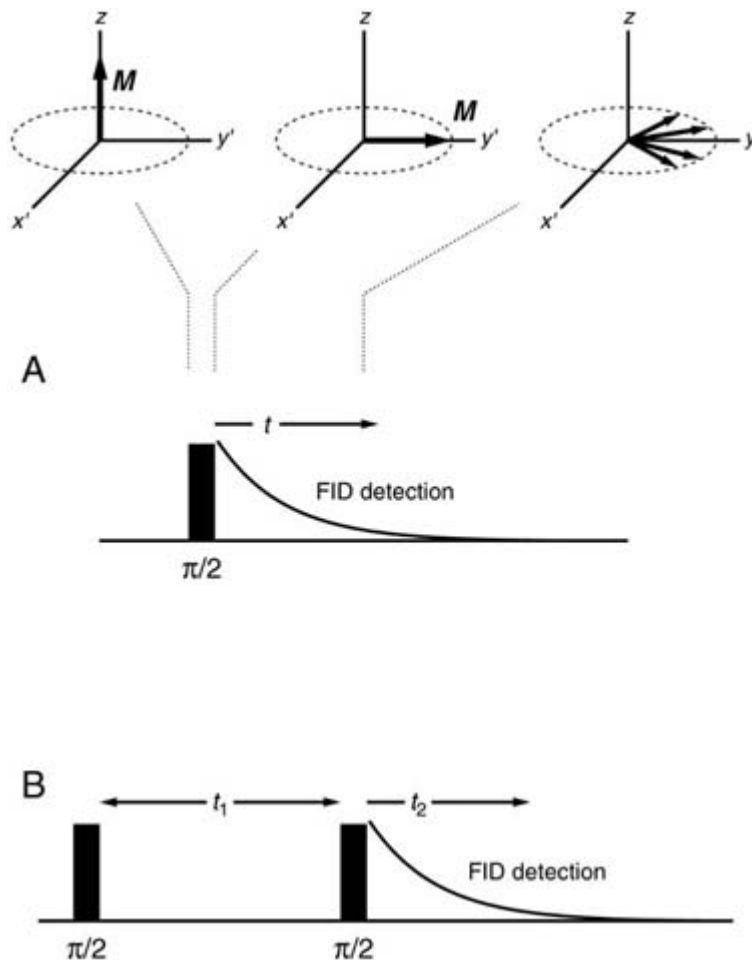


Figure B1.15.10. FT EPR. (A) Evolution of the magnetization during an FT EPR experiment (rotating frame representation). (B) The COSY FT EPR experiment.

In complete analogy to NMR, FT EPR has been extended into two dimensions. Two-dimensional correlation spectroscopy (COSY) is essentially subject to the same restrictions regarding excitation bandwidth and detection deadtime as was described for one-dimensional FT EPR. In 2D-COSY EPR a second time dimension is added to the FID collection time by a preparatory pulse in front of the FID detection pulse and by variation of the evolution time between them (see figure B1.15.10(B)). The FID is recorded during the detection period of duration t_2 , which begins with the second $\pi/2$ -pulse. For each t_1 the FID is collected, then the phase of the first pulse is advanced by 90° , and a second set of FIDs is collected. The two sets of FIDs, whose amplitudes oscillate as functions of t_1 , then undergo a two-dimensional complex Fourier transformation, generating a spectrum over the two frequency variables ω_1 and ω_2 .

The peaks along the leading diagonal $\omega_1 = \omega_2$ correspond to the usual absorption spectrum, whereas the cross-peaks (peaks removed from the diagonal) provide evidence for cross-correlations.

(B) ELECTRON-SPIN ECHO (ESE) METHODS

Under conditions where the rapid decay of the FID following a single excitation pulse is governed by inhomogeneous broadening the dephasing of the individual spin packets in the $x'y'$ -plane can be reversed by the application of a second MW pulse in an ESE experiment (see figure B1.15.11) [34]. As before, the experiment begins with the net electron spin magnetization \mathbf{M} aligned along the magnetic field direction z . At the end of the first $\pi/2$ -pulse, which is applied at the Larmor precession frequency ω_0 , with the amplitude B_1

pointing along the x' -direction, the net magnetic moment is in the equatorial plane. Immediately, the magnetization starts to decay as different spin packets precess about z at their individual Larmor frequencies $\omega_i \neq \omega_0$. In the rotating frame the contributions of different spin packets to the magnetization MD_y , appear to fan out as shown in [figure B1.15.11\(A\)](#) : viewed in the laboratory frame some spins would appear to precess faster ($\omega_i > \omega_0$) and some slower ($\omega_i < \omega_0$) than the average. As a result, the FID decays rapidly and after a short time there is no detectable signal. From this FID an echo can be generated by means of a second MW pulse, applied at time τ after the first pulse. The second pulse is just long enough to turn the magnetization vectors through 180° about the x' -axis. The original precession frequencies and the directions of rotation of the individual components will remain unaltered and, therefore, the magnetizations will rotate toward each other in the $x'y'$ -plane until they refocus after the same time τ into a macroscopic magnetic moment along the $-y'$ -axis. At this point the spin alignment produces a microwave field B_1 in the cavity corresponding to an emission signal that is referred to as an echo [34].

As the spins precess in the equatorial plane, they also undergo random relaxation processes that disturb their movement and prevent them from coming together fully realigned. The longer the time τ between the pulses the more spins lose coherence and consequently the weaker the echo. The decay rate of the two-pulse echo amplitude is described by the phase memory time, T_M , which is the time span during which a spin can remember its position in the dephased pattern after the first MW pulse. T_M is related to the homogeneous linewidth of the individual spin packets and is usually only a few microseconds, even at low temperatures.

The two-pulse sequence $\pi/2-\tau-\pi-\tau$ is not the only sequence which leads to the formation of an echo. A pulse sequence which has proven to have particular value consists of three $\pi/2$ -pulses as depicted in [figure B1.15.11\(B\)](#). In this three-pulse sequence with pulse intervals τ and T a so-called stimulated echo is formed after an interval τ following the third pulse. The mechanism of formation of the stimulated echo is a little more complicated than that of the primary echo and the reader is referred to some excellent review articles [32, 35, 36] for a comprehensive discussion of this topic. Here it is sufficient to mention that with the second $\pi/2$ -pulse the y' -components of the dephased magnetization pattern are temporarily stored in the $x'z'$ -plane where they remain during the waiting time T . The third MW pulse brings the M_z magnetizations back into the $x'y'$ -plane, where they continue their time evolution and give rise to the stimulated echo at time τ after the third pulse. The characteristic time of the three-pulse echo decay as a function of the waiting time T is much longer than the phase memory time T_M (which governs the decay of a two-pulse echo as a function of τ), since the phase information is stored along the z -axis where it can only decay via spin–lattice relaxation processes or via spin diffusion.

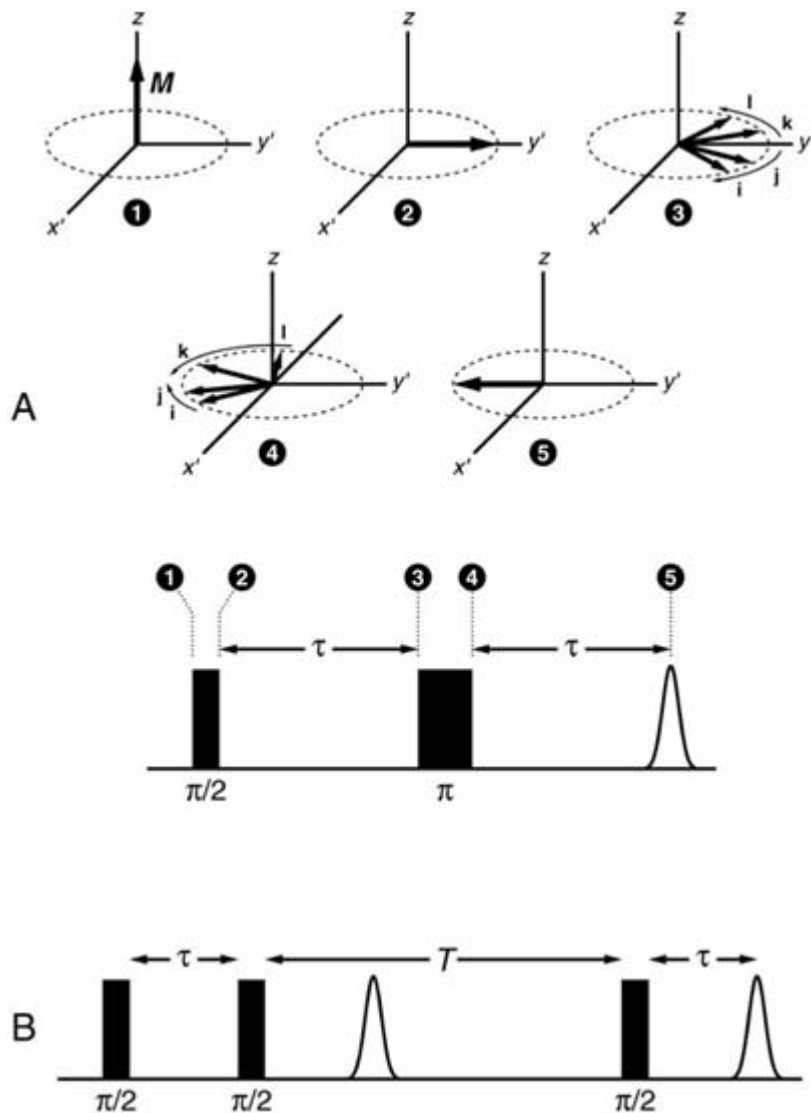


Figure B1.15.11. Formation of electron spin echoes. (A) Magnetization of spin packets i, j, k and l during a two-pulse experiment (rotating frame representation). (B) The pulse sequence used to produce a stimulated echo. In addition to this echo, which appears at τ after the third pulse, all possible pairs of the three pulses produce primary echoes. These occur at times 2τ , $2(\tau+T)$ and $(\tau+2T)$.

In electron-spin-echo-detected EPR spectroscopy, spectral information may, in principle, be obtained from a Fourier transformation of the second half of the echo shape, since it represents the FID of the refocused magnetizations, however, now recorded with much reduced deadtime problems. For the inhomogeneously broadened EPR lines considered here, however, the FID and therefore also the spin echo, show little structure. For this reason, the amplitude of the echo is used as the main source of information in ESE experiments. Recording the intensity of the two-pulse or three-pulse echo amplitude as a function of the external magnetic field defines electron-spin-echo- (ESE-)

detected EPR spectroscopy. Such a field-swept ESE spectrum is similar to the conventional CW EPR spectrum except for the fact that the lines appear in absorption and not in the more familiar first derivative form.

ESE-detected EPR spectroscopy has been used advantageously for the separation of spectra arising from different paramagnetic species according to their different echo decay times. Furthermore, field-swept ESE

spectroscopy is superior to conventional CW EPR when measuring very broad spectral features. This is because the field modulation amplitudes used in CW EPR to detect the first derivative of the signal are often too small compared to the width of the EPR line, so that the gradient of the absorption signal becomes very small. Such features are either invisible in CW EPR or obscured by baseline drifts, while they can be well distinguished in the absorption spectra with ESE-detected EPR.

In electron spin echo relaxation studies, the two-pulse echo amplitude, as a function of the pulse separation time τ , gives a measure of the phase memory relaxation time T_M from which T_2 can be extracted if T_1 -effects are taken into consideration. Problems may arise from spectral diffusion due to incomplete excitation of the EPR spectrum. In this case some of the transverse magnetization may leak into adjacent parts of the spectrum that have not been excited by the MW pulses. Spectral diffusion effects can be suppressed by using the Carr–Purcell–Meiboom–Gill pulse sequence, which is also well known in NMR. The experiment involves using a sequence of π -pulses separated by 2τ and can be denoted as $[\pi/2-(\tau-\pi-\tau\text{-echo})_n]$. A series of echoes separated by 2τ is generated and the decay in their amplitudes is characterized by T_M .

The other important (spin–lattice) relaxation time T_1 is accessible with the help of an additional preparation, e.g. an inversion (i.e. a π -pulse) or saturation pulse (a single long pulse or a chain of short $\pi/2$ -pulses) placed with a variable delay time T in front of a two-pulse ESE sequence. The T_1 -information is then extracted from the dependence of the echo amplitude on the interval T . Experiments of this type are generally called ‘inversion recovery’ or ‘saturation recovery’ experiments. In principle T_1 can also be estimated from the amplitude of the stimulated echo, as the z -magnetization relaxes towards thermal equilibrium during the variable pulse delay time T . Here, inaccuracies in measuring T_1 may again originate from spectral diffusion and the interaction between the electron spin and the nuclear spins which can affect the amplitude of the echo.

The electron-spin echo envelope modulation (ESEEM) phenomenon [37, 38] is of primary interest in pulsed EPR of solids, where anisotropic hyperfine and nuclear quadrupole interactions persist. The effect can be observed as modulations of the echo intensity in two-pulse and three-pulse experiments in which τ or T is varied. In liquids the modulations are averaged to zero by rapid molecular tumbling. The physical origin of ESEEM can be understood in terms of the four-level spin energy diagram for the $S = I = \frac{1}{2}$ model system introduced earlier to describe ENDOR (see [figure B1.15.8](#)). So far, however, only isotropic hyperfine couplings have been considered, leading to an EPR spectrum of this system that comprises the two allowed transitions $1 \leftrightarrow 3$ and $2 \leftrightarrow 4$ with $\Delta M_S = \pm 1$ and $\Delta M_I = 0$. The situation is different for the case where the hyperfine couplings are anisotropic and, in particular, of the same order of magnitude as the nuclear Zeeman couplings. Because of the anisotropic nature of the interactions, the energy levels of the spin system are modified and the nuclear spin states are mixed. As a consequence, the transitions $1 \leftrightarrow 4$ and $2 \leftrightarrow 3$ involving a simultaneous nuclear spin transition (both forbidden for the isotropic case) are now also allowed to some extent. In [figure B1.15.12\(A\)](#) the evolution of this spin system in a two-pulse echo experiment (see [figure B1.15.11\(A\)](#)) is considered. Since there are four transitions, there are four components of the magnetization to keep track of. For simplicity, only the magnetizations of two transitions, ω_{24} and ω_{14} (labelled \mathbf{a} and \mathbf{f} , respectively), originating from the same nuclear spin level in the lower electron spin manifold, are considered. By applying sufficiently short MW pulses both allowed and forbidden transitions are excited simultaneously. After the first $\pi/2$ -pulse, the two sets of electrons

are precessing at two different frequencies separated by the nuclear frequency. If spin packet \mathbf{a} is on resonance ($\omega_{24} = \omega_{MW}$), its component of the magnetization is fixed in the rotating frame, whereas the magnetization component \mathbf{f} precesses with frequency $\omega_{14} - \omega_{MW} = \omega_{12}$. After the time τ , the π -pulse inverts the vector \mathbf{a} into the $-y'$ -direction, and at the same time, because of the branching of transitions, gives rise to a new component \mathbf{f}' . The effect of the π -pulse on the magnetization component \mathbf{f} is a rotation about x' by 180° and the formation of a component \mathbf{a}' according to the ratio of the transition probabilities for allowed and forbidden transitions. \mathbf{a} and \mathbf{a}' will remain unaltered in the rotating frame (because they are on resonance),

whereas the two vectors \mathbf{f} and \mathbf{f}' continue to precess with the off-resonance frequency ω_{12} . At time τ , \mathbf{f} will refocus with \mathbf{a} at the $-y'$ -axis to form an echo, but the vectors \mathbf{a}' and \mathbf{f}' will not contribute, because they are no longer oriented along $-y'$. This results in a reduction of the echo intensity at time τ . The component \mathbf{f}' will only contribute to the echo if, in time τ , it precesses an integral number of times in the $x'y'$ -plane. Therefore, the echo amplitude oscillates in proportion to $\cos(\omega_{12}\tau)$. The same holds for any combination of transitions with energy levels in common. Therefore, one expects the echo intensity to oscillate not only with the nuclear frequencies ω_{12} and ω_{34} but also with the sum and the difference of these frequencies. By FT of the echo envelope an ENDOR-like spectrum is obtained. The amplitudes of the modulation frequencies are determined by the depth parameter $k=4I_a I_f / (B\omega_N / (\omega_{12}\omega_{34}))^2$, where I_a and I_f denote the intensities of the allowed and forbidden transitions, respectively. B/ω_N is a measure of the anisotropy of the hyperfine coupling tensor. Large modulation amplitudes are expected for $\omega_{12}, \omega_{34} \rightarrow 0$. This is in contrast to ENDOR spectroscopy, where the enhancement factor and, therefore, the ENDOR line intensities, decrease for small nuclear transition frequencies. For small hyperfine coupling constants $\omega_{12} \approx \omega_{34} \approx \omega_N$ and $k \propto (B/\omega_N)^2$. Again in contrast to ENDOR, the ESEEM modulation depth will increase for nuclei with smaller γ_n .

The stimulated (three-pulse) echo decay may also be modulated, but only by the nuclear frequencies ω_{12} and ω_{34} and not by their sum and difference frequencies. The qualitative reason for this is that the first pulse generates modulation at the nuclear frequencies; the second pulse additionally incorporates the sum and difference frequencies and the third pulse causes interference of the sum and difference frequencies to leave only the ENDOR frequencies. Apart from the depth parameter k the modulation amplitudes of the ENDOR frequencies ω_{12} and ω_{34} are determined by $\sin^2(\omega_{34}\tau/2)$ and $\sin^2(\omega_{12}\tau/2)$, respectively. As a consequence, so-called blind spots can occur. For example, if ω_{34} is an integral multiple of π , i.e. $\tau=2\pi n/\omega_{34}$, then the modulation at ω_{12} is completely suppressed. Therefore, the dependence on the response on τ should also be examined.

The main advantage of the three-pulse ESEEM experiment as compared to the two-pulse approach lies in the slow decay of the stimulated echo intensity determined by T_1 , which is usually much longer than the phase memory time T_M that limits the observation of the two-pulse ESE.

More sophisticated pulse sequences have been developed to detect nuclear modulation effects. With a five-pulse sequence it is theoretically possible to obtain modulation amplitudes up to eight times greater than in a three-pulse experiment, while at the same time the unmodulated component of the echo is kept close to zero. A four-pulse ESEEM experiment has been devised to greatly improve the resolution of sum-peak spectra.

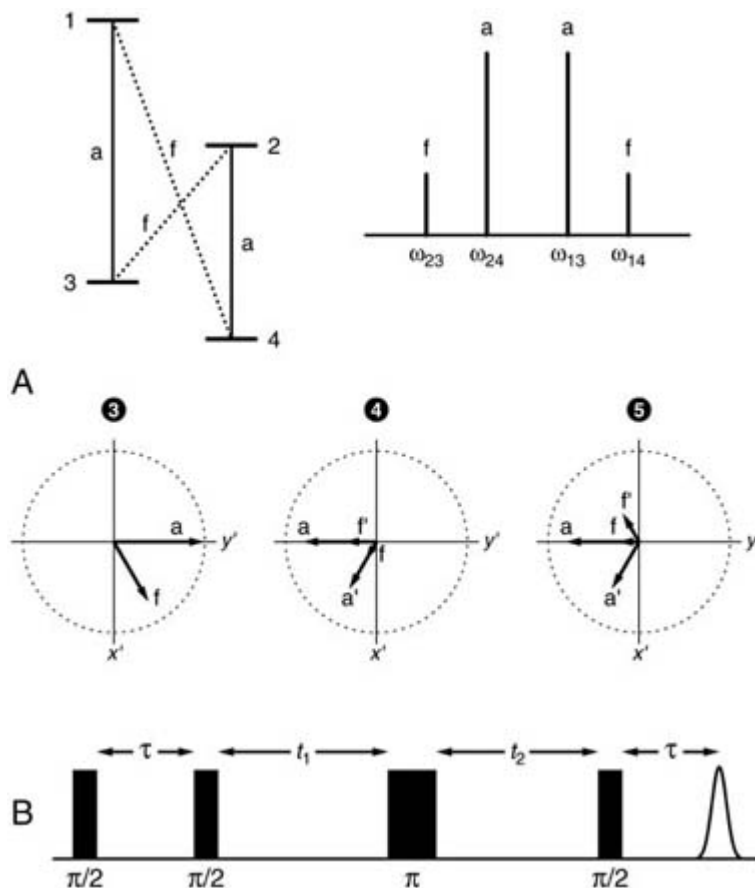


Figure B1.15.12. ESEEM spectroscopy. (A) Top: energy level diagram and the corresponding stick spectrum for the two allowed (**a**) and two forbidden (**f**) transitions. Bottom: time behaviour of the magnetization of an allowed (**a**) spin packet and a forbidden (**f**) spin packet during a two-pulse ESE sequence (see [figure B1.15.11 \(A\)](#)). (B) The HYSORE pulse sequence.

In the 2D three-pulse ESEEM technique both the time intervals $t_1 = \tau$ and $t_2 = T$ of a stimulated echo sequence are independently increased in steps (see [figure B1.15.11\(B\)](#)) [39]. If the spacing between the first two pulses, τ , is varied over a sufficiently broad range, blind spots, which caused problems in one-dimensional spectra, do not arise in this 2D ESEEM method. Combination cross peaks arising from couplings with several inequivalent nuclei can be used to determine the relative signs of the hyperfine splittings. A disadvantage of the 2D three-pulse ESEEM technique is that the echo intensities decay at different rates along the two time axes, with T_M -relaxation along the t_1 -axis and T_1 -relaxation along the t_2 -axis. As a result, the linewidths in the two frequency dimensions can differ by orders of magnitude.

An alternative 2D ESEEM experiment based on the four-pulse sequence depicted in [figure B1.15.12\(B\)](#) has been proposed by Mehring and coworkers [40]. In the hyperfine sublevel correlation (HYSORE) experiment, the decay of the echo intensity as a function of t_1 is governed by T_1 -relaxation, whereas the echo decay along the t_2 -axis is

determined by the T_2 -relaxation of the nuclei. Since both relaxation processes are fairly slow, the resolution along both frequency dimensions is much increased compared to the 2D three-pulse ESEEM experiment. In the HYSORE experiment, too, the positions of the cross-peaks can be used to determine the relative signs of the hyperfine coupling constants.

The ESEEM methods are best suited for the measurement of small hyperfine couplings, e.g. for the case of

nuclear spins with small magnetic moment. Larger hyperfine interactions can be measured best by pulsed versions of ENDOR spectroscopy. These methods will be introduced as a final application of the pulsed excitation scheme introduced earlier. Pulsed ENDOR methods are double-resonance techniques wherein, at some particular time in an ESE pulse sequence, a RF pulse is applied that is swept in frequency to match resonance with the hyperfine-coupled nuclei. The typical pulse schemes for the most commonly used versions of pulsed ENDOR, termed Mims- [41] and Davies-type [42] ENDOR to acknowledge those who originally introduced them, are depicted in [figure B1.15.13](#). In both experiments the ENDOR effect is manifested in a change of the ESE intensity when the RF field is on nuclear resonance. The ENDOR spectrum can thus be recorded by detecting the echo amplitude as a function of the frequency of the RF pulse.

The Davies-ENDOR technique is based on an inversion recovery sequence (see [figure B1.15.13\(A\)](#)). The experiment starts by interchanging the populations of levels 1 and 3 of one of the EPR transitions of the $S=I=\frac{1}{2}$ model spin system by means of a first selective MW π -pulse (with a strength $|\omega_1| \ll |A|/\hbar$). Neglecting relaxation during the time span T , a two-pulse ESE sequence performed after time $t=T$ produces an echo which is inverted with respect to a two-pulse ESE applied to the same spin system at thermal equilibrium. When, during the time span T , a selective RF π -pulse is applied, the two-pulse ESE will disappear as soon as the RF field is on resonance with one of the two transitions $1\leftrightarrow 2$ or $3\leftrightarrow 4$. This is because the populations of the nuclear sublevels are interchanged by the RF pulse, which simultaneously equalizes the populations of the on-resonant EPR transition. The ENDOR effect will not be observable if the preparation pulse (i.e. MW pulse 1) and/or the two-pulse ESE sequence are non-selective.

Mims ENDOR involves observation of the stimulated echo intensity as a function of the frequency of an RF π -pulse applied between the second and third MW pulse. In contrast to the Davies ENDOR experiment, the Mims-ENDOR sequence does not require selective MW pulses. For a detailed description of the polarization transfer in a Mims-type experiment the reader is referred to the literature [43]. Just as with three-pulse ESEEM, blind spots can occur in ENDOR spectra measured using Mims' method. To avoid the possibility of missing lines it is therefore essential to repeat the experiment with different values of the pulse spacing τ . Detection of the echo intensity as a function of the RF frequency and τ yields a real two-dimensional experiment. An FT of the τ -domain will yield cross-peaks in the 2D-FT-ENDOR spectrum which correlate different ENDOR transitions belonging to the same nucleus. One advantage of Mims ENDOR over Davies ENDOR is its larger echo intensity because more spins due to the nonselective excitation are involved in the formation of the echo.

Pulsed ENDOR offers several distinct advantages over conventional CW ENDOR spectroscopy. Since there is no MW power during the observation of the ESE, klystron noise is largely eliminated. Furthermore, there is an additional advantage in that, unlike the case in conventional CW ENDOR spectroscopy, the detection of ENDOR spin echoes does not depend on a critical balance of the RF and MW powers and the various relaxation times. Consequently, the temperature is not such a critical parameter in pulsed ENDOR spectroscopy. Additionally the pulsed technique permits a study of transient radicals.

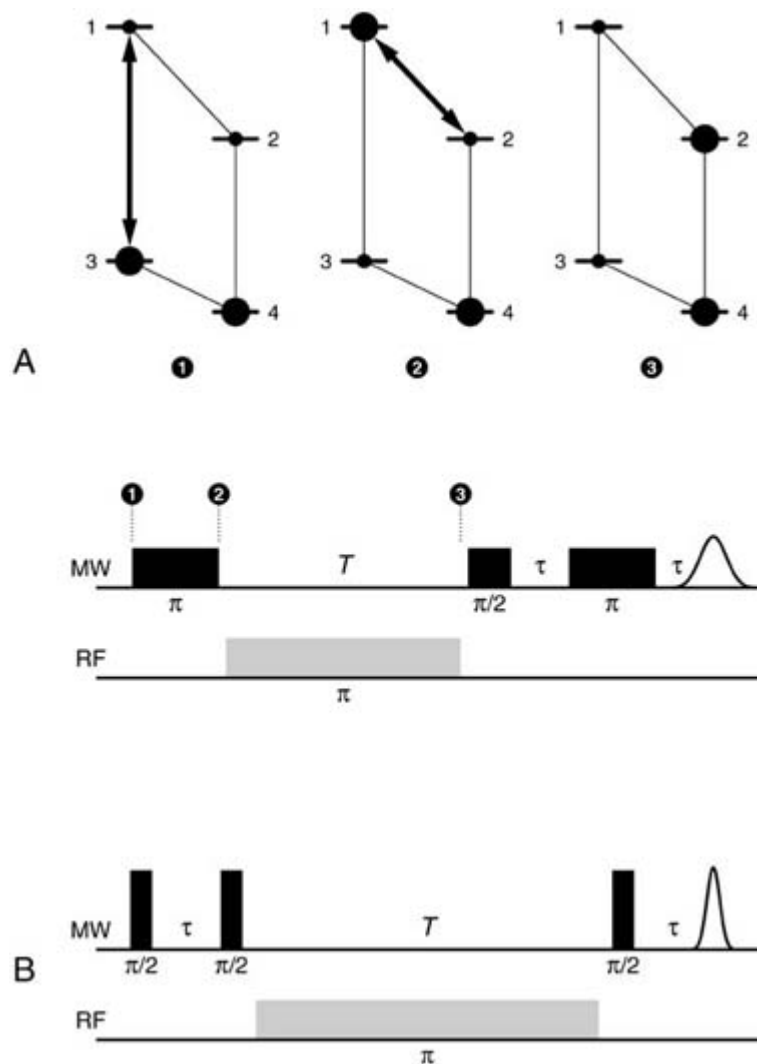


Figure B1.15.13. Pulsed ENDOR spectroscopy. (A) Top: energy level diagram of an $S=I=\frac{1}{2}$ spin system (see also [figure B1.15.8\(A\)](#)). The size of the filled circles represents the relative population of the four levels at different times during the (3+1) Davies ENDOR sequence (bottom). (B) The Mims ENDOR sequence.

More advanced pulsed techniques have also been developed. For a review of pulsed ENDOR techniques the reader is referred to [\[43, 44 and 45\]](#).

B1.15.7 HIGH-FIELD EPR SPECTROSCOPY

Since its discovery in 1944 by Zavoisky, EPR has typically been performed at frequencies below 40 GHz. This limitation was a technical one, but recent developments in millimetre and submillimetre wave frequency technology and magnetic field technology have enabled the exploration of ever higher EPR frequencies [\[46, 47 and 48\]](#). High-field/high-frequency EPR spectroscopy has a number of inherent advantages [\[47\]](#). (1) The spectral resolution of g -factor differences and anisotropies greatly improves since the electron Zeeman interaction scales linearly with the magnetic field (see [equation \(b1.15.18\)](#)). If paramagnetic centres with different g -values or different magnetic sites of rather similar g -values are present, the difference in the spectral field positions of the resonances is proportional to the MW frequency ω

$$\Delta B_0 = \frac{\hbar\omega}{\beta_c} \left(\frac{1}{g_1} - \frac{1}{g_2} \right). \quad (\text{B1.15.40})$$

Even for a single radical the spectral resolution can be enhanced for disordered solid samples if the inhomogeneous linewidth is dominated by unresolved hyperfine interactions. Whereas the hyperfine line broadening is not field dependent, the anisotropic \mathbf{g} -matrix contribution scales linearly with the external field. Thus, if the magnetic field is large enough, i.e. when the condition

$$\frac{\Delta g}{g_{\text{iso}}} B_0 > B_{1/2}^{\text{HFI}} \quad (\text{B1.15.41})$$

is fulfilled, the powder spectrum is dominated by the anisotropic \mathbf{g} -matrix. equation (b1.15.41) may be considered as the high-resolution condition for solid-state EPR spectra, to be fulfilled only at high enough B_0 . From [figure B1.15.14](#) one sees that for a nitroxide spin label in a protein this is fulfilled almost completely at 95 GHz, but not at 10 GHz. In the case of well resolved g -anisotropy the extension to high-field ENDOR and ESEEM has the additional advantage of providing single-crystal-like hyperfine information when transitions are excited at field positions where only specific orientations of the \mathbf{g} -matrix with respect to the external magnetic field contribute to the spectrum. (2) Relaxation times become longer for many systems at higher frequencies. (3) Particularly in studies of small samples, high-field/high-frequency EPR is typically more sensitive compared to EPR at X-band frequencies, by virtue of the increased Boltzmann factor (see [equation \(b1.15.10\)](#)). (4) For high-spin systems with zero-field splittings larger than the MW quantum it is impossible to observe all EPR transitions. Here, higher-frequency experiments are essential for recording the whole spectrum. (5) At high frequencies it becomes possible to violate the high-temperature approximation with standard cryogenic systems. This effect can be exploited to gain information on the absolute sign of parameters of the spin Hamiltonian.

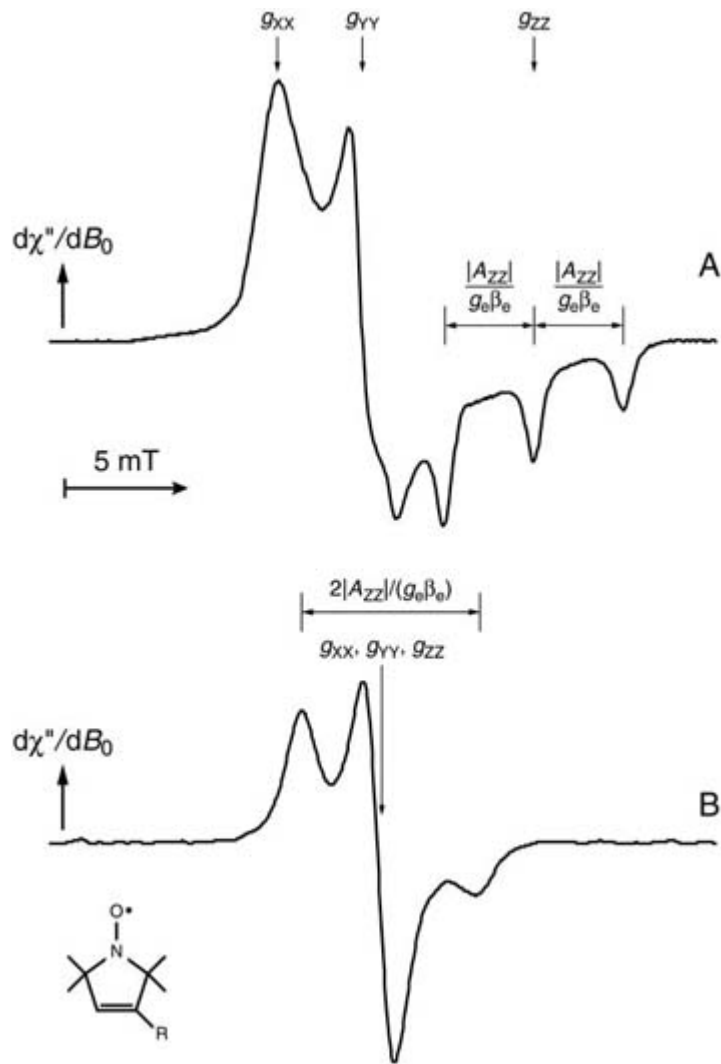


Figure B1.15.14. Comparison of 95.1 GHz (A) and 9.71 GHz (B) EPR spectra for a frozen solution of a nitroxide spin label attached to insulin measured at 170 K.

Disadvantages of high-field/high-frequency EPR are mainly technical ones due to the limited availability of MW components operating at millimetre and submillimetre wavelengths and the high costs of spectrometer development. Furthermore, low-frequency EPR will not be completely superseded by high-field/high-frequency EPR because some experiments that rely on the violation of the high-field approximation no longer work when increasing the EPR frequency. The largest disadvantages occur in studies of proton interactions, where ESEEM is a convenient tool at X-band but cannot be used at W-band frequencies and above because of too small modulation depths for most systems. Nevertheless, these drawbacks are outweighed by the advantages to such an extent that high-field/high-frequency EPR methods will become more and more widespread as one overcomes the technical hurdles.

The early high-field/high-frequency EPR spectrometers were developed mostly on the basis of klystron MW generators. In the past few years, however, solid-state MW sources such as Gunn oscillators or IMPATT diodes were applied more frequently. To improve their mediocre frequency stability they need to be phase locked to a low-frequency stable reference source. If frequencies higher than approximately 150 GHz are required, the output frequency of the solid-state MW generator can be multiplied in a Schottky diode harmonic generator (multiplication factors between 2 and 5). To generate even higher frequencies up to 1 THz more exotic high-power pulsed and CW tube sources such as gyrotrons, extended interaction oscillators

(EIOs), backward wave oscillators (BWOs) or magnetrons are available. Their spectral characteristics may be favourable; however, they typically require highly stabilized high-voltage power supplies. Still higher frequencies may be obtained using far-infrared gas lasers pumped for example by a CO_2 laser [49].

All the waveguide elements become very small at high frequencies as compared to the standard X-band or even Q-band. This produces high losses up to several decibels per metre due to waveguide imperfections. Therefore, if millimetre and submillimetre waves have to be transmitted over long (and straight) distances, oversized or corrugated waveguides are normally used because of their smaller ohmic losses. Corrugated waveguides have narrow grooves, each a quarter-wavelength deep, cut into the guide walls. The effect of the grooves is to destructively average the **E** field near the wall surface which cannot now have a non-zero component perpendicular to the surface. To couple to a resonant cavity, however, these waveguides need to be tapered back to the fundamental-mode waveguide. Very recently developed EPR spectrometers operating at 130 GHz [50], 250 GHz [51] and 360 GHz [52] (see [figure B1.15.15](#) for waveguides for millimetre-wave transmission for the most part. Instead quasi-optic techniques are used [53]. Once millimetre-waves have been converted into Gaussian beams by means of corrugated feedhorns they can be transported and manipulated in free space using quasi-optical elements such as lenses (constructed from Teflon or high-density polyethylene) and off-axis mirrors. The losses in these elements are virtually negligible.

The mechanical specifications of cavities for millimetre waves are highly demanding regarding cavity dimensions, cavity surfaces and precision of coupling mechanisms due to the reduced dimensions at millimetre wavelengths. Furthermore, with increasing MW frequency resonators become more and more difficult to handle. Therefore, high-field/high-frequency EPR measurements are very often carried out without a cavity with the sample placed directly in a transmission waveguide. From the point of view of absolute sensitivity, however, small-volume cavities are evidently preferable. Typically used cavities for millimetre- and submillimetre-wave EPR are multi-mode Fabry–Pérot resonators [54] consisting of a confocal or semiconfocal arrangement of two mirrors placed at a particular distance apart. Fabry–Pérot resonators have been used successfully in the reflection and the transmission mode. A typical Fabry–Pérot resonator is extremely sensitive to displacements of the mirrors by as little as 0.1 μm . Also, the mechanical isolation of the cavity from the modulation coils has to be improved compared to lower frequency designs because of the larger modulation amplitudes and increased interaction forces with the larger static field.

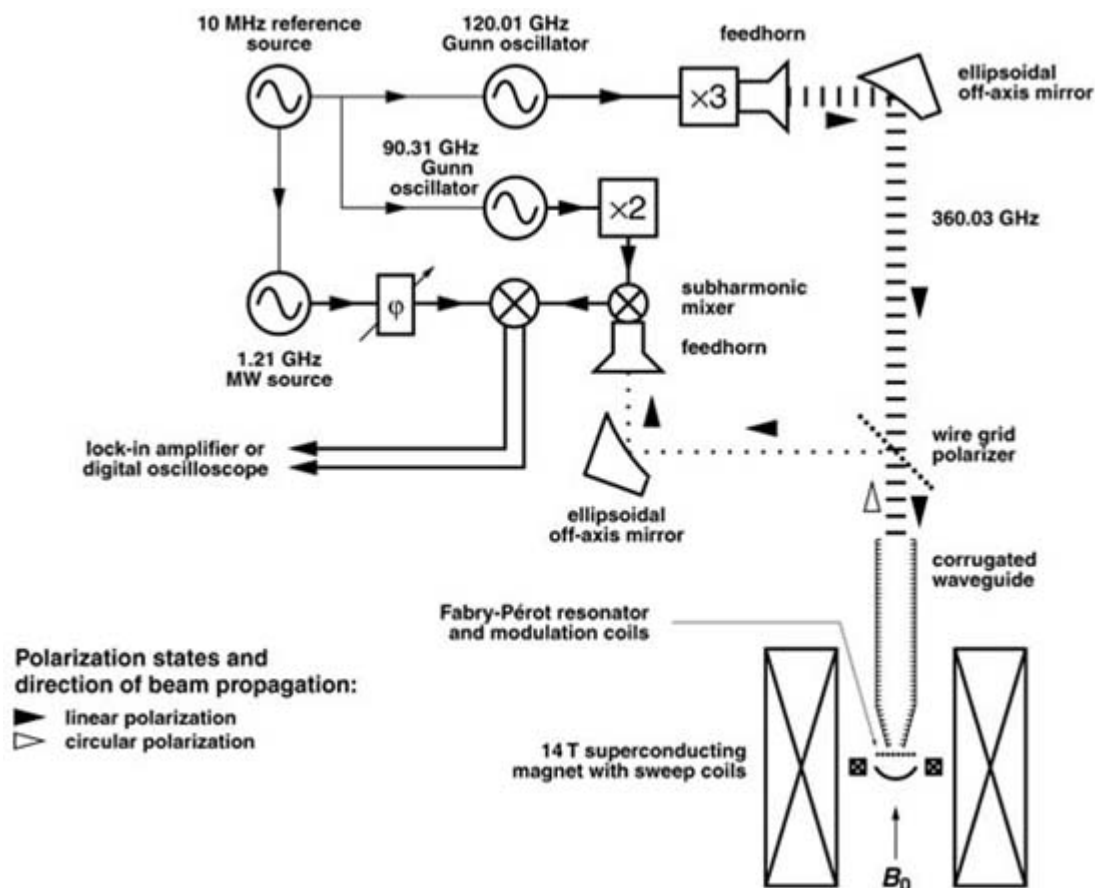


Figure B1.15.15. High-field/high-frequency EPR spectrometer operating at 360.03 GHz b1.15.52. Microwaves at 360.03 GHz are produced by frequency multiplication of the output of a Gunn oscillator and are then passed into a corrugated feedhorn to set up a fundamental Gaussian beam. Refocusing and redirection of the beam is accomplished using off-axis mirrors. The microwaves then travel through a corrugated waveguide and are coupled into a Fabry–Perot-type open cavity. Incident linearly polarized microwaves and circularly polarized EPR signals emitted from the resonator are separated by a wiregrid polarizer. Heterodyne phase-sensitive detection of the EPR signal is achieved in a subharmonic mixer using microwaves of 180.62 GHz to produce a 1.21 GHz intermediate frequency signal which is further down-converted in a quadrature mixer. All MW sources are phase locked to one common reference oscillator.

Many different sorts of millimetre-wave detectors have been developed, each offering its own combination of advantages and drawbacks. For convenience, they may be divided into two general categories: bolometers and mixers (heterodyne detectors). A bolometer is a device which responds to a change in temperature produced when it absorbs incident radiation. The noise figure of an He-cooled bolometer is excellent; however, its small bandwidth limits its application to CW EPR experiments. Heterodyne detection systems transfer a signal band from a high frequency to a lower frequency where low-noise amplifiers are available. This is accomplished, for example, in a mixer by overlaying the signal with a monofrequent and stable local oscillator (LO) frequency to produce a (difference) intermediate frequency (IF) in the lower GHz range. With increasing frequency, however, the Schottky barrier diodes used in mixers become very sensitive to static electricity and mechanical stress, thus limiting their reliability.

Pulsed, or time-domain, EPR spectrometers have also been developed at higher frequencies up to 140 GHz [55, 56]. They are generally low-power units with characteristically long pulse lengths (typically 50 ns for a $\pi/2$ -pulse) due to the limited MW powers available at millimetre wavelengths and the lack of fast-switching

pulse-forming devices at these frequencies. In general, all the experiments outlined in the previous section can be performed, however, with the even more severe restriction to limited excitation bandwidths. Nevertheless, pulsed EPR performed at high frequencies has clear advantages when the poor orientation selection at low frequencies prevents the study of spectral anisotropies of the ESE decay, for example in relaxation measurements. As an example, figure B1.15.16 depicts the decay of a two-pulse ESE of a quinone radical as a function of the external magnetic field. Clearly, the echo decay governed by T_2 is different for selected field positions in the spectrum. This T_2 anisotropy can be analysed in terms of anisotropic motional processes of the radical in its molecular environment. Due to the low g -anisotropy (as is typical in biomolecules) this experiment would not have been successful at lower frequencies.

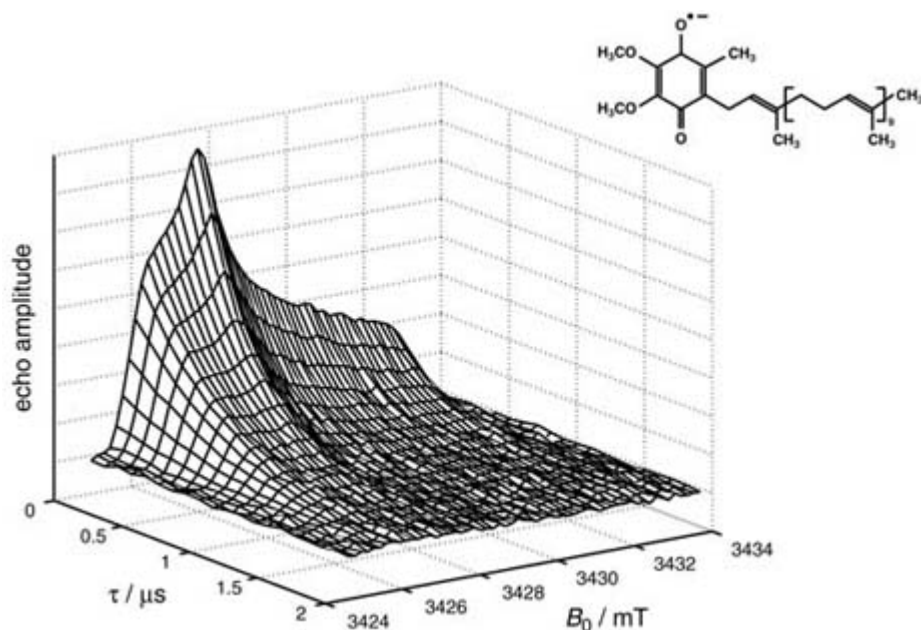


Figure B1.15.16. Two-pulse ESE signal intensity of the chemically reduced ubiquinone-10 cofactor in photosynthetic bacterial reaction centres at 115 K. MW frequency is 95.1 GHz. One dimension is the magnetic field value B_0 ; the other dimension is the pulse separation τ . The echo decay function is anisotropic with respect to the spectral position.

REFERENCES

- [1] Eaton G R, Eaton S S and Salikhov K M 1998 *Foundations of Modern EPR* (Singapore: World Scientific)
- [2] Atherton N M 1993 *Principles of Electron Spin Resonance* (Chichester: Ellis Horwood)
- [3] Weil J A, Bolton J R and Wertz J E 1994 *Electron Paramagnetic Resonance* (New York: Wiley)
- [4] Poole C P 1983 *Electron Spin Resonance: a Comprehensive Treatise on Experimental Techniques* 2nd edn (Mineola, NY: Dover)
- [5] Gordy W 1980 *Theory and Applications of Electron Spin Resonance* (New York: Wiley)
- [6] Alger R S 1968 *Electron Paramagnetic Resonance: Techniques and Applications* (New York: Wiley)
- [7] Carrington A and McLachlan A D 1967 *Introduction to Magnetic Resonance* (New York: Harper and Row)
- [8] Kurreck H, Kirste B and Lubitz W 1988 *Electron Nuclear Double Resonance Spectroscopy of Radicals in Solution: Application to Organic and Biological Chemistry* (Weinheim: VCH)

- [9] Poole C P and Farach H A 1971 *Relaxation in Magnetic Resonance* (New York: Academic)
- [10] Bertini I, Martini G and Luchinat C 1994 Relaxation, background, and theory *Handbook of Electron Spin Resonance* (ed C Poole and H Farach (New York: American Institute of Physics) ch 3, pp 51–77
- [11] Slichter C P 1992 *Principles of Magnetic Resonance* (Berlin: Springer)
- [12] Hyde J S and Froncisz W 1989 Loop gap resonators *Advanced EPR in Biology and Biochemistry* ed A J Hoff (Amsterdam: Elsevier) ch 7, pp 277–305
- [13] Stehlik D, Bock C H and Thurnauer M 1989 Transient EPR-spectroscopy of photoinduced electronic spin states in rigid matrices *Advanced EPR in Biology and Biochemistry* ed A J Hoff (Amsterdam: Elsevier) ch 11, pp 371–403
- [14] McLaughlan K A 1990 Continuous-wave transient electron spin resonance *Modern Pulsed and Continuous-Wave Electron Spin Resonance* ed L Kevan and M Bowman (New York: Wiley) ch 7, pp 285–363
- [15] Kim S S and Weissman S I 1976 Detection of transient electron paramagnetic resonance *J. Magn. Reson.* **24** 167–9
- [16] Furrer R, Fujara F, Lange C, Stehlik D, Vieth H and Vollmann W 1980 Transient ESR nutation signals in excited aromatic triplet states *Chem. Phys. Lett.* **75** 332–9
- [17] Kothe G, Weber S, Bittl R, Ohmes E, Thurnauer M and Norris J 1991 Transient EPR of light-induced radical pairs in plant photosystem I: observation of quantum beats *Chem. Phys. Lett.* **186** 474–80
- [18] Weber S, Ohmes E, Thurnauer M C, Norris J R and Kothe G 1995 Light-generated nuclear quantum beats: a signature of photosynthesis *Proc. Natl Acad. Sci. USA* **92** 7789–93
- [19] Piekara-Sady L and Kispert L D 1994 ENDOR spectroscopy *Handbook of Electron Spin Resonance* ed C P Poole and H A Farach (New York: American Institute of Physics) ch 5, pp 311–57

- [20] Möbius K, Plato M and Lubitz W 1982 Radicals in solution studied by ENDOR and TRIPLE resonance spectroscopy *Phys. Rep.* **87** 171–208
- [21] Möbius K, Lubitz W and Plato M 1989 Liquid-state ENDOR and TRIPLE resonance *Advanced EPR in Biology and Biochemistry* ed A J Hoff (Amsterdam: Elsevier) ch 13, pp 441–99
- [22] Möbius K and Biehl R 1979 Electron–nuclear–nuclear TRIPLE resonance of radicals in solution *Multiple Electron Resonance Spectroscopy* ed M M Dorio and J H Freed (New York: Plenum) ch 14, pp 475–507
- [23] Leniart D S 1979 Instrumentation and experimental methods in double resonance *Multiple Electron Resonance Spectroscopy* ed M M Dorio and J H Freed (New York: Plenum) ch 2, pp 5–72
- [24] Feher G 1956 Observation of nuclear magnetic resonances via the electron spin resonance line *Phys. Rev.* **103** 834–7
- [25] Freed J H 1969 Theory of saturation and double resonance effects in ESR spectra. IV. Electron–nuclear triple resonance *J. Chem. Phys.* **50** 2271–2
- [26] Dinse K P, Biehl R and Möbius K 1974 Electron nuclear triple resonance of free radicals in solution *J. Chem. Phys.* **61** 4335–41
- [27] Biehl R, Plato M and Möbius K 1975 General TRIPLE resonance on free radicals in solution. Determination of relative signs of isotropic hyperfine coupling constants *J. Chem. Phys.* **63** 3515–22

- [28] Hyde J S, Chien J C W and Freed J 1968 Electron–electron double resonance of free radicals in solution *J. Chem. Phys.* **48** 4211–26
- [29] Bruno G and Freed J 1974 ESR lineshapes and saturation in the slow motional region: ELDOR *Chem. Phys. Lett.* pp 328–32
- [30] Freed J 1979 Theory of multiple resonance and ESR saturation in liquids and related media *Multiple Electron Resonance Spectroscopy* ed M M Dorio and J H Freed (New York: Plenum) ch 3, pp 73–142
- [31] Keijzers C P, Reijerse E and Schmidt J 1989 *Pulsed EPR: a New Field of Applications* (Amsterdam: North-Holland)
- [32] Schweiger A 1991 Pulsed electron spin resonance spectroscopy: basic principles, techniques, and examples of applications *Angew. Chem. Int. Edn Engl.* **30** 265–92
- [33] Bowman M K 1990 Fourier transform electron spin resonance *Modern Pulsed and Continuous-Wave Electron Spin Resonance* ed L Kevan and M Bowman (New York: Wiley) ch 1, pp 1–42
- [34] Hahn E L 1950 Spin echoes *Phys. Rev.* **80** 580–94
- [35] Ponti A and Schweiger A 1994 Echo phenomena in electron paramagnetic resonance spectroscopy *Appl. Magn. Reson.* **7** 363–403
- [36] Schweiger A 1990 New trends in pulsed electron spin resonance methodology *Modern Pulsed and Continuous-Wave Electron Spin Resonance* ed L Kevan and M K Bowman (New York: Wiley) ch 2, pp 43–118
- [37] Rowan L G, Hahn E L and Mims W B 1965 Electron-spin echo envelope modulation *Phys. Rev.* **137** A61–A71
-

- [38] Mims W B 1972 Envelope modulation in spin-echo experiments *Phys. Rev. B* **5** 2409–19
- [39] Merks R P J and de Beer R 1979 Two-dimensional Fourier transform of electron spin-echo envelope modulation. An alternative for ENDOR *J. Phys. Chem.* **83** 3319–22
- [40] Höfer P, Grupp A, Nebenführ H and Mehring M 1986 Hyperfine sublevel correlation (HYSCORE) spectroscopy: a 2D ESR investigation of the squaric acid radical *Chem. Phys. Lett.* **132** 279–82
- [41] Mims W B 1965 Pulsed ENDOR experiments *Phys. Rev. S* **452** 452–7
- [42] Davies E R 1974 A new pulse ENDOR technique *Phys. Lett.* **47** 1–2
- [43] Gemperle C and Schweiger A 1991 Pulsed electron-nuclear double resonance methodology *Chem. Rev.* **91** 1481–505
- [44] Grupp A and Mehring M 1990 Pulsed ENDOR spectroscopy in solids *Modern Pulsed and Continuous-Wave Electron Spin Resonance* ed L Kevan and M K Bowman (New York: Wiley) ch 4, pp 195–229
- [45] Dinse K P 1989 Pulsed ENDOR *Advanced EPR in Biology and Biochemistry* ed A J Hoff (Amsterdam: Elsevier) ch 17, pp 615–30
- [46] Lebedev Y S 1990 High-frequency continuous-wave electron spin resonance *Modern Pulsed and Continuous-Wave Electron Spin Resonance* ed L Kevan and M K Bowman (New York: Wiley) ch 8, pp 365–404
- [47] Lebedev Y S 1994 Very-high-field EPR and its applications *Appl. Magn. Reson.* **7** 339–62

- [48] Budil D E, Earle K A, Lynch W B and Freed J H 1989 Electron paramagnetic resonance at 1 millimeter wavelengths *Advanced EPR in Biology and Biochemistry* ed A J Hoff (Amsterdam: Elsevier) ch 8, pp 307–40
- [49] Barra A L, Brunel L and Robert J 1990 EPR spectroscopy at very high field *Chem. Phys. Lett.* **165** 107–9
- [50] Reijerse E J, van Dam P J, Klaassen A A K, Hagen W, van Bentum P J M and Smith G 1998 Concepts in high-frequency EPR—applications to bio-inorganic systems *Appl. Magn. Reson.* **14** 153–67
- [51] Lynch W, Earle K and Freed J 1988 1-mm wave ESR spectrometer *Rev. Sci. Instrum.* **59** 1345–51
- [52] Fuchs M, Weber S, Möbius K, Rohrer M and Prisner T 1998 A submillimeter high-field EPR spectrometer using quasi-optical microwave bridge devices *Magnetic Resonance and Related Phenomena* ed D Ziessow, W Lubitz and F Lendzian (Berlin: Technische Universität Berlin)
- [53] Earle K, Budil D and Freed J 1996 Millimeter wave electron spin resonance using quasioptical techniques *Advances in Magnetic and Optical Resonance* vol 19, ed W Warren (San Diego: Academic) pp 253–323
- [54] Kogelnik H and Li T 1966 Laser beams and resonators *Proc. IEEE* **5** 88–105
- [55] Allgeier J, Disselhorst A, Weber R, Wenckebach W and Schmidt J 1990 High-frequency pulsed electron spin resonance *Modern Pulsed and Continuous-Wave Electron Spin Resonance* ed L Kevan and M K Bowman (New York: Wiley) ch 6, pp 267–83
- [56] Prisner T F 1997 Pulsed high-frequency/high-field EPR *Advances in Magnetic and Optical Resonance* vol 20, ed W Warren (San Diego: Academic) pp 245–99

-1-

B1.16 Chemically-induced nuclear and electron polarization (CIDNP and CIDEP)

Elizabeth J Harbron and Malcolm D E Forbes

B1.16.1 INTRODUCTION

Chemically-induced spin polarization was one of the last truly new physical phenomena in chemistry to be discovered and explained during this century. So unusual were the observations and so ground-breaking the theoretical descriptions that, over a very short time period, the chemist's way of thinking about free radical reactions and how to study them was fundamentally changed. After the earliest experimental reports of unusual phases of electron paramagnetic resonance (EPR) (1963) [1] and nuclear magnetic resonance (NMR) (1967) [2, 3 and 4] transitions in thermal, photolytic and radiolytic reactions involving free radical intermediates, it took several years of theoretical development before the idea of the radical pair mechanism (RPM) was put forward to explain the results [5, 6, 7, 8 and 9]. Gradually, the theory was tested and improved, and additional polarization mechanisms were discovered. The overall physical picture has stood the test of time and now both chemically-induced dynamic nuclear polarization (CIDNP) and its electron analogue (CIDEP) are well understood. The phenomena are exploited by many researchers who are trying to understand the kinetic and magnetic properties (and the links between them) of free radicals, biradicals and radical ion pairs in organic photochemistry, as well as photosynthetic reaction centres and other biologically relevant systems. The high structural resolution of NMR and EPR spectroscopies, combined with recent advances in fast data collection instrumentation and high powered pulsed lasers, has made time-resolved CIDNP and CIDEP experiments some of the most informative in the modern physical chemistry arsenal.

In spectroscopy it is common for transitions to be observed as absorptive lines because the Boltzmann distribution, at equilibrium, ensures a higher population of the lower state than the upper state. Examples where emission is observed, which are by definition non-equilibrium situations, are usually cases where excess population is created in the higher level by infusing energy into the system from an external source. For example, steady-state emission spectroscopy is used to measure fluorescence or phosphorescence from the excited states of organic molecules. The technique requires excitation to the upper energy levels first, then what is observed is a spontaneous emission. Another example is the laser, which is pumped with an external source such as a flash lamp or an electric arc to ensure a population inversion, and stimulated emission then occurs from the upper state upon absorption of another photon. What makes the non-Boltzmann NMR and EPR populations observed in CIDNP and CIDEF experiments so unusual is that *nuclear-spin dependent* chemical reactions (homolytic bond-breaking or forming) are responsible for the process. While it usually requires energy to break the bond, once it is broken the mixing of spin wavefunctions in the resulting radical pair, which will be described in detail in the following, is all that is necessary to make some NMR and EPR transitions appear with enhanced absorption (greater intensity than Boltzmann would predict) or even in emission (higher population in the excited state). The overall phase and magnitude of the polarization is dependent on the nuclear spin projections of the nuclei (usually, but not always, protons) near the free radical site of the molecules in question. For this reason, it is easy to see why a suitable theoretical description of CIDNP took a long time to evolve. The idea that the nuclear spin-state energy level differences, which are much smaller than kT at room temperature, could be responsible for different chemical reaction rates was a revolutionary and somewhat controversial one. As more and

-2-

more experiments were performed to support this idea, it rapidly gained acceptance and, in fact, helped connect the solution dynamics of small molecules to spin quantum mechanics in a very natural and informative fashion.

We make one important note here regarding nomenclature. Early explanations of CIDNP invoked an Overhauser-type mechanism, implying a dynamic process similar to spin relaxation; hence the word ‘dynamic’ in the CIDNP acronym. This is now known to be incorrect, but the acronym has prevailed in its infant form.

The general phenomena of CIDNP and CIDEF are presented in figure B1.16.1 and [figure B1.16.2](#). Figure B1.16.1 shows work by Roth *et al* [10] on radical cation structure in which the bottom trace is a ‘dark’ spectrum and the top trace is the CIDNP spectrum [10]. [Figure B1.16.2](#) shows CIDEF spectra of radicals formed by decomposition of a fluorinated polymer initiator [11]. The NMR spectra in figure B1.16.1 and the EPR spectra in [figure B1.16.2](#) can be recognized as spin-polarized by the presence of lines in emission and enhanced absorption. The origin of the CIDNP and CIDEF phenomena will be explored and explained in the following, and we will return to these examples for further analysis once the theory behind them is understood.

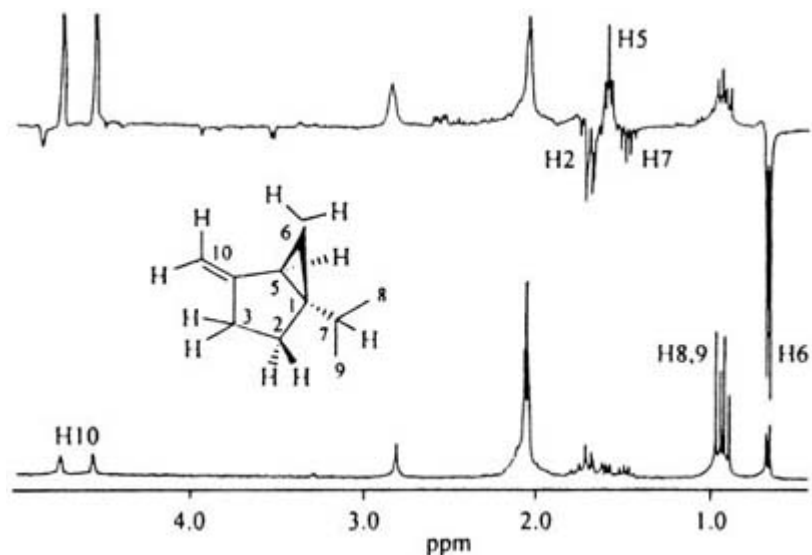


Figure B1.16.1. ^1H CIDNP spectrum (250 MHz; top) observed during irradiation of chloranil with sabinene (**1**) in acetone- d_6 and dark spectrum (bottom). Assignments are based on the 2D ^1H - ^1H COSY spectrum. Reprinted from [10].

-3-

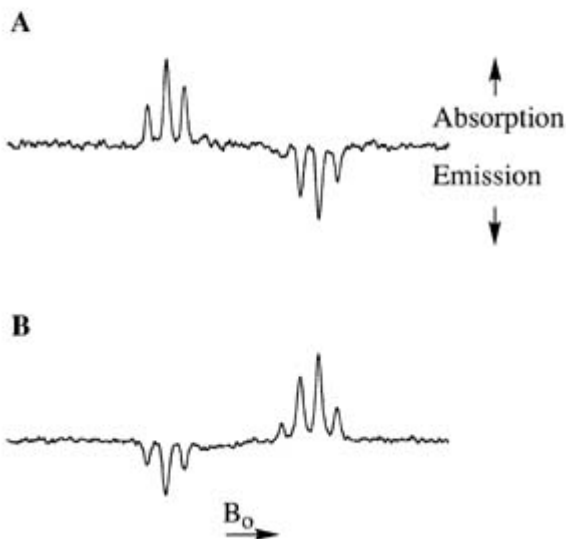
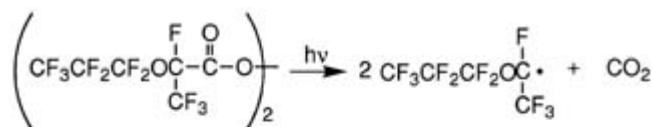


Figure B1.16.2. X-band TREPR spectra obtained at 0.1 μs after 308 nm photolysis of a fluorinated peroxide dimer in Freon 113 at room temperature. Part A is the A/E RPM spectrum obtained upon direct photolysis; part B is the E/A RPM spectrum obtained upon triplet sensitization of this reaction using benzophenone.

B1.16.2 CIDNP

B1.16.2.1 SPIN HAMILTONIAN

CIDNP involves the observation of diamagnetic products formed from chemical reactions which have radical intermediates. We first define the geminate radical pair (RP) as the two molecules which are born in a radical reaction with a well defined phase relation (singlet or triplet) between their spins. Because the spin physics of the radical pair are a fundamental part of any description of the origins of CIDNP, it is instructive to begin with a discussion of the radical-pair spin Hamiltonian. The Hamiltonian can be used in conjunction with an appropriate basis set to obtain the energetics and populations of the RP spin states. A suitable Hamiltonian for a radical pair consisting of radicals 1 and 2 is shown in equation (B1.16.1) below [12].

$$\hat{H}_{\text{RP}} = \beta_e \hbar^{-1} B_0 (g_1 \hat{S}_{1z} + g_2 \hat{S}_{2z}) + \hat{S}_1 \cdot \sum_j a_j \hat{I}_j + \hat{S}_2 \cdot \sum_k a_k \hat{I}_k - J \left(\frac{1}{2} + 2\hat{S}_1 \cdot \hat{S}_2 \right) \quad (\text{B1.16.1})$$

-4-

The first term describes the electronic Zeeman energy, which is the interaction of the magnetic field with the two electrons of the radical pair with the magnetic field, B_0 . The two electron spins are represented by spin operators \hat{S}_{1z} and \hat{S}_{2z} . In this expression, g is the g factor, which is the chemical shift of the unpaired electrons. The other variables in the first term are constants: β_e is the Bohr magneton, and \hbar is Planck's constant divided by 2π . The second term in the Hamiltonian describes the hyperfine interaction between each radical and the nuclei on that radical, where \hat{S} is again the electron spin operator, \hat{I} is the nuclear spin operator, and a_j and a_k are the hyperfine coupling constants. The hyperfine constants describe the coupling between electronic and nuclear spins; coupling between an electron and a proton is designated a_{H} . For most carbon-centred alkyl free radicals, a_{H} for an electron and a proton on the same carbon is negative in sign while a_{H} for a proton β to an electron is positive. The sign of the hyperfine coupling constant will become an important issue in the analysis of CIDNP data below.

The final term in the radical pair Hamiltonian is the exchange interaction (J) between the unpaired electrons. This interaction is a scalar quantity that describes the coupling of the angular momenta of two radicals which are in close proximity. Its magnitude decreases exponentially with increasing inter-radical distance as shown by equation (B1.16.2), where J_0 is the exchange interaction at the point of closest contact, r_0 ; r is the inter-radical distance; and λ is a fall-off parameter generally accepted to be approximately 1 \AA^{-1} for isotropic solutions. The exchange interaction should not be confused with the quantum mechanical term exchange integral, although the two are related [13].

$$J = J_0 e^{-\lambda(r-r_0)}. \quad (\text{B1.16.2})$$

The exchange interaction results in an energy splitting between the singlet and triplet states of the RP as shown in figure B1.16.3 which shows a plot of the RP energy levels versus the inter-radical separation. The S and T levels shown in the lower part of the figure are in the absence of an external magnetic field. When an external field is applied, the triplet level is split into T_+ , T_0 and T_- , as shown in the upper inset. At the high magnetic fields at which most CIDNP experiments are conducted, T_+ and T_- are far away from S and can be neglected in what is known as the high-field approximation.

-5-

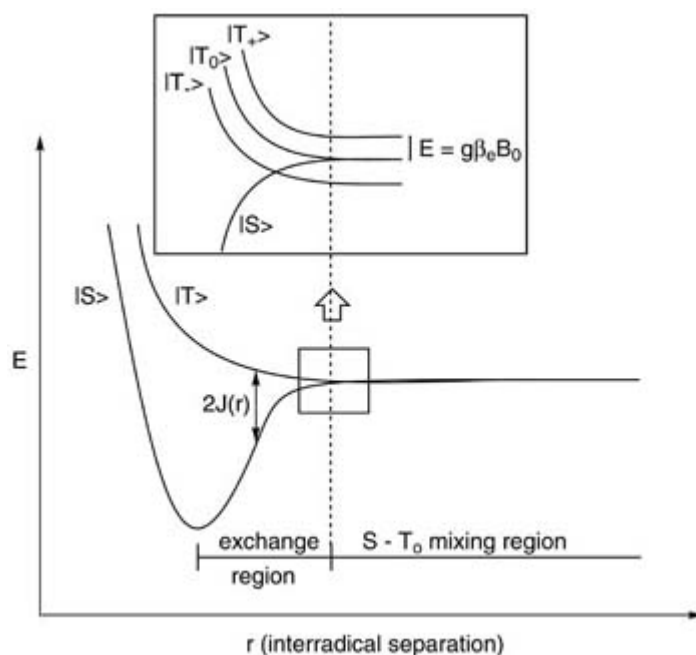


Figure B1.16.3. Energy levels versus inter-radical separation for a radical pair. The lower part of the figure shows the S and T levels in the absence of an external magnetic field while the inset shows the splitting of the triplet levels in the presence of a magnetic field, B_0 .

When the inter-radical distance is very small, J is large, and S and T_0 are far apart and cannot mix. This is called the ‘exchange region’ as the electron spins are constrained by the exchange interaction to remain in their respective spin states. As the radicals diffuse apart, J rapidly falls to zero, and the S and T_0 levels become degenerate and are allowed to mix. If the radicals diffuse back to the region of large J , they must again be in either the S or T_0 state, but some of the RPs which were formerly in one spin state will now be in the other. This phenomenon is known as intersystem crossing; it is a process critical to the understanding of CIDNP and will be explained in greater detail below.

B1.16.2.2 MECHANISM OF INTERSYSTEM CROSSING

Vector representations of the radical pair spin states, shown in [figure B1.16.4A](#), help to explain how intersystem crossing occurs in RPs. The vector diagrams show the magnitude of the spin angular momentum vector and its z component (parallel to B_0) for each of the two spins in the RP. Because the x and y components of spin are unspecified, each spin vector can be thought to precess around a cone. When considering only the influence of the interaction of the electron chemical shift with the magnetic field, the frequency of electron precession is given by the expression for the Larmor frequency, equation (B1.16.3). Additional interactions, such as electron–nuclear hyperfine couplings, must be added to or subtracted from the Larmor frequency in order to determine the actual precessional frequency of a given electron;

$$\omega = g\beta_e\hbar^{-1}B_0. \quad (\text{B1.16.3})$$

As mentioned previously, a geminate RP in close contact is constrained by the exchange interaction to remain in its initial spin state, S or T_0 , and vector representations of these states are shown in [figure B1.16.4A](#). The exchange interaction prevents the two spins in the RP from precessing independently; so long as they precess at the same frequency, the two spins will remain in the same mutual orientation. Once the radicals have diffused to the region where J is zero, however, they are free to precess independently of one another. At this point, differences in g factor and/or hyperfine interaction will cause the radicals to precess at different

frequencies. This difference in precessional frequencies is given by Q , as shown in equation (B1.16.4), where m_{1i} and m_{2j} are the nuclear magnetic quantum numbers for each member of the RP, respectively, and the other variables were defined previously;

$$\omega_1 - \omega_2 = 2Q = (g_1 - g_2)\beta_v\hbar^{-1}B_0 + \sum a_{1i}m_{1i} - \sum a_{2j}m_{2j}. \quad (\text{B1.16.4})$$

Eventually, the two spins will fall out of step with one another and will oscillate between the S state, intermediate states, and the T_0 state. An intermediate state is shown in region II of the vector diagram and can be described as a coherent superposition of the S and T_0 states with the coefficients c_S and c_T delineating the amount of S and T_0 'character'. If the radicals diffuse back together to the region of large J , the RP is again required to be in either the S or T_0 state. The squares of the coefficients c_S and c_T will determine the probability that the RP will jump to the S or T_0 state at this inter-radical distance. Some RPs will now be in a different spin state than they were initially and are said to have undergone intersystem crossing, as mentioned above. While S - T_0 mixing occurs when the radicals in a RP are in the $J=0$ region, the fact that intersystem crossing has occurred cannot be determined unless they diffuse back together and are forced by the exchange interaction to be in the S or T_0 state.

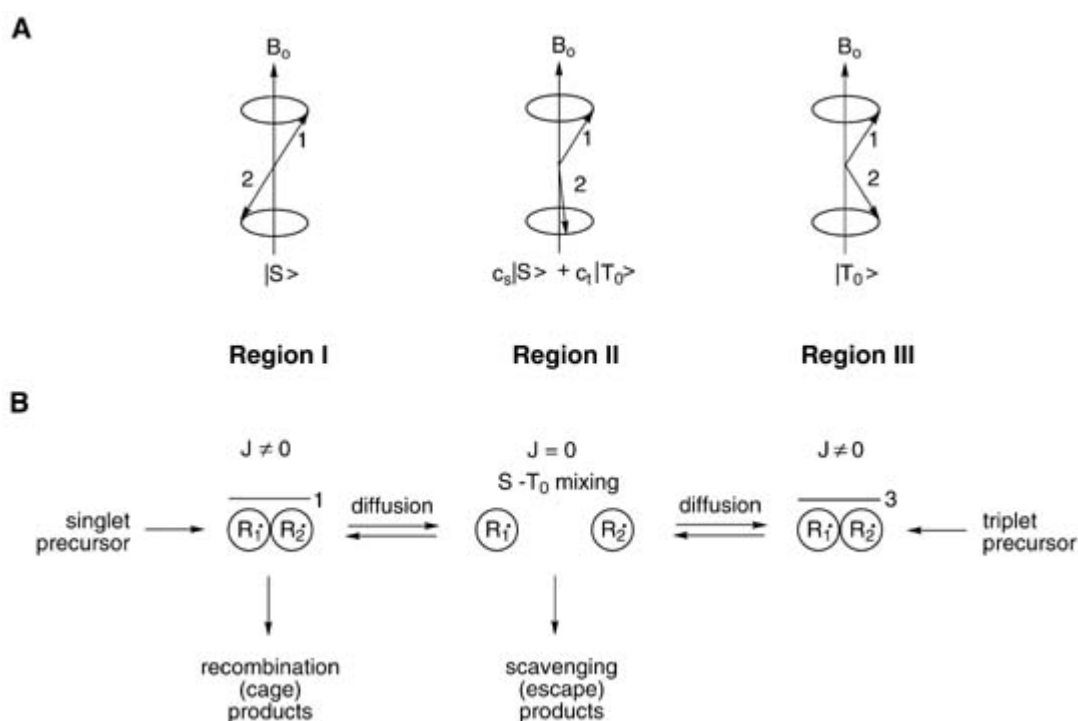


Figure B1.16.4. Part A is the vector representations of the S state, an intermediate state, and the T_0 state of a radical pair. Part B is the radical reaction scheme for CIDNP.

By examining the expression for Q (equation (B1.16.4)), it should now be clear that the nuclear spin state influences the difference in precessional frequencies and, ultimately, the likelihood of intersystem crossing, through the hyperfine term. It is this influence of *nuclear* spin states on *electronic* intersystem crossing which will eventually lead to non-equilibrium distributions of nuclear spin states, i.e. spin polarization, in the products of radical reactions, as we shall see below.

B1.16.2.3 RADICAL REACTION SCHEME

A general reaction scheme for CIDNP is shown in [figure B1.16.4B](#), where the radical dynamics in each region

correspond to the vector diagram for that region shown in [figure B1.16.4A](#). A geminate RP is formed from a singlet or triplet precursor through bond cleavage or an electron transfer reaction. Thermal reactions proceed from the singlet state while photochemical reactions tend to occur from the triplet state; certain species, such as azo compounds, react from the singlet in photochemical reactions [14]. The RP is always formed in the same spin state as its precursor because the RP-forming reaction must conserve angular momentum. For the vast majority of reactions, recombination of singlet RPs is allowed while recombination of triplet RPs is forbidden. While there are exceptions [15], we will consider triplet RPs to be incapable of recombination throughout this explanation. It should also be noted that [figure B1.16.4B](#) shows radical-forming reaction pathways for both singlet and triplet RPs for the purpose of illustration, but it is to be understood that most radical-forming reactions occur from either a singlet (left side) or triplet (right side) precursor but not both.

The geminate RPs in [figure B1.16.4B](#) are indicated by a bar with the spin multiplicity. It is common to speak of geminate RPs as being in a ‘cage’, and this notion is central enough to our discussion of CIDNP to merit some discussion. The idea of the cage effect in radical chemistry stems from early work by Franck and Rabinowitch [16] in which they noted that radicals have an increased probability of recombination in solution as compared to the gas phase. While the term ‘cage’ may encourage one to picture a rigid ensemble of solvent molecules, the cage effect does not describe the influence of a static entity and is actually somewhat difficult to define.

While all agree that the cage is a concept critical to CIDNP, different researchers vary somewhat in their definition of it. Turro *et al* [17] conceive of a RP as guest and a solvent cage as host in an extremely short-lived ‘collision complex’ with a lifetime of about 10^{-11} s. Salikhov *et al* [12] define the cage as a region of effective recombination of two radicals in a RP. They note that radicals may diffuse to the second or third coordination sphere and still come back together to give products. Accordingly, the cage effect describes a twofold influence of condensed media: the two radicals in a RP are not only in close contact for a longer period of time than they would be in the gas phase but are also more likely to re-encounter one another after diffusing apart. Goetz [18] describes the same effects in more abstract terms. He writes of the cage as a region of time and states that two radicals are in the cage so long as they have not lost each other for good. If two radicals in a RP diffuse apart but re-encounter, then they can be said to have been in the cage the entire time. If the same two radicals do not re-encounter, then they are said to have escaped the cage. These different perspectives on the cage are not meant to confuse the reader but are rather intended to present the general idea while conveying the complexity and importance of the concept in CIDNP.

Returning to [figure B1.16.4](#) if the RP is initially in the singlet state, some of the geminate RPs will recombine in the cage in what are called *cage* or *recombination* products. In addition to the reformed radical precursor, recombination products can also include products from disproportionation reactions which occur in cage. A few singlet RPs may escape the cage instead of recombining; as mentioned before, triplet RPs cannot recombine, so they will also escape.

-8-

Escaped radicals diffuse to region II, where J is negligible, and may undergo $S-T_0$ mixing as described previously. From region II, the radicals may follow any of three different pathways.

- (1) The radicals may re-encounter one another following $S-T_0$ mixing. As they diffuse together into the region of large J , the radicals are again constrained to be in either the S or the T_0 state. Some fraction of RPs will have undergone intersystem crossing.
- (2) An individual radical from the RP may encounter a radical from a different RP to form what are known as random RPs or F pairs. F pairs which happen to be in the singlet state have a high probability of recombining, so the remaining F pairs will be in the triplet state. Consequently, the initial condition for F pairs is the triplet state in nearly all cases.

- (3) An individual radical from the RP may be scavenged by a solvent or another chemical species to form diamagnetic products. Because the products are formed following escape from the cage, they are known as *escape* or *scavenging* products.

B1.16.2.4 RADICAL PAIR MECHANISM: NET EFFECT

We have introduced the RP spin Hamiltonian, the mechanism of nuclear spin selective intersystem crossing, and the reaction scheme for RPs has been explained. We now possess all the tools we need to explain qualitatively how nuclear spin polarization arises and manifests itself in a CIDNP spectrum. The RPM may appear in a CIDNP spectrum as a *net effect*, a *multiplet effect*, or a combination of both. The net effect is observed when the g factor difference between radicals R_1 and R_2 (Δg) in a radical pair is large compared with the hyperfine interaction. The simplest example involves a RP with just one hyperfine interaction, as shown in [figure B1.16.5](#). In this example we will set the following conditions: (1) the RP is initially a singlet; (2) the hyperfine coupling constant (a_H) is negative; and (3) the g factor for R_1 is greater than that for R_2 . The recombination product is formed by in-cage recombination of R_1 and R_2 , and a scavenging product is formed by the abstraction of, say, a halogen atom from the solvent following escape from the cage. The scavenging product is formed primarily from escaped triplet RPs although it should be noted that escaped singlets could also form the same product.

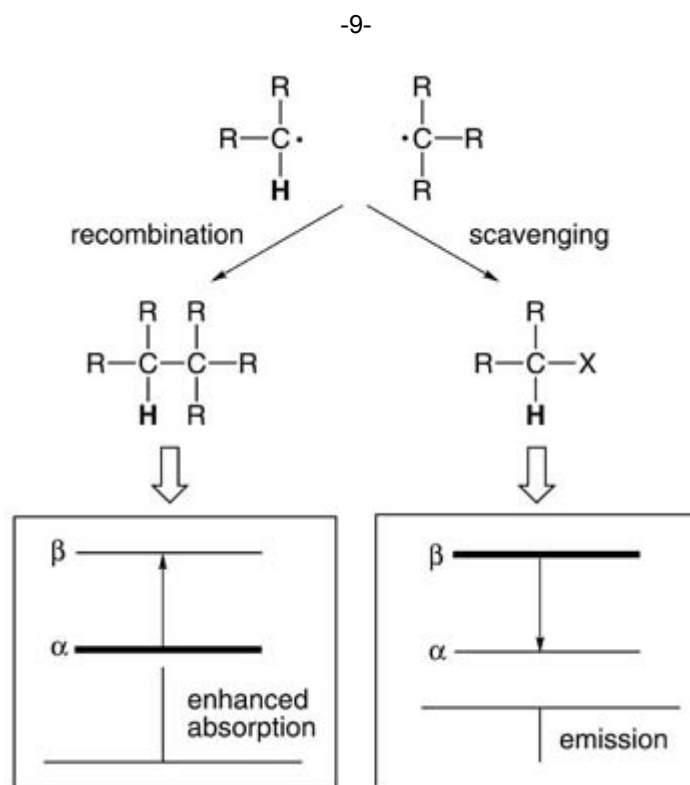


Figure B1.16.5. An example of the CIDNP net effect for a radical pair with one hyperfine interaction. Initial conditions: $g_1 > g_2$; a_H negative; and the RP is initially singlet. Polarized nuclear spin states and schematic NMR spectra are shown for the recombination and scavenging products in the boxes.

In this simple case, there are just two nuclear spin states, α and β . Equation (1.16.5) shows the calculation of the difference in electron precessional frequencies, Q , for nuclear spin states α (equation (B1.16.5a)) and β (equation (B1.16.5b)).

$$2Q = \Delta g \beta_e \hbar^{-1} B_0 + \left(\frac{1}{2}\right)a_H \quad (\text{B1.16.5a})$$

$$2Q = \Delta g \beta_c \hbar^{-1} B_0 + \left(\frac{1}{2}\right) a_H \quad (\text{B1.16.5b})$$

Since Δg is positive and a_H is negative, Q is larger for the β state than for the α state. Radical pairs in the β nuclear spin state will experience a faster intersystem crossing rate than those in the α state with the result that more RPs in the β nuclear spin state will become triplets. The end result is that the scavenging product, which is formed primarily from triplet RPs, will have an excess of spins in the β state while the recombination product, which is formed from singlet RPs, will have an excess of α nuclear spin states.

Relative populations for the α and β states are indicated by the thickness of the lines in the diagrams at the bottom of figure B1.16.5 and the corresponding CIDNP spectrum is shown below each level diagram. The signal for the recombination product, with its excess of α spins, will be in enhanced absorption. This enhanced absorption is distinguished from a typical NMR absorptive signal by its abnormal intensity, which may be as much as 1000 times

-10-

greater than a NMR signal from a Boltzmann population difference. The scavenging product, with its excess of β spins, appears in emission in the CIDNP spectrum. Herein lies an extremely important feature of CIDNP: cage and escape products have opposite phases of polarization. It should be straightforward to see that changing the sign of a_H in the previous example would change the values of Q and would ultimately result in a flipping of the phase of the polarization. A rule for predicting the phase of the polarization for each product will be presented with the next example.

A slightly more complex system exhibiting the RPM net effect is presented in figure B1.16.6 [19]. In this case, radicals R_1 and R_2 each have one hyperfine coupling, so the two protons in the recombination product originate from different radicals. The four nuclear spin states and allowed transitions for the product are shown in figure B1.16.6A along with the NMR spectrum in the absence of spin polarization. Again, we must set some initial conditions for this CIDNP example: the RP is initially in the triplet state, both hyperfine coupling constants are positive and g_1 is greater than g_2 . The values shown on each level in figure B1.16.6B are representative of the absolute values of Q for each nuclear spin state and are proportional to the populations of those states. The phase of each transition can be determined by subtracting the Q value of the upper level from that of the lower. For transitions 1 and 2, this value is $2a_1(\Delta g)$, which is positive and yields absorptive transitions. For transitions 3 and 4, the value $[-2a_2(\Delta g)]$ results in emission transitions. A stick plot of the CIDNP spectrum is shown in the figure. For the CIDNP net effect, each line within a multiplet will always have the same phase.

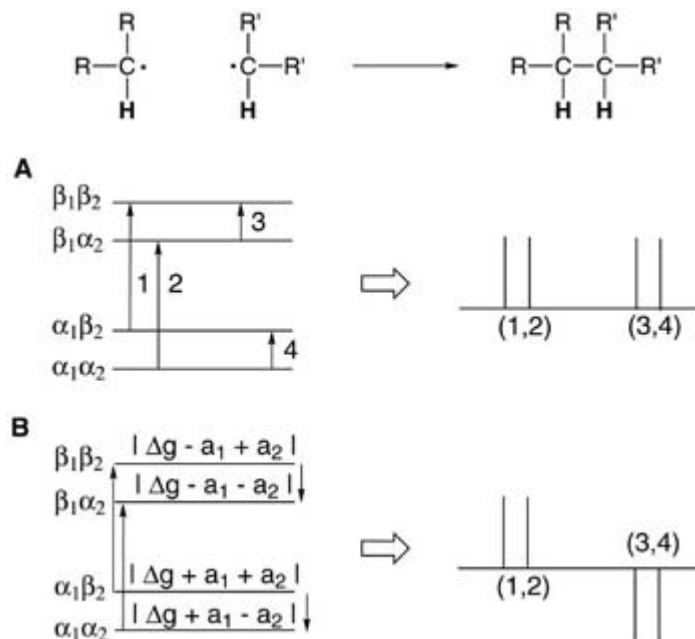


Figure B1.16.6. An example of CIDNP net effect for a radical pair with two hyperfine interactions. Part A shows the spin levels and schematic NMR spectrum for unpolarized product. Part B shows the spin levels and schematic NMR spectrum for polarized product. Populations are indicated on each level. Initial conditions: $g_1 > g_2$; $a_1 > 0$; $a_2 > 0$; spins on different radicals; the RP is initially triplet.

-11-

The phase of a transition in a CIDNP spectrum can be determined using rules developed by Kaptein [20]. The rule for the net effect is shown in equation (B1.16.6). For each term, the sign (+ or -) of that value is inserted, and the final sign determines the phase of the polarization: plus is absorptive and minus is emissive. The variables are defined in the caption to figure B1.16.7.

$$\Gamma_{\text{net}} = \mu \epsilon \Delta g_i a_i \quad \Gamma_{\text{mult}} = \mu \epsilon a_i a_j J_{ij} \sigma_{ij}$$

$\begin{array}{l} + = A \\ - = E \end{array}$
 $\begin{array}{l} + = E/A \\ - = A/E \end{array}$

Figure B1.16.7. Kaptein's rules for net and multiplet RPM of CIDNP. The variables are defined as follows: $\mu = +$ for RP formed from triplet precursor or F pairs and $-$ for RP formed from singlet precursor. $\epsilon = +$ for recombination (or disproportionation)/cage products and $-$ for scavenge/escape products. $\sigma_{ij} = +$ if nuclei i and j were on the same radical and $-$ if nuclei i and j were on different radicals. $\Delta g_i = \text{sign of } (g_1 - g_2)$. $a = \text{sign of hyperfine interaction}$. $J_{ij} = \text{sign of exchange interaction}$.

$$\Gamma_{\text{net}} = \mu \epsilon \Delta g_i a_i \quad (\text{B1.16.6})$$

Kaptein's rule is applied below to each transition in the example in figure B1.16.6. It is important to choose Δg correctly: Δg is equal to $g_1 - g_2$ where g_1 describes the radical containing the nucleus of interest (often a proton) while g_2 is the other radical in the RP. The rule correctly predicts absorptive phase for NMR transitions 1 and 2 and emissive for NMR transitions 3 and 4.

for 1 and 2: $\Gamma_{\text{net}} = (+)(+)(+)(+) = + = A$

for 3 and 4: $\Gamma_{\text{net}} = (+)(+)(-)(+) = - = E$.

B1.16.2.5 RADICAL PAIR MECHANISM: MULTIPLY EFFECT

The other RPM polarization pattern observed in CIDNP spectra is called the multiplet effect. In contrast to the net effect, the multiplet effect occurs when the hyperfine interactions are large compared with Δg . This is best explained by example, and the radical pair for a hypothetical case is shown in [figure B1.16.8](#). We note that only the recombination product will be considered here. Both radicals are identical and have two protons with hyperfine coupling, H_1 and H_2 . The initial conditions for this example are that Δg is zero; the nuclear spin-spin coupling constant, J_{12} is positive; a_1 is negative; a_2 is positive; the RP is initially a singlet; and the nuclear spins are both on the same radical. Values proportional to Q are again shown on each nuclear spin level in [figure B1.16.8](#). Because Δg is zero, the Zeeman term in the equation for Q is zero and, therefore, the value of Q is proportional to the magnitude and the sign of the sum or difference of the hyperfine coupling constants, as shown in [figure B1.16.8](#). Assuming that a_1 and a_2 are roughly equal in magnitude but opposite in sign, it should be clear that the $\alpha\alpha$ and $\beta\beta$ nuclear spin states will have very small Q values while $\alpha\beta$ and $\beta\alpha$ will have larger Q values. Accordingly, $\alpha\beta$ and $\beta\alpha$ will intersystem cross from singlet to triplet more quickly, and these levels will be depleted by escape from the cage relative to the $\alpha\alpha$ and $\beta\beta$ levels. The remaining populations are indicated by the width of the bars in the bottom of [figure B1.16.8](#), and the transitions from more populated levels to less populated ones are shown. As shown in the stick plot CIDNP spectrum, the lines of each multiplet alternate in phase. Because the first line is emissive and the second line is absorptive, this pattern is called E/A for emissive/absorptive.

-12-

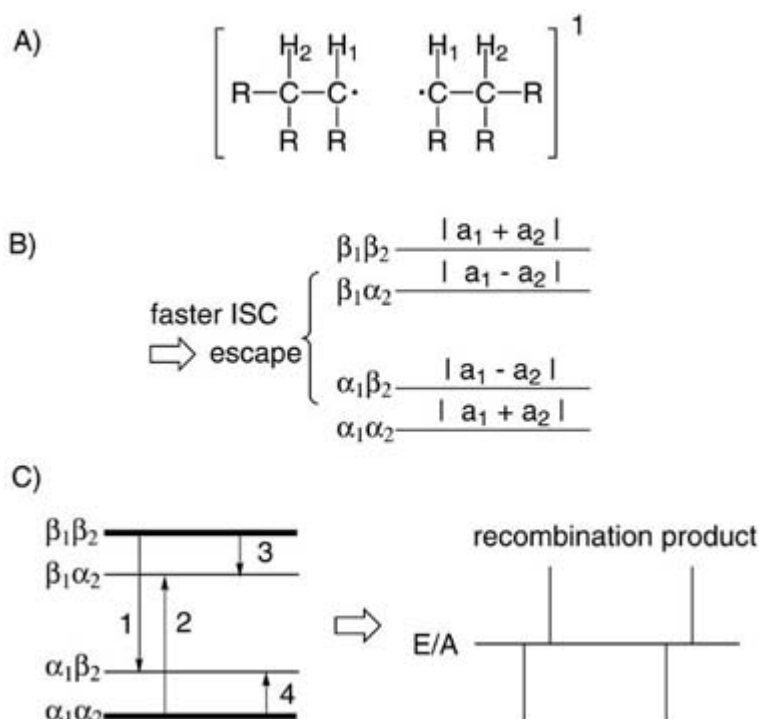


Figure B1.16.8. Example of CIDNP multiplet effect for a symmetric radical pair with two hyperfine interactions on each radical. Part A is the radical pair. Part B shows the spin levels with relative Q values indicated on each level. Part C shows the spin levels with relative populations indicated by the thickness of each level and the schematic NMR spectrum of the recombination product.

Kaptein's rule for the multiplet effect is useful for predicting the phase of each transition, and it is similar to

but has more variables than the rule for the net effect. The variables in equation (B1.16.7) are defined in [figure B1.16.7](#). A final sign of plus predicts E/A phase while minus predicts A/E.

$$\Gamma_{\text{mult}} = \mu \epsilon a_i a_j J_{ij} \sigma_{ij}. \quad (\text{B1.16.7})$$

The application of Kaptein's rule to the example in [figure B1.16.8](#) is shown below, and it correctly predicts E/A multiplets.

$$\Gamma_{\text{mult}} = (-)(+)(-)(+)(+)(+) = + = \text{E/A}.$$

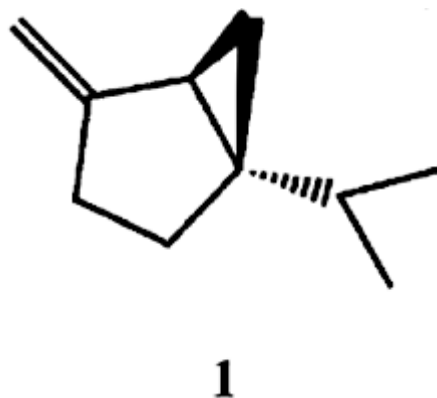
One of the most attractive features of the CIDNP multiplet effect is that it allows determination of the sign of the J coupling, which is often difficult to do by other methods.

-13-

B1.16.2.6 EXAMPLES OF CIDNP

While the stick plot examples already presented show net and multiplet effects as separate phenomena, the two can be observed in the same spectrum or even in the same NMR signal. The following examples from the literature will illustrate 'real life' uses of CIDNP and demonstrate the variety of structural, mechanistic, and spin physics questions which CIDNP can answer.

Roth *et al* [10] have used CIDNP to study the structures of vinylcyclopropane radical cations formed from precursors such as sabinene (**1**).

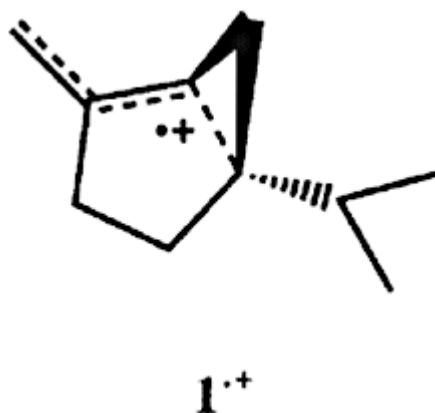


The radical cation of **1** ($\mathbf{1}^{\cdot+}$) is produced by a photo-induced electron transfer reaction with an excited electron acceptor, chloranil. The major product observed in the CIDNP spectrum is the regenerated electron donor, **1**. The parameters for Kaptein's net effect rule in this case are that the RP is from a triplet precursor (μ is +), the recombination product is that which is under consideration (ϵ is +) and Δg is negative. This leaves the sign of the hyperfine coupling constant as the only unknown in the expression for the polarization phase. Roth *et al* [10] used the phase and intensity of each signal to determine the relative signs and magnitudes of the hyperfine coupling constant for each proton in $\mathbf{1}^{\cdot+}$. Signals in enhanced absorption indicated negative hyperfine coupling constants while emissive signals indicated positive hyperfines.

The CIDNP spectrum is shown in [figure B1.16.1](#) from the introduction, top trace, while a dark spectrum is shown for comparison in [figure B1.16.1](#) bottom trace. Because the sign and magnitude of the hyperfine coupling constant can be a measure of the spin density on a carbon, Roth *et al* [10] were able to use the relative spin density of each carbon to determine that the structure of the radical cation $\mathbf{1}^{\cdot+}$ is a delocalized one, shown below. This example demonstrates the use of CIDNP to determine the signs and relative

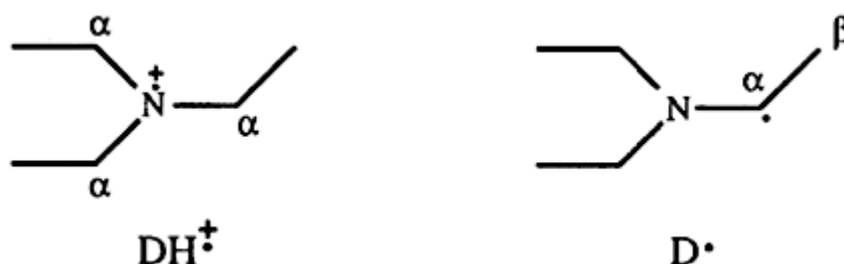
magnitudes of the hyperfine coupling constants and to assign the structure of an intermediate.

-14-



In a case involving both net and multiplet effects, Goetz and Sartorius [21] studied the photoreaction of triethylamine with various triplet sensitizers containing carbonyl functionalities. In a two-step process, the amine (DH) first transfers an electron to the excited sensitizer and forms the aminium cation DH^+ . The aminium cation is then deprotonated to form a neutral α -aminoalkyl radical (D^{\cdot}), which can go on to form products. In this example, triethylamine (DH) was reacted with a variety of sensitizers, and N,N-diethylvinylamine was the polarized product which was studied. N,N-diethylvinylamine can be formed by two different pathways. If the deprotonation step occurs in cage, H^+ is transferred to the sensitizer, and polarization in the product arises from a neutral radical pair, $AH^{\cdot}D^{\cdot}$. If the deprotonation step occurs out of cage, then H^+ will be abstracted by free amine; in this case, polarization is formed from a radical ion pair, $A^{\cdot-}DH^+$. The goal of this work was to determine the intermediates leading to N,N-diethylvinylamine; does the deprotonation step occur in cage or out of cage?

The radical cation and neutral radical derived from triethylamine are shown below.



$DH^{\cdot+}$ has only one non-negligible hyperfine, $a_{H\alpha} = +19.0$ G while D^{\cdot} has two significant hyperfines, $a_{H\alpha} = -13.96$ G and $a_{H\beta} = +19.24$ G. Clearly, these two radicals will lead to very different polarizations in the CIDNP spectrum of both cage and escape products.

Figure B1.16.9 shows background-free, pseudo-steady-state CIDNP spectra of the photoreaction of triethylamine with (a) anthroquinone as sensitizer and (b) and (c) xanthone as sensitizer. Details of the pseudo-steady-state CIDNP method are given elsewhere [22]. In trace (a), no signals from the β protons of products 1 (recombination) or 2 (escape) are observed, indicating that the products observed result from the radical ion pair. Traces (b) and (c) illustrate a useful feature of pulsed CIDNP: net and multiplet effects may be separated on the basis of their radiofrequency (RF) pulse tip angle dependence [23]. Net effects are shown in trace (b) while multiplet effects can

be seen in (c). Both traces show signals from the β protons of products 1 and 2, indicating that these products were formed from a neutral radical pair intermediate. It was ultimately determined that the time scale of the deprotonation step relative to the lifetime of the radical ion pairs determined whether products were formed from radical ion pairs or from neutral radical pairs. The energetics of the system varied with the sensitizer, and results were compiled for a variety of sensitizers. This example illustrates the very common application of CIDNP as a mechanistic tool.

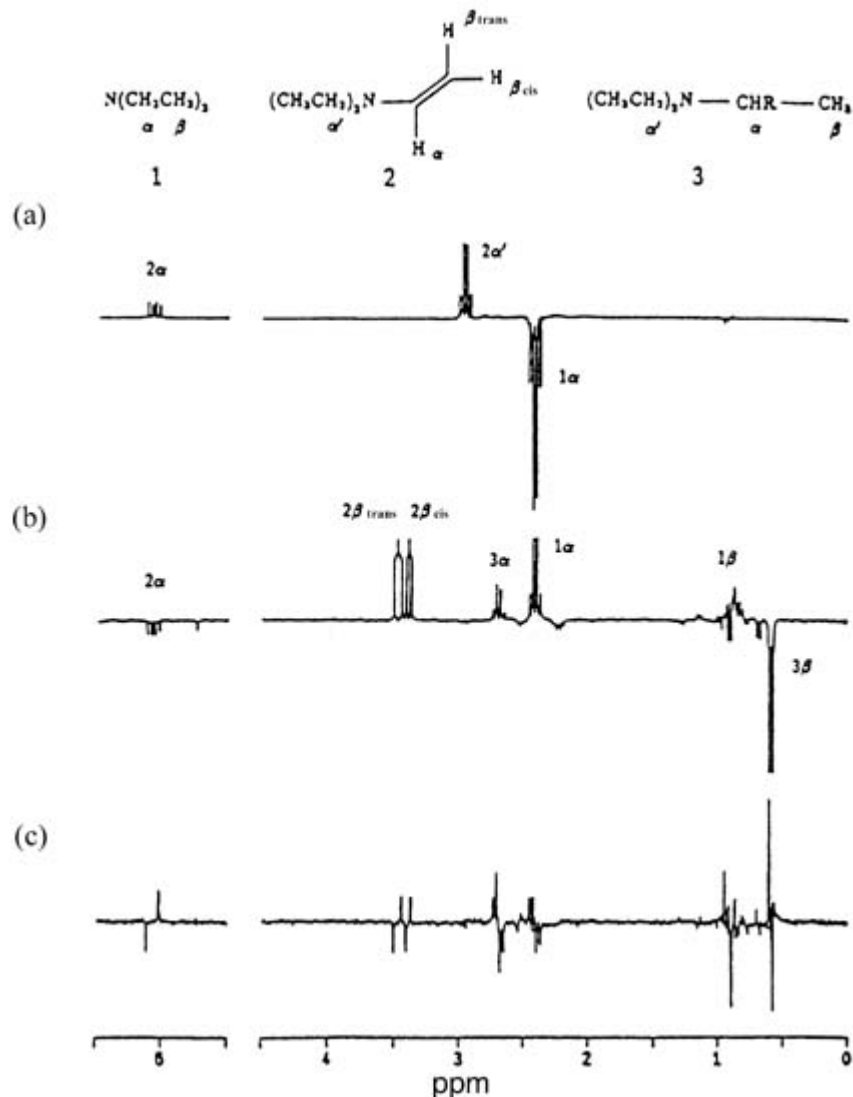
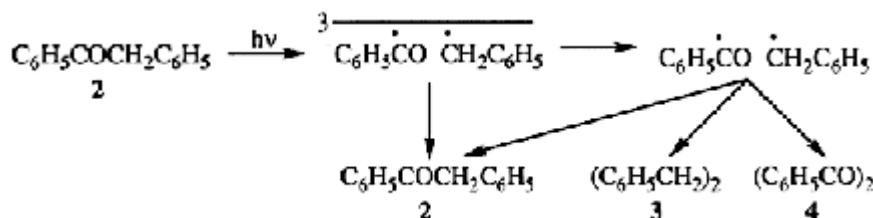


Figure B1.16.9. Background-free, pseudo-steady-state CIDNP spectra observed in the photoreaction of triethylamine with different sensitizers ((a), anthraquinone; (b), xanthone, CIDNP net effect; (c), xanthone, CIDNP multiplet effect, amplitudes multiplied by 1.75 relative to the centre trace) in acetonitrile- d_3 . The structural formulae of the most important products bearing polarizations (1, regenerated starting material; 2, N,N-diethylvinylamine; 3, combination product of amine and sensitizer) are given at the top; R denotes the sensitizer moiety. The polarized resonances of these products are assigned in the spectra. Reprinted from [21].

In an extension of traditional CIDNP methods, Closs and co-workers developed time-resolved CIDNP (TR-CIDNP) in the late 1970s [24, 25 and 26]. The initial time-resolved experiments had a time resolution in the

microsecond range [24], but a nanosecond method was later developed [27]. A typical pulse sequence for time-resolved CIDNP involves a series of saturation pulses to remove background signals from equilibrium polarization followed by a laser pulse to form the radical pairs. After a preset delay time, τ , after the laser flash, a RF pulse is applied, and the FID of the product is acquired. Further details of this experiment are given in [26].

The first application of the time-resolved CIDNP method by Closs and co-workers involved the Norrish 1 cleavage of benzyl phenyl ketone [24, 25]. Geminate RPs may recombine to regenerate the starting material while escaped RPs may form the starting ketone (2), bibenzyl (3), or benzil (4), as shown below.



Closs *et al* [25] plotted the polarizations versus time of the starting ketone (2, trace B, emissive signal) and the bibenzyl escape product (3, trace A, absorptive signal), as shown in figure B1.16.10. Ketone 2 can be formed either by recombination of geminate pairs or the reaction of F pairs; in either case, the polarization will be emissive. At the earliest delay time (1 μs), the emissive signal of 2 is already present while no polarization from F pairs is apparent because there has not been sufficient time for the diffusion of radicals to occur. At later delay times, both absorptive and emissive polarizations grow until they reach a maximum. In order to demonstrate that much of the emissive polarization was due to production of 2 from F pairs, a thiol scavenger was added to trap the escaped benzoyl radicals. As shown in trace C, the emissive polarization is constant from the earliest delay time when the scavenger is present, indicating that much of the emissive polarization from geminate pairs is constant, while that from F pairs grows in with time. This was the first instance in which polarization from F pairs was shown to enhance the polarization of geminate products. In addition to establishing the utility of the time-resolved CIDNP method, this experiment was the first to demonstrate that polarization from cage and escape products could be separated based on the time scale of its production.

-17-

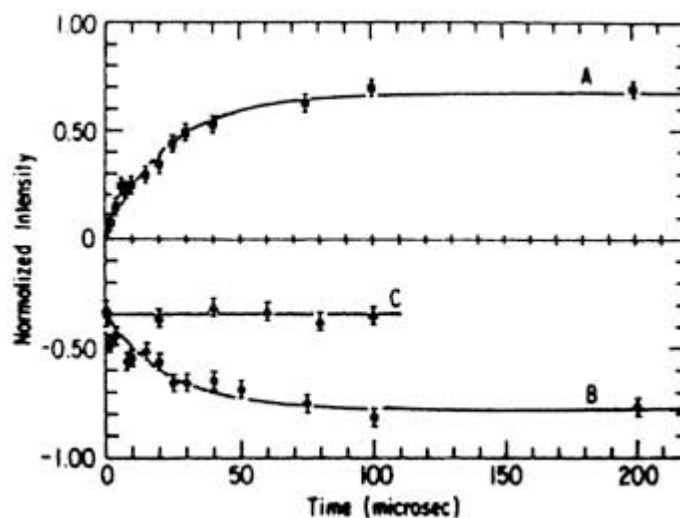


Figure B1.16.10. Intensity of CH_2 resonance as a function of delay time in A dibenzyl, B deoxybenzoin, and C deoxybenzoin in presence of thiol scavenger. Reprinted from [25].

While the earliest TR-CIDNP work focused on radical pairs, biradicals soon became a focus of study. Biradicals are of interest because the exchange interaction between the unpaired electrons is present throughout the biradical lifetime and, consequently, the spin physics and chemical reactivity of biradicals are markedly different from radical pairs. Work by Morozova *et al* [28] on polymethylene biradicals is a further example of how this method can be used to separate net and multiplet effects based on time scale [28]. [Figure B1.16.11](#) shows how the cyclic precursor, 2,12-dihydroxy-2,12-dimethylcyclododecanone, cleaves upon 308 nm irradiation to form an acyl-ketyl biradical, which will be referred to as the primary biradical since it is formed directly from the cyclic precursor. The acyl-ketyl primary biradical decarbonylates rapidly ($k_{CO} \geq 5 \times 10^7 \text{ s}^{-1}$) to form a bis-ketyl biradical, which will be referred to as the secondary biradical. Both the primary and secondary biradicals can form a number of diamagnetic products, as shown in [Figure B1.16.11](#).

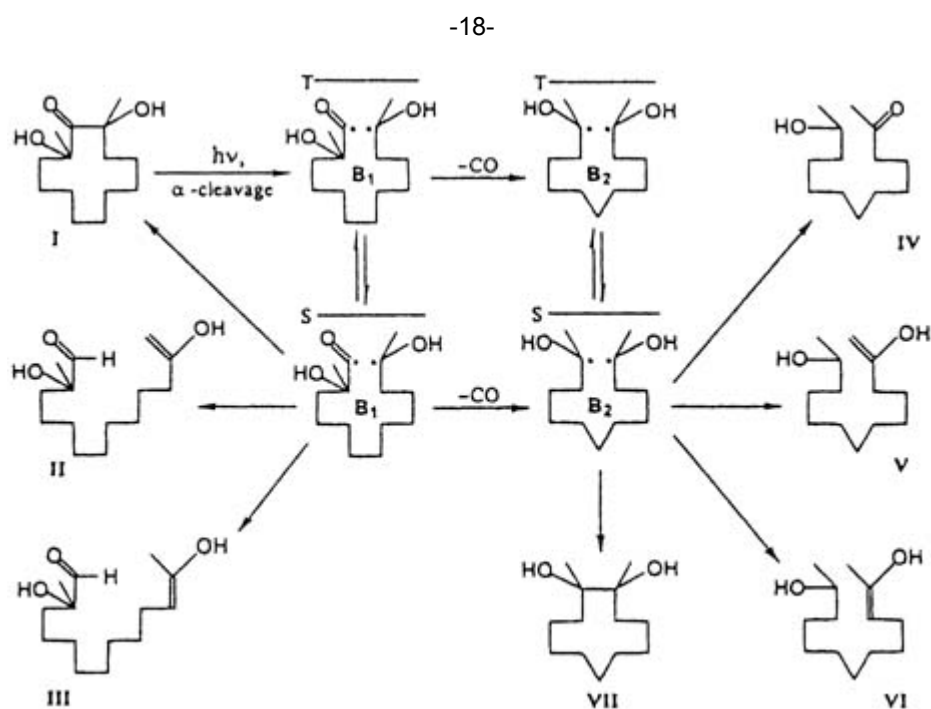


Figure B1.16.11. Biradical and product formation following photolysis of 2,12-dihydroxy-2,12-dimethylcyclododecanone. Reprinted from [28].

In the TR-CIDNP spectrum, the methyl protons of products IV, V, and VI, formed from the secondary biradical, show a combination of net and multiplet polarizations. Morozova *et al* [28] measured separately the time dependence of the net and multiplet polarizations for this group of protons, and the results are shown in [figure B1.16.12](#) and [figure B1.16.13](#) respectively. Clearly, the net and multiplet polarizations develop on different time scales; while the net polarization is constant after approximately 1 μs , the multiplet polarization takes much longer to evolve. It was determined that this difference arises because the net polarization in these products of the secondary biradical is actually inherited from the primary biradical, while the multiplet polarization is generated in the secondary biradical. If chemical transformation (decarbonylation in this case) is fast compared with the rate of intersystem crossing, then a secondary biradical or radical pair may inherit polarization from its precursor. This is known as the *memory effect* in CIDNP, and this work was the first report of the memory effect in biradicals. The polarization inherited from the primary biradical is net because $\Delta g > 0$ in the primary biradical; because the secondary biradical is symmetric, $\Delta g = 0$, and only multiplet polarization can be generated. It was also determined in this study that the kinetics of the net effect reflect the decay of the T_0 level while the multiplet effect corresponds to the decay of the T_+ and T_- levels; the reasons for these observations are beyond the scope of this presentation, but the interested reader is directed to the references for additional details.

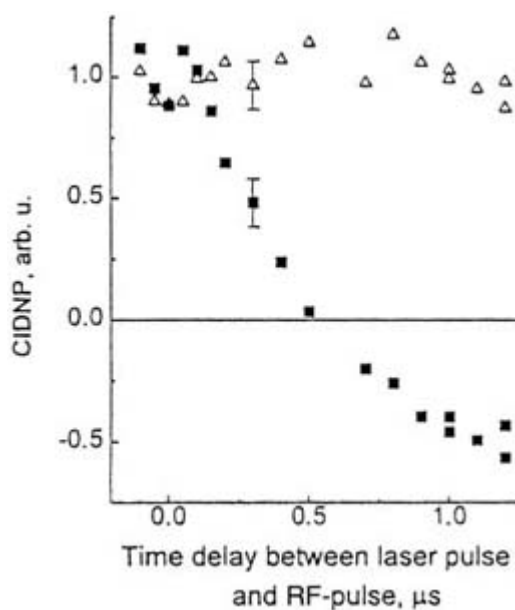


Figure B1.16.12. Experimental kinetics of the CIDNP net effect: (Δ) for the aldehyde proton of the products II and III of primary biradical; (\blacksquare) for the $\text{CH}_3\text{CH}(\text{OH})$ protons of the products IV, V, and VI of secondary biradical. Reprinted from [28].

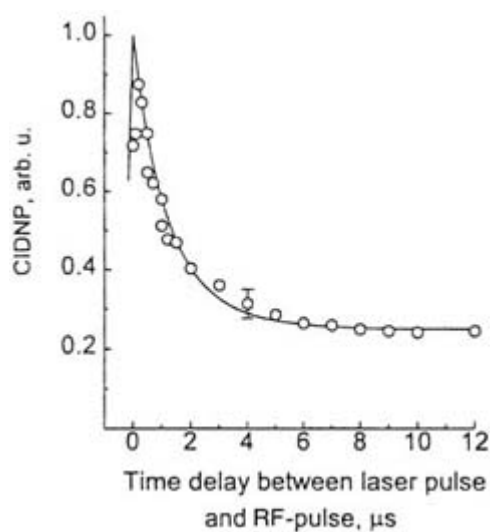


Figure B1.16.13. Kinetics of the CIDNP multiplet effect: (full curve) the calculated CIDNP kinetics for the product of disproportionation of bis-ketyl biradical; (O) experimental kinetics for the $\text{CH}_3\text{CH}(\text{OH})$ protons of the products IV, V and VI of the secondary biradical. Reprinted from [28].

B1.16.3 CIDEP

As the electron counterpart to CIDNP, CIDEP can provide different but complementary information on free radical systems. Whereas CIDNP involves the observation of diamagnetic products, the paramagnetic

intermediates themselves are observed in CIDEP studies. Unlike CIDNP, CIDEP does not require chemical reaction for the formation of polarization, as we shall see below. In addition to the RPM, three other mechanisms may produce electron polarization: the triplet mechanism (TM), the radical-triplet pair mechanism (RTPM) and the spin-correlated radical pair (SCRIP) mechanism. Some of these mechanisms provide information which can lead directly to the structural identification of radical intermediates while all of them supply data which may be used to elucidate mechanisms of radical reactions.

Experimentally, the observation of CIDEP is difficult but not impossible using a commercial steady-state EPR spectrometer with 100 kHz field modulation [29]. The success of this particular experiment requires large steady-state concentrations of radicals and rather slow spin relaxation, as the time response of the instrument is, at best, 10 μ s. Therefore, the steady-state method works well for only a limited number of systems and generally requires a very strong CW lamp for irradiation. The time response can be shortened if the radicals are produced using a pulsed laser and the 100 kHz modulation is bypassed. The EPR signal is then taken directly from the preamplifier of the microwave bridge and passed to a boxcar signal averager [30] or transient digitizer [31]. At the standard X-band frequency the time response can be brought down to about 60 ns in this fashion. In this case the overall response becomes limited by the resonant microwave cavity quality factor [32]. At higher frequencies such as Q-band, the time response can be limited by laser pulse width or preamplifier rise time (< 10 ns) [33]. Using the boxcar or digitizer method, CIDEP is almost always observable, and this so-called time-resolved (CW) EPR spectroscopy is the method of choice for many practitioners. The ‘CW’ in the name is used to indicate that the microwaves are always on during the experiment, even during the production of the radicals, as opposed to pulsed microwave methods such as electron-spin-echo or Fourier-transform (FT) EPR. Significant advantages in sensitivity with similar time response are available with FT-EPR, but there are also disadvantages in terms of the spectral width of the excitation that limit the application of this technique [34]. The TR (CW) method is the most facile and cost effective method for the observation of complete EPR spectra exhibiting CIDEP on the sub-microsecond time scale.

B1.16.3.1 RADICAL PAIR MECHANISM

The RPM has already been introduced in our explanation of CIDNP. The only difference is that for the electron spins to become polarized, product formation from the geminate RP is not required. Rather, in a model first introduced by Adrian, so-called ‘grazing encounters’ of geminate radical pairs are all that is required [35]. Basically the RP must diffuse from a region where the exchange interaction is large to one where it is small, then back again. The spin wavefunction evolution that mixes the S and T_0 electronic levels in the region of small J leads to unequal populations of the S and T_0 states in the region of large J . The magnitude of the RPM CIDEP is proportional to this population difference.

-21-

As for CIDNP, the polarization pattern is multiplet (E/A or A/E) for each radical if Δg is smaller than the hyperfine coupling constants. In the case where Δg is large compared with the hyperfines, net polarization (one radical A and the other E or *vice versa*) is observed. A set of rules similar to those for CIDNP have been developed for both multiplet and net RPM in CIDEP (equation (B1.16.8) and equation (B1.16.9)) [36]. In both expressions, μ is positive for triplet precursors and negative for singlet precursors. J is always negative for neutral RPs, but there is evidence for positive J values in radical ion reactions [37]. In equation (B1.16.8), $\Gamma_{\text{mult}} = +$ predicts E/A while $\Gamma_{\text{mult}} = -$ predicts A/E. For the net effect in equation (B1.16.9), $\Gamma_{\text{net}} = +$ predicts A while $\Gamma_{\text{net}} = -$ predicts E.

$$\Gamma_{\text{mult}} = -J\mu \quad (\text{B1.16.8})$$

$$\Gamma_{\text{net}} = +J\Delta g\mu. \quad (\text{B1.16.9})$$

Because the number of grazing encounters is a function of the diffusion coefficient, CIDEP by the RPM mechanism is a strong function of the viscosity of the solvent and, in general, the RPM becomes stronger with increasing viscosity. Pedersen and Freed [39] have developed analytical techniques for the functional form of the viscosity dependence of the RPM.

A typical example of RPM multiplet effects is shown in [figure B1.16.2](#). Upon 308 nm laser irradiation, the O–O bond of a fluorinated peroxide dimer is cleaved to yield two radicals plus CO₂ as shown in [figure B1.16.2](#). The radical signal is split into a doublet by the α -fluorine atom, and each line in the doublet is split into a quartet by the adjacent CF₃ group (although not all lines are visible in these spectra). The A/E pattern of the spectrum in [figure B1.16.2A](#) indicates that the RP is formed from a singlet precursor. When benzophenone, a triplet sensitizer, is added to the system, the precursor becomes a triplet, and the polarization pattern is now E/A, as shown in [figure B1.16.2B](#). These spectra demonstrate the utility of RPM CIDEP in determining the spin multiplicity of the precursor.

B1.16.3.2 TRIPLET MECHANISM

A second mechanism of CIDEP is the triplet mechanism (TM) [40, 41]. As the name implies, this polarization is generated only when the RP precursor is a photoexcited triplet state. The polarization is produced during the intersystem crossing process from the first excited singlet state of the molecular precursor. It should be noted that this intersystem crossing, which will be explained in detail later, is to be distinguished from that described for RPs described above. Because the TM polarization is present before the triplet reacts to produce the RP, its phase is either net E or net A for both radicals. The origin of the polarization is as follows: in the intersystem crossing process, which is dominated by spin–orbit interactions, the most suitable basis set is one where the canonical orientations of the triplet state are represented. An example is shown in [figure B1.16.14](#) where the directions of the triplet T_x , T_y , and T_z basis functions for naphthalene are indicated in their usually defined orientations. These are also sometimes called the ‘zero-field’ basis functions. As the electrons undergo the intersystem crossing process, these are the orbital directions they ‘see’. This is called the molecular frame of reference.

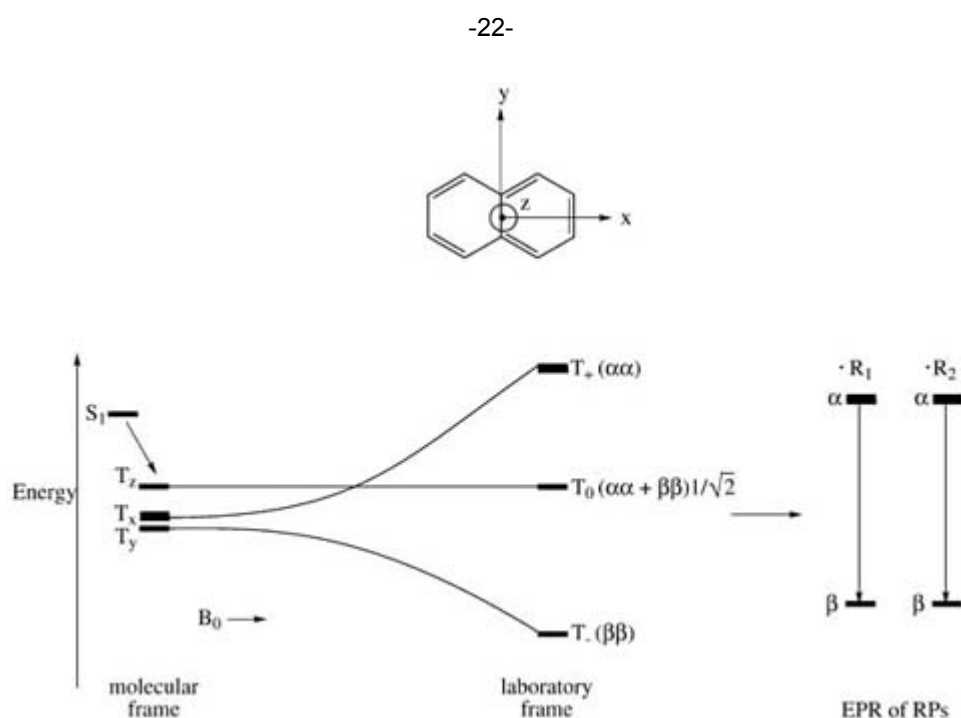


Figure B1.16.14. Top, the canonical axes for triplet naphthalene. The z-axis is directed out of the plane of the paper. Bottom, energy levels and relative populations during the CIDEP triplet mechanism process. See text

for further details.

Because the spin–orbit interaction is anisotropic (there is a directional dependence of the ‘view’ each electron has of the relevant orbitals), the intersystem crossing rates from S_1 to each triplet level are different. Therefore, unequal populations of three triplet levels results. The T_x , T_y and T_z basis functions can be rewritten as linear combinations of the familiar α and β spin $\frac{1}{2}$ functions and, consequently, they can also be rewritten as linear combinations of the high-field RP spin wavefunctions T_+ , T_0 and T_- , which we have already described above. The net polarization generated in the zero-field basis is carried over to the high-field basis set and, consequently, the initial condition for the geminate RP is that the population of the triplet levels is not strictly equal ($\frac{1}{3}$ each). Exactly which triplet level is overpopulated depends on the sign of the zero-field splitting parameter D in the precursor triplet state. A representative energy level diagram showing the flow of population throughout the intersystem crossing, RP and free radical stages is shown in figure B1.16.14.

The absolute magnitude of the TM polarization intensity is governed by the rate of rotation of the triplet state in the magnetic field. If the anisotropy of the zero-field states is very rapidly averaged (low viscosity), the TM is weak. If the experiment is carried out in a magnetic field where the Zeeman interaction is comparable with the D value and the molecular tumbling rate is slow (high viscosity), the TM is maximized. Additional requirements for a large TM polarization are: (1) the intersystem crossing rates from S_1 to T_x , T_y and T_z must be fast relative to the RP production step, and (2) the spin relaxation time in the excited triplet state should not be too short in order to ensure a large TM polarization. An example of the triplet mechanism from work by Jent *et al* [42] is shown in figure B1.16.15. Upon laser flash photolysis, dimethoxyphenylacetophenone (DMPA, **5**) forms an excited singlet and undergoes fast intersystem crossing and subsequent photocleavage of the triplet to form radicals **6** and **7** as shown below.

-23-

Radical **7** can subsequently fragment to form methyl radical (**8**) and methylbenzoate (**9**).

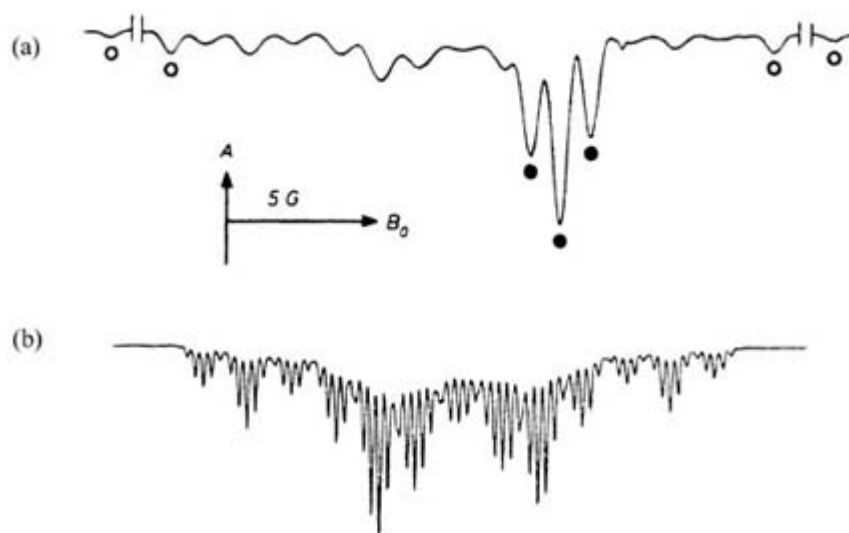
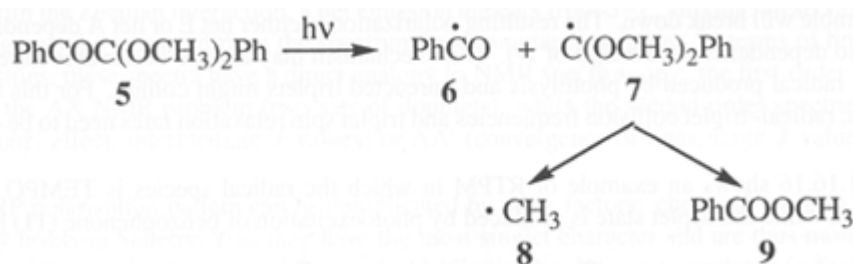


Figure B1.16.15. TREPR spectrum after laser flash photolysis of 0.005 M DMPA (**5**) in toluene. (a) 0.7 μ s, 203 K, RF power 10 mW; O, lines $\cdot\text{CH}_3$ (**8**), spacing 22.8 G; \bullet , benzoyl (**6**), remaining lines due to (**7**). (b) 2.54 μ s, 298 K, RF power 2 mW to avoid nutations, lines of **7** only. Reprinted from [42].



Contradictory evidence regarding the reaction to form **8** and **9** from **7** led the researchers to use TREPR to investigate the photochemistry of DMPA. Figure B1.16.15A shows the TREPR spectrum of this system at 0.7 μs after the laser flash. Radicals **6**, **7** and **8** are all present. At 2.54 μs , only **7** can be seen, as shown in figure B1.16.15B. All radicals in this system exhibit an emissive triplet mechanism. After completing a laser flash intensity study, the researchers concluded that production of **8** from **7** occurs upon absorption of a second photon and not thermally as some had previously believed.

B1.16.3.3 RADICAL-TRIPLET PAIR MECHANISM

In the early 1990s, a new spin polarization mechanism was postulated by Paul and co-workers to explain how polarization can be developed in transient radicals in the presence of excited triplet state molecules (Blättler *et al* [43], Blättler and Paul [44], Goudsmit *et al* [45]). While the earliest examples of the radical-triplet pair mechanism (RTPM) involved emissive polarizations similar in appearance to triplet mechanism polarizations, cases have since been discovered in which absorptive and multiplet polarizations are also generated by RTPM.

Polarization obtained by RTPM is related to the RPM in that diffuse encounters are still required, but differs in that it involves the interaction of a photoexcited triplet state with a doublet state radical. When a doublet state (electron spin $\frac{1}{2}$) radical is present in high concentration upon production of a photoexcited triplet, the doublet and triplet interact to form new quartet and doublet states. When the two species find themselves in regions of effective exchange ($|J| > 0$), a fluctuating dipole–dipole interaction (D) induces transitions between states, leading to a population redistribution that is non-Boltzmann, i.e. CIDEP. This explanation of RTPM is only valid in regions of moderate viscosity. If the motion is too fast, the assumption of a static ensemble will break down. The resulting polarization is either net E or net A depending on the sign of J (there is no dependence on the sign of D). This mechanism may also be observable in reactions where a doublet state radical produced by photolysis and unreacted triplets might collide. For this to happen, the triplet lifetimes, radical–triplet collision frequencies and triplet spin relaxation rates need to be of comparable time scales.

Figure B1.16.16 shows an example of RTPM in which the radical species is TEMPO (**10**), a stable nitroxide radical, while the triplet state is produced by photoexcitation of benzophenone (**11**) [45].

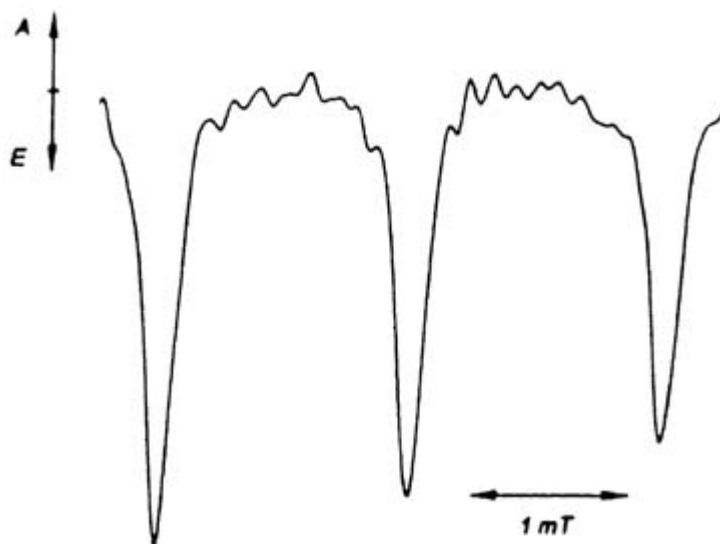
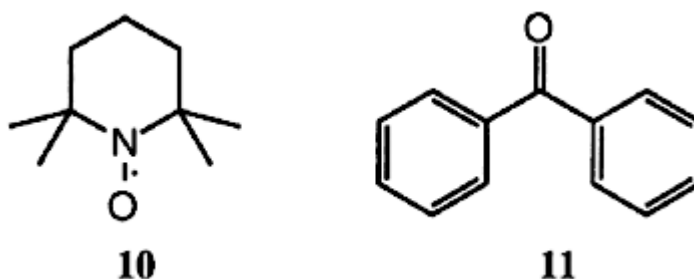


Figure B1.16.16. TREPR spectrum of TEMPO radicals in 1,2-epoxypropane solution with benzophenone, 1 μ s after 308 nm laser flash. Reprinted from [45].



-25-

The three-line spectrum with a 15.6 G hyperfine reflects the interaction of the TEMPO radical with the nitrogen nucleus ($I = 1$); the benzophenone triplet cannot be observed because of its short relaxation times. The spectrum shows strong net emission with weak E/A multiplet polarization. Quantitative analysis of the spectrum was shown to match a theoretical model which described the size of the polarizations and their dependence on diffusion.

B1.16.3.4 SPIN CORRELATED RADICAL PAIR MECHANISM

The fourth and final CIDEP mechanism results from the observation of geminate radical pairs when they are still interacting, i.e. there is a measurable dipolar or exchange interaction between the components of the RP at the time of measurement. It is called the spin correlated radical pair (SCRCP) mechanism and is found under conditions of restricted diffusion such as micelle-bound RPs [46, 47] or covalently linked biradicals [48] and also in solid state structures such as photosynthetic reaction centres [49] and model systems [50]. In this mechanism additional lines in the EPR spectrum are produced due to the interaction. If the D or J value is smaller than the hyperfines, then the spectrum is said to be first order, with each individual hyperfine line split by $2J$ or $2D$ into a doublet. The most unusual and immediately recognizable feature of the SCRCP mechanism with small interactions is that each component of the doublet receives an opposite phase. For triplet precursors and negative J values, which is the common situation, the doublets are E/A. The level diagram in figure B1.16.17 shows the origin of the SCRCP polarization for such a system, considering only one hyperfine line.

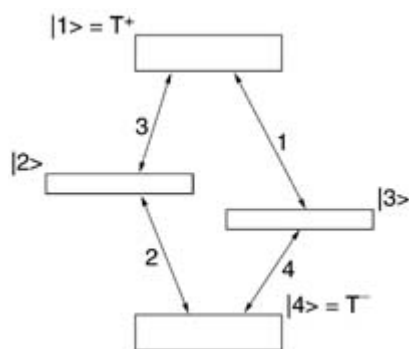


Figure B1.16.17. Level diagram showing the origin of SCRCP polarization.

When the J or D coupling exceeds the hyperfine couplings, the spectrum becomes second order and is much more complex. Lines are alternating with E or A phase, and, if J or D becomes even larger or becomes comparable with the Zeeman interaction, a net emission appears from $S-T$ mixing. In second-order spectra the J coupling must be extracted from the spectrum by computer simulation. In terms of line positions and relative intensities, these spectra have a direct analogy to NMR spectroscopy: the first-order spectrum is the equivalent of the AX NMR problem (two sets of doublets), while the second-order spectrum is analogous to the AB ('roof' effect, intermediate J values) or AA' (convergence of lines, large J values) nuclear spin system [51].

-26-

The SCRCP polarization pattern can be complicated by other factors: chemical reaction can deplete the middle energy levels in Scheme Y as they have the most singlet character and are thus more likely to react upon encounter. Spin relaxation via either correlated (dipole-dipole) or uncorrelated (g factor or hyperfine anisotropy) mechanisms can also redistribute populations on the TREPR time scale [52]. The most important relaxation mechanism is that due to modulation of the exchange interaction caused by conformational motion which changes the inter-radical distance on the EPR time scale. Here both the T_1 and T_2 relaxation processes are important [53]. Interestingly, J modulation is also the process by which RPM is produced, and recently it has been demonstrated that at certain viscosities, both RPM and SCRCP polarization patterns can be observed simultaneously in both micellar [54] and biradical-type RPs [55]. The presence of SCRCP polarization in biradicals has enabled much information to be obtained regarding weak electronic couplings in flexible systems as a function of molecular structure, solvent, and temperature [56, 57, 58 and 59]. The spin polarization observed in EPR spectra of photosynthetic reaction centres has also proven informative in relating structure to function in those systems, especially in comparing structural parameters measured magnetically to those found by other methods such as x-ray crystallography [60].

Most observations of SCRCP have been from triplet precursors, but Fukuju *et al* [61] have observed singlet-born SCRCP upon photolysis of tetraphenylhydrazine in sodium dodecyl sulfate (SDS) micelles. The tetraphenylhydrazine (**12**) cleaves to form two diphenylamino (DPA) radicals, as shown below.

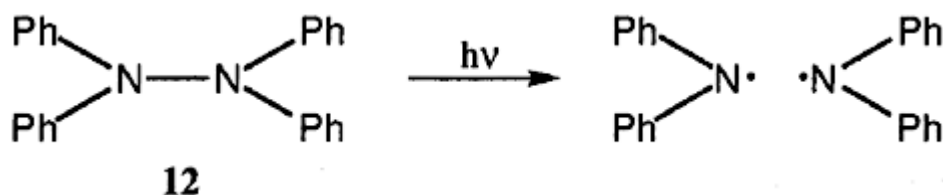


Figure B1.16.18 shows TREPR spectra of this system in SDS micelles at various delay times. The A/E/A/E pattern observed at early delay times is indicative of a singlet-born SCRCP. Over time, a net absorptive

component develops and, eventually, the system inverts to an E/A/E/A pattern at late delay times. The long lifetime of this SCRIP indicates that the DPA radicals are hydrophobic enough that they prefer to remain in the micelle rather than escape. The time dependence of the spectra can be described by a kinetic model which considers the recombination process and the relaxation between all states of the RP. The reader is directed to the reference for further details.

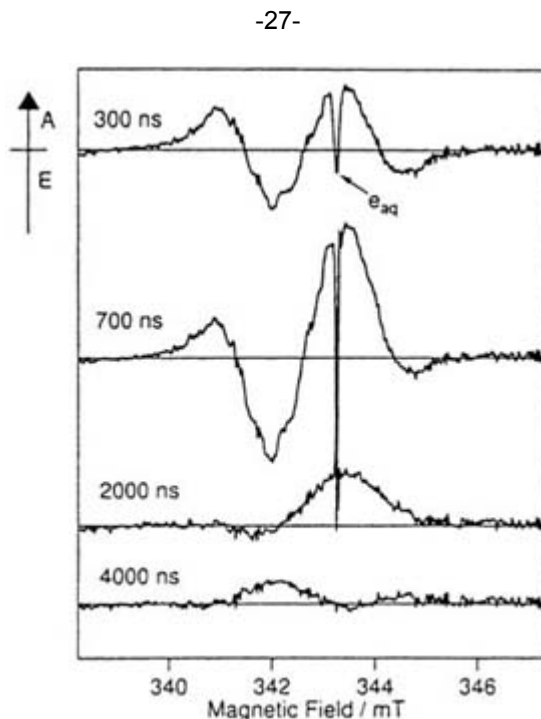


Figure B1.16.18. TREPR spectra observed after laser excitation of tetraphenylhydrazine in an SDS micelle at room temperature. Reprinted from [61].

B1.16.3.5 FURTHER EXAMPLES OF CIDEP

While each of the previous examples illustrated just one of the electron spin polarization mechanisms, the spectra of many systems involve polarizations from multiple mechanisms or a change in mechanism with delay time.

Work by Koga *et al* [62] demonstrates how the polarization mechanism can change upon alteration of the chemical environment. Upon laser flash photolysis, excited xanthone abstracts a proton from an alcohol solvent, cyclohexanol in this case. The xanthone ketyl radical ($\cdot\text{XnH}$) and the alcoholic radical ($\cdot\text{ROH}$) exhibit E/A RPM polarization with slight net emission, as shown in [figure B1.16.19\(a\)](#). Upon addition of HCl to the cyclohexanol solution, the same radicals are observed, but the polarization is now entirely an absorptive triplet mechanism, as shown in [figure B1.16.19 \(b\)](#). It was determined that both H^+ and Cl^- or HCl molecules must be present for this change in mechanism to occur, and the authors postulated that the formation of a charge transfer complex between xanthone and HCl in their ground state might be responsible for the observed change in polarization. This curious result demonstrates how a change in the CIDEP mechanism can yield information about chemical changes which may be occurring in the system.

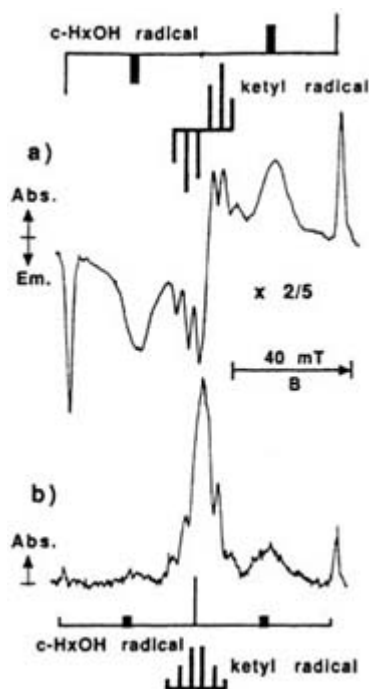


Figure B1.16.19. (a) CIDEP spectrum observed in the photolysis of xanthone (1.0×10^{-2} M) in cyclohexanol at room temperature. The stick spectra of the ketyl and cyclohexanol radicals with RPM polarization are presented. (b) CIDEP spectrum after the addition of hydrochloric acid (4.1 vol%; HCl 0.50 M) to the solution above. The stick spectra of the ketyl and cyclohexanol radicals with absorptive TM polarization are presented. The bold lines of the stick spectra of the cyclohexanol radical show the broadened lines due to ring motion of the radical. Reprinted from [62].

Utilizing FT-EPR techniques, van Willigen and co-workers have studied the photoinduced electron transfer from zinc tetrakis(4-sulfonatophenyl)porphyrin (ZnTPPS) to duroquinone (DQ) to form ZnTPPS^{3-} and DQ^- in different micellar solutions [34, 63]. Spin-correlated radical pairs [$\text{ZnTPPS}^{3-} \dots \text{DQ}^-$] are formed initially, and the SCRPs lifetime depends upon the solution environment. The ZnTPPS^{3-} is not observed due to its short T_2 relaxation time, but the spectra of DQ^- allow for the determination of the location and stability of reactant and product species in the various micellar solutions. While DQ is always located within the micelle, the location of ZnTPPS and free DQ^- depends upon the micellar environment.

Figure B1.16.20 shows spectra of DQ^- in a solution of TX100, a neutral surfactant, as a function of delay time. The spectra are qualitatively similar to those obtained in ethanol solution. At early delay times, the polarization is largely TM while RPM increases at later delay times. The early TM indicates that the reaction involves ZnTPPS triplets while the A/E RPM at later delay times is produced by triplet excited-state electron transfer. Calculation of relaxation times from spectral data indicates that in this case the ZnTPPS porphyrin molecules are in the micelle, although some may also be in the hydrophobic mantle of the micelle. Further, lineshape and polarization decay analyses indicate that free DQ^- radicals move from the micelle into the aqueous phase. The lack of observation of spin-correlated radical pairs indicates that they have dissociated prior to data acquisition. Small out-of-phase signal contributions at the earliest delay times show that the radical pair lifetime in TX100 solution is approximately 100 ns.

FT-EPR spectra of the ZnTPPS/DQ system in a solution of cetyltrimethylammonium chloride (CTAC), a cationic surfactant, are shown in figure B1.16.21. As in the TX100 solution, both donor and acceptor are associated with the micelles in the CTAC solution. The spectra of DQ^- at delays after the laser flash of less than 5 μs clearly show polarization from the SCRPs mechanism. While SCRPs were too short-lived to be observed in TX100 solution, they clearly have a long lifetime in this case. Van Willigen and co-workers

determined that the anionic radicals ZnTPPS^{3-} and DQ^- remain trapped in the cationic micelles, i.e. an electrostatic interaction is responsible for the extremely long lifetime of the $[\text{ZnTPPS}^{3-} \dots \text{DQ}^-]$ spin-correlated radical pair. The spectrum at delay times greater than $5 \mu\text{s}$ is again due to free DQ^- . Linewidth analysis and relaxation time calculations indicate that the free DQ^- remains trapped in the cationic micelles. These results demonstrate the use of the CIDEP mechanisms in helping to characterize the physical environments of free radicals. In both cases shown here, spectral analysis allowed a determination of the lifetime of the initial SCRPs and the location of the porphyrin and the free DQ^- .

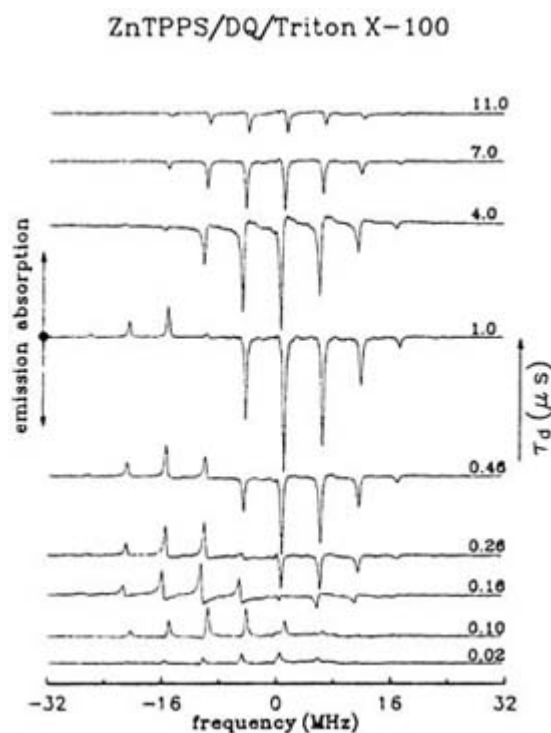


Figure B1.16.20. FTEPR spectra of photogenerated DQ^- in TX100 solution for delay times between laser excitation of ZnTPPS and microwave pulse ranging from 20 ns to 11 μs . The central hyperfine line ($M = 0$) is at ≈ -4.5 MHz. Reprinted from [63].

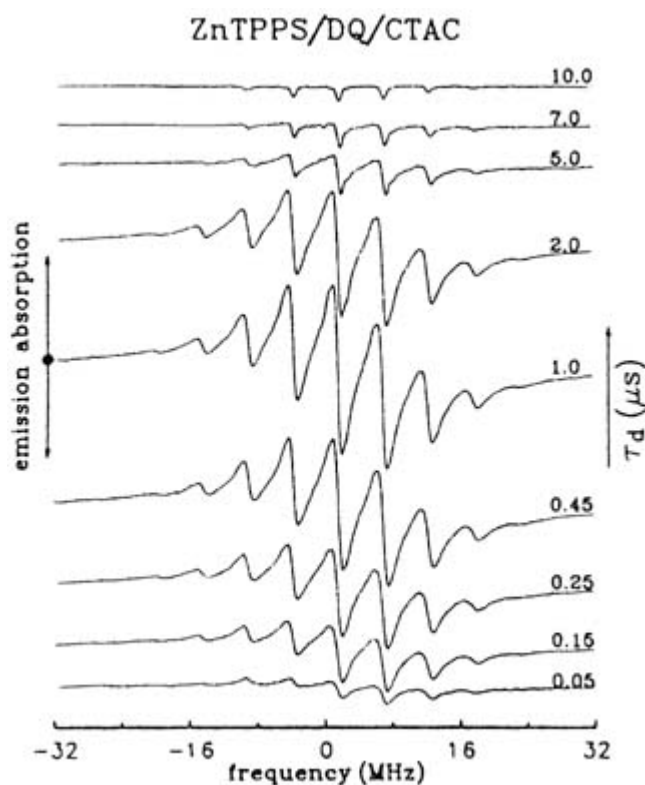


Figure B1.16.21. FTEPR spectra photogenerated DQ^- in CTAC solution for delay times between laser excitation of ZnTPPS and microwave pulse ranging from 50 ns to 10 μ s. The central hyperfine line ($M = 0$) is at ≈ 7 MHz. Reprinted from [63].

Figure B1.16.22 shows a stick plot summary of the various CIDEP mechanisms and the expected polarization patterns for the specific cases detailed in the caption. Each mechanism clearly manifests itself in the spectrum in a different and easily observable fashion, and so qualitative deductions regarding the spin multiplicity of the precursor, the sign of J in the RP and the presence or absence of SCRPs can immediately be made by examining the spectral shape. Several types of quantitative information are also available from the spectra. For example, if the molecular structure of one or both members of the RP is unknown, the hyperfine coupling constants and g -factors can be measured from the spectrum and used to characterize them, in a fashion similar to steady-state EPR. Sometimes there is a marked difference in spin relaxation times between two radicals, and this can be measured by collecting the time dependence of the CIDEP signal and fitting it to a kinetic model using modified Bloch equations [64].

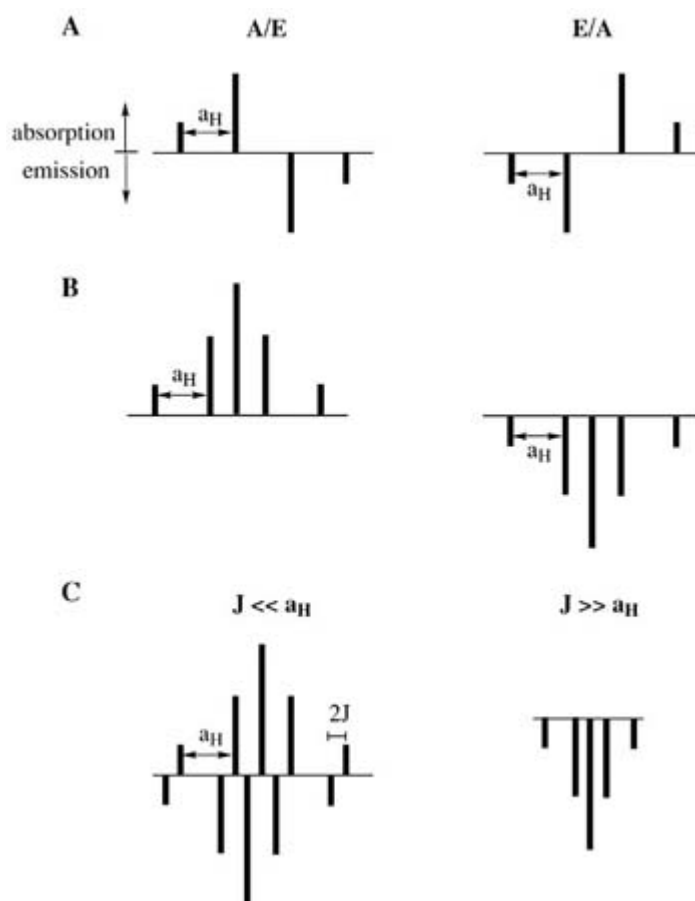


Figure B1.16.22. Schematic representations of CIDEP spectra for hypothetical radical pair $\cdot\text{CH}_3 + \cdot\text{R}$. Part A shows the A/E and E/A RPM. Part B shows the absorptive and emissive triplet mechanism. Part C shows the spin-correlated RPM for cases where $J \ll a_H$ and $J \gg a_H$.

If the rate of chemical decay of the RP is desired, the task is complex because the majority of the CIDEP signal decays via relaxation pathways on the 1–10 μs time scale, as opposed to chemical reaction rates which are nominally about an order of magnitude longer than this. There are two ways around this problem. The first is to use a transient digitizer or FT-EPR and signal average many times to improve the signal-to-noise ratio at long delay times where chemical reaction dominates the decay trace. The second is to return to the steady-state method described above and run what is called a ‘kinetic EPR’ experiment, where the light source is suddenly interrupted and the EPR signal decay is collected over a very long time scale. The beginning of the trace may contain both relaxation of CIDEP intensity as well as chemical decay; however, the tail end of this trace should be dominated by the chemical reaction rates. Much use has been made of kinetic EPR in measuring free radical addition rates in polymerization reactions [65, 66].

From SCRIP spectra one can always identify the sign of the exchange or dipolar interaction by direct examination of the phase of the polarization. Often it is possible to quantify the absolute magnitude of D or J by computer simulation. The shape of SCRIP spectra are very sensitive to dynamics, so temperature and viscosity dependencies are informative when knowledge of relaxation rates of competition between RPM and SCRIP mechanisms is desired. Much use of SCRIP theory has been made in the field of photosynthesis, where structure/function relationships in reaction centres have been connected to their spin physics in considerable detail [67, 68].

REFERENCES

- [1] Fessenden R W and Schuler R H 1963 Electron spin resonance studies of transient alkyl radicals *J. Chem. Phys.* **39** 2147–95
- [2] Bargon J, Fischer H and Johnsen U 1967 Kernresonanz-Emissionslinien während rascher Radikalreaktionen *Z. Naturf. a* **20** 1551–5
- [3] Ward H R and Lawler R G 1967 Nuclear magnetic resonance emission and enhanced absorption in rapid organometallic reactions *J. Am. Chem. Soc.* **89** 5518–19
- [4] Lawler R G 1967 Chemically induced dynamic nuclear polarization *J. Am. Chem. Soc.* **89** 5519–21
- [5] Closs G L 1969 A mechanism explaining nuclear spin polarizations in radical combination reactions *J. Am. Chem. Soc.* **91** 4552–4
- [6] Closs G L and Trifunac A D 1969 Chemically induced nuclear spin polarization as a tool for determination of spin multiplicities of radical-pair precursors *J. Am. Chem. Soc.* **91** 4554–5
- [7] Closs G L and Trifunac A D 1970 Theory of chemically induced nuclear spin polarization. III. Effect of isotropic g shifts in the components of radical pairs with one hyperfine interaction *J. Am. Chem. Soc.* **92** 2183–4
- [8] Kaptein R and Oosterhoff J L 1969 Chemically induced dynamic polarization II (relation with anomalous ESR spectra) *Chem. Phys. Lett.* **4** 195–7
- [9] Kaptein R and Oosterhoff J L 1969 Chemically induced dynamic nuclear polarization III (anomalous multiplets of radical coupling and disproportionation products) *Chem. Phys. Lett.* **4** 214–16
- [10] Roth H D, Weng H and Herbertz T 1997 CIDNP study and *ab initio* calculations of rigid vinylcyclopropane systems: evidence for delocalized 'ring-closed' radical cations *Tetrahedron* **53** 10 051–70
- [11] Dukes K E 1996 *PhD Dissertation* University of North Carolina
- [12] Salikhov K M, Molin Yu N, Sagdeev R Z and Buchachenko A L 1984 *Spin Polarization and Magnetic Effects in Radical Reactions* (Amsterdam: Elsevier)
- [13] Michl J and Bonacic-Koutecky V 1990 *Electronic Aspects of Organic Photochemistry* (New York: Wiley)
- [14] Porter N A, Marnett L J, Lochmüller C H, Closs G L and Shobtaki M 1972 Application of chemically induced dynamic nuclear polarization to a study of the decomposition of unsymmetric azo compounds *J. Am. Chem. Soc.* **94** 3664–5

- [15] Closs G L and Czeropski M S 1977 Amendment of the CIDNP phase rules. Radical pairs leading to triplet states *J. Am. Chem. Soc.* **99** 6127–8
- [16] Franck J and Rabinowitsch E 1934 Some remarks about free radicals and the photochemistry of solutions *Trans. Faraday Soc.* **30** 120–31
- [17] Turro N J, Buchachenko A L and Tarasov V F 1995 How spin stereochemistry severely complicates the formation of a carbon-carbon bond between two reactive radicals in a supercage *Acc. Chem. Res.* **28** 69–80
- [18] Goetz M 1995 An introduction to chemically induced dynamic nuclear polarization *Concepts Magn. Reson.* **7** 69–86
- [19] Pedersen J B 1979 *Theories of Chemically Induced Magnetic Polarization* (Odense: Odense University Press)

- [20] Kaptein R 1971 Simple rules for chemically induced dynamic nuclear polarization *J. Chem. Soc. Chem. Commun.* 732–3
- [21] Goetz M and Sartorius I 1993 Photo-CIDNP investigation of the deprotonation of aminium cations *J. Am. Chem. Soc.* **115** 11 123–33
- [22] Goetz M 1995 Pulse techniques for CIDNP *Concepts Magn. Reson.* **7** 263–79
- [23] Hany R, Vollenweider J-K and Fischer H 1988 Separation and analysis of CIDNP spin orders for a coupled multiproton system *Chem. Phys.* **120** 169–75
- [24] Closs G L and Miller R J 1979 Laser flash photolysis with NMR detection. Microsecond time-resolved CIDNP: separation of geminate and random-phase polarization *J. Am. Chem. Soc.* **101** 1639–41
- [25] Closs G L, Miller R J and Redwine O D 1985 Time-resolved CIDNP: applications to radical and biradical chemistry *Acc. Chem. Res.* **18** 196–202
- [26] Miller R J and Closs G L 1981 Application of Fourier transform-NMR spectroscopy to submicrosecond time-resolved detection in laser flash photolysis experiments *Rev. Sci. Instrum.* **52** 1876–85
- [27] Closs G L and Redwine O D 1985 Direct measurements of rate differences among nuclear spin sublevels in reactions of biradicals *J. Am. Chem. Soc.* **107** 6131–3
- [28] Morozova O B, Tsentelovich Y P, Yurkovskaya A V and Sagdeev R Z 1998 Consecutive biradicals during the photolysis of 2,12-dihydroxy-2,12-dimethylcyclododecanone: low- and high-field chemically induced dynamic nuclear polarizations (CIDNP) study *J. Phys. Chem. A* **102** 3492–7
- [29] Smaller B, Remko J R and Avery E C 1968 Electron paramagnetic resonance studies of transient free radicals produced by pulse radiolysis *J. Chem. Phys.* **48** 5174–81
- [30] Trifunac A D, Thurnauer M C and Norris J R 1978 Submicrosecond time-resolved EPR in laser photolysis *Chem. Phys. Lett.* **57** 471–3
- [31] Fessenden R W and Verma N C 1976 Time resolved electron spin resonance spectroscopy. III. Electron spin resonance emission from the hydrated electron. Possible evidence for reaction to the triplet state *J. Am. Chem. Soc.* **98** 243–4
- [32] Forbes M D E, Peterson J and Breivogel C S 1991 Simple modification of Varian E-line microwave bridges for fast time-resolved EPR spectroscopy *Rev. Sci. Instrum.* **66** 2662–5

- [33] Forbes M D E 1993 A fast 35 GHz time-resolved EPR apparatus *Rev. Sci. Instrum.* **64** 397–402
- [34] van Willigen H, Levstein P R and Ebersole M H 1993 Application of Fourier transform electron paramagnetic resonance in the study of photochemical reactions *Chem. Rev.* **93** 173–97
- [35] Adrian F J 1971 Theory of anomalous electron spin resonance spectra of free radicals in solution. Role of diffusion-controlled separation and reencounter of radical pairs *J. Chem. Phys.* **54** 3918–23
- [36] Hore P J 1989 Analysis of polarized electron paramagnetic resonance spectra *Advanced EPR: Applications in Biology and Biochemistry* ed A J Hoff (Amsterdam: Elsevier) ch 12
- [37] Sekiguchi S, Kobori Y, Akiyama K and Tero-Kubota S 1998 Marcus free energy dependence of the sign of exchange interactions in radical ion pairs generated by photoinduced electron transfer reactions *J. Am. Chem. Soc.* **120** 1325–6
- [38] Pedersen J B and Freed J H 1973 Theory of chemically induced dynamic electron polarization. I *J.*

Chem. Phys. **58** 2746–62

- [39] Pedersen J B and Freed J H 1973 Theory of chemically induced dynamic electron polarization. II *J. Chem. Phys.* **59** 2869–85
- [40] Wong S K, Hutchinson D A and Wan J K S 1973 Chemically induced dynamic electron polarization. II. A general theory for radicals produced by photochemical reactions of excited triplet carbonyl compounds *J. Chem. Phys.* **58** 985–9
- [41] Atkins P W and Evans G T 1974 Electron spin polarization in a rotating triplet *Mol. Phys.* **27** 1633–44
- [42] Jent F, Paul H and Fischer H 1988 Two-photon processes in ketone photochemistry observed by time-resolved ESR spectroscopy *Chem. Phys. Lett.* **146** 315–19
- [43] Blättler C, Jent F and Paul H 1990 A novel radical-triplet pair mechanism for chemically induced electron polarization (CIDEP) of free radicals in solution *Chem. Phys. Lett.* **166** 375–80
- [44] Blättler C and Paul H 1991 CIDEP after laser flash irradiation of benzil in 2-propanol. Electron spin polarization by the radical-triplet pair mechanism *Res. Chem. Intermed.* **16** 201–11
- [45] Goudsmit G-H, Paul H and Shushin A I 1993 Electron spin polarization in radical-triplet pairs. Size and dependence on diffusion *J. Phys. Chem.* **97** 13 243–9
- [46] Closs G L, Forbes M D E and Norris J R 1987 Spin-polarized electron paramagnetic resonance spectra of radical pairs in micelles. Observation of electron spin–spin interactions *J. Phys. Chem.* **91** 3592–9
- [47] Buckley C D, Hunger D A, Hore P J and McLauchlan K A 1987 Electron spin resonance of spin-correlated radical pairs *Chem. Phys. Lett.* **135** 307–12
- [48] Closs G L and Forbes M D E 1991 EPR spectroscopy of electron spin polarized biradicals in liquid solutions. Technique, spectral simulation, scope and limitations *J. Phys. Chem.* **95** 1924–33
- [49] Norris J R, Morris A L, Thurnauer M C and Tang J 1990 A general model of electron spin polarization arising from the interactions within radical pairs *J. Chem. Phys.* **92** 4239–49
- [50] Levanon H and Möbius K 1997 Advanced EPR spectroscopy on electron transfer processes in photosynthesis and biomimetic model systems *Ann. Rev. Biophys. Biomol. Struct.* **26** 495–540

- [51] Friebolin H 1993 *Basic One- and Two-Dimensional NMR Spectroscopy* (New York: VCH)
- [52] De Kanter F J J, den Hollander J A, Huizer A H and Kaptein R 1977 Biradical CIDNP and the dynamics of polymethylene chains *Mol. Phys.* **34** 857–74
- [53] Avdievich N I and Forbes M D E 1995 Dynamic effects in spin-correlated radical pair theory: J modulation and a new look at the phenomenon of alternating line widths in the EPR spectra of flexible biradicals *J. Phys. Chem.* **99** 9660–7
- [54] Forbes M D E, Schulz G R and Avdievich N I 1996 Unusual dynamics of micellized radical pairs generated from photochemically active amphiphiles *J. Am. Chem. Soc.* **118** 10 652–3
- [55] Forbes M D E, Avdievich N I, Schulz G R and Ball J D 1996 Chain dynamics cause the disappearance of spin-correlated radical pair polarization in flexible biradicals *J. Phys. Chem.* **100** 13 887–91
- [56] Maeda K, Terazima M, Azumi T and Tanimoto Y 1991 CIDNP and CIDEP studies on intramolecular hydrogen abstraction reaction of polymethylene-linked xanthone and xanthene. Determination of the

exchange integral of the intermediate biradicals *J. Phys. Chem.* **95** 197–204

- [57] Forbes M D E, Closs G L, Calle P and Gautam P 1993 The temperature dependence of the exchange coupling in polymethylene biradicals. Conclusions regarding the mechanism of the coupling *J. Phys. Chem.* **97** 3384–9
- [58] Forbes M D E 1993 The effect of localized unsaturation on the scalar exchange coupling in flexible biradicals *J. Phys. Chem.* **97** 3390–5
- [59] Forbes M D E 1993 The effect of π -system spacers on exchange couplings and end-to-end encounter rates in flexible biradicals *J. Phys. Chem.* **97** 3396–400
- [60] Bittl R, van der Est A, Kamlowski A, Lubitz W and Stehlik D 1994 Time-resolved EPR of the radical pair $P_{665}^+ Q_A^-$ in bacterial reaction centers. Observation of transient nutations, quantum beats and envelope modulation effects *Chem. Phys. Lett.* **226** 249–58
- [61] Fukuju T, Yashiro H, Maeda K, Murai H and Azumi T 1997 Singlet-born SCRIP observed in the photolysis of tetraphenylhydrazine in an SDS micelle: time dependence of the population of the spin states *J. Phys. Chem. A* **101** 7783–6
- [62] Koga T, Ohara K, Kuwata K and Murai H 1997 Anomalous triplet mechanism spin polarization induced by the addition of hydrochloric acid in the photochemical system of xanthone in alcohol *J. Phys. Chem. A* **101** 8021–5
- [63] Levstein P R and van Willigen H 1991 Photoinduced electron transfer from porphyrins to quinones in micellar systems: an FT-EPR study *Chem. Phys. Lett.* **187** 415–22
- [64] Verma N C and Fessenden R W 1976 Time resolved ESR spectroscopy. IV. Detailed measurement and analysis of the ESR time profile *J. Chem. Phys.* **65** 2139–60
- [65] Héberger K and Fischer H 1993 Rate constants for the addition of the 2-cyano-2-propyl radical to alkenes in solution *Int. J. Chem. Kin.* **25** 249–63
- [66] Héberger K and Fischer H 1993 Rate constants for the addition of 2-hydroxy-2-propyl radical to alkenes in solution *Int. J. Chem. Kin.* **25** 913–20
- [67] Thurnauer M C and Norris J R 1980 An electron spin echo phase shift observed in photosynthetic algae. Possible evidence for dynamic radical pair interactions *Chem. Phys. Lett.* **76** 557–61

-36-

- [68] Prisner T F, van der Est A, Bittl R, Lubitz W, Stehlik D and Möbius K 1995 Time-resolved W-band (95 GHz) EPR spectroscopy of Zn-substituted reaction centers of *Rhodobacter sphaeroides* R-26 *Chem. Phys.* **194** 361–70

FURTHER READING

Salikhov K M, Molin Yu N, Sagdeev R Z and Buchachenko A L 1984 *Spin Polarization and Magnetic Effects in Radical Reactions* (Amsterdam: Elsevier)

A detailed description of spin polarization theory.

Lepley A R and Closs G L (eds) 1973 *Chemically Induced Magnetic Polarization* (New York: Wiley)

An early summary of research in the field of CIDNP.

Muus L T (ed) 1977 *Chemically Induced Magnetic Polarization: Proc. NATO Advanced Study Institute (Sogesta, Urbino, Italy, April 17–30, 1977)* (Boston, MA: Reidel)

The proceedings of the first international meeting on spin polarization phenomena.

Carrington A and McLachlan A D 1979 *Introduction to Magnetic Resonance, with Applications to Chemistry and Chemical Physics* (New York: Wiley)

An excellent beginner's introduction to magnetic resonance, spin operators and their manipulation to predict and analyze spectra.

Weil J A, Bolton J R and Wertz J E 1994 *Electron Paramagnetic Resonance: Elementary Theory and Practical Applications* (New York: Wiley)

The standard monograph for those seeking an introduction to EPR spectroscopy.

Friebolin H 1993 *Basic One- and Two-Dimensional NMR Spectroscopy* (New York: VCH)

A basic introduction to NMR spectral analysis.

-1-

B1.17 Microscopy: electron (SEM and TEM)

Rasmus R Schröder and Martin Müller

ABBREVIATIONS

2D	two-dimensional
3D	three-dimensional
ssCCD	slow-scan charge coupled device
BSE	backscattered electrons
CTF	contrast transfer function
DQE	detection quantum efficiency
E	electron energy
E_0	electron rest energy
EDX	energy dispersive x-ray detection spectroscopy
EELS	electron energy loss spectroscopy
EFTEM	energy filtering transmission electron microscope
EM	electron microscope/microscopy
EPMA	electron probe micro analysis

ESEM	environmental scanning electron microscope
ESI	electron spectroscopic imaging
ESR	electron spin resonance
FEG	field emission gun
IP	imaging plate
LVSEM	low-voltage scanning electron microscope
MTF	modulation transfer function
NMR	nuclear magnetic resonance

-2-

PSF	point spread function
SE	secondary electron
SEM	scanning electron microscope
STEM	scanning transmission electron microscope
TEM	transmission electron microscope

B1.17.1 INTRODUCTION

The electron microscope (EM) was developed in the 1930s primarily as an imaging device which exceeded the resolution power of the light microscope by several orders of magnitude. With the evolution towards dedicated instruments designed to answer specific structural and analytical questions, electron microscopy (EM) has grown into a heterogeneous field of electron beam devices. These allow the study of the interaction of electrons with the sample, which can subsequently be interpreted as information about object structure or chemical composition. Therefore, EM must be compared to other high-resolution diffraction methods, such as x-ray or neutron scattering, or to spectroscopic techniques such as electron spin resonance spectroscopy (ESR) and nuclear magnetic resonance spectroscopy (NMR). More recent, non-diffractive techniques include scanning tunnelling microscopy (STM) and atomic force microscopy (AFM) (for a detailed discussion see [chapter B1.19](#)).

All these methods are used today to obtain structural and analytical information about the object (see also the specific chapters about these techniques). In the case of structural studies, x-ray crystallography is the method of choice if suitable three-dimensional (3D) crystals of the object are available. In fact, x-ray crystallography has provided a vast number of atomic structures of inorganic compounds or organic molecules. The main advantage of EM, however, is the possibility of directly imaging almost any sample, from large biological complexes consisting of many macromolecules down to columns of single atoms in a solid material. With modern instruments and applying specific preparation techniques, it is even possible to visualize beam-sensitive organic material at molecular resolution (of the order of 3–4 Å). Imaging at atomic resolution is almost routine in material science. Today, some challenges remain, including the combination of sub-Ångström resolution with chemical analysis and the ability to routinely reconstruct the complete 3D spatial structure of a given sample.

The history of EM (for an overview see [table B1.17.1](#)) can be interpreted as the development of two concepts: the electron beam either illuminates a large area of the sample ('flood-beam illumination', as in the typical transmission electron microscope (TEM) imaging using a spread-out beam) or just one point, i.e. focused to the smallest spot possible, which is then scanned across the sample (scanning transmission electron microscopy (STEM) or scanning electron microscopy (SEM)). In both situations the electron beam is considered as a matter wave interacting with the sample and microscopy simply studies the interaction of the scattered electrons.

Table B1.17.1. Instrumental development of electron microscopy.

Year	Event	Reference
1926	Busch focuses an electron beam with a magnetic lens	
1931	Ruska and colleagues build the first TEM prototype	Knoll and Ruska (1932) [79]
1935	Knoll proves the concept of SEM	
1938	von Ardenne builds the first SEM prototype	
1939	Siemens builds the first commercial TEM	
1965	Cambridge Instruments builds the first commercial SEM	
1968	Crewe and colleagues introduce the FEG as electron beam source	Crewe <i>et al</i> (1968) [80]
1968	Crewe and colleagues build the first & STEM prototype	Crewe <i>et al</i> (1968) [81]
1995	Zach proves the concept of a corrected LVSEM	Zach (1995) [10]
1998	Haider and colleagues prove the concept of the TEM spherical aberration corrector	Haider <i>et al</i> (1998) [55]

In principle, the same physico-chemical information about the sample can be obtained from both illumination principles. However, the difference of the achievable spatial resolutions of both illumination principles illustrates the general difference of the two approaches. Spatial resolution in the case of flood-beam illumination depends on the point spread function (PSF) of the imaging lens and detector system and—for a typical TEM sample—on the interference effects representing phase shifts of the scattered electron wave. In general, resolution is described as a global phenomenon. For the scanned beam, all effects are local and confined to the illuminated spot. Thus, spatial resolution of any event is identical to the achievable spot diameter. Therefore, all steps in the further development of EM involve either the improvement of the PSF (either by improving sample preparation, imaging quality of the electron lenses, or improving spatial resolution of the electron detection devices) or the improvement of the spot size (by minimizing the size of the electron source using a field emission gun (FEG) and by improving the imaging quality of the electron lenses).

In general, EM using a focused beam provides higher spatial resolution if the spot size of the scanning beam is smaller than the delocalization of the event studied: consider, for example, inelastic scattering events which for signal-to-noise ratio (SNR) reasons can be imaged in energy filtering TEMs (EFTEM) at 1–2 nm resolution. In a dedicated STEM with an FEG electron source, a localization of the inelastic event comparable

to the actual probe size of e.g. 2 Å can be expected. Moreover, this effect is enhanced by the electron detectors used in TEM and S(T)EM. For scanned-beam microscopy, detectors do not resolve spatial information, instead they are designed for highest detection quantum efficiency (DQE almost ideal). Such an ideal detection has only recently been reached for TEM by the use of slow-scan charge-coupled device (ssCCD) cameras or imaging plates (IP). However, their spatial resolution is only moderate, compared to conventional, electron-sensitive photographic material.

-4-

For several reasons, such as ease of use, cost, and practicability, TEM today is the standard instrument for electron diffraction or the imaging of thin, electron-transparent objects. Especially for structural imaging at atomic level (spatial resolution of about 1 Å) the modern, aberration-corrected TEM seems to be the best instrument. SEM provides the alternative for imaging the surface of thick bulk specimens. Analytical microscopy can either be performed using a scanning electron probe in STEM and SEM (as for electron probe micro-analysis (EPMA), energy-dispersive x-ray spectroscopy (EDX) and electron energy loss spectroscopy (EELS)) or energy-selective flood-beam imaging in EFTEM (as for image-EELS and electron spectroscopic imaging (ESI)). The analytical EM is mainly limited by the achievable probe size and the detection limits of the analytical signal (number of inelastically scattered electrons or produced characteristic x-ray quanta). The rest of this chapter will concentrate on the structural aspects of EM. Analytical aspects are discussed in more detail in specialized chapters (see, for example, [B1.6](#)).

It is interesting to note the analogy of developments in light microscopy during the last few decades. The confocal microscope as a scanning beam microscope exceeds by far the normal fluorescence light microscope in resolution and detection level. Very recent advances in evanescent wave and interference microscopy seem to promise to provide even higher resolution ([B1.18](#)).

EM has been used in a wide variety of fields, from material sciences to cell and structural biology or medical research. In general, EM can be used for any high-resolution imaging of objects or their analytical probing. Modern instrumentation of STEM and SEM provides high-resolution instruments capable of probe sizes, in the case of TEM, of a few Ångström or sub-Ångström information limit. However, specimen properties and sample preparation limit the achievable resolution. Typical resolution obtained today range from atomic detail for solid materials, molecular detail with resolution in the order of 3–5 Å for crystalline biological samples, and about 1–2 nm for individual particles without a certain intrinsic symmetry. Recent publications on the different aspects of EM include Williams and Carter ([\[1\]](#), general text covering all the modern aspects of EM in materials science), the textbooks by Reimer ([\[2,3 and 4\]](#), detailed text about theory, instrumentation, and application), or—for the most complete discussion of all electron-optical and theoretical aspects—Hawkes and Kasper [\[5\]](#). Additional research papers on specialized topics are referenced in text.

B1.17.2 INTERACTION OF ELECTRONS WITH MATTER AND IMAGING OF THE SCATTERING DISTRIBUTION

The interaction of electrons with the specimen is dominated by the Coulomb interaction of charged particles. For a summary of possible charge–charge interactions see [figure B1.17.1](#). Elastic scattering by the Coulomb potential of the positively charged atomic nucleus is most important for image contrast and electron diffraction. This scattering potential is well localized, leads to large scattering angles and yields high-resolution structural information from the sample. In contrast, interactions with the atomic electrons lead to an energy loss of the incident electron by the excitation of different energy states of the sample, such as phonon excitation, plasmon excitation or inner-shell ionization (see [table B1.17.2](#)). Inelastic scattering processes are not as localized as the Coulomb potential of the nucleus, leading to smaller scattering angles (inelastic forward scattering), and are in general not used to obtain high-resolution structural information. Instead, inelastic scattering provides analytical information about the chemical composition and state of the sample.

Table B1.17.2. Electron–specimen interactions.

	Elastic scattering	Inelastic scattering
Where	Coulomb potential of nucleus	At atomic shell electrons
Scattering potential	Localized	Less localized
Scattering angles $E = 100 \text{ keV}$	Large ($> 10 \text{ mrad}$)	Smaller ($< 10 \text{ mrad}$)
Application	High-resolution signal (TEM, STEM) Back-scattering of electrons (BSE signal in SEM)	Analytical signal (TEM, STEM, SEM) Emission of secondary electrons (SE signal in SEM)
Used effects		Phonon excitation (20 meV–1 eV) Plasmon and interband excitations (1–50 eV) Inner-shell ionization ($\Delta E =$ ionization energy loss) Emission of x-ray (continuous/characteristic, analytical EM)

The ratio of elastically to inelastically scattered electrons and, thus, their importance for imaging or analytical work, can be calculated from basic physical principles: consider the differential elastic scattering cross section

$$\frac{d\sigma_{el}}{d\Omega} = \frac{4Z^2 R^4 (1 + E/E_0)}{a_H^2} \frac{1}{[1 + (\theta/\theta_0)^2]} \quad (\text{B1.17.1})$$

with the characteristic screening angle $\theta_0 = \lambda/(2\pi R)$ where θ denotes the scattering angle, E is the electron energy, E_0 the electron rest energy, $+eZ$ is the charge of the nucleus, $R = a_H Z^{1/3}$, and $a_H = 0.0529 \text{ nm}$ is the Bohr radius. Compare this to the inelastic differential scattering cross section

$$\frac{d\sigma_{incl}}{d\Omega} = \frac{\lambda^4 (1 + E/E_0)^2}{4\pi^4 a_H^2} \frac{Z \left\{ 1 - \frac{1}{[1 + (\theta/\theta_0)^2]} \right\}}{(\theta^2 + \theta_E^2)^2} \quad (\text{B1.17.2})$$

with the characteristic inelastic scattering angle $\theta_E = \Delta E/E \cdot (E+E_0)/(E+2E_0)$ for a given energy loss ΔE .

For large scattering angles, $\theta \gg \theta_0, \theta_E$ the ratio

$$\frac{d\sigma_{\text{inel}}/d\Omega}{d\sigma_{\text{el}}/d\Omega} = \frac{1}{Z} \quad (\text{B1.17.3})$$

-6-

only depends on the atomic number Z , whereas for small angles θ

$$d\sigma_{\text{inel}}/d\Omega > d\sigma_{\text{el}}/d\Omega \quad (\text{B1.17.4})$$

for all Z . The total scattering cross sections are found by integrating the above equations [6]

$$\nu = \frac{\sigma_{\text{inel}}}{\sigma_{\text{el}}} \approx \frac{26}{Z} \quad (\text{B1.17.5})$$

or experimentally [7]

$$\nu \approx \frac{18}{Z}. \quad (\text{B1.17.6})$$

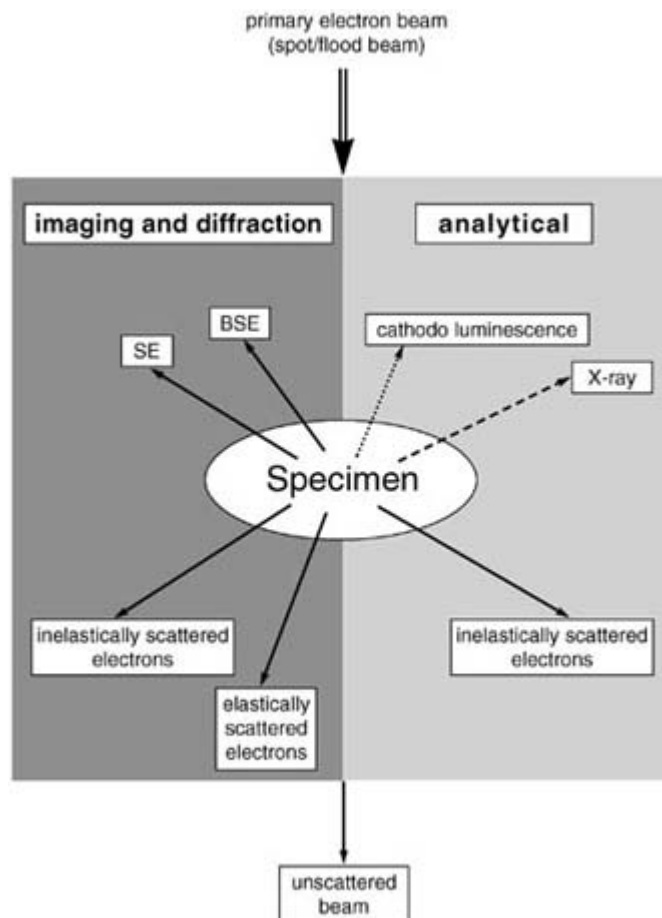


Figure B1.17.1. Schema of the electron specimen interactions and their potential use for structural and analytical studies.

Equation (B1.17.1), Equation (B1.17.2), Equation (B1.17.3), Equation (B1.17.4), Equation (B1.17.5) and Equation (B1.17.6) indicate that inelastic scattering is most important for light atoms, whereas elastic scattering dominates for large scattering angles and heavy atoms. Therefore, the high resolution image contrast—given by electrons scattered to large angles—for TEM and STEM is dominated by the elastic scattering process. Inelastically scattered electrons are treated either as background, or separated from the elastic image by energy-dispersive spectrometers.

In the case of SEM, both elastic and inelastic processes contribute to image contrast: elastic scattering to large angles (multiple elastic scattering resulting in a large scattering angle) produces backscattered electrons (BSE), which can be detected above the surface of the bulk specimen. Inelastic collisions can excite atomic electrons above the Fermi level, i.e. more energy is transferred to the electron than it would need to leave the sample. Such secondary electrons (SE) are also detected, and used to form SEM images. As will be discussed in B1.17.4 (specimen preparation), biological material with its light atom composition is often stained or coated with heavy metal atoms to increase either the elastic scattering contrast in TEM or the BSE signal in SEM. Unstained, native biological samples generally produce only little image contrast.

Inelastic scattering processes are not used for structural studies in TEM and STEM. Instead, the signal from inelastic scattering is used to probe the electron-chemical environment by interpreting the specific excitation of core electrons or valence electrons. Therefore, inelastic excitation spectra are exploited for analytical EM.

Next we will concentrate on structural imaging using only elastically scattered electrons. To obtain the structure of a scattering object in TEM it is sufficient to detect or image the scattering distribution. Consider the scattering of an incident plane wave $\psi_0(\mathbf{r}) \equiv 1$ on an atomic potential $V(\mathbf{r})$. The scattered, outgoing wave will be described by the time-independent Schrödinger equation. Using Green's functions the wave function can be written as

$$\psi(\mathbf{r}) = \exp(i\mathbf{k}\mathbf{r}) - \frac{m}{2\pi\hbar^2} \int \frac{\exp(i\mathbf{k}|\mathbf{r} - \mathbf{r}'|)}{|\mathbf{r} - \mathbf{r}'|} V(\mathbf{r}')\psi(\mathbf{r}') d\mathbf{r}' \quad (\text{B1.17.7})$$

where \mathbf{k} denotes the initial wave vector, m is equal to the electron mass. For large distances from the scattering centre $r \gg r'$. This can be approximated by

$$\psi(\mathbf{r}) = \exp(i\mathbf{k}\mathbf{r}) + f(\theta) \frac{\exp(ikr)}{r} \quad (\text{B1.17.8})$$

where θ denotes the scattering angle.

In this approximation, the wave function is identical to the incident wave (first term) plus an outgoing spherical wave multiplied by a complex scattering factor

$$f(\theta) = |f(\theta)| \exp(i\eta(\theta)) \quad (\text{B1.17.9})$$

which can be calculated as

$$f(\theta) = -\frac{m}{2\pi\hbar^2} \int \exp(ik'\mathbf{r}') V(\mathbf{r}') \psi(\mathbf{r}') d\mathbf{r}' \quad (\text{B1.17.10})$$

where $\mathbf{k}' = k\mathbf{r}'/r$. For a weak potential $V(\mathbf{r})$ it is possible to use the first Born approximation, i.e. $\psi(\mathbf{r}')$ in equation (B1.17.10) can be replaced by the incident wave resulting in:

$$f(\theta) = -\frac{m}{2\pi\hbar^2} \int \exp(i(\mathbf{k} - \mathbf{k}')\mathbf{r}') V(\mathbf{r}') d\mathbf{r}'. \quad (\text{B1.17.11})$$

This equation describes the Fourier transform of the scattering potential $V(\mathbf{r})$. It should be noted that, in the Born approximation the scattering amplitude $f(\theta)$ is a real quantity and the additional phase shift $\eta(\theta)$ is zero. For atoms with high atomic number this is no longer true. For a rigorous discussion on the effects of the different approximations see [2] or [5].

In a diffraction experiment a quantity $|F(\mathbf{S})|^2$ can be measured which follows from equation (B1.17.8) and equation (B1.17.9) in Fourier space as

$$F(\mathbf{S}) = \delta_0 + i|f(\mathbf{S})| \exp(i\eta(\mathbf{S})) \quad (\text{B1.17.12})$$

where $\mathbf{S} = \mathbf{k} - \mathbf{k}'$ denotes the scattering vector. Combining equation (B1.17.11) and equation (B1.17.12) leads to the conclusion that in a diffraction experiment the squared amplitude of the Fourier transform of the scattering potential $V(\mathbf{r})$ is measured. Similar formulae can be deduced for whole assemblies of atoms, e.g. macromolecules, resulting in the molecular transform instead of simple atomic scattering factors (for an introduction to the concept of molecular transforms see e.g. [8]). Such measurements are performed in x-ray crystallography, for example. To reconstruct the original scattering potential $V(\mathbf{r})$ it is necessary to determine the phases of the structure amplitudes to perform the reverse Fourier transform. However, if lenses are available for the particles used as incident beam—as in light and electron microscopy—a simple microscope can be built: the diffracted wave is focused by an objective lens into the back focal plane, where the scattered and unscattered parts of the wave are separated. Thus, the objective lens can simply be understood as a Fourier transform operator. In a subsequent imaging step by one additional lens, the scattered and unscattered waves are allowed to interfere again to form a direct image of the scattering potential. This can be understood as a second Fourier transform of the scattering factor (or molecular transform) recovering the spatial distribution of the scattering centres. The small angular scattering distribution of only 10–20 mrad results in a complication in the case of EM. The depth of focus is very large, i.e. it is not possible to recalculate the 3D distribution of the scattering potential but only its 2D projection along the incident beam. All scattering distributions, images, or diffraction patterns are always produced by the 2D projecting transmission function of the actual 3D object. Using a variety of tomographic data collections it is then possible to reconstruct the true 3D object (see below).

The above theory can also be applied to STEM, which records scattering distributions as a function of the scanning probe position. Images are then obtained by plotting the measured scattering intensities (i.e. in the case of the elastic scattering, the direct measurement of the scattering factor amplitudes) according to the probe position. Depending on the signal used, this leads to a conventional elastic dark-field image, or to STEM phase-contrast images [9].

In the case of the scanning electron microscope (SEM), images are formed by recording a specific signal resulting from the electron beam/specimen interaction as a function of the scanning probe position. Surface structures are generally described with the SE (secondary electron) signal. SEs are produced as a consequence

of inelastic events. They have very low energies and, therefore, can leave the specimen and contribute to the imaging signal only when created very close to the specimen surface. The escape depth for the secondary electrons depends on the material. It is relatively large (tens of nanometres) for organic and biological material and small for heavy metals (1–3 nm). High-resolution topographic information (limited mainly by the diameter of the scanning electron beam) requires that the source of the signal is localized very close to the specimen surface. In the case of organic materials this localization can be achieved by a very thin metal coating (W, Cr, Pt; thickness = approximate SE escape depth).

The BSE signal is also frequently used for imaging purposes. BSE are electrons of the primary beam (scanning probe) that have been elastically or inelastically scattered in the sample. Their energy depends on their scattering history. When scattered from the surface, they may have lost no or very little energy and provide high topographic resolution. When multiply scattered inside the sample they may have lost several keV and transfer information from a large volume. This volume depends on the material as well as on the energy of the primary beam. BSE produce SE when passing through the SE escape zone. These SE are, however, not correlated with the position of the scanning probe and contribute a background noise which can obscure the high resolution topographic SE signal produced at the point of impact of the primary beam. The interpretation of high-resolution topographic images therefore depends on optimized handling of specimen properties, energy of the electron probe, metal coating and sufficient knowledge of the signal properties.

The discussion of electron–specimen interactions shows that, for a given incident electron dose, a certain quantity of resulting scattered electrons and secondary electrons or photons is produced. The majority of energy transfer into the specimen leads to beam damage and, finally, to the destruction of the sample structure. Therefore it is desirable to simultaneously collect as much information from the interactions as possible. This concept could lead to an EM instrument based on the design of a STEM but including many different detectors for the elastic dark field, phase contrast, inelastically scattered electrons, BSE, SE, and EDX. The complexity of such an instrument would be enormous. Instead, specific instruments developed in the past can coarsely be categorized as TEM for structural studies on thin samples, STEM for analytical work on thin samples and SEM for analytical and surface topography studies.

B1.17.3 INSTRUMENTATION

B1.17.3.1 ELECTRON BEAM INSTRUMENTS

The general instrumentation of an EM very much resembles the way an ordinary, modern light microscope is built. It includes an electron beam forming source, an illumination forming condenser system and the objective lens as the main lens of the microscope. With such an instrumentation, one forms either the conventional bright field microscope with a large illuminated sample area or an illumination spot which can be scanned across the sample. Typical electron sources are conventional heated tungsten hairpin filaments, heated LaB₆, or CeB₆-single-crystal electron emitters, or—as the most sophisticated source—FEGs. The latter sources lead to very coherent electron beams, which are necessary to obtain high-resolution imaging or very small electron probes.

Modern EMs use electromagnetic lenses, shift devices and spectrometers. However, electrostatic devices have always been used as electron beam accelerators and are increasingly being used for other tasks, e.g. as the objective lens (LVSEM, [10]).

EM instruments can be distinguished by the way the information, i.e. the interacting electrons, is detected. [Figure B1.17.2](#) shows the typical situations for TEM, STEM, and SEM. For TEM the transmitted electron beam of the brightfield illumination is imaged simply as in an light microscope, using the objective and

projective lenses as conventional imaging system. Combining such TEMs with energy-dispersive imaging elements (filters, spectrometers; see [14]) the modern generation of EFTEMs has been introduced in the last decade. In the case of STEM, the transmitted electrons are not again imaged by lenses, instead the scattered electrons are directly recorded by a variety of detectors. For SEM, the situation is similar to that of STEM. However, only the surface of a bulk specimen is scanned and the resulting backscattered or secondary electrons are recorded by dedicated detectors.

As a special development in recent years, SEMs have been designed which no longer necessitate high vacuum (environmental SEM, ESEM; variable pressure SEM, VPSEM). This development is important for the imaging of samples with a residual vapour pressure, such as aqueous biological or medical samples, but also samples in materials science (wet rock) or organic chemistry (polymers).

-11-

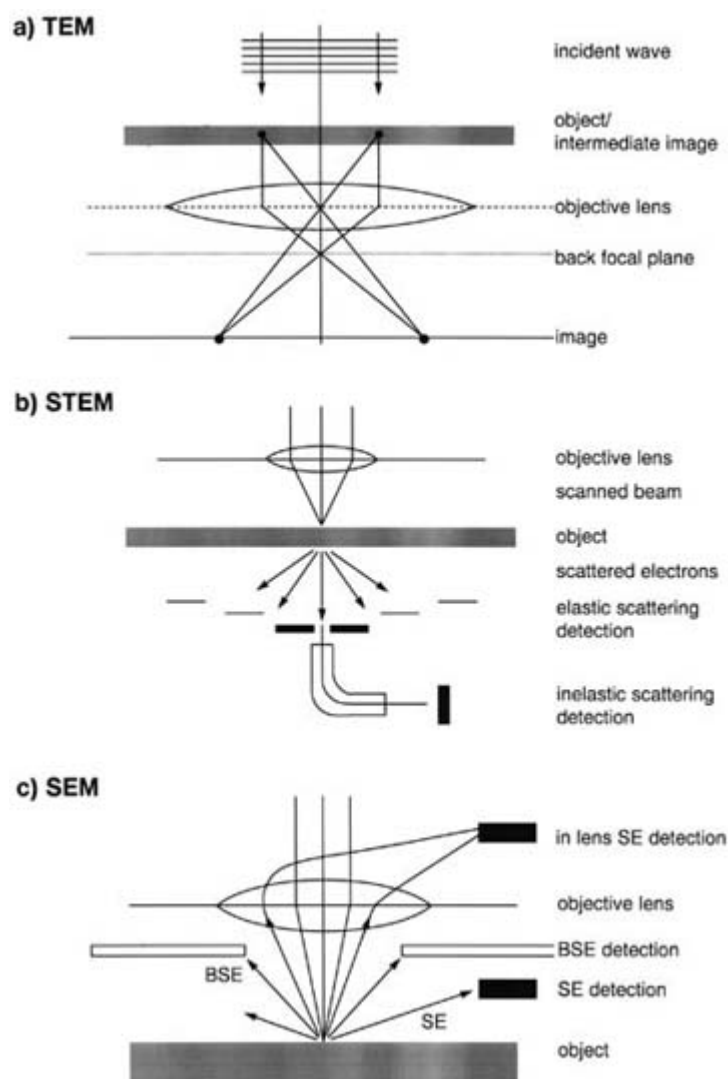


Figure B1.17.2. Typical electron beam path diagrams for TEM (a), STEM (b) and SEM (c). These schematic diagrams illustrate the way the different signals can be detected in the different instruments.

B1.17.3.2 ELECTRON DETECTORS IN TEM, STEM, AND SEM

Detectors in EM can be categorized according to their different spatial resolution or in relation to the time it takes to actually see and process the signal (real-time/on-line capability).

Historically, TEM—as an offspring of the electron oscillograph—uses a fluorescent viewing screen for the direct observation of impinging electrons by green fluorescent light. The spatial resolution of such screens is of the order of 30–50 μm . Coupled to a TV camera tube and computer frame-grabber cards, fluorescent screens are still used for the real-time recording of the image. Often the cameras are combined with silicon intensifier targets (SIT), which allows

-12-

the detection of single electrons. This is of special importance for the study of beam-sensitive samples. In general, the spatial resolution of such combinations is very poor and dynamic range of the signal is limited to about 256 grey levels (8 bit).

The most common electron detector for TEM has been photographic emulsion. Its silver halide particles are sensitive to high-energy electrons and record images at very high resolution (grain size smaller than 10 μm) when exposed to the electron beam in the TEM. The resolution recorded on film depends on the scattering volume of the striking electron. For typical electron energies of 100–200 keV, its lateral spread is 10–20 μm . The dynamic range of photographic film is up to 4000 grey levels (12 bit). The most important advantage of film is its large detector size combined with high resolution: with a film size of $6 \times 9 \text{ cm}^2$, up to 10^7 image points can be recorded.

A recent development is the adaptation of IP to EM detection. IPs have been used for detecting x-rays, which generate multiple electron–electron hole pairs in a storage layer of BaFBr:Eu^{2+} . The pairs are trapped in Eu–F centres and can be stimulated by red light to recombine, thereby emitting a blue luminescent signal. Exposing IPs to high energetic electrons also produces electron–electron hole pairs. Scanning the IP with a red laser beam and detecting the blue signal via a photo-multiplier tube results in the readout of the latent image. The large detector size and their extremely high dynamic range of more than 10^7 makes IPs the ideal detector for electron diffraction.

The only on-line detector for TEM with moderate-to-high spatial resolution is the slow-scan CCD camera. A light-sensitive CCD chip is coupled to a scintillator screen consisting of plastic, an yttrium–aluminium garnet (YAG) crystal, or phosphor powder. This scintillator layer deteriorates the original resolution of the CCD chip elements by scattering light into neighbouring pixels. Typical sizes of chips at present are 1024×1024 or 2048×2048 pixels of $(19\text{--}24 \mu\text{m})^2$; the achievable dynamic range is about 10^5 grey levels.

For all the detectors with spatially distinct signal recording, the numeric pixel size (such as scanning pixel size for photographic film and IP, or chip-element size for ssCCD) must be distinguished from the actual obtainable resolution. This resolution can be affected by the primary scattering process of electrons in the detecting medium, or by the scattering of a produced light signal or a scanning light spot in the detecting medium. Therefore, a point signal is delocalized, mathematically described by the PSF. The Fourier transform of the PSF is called the modulation transfer function (MTF), describing the spatial frequency response of the detector. Whereas the ideal detector has a $\text{MTF} = 1$ over the complete spatial frequency range, real detectors exhibit a moderate to strong fall-off of the MTF at the Nyquist frequency, i.e. their maximal detectable spatial resolution. In addition to spatial resolution, another important quantity characterizing a detector is the detection quantum efficiency (DQE). It is a measure of the detector noise and gives an assessment for the detection of single electrons.

For all TEM detectors the ssCCD has the best DQE. Depending on the scintillator, it is in the range of $\text{DQE} = 0.6 - 0.9$, comparable to IPs which show a DQE in the order of 0.5–0.8. The DQE of photographic emulsion is strongly dependent on electron dose and does not exceed $\text{DQE} = 0.2$. For a complete and up-to-date discussion on TEM electron detectors see the special issue of *Microscopy Research and Technique* (vol 49, 2000).

In SEM and STEM, all detectors record the electron current signal of the selected interacting electrons (elastic scattering, secondary electrons) in real time. Such detectors can be designed as simple metal-plate detectors, such as the elastic dark-field detector in STEM, or as electron-sensitive PMT. For a rigorous discussion of SEM detectors see [3].

-13-

Except for the phase-contrast detector in STEM [9], STEM and SEM detectors do not track the position of the recorded electron. The spatial information of an image is formed instead by assigning the measured electron current to the known position of the scanned incident electron beam. This information is then mapped into a 2D pixel array, which is depicted either on a TV screen or digitalized in a computer.

For the parallel recording of EEL spectra in STEM, linear arrays of semiconductor detectors are used. Such detectors convert the incident electrons into photons, using additional fluorescent coatings or scintillators in the very same way as the TEM detectors described above.

B1.17.4 SPECIMEN PREPARATION

The necessity to have high vacuum in an electron beam instrument implies certain constraints on the specimen. In addition, the beam damage resulting from the interaction of electrons with the specimen (radiation damage) requires specific procedures to transfer the specimen into a state in which it can be analysed. During these procedures, which are more elaborate for organic than for inorganic solid-state materials, structural and compositional aspects of the specimen may be altered and consequently the corresponding information may be misleading or completely lost. It must be mentioned here that the EM is only the tool to extract information from the specimen. As well as having its own physical problems (CTF, beam damage, etc) it is—like any other microscopy—clearly not capable of restoring information that has been lost or altered during specimen preparation.

Specimens for (S)TEM have to be transparent to the electron beam. In order to get good contrast and resolution, they have to be thin enough to minimize inelastic scattering. The required thin sections of organic materials can be obtained by ultramicrotomy either after embedding into suitable resins (mostly epoxy- or methacrylate resins [11]) or directly at low temperatures by cryo-ultramicrotomy [12].

Ultramicrotomy is sometimes also used to produce thin samples of solid materials, such as metals [13] which are, however, preferentially prepared by chemical- or ion-etching (see [1]) and focused ion beam (FIB) techniques [14].

Bulk specimens for SEM also have to resist the impact of the electron beam instrument. While this is generally a minor problem for materials science specimens, organic and aqueous biological samples must be observed either completely dry or at low enough temperatures for the evaporation/sublimation of solvents and water to be negligible. Internal structures of aqueous biological samples can be visualized by cryosectioning or cryofracturing procedures [15,16]. Similar procedures are used in the preparation of polymers and composites [17]. Fracturing and field-ion beam procedures are used to expose internal structures of semiconductors, ceramics and similar materials.

The preparation of biological specimens is particularly complex. The ultrastructure of living samples is related to numerous dynamic cellular events that occur in the range of microseconds to milliseconds [18]. Interpretable high-resolution structural information (e.g. preservation of dimensions, or correlation of the structural detail with a physiologically or biochemically controlled state) is therefore obtained exclusively from samples in which life has been stopped very quickly and with a sufficiently high time resolution for the cellular dynamics [19]. Modern concepts for specimen preparation therefore try to avoid traditional, chemical

fixation as the life-stopping step because it is comparatively slow (diffusion limited) and cannot preserve all cellular components. Cryotechniques, often in

-14-

combination with microscopy at low temperatures, are used instead. Very high cooling rates ($> 10\,000\text{ K s}^{-1}$) are required to prevent the formation and growth of ice crystals, which would affect the structural integrity. Such high cooling rates, at the same time, result in a rapid arrest of the physiological events, i.e. produce a very high temporal resolution (microseconds to milliseconds) [20], in capturing dynamic processes in the cell [21,22].

Despite the drawbacks of chemical-fixation based procedures [23,24], most of our current knowledge on biological ultrastructure relies on this approach. In contrast to cryopreparative procedures, chemical fixation does not require special skills and instrumentation.

Cryoimmobilization procedures that lead to vitrification (immobilization of the specimen water in the amorphous state) are the sole methods of preserving the interactions of the cell constituents, because the liquid character of the specimen water is retained (reviewed in [25]).

Vitrification at ambient pressure requires very high cooling rates. It can be accomplished by the 'bare grid' approach for freezing thin ($> 100\text{ nm}$) aqueous layers of suspensions containing isolated macromolecules, liposomes, viruses, etc. This technique was used to produce [figure B1.17.6](#). It has developed into a powerful tool for structural biology, now providing subnanometre resolution of non-crystalline objects [26,27]. The bare-grid technique permits imaging of macromolecules in functional states with sufficient resolution to allow the correlation with atomic data from x-ray diffraction of crystals [28,29] (see also *Journal of Structural Biology*, vol 125, 1999).

High-pressure freezing is at present the only practicable way of cryoimmobilizing larger non-pretreated samples [30,32]. At a pressure of 2100 bars, about the 10-fold greater thickness can be vitrified, as compared to vitrification at ambient pressure [33] and a very high yield of adequately frozen specimens (i.e. no detectable effects of ice crystal damage visible after freeze substitution) has been demonstrated by TEM of suspensions of micro-organisms, as well as for plant and animal tissue, provided that the thickness of the aqueous layer did not exceed $200\text{ }\mu\text{m}$.

Biological material, immobilized chemically or by rapid freezing, must be transferred into an organic solvent that is compatible with the most frequently used hydrophobic resins. Chemically fixed materials are dehydrated in graded series of alcohol or acetone at room temperature. The ice of the frozen sample is dissolved at low temperature by a freeze-substitution process [34]. For TEM, the samples are embedded in resin [11], for SEM they are dried, most frequently by the critical-point drying technique, which avoids deleterious effects of the surface tension of the solvents. Dehydration and complete drying results in non-isotropic shrinkage of biological materials.

The information that can be extracted from inorganic samples depends mainly on the electron beam/specimen interaction and instrumental parameters [1], in contrast to organic and biological materials, where it depends strongly on specimen preparation.

For analytical SEM and non-destructive imaging (e.g. semiconductor, critical-dimension measurements (CD) and other quality control) adequate electron energies have to be selected in order to minimize charging up of the specimen. For high-resolution imaging of surfaces using the SE signal, the signal source often must be localized at the specimen surface by a thin metal coating layer [35].

B1.17.5 IMAGE FORMATION AND IMAGE CONTRAST

Whereas electron optics, sample preparation and the interaction of electrons with the sample follow a common set of rules for all different kinds of EM (TEM, STEM, SEM), the image formation and image contrast in EM images is very technique-specific. In general, analytical imaging is distinguished from structural imaging, the latter being further classified either into the imaging of the projected 2D scattering potential for TEM and STEM or into the topographical imaging of a surface in SEM. The complete understanding of the electron–sample interaction and the increasingly better understanding of the sample preparation and reconstruction of the object from image contrast allows a quantitative interpretation of EM data. The electron microscope has evolved, over a long period, from a simple imaging microscope to a quantitative data collection device.

B1.17.5.1 IMAGING OF PROJECTED STRUCTURE—THE CONTRAST TRANSFER FUNCTION (CTF) OF TEM

The discussion of the electron–specimen interaction has already provided the necessary physical principles leading to amplitude and phase changes of the scattered electron wave. Consider again the elastic scattering as described by [equation \(B1.17.1\)](#). In STEM the elastic scattering is measured by an angular detector, integrating over all electrons scattered to high angles (see [figure B1.17.3\(a\)](#)). For thin samples, the measured image contrast can be directly interpreted as the spatial distribution of different atomic composition. It corresponds to the pointwise measurement of the sample’s scattering factor amplitude (see [equation \(B1.17.9\)](#))

$$f_{\text{int}} = \int_{\theta_{\text{det. min}}}^{\theta_{\text{det. max}}} \int_0^{2\pi} f(\theta) \, d\varphi \, d\theta \quad (\text{B1.17.13})$$

where θ denotes the scattering angle, constrained by the geometry of the angular detector; φ denotes the azimuthal angle. According to the properties of the elastic scattering distribution ([equation \(B1.17.1\)](#)), the detected signal for a given detection angle interval depends strongly on the atomic number Z of the scattering atom. This results in different contrast for different atomic composition (Z -contrast). With a different, position selective, detector it is also possible to measure the phase part of the scattering factor. The geometry of these detectors is illustrated in [figure B1.17.3\(b\)](#) [9].

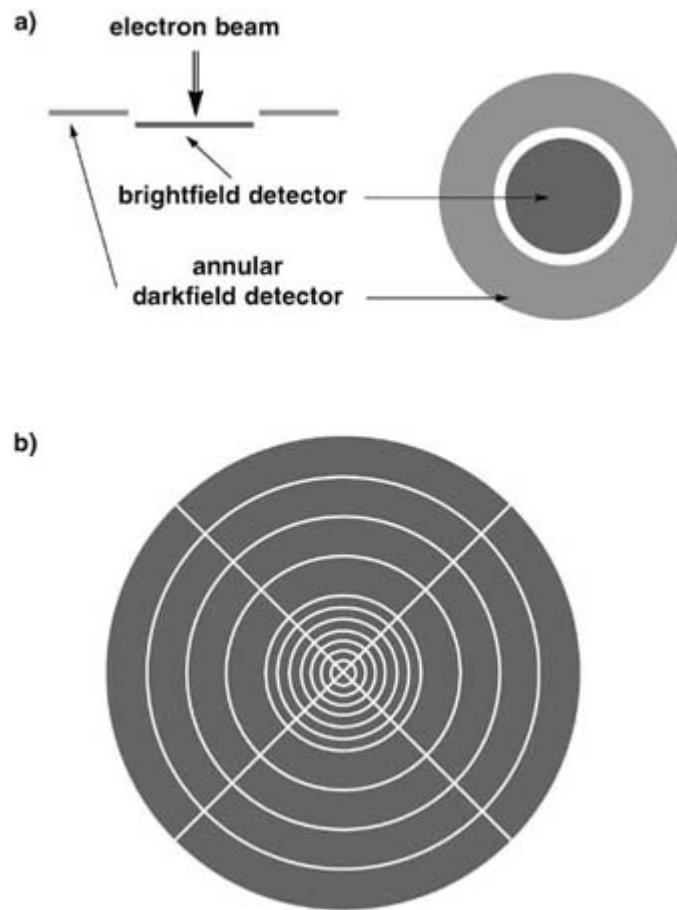


Figure B1.17.3. STEM detectors: (a) conventional bright and dark-field detectors, electrons are detected according to their different scattering angles, all other positional information is lost; (b) positional detector as developed by Haider and coworkers (Haider *et al* 1994).

It should be noted again that STEM provides lateral spatial discrimination of the 2D projected sample by the scanning of a point-like electron beam. Spatial resolution is thereby given by the focus of the incident beam, which is at present limited to a typical diameter of a few Ångströms. Modern TEM, using the very coherent electron wave of a FEG and higher electron energies (see B1.17.3.1), delivers higher resolution, i.e. its information limit can be improved into the sub-Ångström regime. However, the correspondence of image contrast and scattering factor $f(\theta)$ (equation (B1.17.9)) is more complicated than in STEM. In TEM, image contrast can be understood either by interference effects between scattered and unscattered parts of the electron wave, or by simple removal of electrons scattered to higher angles (scattering contrast). The latter is important for imaging of strongly scattering objects consisting of heavy metal atoms. Electrons scattered to high angles are easily removed by a circular aperture (see figure B1.17.4(a)). Elastically scattered electrons not removed by such an aperture can form interference patterns with the unscattered part of the incident electron wave (see figure B1.17.4(b)). Such interference patterns lead to either diffraction patterns or images of the sample, depending on the imaging conditions of the microscope. Therefore, any wave aberrations of the electron wave, both by the imaged sample (desired signal) and by lens aberrations or defocused imaging, result in a change of image

contrast. Mathematically, this behaviour is described by the concept of contrast transfer in spatial frequency space (Fourier space of image), modelled by the CTF.

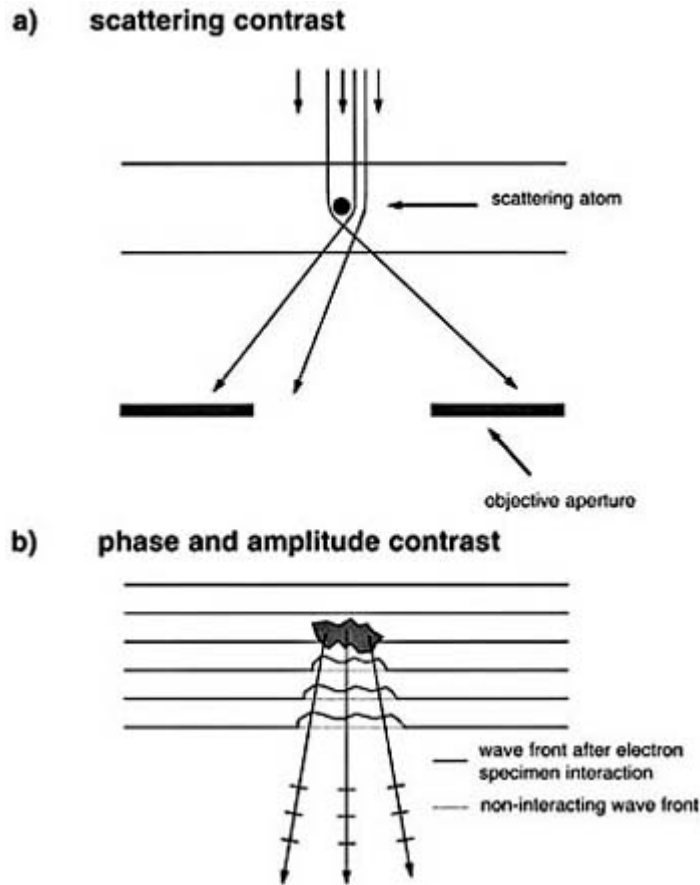


Figure B1.17.4. Visualization of image contrast formation methods: (a) scattering contrast and (b) interference contrast (weak phase/weak amplitude contrast).

In the simplest case of bright-field imaging, the CTF can easily be deduced: the elastically scattered electron wave can be described using a generalized phase shift Φ_{gen} by

$$\psi_{\text{scattered}}(\mathbf{r}) = \psi_0(\mathbf{r}) \exp(i\Phi_{\text{gen}}(\mathbf{r})) \quad (\text{B1.17.14})$$

where

$$\exp(i\Phi_{\text{gen}}(\mathbf{r})) = \exp(i\varphi_{\text{el}}(\mathbf{r}) + \mu_{\text{el}}(\mathbf{r})) \quad (\text{B1.17.15})$$

and $\varphi_{\text{el}}(\mathbf{r})$ and $\mu_{\text{el}}(\mathbf{r})$ denote the elastic phase and amplitude contrast potential. In the notation of [equation \(B1.17.15\)](#), $\varphi_{\text{el}}(\mathbf{r})$ denotes a positive phase potential whereas $\mu_{\text{el}}(\mathbf{r})$ denotes a negative absorption potential. Assuming a weak phase/weak amplitude object—compared to the unscattered part of the electron wave—the generalized phase shift can be reduced to

$$\exp(i\Phi_{\text{gen}}(\mathbf{r})) = 1 + i\varphi_{\text{el}}(\mathbf{r}) + \mu_{\text{el}}(\mathbf{r}). \quad (\text{B1.17.16})$$

After propagation into the back focal plane of the objective lens, the scattered electron wave can be expressed in terms of the spatial frequency coordinates k as

$$\tilde{\psi}_{\text{scattered}}(\mathbf{k}) = (\delta(\mathbf{k}) + i\tilde{\varphi}_{\text{el}}(\mathbf{k}) + \tilde{\mu}_{\text{el}}(\mathbf{k})) \times \exp(-iW(\mathbf{k})). \quad (\text{B1.17.17})$$

Here $W(\mathbf{k})$ denotes the wave aberration

$$W(\mathbf{k}) = \frac{\pi}{2}(C_s\lambda^3k^4 - 2\Delta z\lambda k^2) \quad (\text{B1.17.18})$$

with the objective lens spherical aberration C_s , the electron wave length λ and the defocus Δz . It should be noted that, in the above formulae, the effect of inelastic scattering is neglected. For a rigorous discussion of image contrast, including inelastic scattering, see [36].

In the usual approximation of the object as a weak phase/weak amplitude object, this scattered wave can be used to calculate the intensity of the image transform as

$$\tilde{I}(\mathbf{k}) = \tilde{\psi}_{\text{scattered}}(\mathbf{k}) \otimes \tilde{\psi}_{\text{scattered}}^*(\mathbf{k}). \quad (\text{B1.17.19})$$

Calculating the convolution using equation (B1.17.17) and regrouping the terms yields the final equation for the image transform:

$$\tilde{I}(\mathbf{k}) = \delta(\mathbf{k}) - 2 \times (\tilde{\varphi}_{\text{el}}(\mathbf{k})) \sin(W(\mathbf{k})) + \tilde{\mu}_{\text{el}}(\mathbf{k}) \cos(W(\mathbf{k})). \quad (\text{B1.17.20})$$

The power spectrum of the image $PS(\mathbf{k})$ is then given by the expectation value $\langle \tilde{I}(\mathbf{k}) \times \tilde{I}(\mathbf{k}) \rangle$, normally calculated as the squared amplitude of the image transform. More detailed discussions of the above theory are found in [1,2].

In the conventional theory of elastic image formation, it is now assumed that the elastic atomic amplitude scattering factor is proportional to the elastic atomic phase scattering factor, i.e.

$$\tilde{\mu}_{\text{el}}(\mathbf{k}) = -A(\mathbf{k})\tilde{\varphi}(\mathbf{k}) \equiv -A\tilde{\varphi}_{\text{el}}(\mathbf{k}). \quad (\text{B1.17.21})$$

The factor A has been measured for a variety of samples, indicating that the approximation can be applied up to quasi-atomic resolution. In the case of biological specimens typical values of A are of the order of 5–7%, as determined from images with a resolution of better than 10 Å [37,38]. For an easy interpretation of image contrast and a retrieval of the object information from the contrast, such a combination of phase and amplitude information is necessary.

Figure B1.17.5 shows typical examples of the CTF for weak phase, weak amplitude or combined samples. The resulting effect on image contrast is illustrated in Figure B1.17.6, which shows image averages of a protein complex embedded in a vitrified aqueous layer recorded at different defocus levels. The change in contrast and visible details is clear, but a direct interpretation of the image contrast in terms of object structure is not possible. To reconstruct the imaged complex, it is necessary to combine the information from the different images recorded with different defocus levels. This was first suggested by Schiske [39] and is normally applied to high-resolution images in materials science or biology. Such correction procedures are necessary to rectify the imaging aberrations from imperfect electron-optical systems, which result in a delocalized contrast of object points. In recent years, improvements of the electron-optical lenses have been made, and high-resolution imaging with localized object contrast will be possible in the future (see B1.17.5.3

and [figure B1.17.9](#)). It should be noted that any high-resolution interpretation of an EM image strongly depends on the correction of the CTF. In high-resolution TEM in materials science, sophisticated methods for this correction have been developed and are often combined with image simulations, assuming a certain atomic model structure of the sample. Three mainstream developments in this field are: (1) the use of focal series and subsequent image processing [[40,41](#)], (2) electron holography [[42,43](#) and [44](#)] and (3) the development of corrected TEMs, which prohibit contrast delocalization (see [B1.17.5.3](#)).

-20-

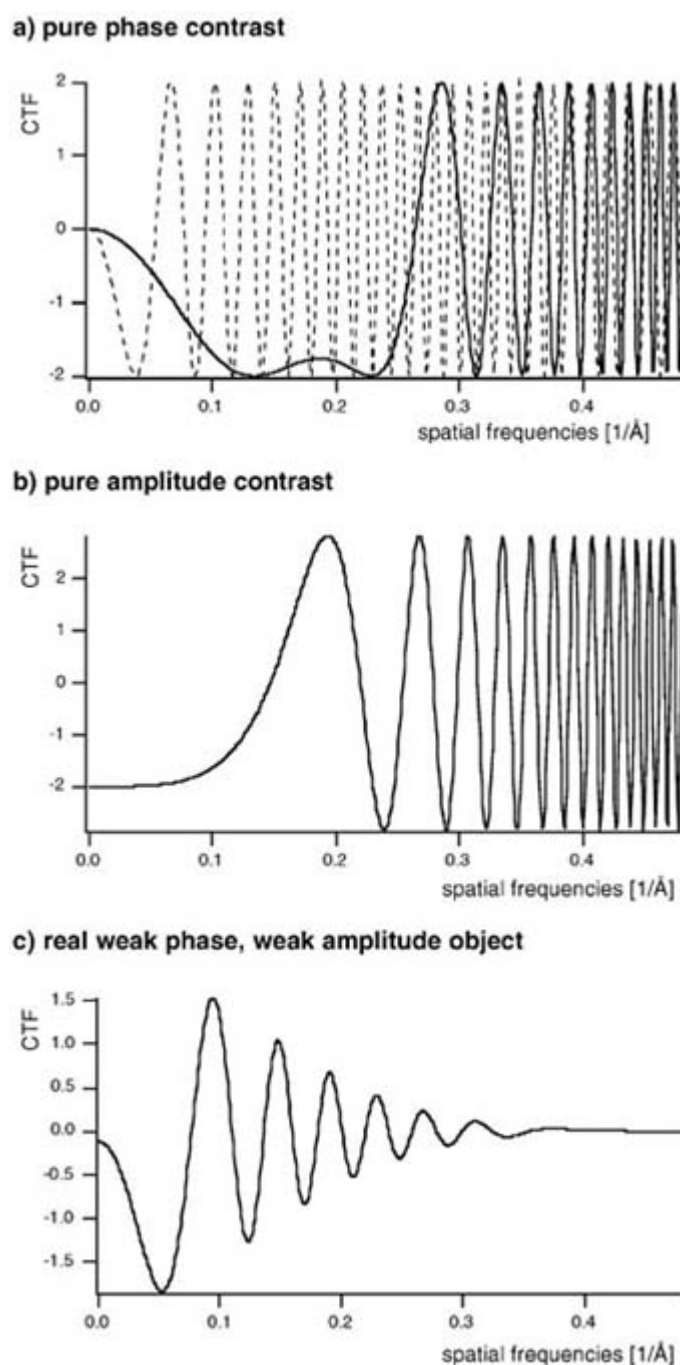


Figure B1.17.5. Examples of CTFs for a typical TEM (spherical aberration $C_s = 2.7$ mm, 120 keV electron energy). In (a) and (b) the idealistic case of no signal decreasing envelope functions [[77](#)] are shown. (a) Pure phase contrast object, i.e. no amplitude contrast; two different defocus values are shown (Scherzer focus of 120 nm underfocus (solid curve), 500 nm underfocus (dashed curve)); (b) pure amplitude object (Scherzer focus of 120 nm underfocus); (c) realistic case including envelope functions and a mixed weak

amplitude/weak phase object (500 nm underfocus).

-21-

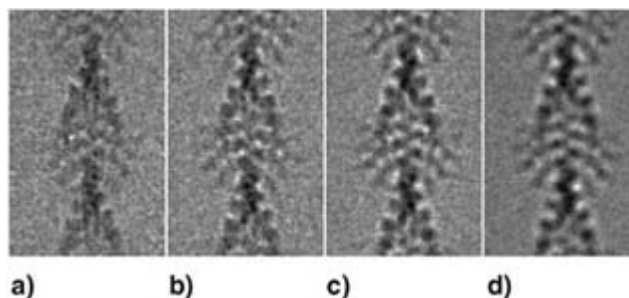


Figure B1.17.6. A protein complex (myosin S1 decorated filamentous actin) embedded in a vitrified ice layer. Shown is a defocus series at (a) 580 nm, (b) 1130 nm, (c) 1700 nm and (d) 2600 nm underfocus. The pictures result from averaging about 100 individual images from one electron micrograph; the decorated filament length shown is 76.8 nm.

B1.17.5.2 IMAGING OF SURFACE TOPOLOGY

SEMs are ideally suited to study highly corrugated surfaces, due to the large depth of focus. They are generally operated with lower beam energies (100 eV–30 keV) in order to efficiently control the volume in which the electron beam interacts with the sample to produce various specific signals (see [figure B1.17.1](#) for imaging and compositional analysis. Modern SEM instruments (equipped with a field emission electron source) can scan the sample surface with a beam diameter of 1 nm or smaller, thus providing high-resolution structural information that can complement the information obtained from atomic force microscopy (AFM). In contrast to AFM, which directly provides accurate height information in a limited range, quantitative assessment of the surface topography by SEM is possible by measuring the parallax of stereo pairs [45].

High-resolution topographic information is obtained by the secondary-electron signal (SE 1, see [figure B1.17.7](#) produced at the point of impact of the primary beam (PE). The SE 1 signal alone is related to the position of the scanning beam. It depends on the distance the primary beam travels through the SE escape zone, where it releases secondary electrons that can leave the specimen surface, i.e. it depends on the angle of impact of the primary beam (see [figure B1.17.7](#)). The high-resolution topographic signal is obscured by other SE signals (SE 2, SE 3, [figure B1.17.7](#)) that are created by BSE (electrons of the primary beam, multiply scattered deeper inside the specimen) when they leave the specimen and pass through the SE escape zone (SE 2) and hit the pole pieces of the objective lens and/or the walls of the specimen chamber (SE 3). Additional background signal is produced by the primary beam striking the objective lens aperture.

-22-

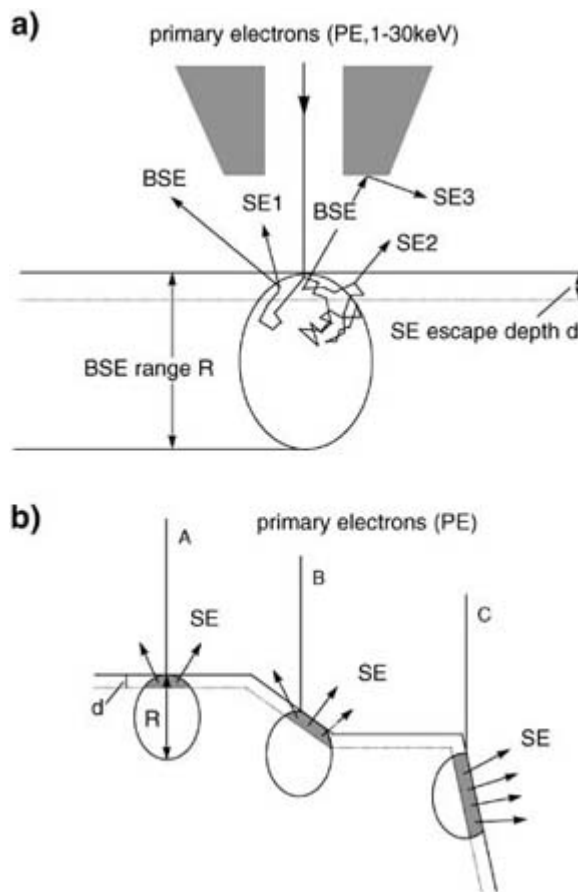


Figure B1.17.7. (a) Classification of the secondary electron signals. High-resolution topographic information is obtained by the SE 1 signal. It might be obscured by SE 2 and SE 3 signals that are created by the conversion of BSE. d : SE escape depth, R : range out of which BSE may leave the specimen. (b) SE signal intensity, $R > d$: the SE 1 signal depends on the angle of impact of the electron beam (PE). SE 1 can escape from a larger volume at tilted (B) surfaces and edges (C) than at orthogonal surfaces (A).

High-resolution topographic imaging by secondary electrons therefore demands strategies (instrumentation, specimen preparation [35] and imaging conditions) that aim at enhancing the SE 1 signal and suppressing the background noise SE signals (e.g. [46]). Basically, the topographic resolution by SE 1 depends on the smallest spot size available and on the SE escape zone, which can be up to 100 nm for organic materials and down to 1–2 nm for metals.

Non-conductive bulk samples, in particular, are frequently rendered conductive by vacuum coating with metals using sputter or evaporation techniques. The metal coating should be of uniform thickness and significantly thinner than the smallest topographic details of interest. Metal coating provides the highest resolution images of surface details. It may, however, irreversibly destroy the specimen. An example of such a metal-coated sample is shown in [figure B1.17.8](#).

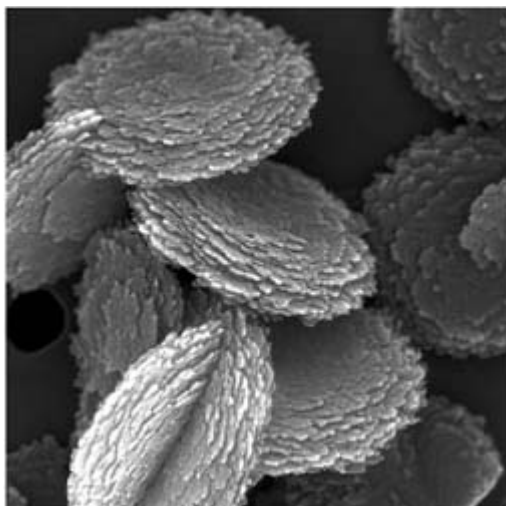


Figure B1.17.8. Iron oxide particles coated with 4 nm of Pt in an m-planar magnetron sputter coater (Hermann and Müller 1991). Micrographs were taken in a Hitachi S-900 ‘in-lens’ field emission SEM at 30,000 primary magnification and an acceleration voltage of 30 kV. Image width is 2163 nm.

SEM with low acceleration voltage (1–10 kV) (LVSEM) can be applied without metal coating of the sample, e.g. for quality control purposes in semiconductor industries, or to image ceramics, polymers or dry biological samples. The energy of the beam electrons (the acceleration voltage) should be selected so that charge neutrality is approached, i.e. the amount of energy that enters the sample also leaves the sample in the form of SE and BSE. Modern SEM instruments, equipped with FEGs provide an adequate spot size, although the spot size increases with decreasing acceleration voltage. The recent implementation of a cathode lens system [47] with very low aberration coefficients will allow the surfaces of non-metal coated samples at beam energies of only a few electronvolts to be imaged without sacrificing spot size. New contrast mechanisms and new experimental possibilities can be expected.

The fact that electron beam instruments work under high vacuum prohibits the analysis of aqueous systems, such as biological materials or suspensions, or emulsions without specimen preparation as outlined above. These preparation procedures are time consuming and are often not justified in view of the only moderate resolution required to solve a specific practical question (e.g. to analyse the grain size of powders, bacterial colonies on agar plates, to study the solidification of concrete, etc). Environmental SEM (ESEM) and ‘high-pressure SEM’ instruments are equipped with differentially pumped vacuum systems and Peltier-cooled specimen stages, which allow wet samples to be observed at pressures up to 5000 Pa [48]. Evaporation of water from the specimen or condensation of water onto the specimen can thus be efficiently controlled. No metal coating or other preparative steps are needed to control charging of the specimen since the interaction of the electron beam with the gas molecules in the specimen chamber produces positive ions that can compensate surface charges. ‘High-pressure SEM’, therefore, can study insulators without applying a conductive coating. The high gas pressure in the vicinity of the specimen leads to a squirting of the electron beam. Thus the resolution-limiting spot size achievable on the specimen surface depends on the acceleration voltage, the gas pressure, the scattering cross section of the gas and the distance the electrons have to travel through the high gas pressure zone [49]. High-pressure SEM and ESEM is still under development and the scope of applications is expanding. Results to date consist mainly of analytical and low-resolution images (e.g. [50]).

B1.17.5.3 MODERN DEVELOPMENTS OF INSTRUMENTS AFFECTING IMAGE CONTRAST AND RESOLUTION

As was discussed above, the image contrast is significantly affected by the aberrations of the electron-optical

lenses. The discussion on the CTF showed that the broadening PSF of the TEM delocalizes information in an TEM image. This necessitates additional techniques to correct for the CTF, in order to obtain interpretable image information. Furthermore, it was discussed that the resolution of STEM and SEM depends on the size of the focused beam, which is also strongly dependent on lens aberrations. The leading aberrations in state-of-the-art microscopes are the spherical and chromatic aberrations in the objective lens. Correction of such aberrations was discussed as early as 1947, when Scherzer suggested the correction of electron optical lenses [51], but it was not until 1990 that a complete concept for a corrected TEM was proposed by Rose [52]. It was in the last decade that prototypes of such corrected microscopes were presented.

The first corrected electron-optical SEM was developed by Zach [10]. For low-voltage SEM (LVSEM, down to 500 eV electron energy instead of the conventional energies of up to 30 keV) the spot size is extremely large without aberration correction. Combining C_s and C_c correction and an electrostatic objective lens, Zach showed that a substantial improvement in spot size and resolution is possible. The achievable resolution in a LVSEM is now of the order of 1–2 nm. More recently, Krivanek and colleagues succeeded in building a C_s corrected STEM [53,54].

The construction of an aberration-corrected TEM proved to be technically more demanding: the point resolution of a conventional TEM today is of the order of 1–2 Å. Therefore, the aim of a corrected TEM must be to increase the resolution beyond the 1 Å barrier. This implies a great number of additional stability problems, which can only be solved by the most modern technologies. The first C_s corrected TEM prototype was presented by Haider and coworkers [55]. Figure B1.17.9 shows the improvement in image quality and interpretability gained from the correction of the spherical aberration in the case of a materials science sample.

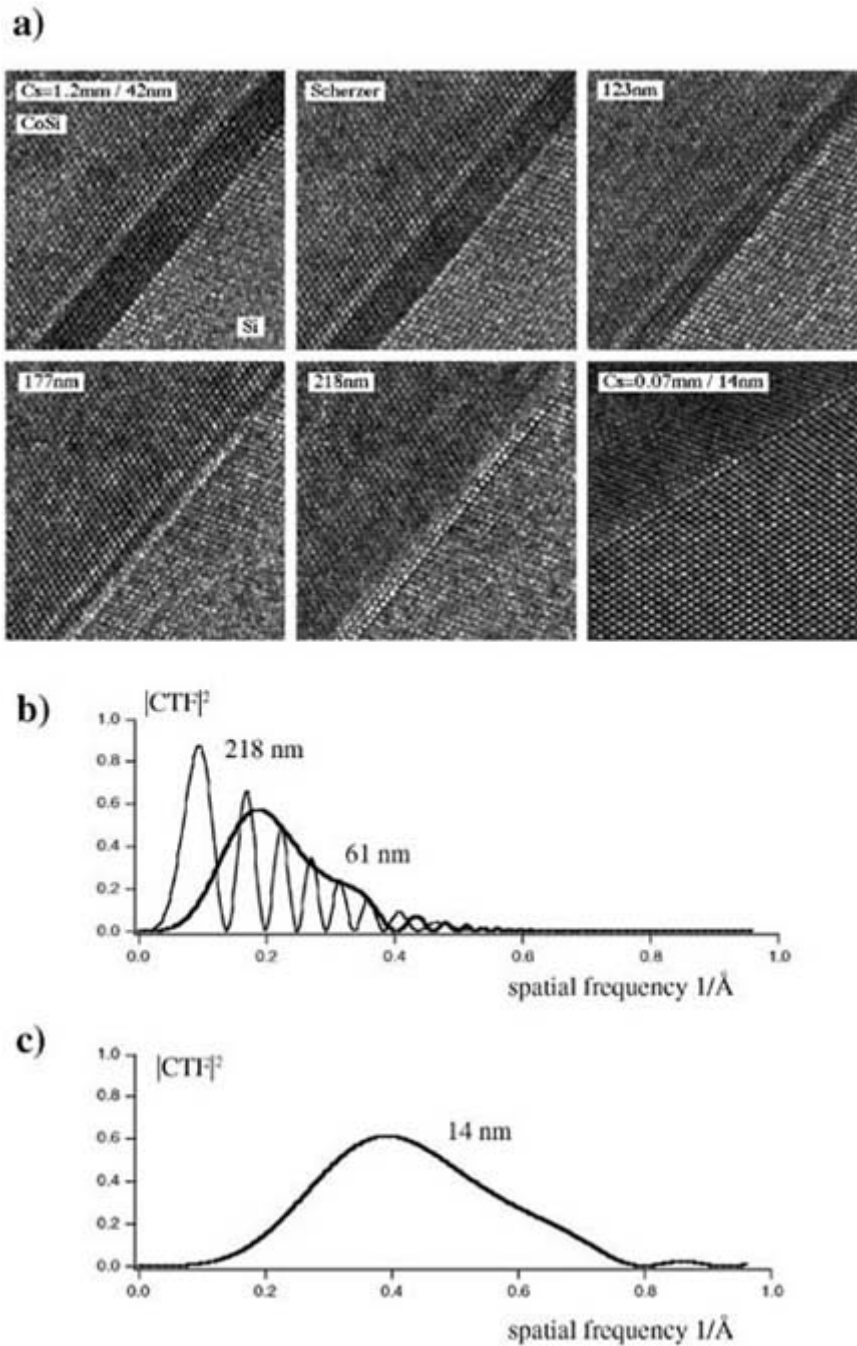


Figure B1.17.9. A CoSi grain boundary as visualized in a spherical-aberration-corrected TEM (Haider *et al* 1998). (a) Individual images recorded at different defocus with and without correction of C_s ; (b) CTFs in the case of the uncorrected TEM at higher defocus; (c) CTF for the corrected TEM at only 14 nm underfocus. Pictures by courtesy of M Haider and Elsevier.

The field of corrected microscopes has just begun with the instruments discussed above. The progress in this field is very rapid and proposals for a sub-Ångström TEM (SATEM) or even the combination of this instrument with a corrected energy filter to form a sub-electronvolt/sub-Ångström TEM (SESAM) are underway.

B1.17.6 ANALYTICAL IMAGING, SPECTROSCOPY, AND MASS MEASUREMENTS

For a detailed discussion on the analytical techniques exploiting the amplitude contrast of inelastic images in ESI and image-EELS, see [chapter B1.6](#) of this encyclopedia. One more recent but also very important aspect is the quantitative measurement of atomic concentrations in the sample. The work of Somlyo and colleagues [56], Leapman and coworkers [57,58], and Door and Gängler [59] introduce techniques to convert measured intensities of inelastically scattered electrons directly into atomic concentrations.

For bio-medical or cell-biological samples, in particular, this provides a direct measurement of physiological ion concentrations. The main disadvantage of such methods is the almost unavoidable delocalization of free ions and the resulting change in concentration during sample preparation steps. The discussed rapid freezing and the direct observation of samples without any chemical treatment provides a very good compromise for organic samples. In materials science, delocalization does not seem to pose a major problem.

Another specialized application of EM image contrast is mass measurement. Using the elastic dark-field image in the STEM or the inelastic image in the EFTEM, a direct measurement of the scattering mass can be performed. For reviews on this technique see [60,61].

B1.17.7 3D OBJECT INFORMATION

EM images are always either 2D projections of an interaction potential (see equation (B1.17.11) and equation (B1.17.12)) or a surface topology encoded in grey levels of individual image points (image contrast). The aim of EM image processing is to reconstruct the 3D object information from a limited number of such projections. This problem does not arise for all applications of EM. Very often in materials science the sample is prepared in such a way that one single projection image contains all the information necessary to answer a specified question. As an example, consider [figure B1.17.9](#) which shows a Co–Si interface. The orientation of the sample is chosen to give a perfect alignment of atoms in the direction of the grain boundary. Imaging at atomic resolution allows the direct interpretation of the contrast as images of different atoms. One single exposure is, in principle, sufficient to collect all the information needed. It should be noted here again that this is only true for the spherical-aberration, C_s corrected EM with its non-oscillating CTF (bottom right panel in [figure B1.17.9\(a\)](#)). As is obvious from the Co–Si interface ([figure B1.17.9\(a\)](#)) finite C_s imaging) and the defocus series of the biological sample ([figure B1.17.6](#)) more than one image has to be combined for conventional EMs. However, such a direct interpretation of one projected image to obtain the 3D structure information works only for samples that are ordered crystallographically at the atomic level. For other samples it is necessary to combine projections from different angles. Such samples are unique, non-crystallographic structures, e.g. structural defects in materials science, cellular compartments in cell biology, or macromolecules (protein complexes or single biological molecules) in high-resolution molecular imaging.

The generalized problem has been solved by tomography. In EM it is possible to tilt the sample along an axis which is perpendicular to the electron beam (single-axis tomography, see [figure B1.17.10](#)). If the sample can withstand a high cumulative electron dose, then an unlimited number of exposures at different defocus values and tilting angles can be recorded. If kinematical scattering theory can be applied (single elastic scattering, compare [equation \(B1.17.8\)](#)), then it is possible to correct all the effects of the CTF and each corrected projection image at one particular tilting angle corresponding to a section of the Fourier-transformed scattering potential [equation \(B1.17.11\)](#) and [equation \(B1.17.12\)](#). The combination of information from different tilting angles provides the determination of the structure factors in the complete 3D Fourier space. Finally, a simple mathematical inverse Fourier-transform produces a complete 3D reconstructed object. A

geometric equivalent of the projection and reconstruction process is found in the sectioning of the Ewald-sphere with the 3D Fourier-transform of the scattering potential. For an introduction to the concepts of the Ewald-sphere and Fourier techniques in structure determination see [62].

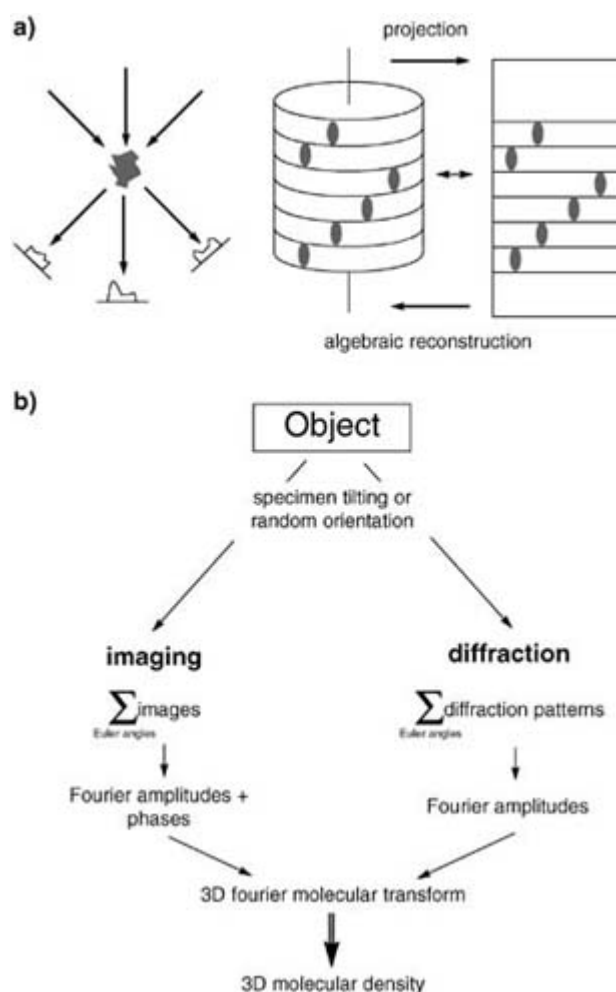


Figure B1.17.10. Principles of 3D reconstruction methods. (a) Principle of single axis tomography: a particle is projected from different angles to record corresponding images (left panel); this is most easily realized in the case of a helical complex (right panel). (b) Principle of data processing and data merging to obtain a complete 3D structure from a set of projections.

The basic reconstruction algorithms involved are mathematically well known and are well established [63]. A variety of concepts has evolved for single- and multi-axis tomography, combining projection information, calculating 3D object densities either with Fourier- or real-space algorithms (see figure B1.17.10 which shows some examples of the geometry used for the single- and multi-axis tomography). For a complete reference to the methods used see Frank [64] and the special issue of *Journal of Structural Biology* (vol 120, 1997).

The main disadvantage of the tomographic approach is the beam-induced destruction of the sample. In practice, one can record only a limited number of images. Therefore, it is not possible to correct the CTF completely or to obtain an infinite sampling of the projection angles from only one specimen. Two major approaches are used today: the single- and double-axis tomography of one individual object (e.g. cell organelles, see Baumeister and coworkers [65]) and, second, the imaging of many identical objects under different projection angles (see [64]; random conical tilt, angular reconstitution).

For single-axis tomography, with its limited number of images and the subsequent coarse sampling in reciprocal Fourier space, only a moderate resolution can be expected. For chemically fixed samples with high image contrast from heavy atom staining it is possible to obtain a resolution of about 4 nm ([66], reconstruction of the centrosome). For native samples, true single-axis tomography without averaging over different samples results in even lower resolution. Today, sophisticated EM control software allows a fully automatic collection of tilted images [67], making single-axis tomography a perfect reconstruction tool for unique objects.

If many identical copies of the object under study are available, other procedures are superior. They rely on the fact that the individual molecules are oriented with respect to the incident electron beam. Such a situation is found mainly for native ice-embedded samples (compare the paragraph about preparation). In ice layers of sufficient thickness, no special orientation of the molecule is preferred. The obtained projection images from one, untilted image can then be classified and aligned in an angular reconstitution reconstruction process. By averaging large numbers of projection images, it is possible to correct for CTF effects [68] and to obtain an almost complete coverage of reciprocal Fourier space. If—for some reason—the object still shows a limited number of preferred orientations, an additional tilting of the sample again gives complete coverage of possible projection angles (random conical tilt method). Both methods have been successfully applied to many different biological samples (for an overview, see [64]).

An important point for all these studies is the possible variability of the single molecule or single particle studies. It is not possible, *a priori*, to exclude ‘bad’ particles from the averaging procedure. It is clear, however, that high structural resolution can only be obtained from a very homogeneous ensemble. Various classification and analysis schemes are used to extract such homogeneous data, even from sets of mixed states [69]. In general, a typical resolution of the order of 1–3 nm is obtained today.

The highest resolutions of biological samples have been possible for crystalline samples (electron crystallography). 2D crystals of membrane proteins and one cytoskeletal protein complex have been solved at the 3–4 Å level combining imaging and electron diffraction, as pioneered by Henderson and coworkers [70,71 and 72], also see [figure B1.17.11](#). Icosahedral virus particles are reconstructed from images to 8–9 Å resolution [26,27], allowing the identification of alpha helices. Compared to single particles, these samples give much higher resolution, in part because much higher numbers of particles are averaged, but it is also possible that a crystallization process selects for the uniformity of the crystallizing object and leads to very homogeneous ensembles.

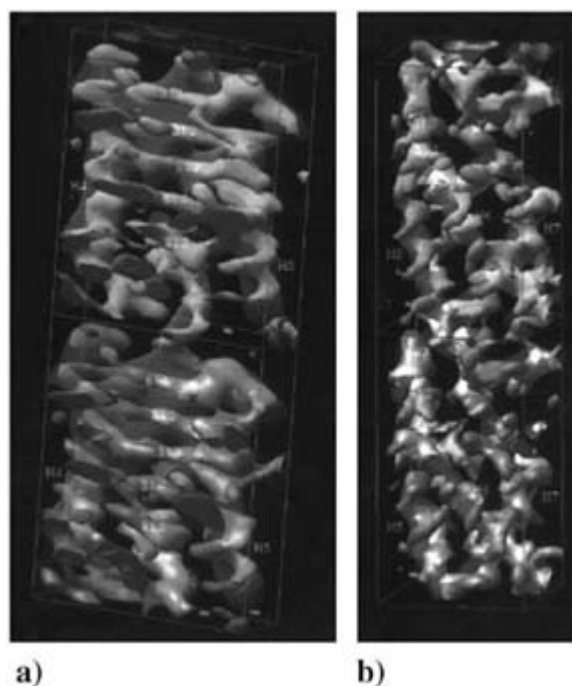


Figure B1.17.11. Reconstructed density of an α,β -tubulin protein dimer as obtained from electron crystallography (Nogales *et al* 1997). Note the appearance of the β -sheets ((a), marked B) and the α -helices ((b), marked H) in the density. In particular the right-handed α -helix H6 is very clear. Pictures by courtesy of E Nogales and Academic Press.

For electron crystallography, the methods to obtain the structure factors are comparable to those of conventional x-ray crystallography, except that direct imaging of the sample is possible. This means that both electron diffraction and imaging can be used, i.e. structure amplitudes are collected by diffraction, structure factor phases by imaging. For a general overview in the structure determination by electron crystallography, see [73]. The 3D structure of the sample is obtained by merging diffraction and imaging data of tilt series from different crystals. It is, therefore, a form of tomography adapted for a diffracting object.

Even though it is easy to get the phase information from imaging, in general, imaging at the desired high resolution (for structure determination work of the order of 3–4 Å) is very demanding. Specialized instrumentation (300 kV, FEG, liquid He sample temperature) have to be used to avoid multiple scattering, to allow better imaging (less imaging aberrations, less specimen charging which would affect the electron beam) and to reduce the effects of beam damage.

B1.17.8 TIME-RESOLVED AND *IN SITU* EM STUDIES: VISUALIZATION OF DYNAMICAL EVENTS

As a result of the physical conditions in electron microscopes such as the high vacuum, the high energy load on the sample by inelastic scattering, or the artificial preparation of the sample by sectioning or thinning, it has become customary to think about samples as static objects, precluding the observation of their native structural changes during a reaction. In some studies, however, the dynamics of reactions have been studied for biological systems as well as in

materials science. *In situ* microscopy was widely used in materials science in the 1960s and 1970s, when, for example, metal foils were studied, heated up in the EM, and reactions followed in a kind of time-lapse

microscopy [74]. In recent years, similar experiments have been performed on semiconductors and ceramics, and a general new interest in *in situ* microscopy has developed.

Time-resolved EM in biological systems is a comparatively new and limited field. Simple time-lapse fixation of different samples of a reacting biological tissue has long been used, but the direct, temporal monitoring of a reaction was developed only with the invention of cryo-fixation techniques. Today, time-lapse cryo-fixation studies can be used in the case of systems with slow kinetics, i.e. reaction times of the order of minutes or slower. Here, samples of a reacting system are simply taken in certain time intervals and frozen immediately. For the study of very fast reactions, two approaches have been developed that couple the initiation of the reaction and the fixation of the system on a millisecond time scale. The reaction itself can be started either by a rapid mixing procedure [75] or by the release of a masked reaction partner photolysing caged compounds (see figure B1.17.12 [76]. For a review of time-resolved methods used in biological EM, see [19].

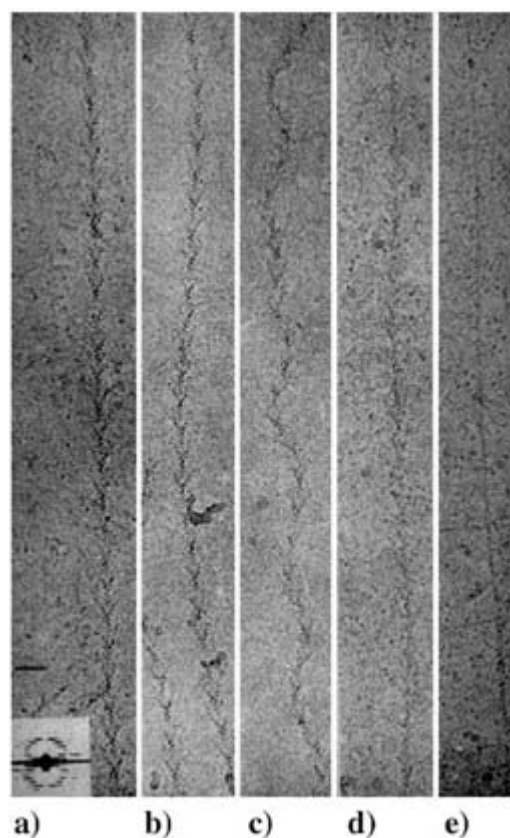


Figure B1.17.12. Time-resolved visualization of the dissociation of myosin S1 from filamentous actin (see also figure B1.17.6). Shown are selected filament images before and after the release of a nucleotide analogue (AMPPNP) by photolysis: (a) before flashing, (b) 20 ms, (c) 30 ms, (d) 80 ms and (e) 2 s after flashing. Note the change in obvious order (as shown by the diffraction insert in (a)) and the total dissociation of the complex in (e). The scale bar represents 35.4 nm. Picture with the courtesy of Academic Press.

REFERENCES

- [1] Williams D B and Carter C B 1996 *Transmission Electron Microscopy, A Textbook for Material Science* (New York: Plenum)
- [2] Reimer L 1993 *Transmission Electron Microscopy* (Berlin: Springer)

- [3] Reimer L 1998 *Scanning Electron Microscopy* (Berlin: Springer)
- [4] Reimer L 1995 *Energy-Filtering Transmission Electron Microscopy* (Berlin: Springer)
- [5] Hawkes P W and Kasper E 1989 *Principles of Electron Optics* vol 1 (London: Academic)
 Hawkes P W and Kasper E 1989 *Principles of Electron Optics* vol 2 (London: Academic)
 Hawkes P W and Kasper E 1994 *Principles of Electron Optics* vol 3 (London: Academic)
- [6] Lenz F 1954 Zur Streuung mittelschneller Elektronen in kleinste Winkel *Z. Naturf.* a **9** 185–204
- [7] Egerton R F 1976 Measurement of inelastic/elastic scattering ratio for fast electrons and its use in the study of radiation damage *Phys. Status Solidi* a **37** 663–8
- [8] Cantor C R and Schimmel P R 1980 *Biophysical Chemistry, Part II: Techniques for the Study of Biological Structure and Function* (San Francisco: Freeman)
- [9] Haider M, Epstein A, Jarron P and Boulin C 1994 A versatile, software configurable multichannel STEM detector for angle-resolved imaging *Ultramicroscopy* **54** 41–59
- [10] Zach J 1989 Design of a high-resolution low-voltage scanning electron microscope *Optik* **83** 30–40
- [11] Luft J H 1961 Improvements in epoxy resin embedding methods *J. Biophys. Biochem. Cytol.* **9** 409–14
- [12] Michel M, Gnägi H and Müller M 1992 Diamonds are a cryosectioner's best friend *J. Microsc.* **166** 43–56
- [13] Malis T F and Steele D 1990 Specimen preparation for TEM of materials II *Mat. Res. Soc. Symp. Proc.* **199** 3
- [14] Dravid V P 1998 *Hitachi Instrument News* 34th edn
- [15] Walther P, Hermann R, Wehrli E and Müller M 1995 Double layer coating for high resolution low temperature SEM *J. Microsc.* **179** 229–37
- [16] Walther P and Müller M 1999 Biological structure as revealed by high resolution cryo-SEM of block faces after cryo-sectioning *J. Microsc.* **196** 279–87
- [17] Roulin-Moloney A C 1989 *Fractography and failure mechanisms of polymers and composites* (London: Elsevier)
- [18] Plattner H 1989 *Electron Microscopy of Subcellular Dynamics* (London: CRC)
- [19] Knoll G 1995 Time resolved analysis of rapid events *Rapid Freezing, Freeze-fracture and Deep Etching* ed N Sievers and D Shotton (New York: Wiley-Lyss) p 105
- [20] Jones G J 1984 On estimating freezing times during tissue rapid freezing *J. Microsc.* **136** 349–60

- [21] van Harreveld A and Crowell J 1964 Electron microscopy after rapid freezing on a metal surface and substitution fixation *Anat. Rec.* **149** 381–6
- [22] Knoll G and Plattner H 1989 Ultrastructural analysis of biological membrane fusion and a tentative correlation with biochemical and biophysical aspects *Electron Microscopy of Subcellular Dynamics* ed H Plattner (London: CRC) pp 95–117
- [23] Hyatt M A 1981 Changes in specimen volume *Fixation for Electron Microscopy* ed M A Hyatt (New York: Academic) pp 299–306

- [24] Coetzee J and van der Merwe F 1984 Extraction of substances during glutaraldehyde fixation of plant cells *J. Microsc.* **135** 147–58
- [25] Dubochet J, Adrian M, Chang J, Homo J-C, Lepault J, McDowell A W and Schultz P 1988 Cryo-electron microscopy of vitrified specimens *Q. Rev. Biophys.* **21** 129–228
- [26] Böttcher B, Wynne S A and Crowther R A 1997 Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy *Nature* **386** 88–91
- [27] Conway J F, Cheng N, Zlotnick A, Wingfield P T, Stahl S J and Steven A C 1997 Visualization of a 4-helix bundle in the hepatitis B virus capsid by cryo-electron microscopy *Nature* **386** 91–4
- [28] Rayment I, Holden H M, Whittaker M, Yohn C B, Lorenz M, Holmes K C and Milligan R A 1993 Structure of the actin–myosin complex and its implications for muscle contraction *Science* **261** 58–65
- [29] Schröder R R, Jahn W, Manstein D, Holmes K C and Spudich J A 1993 Three-dimensional atomic model of F-actin decorated with *Dictyostelium* myosin S1 *Nature* **364** 171–4
- [30] Riehle U and Höchli M 1973 The theory and technique of high pressure freezing *Freeze-Etching Technique and Applications* ed E L Benedetti and P Favard (Paris: Société Française de Microscopie Electronique) pp 31–61
- [31] Müller M and Moor H 1984 Cryofixation of thick specimens by high pressure freezing *The Science of Biological Specimen Preparation* ed J-P Revel, T Barnard and G H Haggis (O'Hare, IL: SEM, AMF 60666) pp 131–8
- [32] Moor H 1987 Theory and practice of high pressure freezing *Cryotechniques in Biological Electron Microscopy* ed R A Steinbrecht and K Zierold (Berlin: Springer) pp 175–91
- [33] Sartori N, Richter K and Dubochet J 1993 Vitrification depth can be increased more than 10-fold by high pressure freezing *J. Microsc.* **172** 55–61
- [34] Steinbrecht R A and Müller M 1987 Freeze-substitution and freeze-drying *Cryotechniques in Biological Electron Microscopy* ed R A Steinbrecht and K Zierold (Berlin: Springer) pp 149–72
- [35] Hermann R and Müller M 1991 High resolution biological scanning electron microscopy: a comparative study of low temperature metal coating techniques *J. Electron. Microsc. Tech.* **18** 440–9
- [36] Angert I, Majorovits E and Schröder R R 2000 Zero-loss image formation and modified contrast transfer theory of EFTEM *Ultramicroscopy* **81** 203–22
- [37] Toyoshima C, Yonekura K and Sasabe H 1993 Contrast transfer for frozen-hydrated specimens II. Amplitude contrast at very low frequencies. *Ultramicroscopy* **48** 165–76
- [38] Toyoshima C and Unwin P N T 1988 Contrast transfer for frozen-hydrated specimens: determination from pairs of defocus images *Ultramicroscopy* **25** 279–92
- [39] Schiske P 1968 Zur Frage der Bildrekonstruktion durch Fokusreihen *Proc. 14th Eur. Conf. on Electron Microscopy* p 145–6

- [40] Frank J and Penczek P 1995 On the correction of the contrast transfer function in biological electron microscopy *Optik* **98** 125–9
- [41] Thust A and Rosenfeld R 1998 State of the art of focal-series reconstruction in HRTEM *Electron Microscopy 1998: 14th Int. Conf. on Electron Microscopy (Cancun)* vol 1 (Bristol: Institute of Physics Publishing) pp 119–20
- [42] Tonomura A 1995 Recent developments in electron holography for phase microscopy *J. Electron Microsc.* **44** 425–35
- [43] Lichte H 1998 Gottfried Möllenstedt and his electron biprism: four decades of challenging and exciting electron physics *J. Electron Microsc.* **47** 387–94
- [44] Lehmann M, Lichte H, Geiger D, Lang G and Schweda E 1999 Electron holography at atomic dimensions—present state *Mater. Character.* **42** 249–63
- [45] Boyde A 1970 Practical problems and methods in the three-dimensional analysis of SEM images *Scanning Electron Microsc.* **105** 112
- [46] Peters K-R 1986 Working at higher magnifications in scanning electron microscopy with secondary and backscattered electrons on metal coated biological specimens and imaging macromolecular cell membrane structures *Science of Biological Specimen Preparation 1985* ed M Müller *et al* (O'Hare, IL: SEM, AMF 60666) pp 257–82
- [47] Frank L, Müllerova I, Faulian K and Bauer E 1999 The scanning low-energy electron microscope: first attainment of diffraction contrast in the scanning electron microscope *Scanning* **21** 1–13
- [48] Danilatos G D 1990 Design and construction of an environmental SEM *Scanning* **12** 23–7
- [49] Adamaik B and Mathieu C 2000 The reduction of the beam gas interactions in the variable pressure scanning electron microscope with the use of helium gas *Scanning* **21** 178
- [50] Manero J M, Masson D V, Marsal M and Planell J L 1999 Application of the technique of environmental scanning electron microscopy to the paper industry *Scanning* **21** 36–9
- [51] Scherzer O 1947 Sphärische und chromatische Korrektur von Elektronen-Linsen *Optik* **2** 114–32
- [52] Rose H 1990 Outline of a spherically corrected semiaplanatic medium-voltage transmission electron microscope *Optik* **85** 19–24
- [53] Krivanek O L, Dellby N, Spence A J and Brown L M 1998 Spherical aberration correction in dedicated STEM *Electron Microscopy 1998: 14th Int. Conf. on Electron Microscopy (Cancun)* vol 1 (Bristol: Institute of Physics Publishing) pp 55–6
- [54] Lupini A R and Krivanek O L 1998 Design of an objective lens for use in Cs-corrected STEM *Electron Microscopy 1998: 14th Int. Conf. on Electron Microscopy (Cancun)* vol 1 (Bristol: Institute of Physics Publishing) pp 59–60
- [55] Haider M, Rose H, Uhlemann S, Schwab E, Kabius B and Urban K 1998 A spherical-aberration-corrected 200 keV transmission electron microscope *Ultramicroscopy* **75** 53–60
- [56] Somlyo A V, Gonzalez-Serratos Y, Shuman H, McClellan G and Somlyo A P 1981 Calcium release and ionic changes in the sarcoplasmic reticulum of tetanized muscle: an electron-probe study *J. Cell Biol.* **90** 577–94
- [57] Leapman R D and Swyt C R 1988 Separation of overlapping core edges in electron energy loss spectra by multiple-least squares fitting *Ultramicroscopy* **26** 393–404
- [58] Leapman R D, Hunt J A, Buchanan R A and Andrews S B 1993 Measurement of low calcium concentrations in cryosectioned cells by parallel-EELS mapping *Ultramicroscopy* **49** 225–34

- [59] Door R and Gängler D 1995 Multiple least-squares fitting for quantitative electron energy-loss spectroscopy—an experimental investigation using standard specimens *Ultramicroscopy* **58** 197–210
-

-34-

- [60] Engel A and Colliex C 1993 Application of scanning transmission electron microscopy to the study of biological structure *Curr. Opin. Biotechnol.* **4** 403–11
- [61] Feja B and Aebi U 1999 Molecular mass determination by STEM and EFTEM: a critical comparison *Micron* **30** 299–307
- [62] Holmes K C and Blow D M 1965 The use of x-ray diffraction in the study of protein and nucleic acid structure *Meth. Biochem. Anal.* **13** 113–239
- [63] Klug A and Crowther R A 1972 Three-dimensional image reconstruction from the viewpoint of information theory *Nature* **238** 435–40
- [64] Frank J 1996 *Three-Dimensional Electron Microscopy of Macromolecular Assemblies* (New York: Academic)
- [65] Baumeister W, Grimm R and Walz T 1999 Electron tomography of molecules and cells *Trends Cell Biol.* **9** 81–5
- [66] Ruiz T 1998 Conference talk *Gordon Conf. on Three-Dimensional Electron Microscopy*
- [67] Koster A J, Grimm R, Typke D, Hegerl R, Stoschek A, Walz J and Baumeister W 1997 Perspectives of molecular and cellular electron tomography *J. Struct. Biol.* **120** 276–308
- [68] Zhu J, Penczek P, Schröder R R and Frank J 1997 Three-dimensional reconstruction with contrast transfer function correction from energy-filtered cryoelectron micrographs: procedure and application to the 70S *Escherichia coli* ribosome *J. Struct. Biol.* **118** 197–219
- [69] Agrawal R K, Heagle A B, Penczek P, Grassucci R A and Frank J 1999 EF-G-dependent GTP hydrolysis induces translocation accompanied by large conformational changes in the 70S ribosome *Nature Struct. Biol.* **6** 643–7
- [70] Henderson R, Baldwin J M, Ceska T, Zemlin F, Beckmann E and Downing K 1990 Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy *J. Mol. Biol.* **213** 899–929
- [71] Kühlbrandt W, Wang D N and Fujiyoshi Y 1994 Atomic model of plant light-harvesting complex by electron crystallography *Nature* **367** 614–21
- [72] Nogales E, Wolf S G and Downing K 1998 Structure of α,β -tubulin dimer by electron crystallography *Nature* **391** 199–202
- [73] Walz T and Grigorieff N 1998 Electron crystallography of two-dimensional crystals of membrane proteins *J. Struct. Biol.* **121** 142–61

- [74] Butler E P and Hale K F 1981 *Practical Methods in Electron Microscopy* vol 9, ed A Glauert (New York: North-Holland)
- [75] Berriman J and Unwin N 1994 Analysis of transient structures by cryo-electron microscopy combined with rapid mixing of spray droplets *Ultramicroscopy* **56** 241–52
- [76] Menetret J-F, Hofmann W, Schröder R R, Rapp G and Goody R S 1991 Time-resolved cryo-electron microscopic study of the dissociation of actomyosin induced by photolysis of photolabile nucleotides *J. Mol. Biol.* **219** 139–43
- [77] Frank J 1973 The envelope of electron microscopic transfer functions for partially coherent illumination *Optik* **38** 519–27
- [78] Nogales E and Downing K C 1997 Visualizing the secondary structure of tubulin: three-dimensional map at 4Å *J. Struct. Biol.* **118** 119–27
- [79] Knoll M and Ruska E 1932 Beitrag zur geometrischen Elektronenoptik I *Ann. Phys.* **12** 607–40
- [80] Crewe A V, Eggenberger D N, Wall J and Welter L M 1968 Electron gun using a field emission source *Rev. Sci. Instrum.* **39** 576–86

-35-

- [81] Crewe A V, Wall J and Welter L M 1968 A high resolution scanning transmission electron microscope *J. Appl. Phys.* **39** 5861–8
-

-1-

B1.18 Microscopy: light

H Kiess

B1.18.1 INTRODUCTION

Light microscopy is of great importance for basic research, analysis in materials science and for the practical control of fabrication steps. When used conventionally it serves to reveal structures of objects which are otherwise invisible to the eye or magnifying glass, such as micrometre-sized structures of microelectronic devices on silicon wafers. The lateral resolution of the technique is determined by the wavelength of the light

and the objective of the microscope. However, the quality of the microscopic image is not solely determined by resolution; noise and lack of contrast may also prevent images of high quality being obtained and the theoretical resolution being reached even if the optical components are ideal. The working range of the light microscope in comparison to other microscopic techniques is depicted schematically in table B1.18.1. Clearly, the light microscope has an operating range from about half a micrometer up to millimetres, although recent developments in improving resolution allow the lower limit to be pushed below half a micrometer.

Table B1.18.1 Overview of working ranges of various microscopic techniques (in μm).

Light microscope	0.5 \leftrightarrow 1000
Scanning electron microscope	0.05 \leftrightarrow 1000
Transmission electron microscope	0.001 \leftrightarrow 10
Scanning probe microscope	0.0001 \leftrightarrow 100

Microscopes are also used as analytical tools for strain analysis in materials science, determination of refractive indices and for monitoring biological processes *in vivo* on a microscopic scale etc. In this case resolution is not necessarily the only important issue; rather it is the sensitivity allowing the physical quantity under investigation to be accurately determined.

Light microscopy allows, in comparison to other microscopic methods, quick, contact-free and non-destructive access to the structures of materials, their surfaces and to dimensions and details of objects in the lateral size range down to about 0.2 μm . A variety of microscopes with different imaging and illumination systems has been constructed and is commercially available in order to satisfy special requirements. These include stereo, darkfield, polarization, phase contrast and fluorescence microscopes.

-2-

The more recent scanning light microscopes are operated in the conventional and/or in the confocal mode using transmitted, reflected light or fluorescence from the object. Operation in the confocal mode allows samples to be optically sectioned and 3D images of objects to be produced—an important aspect for imaging thick biological samples. The breakthrough for confocal microscopes was intimately connected with the advent of computers and data processing. The conventional microscope is then replaced by a microscopic *system* comprising the microscope, the scanning, illumination and light detection systems, the data processor and computer.

This overview will first deal with the optical aspects of conventional microscopes and the various means to improve contrast. Confocal microscopy, which in the last decade has become an important tool, especially for biology, is discussed in the final section.

B1.18.2 MAGNIFICATION, RESOLUTION AND DEPTH OF FOCUS

B1.18.2.1 MAGNIFICATION

Microscopes are imaging systems and, hence, the image quality is determined by lens errors, by structures in the image plane (e.g., picture elements of CCD cameras) and by diffraction. In addition, the visibility of objects with low contrast suffers from various noise sources such as noise in the illuminating system (shot noise), scattered light and by non-uniformities in the recording media. Interest often focuses on the achievable resolution, and discussions on limits to microscopy are then restricted to those imposed by diffraction (the so-called Abbe limit), assuming implicitly that lenses are free of errors and that the visual system or the image sensors are ideal. However, even under these conditions the Abbe limit of the resolution may not be reached if the contrast is insufficient and noise is high.

Before discussing the limits imposed by diffraction and the influence of contrast and noise on resolution, it is important to recall the basic principle of the light microscope: The objective lens provides a magnified real image of the object in the focal plane of the eyepiece. This image is then focused by the eyepiece onto the retina of the eye and is seen by the observer as a virtual image at about 25 cm distance, the normal distance for distinct vision (figure B1.18.1). The object is illuminated by the light of a lamp, either from below through the stage of the object holder if the object is transparent, or from the top if the object is non-transparent and reflecting. Organic objects containing fluorescent molecules are often investigated with an illuminating light beam that causes the sample to fluoresce. The exciting light is ‘invisible’ and the object is imaged and characterized by the emitted light.

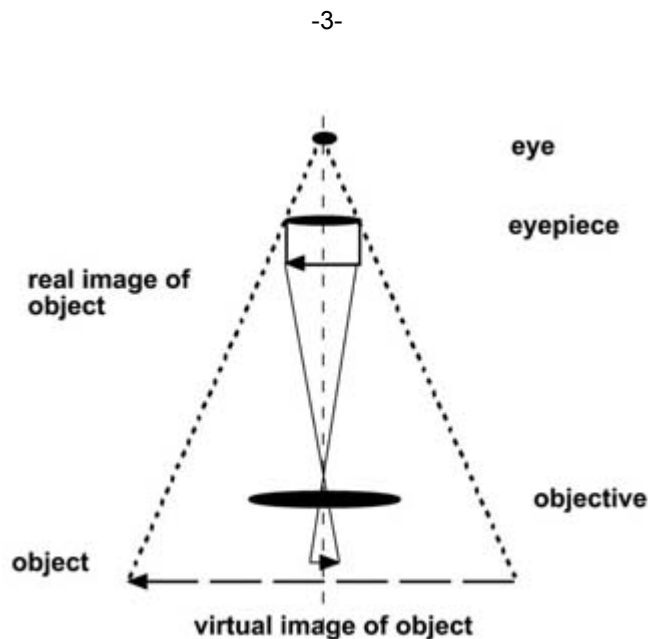


Figure B1.18.1. Light rays and imaging lenses in the microscope. The illumination system is not included. The real image is seen by the eye as a virtual image at 25 cm distance, the normal distance for distinct vision.

The magnification $M_{\text{microscope}}$ achievable by a microscope is the ratio of the scales of the virtual image and of the object. It can be easily seen from figure B1.18.1 that this ratio is given by

$$M_{\text{microscope}} = M_{\text{obj}} M_{\text{eyepiece}}$$

M_{obj} is the scale of magnification by the objective under the geometric conditions given by the microscope and $M_{\text{eyepiece}} = \ell / f_{\text{eyepiece}}$ is the magnification by the eyepiece, with the focal length f_{eyepiece} and $\ell = 25$ cm, the normal distance for distinct vision. The objectives are marked with the scale of magnification (e.g. 40:1) by the manufacturer and similarly, the eyepiece by its magnification under the given conditions (e.g. 5 \times). Multiplication of both numbers gives the magnification of the microscope. For practical reasons the magnification of the objective is not so high as to resolve all the details in the real image with the naked eye.

The magnification is rather chosen to be about $500 A_{\text{obj}}$ to $1000 A_{\text{obj}}$, where A_{obj} is the numerical aperture of the objective (see the next section) The eyepiece is then necessary to magnify the real image so that it can conveniently be inspected.

B1.18.2.2 LATERAL RESOLUTION: DIFFRACTION LIMIT

The performance of a microscope is determined by its objective. It is obvious that details of the object that are not contained in the real image (figure B1.18.1) cannot be made visible by the eyepiece or lens systems, whatever quality or magnification they may have. The performance is defined here as the size of the smallest lateral structures of the object that can be resolved and reproduced in the image. To fully assess resolution *and* image fidelity, the modulation transfer function of the imaging system has to be known (or for scanning microscopes more conveniently the point spread function). The resolution is then given by the highest lateral frequency of an object which can just be transmitted by the optical system.

-4-

Alternatively, one may consider the separation of two structure elements in the plane of the object, which are just discernible in the image [1, 2 and 3]. Since an exact correlation exists between the pattern generated by the object in the exit pupil of the objective and the image, the limit on the resolution can be estimated simply. If the diffracted beams of zeroth and \pm first order are collected by the lens, then an image of low fidelity of the structure, with the zeroth order only a grey area, is obtained. Hence, the limit to resolution is given whenever the zeroth- and first-order beams are collected (figure B1.18.2). If the diffracted light enters a medium of refractive index n , the minimal discernible separation a_{min} of two structure elements is given by $n \sin \alpha = \lambda/a_{\text{min}}$. The expression $n \sin \alpha$ can be called the numerical aperture of the diffracted beam of first order which, for microscopes, is identical to the numerical aperture A_{obj} of the objective lens. The numerical aperture of a lens is the product of the refractive index of the medium in front of the objective and of the sine of half of the angle whose vertex is located on the optical axis and being the starting point of a light cone of angle α which is just collected by the lens.

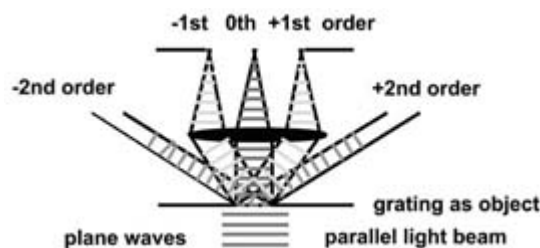


Figure B1.18.2. Diffraction figure of a grating: If only the zeroth-order beam were collected by the lens, only a bright area would be visible without any structure indicating the presence of the grating. If the zeroth- and \pm first-order beams are collected, as indicated in the figure, the grating can be observed, albeit with incomplete object fidelity.

The smallest resolvable structure is thus $a_{\text{min}} = \lambda/A_{\text{obj}}$. If, in addition, the aperture of the illumination system is taken into account, one finds:

$$a_{\text{min}} = \lambda/(A_{\text{obj}} + A_{\text{ill}}).$$

The highest resolution is obtained if $A_{\text{ill}} = A_{\text{obj}}$. In this case a_{min} equals about half the wavelength used for the illumination divided by the numerical aperture. Using blue or ultraviolet light for illumination, a_{min} can reach values of 0.2 to 0.15 μm with a numerical aperture of the microscope of about 0.9.

The diffraction limit for resolution does not imply that objects of dimensions smaller than a_{\min} are not detectable. Single light-emitting molecules or scattering centres of atomic dimensions can be observed even though their size is below the resolution. For their detectability it is required that the separation of the centres is greater than the resolution and that the emitted signal is sufficiently high to become detectable by a light-sensitive device. Microscopy, for which these assumptions are fulfilled, has sometimes been called ultramicroscopy.

-5-

B1.18.2.3 CONTRAST, NOISE AND RESOLUTION

The resolution limited by diffraction assumes that illumination and contrast of the object are optimal. Here we discuss how noise affects the discernibility of small objects and of objects of weak contrast [4]. Noise is inherent in each light source due to the statistical emission process. It is, therefore, also a fundamental property by its very nature and limits image quality and resolution, just as diffraction is also responsible for the fundamental limit. Light passing through a test element in defined time slots Δt will not contain the same number of photons in a series of successive runs. This is due to the stochastic emission process in the light source. Other sources of noise, such as inhomogeneities of recording media, will not be considered here.

It is assumed that the image can be divided up into a large number of picture elements, whose number will be of the order of 10^6 . If the contrast due to a structure in the object between two adjacent elements is smaller than the noise-induced fluctuations, the structure cannot be discerned, even if diffraction would allow this. Similarly, if the statistical excursion of the photon number in one or several of the elements is larger than or equal to the signal, then the noise fluctuations might be taken as true signals and lead to misinterpretations.

If viewed in transmission, the background brightness B_b is higher than the light B_o transmitted by an absorbing object. The contrast can then be defined as $C = (B_b - B_o)/B_o$ with $B_b \geq B_o \geq 0$ and $1 \geq C \geq 0$. It has been shown that density of photons R (photons cm^{-2}) required to detect the contrast C is

$$R = Nk^2/C^2.$$

Here, N is the density of picture elements (cm^{-2}); if N is high, the resolution is high, requiring, however, an increase of photon density over images with lower resolution. Also, low contrast requires greater photon density than high contrast in order to overcome false signals by noise fluctuations of adjacent picture elements. The factor k reflects the random character of the photons (noise) and has to be chosen so as to protect against misinterpretations caused by noisy photon flux. k depends somewhat on how well the image should be freed from noise-induced artefacts, a reasonable value being $k = 5$.

A summary of the diffraction- and noise-induced limitations of the resolution is qualitatively depicted in [figure B1.18.3](#). With noise superimposed, the rectangular structure depicted in [figure B1.18.3\(a\)](#) becomes less defined with decreasing spacing and width of the rectangles. In [figure B1.18.3\(b\)](#), an assumed modulation transfer function of an objective is shown: that is, the light intensity in the image plane as a function of spatial frequency obtained by an object which sinusoidally modulates the transmitted light intensity. At low spatial frequencies, the amplitude is independent of frequency; at higher frequencies it drops linearly with increasing frequency. The root mean square (rms) noise, due to the statistical nature of the light, increases with spatial frequency. The intersection of the rms noise with the modulation transfer function gives the frequency at which noise becomes equal to ($k = 1$) or 1/25th ($k = 5$) of the signal. At high contrast, the decrease in image amplitude is usually determined by diffraction; at lower contrast, noise is predominant.

-6-

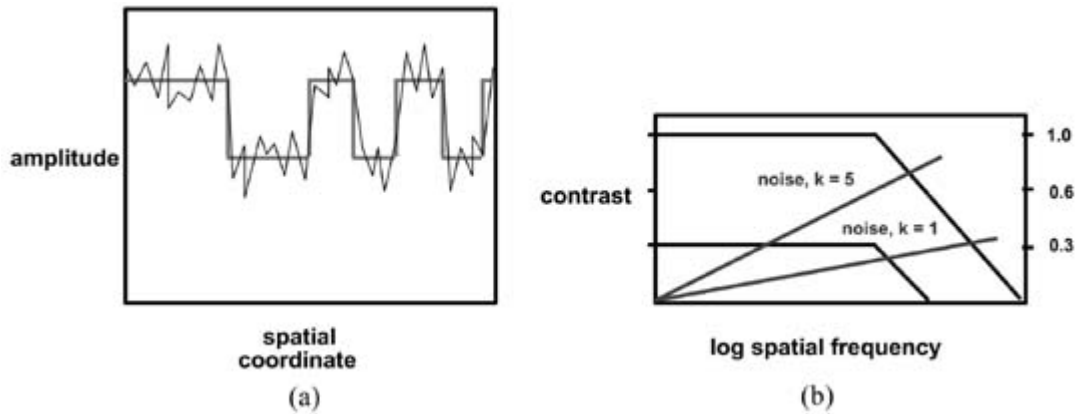


Figure B1.18.3. (a) Rectangular structure with noise superimposed: in the left half of the graph the rectangular structure is recognizable; in the right half, with the narrower spacing the rectangular structure would barely be recognizable without the guiding lines in the figure. (b) Modulation transfer function for a sinusoidal signal of constant amplitude as a function of frequency: at low frequency the amplitude of the transfer function is independent of the frequency. However, beyond a certain frequency the amplitude decreases with increasing frequency. This drop corresponds to a limitation of the resolution by diffraction. Noise increases with frequency and the crossings with the transfer function indicate where noise limits the resolution. At a contrast of 30% this cutoff frequency is exclusively determined by noise at point A; at 100% contrast the amplitude of the signal drops before the final signal is cut off at point B by noise. For $k = 5$, the noise at the crossing is 1/25th of the signal; for $k = 1$, it is equal to the signal.

Under appropriate contrast and high light intensity, the resolution of planar object structures is diffraction limited. Noise in the microscopic system may also be important and may reduce resolution, if light levels and/or the contrasts are low. This implies that the illumination of the object has to be optimal and that the contrast of rather transparent or highly reflecting objects has to be enhanced. This can be achieved by an appropriate illumination system, phase- and interference-contrast methods and/or by data processing if electronic cameras (or light sensors) and processors are available. Last but not least, for low-light images, efforts can be made to reduce the noise either by averaging the data of a multitude of images or by subtracting the noise. Clearly, if the image is inspected by the eye, the number of photons, and hence the noise, are determined by the integration time of the eye of about 1/30 s; signal/noise can then only be improved, if at all possible, by increasing the light intensity. Hence, electronic data acquisition and processing can be used advantageously to improve image quality, since integration times can significantly be extended and noise suppressed.

B1.18.2.4 DEPTH OF FOCUS

The depth of focus is defined as how far the object might be moved out of focus before the image starts to become blurred. It is determined (i) by the axial intensity distribution which an ideal object point suffers by imaging with the objective (point spread function), (ii) by geometrical optics and (iii) by the ability of the eye to adapt to different distances. In case (iii), the eye adapts and sees images in focus at various depths by successive scanning. Obviously, this mechanism is inoperative for image sensors and will not be considered here. The depth of focus caused by spreading the light intensity in axial direction (i) is given by

-7-

$$t_{\text{PSF}} = n\lambda / (A)^2$$

with n the refractive index, λ the wavelength of the light and A the numerical aperture.

The focal depth in geometrical optics is based on the argument that ‘points’ of a diameter smaller than 0.15 mm in diameter cannot be distinguished. This leads to a focal depth of

$$t_{GO} = 0.15n / (AM_{\text{microscope}}).$$

The total depth of focus is the sum of both. It increases with the wavelength of the light, depends on the numerical aperture and the magnification of the microscope. For $\lambda = 550$ nm, a refractive index of 1 and a numerical aperture of 0.9, the depth of focus is in the region of $0.7 \mu\text{m}$; with a numerical aperture of 0.4 it increases to about $5 \mu\text{m}$. High-resolution objectives exclude the observation of details in the axial direction beyond their axial resolution. This is true for conventional microscopy, but not for scanning confocal microscopy, since optical sectioning allows successive layers in the bulk to be studied. Similarly, the field of view decreases with increasing resolution of the objective in conventional microscopy, whereas it is independent of resolution in scanning microscopy.

B1.18.3 CONTRAST ENHANCEMENT

In transmission microscopy, a transparent object yields low contrast. Molecular biological samples may be dyed in order to enhance contrast. However, this is in many cases neither possible nor desirable for various reasons, meaning that the object is only barely visible in outline and with practically no contrast. Similarly, if inorganic samples are to be investigated which are composites of materials of practically equal indices of refraction, the different components can only be distinguished in the microscope with great difficulty. This is all equally true for reflected light microscopy: the visibility and resolution of microscopic images suffer, if contrast is low. In order to cope with this, different illumination techniques are applied in order to enhance the contrast.

B1.18.3.1 KÖHLER'S BRIGHT-FIELD ILLUMINATION SYSTEM

Köhler's illumination system [5], which allows the field of view to be precisely illuminated, is schematically depicted in [figure B1.18.4](#). The object is illuminated through a substage: the filament of a lamp is imaged by a collector lens into the focal plane of the condenser, where the condenser iris is located. Light from each point in the condenser iris passes through the object as a parallel beam inclined to the axis of the microscope at an angle depending on the position of the point in the iris. The parallel beams come to a focus at corresponding points in the focal plane of the objective. The collector iris allows the area illuminated in the object plane to be varied. The condenser iris is the aperture of the illumination and its opening should be adjusted to the aperture of the objective lens and contrast properties of the object: if the apertures are equal, the highest resolution is achieved; if the illumination aperture is reduced, the contrast is enhanced. In practice, the aperture of the condenser is in most cases chosen to be smaller than the aperture of the objective lens in order to improve contrast.

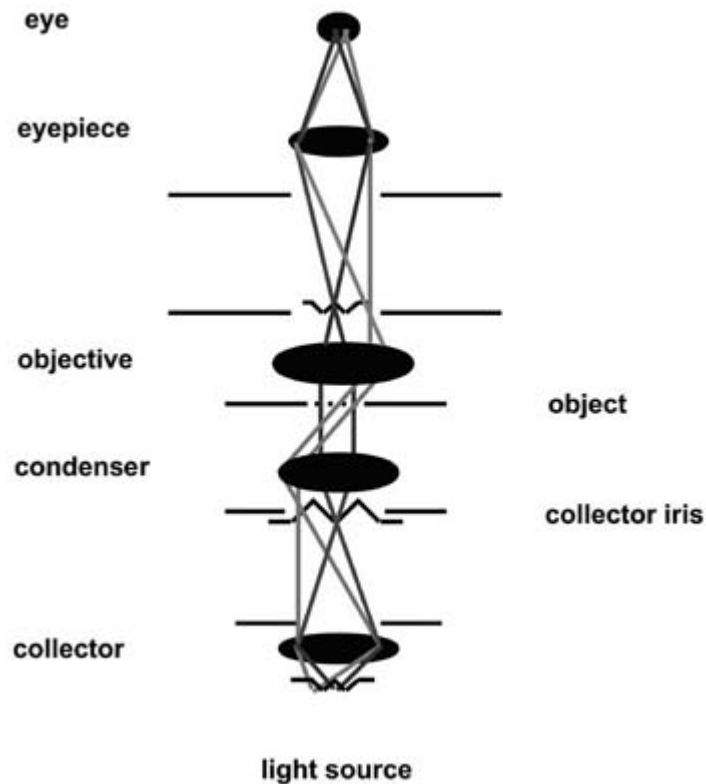


Figure B1.18.4. The most frequently used illumination system in bright-field microscopy.

B1.18.3.2 ENHANCED CONTRAST BY DARK-FIELD ILLUMINATION

Dark-field microscopy utilizes only those light beams from the condenser that have an aperture greater than that of the objective. This is in contrast to Köhler's illumination principle, where care is taken to adjust the aperture of the condenser by an iris to become equal to or smaller than that of the objective. A ring-type diaphragm is used to allow light beams to pass the condenser lens at an aperture greater than that of the objective lens. This is shown schematically in [figure B1.18.5](#). In this arrangement, no direct light beams pass through the objective but only those which are diffracted or scattered by the object. If the direct light beam is blocked out, the background appears black instead of bright, thus increasing the contrast. Special condensers have been designed for dark-field illumination. Dark-field illumination has often been used in reflection.

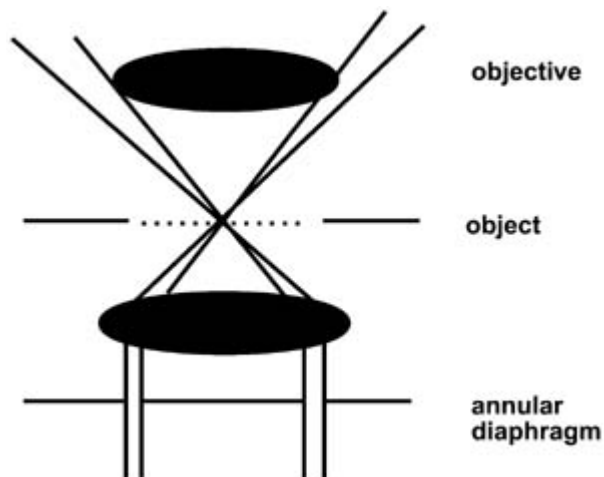


Figure B1.18.5. Dark-field illumination: the aperture of the objective is smaller than the aperture of the beams allowed by the annular diaphragm.

B1.18.3.3 ZERNIKE'S PHASE CONTRAST MICROSCOPY

Phase contrast microscopy [6, 7] is more sophisticated and universal than the dark-field method just described. In biology, in particular, microscopic objects are viewed by transmitted light and phase contrast is often used. Light passing through transparent objects has a different phase from light going through the embedding medium due to differences in the indices of refraction. The image is then a so-called phase image in contrast to an amplitude image of light absorbing objects. Since the eye and recording media in question respond to the intensity (amplitude) of the light and not to changes of the light phase, phase images are barely visible unless means are taken to modify the interference of the diffracted beams. The diffraction pattern of a phase grating is like that of an amplitude grating except that the zeroth-order beam is especially dominant in intensity. Zernike realized that modification of the zeroth-order beam will change the character of the image very effectively, by changing its phase and its intensity. For each object, depending on its character concerning the phase and amplitude, a 'Zernike diaphragm' (i.e. a diaphragm that affects the phase and the amplitude of the zeroth-order beam) can be constructed with an appropriate absorption and phase shift, which allows the weak-contrast image of the object to be transformed into an image of any desired contrast.

The principle of phase contrast microscopy is explained by [figure B1.18.6](#). The object is assumed to be a linear phase grating. The diaphragm is annular, which means that only a small fraction of the diffracted light is covered by the Zernike diaphragm, as indicated in the figure for the first-order beams. In general, the Zernike diaphragm shifts the phase of the zeroth order by $\pi/2$ with respect to the diffracted beams. Since the intensity of the diffracted beams is much lower than that of the direct beam, the intensity of the zeroth-order beam is usually attenuated by adding an absorbing film. Clearly, all these measures indicate that images of high contrast cannot be combined with high intensity using this technique; a compromise between both has to be found depending on the requirements. The image fidelity of phase contrast imaging depends, therefore, on the width and light absorption of the Zernike diaphragm, in addition to the size and optical path difference created by the object under study and, finally, on the magnification.

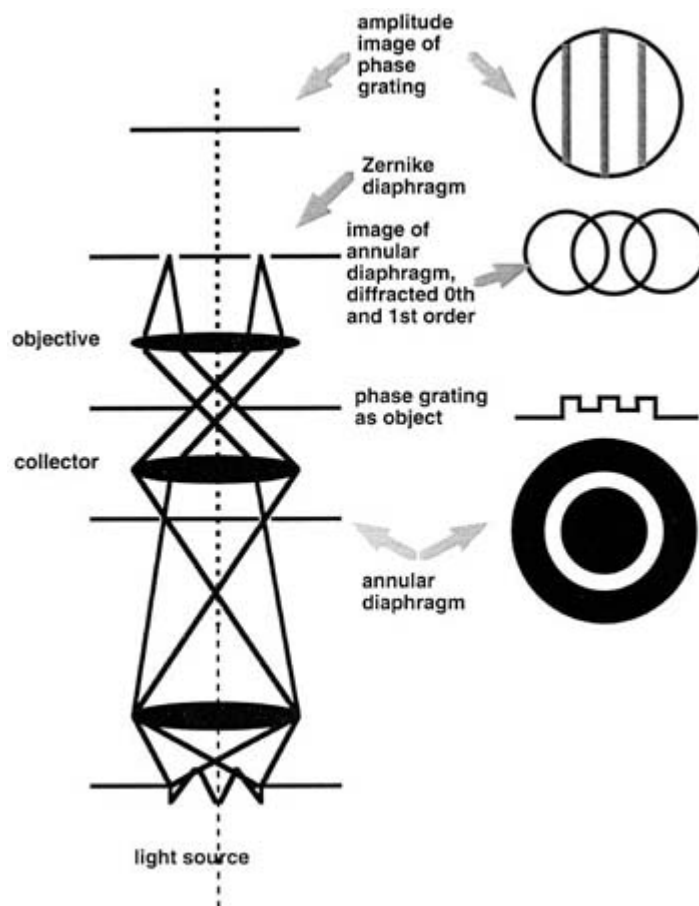


Figure B1.18.6. Schematic representation of Zernike's phase contrast method. The object is assumed to be a relief grating in a transparent material of constant index of refraction. Phase and amplitude are varied by the Zernike diaphragm, such that an amplitude image is obtained whose contrast is, in principle, adjustable.

At this point it is worth comparing the different techniques of contrast enhancements discussed so far. They represent spatial filtering techniques which mostly affect the zeroth order: dark field microscopy, which eliminates the zeroth order, the Schlieren method (not discussed here), which suppresses the zeroth order and one side band and, finally, phase contrast microscopy, where the phase of the zeroth order is shifted by $\pi/2$ and its intensity is attenuated.

B1.18.3.4 INTERFERENCE MICROSCOPY

As already discussed, transparent specimens are generally only weakly visible by their outlines and flat areas cannot be distinguished from the surroundings due to lack of contrast. In addition to the phase contrast techniques, light interference can be used to obtain contrast [8, 9].

Transparent, but optically birefringent, objects can be made visible in the polarizing microscope if the two beams generated by the object traverse about the same path and are brought to interfere. In the case of optically isotropic bodies, the illuminating light beam has to be split into two beams: one that passes through the specimen and suffers phase shifts in the specimen which depend on the thickness and refraction index, and a second beam that passes through a reference object on a separate path (see figure B1.18.7). By superposing the two beams, phase objects appear in dark-bright contrast.

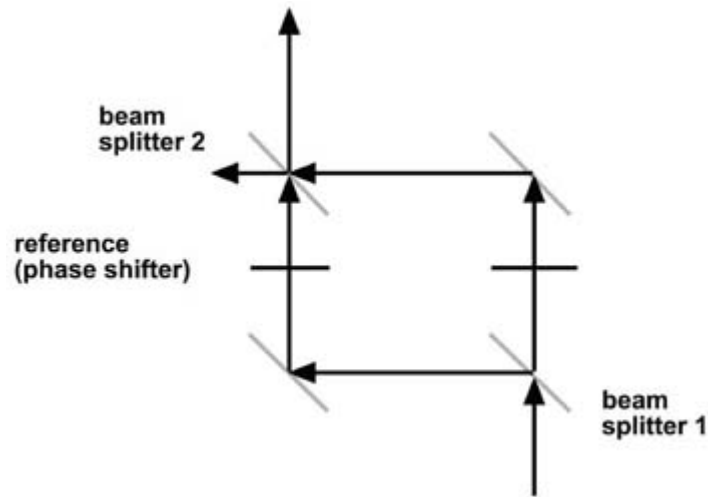


Figure B1.18.7. Principle for the realization of interference microscopy. The illuminating beam is split by beamsplitter 1 before passing the object so that the reference beam is not affected by the object. The separated beams interfere behind beamsplitter 2.

The differential interference contrast method utilizes the fact that, using a Wollaston prism, linearly polarized light can be split into two light beams of perpendicular polarization (figure B1.18.8). Since they are slightly parallel shifted, the two beams pass through the object at positions having different thickness and/or refractive index. The splitting of the beams is chosen to be sufficiently small not to affect the resolution. They are brought together again by a second Wollaston prism, and pass through the analyser. Since the beams are parallel and their waves planar in the object plane, the beams in the image plane are also parallel and the waves planar. Hence, the interference of the beams does not give rise to interference lines but to contrast, whose intensity depends on the phase difference caused by small differences of the refractive index. The image appears as a relief contrast which can be modified by changing the phase difference: for example, by moving the Wollaston prism perpendicularly to the optical axis.

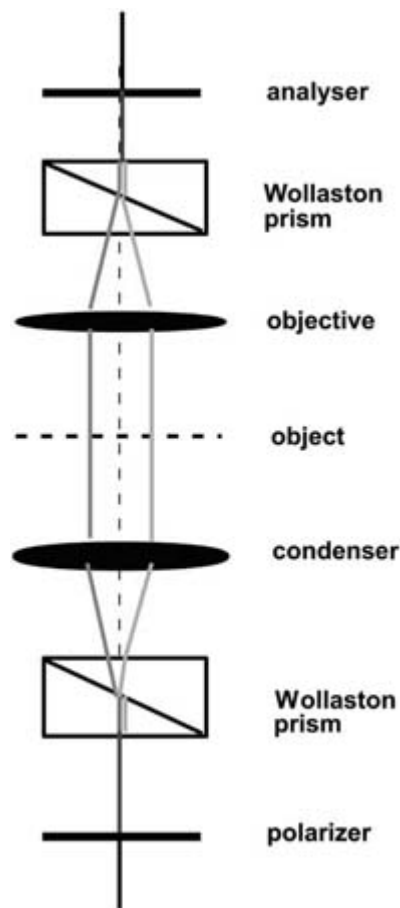


Figure B1.18.8. Differential interference contrast: the light beam is split into two beams by a Wollaston prism. The two beams pass the object at closely spaced positions and give, after interference, a contrast due to the phase difference.

The interpretation of such images requires some care, because the appearance of a relief structure may be misleading; it does not necessarily mean that the surface or thickness of the object is relief-like. Obviously, such a relief may also appear if samples are homogeneously thick, but composed of elements of different indices of refraction. Also, edges of the object may be missed if they are inappropriately oriented with respect to the polarization direction of the beams.

Interference microscopy is also possible in reflection. The surface structure of highly reflecting objects such as metals or metallized samples is frequently investigated in this way. Using multiple-beam interference [10], surface elevations as small as a few nanometres in height or depth can be measured. This is due to the fact that the interference lines become very sharp if the monochromaticity of the light and the number of interfering beams are high.

B1.18.3.5 FLUORESCENCE MICROSCOPY

Fluorescence microscopy has been a very popular method of investigating biological specimens and obtaining contrast in otherwise transparent organic objects. The samples are stained with fluorescent dyes and illuminated with light capable of exciting the dye to fluoresce. The wavelength of the emitted light is Stokes-shifted to wavelengths longer than that of the primary beam. Since the quantum efficiency (the ratio of the numbers of emitted to exciting photons) of the dyes is often low and since the light is emitted in all directions, the image is of low intensity. Nevertheless, this technique allows images of high contrast and of high signal-

to-noise ratio to be obtained. The principle of fluorescence microscopy is illustrated in figure B1.18.9 for the epifluorescence microscope. The primary excitation does not in principle directly enter the detector and thus provides the desired contrast between stained and unstained areas, which appear completely dark. It is obvious from the very nature of the preparation technique that, in addition to morphological structures, chemical and physicochemical features of the sample can be revealed if the dyes adsorb only at special chemical sites.

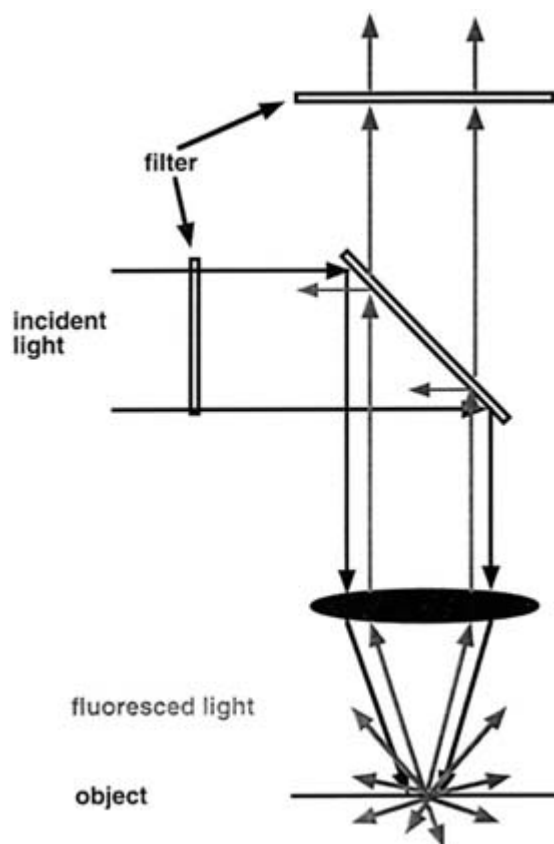


Figure B1.18.9. Epifluorescence microscope: the object is excited from the top and the fluorescent light is emitted in all directions, as indicated by the multitude of arrows in the object plane. The fluorescent light within the aperture of the objective gives rise to the image, showing that much of the fluorescent light is lost for imaging.

B1.18.4 SCANNING MICROSCOPY

In scanning microscopy, the object is successively scanned by a light spot, in contrast to conventional microscopy in which the entire object field is processed simultaneously. Thus, scanning represents a serial (and conventional microscopy a parallel) processing system. The requirements for the optical lenses are relaxed for the scanning microscope, because the whole field of view is no longer imaged at once, but the price that is paid is the need for reconstruction of the image from a set of data and the required precision for the scanning.

A point light source is imaged onto the specimen by the objective and the transmitted light collected by the collector lens and detected by a broad-area detector; in the case of reflection microscopy, the objective lens also serves simultaneously as a collector (see figure B1.18.10). The resolution is solely determined by the objective lens, because the collector has no imaging function and only collects the transmitted light. The

scanning is assumed in [figure B1.18.10](#) to be based on the mechanical movement of the sample through the focal point of the objective. In this case, off-axis aberrations of the objective are avoided, the area to be imaged is not limited by the field of view of the objective and the image properties are identical and only determined by the specimen. The drawback of stage movement is the lower speed compared with beam scanning and the high mechanical precision required for the stage. Beam scanning allows the image to be reconstructed from the serially available light intensity data of the spots in real time [11, 12]. If a framestore is available, the image can be taken, stored, processed if desired and displayed. Processing of the electrical signal offers advantages. There is, for example, no need to increase contrast by stopping down the collector lens or by dark-field techniques which, in contrast to electronic processing, modify the resolution of the image.

Scanning gives many degrees of freedom to the design of the optical system and the confocal arrangement is one of the most prominent, having revolutionized the method of microscopic studies, in particular of biological material. Since confocal microscopy has in recent years proved to be of great importance, it is discussed in some detail here.

-15-

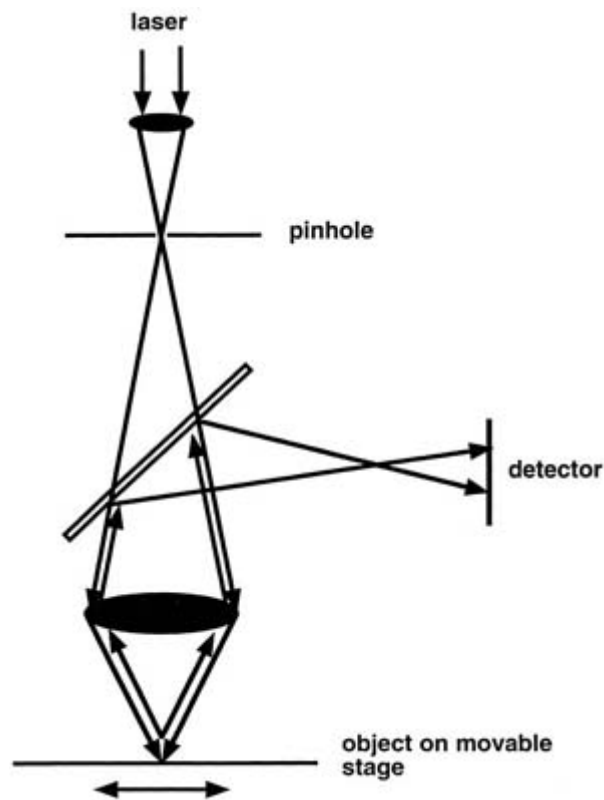


Figure B1.18.10. Scanning microscope in reflection: the laser beam is focused on a spot on the object. The reflected light is collected and received by a broad-area sensor. By moving the stage, the object can be scanned point by point and the corresponding reflection data used to construct the image. Instead of moving the stage, the illuminating laser beam can be used for scanning.

B1.18.5 CONFOCAL SCANNING MICROSCOPY

B1.18.5.1 PRINCIPLE AND ADVANTAGES OF CONFOCAL MICROSCOPY

The progress that has been achieved by confocal microscopy [13, 14, 15 and 16] is due to the rejection of

object structures outside the focal point, rejection of scattered light and slightly improved resolution. These improvements are obtained by positioning pointlike diaphragms in optically conjugate positions (see [figure B1.18.11](#)). The rejection of structures outside the focal point allows an object to be optically sectioned and not only images of the surface are obtained by scanning but also of sections deep in a sample, so that three-dimensional microscopic images can be prepared as well as images of sections parallel to the optical axis. Therefore, internal structures in biological specimens can be made visible on a microscopic scale without major interference with the biological material by preparational procedures (fixation, dehydration etc) and without going through the painstaking procedure of mechanical sectioning. In addition, time-dependent studies of microscopic processes are possible. Obviously, there is a price to be paid:

-16-

Confocal microscopy requires serial data acquisition and processing and hence comprises a complete system whose cost exceeds that of a conventional microscope.

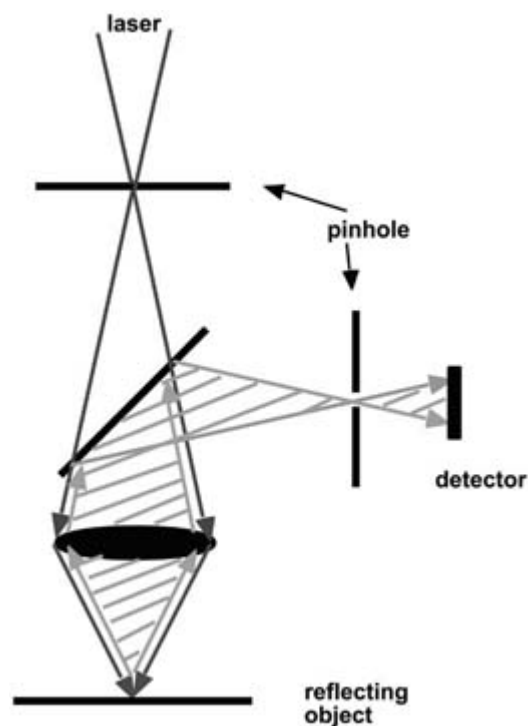


Figure B1.18.11. Confocal scanning microscope in reflection: the pinhole in front of the detector is in a conjugate position to the illumination pinhole. This arrangement allows the object to be optically sectioned. The lens is used to focus the light beam onto the sample and onto the pinhole. Thus, the resulting point spread function is sharpened and the resolution increased.

Figure B1.18.11 shows the basic arrangement of a confocal instrument. The important points are more easily presented for the reflection microscope, although everything also applies to transmission if modified appropriately. The broad-area detector is replaced by a point detector implemented by a pinhole placed in front of the detector at the conjugate position to the pinhole on the illumination side. This arrangement ensures that only light from the small illuminated volume is detected and light that stems from outside the focal plane is strongly reduced in intensity. This is illustrated in [figure B1.18.12](#) where a reflecting object is assumed to be below the focal plane: only a small fraction of the reflected light reaches the detector, since it is shielded by the pinhole. The intensity drops below detection threshold and no image can be formed.

-17-

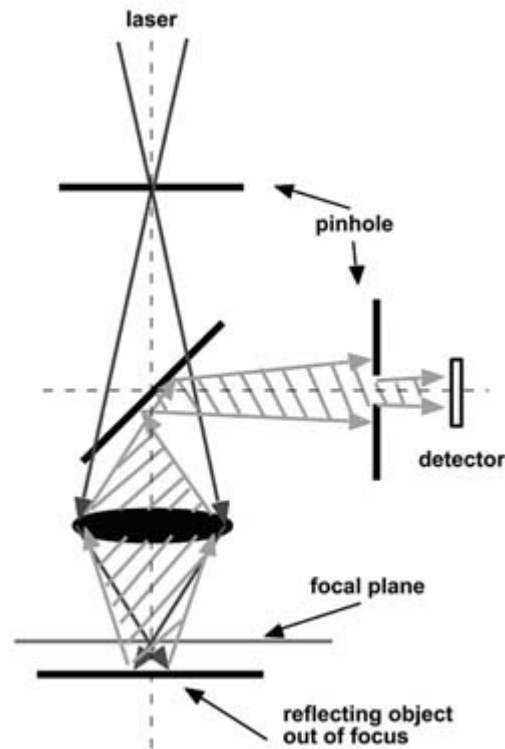


Figure B1.18.12. Illustration that only light reflected from object points in the focal plane contributes to the image. If the light is reflected from areas below the focal plane, only a small fraction can pass through the pinhole so that light from those areas does not contribute to the image. The pinhole in front of the detector is exaggerated in size for the sake of presentation.

B1.18.5.2 OPTICAL SECTIONING, SMALLEST SLICE THICKNESS AND AXIAL RESOLUTION

The fact that only points in the focal plane contribute to the image, whereas points above or below do not, allows optical sectioning. Thus, the object can be imaged layer by layer by moving them successively into the focal plane. For applications, it is important how thin a slice can be made by optical sectioning. A point-like object imaged by a microscope has a finite volume [13] which is sometimes called voxel, in analogy to pixel in two dimensions. Its extension in the axial direction determines the resolution in this direction (z) and the smallest thickness of a layer that can be obtained by optical sectioning. The intensity variation of the image of a point along the optical axis for the confocal arrangement is given by

$$I(u) = \{\sin(u/4)/u/4\}^4$$

with $u = (8\pi/\lambda) z \sin^2(\alpha/2)$. λ is the wavelength of the light, and $n \sin \alpha$ is the numerical aperture of the lens. The function $I(u)$ is zero at $u = \pi, 2\pi \dots$. If we take the spread of the function $I(u)$ between $u = \pm \pi$ as the smallest slice thickness t , one obtains

$$t = \lambda/(4 \sin^2 \alpha/2).$$

The slice thickness is proportional to the wavelength of the light and a function of the aperture angle. For $\lambda = 0.5 \mu\text{m}$, the slice thickness is about $0.25 \mu\text{m}$ for $\alpha = \pi/2$. Obviously, the point spread function serves also to

determine the smallest separation that two points in the axial direction may have in order to be resolved. If the Rayleigh criterion is applied—intensity between the image points to be half of the intensity at maximum—then the resolution is in the range of 0.15–0.2 μm .

B1.18.5.3 RANGE OF DEPTH FOR OPTICAL SECTIONING

The greatest depth at which a specimen can be optically sectioned is also of interest. This depth is limited by the working distance of the objective, which is usually smaller for objectives with greater numerical aperture. However, the depth imposed by the working distance of the objective is rarely reached, since other mechanisms provide constraints as well. These are light scattering and partial absorption of the exciting and emitted light, in the case of fluorescence microscopy. The exciting beam is partially absorbed by fluorophores until it reaches the focused volume. Hence, less light is emitted from a focused volume that is deep in the bulk of a sample. Thus, the intensity of the light reaching the detector decreases with increasing depth, so that for image formation laser power and/or integration time would have to be increased. Though technically possible, both cannot be increased beyond thresholds at which the samples, especially biological materials, are damaged.

B1.18.5.4 LATERAL RESOLUTION

The extension of the voxel in a radial direction gives information on the lateral resolution. Since the lateral resolution has so far not been discussed in terms of the point spread function for the conventional microscope, it will be dealt with here for both conventional and confocal arrangements [13]. The radial intensity distribution in the focal plane (perpendicular to the optical axis) in the case of a conventional microscope is given by

$$I_m(v) = (2J_1(v)/v)^2$$

with $v = (2\pi/\lambda)rn \sin \alpha$, λ is the wavelength of the light, r is the radial coordinate, $n \sin \alpha$ is the numerical aperture, and $J_1(v)$ is the first-order Bessel function of the first kind. Zero intensity is at $v = 1.22 \pi, 2.23\pi, 3.42\pi \dots$

For the confocal arrangement in transmission, the objective and the collector are used for imaging; in reflection the objective is used twice. Therefore, the radial intensity distribution in the image is the square of that of the conventional microscope:

$$I_{\text{conf}}(v) = (2J_1(v)/v)^4.$$

$I_{\text{conf}}(v)$ has the same zero points as $I_m(v)$. However, in the confocal case the function is sharpened and the sidelobes are suppressed. The light intensity distributions for the conventional and the confocal case are depicted in [figure B1.18.13](#). If the Rayleigh criterion for the definition of resolution is applied, one finds that the lateral resolution in

the confocal case is improved in comparison with conventional microscopy: obviously, the sharpened function in the confocal case allows two closely spaced points at smaller separation to be distinguished.

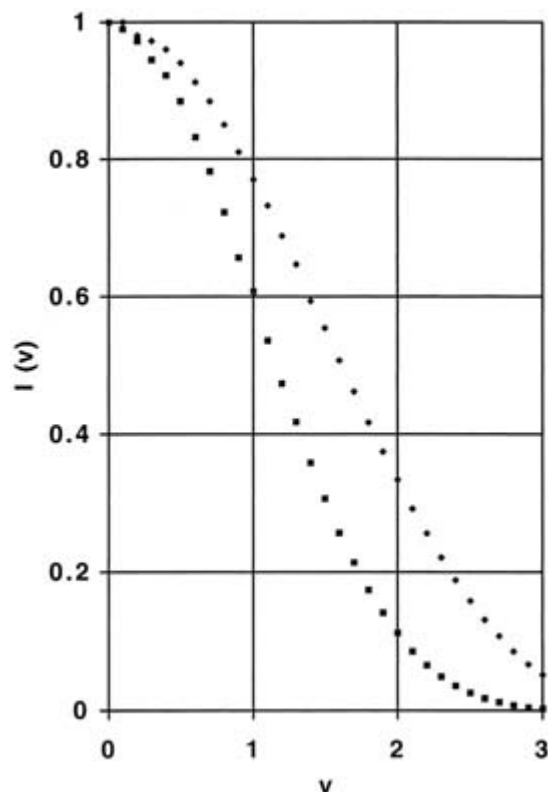


Figure B1.18.13. The point spread function in the confocal arrangement (full square) is sharpened in comparison with the conventional arrangement (full diamond). Therefore, the resolution is improved.

B1.18.5.5 CONTRAST ENHANCEMENT AND PRACTICAL LIMITS TO CONFOCAL ONE-PHOTON-EXCITATION FLUORESCENCE MICROSCOPY

The methods to improve contrast described for conventional microscopy can also be applied to confocal microscopy [17]. However, because the images are obtained by scanning and data processing, the tools of image manipulation are also advantageously utilized to improve image quality. Nevertheless, all these methods have their limitations, as will be explained in the following example. Biological studies are often made with fluorescence. Usually the fluorophore is excited by one photon from the ground state to the excited state; the ratio of the number of photons emitted by the fluorophore to the number of exciting photons is, as a rule, significantly below one. Therefore, the number of photons collected per voxel is low, depending on the density of fluorophores, on the exciting light intensity and on the scan rate. The density of fluorophores is, in general, determined by the requirements of the experiment and cannot be significantly varied. In order to increase the signal, the scan rate would have to be lowered, and the number of scans and the exciting light intensity increased. However, extended exposure of the dyes leads to bleaching in the whole cone of illumination and hence to the number of layers to be sectioned. The number of layers is even more reduced if, in addition to bleaching the fluorophore, the excitation produces toxic products that modify or destroy the properties of

living cells or tissues. Lasers could, in principle, supply higher light intensities, but saturation of emission of the dyes additionally limits the applicable power. In these circumstances, real improvement can only be reached if the voxel at the focal point could exclusively be excited by the incident light.

B1.18.5.6 CONFOCAL MICROSCOPY WITH MULTIPHOTON-EXCITATION FLUORESCENCE

Usually a fluorophore is excited from its ground to its first excited state by a photon of an energy which corresponds to the energy difference between the two states. Photons of smaller energy are generally not absorbed. However, if their energy amounts to one-half or one-third (etc) of the energy difference, a small probability for simultaneous absorption of two or three (etc) photons exists since the energy condition for absorption is fulfilled. However, due to this small probability, the photon density has to be sufficiently high if two-, three- or n -photon absorption is to be observed [18]. In general, these densities can only be achieved by lasers of the corresponding power and with appropriate pulse width, since absorption by multiphoton processes increases with the n th power of the photon density (figure B1.18.14).

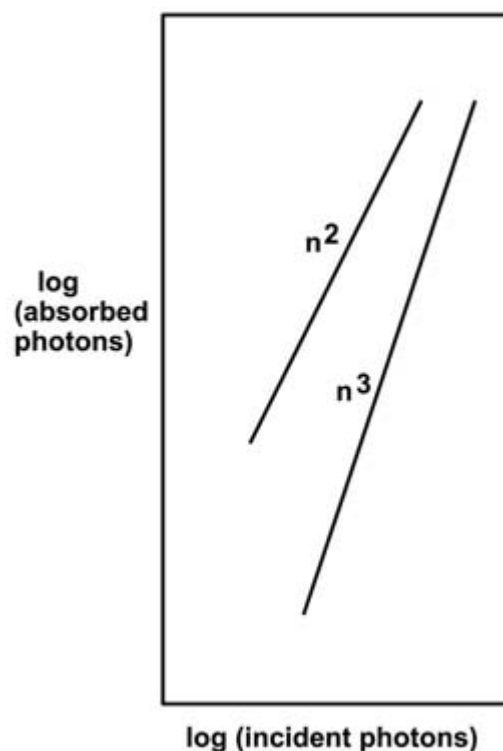


Figure B1.18.14. Schematic representation of the increase of absorption with photon density for two- and three-photon absorption.

One-photon excitation has limitations due to the unwanted out-of-focus fluorophore absorption and bleaching, and light scattering. These drawbacks can be circumvented if multiphoton excitation of the fluorophore is used. Since it increases with the n th power of the photon density, significant absorption of the exciting light will only occur at the focal point of the objective where the required high photon density for absorption is reached. Consequently, only

-21-

fluorescent light will be emitted from a volume element whose size will be determined by the intensity and power law dependence of the exciting radiation. Though confocal arrangement is not needed, it was shown that, in the confocal arrangement, the effective point spread function is less extended and, hence, the resolution improved [19, 20].

Thus, multiphoton excitation eliminates unwanted out-of-focus excitation, unnecessary phototoxicity and bleaching. However, efficient power sources are required and, since the efficiency of multiphoton excitation is usually low, the times needed to generate images are increased.

MICROSCOPY?

As light microscopy has many advantages over other microscopic techniques, the desire is to overcome the limit due to the extension of the point spread function or to reduce the emitting volume by multiphoton excitation. One proposal was made on the basis of fluorescence microscopy [21]. As discussed, in fluorescence microscopy, molecules are excited to emit light which is then used to form the microscopic image. If the exciting light is imaged onto a small volume of the sample, light emitted from this volume determines the spatial resolution (i.e. both in depth and the lateral direction). If the light-emitting volume can be reduced, resolution will be improved. This is achievable, in principle, by stimulated emission: if the stimulated emission rate is higher than the fluorescence decay and slower than the decay rate of intrastate vibrational relaxation, the emitting volume in the focal region shrinks. Estimates predict a resolution of 0.01–0.02 μm for continuous illumination of 1 mW and picosecond excitations of 10 MW cm^{-2} at a rate of 200 kHz. If this idea can be reduced into practice, the diffraction limit would be overcome. The high resolution combined with the advantages of light microscopy over other microscopic methods would indeed represent a major breakthrough in this field.

REFERENCES

- [1] Tappert J 1957 Bildentstehung im Mikroskop *Wissenschaft und Fortschritt* **7** 361
 - [2] Michel K 1981 *Die Grundzüge der Theorie des Mikroskops* 3rd edn (Stuttgart: Wissenschaftliche)
 - [3] Pluta M 1988 *Advanced Light Microscopy, vol 1: Principles and Basic Properties* (Amsterdam: Elsevier)
 - [4] Rose A 1973 *Vision, Human and Electronic* (New York: Plenum)
 - [5] Köhler A 1893 Ein neues Beleuchtungsverfahren für mikrophotographische Zwecke *Z. Wissensch. Mikr.* **10** 433
 - [6] Zernike F 1935 Das Phasenkontrastverfahren bei der mikroskopischen Beobachtung *Z. Phys.* **36** 848
 - [7] Zernike F 1942 Phase contrast, a new method for the microscopic observation of transparent objects I *Physica* **9** 686
Zernike F 1942 Phase contrast, a new method for the microscopic observation of transparent objects II *Physica* **9** 974
 - [8] Pluta M 1989 *Advanced Light Microscopy, vol 2: Advanced Methods* (Amsterdam: Elsevier)
 - [9] Beyer H 1974 *Interferenzmikroskopie* (Leipzig: Academic)
 - [10] Tolansky S 1960 *Surface Microtopography* (New York: Interscience)
-

- [11] Shaw S L, Salmon E D and Quatrano R S 1995 Digital photography for the light microscope: results with a gated, video-rate CCD camera and NIH-image software *BioTechniques* **19** 946–55
- [12] Kino G S and Xiao G Q 1990 Real time scanning optical microscope *Confocal Microscopy* ed T Wilson (New York: Academic)
- [13] Wilson T (ed) 1990 Confocal microscopy *Confocal Microscopy* (New York: Academic)
- [14] Hell S W and Stelzer E H K 1992 Fundamental improvement of resolution with a four Pi-confocal

fluorescence microscope using two-photon excitation *Opt. Commun.* **93** 277–82

- [15] Lindek St, Cremer Chr and Stelzer E H K 1996 Confocal theta fluorescence microscopy using two-photon absorption and annular apertures *Optik* **102** 131–4
- [16] van Oijen A M, Kohler J, Schmidt J, Muller M and Brakenhoff G J 1998 3-Dimensional super-resolution by spectrally selective imaging *Chem. Phys. Lett.* **292** 183–7
- [17] Török P, Sheppard C J R and Laczik Z 1996 Dark field and differential phase contrast imaging modes in confocal microscopy using a half aperture stop *Optik* **103** 101–6
- [18] Wokosin D L, Centonze V, White J G, Armstrong D, Robertson G and Ferguson A I 1996 All-solid-state ultrafast lasers facilitate multiphoton excitation fluorescence imaging *IEEE J. Sel. Top. Quantum Electron.* **2** 1051–65
- [19] Schrader M, Bahlmann K and Hell S W 1997 Three-photon-excitation microscopy: Theory, experiment, and applications *Optik* **104** 116–24
- [20] Sako Y, Sekihata A, Yanagisawa Y, Yamamoto M, Shimada Y, Ozaki K and Kusumi A 1997 Comparison of two-photon excitation laser scanning microscopy with UV-confocal laser scanning microscopy in three-dimensional calcium imaging using the fluorescence indicator Indo-1 *J. Microsc.* **185** 9–20
- [21] Hell S W and Kroug M 1995 Ground-state-depletion fluorescence microscopy: a concept for breaking the diffraction resolution limit *Appl. Phys. B* **60** 495–7

FURTHER READING

Books

v Amelinck S, van Dyck D, van Landuyt J and van Trendelo G (eds) 1996 *Handbook of Microscopy, Application in Materials Science, Solid State Physics and Chemistry* 3 vols (Weinheim: VCH)

Pluta M 1988 *Advanced Microscopy* 3 vols (Amsterdam: Elsevier)

de Hoff R and Rhines F N 1991 *Quantitative Microscopy* (Lake Grove: Tech. Books)

Beyer H (ed) 1997 *Handbuch der Mikroskopie* (VEB-Verlag Technik)

Robenek H (ed) 1995 *Mikroskopie in Forschung und Technik* (GIT-Verlag)

Herman B and Jacobsen K 1990 *Optical Microscopy for Biology* (Wiley)

Brabury S and Everett B 1996 *Contrast Techniques in Light Microscopy, Microscopy Handbooks* 34 (Oxford: BIOS Scientific Publishers)

v Kriete (ed) 1992 *Visualization in Biomedical Microscopy, 3-d Imaging and Computer Visualization* (Weinheim: VCH)

Wilson T (ed) 1996 *Confocal Microscopy* (New York: Academic)

Cork T and Kino G S 1996 *Confocal Scanning Optical Microscopy and Related Imaging Systems* (New York: Academic)

Gu Min 1996 *Principles of Three Dimensional Imaging in Confocal Microscopes* (Singapore: World Scientific)

Reviews

- Sheppard C J R 1987 Scanning optical microscopy *Adv. Opt. Electron Microscopy* **10** 1-98
- Cooke P M 1996 Chemical microscopy *Anal. Chem.* **68** 333-78
- Kapitza H G 1996 Confocal laser scanning microscopy for optical measurement of the microstructure of surfaces and layers *Tech. Mess.* **63** 136-41
- Schroth D 1997 The confocal laser scanning microscopy. A new tool in materials testing *Materialpruefung* **39** 264
- Chestnut M H 1997 Confocal microscopy of colloids *Curr. Opin. Colloid Interface Sci.* **2** 158-61
- van Blaaderen A 1997 Quantitative real-space analysis of colloidal structures and dynamics with confocal scanning light microscopy *Prog. Colloid Polym. Sci.* **104** 59-65
- Ribbe A E 1997 Laser scanning confocal microscopy in polymer science *Trends Polym. Sci.* **5** 333-7
- Oliveira M J and Hemsley D A 1996 Optical microscopy of polymers *Sonderb. Prakt. Metallogr.* **27** 13-22
- Nie Sh and Zare R N 1997 Optical detection of single molecules *Ann. Rev. Biophys. Biomol. Struct.* **26** 567-96
- Masters B R 1994 Confocal redox imaging of cells *Adv. Mol. Cell Biol.* **8** 1-19
- Ojcius D M, Niedergang F, Subtil A, Hellio R and Dautry-Varsat A 1996 Immunology and the confocal microscope *Res. Immunol.* **147** 175-88
- Lemasters J J 1996 Confocal microscopy of single living cells *Chem. Anal., NY* **137** 157-77
- Schrof W, Klingler J, Heckmann W and Horn D 1998 Confocal fluorescence and Raman microscopy in industrial research *Colloid Polym. Sci.* **276** 577-88
- Sabri S, Richelme F, Pierres A, Benoliel A M and Bongrand P 1997 Interest of image processing in cell biology and immunology *J. Immunol. Methods* **208** 1-27
- van Der Oord C J R, Jones G R, Shaw D A, Munro I H, Levine Y K and Gerritsen H C 1996 High-resolution confocal microscopy using synchrotron radiation *J. Microsc.* **182** 217-24

Applications

- Booker Gr, Laczik Z and Toeroek P 1995 *Applications of Scanning Infra-red Microscopy to Bulk Semiconductors (Inst. Phys. Conf. Ser., 146)* pp 681-92
- Bhawalkar J D, Swiatkiewicz J, Pan S J, Samarabandu J K, Liou W S, He G S, Berezney R, Cheng P C and Prasad P N 1996 Three-dimensional laser scanning two-photon fluorescence confocal microscopy of polymer materials using a new, efficient upconverting fluorophore *Scanning* **18** 562-6
- Ling X, Pritzker M D, Byerley J J and Burns C M 1998 Confocal scanning laser microscopy of polymer coatings *J. Appl. Polym. Sci.* **67** 149-58
- Carlsson K and Liljeborg A 1997 Confocal fluorescence microscopy using spectral and lifetime information to simultaneously record four fluorophores with high channel separation *J. Microsc.* **185** 37-46
- Wokosin D L and White J G 1997 Optimization of the design of a multiple-photon excitation laser scanning fluorescence imaging system *Proc. SPIE* **2984** 25-9
- Fleury L, Gruber A, Draebenstedt A, Wrachtrup J and von Borczyskowski C 1997 Low-temperature confocal microscopy on individual molecules near a surface *J. Phys. Chem. B* **101** 7933-8

Peachey L D, Ishikawa H and Murakami T 1996 Correlated confocal and intermediate voltage electron microscopy imaging of the same cells using sequential fluorescence labeling fixation and critical point dehydration *Scanning Microsc. (Suppl)* **10** 237–47

Wolleschensky R, Feurer T, Sauerbrey R and Simon U 1998 Characterization and optimization of a laser-scanning microscope in the femtosecond regime *Appl. Phys. B* **67** 87–94

Llorca-Isern N and Espanol M 1997 Advanced microscopic techniques for surface characterization *Surf. Modif. Technol. XI (Proc. 11th Int. Conf.)* pp 722–35

Leonas K K 1998 Confocal scanning laser microscopy: a method to evaluate textile structures *Am. Dyest. Rep.* **87** 15–18

Wilson K R *et al* 1998 New ways to observe and control dynamics *Proc. SPIE* **3273** 214–18

Kim Ki H, So P T C, Kochevar I E, Masters B R and Gratton E 1998 Two-photon fluorescence and confocal reflected light imaging of thick tissue structures *Proc. SPIE* **3260** 46–57

-1-

B1.19 Scanning probe microscopies

Nicholas D Spencer and Suzanne P Jarvis

B1.19.1 INTRODUCTION

The development of the scanning tunnelling microscope (STM) [1] was a revelation to the scientific community, enabling surface atomic features to be imaged in air with remarkably simple apparatus. The STM earned Binnig and Rohrer the Nobel prize for physics in 1986, and set the stage for a series of scanning probe microscopies (SPMs) based on a host of different physical principles, many of the techniques displaying nanometre resolution or better.

The methods have in turn launched the new fields of nanoscience and nanotechnology, in which the manipulation and characterization of nanometre-scale structures play a crucial role. STM and related methods have also been applied with considerable success in established areas, such as tribology [2], catalysis [3], cell biology [4] and protein chemistry [4], extending our knowledge of these fields into the nanometre world; they have, in addition, become a mainstay of surface analytical laboratories, in the worlds of both academia and industry.

Central to all SPMs (or ‘local probe methods’, or ‘local proximal probes’ as they are sometimes called) is the presence of a tip or sensor, typically of less than 100 nm radius, that is rastered in close proximity to—or in ‘contact’ with—the sample’s surface. This set-up enables a particular physical property to be measured and imaged over the scanned area. Crucial to the development of this family of techniques were both the ready availability of piezoelements, with which the probe can be rastered with subnanometre precision, and the highly developed computers and stable electronics of the 1980s, without which the operation of SPMs as we know them would not have been possible.

A number of excellent books have been written on SPMs in general. These include the collections edited by Wiesendanger and Güntherodt [5] and Bonnell [6] as well as the monographs by Wiesendanger [7], DiNardo [8] and Colton [9].

B1.19.2 SCANNING TUNNELLING MICROSCOPY

B1.19.2.1 PRINCIPLES AND INSTRUMENTATION

Tunnelling is a phenomenon that involves particles moving from one state to another through an energy barrier. It occurs as a consequence of the quantum mechanical nature of particles such as electrons and has no explanation in classical physical terms. Tunnelling has been experimentally observed in many physical systems, including both semiconductors [10] and superconductors [11].

In STM, a sharp metal tip [12] is brought within less than a nanometre of a conducting sample surface, using a piezoelectric drive (figure B1.19.1). At these separations, there is overlap of the tip and sample wavefunctions at the

-2-

gap, resulting in a tunnelling current of the order of nanoamps when a bias voltage ($\pm 10^{-3}$ –4 V) is applied to the tip [13]. The electrons flow from the occupied states of the tip to the unoccupied states in the sample, or *vice versa*, depending on the sign of the tip bias. The current is exponentially dependent on the tip–sample distance [14],

$$I = C\rho_t\rho_s e^{-s\sqrt{\phi}} \quad (\text{B1.19.1})$$

where s is the sample–tip separation, ϕ is a parameter related to the barrier between the sample and the tip, ρ_t is the electron density of the tip, ρ_s is the electron density of the sample and C , a constant, is a linear function of voltage. The exponential dependence on distance has several very important consequences. Firstly, it enables the local tip–sample spacing to be controlled very precisely ($< 10^{-2}$ Å) by means of a feedback loop connected to the z -piezo, using the tunnelling current as a control parameter. Secondly, it means that despite the fact that the tip may be many tens of nanometres in radius, the effective radius—through which most of the tunnelling takes place—is of atomic dimensions, yielding subnanometre spatial resolution. This tip may be rastered over an area that can range from hundredths of square nanometres to hundreds of square microns, and the surface topography—or more specifically the spatial distribution of particular electronic states—may thereby be imaged. Imaging (which may be done in air, in vacuum, or even under liquids) may be achieved either by monitoring the tunnelling current, in order to maintain a constant tip–sample separation and displaying the z -voltages as a function of x and y position, or by simply rastering the tip above the surface at a constant height, and plotting the tunnelling current on the z -axis. The former is known as *constant-current mode*, the latter as *constant-height mode* (figure B1.19.2). While constant-current mode is more stable for relatively rough surfaces, it is also somewhat slower than constant-height mode, because of its reliance on the feedback system, which sets a limit on the maximum scan speed.

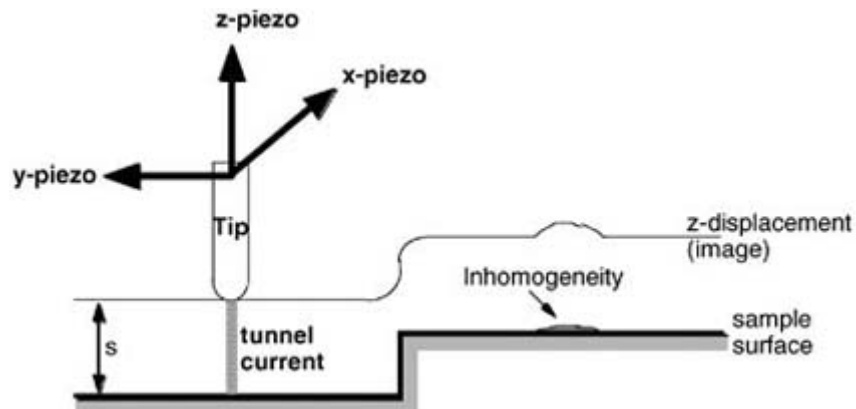


Figure B1.19.1. Principle of operation of a scanning tunnelling microscope. The x - and y -piezodrives scan the tip across the surface. In one possible mode of operation, the current from the tip is fed into a feedback loop that controls the voltage to the z -piezo, to maintain constant current. The line labelled z -displacement shows the tip reacting both to morphological and chemical (i.e. electronic) inhomogeneities. (Taken from [213].)

-3-

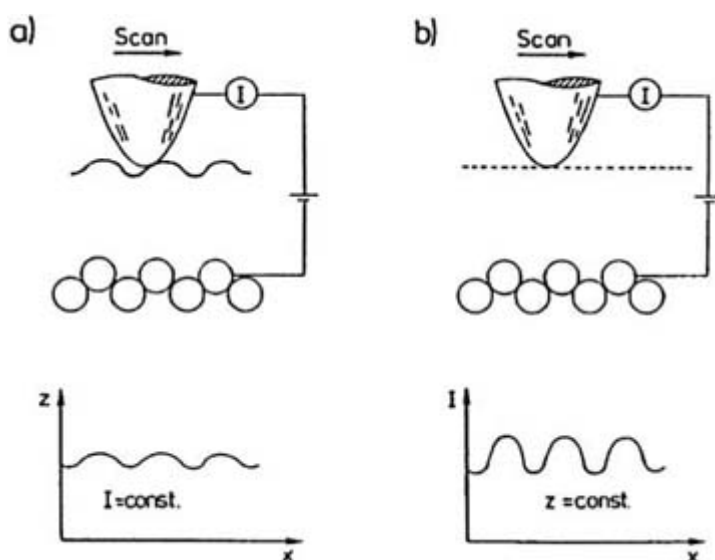


Figure B1.19.2. The two modes of operation for scanning tunnelling microscopes: (a) constant current and (b) constant height. (Taken from [214], figure 1.)

The image obtained in a STM experiment is conventionally displayed on the computer screen as grey scales or false colour, with the lightest shades corresponding to peaks (or highest currents) and darkest shades corresponding to valleys (or lowest currents). With such graphic methods of data display, it is particularly tempting to interpret atomic-scale STM images as high-resolution topographs. However, it must be remembered that only electrons near the Fermi energy contribute to the tunnelling current, whereas all electrons contribute to the surface charge density. Since topography can reasonably be defined as a contour of constant surface charge density [15], STM images are intrinsically different from surface topographs.

In addition to its strong dependence on tip-sample separation, the tunnelling current is also dependent on the electron density of states (DOS) of both tip and sample (equation B1.19.1). This dependence can be exploited to produce a map of the local DOS under the tip by varying the applied voltage and measuring the tunnelling current. Both occupied and unoccupied electronic states can be probed by this method, which is known as scanning tunnelling spectroscopy (STS) [16]. The traditional method of mapping DOS is to use ultraviolet photoelectron spectroscopy (UPS) to measure occupied states and inverse photoemission spectroscopy (IPS) to measure empty states. However, it is important to remember that these data do not correspond exactly to those derived from STS measurements. Firstly, the STS spectrum is a convolution of tip and sample properties (a potential problem, should the tip become contaminated during the experiment). Secondly, since states near the upper edge of the energy range investigated see a lower barrier than those near the lower end, they contribute a greater tunnelling current, so that sensitivity to occupied states falls off with increasing energy below the Fermi level. Thirdly, STS is a much more surface- (and above-surface-) sensitive technique than UPS or IPS, meaning that surface electronic states contribute far more to the STS spectrum. This also means that the sensitivity to s, p, and d states is different in STS, due to the different degrees to which the electron density associated with these states extends out of the surface.

-4-

(A) SEMICONDUCTORS

STM found one of its earliest applications as a tool for probing the atomic-level structure of semiconductors. In 1983, the 7×7 reconstructed surface of Si(111) was observed for the first time [17] in real space; all previous observations had been carried out using diffraction methods, the 7×7 structure having, in fact, only been hypothesized. By capitalizing on the spectroscopic capabilities of the technique it was also proven [18] that STM could be used to probe the electronic structure of this surface (figure B1.19.3).

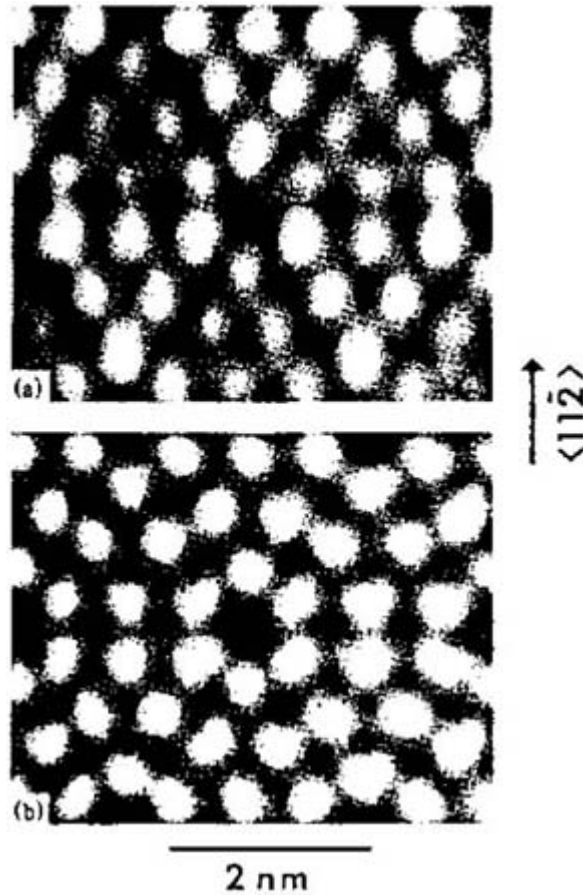


Figure B1.19.3. STM images of Si(111)-(7×7) measured with (a) -2 V and (b) $+2$ V applied to the sample. (Taken from [18], figure 1.)

A complete STS spectrum of the 7×7 reconstructed Si(111) surface displays remarkable correlation with the corresponding UPS and IPS spectra [19] (figure B1.19.4), showing the potential value of this approach. The high spatial resolution of the STS technique has also been demonstrated using a silicon surface containing impurity atoms [20] (figure B1.19.5), where the absence or presence of a band gap over *an individual atom* shows whether it belongs, respectively, to the silicon or to a metallic impurity.

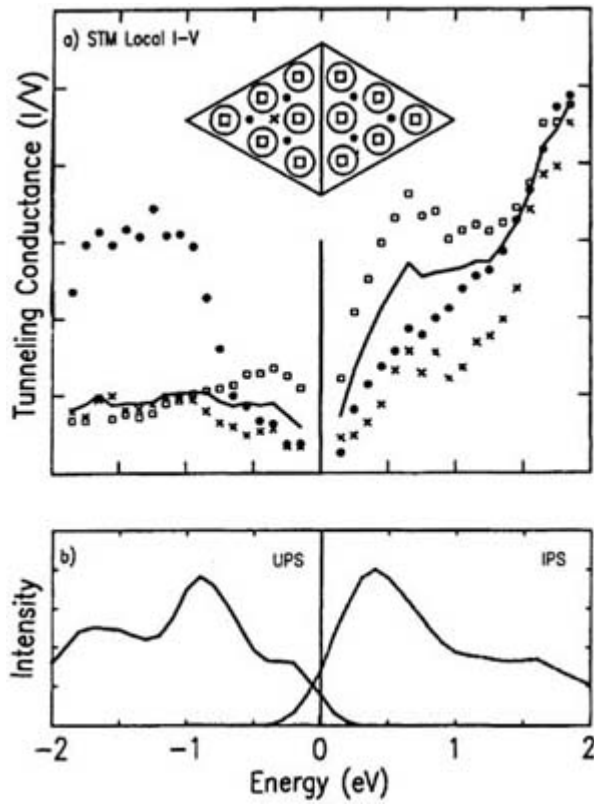


Figure B1.19.4. (a) Local conductance STS measurements at specific points within the Si(111)-(7 × 7) unit cell (symbols) and averaged over whole cell. (b) Equivalent data obtained by ultraviolet photoelectron spectroscopy (UPS) and inverse photoemission spectroscopy (IPS). (Taken from [19], figure 2.)

-6-

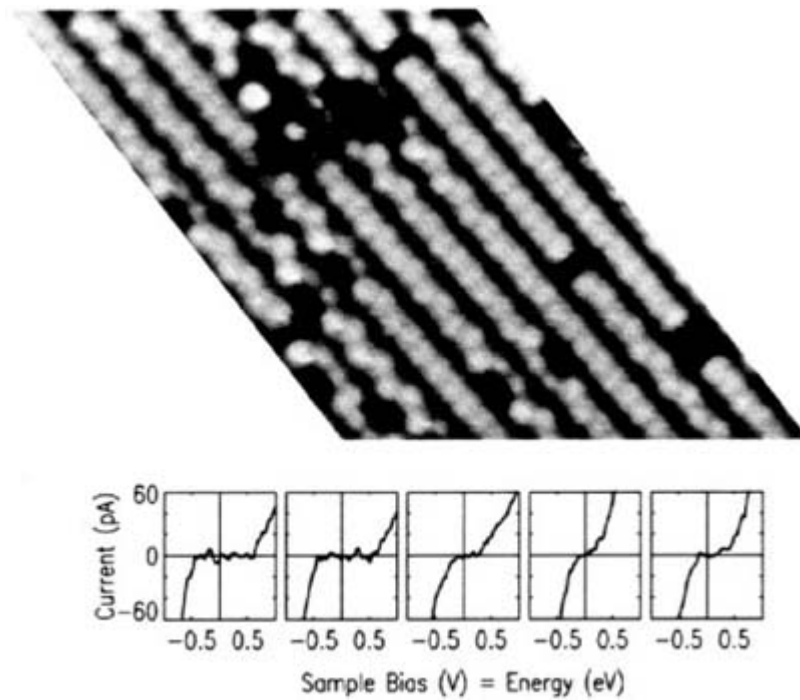


Figure B1.19.5. Tunnelling $I-V$ curves acquired across a defect on Si(100). Away from the defect a bandgap can be seen. Over the defect itself, the bandgap disappears, suggesting that it possesses metallic character.

This analysis was performed with spatial resolution of better than 1 nm. (Taken from [20], figure 2 and figure 6.)

Chemical reactions of ammonia with the silicon surface have also been clearly observed using STS [21], where the disappearance of the π and π^* states characteristic of the clean surface coincides with the formation of Si–H antibonding states corresponding to the dissociation of the ammonia on the Si surface.

Other semiconductors have also proved to be a fruitful ground for STM investigation. Zheng *et al* [22] have used the spatial resolution and electronic state sensitivity of STM to spatially display the electronic characteristics of single Zn impurity atoms in Zn-doped GaAs, both in filled and empty states, which show spherical and triangular symmetry, respectively. Upon imaging a number of Zn-induced features, a variety of different heights were recorded, corresponding to the depth of the impurity atoms within the sample. Thus STM was used to probe both the chemical nature and the 3D spatial location of the impurity atoms—an achievement that would have been inconceivable before the advent of STM.

STM has not as yet proved to be easily applicable to the area of ultrafast surface phenomena. Nevertheless, some success has been achieved in the direct observation of dynamic processes with a larger timescale. Kitamura *et al* [23], using a high-temperature STM to scan single lines repeatedly and to display the results as a time-*versus*-position pseudoimage, were able to follow the diffusion of atomic-scale vacancies on a heated Si(001) surface in real time. They were able to show that vacancy diffusion proceeds exclusively in one dimension, along the dimer row.

-7-

(B) METALS

STM has been applied with great success to the study of metals and adsorbate–metal systems [24]. This has naturally brought the technique into the mainstream of surface science, where structural information at the atomic level could previously only be obtained via diffraction methods such as low-energy electron diffraction (LEED) [25]. The STM can also provide a level of electronic information and visualization of the quantum mechanical behaviour of electrons that is unavailable from other methods: the images of copper and silver surfaces obtained by the groups of Eigler [26] and Avouris [27], showing standing waves produced by the defect-induced scattering of the 2D electron gas in surface states, bear eloquent testament to this (figure B1.19.6).

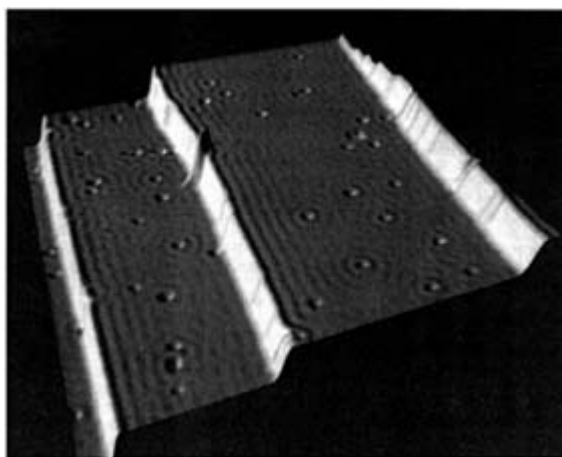


Figure B1.19.6. Constant current 50 nm × 50 nm image of a Cu(111) surface held at 4 K. Three monatomic steps and numerous point defects are visible. Spatial oscillations (electronic standing waves) with a

periodicity of ~ 1.5 nm are evident. (Taken from [26], figure 1.)

Surface reconstructions have been observed by STM in many systems, and the technique has, indeed, been used to confirm the ‘missing row’ structure in the 1×2 reconstruction of Au(110) [28]. As the temperature was increased within 10 K of the transition to the disordered 1×1 phase (700 K), a drastic reduction in domain size to ~ 20 – 40 Å (i.e. less than the coherence width of LEED) was observed. In this way, the STM has been used to help explain and extend many observations previously made by diffraction methods.

STM studies of simple adsorbates on metal surfaces have proved challenging, partly due to the significant mobility of most small species on metals at room temperature, which therefore generally necessitates low-temperature operation. Additionally, since adsorbates can change the local density of states in the metal surface, particular care must be taken not to interpret STM images of adsorbate–metal systems as simple topographs, but rather to capitalize on the technique’s capability for observing unoccupied and occupied energy states. In this way, the bond between adsorbate and substrate can be investigated on a local level, subject to the restrictions on energy range mentioned above. By observing the electronic changes in the neighbourhood of an adsorption site, much can be learned about the range over which chemical bonds can act and influence each other.

-8-

Metal surfaces in motion have also been characterized by STM, one of the clearest examples being the surface diffusion of gold atoms on Au(111) [29] (figure B1.19.7). Surface diffusion of adsorbates on metals can be followed [30] provided that appropriate cooling systems are available, and STM has been successfully employed to follow the 2D dendritic growth of metals on metal surfaces [31].

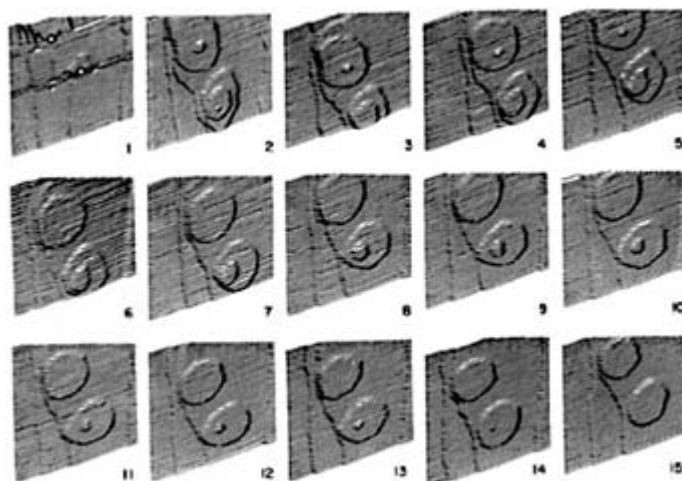


Figure B1.19.7. A series of time-lapse STM topographic images at room temperature showing a 40 nm \times 40 nm area of Au(111). The time per frame is 8 min, and each took about 5 min to scan. The steps shown are one atomic unit in height. The second frame shows craters left after tip–sample contact, which are two and three atoms deep. During a 2 h period the small craters have filled completely with diffusing atoms, while the large craters continue to fill. (Taken from [29], figure 1.)

(C) ORGANIC SURFACES

The operation of the STM depends on the conduction of electrons between tip and sample. This means, of course, that insulating samples are, in general, not accessible to STM investigations. Nevertheless, a large body of work [32] dealing with STM characterization of thin organic films on conducting substrates is now in

the literature, and the technique provides local structural and electronic information that is essentially inaccessible by any other method.

STM of thin organic layers involves the tunnelling of current between the tip and the conducting substrate, underneath the organic layer. By choosing the tunnelling parameters appropriately [32] ($\approx 0.3\text{--}1\text{ V}$, $0.05\text{--}1\text{ nA}$ for adsorbate, $0.1\text{--}0.3\text{ V}$, $0.3\text{--}10\text{ nA}$ for substrate), the method can be used to image either the substrate or the adsorbate—or both simultaneously, if a suitable voltage programme is used—repeating each line scanned at both voltages. There is some evidence that the tip can damage the organic layer during the imaging process [33]. The precise mechanism by which insulating molecules are imaged remains a topic of much discussion. Although single organic molecules have been successfully imaged by STM [34], the majority of STM studies of organic species has concerned a single monolayer of molecules deposited by evaporation, or by self-assembly, or by Langmuir–Blodgett techniques [35]. Often these images corroborate what had already been deduced from painstaking LEED investigations: an example is the imaging of co-adsorbed arrays of benzene and mobile CO, as seen by Ohtani *et al* [36].

-9-

One class of large molecules that was investigated relatively early was liquid crystals [37, 38], and in particular the group 4-*n*-alkyl-4'-cyanobiphenyl (*m*CB). These molecules form a highly crystalline surface adlayer, and STM images clearly show the characteristic shape of the molecule (figure B1.19.8).

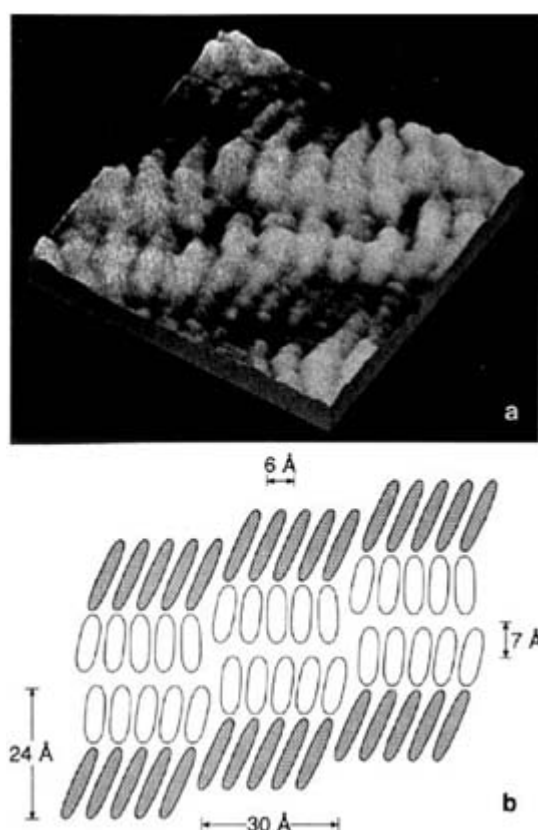


Figure B1.19.8. (a) STM image ($5.7\text{ nm} \times 5.7\text{ nm}$) of 10-alkylcyanobiphenyl on graphite; (b) model showing the packing of the molecules. The shaded and unshaded segments represent the alkyl tails and the cyanobiphenyl head groups, respectively. (Taken from [38], figure 2.)

The self-assembly of alkanethiols on gold has been an important topic in surface chemistry over the last few years [39] and STM has contributed significantly to our understanding of these systems. In particular, the

formation of etch pits on the surface of Au(111) following treatment with alkanethiols is a phenomenon that was first observed by STM [40]. The segregation of thiols of different molecular weight or functionality is proving to be a relevant issue in their application. Stranick *et al* [41] have used STM to show the segregation of thiols with only very slight molecular differences into domains of size 10–100 Å and their subsequent coalescence.

The STM study of biological macromolecules has also been an area of great activity, and the imaging of DNA has been one of the challenges of the STM technique [42] (figure B1.19.9). The elimination of artifacts has been a major issue in this story, and the work of Beebe *et al* [43] showing that ‘DNA-like’ structures were to be seen on the surface of clean graphite (HOPG) substrates was something of a milestone (figure B1.19.10).

-10-

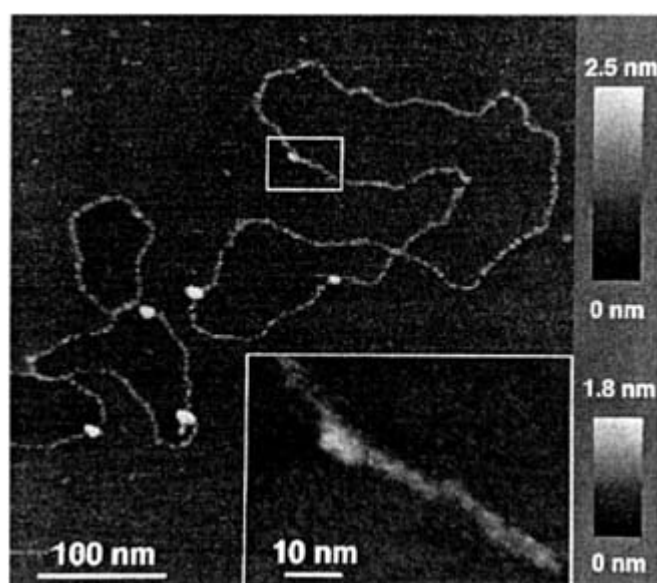


Figure B1.19.9. Plasmid DNA (pUC18) on mica imaged by STM at high resolution. The inset is a cut-out of a zoomed-in image taken immediately after the overview. (Taken from [42], figure 2.)

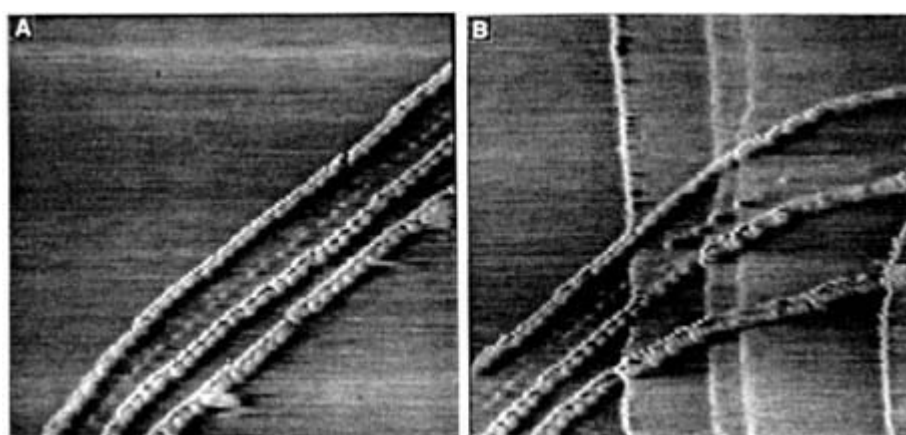


Figure B1.19.10. These images illustrate graphite (HOPG) features that closely resemble biological molecules. The surface features not only appear to possess periodicity (A), but also seem to meander across the HOPG steps (B). The average periodicity was 5.3 ± 1.2 nm. Both images measure 150 nm × 150 nm. (Taken from [43], figure 4.)

Other biomolecules imaged have included all DNA bases [44], polysaccharides [45] and proteins [46, 47]. In many cases there is strong evidence that the imaging process is facilitated by the presence of ultrathin (conducting) water films on the surface of the sample [48, 49 and 50].

Lastly, STM has also been applied to the molecular-level imaging of polymer structures. In some cases these materials were deposited by Langmuir–Blodgett techniques [51], and in some cases by *in situ* polymerization [52]. Fujiwara *et al* [51] have used molecular dynamics simulations to interpret the images obtained from STM experiments. The combined use of these two techniques is proving to be a very powerful tool for understanding the conformation of polymer films on surfaces. They showed that the individual polyimide strands observed were aligned parallel to the deposition direction of the Langmuir–Blodgett film.

(D) ELECTROCHEMISTRY

The molecular-level observation of electrochemical processes is another unique application of STM [53, 54]. There are a number of experimental difficulties involved in performing electrochemistry with a STM tip and substrate, although many of these have been essentially overcome in the last few years.

If the scanning tip is to be involved in electrochemical reactions, it is important to remember that at micrometre separations (i.e. when the tip is too far from the substrate for tunnelling to occur), the faradaic current is given by the equation [54]: $I_f \approx 4nFD_O C_O r$, where D_O is the diffusion coefficient of a particular species, F is Faraday's constant, C_O is the concentration of the species in solution, r is the radius of a disc of area equal to the effective exposed area of the tip and n is the number of electrons involved in the reaction. The total tip current, I , when the separation is small enough for tunnelling to occur, is given by $I = I_f + I_t$, where I_t is the tunnelling current, which is virtually independent of the total tip area exposed. In order to minimize I_f so as to be able to perform meaningful STM experiments, the exposed tip must be made as small as possible, and a plethora of techniques has been developed [53] for insulating all but the very end of the tip.

Several designs for STM electrochemical cells have appeared in the literature [55]. In addition to an airtight liquid cell and the tip insulation mentioned above, other desirable features include the incorporation of a reference electrode (e.g. Ag/AgCl in saturated KCl) and a bipotentiostat arrangement, which allows the independent control of the two working electrodes (i.e. tip and substrate) [56] (figure B1.19.11).

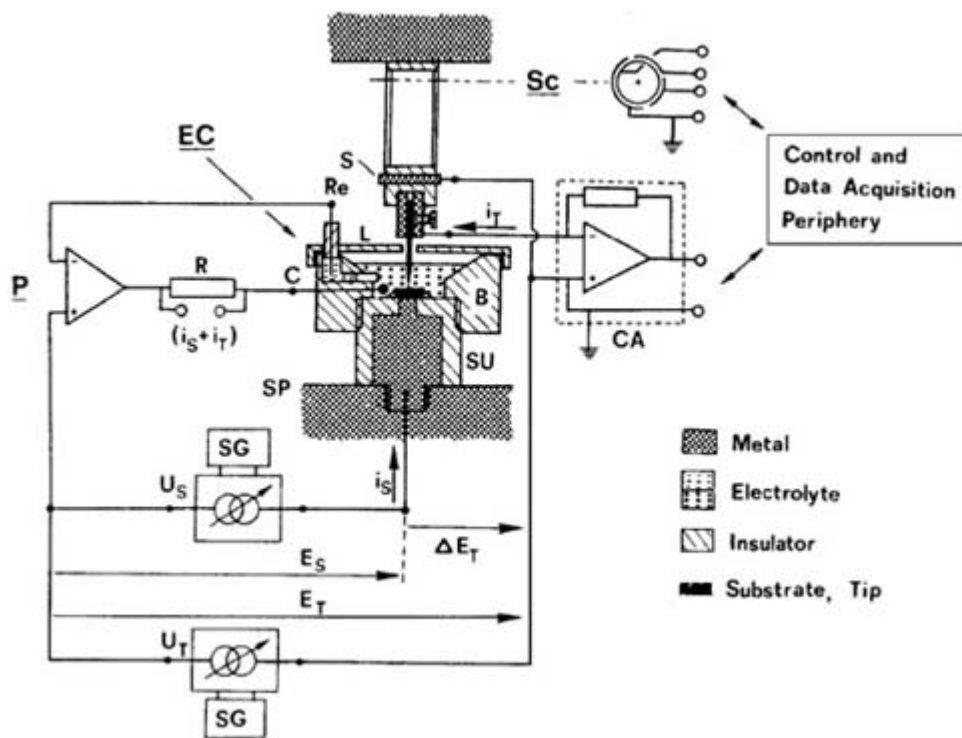


Figure B1.19.11. Schematic presentation of a potentiostatic STM system, with individual potential control of substrate and tip. Piezoelectric single-tube scanner *Sc* with titanium spacer plate *S*; electrochemical cell *EC* consisting of plexiglass beaker *B* with Pt counterelectrode *C* and Ag/AgCl reference electrode *Re* in 0.1 M NaCl; plexiglass lid *L*; PTFE support unit *SU* with epoxy-sealed substrate, mounted on support plate *SP*. Low-noise potentiostat *P* with low-impedance voltage units U_S and U_T , both equipped with low-pass filter and signal generator *SG*; precision resistor *R* for measuring $(i_S + i_T)$; low-noise current amplifier *CA* for measuring i_T . (Taken from [56], figure 1.)

Examining electrodes and how they change under conditions of electrochemical reaction has been a major part of the electrochemical STM work performed until now. Many studies have revealed changes in surface reconstructions on silver and gold electrodes during electrochemical reactions [57], as well as increasing or decreasing surface roughness, depending on the conditions and electrolyte employed. Another field of activity has been the monitoring of metal deposition on electrodes [58], which is, of course, of tremendous practical importance. Since STM can image both periodic and non-periodic structures, it is of great utility, both in determining the geometric relationships between deposited metal and substrate, as well as in assessing the role of steps and defects in the deposition process [57].

Corrosion is another economically significant process that can be investigated on a molecular level, thanks to electrochemical STM. In addition to a number of academically interesting studies of systems such as the selective dissolution of copper from Cu–Au alloys [59], STM has also been used to investigate the properties of iron and steel under a variety of conditions designed to induce either passivation, corrosion, or electrochemical anodization [60, 61]. In the case of corrosion, STM has been used to monitor the growth of magnetite crystallites on the surface of the sample as it is taken through several successive cyclic voltammograms [61].

The technique of scanning electrochemical microscopy (SECM) [62] uses the same apparatus as in electrochemical STM, but instead of measuring tunnelling currents, the reaction $O + ne \rightarrow R$ (where *O* and *R*

are oxidized and reduced species, respectively) is followed, by measuring the Faradaic current, I_f , at distances further from the substrate than those at which tunnelling will readily occur. The current, I_f , at distances far from the substrate surface, corresponds to the hemispherical diffusion of O to the tip surface (figure B1.19.12). As the tip nears the surface, this current is perturbed, either by hindered diffusion (lower current) or by reoxidation of R on the surface (higher current). The conductivity, potential, and electrochemical activity will therefore all influence I_f , which can thus be used to produce an electrochemical image of the surface—if plotted as a function of x and y —as the tip is rastered over the surface. The technique has been used to image metals, polymers, biological materials and semiconductors.

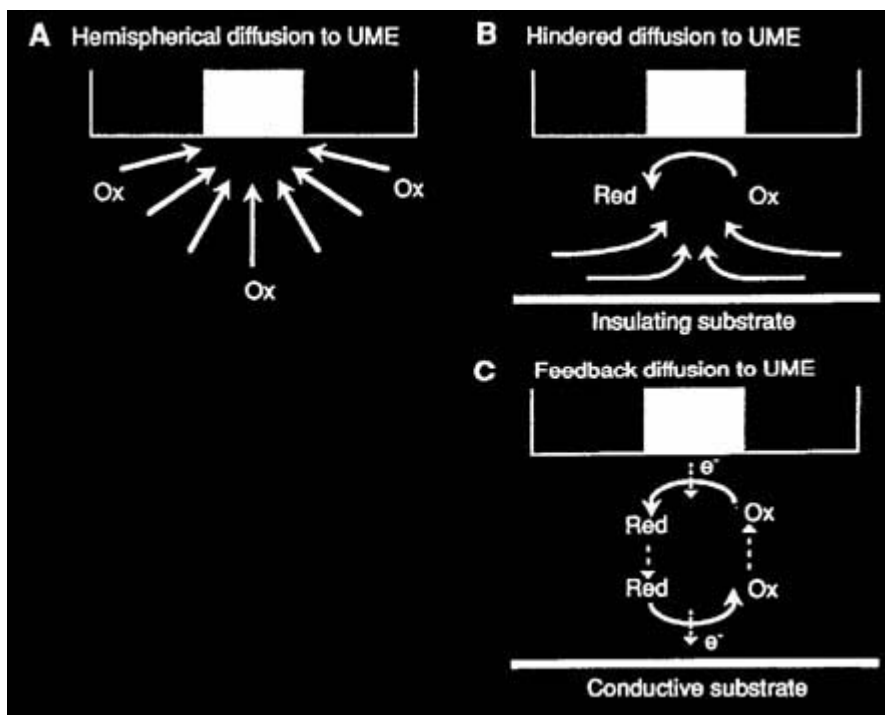


Figure B1.19.12. Basic principles of SECM. (a) With ultramicroelectrode (UME) far from substrate, diffusion leads to a steady-state current, i_T, ∞ . (b) UME near an insulating substrate. Hindered diffusion leads to $i_T < i_T, \infty$. (c) UME near a conductive substrate. Positive feedback leads to $i_T > i_T, \infty$. (Taken from [62], figure 2.)

(E) CATALYSIS

It has long been the goal of many catalytic scientists to be able to study catalysts on a molecular level under reaction conditions. Since the vast majority of catalytic reactions take place at elevated temperatures, the use of STM for such *in situ* catalyst investigations was predicated upon the development of a suitable STM reaction cell with a heating stage. This has now been done [3] by McIntyre *et al*, whose cell-equipped STM can image at temperatures up to 150 °C and in pressures ranging from ultrahigh vacuum up to several atmospheres. The set-up has been used for a number of interesting studies. In one mode of operation [63] (figure B1.19.13(a)), a Pt–Rh tip was first used to image clusters of carbonaceous species formed on a clean Pt(111) surface by heating a propylene adlayer to 550 K, and later to catalyze

the rehydrogenation of the species (in a propylene/hydrogen atmosphere) at room temperature. The catalytic activity of the tip was induced by applying a voltage pulse, which presumably cleaned the surface of deactivating debris (figure B1.19.13(b)).

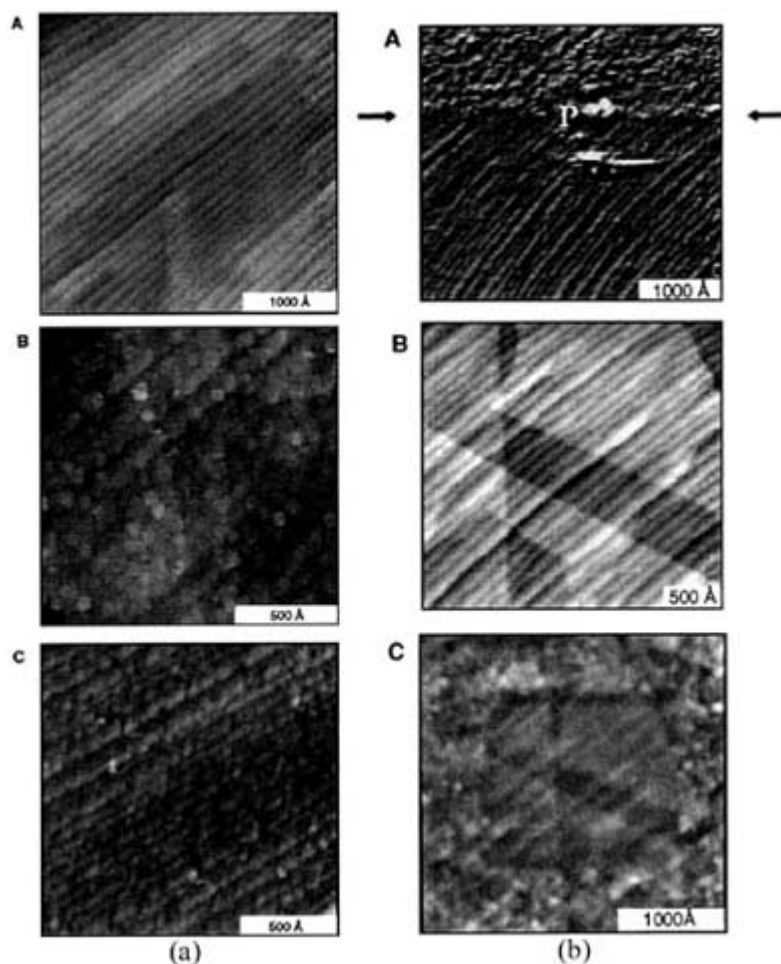


Figure B1.19.13. (a) Three STM images of a Pt(111) surface covered with hydrocarbon species generated by exposure to propene. Images taken in constant-height mode. (A) after adsorption at room temperature. The propylidyne ($\equiv\text{C}-\text{CH}_2-\text{CH}_3$) species that formed was too mobile on the surface to be visible. The surface looks similar to that of the clean surface. Terraces (~ 10 nm wide) and monatomic steps are the only visible features. (B) After heating the adsorbed propylidyne to 550 K, clusters form by polymerization of the C_xH_y fragments. The clusters are of approximately round shape with a diameter equal to the terrace width. They form rows covering the entire image in the direction of the step edges. (C) Rows of clusters formed after heating to 700 K. At this higher temperature, the carbonaceous clusters are more compact and slightly smaller in size, as they evolve to the graphitic form when H

-15-

is lost completely. (b) The catalytic action of the STM Pt-Rh tip on a surface covered by carbonaceous clusters, as in figure B1.19.13(a). (B) Imaging was performed in 1 bar of a propene (10%) and hydrogen (90%) mixture at room temperature. (A) Carbon clusters were imaged in the top third of the image while the tip was inactive. A voltage pulse of 0.9 V was applied to the position marked P, leaving a mound of material 1.5 nm high. This process produced a chemically active Pt-Rh tip, which catalyzed the removal of all clusters in the remaining two-thirds of the image. Only the lines corresponding to the steps are visible. This image was illuminated from a near-incident angle to enhance the transition region where the tip was switched to its active state. (B) While the tip was in this catalytically active state, another area was imaged, and all of the clusters were again removed. (C) A slightly larger image of the area shown in (B) (centre square of this image), obtained after the tip was deactivated, presumably by contamination. The active-tip lifetime was of the order of minutes. (Taken from [63], figure 1 and figure 2.)

(F) STM AS A SURFACE MODIFICATION METHOD

Within a few years of the development of STM as an imaging tool, it became clear that the instrument could also find application in the manipulation of individual or groups of atoms on a surface [64]. Perhaps the most dramatic image originated from Eigler and Schweizer [65], who manipulated single physisorbed atoms of xenon on a Ni(110) surface, held at liquid helium temperature (figure B1.19.14). The tip–Xe distance was reduced (by raising the setpoint for the tunnelling current) until the tip–sample interaction became strong enough for the tip to be able to pick up the atom. After being moved to the desired location, the atom was removed by reversing the procedure. Using a similar experimental set-up, Crommie *et al* [66] have managed to shape the spatial distribution of electrons on an atomic scale, by building a ring of 48 iron adatoms (a ‘quantum corral’) on a Cu(111) surface, which confines the surface-state electrons of the copper by virtue of the scattering effect of the Fe atoms (figure B1.19.15). STS measurements of the local densities of states for the confined electrons correspond to the expected values for a ‘particle-in-a-box’, where the box is round and two-dimensional. In a similar way, Yokoyama *et al* [67] formed a pair of long straight chains of Al on the Si(001)-c(4 × 2) surface to create well defined 1D quantum wells. The electrons in the Π^* surface states can propagate only in the dimer-row direction of Si(001)-c(4 × 2) because of nearly flat dispersion in the perpendicular direction. The STM/STS measurements of the standing-wave patterns and their discrete energy levels could be interpreted according to the ‘1D particle-in-a-box model’. This technique shows considerable promise for the further investigation of confined electrons and waveguides. There are numerous other means for moving atoms in surfaces, including voltage-pulsing techniques, which show promise as potential lithographic methods for silicon [68].

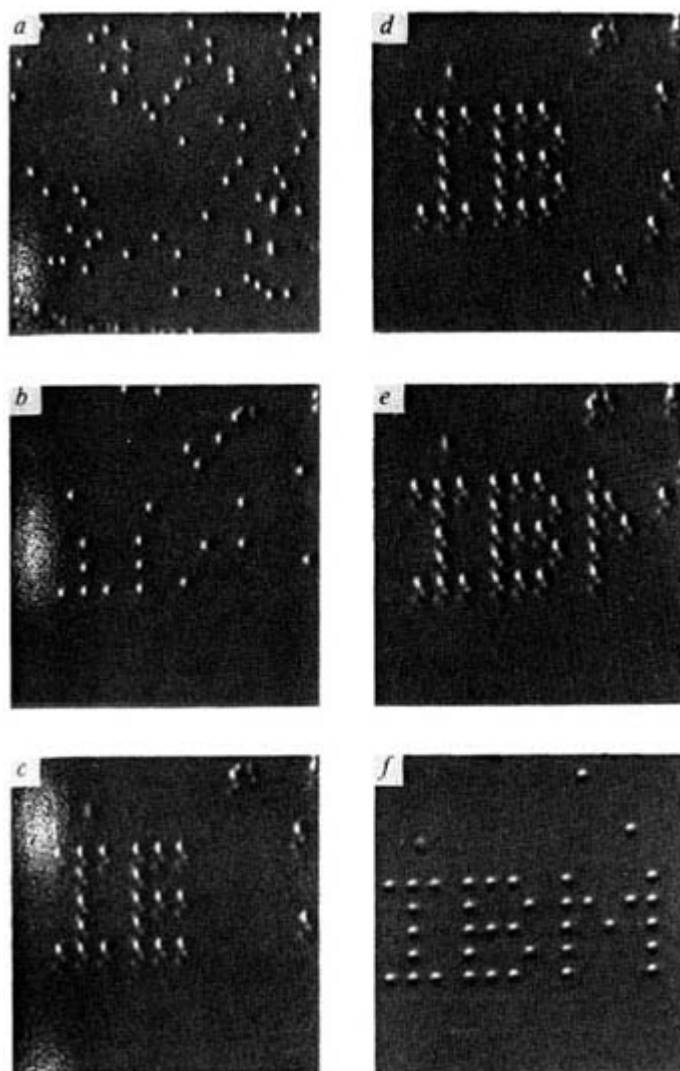


Figure B1.19.14. A sequence of STM images taken during the construction of a patterned array of xenon atoms on a Ni(100) surface. Grey scale is assigned according to the slope of the surface. The atomic structure of the nickel surface is not resolved. Each letter is 5 nm from top to bottom. (Taken from [65], figure 1.)

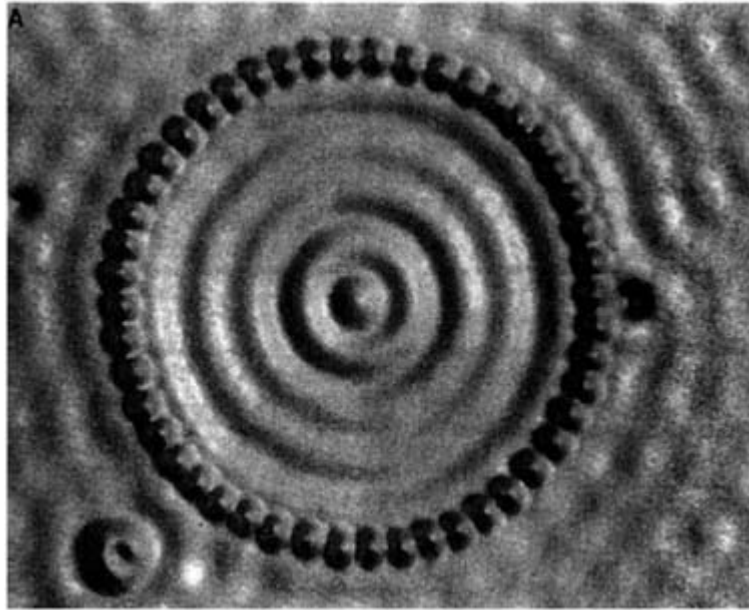


Figure B1.19.15. Spatial image of the eigenstates of a quantum corral. 48-atom Fe ring constructed on a Cu (111) surface. Average diameter of ring is 14.3 nm. The ring encloses a defect-free region of the surface. (Taken from [66], figure 2.)

Finally, a technique that combines chemical vapour deposition (CVD) with STM has been devised by Kent *et al* [69]. The CVD gas used was iron pentacarbonyl, which is known to decompose under electron bombardment. Decomposition between tip and sample was found to occur at bias voltages above 5 V, forming iron clusters as small as 10 nm in diameter on the Si(111) substrate. Of particular practical interest is that arrays of 20 nm diameter dots have been shown to be magnetic, presenting a whole new range of possibilities for high-density data storage, as well as providing a convenient laboratory for nanometre-scale experiments in quantum magnetism.

B1.19.3 FORCE MICROSCOPY

B1.19.3.1 PRINCIPLES

(A) BACKGROUND

A major limitation of the scanning tunnelling microscope is its inability to analyse insulators, unless they are present as ultrathin films on conducting substrates. Soon after the development of the STM, work started on the development of an equivalent nanoscale microscope based on force instead of current as its imaging parameter [70]. Such an instrument would be equally adept at analysing both conducting and insulating samples. Moreover, the instrument already existed on a micro- and macro-scale as the stylus profilometer [71]; this is typically used to measure surface roughness in one dimension, although it had been extended into a three-dimensional imaging technique, with moderate resolution (0.1 μm lateral and 1 nm vertical), by Teague *et al* [72].

The concept that Binnig and co-workers [73] developed, which they named the atomic force microscope (AFM, also known as the scanning force microscope, SFM), involved mounting a stylus on the end of a cantilever with a spring constant, k , which was lower than that of typical spring constants between atoms. This sample surface was then rastered below the tip, using a piezo system similar to that developed for the STM, and the position of the tip monitored [74]. The sample position (z -axis) was altered in an analogous way to STM, so as to maintain a constant displacement of the tip, and the z -piezo signal was displayed as a function of x and y coordinates (figure B1.19.16). The result is a force map, or image of the sample's surface [75], since displacements in the tip can be related to force by Hooke's Law, $F = -kz$, where z is the cantilever displacement. In AFM, the displacement of the cantilever by the sample is very simply considered to be the result of long-range van der Waals forces and Born repulsion between tip and sample. However, in most practical implementations, meniscus forces and contaminants often dominate the interaction with interaction lengths frequently exceeding those predicted [76]. In addition, an entire family of force microscopies has been developed, where magnetic, electrostatic, and other forces have been measured using essentially the same instrument.

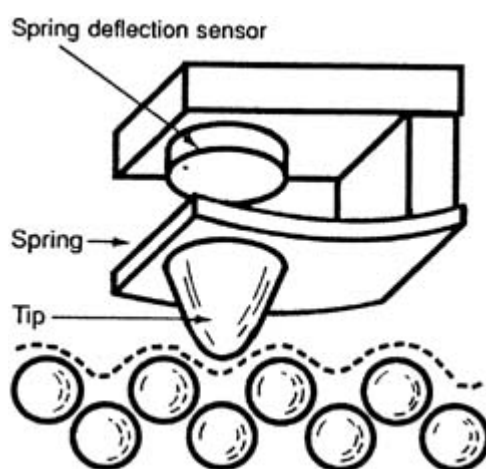


Figure B1.19.16. Schematic view of the force sensor for an AFM. The essential features are a tip, shown as a rounded cone, a spring, and some device to measure the deflection of the spring. (Taken from [74], figure 6.)

(B) AFM INSTRUMENTATION

The first AFM used a diamond stylus, or 'tip' attached to a gold-foil cantilever, and much thought was given to the choice of an appropriate k -value [73]. While on the one hand a soft spring was necessary (in order to obtain the maximum deflection for a given force), it was desirable to have a spring with a high resonant frequency (10–100 kHz) in order to avoid sensitivity to ambient noise. The resonant frequency, f_0 , is given by the equation

$$f_0 = (1/2\pi)\sqrt{k/m_0}$$

where m_0 is the effective mass loading the spring. Thus, as k is reduced to soften the spring, the mass of the cantilever must be reduced in order to keep the k/m_0 ratio as large as possible. Nowadays, cantilevers and integrated tips are

routinely microfabricated out of silicon or silicon nitride. Typical dimensions of a cantilever are of the order

of $1 \times 10 \times 100 \mu\text{m}^3$ [77], the exact dimensions depending on the intended use. Cantilevers designed to operate in contact with the surface, in a similar way to a surface profilometer, have low spring constants (usually less than 1 N m^{-1}) and correspondingly low resonant frequencies. Such levers are often fabricated in a V-shape configuration (figure B1.19.17), which makes for a greater stability towards lateral motion. If the tip is to come into hard contact with the surface, high-aspect-ratio tips are often desirable, with, typically, a radius of curvature of 10–30 nm. It is interesting to note that since experiments are generally carried out with contact forces on the order of nanonewtons, contact pressures in these experiments can be in the gigapascal range. In a stylus profilometer, the force exerted on the sample is some five orders of magnitude greater, but it is exerted over a larger contact area, leading to pressures in the tens of megapascals.



Figure B1.19.17. Commercially produced, microfabricated, V-shaped Si_3N_4 cantilever and tip for AFM (Taken from [215].)

When the lever is intended for use with the tip separated from the surface, the lever stiffness is usually greater than 10 N m^{-1} with a high resonant frequency. In this case, more care is taken to prepare tips with small radii of curvature—sometimes as low as 2 nm. However, in reality, most experiments are performed with tips of unknown radii or surface composition, apart from rare cases where the AFM has been combined with field ion microscopy [78] or a molecule or nanotube of known dimensions and composition has been attached to the tip [79]. It is likely that in most AFMs, microasperities and contaminants mediate the contact.

Detection of cantilever displacement is another important issue in force microscope design. The first AFM instrument used an STM to monitor the movement of the cantilever—an extremely sensitive method. STM detection suffers from the disadvantage, however, that tip or cantilever contamination can affect the instrument's sensitivity, and that the topography of the cantilever may be incorporated into the data. The most common methods in use today are optical, and are based either on the deflection of a laser beam [80], which has been bounced off the rear of the cantilever onto a position-sensitive detector (figure B1.19.18), or on an interferometric principle [81].

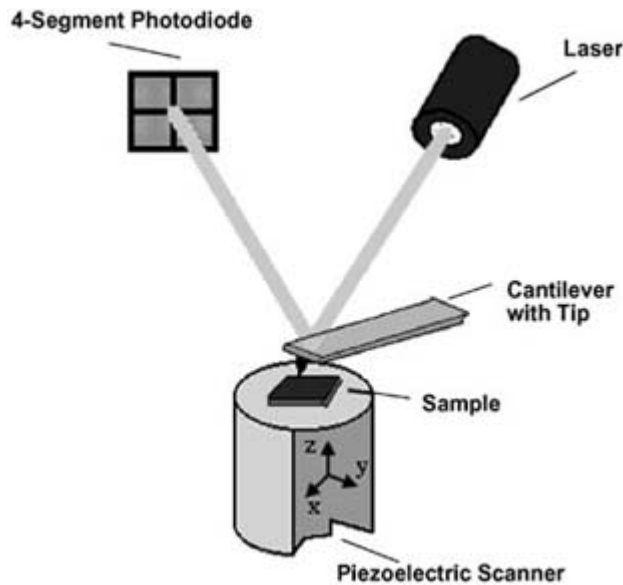


Figure B1.19.18. Schematic of an atomic force microscope showing the optical lever principle.

Lateral resolution in AFM is usually better than 10 nm and, by utilizing dynamic measurement techniques in ultrahigh vacuum, true atomic resolution can be obtained [82]. In hard contact with the surface, the atomic-scale structure may still appear to be present, but atomic-scale defects will no longer be visible, suggesting that the image is actually averaged over several unit cells. The precise way in which this happens is still the subject of debate, although the ease with which atomic periodicity can be observed with layered materials is probably due to the Moiré effect suggested by Pethica [83]. In this case a periodic image is formed by the sliding of planes directly under the tip caused by the lateral tip motion as the force varies in registry with unit lattice shear.

A further issue that should be considered when interpreting AFM images is that they are convolutions of the tip shape with the surface (figure B1.19.19). This effect becomes critical with samples containing ‘hidden’ morphology (or ‘dead zones’) on the one hand (such as deep holes into which the tip does not fit, or the underside of spherical features), or structure that is comparable in size to that of the tip on the other. While the hidden morphology cannot be regenerated, there have been several attempts to deconvolute tip shape and dimensions from AFM images (‘morphological restoration’) [84]. Some of these methods involve determining tip parameters by imaging a known sample, such as monodisperse nanospheres [85] or faceted surfaces [86]. Another approach is to analyse the AFM image as a whole, extracting a ‘worst case’ tip shape from common morphological features that appear in the image [87].

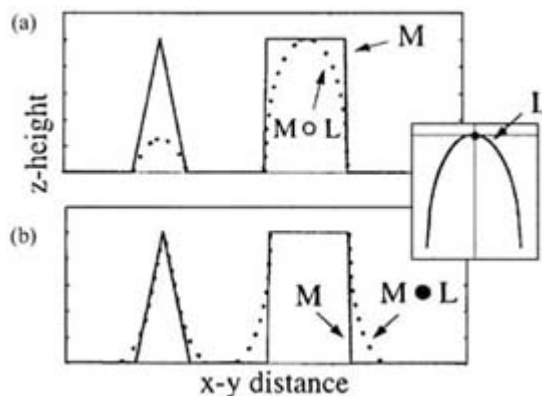


Figure B1.19.19. Examples of inaccessible features in AFM imaging. L corresponds to the AFM tip. The dotted curves show the image that is recorded in the case of (a) depressions on the underside of an object and (b) mounds on the top surface of an object. $M \cdot L$ and $M \oslash L$ correspond to convolutions of the surface features with the tip shape. (Taken from [85], figure 2.)

As with STM, the AFM can be operated in air, in vacuum or under liquids, providing a suitable cell is provided. Liquid cells (figure B1.19.20) are particularly useful for the examination of biological samples.

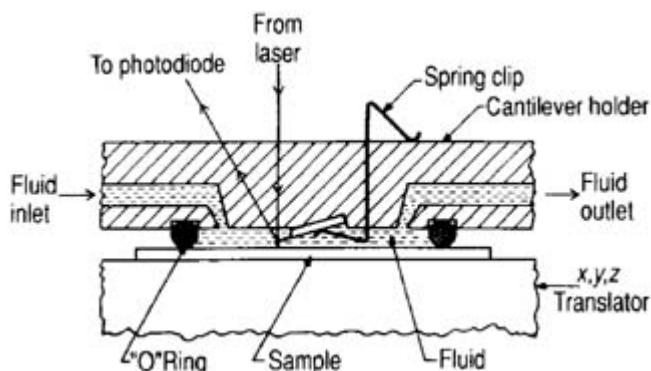


Figure B1.19.20. Cross section of an AFM fluid cell. (Taken from [216], figure 1.)

Analogously to STM, the image obtained in a force microscopy experiment is conventionally displayed on the computer screen as grey scales or false colour, with the lightest shades corresponding to peaks (or highest forces) and darkest shades corresponding to valleys (or lowest forces).

(C) FORCES IN AFM

Although imaging with force microscopy is usually achieved by means of rastering the sample in close proximity to the tip, much can be learned by switching off the x - and y -scanning piezos and following the deflection of the cantilever as function of sample displacement, from large separations down to contact with the surface, and then back out to large separations. The deflection–displacement diagram is commonly known as a ‘force curve’ and this technique as

‘force spectroscopy’, although strictly speaking it is displacement that is actually measured, and further processing [88], reliant on an accurate spring constant, is necessary in order to convert the data into a true force–distance curve. In order to obtain sensitive force curves, it is preferable to make dynamic force-gradient measurements with the force and energy being found by integration [89] or alternatively to measure frequency shift of the cantilever resonant frequency as a function of displacement. This method has become particularly popular in combination with non-contact mode AFM (see below). Although with this method it is not straightforward to relate frequency shifts to forces, it appears to be a promising technique for distinguishing between materials on the nanometre scale [90].

As the tip is brought towards the surface, there are several forces acting on it. Firstly, there is the spring force due to the cantilever, F_S , which is given by $F_S = -kz$. Secondly, there are the sample forces, which, in the case of AFM, may comprise any number of interactions including (generally attractive) van der Waals forces, chemical bonding interactions, meniscus forces or Born (‘hard-sphere’) repulsion forces. The total force

gradient as the tip approaches the sample is the convolution of spring and sample force gradients, or

$$\frac{\partial F}{\partial D} = \frac{k(\partial^2 U / \partial D^2)}{k + \partial^2 U / \partial D^2}$$

where U is the sample potential and D the tip–sample separation. If the spring constant of the cantilever is comparable to the gradient of the tip–surface interaction, then at some point where $\partial^2 U / \partial D^2$ (negative for attraction) equals k , the total force gradient becomes instantaneously infinite, and the tip jumps towards the sample [91] (figure B1.19.21). This ‘jump to contact’ is analogous to the jump observed when two attracting magnets are brought together. The kinetic energy involved is often sufficient to damage the tip and sample, thus reducing the maximum possible resolution during subsequent imaging. Once a jump to contact occurs, the tip and sample move together (neglecting sample deformation for the time being) until the direction of sample travel is reversed. The behaviour is almost always hysteretic, in that the tip remains in contact with the sample due to adhesion forces, springing back to the equilibrium position when these have been exceeded by the spring force of the cantilever. The adhesion forces add to the total force exerted on the sample, and are often caused by tip contamination. It has been found that pretreating the tip in ozone and UV light, in order to remove organic contamination, reduces the adhesion observed, and improves image quality [92]. Image quality can also be enhanced by tailoring the imaging medium (i.e. in a liquid cell) to have a dielectric constant intermediate between those of the tip and the sample. This leads to a small, repulsive van der Waals force, which eliminates the jump to contact, and has been shown to improve resolution in a number of cases [93, 94], probably due to the fact that the tip is not damaged during the approach. It should be noted that the jump to contact may also be eliminated if a stiff cantilever is chosen, such that k , the force constant of the cantilever, is greater than $\partial^2 U / \partial D^2$ at all separations. This condition for jump-to-contact is insufficient if the stiffness of the cantilever is artificially enhanced using feedback [95] or if dynamic measurements are made [96, 97].

Since the AFM is commonly used under ambient conditions, it must be borne in mind that the sample is likely to be covered with multilayers of condensed water. Consequently, as the tip approaches the surface, a meniscus forms between tip and surface, introducing an additional attractive capillary force. Depending on the tip radius, the magnitude of this force can be equal to or greater than that of the van der Waals forces and is observed clearly in the approach curve [98]. In fact, this effect has been exploited for the characterization of thin liquid lubricant films on surfaces [95]. The capillary forces may be eliminated by operation in ultrahigh vacuum, provided both tip and sample are baked, or, most simply, by carrying out the experiment under a contamination-free liquid environment, using a liquid cell [99].

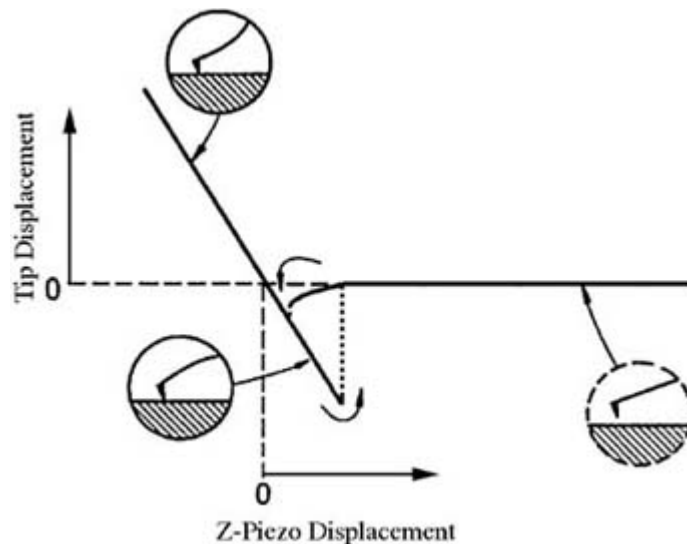


Figure B1.19.21. A plot of cantilever displacement as a function of tip sample separation during approach and retraction with an AFM. Note the adhesive forces upon retraction from the surface.

(D) NON-CONTACT AFM

Non-contact AFM (NC-AFM) imaging is now a well established true-atomic-resolution technique which can image a range of metals, semiconductors and insulators. Recently, progress has also been made towards high-resolution imaging of other materials such as C_{60} , DNA and polypropylene. A good overview of recent progress is the proceedings from the First International Conference on NC-AFM [100].

Most NC-AFMs use a frequency modulation (FM) technique where the cantilever is mounted on a piezo and serves as the resonant element in an oscillator circuit [101, 102]. The frequency of the oscillator output is instantaneously modulated by variations in the force gradient acting between the cantilever tip and the sample. This technique typically employs oscillation amplitudes in excess of 20 nm peak to peak. Associated with this technique, two different imaging methods are currently in use: namely, fixed excitation and fixed amplitude. In the former, the excitation amplitude to the lever (via the piezo) is kept constant, thus, if the lever experiences a damping close to the surface the actual oscillation amplitude falls. The latter involves compensating the excitation amplitude to keep the oscillation amplitude of the lever constant. This mode also readily provides a measure of the dissipation during the measurement [100].

Although both methods have produced true-atomic-resolution images it has been very problematic to extract quantitative information regarding the tip–surface interaction as the tip is expected to move through the whole interaction potential during a small fraction of each oscillation cycle. For the same reason, it has been difficult to conclusively identify the imaging mechanism or the minimum tip–sample spacing at the turning point of the oscillation.

Many groups are now trying to fit frequency shift curves in order to understand the imaging mechanism, calculate the minimum tip–sample separation and obtain some chemical sensitivity (quantitative information on the tip–sample interaction). The most common methods appear to be perturbation theory for considering the lever dynamics [103], and quantum mechanical simulations to characterize the tip–surface interactions [104]. Results indicate that the

interaction curve measured as a function of frequency shift does not correspond directly to the force gradient as first believed.

(E) INTERMITTENT CONTACT AFM

A further variation is intermittent contact mode or ‘TappingMode’ [105] (TappingMode® is a trademark of Digital Instruments, Santa Barbara, CA.), where the tip is oscillated with large amplitudes (20–100 nm) near its resonant frequency, and the amplitude used to control the feedback loop. The system is adjusted so that the tip contacts the sample once within each vibrational cycle. Since the force on the sample in this mode is both small (<5 nN) and essentially normal to the surface, it is far less destructive than contact AFM, with its inherently large shear forces. This is of great importance when imaging biological materials; a further development of the intermittent-mode AFM, which allows it to be operated under *liquids* [106], extends the possibilities in this area even further.

(F) MAGNETIC FORCE MICROSCOPY

Magnetic forces may be exploited for the imaging of samples containing magnetic structure. Resolutions as high as 10 nm have been reported [107]. The central modification of the AFM needed to perform magnetic force microscopy (MFM) is the use of a magnetic tip, which often consists of an electrochemically etched ferromagnetic material, or a non-magnetic tip that has been coated with a magnetic thin film [108]. The experiment is run in non-contact mode, with the tip some 10 nm away from the surface. Detection at long range helps distinguish between magnetic and non-magnetic interactions. Greater sensitivity is obtained when the cantilever is oscillated and magnetic force gradients detected by changes in the resonant frequency as the tip approaches the magnetic surface. The method is unique in its ability to image magnetic structure in surfaces (figure B1.19.22) [109], which lies at the heart of magnetic data storage technology.

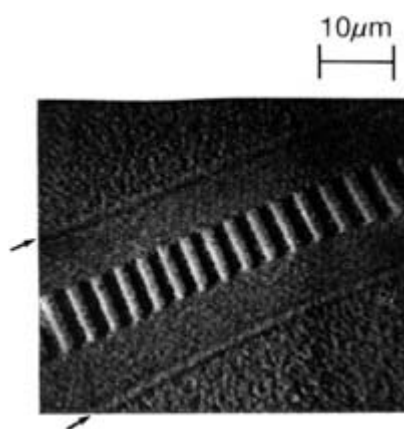


Figure B1.19.22. Magnetic force microscopy image of an 8 μm wide track on a magnetic disk. The bit transitions are spaced every 2 μm along the track. Arrows point to the edges of the DC-erased region. (Taken from [109], figure 7.)

(G) LATERAL FORCE MICROSCOPY

Lateral force microscopy (LFM) has provided a new tool for the investigation of tribological (friction and wear) phenomena on a nanometre scale [110]. Alternatively known as friction force microscopy (FFM), this variant of AFM focuses on the lateral forces experienced by the tip as it traverses the sample surface, which

correspond to the local coefficients of dynamic friction. LFM can therefore provide a frictional map of the surface with sub-nanometre resolution. It therefore has the potential to reveal chemical differences between regions of similar morphology, virtually down to the atomic scale.

The LFM method is an inherently contact-mode technique, and can be performed with an AFM, provided that there is some means of measuring the lateral tip displacement. Mate *et al* [111] were the first to modify their AFM in order to detect lateral forces and to observe frictional behaviour on the atomic scale. Their detection system was interferometric, and their cantilever and tip consisted of a shaped tungsten wire. Later developments using the laser beam-deflection method [112, 113] with two sets of position-sensing detectors (figure B1.19.23), enabled both lateral and normal forces to be measured simultaneously. Clearly, these two sets of forces are not entirely independent, since a lateral force will be felt as the tip is scanned over a step, for example, irrespective of the frictional coefficient at the step. However, by measuring the lateral force as the tip is scanned in both directions along the same line (producing a ‘friction loop’, figure B1.19.24), and subtracting one trace from the other, the frictional information can be separated from the purely morphological.

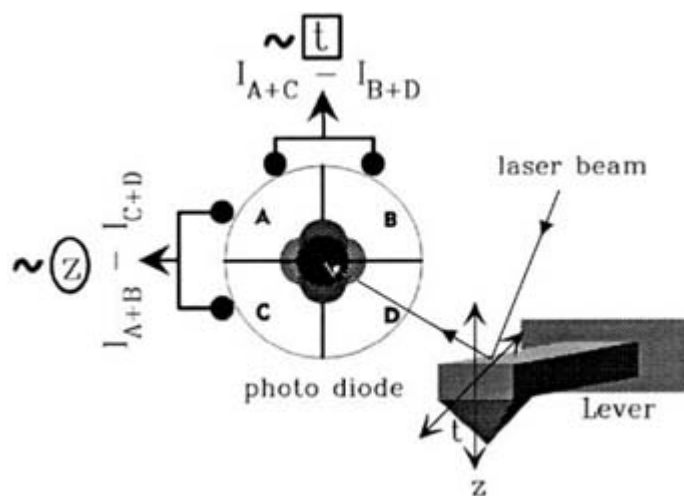


Figure B1.19.23. Principle of simultaneous measurement of normal and lateral (torsional) forces. The intensity difference of the upper and lower segments of the photodiode is proportional to the z-bending of the cantilever. The intensity difference between the right and left segments is proportional to the torsion, t , of the force sensor. (Taken from [110], figure 2.)

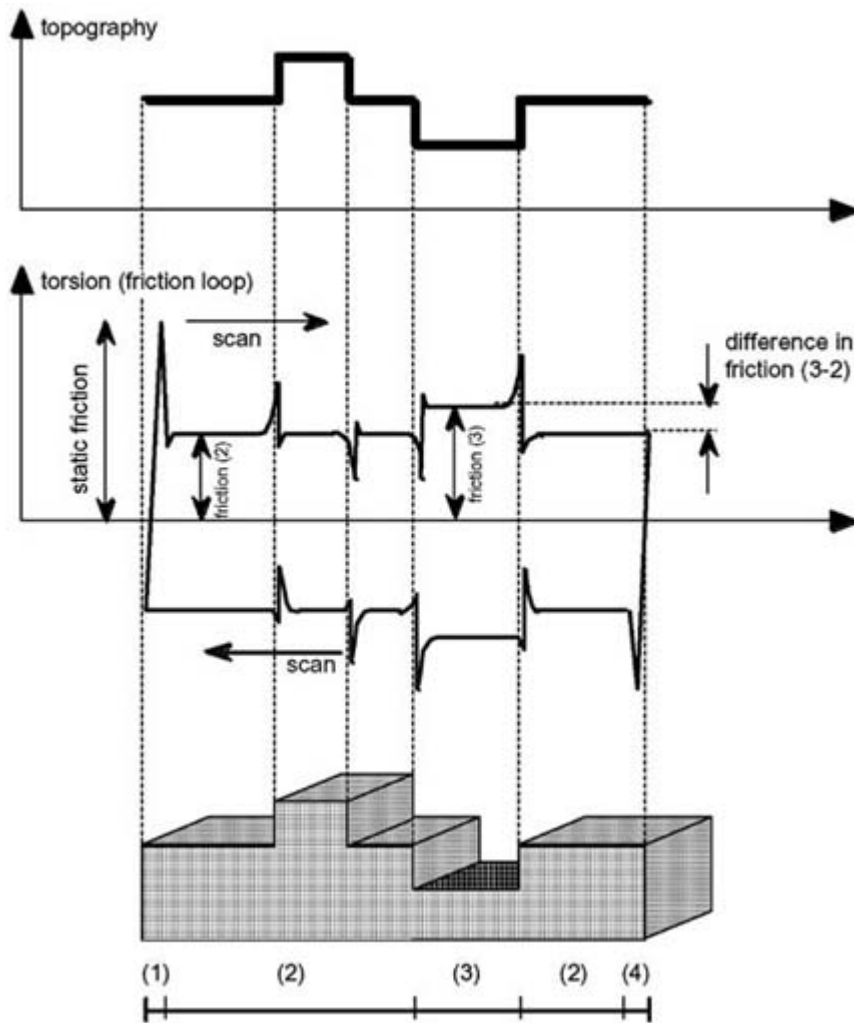


Figure B1.19.24. Friction loop and topography on a heterogeneous stepped surface. Terraces (2) and (3) are composed of different materials. In regions (1) and (4), the cantilever sticks to the sample surface because of static friction F_{ST} . The sliding friction is t_1 on part (2) and t_3 on part 3. In a torsional force image, the contrast difference is caused by the relative sliding friction, $\Delta F_{SL} = t_1 - t_3$. Morphological effects may be distinguished from frictional ones by their non-inverted behaviour upon scanning in the opposite direction. (Adapted from [110], figure 2.)

(H) MECHANICAL IMAGING WITH FORCE MICROSCOPY

In AFM, the relative approach of sample and tip is normally stopped after ‘contact’ is reached. However, the instrument may also be used as a nanoindenter, measuring the penetration depth of the tip as it is pressed into the surface of the material under test. Information such as the elastic modulus at a given point on the surface may be obtained in this way [114], although producing enough points to synthesize an elastic modulus image is very time consuming.

Pulsed-force mode AFM (PFM-AFM) is a method introduced for fast mapping of local stiffness and adhesion with lower required data storage than recording force–distance curves at each point on the x – y plane [115]. A sinusoidal or triangular modulation is applied between the tip and sample (either *via* lever or sample piezo) at a lower frequency than that of either the piezo or cantilever resonance frequency. Tip and sample then come

into contact for part of each oscillation cycle. The deflection signal of the cantilever is put into sample-and-hold circuits. The peak displacement of the lever during the approach cycle is usually chosen for the feedback signal to the piezo, in order to maintain a constant force during scanning. Other sampling points can be chosen at any arbitrary timings within one cycle depending on the required property. For example, the local stiffness can be calculated from the slope obtained from subtracting the sample-and-hold signals at two different points on the linear part of the deflection–displacement curve. The adhesion force can be calculated from the difference between the largest negative deflection signal and the zero-deflection point.

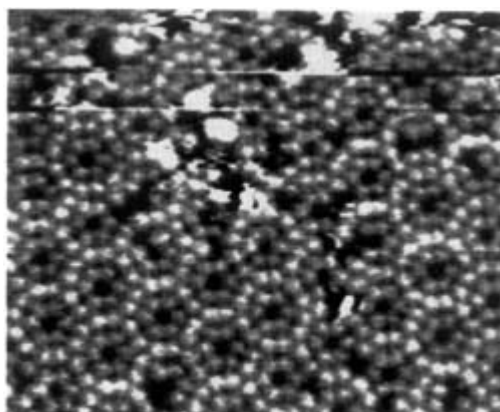
Another method developed for imaging mechanical properties is ultrasonic force microscopy [116] (UFM). This technique involves carrying out contact AFM while oscillating the sample at high frequency: typically 200–700 kHz, chosen to be above the highest tip–sample resonance. If operating in the purely contact mode (i.e. if the tip does not leave the surface), the amplitude of the tip oscillations is determined by the elastic modulus of the tip–sample system, independent of the spring stiffness of the cantilever. The technique can be thought of as a fast-indentation system, and it samples a volume that has a radius some ten times the indentation depth. The overall spatial resolution is typically a few nanometres.

B1.19.3.2 APPLICATIONS OF FORCE MICROSCOPY

(A) INORGANIC SURFACES

True atomic resolution has been obtained on a wide range of inorganic surfaces including metals, semiconductors and insulators. Initially, imaging concentrated on Si (111) 7×7 as a means of demonstrating the true-atomic-resolution imaging capability of the technique [81]. Even with such a well understood surface, surprising results were obtained in the form of additional contrast revealed between different surface atoms. In the case of Erlandsson *et al* [117] their results showed that centre adatoms appeared to be 0.13 Å higher than the corner adatoms. They suggest that the additional contrast may be due to variation in chemical reactivity of the adatoms or to tip-induced, atomic relaxation effects reflecting the stiffness of the surface lattice (figure B1.19.25). Nakagiri *et al* [118] also saw additional contrast in their images of Si(111) 7×7 . However, they observed the six atoms in one half of the unit cell to be brighter than in the other half. The two halves correspond to faulted and unfaulted halves of the unit cell according to the dimer–adatom stacking fault model [119]. At present they are not able to distinguish which atoms correspond to which half. The fact that this additional contrast varied depending on the precise experimental technique used indicates that different imaging mechanisms could be responsible as a result of the different tip material or height of the tip with respect to the surface.

-28-



40 Å

Figure B1.19.25. AFM image of Si(111)-(7 × 7) taken in the AC mode. Contrast can be observed between inequivalent adatoms. Image courtesy of R Erlandsson. (Taken from [217], figure 4.)

Of particular interest are those surfaces where AFM has provided complementary information or revealed surface structure which could not be obtained by STM. One obvious application is the imaging of insulators such as NaCl(001) [120]. In this case it was possible to observe point defects and thermally activated atomic jump processes, although it was not possible to assign the observed maxima to anion or cation.

Another area where AFM has provided new information is the imaging of metal oxides such as TiO₂ [121, 122]. Although the surface of TiO₂(110) is observable with STM, only NC-AFM was able to image the bridging oxygen rows which are the outermost atoms on the surface (figure B1.19.26). True-atomic-resolution imaging is still a relatively recent development and the full power of the technique in imaging insulators such as Al₂O₃ or SiO₂ has yet to be demonstrated.

-29-

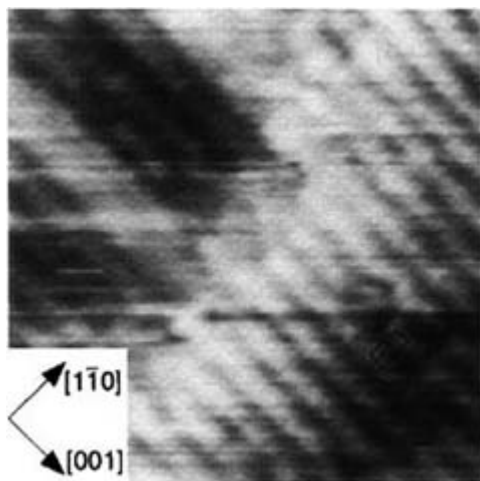


Figure B1.19.26. Highly resolved, non-contact AFM image of the TiO₂(110)-(1 × 1) surface (8.5 × 8.5 nm²) with a single step. The two-dimensional order of the bright spots (0.65 × 0.3 nm²) reproduces the alignment of the bridging oxygen atoms. (Taken from [121], figure 3.)

Even without atomic resolution, AFM has proved its worth as a technique for the local surface structural determination of a number of bio-inorganic materials, such as natural calcium carbonate in clam and sea-urchin shells [123], minerals such as mica [124] and molybdenite [125] as well as the surfaces of inorganic crystals, such as silver bromide [126] and sodium decatungstocerate [127]. This kind of information can prove invaluable in the understanding of phenomena such as biomineralization, the photographic process or catalysis, where the surface crystallography, especially the presence of defects and superstructures, can play an important role, but is difficult to determine by other methods. AFM has the considerable advantage that it can be used to examine powdered samples, either pressed into a pellet, if the contact mode is employed, or loosely dispersed on a surface, if intermittent or non-contact AFM is available.

AFM has also provided insights into the growth of metal clusters and films on mica. In the case of palladium [128], for example, it was found that clusters in the 50 nm range exhibited truncated triangular shapes. Epitaxial growth of silver on a mica surface [129] is seen to depend in a complex way on both substrate temperature and film thickness (figure B1.19.27), with island morphology giving way to channels, which become holes and then networks as the film thickness increases, the changes progressing more gradually as the substrate temperature is increased. The issue of surface roughness of silver films is central to the technique of surface-enhanced Raman spectroscopy, and AFM has been used to characterize films produced for this

purpose [130]. AFM possesses a considerable advantage over electron microscopy in this kind of application, in that it can, if calibrated, automatically yield *quantitative* morphological and roughness data [131].

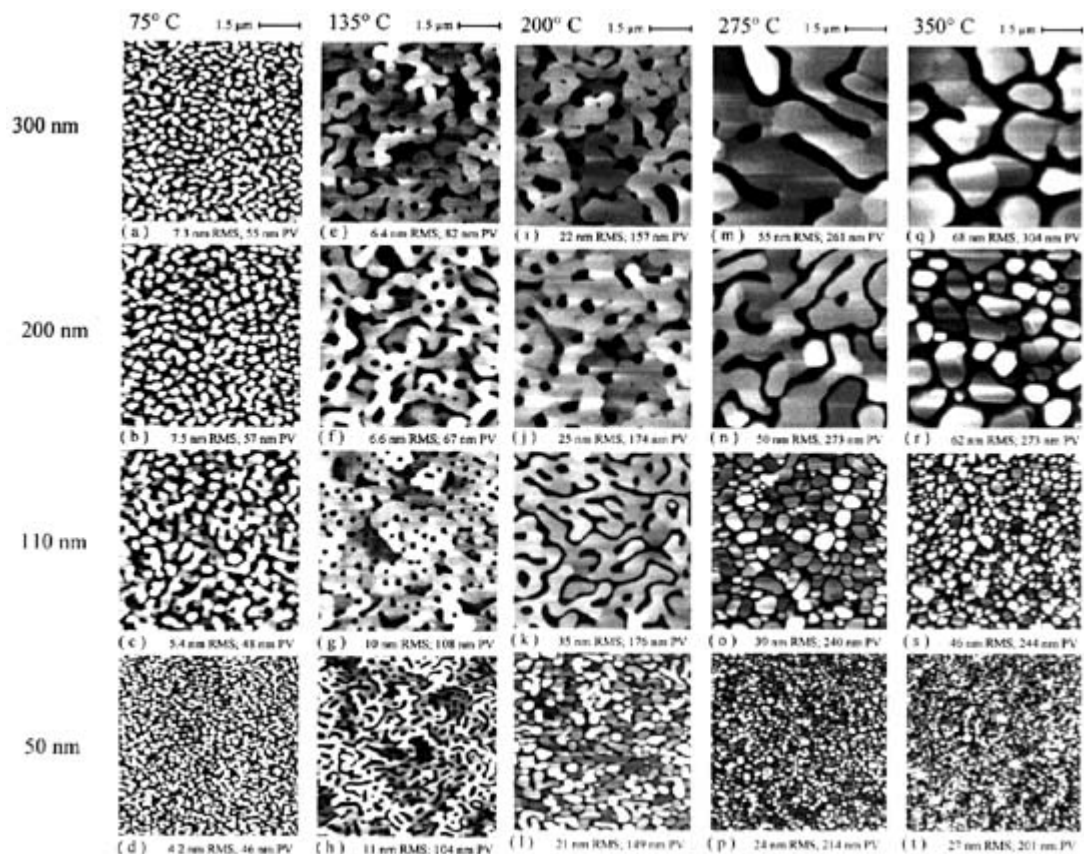


Figure B1.19.27. AFM topographic images ($7 \times 7 \mu\text{m}^2$) of 20 epitaxial Ag films on mica prepared at five substrate temperatures (75, 135, 200, 275, and 350 °C) and four film thicknesses (50, 110, 200, and 300 nm) using metal deposition rates of 0.1 to 0.2 nm s⁻¹. The vertical RMS and peak-to-valley roughness are indicated for each image. (Taken from [129], figure 1.)

(B) ORGANIC SURFACES

AFM has been used to image several surfaces of organic crystals—such as tetracene [132] and pyrene [133]—and has produced images that can be compared to the unit cells expected from previous x-ray crystallographic studies. In the case of tetracene, the surface is merely a truncation of that expected from the bulk data. However, in the case of pyrene, where the bulk consists of dimer pairs, a surface reconstruction is evident in the AFM image, corresponding to the presence of monomer species. Reconstructions are common phenomena in metal surfaces, where LEED has frequently been used to detect them [25]. LEED has scarcely been used to analyse organic crystal surfaces, however, due to problems associated with charging and/or degradation of the sample in an electron beam. AFM nicely fills this gap in the surface analytical arsenal.

The overwhelming majority of AFM studies on organic surfaces has concerned organic thin films on inorganic substrates and, in particular, those deposited *via* Langmuir–Blodgett or self-assembly processes [35]. These films

have been an active research area for several years, frequently serving as models for complex systems, such as membranes. Thin organic films are also being developed for their nonlinear optical properties, as microlithographic resists and as sensor components [32].

Two Langmuir–Blodgett film systems that have been much studied by AFM are the calcium and barium arachidates, adsorbed as double layers on a silicon substrate, which has been pretreated so as to be hydrophobic [32]. These structures are formed by dipping the silicon into an arachidate film on a water trough and then removing the substrate again through the film. Repeating the process simply adds another double layer to the structure. As with organic crystals, more traditional surface-structure-determining approaches are often too destructive to allow analysis of systems such as these; possibly the most important and unique aspect of AFM in this context is that it is a *local probe*, and therefore capable of showing the enormous variety of defects in the films, both on a microscale (pores and islands in the films) and a nanoscale [134, 135] (twinning, defects and complexities in molecular packing) (figure B1.19.28). This topic has been dealt with at length by Frommer in her excellent review article [32].

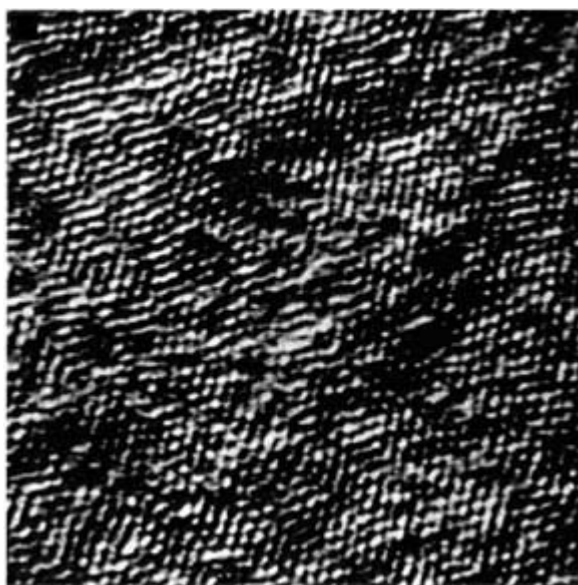


Figure B1.19.28. Molecular-scale image (2 nm × 20 nm) of a barium arachidate bilayer. Image was produced by averaging six images, but without filtering data. (Taken from [135], figure 1.)

Self-assembly of long-chain alkanethiols on the Au(111) surface has been studied by a number of different techniques including AFM, and it has been consistently shown that the molecules form a commensurate ($\sqrt{3} \times \sqrt{3}$)R30° structure. AFM studies can also provide additional information on the mechanical properties of the organic layers [136], which are interesting in that they serve as a model system for lubricants [137]. Above a critical applied load of 280 nN for the C₁₈ thiols, it has been found that the monolayers were disrupted, and that the subsequent image corresponded to that of the Au(111) substrate. However, on reducing the load to substantially below the critical value, the surface apparently healed, and the characteristic periodicity of the thiol overlayer returned. The exact way in which this phenomenon occurs is not completely understood [138]. Possibilities include displacement of the thiols by the tip, binding of the thiols to the tip, or desorption of the thiols into a liquid phase.

(C) POLYMER SURFACES

AFM is contributing significantly to our understanding of the surface structure of polymers, both on a microscale and on a molecular level. Segregation in the surface of block copolymers and polymer blends is often critical in determining technologically important properties, such as wettability or biocompatibility. It is also often difficult to image by optical or electron microscopy. AFM, on the other hand, offers a method for scrutinizing these materials down to the molecular level, without the need for surface preparation. AFM's ability to operate in a liquid environment makes it particularly useful in analysing polymers for medical applications, since these materials are designed to function surrounded by body fluids, which can influence the surface microstructure and nanostructure.

Several studies have concerned the microstructure of lamellae in materials such as the block copolymers polystyrene-*block*-poly-2-vinylpyridine [139] and polystyrene-*block*-polybutadiene [140], as well as single crystals of poly-*para*-xylylene [139], and reveal features (such as intersecting lamellae (figure B1.19.29)) that had not been previously observed.

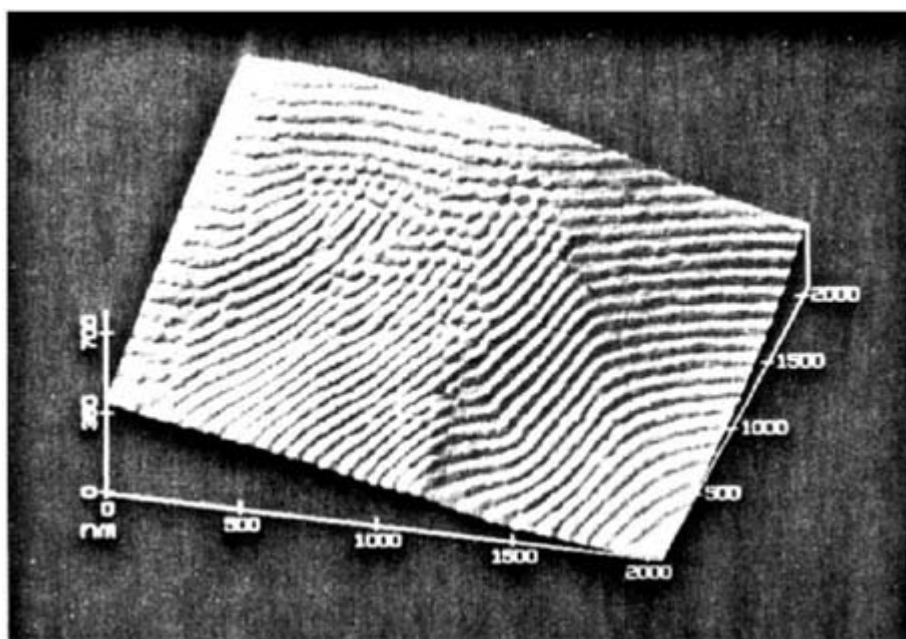


Figure B1.19.29. AFM image of polystyrene/polybutadiene copolymer, showing lamellar structure. (Taken from [140], figure 1.)

LFM has also proved useful in the examination of polymer blends, since its ability to image effective frictional coefficients imparts a certain chemical sensitivity to the method [141]. A novel approach to discrimination between components of a polymer blend was adopted by Feldman *et al* [142], who used the low-refractive-index liquid, perfluorodecalin, as a medium for measurement of tip-polymer 'pull-off', thereby enhancing the London component of the Hamaker constant, improving the signal-to-noise ratio of the measurement. Using the same medium for lateral force (frictional) imaging, a striking contrast reversal was found between polystyrene and PMMA blend components, when the tip surface was changed from (hydrocarbon-covered) gold to silica (figure B1.19.30).

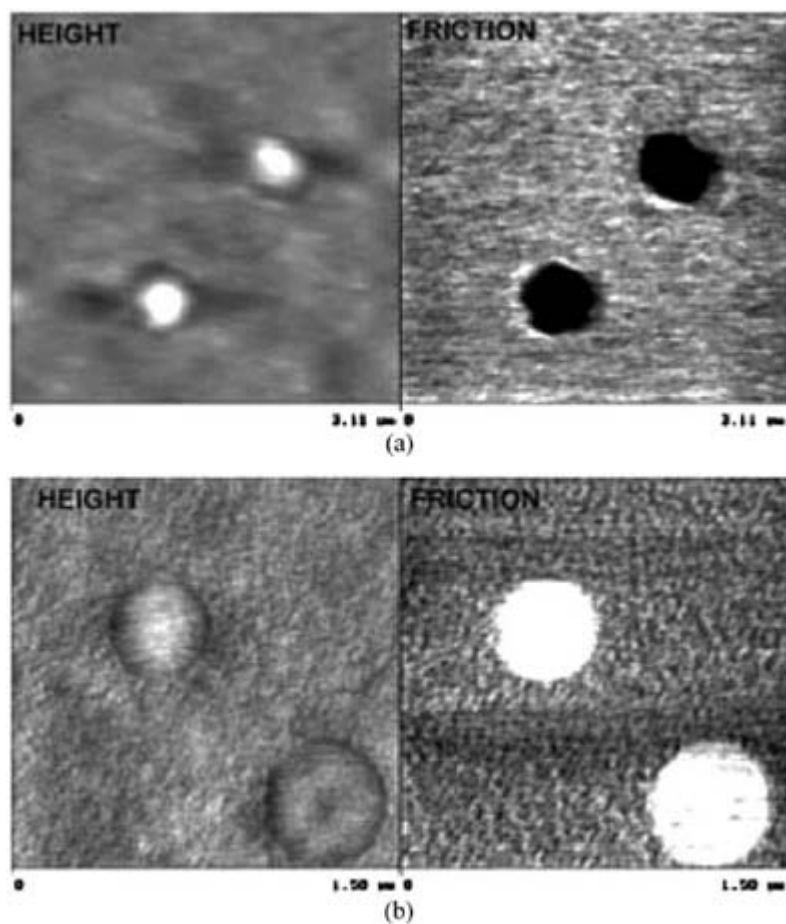


Figure B1.19.30. Height and friction images of a spin-cast polystyrene-poly(methyl methacrylate) blend obtained with (a) gold and (b) silica probes under perfluorodecalin. Note the reversal of frictional contrast and the high spatial resolution. (Taken from [142], figure 7.)

On the molecular level, spectacular AFM images have been obtained for a number of systems. In the case of isotactic polypropylene, for example, Snétiy and Vancso [143] have succeeded in imaging individual methyl groups on the polymer chain, and distinguishing between left- and right-handed helices in the crystalline *i*-polypropylene matrix (figure B1.19.31). The same group has also used AFM to image the phenylene groups in poly(*p*-phenyleneterephthalamide) fibres, and have used this data to show the existence of a new polymorphic form that had previously only been suggested by computer simulations.

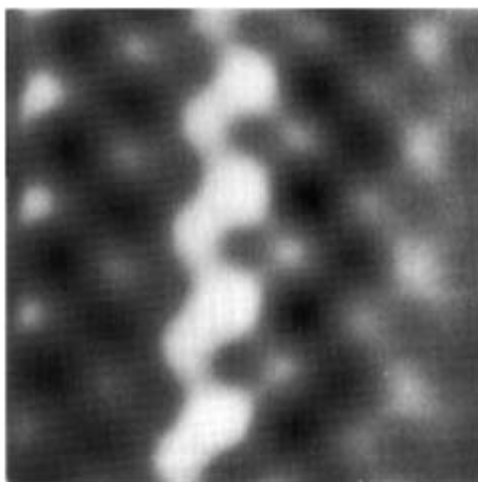


Figure B1.19.31. AFM image of a 2.7 nm × 2.7 nm area of a polypropylene surface, displaying methyl groups and right- and left-handed helices. (Taken from [143], figure 10.)

(D) BIOLOGICAL SURFACES

The AFM is now firmly established as a unique tool for the *in situ* investigation of biological surfaces [144], whether these be biomolecules, cell structures, or even viruses [145]. Often, a successful immobilization strategy has been key to successful imaging of the biological surface [146].

Lured by the promise of a new way to sequence the genetic code, and the prospect of nanomanipulation of nucleic acids, investigators have produced a plethora of papers in the area of AFM imaging of DNA. Notable among these is the work of Bustamente *et al* [147], who developed one of the first reproducible methods for imaging nucleic acids. Their approach involved three important components: (1) the use of extremely sharp tips [148] (radius of curvature ≈ 10 nm), which are prepared by electron-beam deposition of a carbon whisker on the tip apex in an electron microscope [149], (2) the use of mica substrates that have been ion-exchanged with magnesium, in order to promote interaction with the phosphate groups on the DNA and (3) careful control of the relative humidity [150] (or operation under liquids), in order to prevent tip-induced sample movement. Using this approach, and imaging under 2-propanol, high-resolution images of plasmid DNA have been obtained [151] (figure B1.19.32), and the molecules dissected by momentarily increasing the AFM force [152]. Single- and double-stranded DNA [153] and RNA-polymerase-DNA complexes [154] have also been imaged using this approach, the helical pitch of the DNA deciphered [155], AFM used to determine local chirality of the DNA supercoiling [156] and even individual base pairs resolved [157].

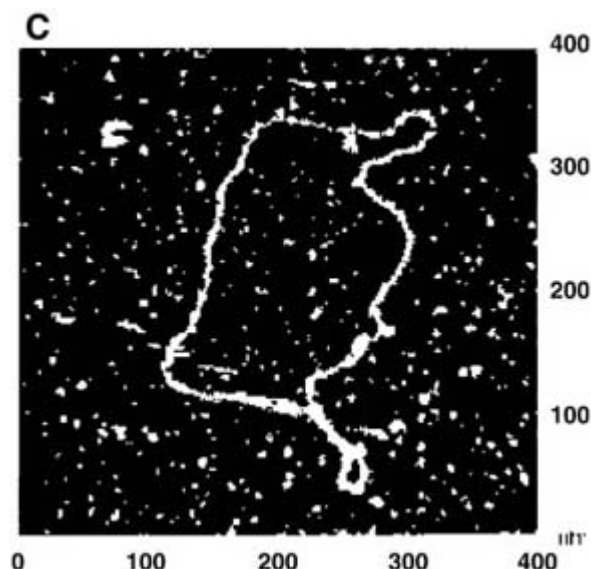


Figure B1.19.32. AFM image of Blue Script II plasmid (400 nm × 400 nm) in propanol, taken with ‘super tip’, prepared by carbon deposition on normal tip in SEM, followed by ion milling. (Taken from [152], figure 1.)

Proteins have also been investigated extensively by AFM. Radmacher *et al* [158] monitored height changes in an adsorbed layer of lysozyme using intermittent-contact AFM, as the enzyme was exposed to a substrate molecule. The height changes were variously interpreted as due to conformational adaptations of the lysozyme, or to the different height of the enzyme–substrate complex. Many other proteins have been imaged, using various immobilization methods, and this area has been comprehensively reviewed in the literature [159]. Among the highest-resolution (<1 nm) examples have been those of Müller *et al* [160], who have studied the inner surface of the hexagonally packed intermediate (HPI) layers of cell envelope proteins, such as those in the bacterium *Deinococcus radiodurans*, where protein conformational changes can be observed (figure B1.19.33). AFM has also provided a new window into the channels present in the surfaces of living cells, and Lal *et al* [161] have imaged the channels formed when the responsible cell proteins (porins) are reconstituted as crystalline arrays. The resolution obtained was such that individual polar head groups of the lipid molecules could be discerned. Several mechanical studies on proteins using AFM have also been reported, notably the hysteretic unfolding of the giant muscle protein, titin, reported by Rief *et al* [162], where the importance of this protein as a ‘strength reservoir’ during muscle stretching was demonstrated for the first time.

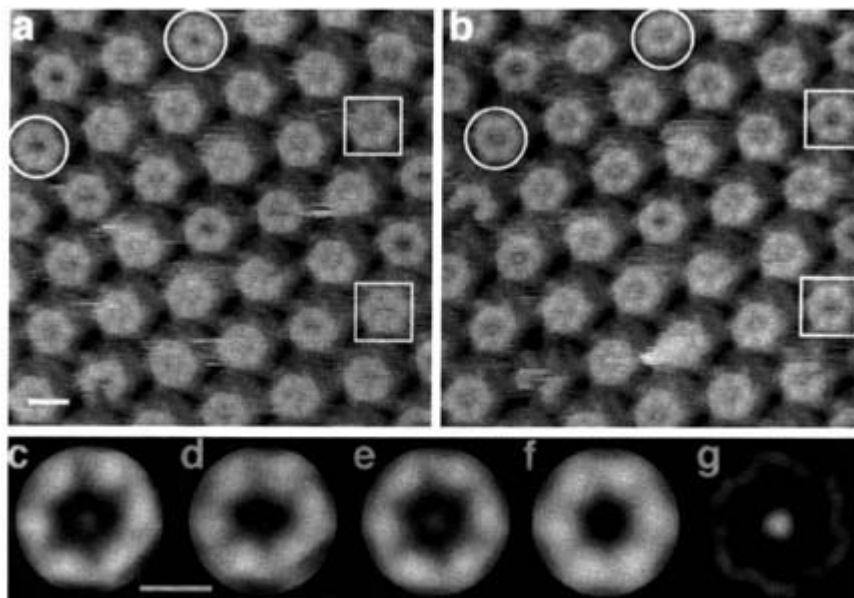


Figure B1.19.33. Conformational changes of the inner surface of an HPI layer. (a) The protruding cores are clearly visible, with some pores in an open conformation and others in an obstructed conformation. (b) Area shown in panel a imaged 5 min later. Some pores that were open earlier are now closed (circles), while closed ones have opened during this time interval (squares). Units were aligned and divided into two classes. The class averages exhibit a plugged (c) and an open hexamer (d). The difference map (g) represents the modulus of the height difference between the sixfold-symmetrized class averages ((e) and (f)). The full grey-level range corresponds to a vertical distance of 6 nm ((a) and (b)) and 3 nm ((c) to (f)). (Taken from [160], figure 3.)

In the area of cell and cell-structure imaging, AFM offers an advantage over optical and scanning electron microscopies in that it permits high-resolution imaging of living cells, and even the observation of dynamic phenomena. Henderson *et al* [163] observed the motion of filamentous actin in living glial cells (i.e., structures beneath the plasma membrane were imaged), and even performed nanosurgery on the cell with the AFM tip. Brandow *et al* [164] also cut into lipid membranes on a graphite surface using deliberately high force from the AFM tip, and found that the membranes healed themselves after sufficient time, but that the healing could be accelerated by rubbing the AFM tip perpendicular to the cut with a controlled force. The same group was also able to manipulate living glial cells [165] and even peel them away from a surface, if the normal force was appropriate. Thus, depending on the amount of force applied, the AFM could be used to cut, anneal, peel or image the sample.

Several groups have focused on the biochemical receptor–ligand interaction, measuring forces between individual molecular binding pairs, such as biotin and streptavidin [166, 167]. One approach involves coating the tip with either receptor or ligand, the sample with the complementary molecule, and then carefully monitoring tip–sample separation after contact. If the cantilever stiffness is appropriately chosen, the force–distance curve during separation appears to be ‘quantized’ in units corresponding to a single molecular-pair interaction. Lee *et al* [168] have also demonstrated and measured the interaction between single complementary strands of DNA base pairs using AFM. Several other examples of ligand–receptor interactions have been summarized in a review by Bongrand [169].

Due to its sensitivity to small forces and its ability to operate in liquids, the AFM has opened up a new avenue of investigation into colloidal systems. A frequently used approach, developed by Ducker *et al* [170], involves the cementing of a colloidal-sized particle onto an AFM cantilever in place of the usual tip, and then monitoring interactions with some appropriate flat surface, or even another colloidal particle [171], under a variety of conditions. Larson *et al* [172] used this technique to investigate the interactions between a titania (rutile) particle ($\approx 9 \mu\text{m}$ diameter) and a rutile single-crystal surface under various conditions of pH and ionic strength. From their experiments they were able to measure the van der Waals interaction between rutile surfaces directly, and to calculate the non-retarded Hamaker constant for the system. Biggs *et al* have applied a similar approach to the venerable subject of gold colloid stability [173], by immobilizing a $\sim 6 \mu\text{m}$ gold sphere on a cantilever and measuring its interaction with a polished gold plate under solutions containing combinations of gold, citrate, chloride and a number of other ions of relevance to the colloid system. The authors were able to demonstrate the presence of a repulsive interaction between the gold surfaces due to adsorbed citrate or chloride; they showed that citrate adsorbed preferentially, and succeeded in measuring the surface potential of the gold as a function of anion concentrations.

(F) TRIBOLOGY

Force microscopy, and lateral (or frictional) force microscopy in particular, are having a tremendous impact in tribology—the science and technology of friction, wear, and lubrication. The interaction between moving surfaces (the central issue in tribology) is thought to consist of separate interactions between the many peaks in one surface with the many peaks in the countersurface. These peaks are known as asperities. It is this mode of interaction that leads to Amontons' empirical 'law' of friction (this law is generally attributed to Amontons, although initially observed by Leonardo da Vinci a century earlier), $F = \mu N$, where F is the frictional force, N is the normal force, and μ is the coefficient of friction. Notable by its absence in this equation is the *apparent* contact area between the two sliding bodies. In fact, as one might intuitively believe, the *actual* contact area, A , between the asperities is all-important, and the frictional force is proportional to A . However, A , in turn, increases in proportion to the normal force [174], (both due to 'flattening' of the asperities by the load and the creation of new load-bearing asperities, as the higher asperities are flattened) so that A can be cancelled out of the resulting equation, leaving behind only the measurable quantities, F and N . The importance of AFM in fundamental tribology research is that the AFM measurement can be thought of as a single asperity contact, i.e. the fundamental interaction in frictional behaviour.

Carpick *et al* [84] used AFM, with a Pt-coated tip on a mica substrate in ultrahigh vacuum, to show that if the deformation of the substrate and the tip–substrate adhesion are taken into account (the so-called JKR model [175] of elastic adhesive contact), then the frictional force is indeed proportional to the contact area between tip and sample. However, under these single-asperity conditions, Amontons' law does not hold, since the 'statistical' effect of more asperities coming into play no longer occurs, and the contact area is not simply proportional to the applied load.

Mate *et al* [111], who pioneered LFM, used their instrument to show atomic-scale structure in the frictional force between a tungsten tip and surface graphite atoms. In fact, what these authors observed was stick–slip behaviour: an effect normally associated with macroscopic phenomena, such as the vibration induced in a violin string by the bow, or the squealing of an automobile's brakes. They also found that the frictional coefficient varied slightly with applied load, i.e. a deviation from Amontons' law. The same group went on to study mica surfaces [124], where a similar stick–slip behaviour was observed, the frictional coefficient varying with the unit cell periodicity of the mica cleavage plane (figure B1.19.34). Hu *et al* [176] examined mica surfaces by LFM using silicon nitride tips, and found that above

normal loads of 10 nN, Amontons' law was obeyed, while at low loads, deviations from linearity were observed. They also showed that friction decreased substantially as a function of humidity, decreasing by an order of magnitude under liquid water, but was essentially invariant with scanning direction across the mica surface. Wear was also observed by these authors, who found that a threshold load value needed to be exceeded before layer-by-layer wear was initiated. Below this value, even multiple scans were not found to produce visible damage to the surface. A similar effect was observed for the system AgBr on NaCl(001) (SiO₂-coated tip) by Lüthi *et al* [177], where a wear onset was observed at around 14 nN. Interestingly, frictional coefficients were measured over a range of loads in this latter work and the μ measured on NaCl was found to be an order of magnitude lower than that measured on AgBr.

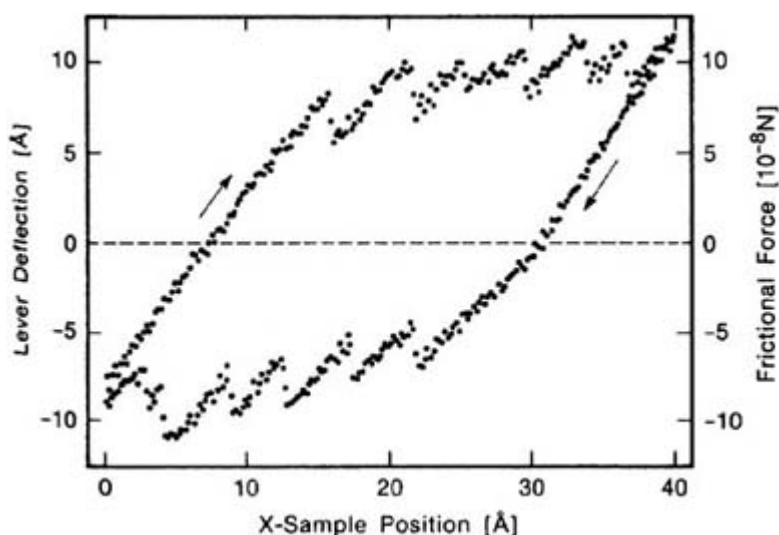


Figure B1.19.34. Cantilever deflection and corresponding frictional force in the x -direction as a function of sample position as a mica sample is scanned back and forth under a tungsten tip. (Taken from [124], figure 2.)

Numerous groups have applied LFM to address the issue of friction on thin organic films. These systems serve as useful models for the important macroscopic tribological issue of boundary lubrication. Overney *et al* [178] used LFM to examine mixed Langmuir–Blodgett films of perhydro arachidic acid and partially fluorinated carboxylic acid that had been transferred onto a silicon surface as a bilayer system, using poly(4-vinyl-N-methylpyridinium) as a counteraction. The images clearly showed a difference in frictional coefficient between the two components, which segregated into submicron domains. Surprising, however, was the observation that the apparent frictional coefficient on the fluorinated component was a factor of four higher than that measured on the non-fluorinated one. The authors attributed this to the greater shear strength of fluorinated films, although more recent measurements using LB-deposited straight-chain acids of different lengths [179] suggest that the length of the molecule itself has a significant effect on mechanical properties, and therefore on the frictional coefficients measured.

Kim *et al* [180], using specially synthesized, end-functionalized alkanethiols, investigated mechanisms of friction by producing gold-supported monolayers containing varying quantities of various bulky endgroups. They found that the differences in friction were apparently due to differences in the size of the terminal groups, larger terminal groups (whether F-containing or not) giving rise to increasing interactions that provided pathways for energy dissipation, and therefore higher frictional losses.

The issues of the correlation of adhesion and of viscoelastic relaxation with friction are currently being investigated using AFM and LFM. Although friction does not correlate with the adhesion energy between two

surfaces, there is increasing evidence that it is proportional to adhesion *hysteresis* [181]: i.e., the energy dissipated during the making and breaking of an adhesive bond. Marti *et al* [189] have shown that the adhesion hysteresis measured during force-curve measurements (Si_3N_4 tip) on silica and alumina surfaces under electrolytes is proportional to the LFM friction force measurements made under the same conditions. Haugstad *et al* [182] characterized fundamental aspects of friction on polymer surfaces by measuring frictional force on a gelatin film as a function of scanning frequency, i.e. velocity. The friction was found to increase at lower velocities, which can be explained by the onset of rubbery-state-type molecular relaxations, the viscoelastic dissipation correlating with frictional dissipation. Following intensive scanning at a particular site on the gelatin surface, a high-friction ‘signature’ could be observed in subsequent LFM measurements, and was explained by the perturbative effect of the previous frictional measurement. In this case, the LFM images were providing a map of molecular relaxation across the film.

(G) MEASUREMENT OF MECHANICAL PROPERTIES

The technological importance of thin films in such areas as semiconductor devices and sensors has led to a demand for mechanical property information for these systems. Measuring the elastic modulus for thin films is much harder than the corresponding measurement for bulk samples, since the results obtained by traditional indentation methods are strongly perturbed by the properties of the substrate material. Additionally, the behaviour of the film under conditions of low load, which is necessary for the measurement of thin-film properties, is strongly influenced by surface forces [75]. Since the force microscope is both sensitive to surface forces and has extremely high depth resolution, it shows considerable promise as a technique for the mechanical characterization of thin films.

A striking example of the use of the AFM as an indenter is provided by Burnham and Colton [184] (figure B1.19.35), where the differences (at <100 nm indentation) between the plastic behaviour of gold at $20 \mu\text{N}$ load and the elastic behaviour of graphite and elastomer at lower loads are clearly observable. The importance of surface forces in measuring mechanical properties at low load has been demonstrated by Salmeron *et al* [184], who showed that tip-sample adhesion can seriously perturb hardness measurements, not only for clean surfaces, but also for those covered by a layer of contamination. These authors also suggested that initial passivation of the surface (e.g., by sulfidation) might be an effective approach to overcoming these artefacts.

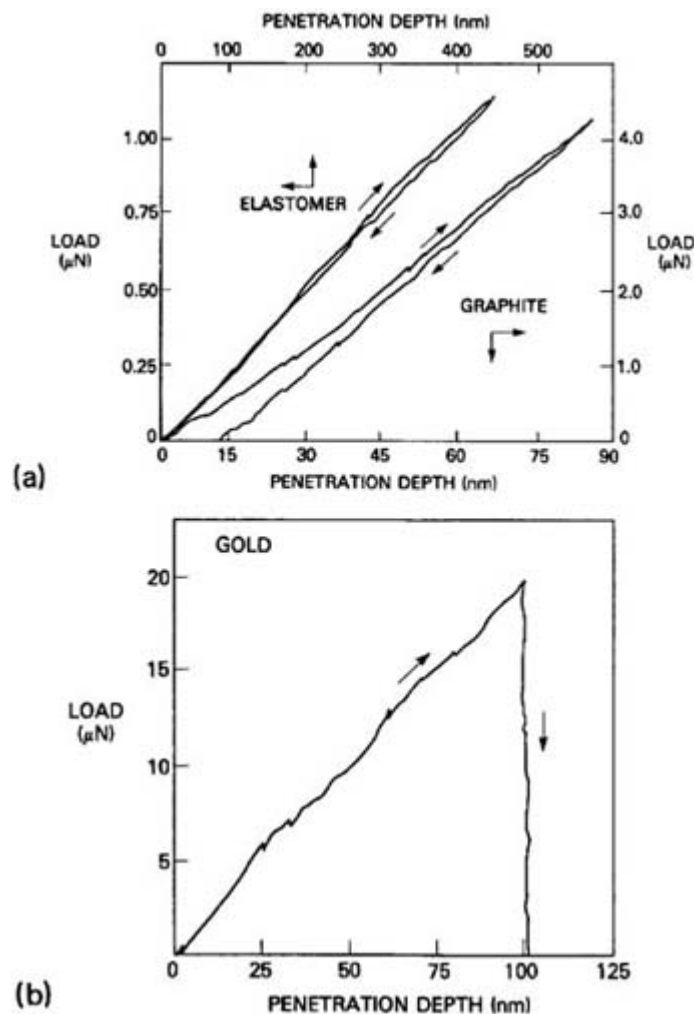


Figure B1.19.35. Experimental nanoindentation curves obtained with the AFM showing the loading and unloading behaviour of (a) an elastomer and highly oriented pyrolytic graphite and (b) a gold foil. (Taken from [183], figure 4.)

(H) CHEMICAL IMAGING

The imaging of surfaces according to chemical species is immensely useful in understanding the mechanisms of many complex technological systems on a fundamental level. Over the past two decades, Auger electron spectroscopy, x-ray photoelectron spectroscopy and secondary ion mass spectroscopy have been extended into surface-sensitive chemical imaging methods [185]; nowadays the use of such techniques for troubleshooting has become virtually routine in the semiconductor industry, and has contributed significantly to our knowledge of catalyst systems, corrosion mechanisms and many other areas. However, these methods are not universally applicable since they are limited in spatial resolution (especially for insulating samples) and require the sample to be analysed in a vacuum. A chemically sensitive scanning force microscopy that could image the distribution of chemical species on an insulating surface with nanometre resolution under ambient conditions, or even under liquids, is therefore a highly desirable goal, and many research groups are active in this area.

A promising approach, still in the early stages of development, involves the functionalization of the AFM tip.

In many cases this has been limited to the demonstration of the sensitivity of force curves to the chemical species present on the surface—a necessary first step for the development of an imaging methodology. Akari *et al* [186] have functionalized a gold-coated tip with COOH-terminated long-chain hydrocarbons, using a self-assembled monolayer procedure; they have shown that the contrast is considerably enhanced over that of the uncoated tip, when imaging a patterned mixed monolayer consisting of CH₃- and COOH-terminated molecules of similar length.

LFM coupled with tip functionalization is a potentially important chemical imaging technique, since the tip functionality can be tailored so as to produce maximum contrast (i.e. maximum difference in frictional coefficient) between different chemical species on the surface. Frisbie *et al* [187] have examined a patterned surface of COOH and CH₃ groups, and have shown that the pattern could be readily imaged by a COOH- or CH₃-coated tip (figure B1.19.36) running in LFM mode, since the imaged frictional coefficients depended on the particular tip–surface species interaction. A potential pitfall with this technique is that both local chemistry and local mechanical properties, due to differences in molecular packing, contribute to the imaging contrast. This problem was elegantly sidestepped by McKendry *et al* who investigated chiral discrimination by chemical force microscopy [188]. In this case the interaction between the tip and surface can be changed without alteration of the mechanical properties or wetting energies of tip or surface. The chemical force microscope proved to be sufficiently sensitive to permit discrimination between enantiomers of simple chiral molecules in a friction image with more quantitative differences being obtained from adhesion force or ‘pull-off’ force histograms.

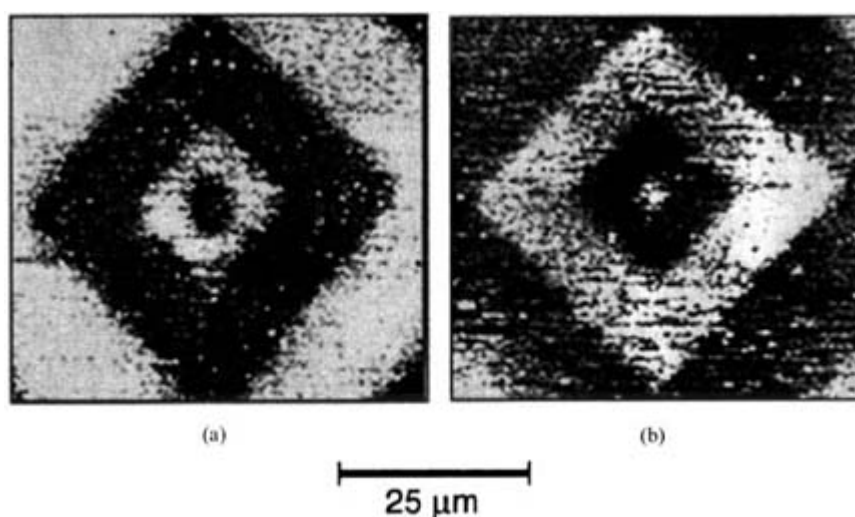


Figure B1.19.36. Image of the frictional force distribution of a pattern consisting of areas of CH₃-terminated and areas of COOH-terminated molecules attached to gold-coated silicon. The tip was also functionalized in (a) with CH₃ species and in (b) with COOH species. The bright regions correspond to the higher friction force, which in (a) is observed on the CH₃ areas and in (b) on the COOH areas. (Taken from [187], figure 3.)

One potentially powerful approach to chemical imaging of oxides is to capitalize on the tip–surface interactions caused by the surface charge induced under electrolyte solutions [189]. The sign and the amount of the charge induced on, for example, an oxide surface under an aqueous solution is determined by the pH and ionic strength of the solution, as well as by the isoelectric point (IEP) of the sample. At pH values above the IEP, the charge is negative; below this value,

the charge is positive. The same argument applies to a Si₃N₄ tip (normally an oxynitride at the surface), so

that at every pH, either an attractive or a repulsive electrostatic tip–sample force will be superimposed upon the forces that are normally encountered in AFM (figure B1.19.37). By varying the pH and determining the value at which the charges switch sign, the IEP of the sample may be determined. Since this value is characteristic for a particular oxide, and the sample area probed is on the nanometre-squared scale, this appears to be a promising direction for the high-resolution chemical imaging of mixed oxide, oxide-covered alloy [190] or even protein [191] surfaces.

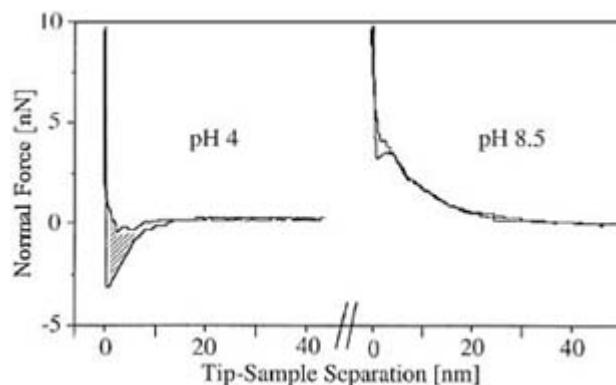


Figure B1.19.37. Normal force versus tip–sample distance curves for a Si_3N_4 tip on a SiO_2 surface under 1 mM NaCl solution at pH 4 and pH 9. (Taken from [189], figure 2.)

Finally, Berger *et al* [192] have developed a technique whereby an array of force curves is obtained over the sample surface (‘force-curve mapping’), enabling a map of the tip–sample adhesion to be obtained. The authors have used this approach to image differently oriented phase domains of Langmuir–Blodgett-deposited lipid films.

B1.19.4 SCANNING NEAR-FIELD OPTICAL MICROSCOPY AND OTHER SPMS

Since the invention of the scanning tunnelling microscope, many other related techniques have been developed that combine the principles of piezo positioning with a feedback system, but rely on a surface interaction other than electron tunnelling or force sensing in order to produce images. The overwhelming majority of these scanning techniques are still in the very early stages of development and, in contrast to STM and AFM, have as yet revealed little that could not be better determined by other methods. Nevertheless, this is an area with tremendous promise, and a selection of these methods is therefore described below.

B1.19.4.1 SCANNING NEAR-FIELD OPTICAL MICROSCOPY

Of the methods described in this section, scanning near-field optical microscopy (SNOM or NSOM) is the closest to being able to provide useful information that is unobtainable by other means. Indeed, this technique has already been made available as a commercial instrument. A detailed review of SNOM has been written by Pohl [193].

While the spatial resolution in classical microscopy is limited to approximately $\lambda/2$, where λ is the optical wavelength (the so-called Abbé Limit [194], $\sim 0.2 \mu\text{m}$ with visible light), SNOM breaks through this barrier by monitoring the *evanescent* waves (of high spatial frequency) which arise following interaction with an

object, rather than the *propagating* waves (of low spatial frequency), which are observed under far-field conditions. While the field intensity of propagating waves decays with the well known inverse-square dependence on distance from the source, evanescent waves decay much more rapidly, with an inverse-fourth power relationship. This means that the evanescent waves are almost completely damped out within a few nanometres of the source, and play no part in far-field (i.e. classical) optical measurements. Synge [195] was the first to discuss the possibility of exceeding the Abbé limit, as long ago as 1928, and suggested that by scanning a nanometre-sized aperture over the surface of the sample, a resolution higher than $\lambda/2$ could be obtained. This is analogous to the use of a stethoscope by a physician, where spatial information can be obtained with a resolution far greater than the acoustical wavelength [196]. This forms the basis of SNOM. In 1972, Ash *et al* [197] were able to show that this principle could indeed be demonstrated for microwave wavelengths (3 cm), but SNOM with visible light was not developed until the 1980s, when the technologies surrounding the STM became available.

The design of the SNOM in the first experiments consisted of a minute aperture, formed by a metallized glass fibre tip, which was rastered across a sample that was illuminated by a laser beam from behind (figure B1.19.38). The tip was maintained at a constant distance from the sample by using the tip-sample tunnelling current in a feedback loop. The resolution obtained was 25 nm ($\lambda/20$) [196, 198]. Subsequent variations to the experimental set-up have included the use of force interactions to maintain tip-sample separation [199] (enabling the imaging of insulators), as well as operation in transmission mode [200]. Recent applications have included the detection and spectroscopy of single molecules [201], where spectral differences corresponding to the different chemical environments of individual molecules can be discerned, and the orientation of each molecular dipole determined. In general, the combination of high lateral resolution in the optical near-field and spectroscopic information has been restricted to fluorescence and luminescence experiments. Recently, vibrational spectroscopy in the optical near-field has also been attempted [202, 203, 204 and 205] with surface-enhanced Raman scattering in particular, appearing to be a promising method for obtaining spectral, spatial and chemical information of molecular adsorbates with subwavelength lateral resolution. This has been implemented in illumination mode, where the incident light emerges from the fibre probe, to obtain surface-enhanced Raman spectra from single Ag colloid nanoparticles [206]. Raman chemical imaging on a scale of 100 nm has also been demonstrated by Deckert *et al* on dye-labelled DNA. On the nanometre scale a strong dependence of the enhanced Raman signal on substrate morphology is expected. It is therefore particularly useful to correlate near-field Raman spectra with topographic information on a nanoscale, as in the experiment of Zeisel *et al* [207] who investigate the near-field surface-enhanced Raman spectroscopy of dye molecules adsorbed on silver island films.

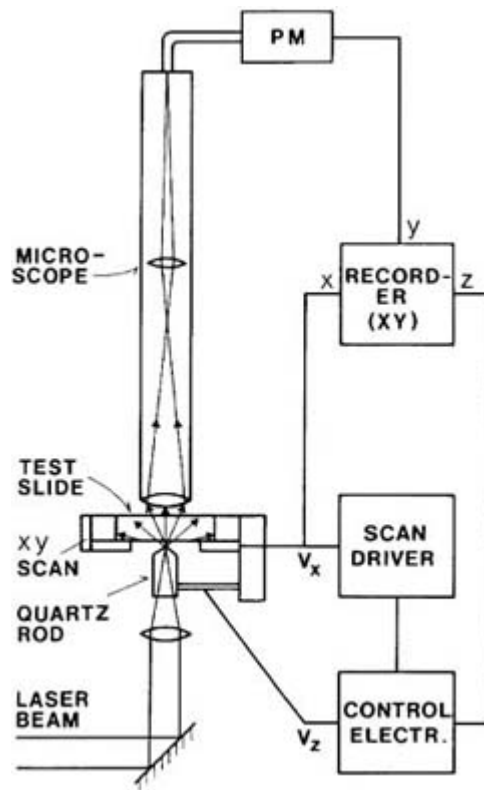


Figure B1.19.38. Schematic of a scanning near-field optical microscope (SNOM). (Taken from [196], figure 2.)

B1.19.4.2 SCANNING NEAR-FIELD ACOUSTIC MICROSCOPY (SNAM)

This corresponds to the physician's stethoscope case mentioned above, and has been realized [208] by bringing one leg of a resonating 33 kHz quartz tuning fork close to the surface of a sample, which is being rastered in the x - y plane. As the fork-leg nears the sample, the fork's resonant frequency and therefore its amplitude is changed by interaction with the surface. Since the behaviour of the system appears to be dependent on the gas pressure, it may be assumed that the coupling is due to hydrodynamic interactions within the fork-air-sample gap. Since the fork tip-sample distance is approximately $200\ \mu\text{m}$ ($\sim\lambda/20$), the technique is sensitive to the near-field component of the scattered acoustic signal. $1\ \mu\text{m}$ lateral and $10\ \text{nm}$ vertical resolutions have been obtained by the SNAM.

B1.19.4.3 SCANNING THERMAL PROFILER (STP)

This technique involves the scanning of a heated thermocouple tip above the surface of the sample (figure B1.19.39). Since the heat loss from the tip is highly dependent on the tip-sample spacing, the temperature of the tip can be used as a control parameter to monitor sample morphology and/or thermal properties [209]. The lateral resolution obtained with this method is of the order of $0.1\ \mu\text{m}$, with a vertical resolution of about $3\ \text{nm}$. The temperature sensitivity of the tip is $\sim 0.1\ \text{mK}$. An advantage of the technique is that morphology can be imaged at distances approximately equal to the desired lateral resolution.

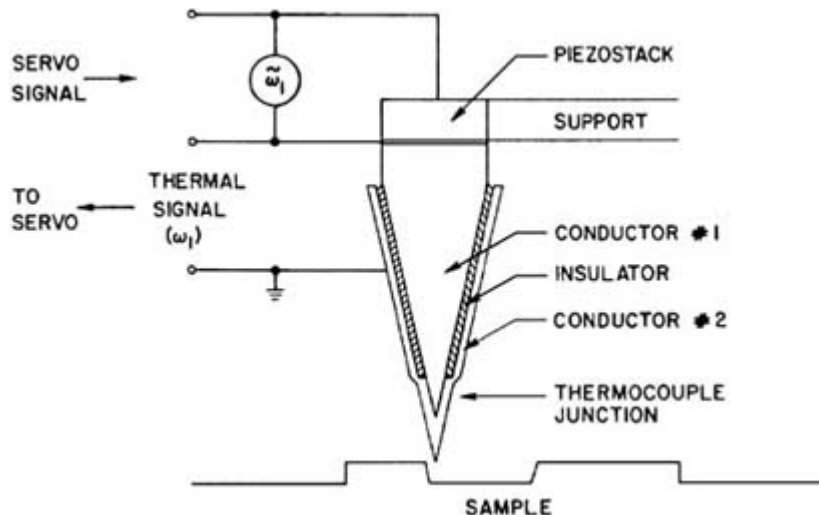


Figure B1.19.39. Schematic of the thermocouple probe in a scanning thermal profiler. The probe is supported on a piezoelectric element for modulation of tip–sample distance at frequency ω_1 and to provide positioning. The AC thermal signal at ω_1 is detected, rectified, and sent to the feedback loop, which supplies a voltage to the piezostack to maintain the average tip–sample spacing constant. (Taken from [209], figure 1.)

An extension of this technique, known as scanning thermal microscopy [210] (S_{Th}M) combines the thermal profiling technique with modulated-temperature differential scanning calorimetry, by applying a sinusoidal modulation to the tip temperature. Using this technique, the spatial distribution of thermal properties of materials (such as conductivity and diffusivity) can be monitored, and subsurface features imaged (since the penetration depth of the evanescent temperature wave is frequency-dependent). Additionally, local calorimetric analysis can be used to probe thermally activated near-surface processes, such as glass transitions, melting, crystallization or cure reactions. The technique has been used to great effect with polymer blends, where imaging contrast is caused by differences in the thermal properties of the individual components.

B1.19.4.4 SCANNING ION CONDUCTANCE MICROSCOPY (SICM)

This method relies on the simple principle that the flow of ions into an electrolyte-filled micropipette as it nears a surface is dependent on the distance between the sample and the mouth of the pipette [211] (figure B1.19.40). The probe height can then be used to maintain a constant current flow (of ions) into the micropipette, and the technique functions as a non-contact imaging method. Alternatively, the height can be held constant and the measured ion current used to generate the image. This latter approach has, for example, been used to probe ion flows through channels in membranes. The lateral resolution obtainable by this method depends on the diameter of the micropipette. Values of 200 nm have been reported.

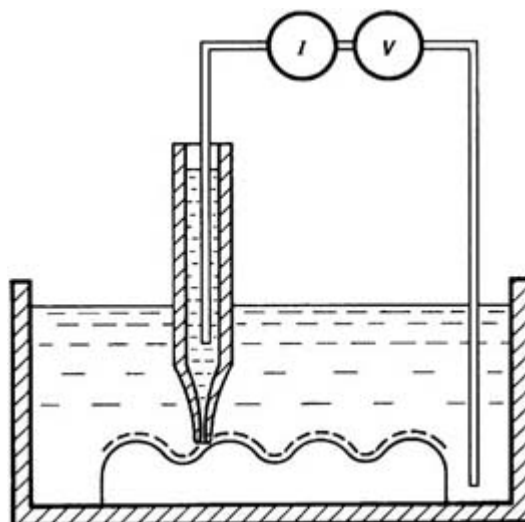


Figure B1.19.40. The scanning ion-conductance microscope (SICM) scans a micropipette over the contours of a surface, keeping the electrical conductance through the tip of the micropipette constant by adjusting the vertical height of the probe. (Taken from [211], figure 1.)

B1.19.4.5 SCANNING MICROPIPETTE MOLECULE MICROSCOPY (SMMM)

The apparatus involved in this method is related to that of SICM, except that the micropipette is blocked by a permeable polymer plug, and connected to the inlet of a differentially pumped quadrupole mass spectrometer [212] (figure B1.19.38). Unlike most other scanning techniques, SMMM relies on a light microscope for positioning. Nevertheless, it is a unique spatially resolved sampling method for desorbing surface species under solution, and has numerous potential applications in biology and medicine. The diffusion of water through pores in a polymer membrane has been followed by using the set-up in figure B1.19.41 where diffusing HDO is converted to HD (with the unambiguous mass of 3 amu) prior to mass spectrometric detection by means of a uranium reduction furnace.

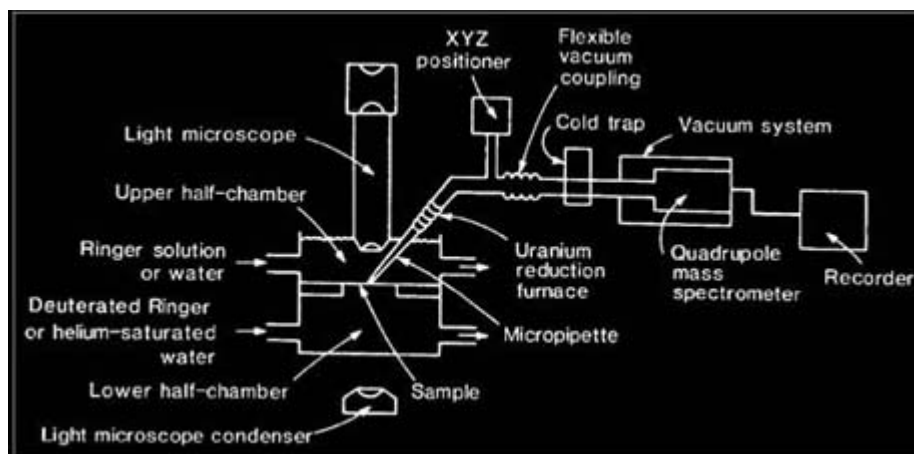


Figure B1.19.41. Schematic of the scanning micropipette molecule microscope. (Taken from [212], figure 1.)

STM and its many related methods have opened new windows onto the nanometre-scale world. Within a short period of time, SPMs have become common fixtures, not only in surface science laboratories, but also in research groups working in areas as diverse as ceramics, polymers, cell biology, robotics, catalysis, and tribology. Clearly this trend will continue, as the concept of the proximal probe becomes combined with more and more of the macroscale analytical tools that we know today. It is also clear that the nanoscale chemical analysis of surfaces by means of SPM will become increasingly viable, with biological surfaces providing some of the most challenging and potentially fruitful analytical problems. Other suggested applications of SPMs are as high-density information storage systems, as selective molecular manipulators and as aids in microsurgery. With the field barely into its second decade, the technological and scientific possibilities of the SPM approach are immense.

REFERENCES

- [1] Binnig G, Rohrer H, Gerber Ch and Weibel E 1982 Tunnelling through a controllable vacuum gap *Appl. Phys. Lett.* **40** 178
 - [2] Salmeron M B 1993 Use of the atomic force microscope to study mechanical properties of lubricant layers *MRS Bulletin XVIII-5* 20
Overney R and Meyer E 1993 Tribological investigations using friction force microscopy *MRS Bulletin XVIII-5* 20
 - [3] McIntyre B J, Salmeron M and Somorjai G A 1993 A variable pressure/temperature scanning tunnelling microscope for surface science and catalysis studies *Rev. Sci. Instrum.* **64** 687
 - [4] Guckenberger R, Hartmann T, Wiegräbe W and Baumeister W 1995 The scanning tunnelling microscope in biology *Scanning Tunnelling Microscopy* vol II, ed R Wiesendanger and H-J Güntherodt (Berlin: Springer) ch 3
 - [5] Wiesendanger R and Güntherodt H-J (eds) 1995 *Scanning Tunnelling Microscopy* (Berlin: Springer) vols I–III
 - [6] Bonnell D A (ed) 1993 *Scanning Tunnelling Microscopy and Spectroscopy* (Weinheim: VCH)
 - [7] Wiesendanger R 1994 *Scanning Probe Microscopy and Spectroscopy* (Cambridge: Cambridge University Press)
 - [8] DiNardo N J 1994 *Nanoscale Characterization of Surfaces and Interfaces* (Weinheim: VCH)
 - [9] Colton R J *et al* (eds) 1998 *Procedures in Scanning Probe Microscopies* (New York: Wiley)
 - [10] Esaki L 1958 New phenomenon in narrow germanium p–n junction *Phys. Rev.* **109** 603
 - [11] Josephson B D 1962 Possible new effects in superconductive tunnelling *Phys. Lett.* **1** 251
 - [12] Rohrer G 1993 The preparation of tip and sample surfaces for STM experiments *Scanning Tunnelling Microscopy and Spectroscopy* ed D A Bonnell (Weinheim: VCH) ch 6
 - [13] Wiesendanger R and Güntherodt H-J 1995 Introduction *Scanning Tunnelling Microscopy II* ed R Wiesendanger and H-J Güntherodt (Berlin: Springer) ch 1
-

- [14] Bonnell D A 1993 Microscope design and operation *Scanning Tunnelling Microscopy and Spectroscopy* ed D A Bonnell (Weinheim: VCH) ch 2
- [15] Tersoff J 1993 Theory of scanning tunnelling microscopy *Scanning Tunnelling Microscopy and Spectroscopy* ed D A Bonnell (Weinheim: VCH) ch 3
- [16] Hamers R 1993 Methods of tunnelling spectroscopy with the STM *Scanning Tunnelling Microscopy and Spectroscopy* ed D A Bonnell (Weinheim: VCH) ch 4

- [17] Binnig G, Rohrer H, Gerber Ch and Weibel E 1983 7×7 reconstruction on Si(111) resolved in real space *Phys. Rev. Lett.* **50** 120
- [18] Tromp R M, Hamers R J and Demuth J E 1986 Atomic and electronic contributions to Si(111)-(7×7) scanning-tunnelling-microscopy images *Phys. Rev. B* **34** 1388
- [19] Hamers R J, Tromp R M and Demuth J E 1986 Surface electronic structure of Si(111)-(7×7) resolved in real space *Phys. Rev. Lett.* **56** 1972
- [20] Hamers R J and Köhler U K 1989 Determination of the local electronic structure of atomic-sized defects on Si(001) by tunnelling spectroscopy *J. Vac. Sci. Technol. A* **7** 2854
- [21] Hamers R, Avouris P and Boszo F 1987 Imaging of chemical-bond formation with the scanning tunnelling microscope: NH_3 dissociation on Si(001) *Phys. Rev. Lett.* **59** 2071
- [22] Zheng Z F, Salmeron M B and Weber E R 1994 Empty state and filled state image of Zn_{Ga} acceptor in GaAs studied by scanning tunnelling microscopy *Appl. Phys. Lett.* **64** 1836
- [23] Kitamura N, Lagally M G and Webb M B 1993 Real-time observations of vacancy diffusion on Si(001)-(2×1) by scanning tunnelling microscopy *Phys. Rev. Lett.* **71** 2082
- [24] Kuk Y 1994 STM on metals *Scanning Tunnelling Microscopy I* ed R Wiesendanger and H-J Güntherodt (Berlin: Springer) ch 3
- [25] Somorjai G A 1994 *Introduction to Surface Chemistry and Catalysis* (New York: Wiley)
- [26] Crommie M F, Lutz C P and Eigler D M 1993 Imaging standing waves in a two-dimensional electron gas *Nature* **363** 524
- [27] Hasegawa Y and Avouris Ph 1993 Direct observation of standing wave formation at surface steps using scanning tunnelling spectroscopy *Phys. Rev. Lett.* **71** 1071
- [28] Campuzano J C, Foster M S, Jennings G, Willis R F and Unertl W 1985 Au(110) (1×2)-to-(1×1) phase transition: a physical realisation of the two-dimensional Ising model *Phys. Rev. Lett.* **54** 2684
- [29] Jaklevic R C and Elie L 1988 Scanning-tunnelling-microscope observation of surface diffusion on an atomic scale: Au on Au(111) *Phys. Rev. Lett.* **60** 120
- [30] Winterlin J and Behm R J 1994 Adsorbate covered metal surfaces and reactions on metal surfaces *Scanning Tunnelling Microscopy I* ed R Wiesendanger and H-J Güntherodt (Berlin: Springer) ch 4
- [31] Hwang R Q, Schröder J, Günther C and Behm R J 1991 Fractal growth of two-dimensional islands: Au on Ru(0001) *Phys. Rev. Lett.* **67** 3279
- [32] Frommer J E 1992 Rastertunnel- und Kraftmikroskopie in der Organischen Chemie *Angew. Chem.* **104** 1325
- [33] Yackoboski K, Yeo Y H, McGonigal G C and Thomson D J 1992 Molecular position at the liquid/solid interface measured by voltage-dependent imaging with the STM *Ultramicroscopy* **42-44** 963

- [34] Weiss P S and Eigler D M 1993 Site dependence of the apparent shape of a molecule in scanning tunnelling microscope images: benzene on Pt(111) *Phys. Rev. Lett.* **71** 3139
- [35] Ulman A 1991 *Ultrathin Organic Films* (London: Academic)
- [36] Ohtani H, Wilson R J, Chiang S and Mate C M 1988 Scanning tunnelling microscopy observations of benzene molecules on the Rh(111)-(3×3) ($\text{C}_6\text{H}_6 + 2\text{CO}$) surface *Phys. Rev. Lett.* **60** 2398
- [37] Forster J S and Frommer J E 1988 Imaging of liquid crystals using a tunnelling microscope *Nature* **333** 542
- [38] Smith D P E, Hörber H, Gerber Ch and Binnig G 1989 Smectic liquid crystal monolayers on graphite observed by scanning tunnelling microscopy *Science* **245** 43

- [39] Nuzzo R G and Allara D L 1983 Adsorption of bifunctional organic disulfides on gold surfaces *J. Am. Chem. Soc.* **105** 4481
 Nuzzo R G, Zegarski D R and Dubois L H 1987 Fundamental studies of the chemisorption of organosulfur compounds on Au(111). Implications for molecular self-assembly on gold surfaces *J. Am. Chem. Soc.* **109** 733
 Bain C D, Troughton E B, Tao T, Evall J, Whitesides G M and Nuzzo R G 1989 Formation of monolayer films by the spontaneous assembly of organic thiols from solution onto gold *J. Am. Chem. Soc.* **111** 321
- [40] Poirier G E and Tarlov M J 1994 The $c(4 \times 2)$ superlattice of n-alkanethiol monolayers self-assembled on Au(111) *Langmuir* **10** 2853
- [41] Stranick S J, Parikh A N, Tao Y-T, Allara D L and Weiss P S 1994 Phase separation of mixed-composition self-assembled monolayers into nanometer scale molecular domains *J. Phys. Chem.* **98** 7636
- [42] Guckenberger R, Heim M, Cevc G, Knapp H F, Wiegräbe W and Hillebrand A 1994 Scanning tunnelling microscopy of insulators and biological specimens based on lateral conductivity of ultrathin water films *Science* **266** 1538
- [43] Clemmer C R and Beebe T P 1991 Graphite: a mimic for DNA and other biomolecules in scanning tunnelling microscope studies *Science* **25** 640
- [44] Heckl W M, Smith D P E, Binnig G, Klagges H, Hänsch T W, and Maddocks J 1991 Two-dimensional ordering of the DNA base guanine observed by scanning tunnelling microscopy *Proc. Natl Acad. Sci., USA* **88** 8003
 Heckl W M and Engel A 1995 *Visualisation of Nucleic Acids* ed G Morel (Boca Raton, FL: CRC Press)
 Allen M J, Balooch M, Subbiah S, Tench R J, Balhorn R and Siekhaus W 1991 Scanning tunnelling microscope images of adenine and thymine at atomic resolution *Scanning Microsc.* **5** 625
 Allen M J, Balooch M, Subbiah S, Tench R J, Balhorn R and Siekhaus W 1992 Analysis of adenine and thymine adsorbed on graphite by scanning tunnelling and atomic force microscopy *Ultramicrosc.* **42–44** 1049
- [45] Rabe J P, Buchholz S and Ritcey A M 1990 Reactive graphite etch of an adsorbed organic monolayer—a scanning tunnelling microscopy study *J. Vac. Sci. Technol. A* **8** 679
 Miles M J, Lee I and Atkins E D T 1991 Molecular resolution of polysaccharides by scanning tunnelling microscopy *J. Vac. Sci. Technol. B* **9** 1206
 Tang S L, McGhie A J and Suna A 1993 Molecular-resolution imaging of insulating macromolecules with the scanning tunnelling microscope via a nontunnelling, electric-field-induced mechanism *Phys. Rev. B* **47** 3850
- [46] Guckenberger R, Hartmann T and Knapp H F 1995 Recent developments *Scanning Tunnelling Microscopy II* ed R Wiesendanger and H-J Güntherodt (Berlin: Springer) ch 9
- [47] Maaloum M, Chrétien D, Karsenti E and Hörber J K H 1994 Approaching microtubule structure with the scanning tunnelling microscope (STM) *J. Cell Sci.* **107** part II 3127
- [48] Yuan J-Y, Shao Z and Gao C 1991 Alternative method of imaging surface topologies of nonconducting bulk specimens by scanning tunnelling microscopy *Phys. Rev. Lett.* **67** 863

- [49] Guckenberger R, Heim M, Cevc G, Knapp H F, Wiegräbe W and Hillebrand A 1994 Scanning tunnelling microscopy of insulators and biological specimens based on lateral conductivity of ultrathin water films *Science* **266** 1538
- [50] Fan F-R F and Bard A J 1995 STM on wet insulators: electrochemistry or tunnelling? *Science* **270** 1849
- [51] Fujiwara I, Ishimoto C and Seto J 1991 Scanning tunnelling microscopy study of a polyimide Langmuir–Blodgett film *J. Vac. Sci. Technol. B* **9** 1148
 Sotobayashi H, Schilling T and Tesche B 1990 Scanning tunnelling microscopy of polyimide monolayers prepared by the Langmuir–Blodgett technique *Langmuir* **6** 1246
- [52] Grunze M, Unertl W N, Ganarajan S and French J 1988 Chemistry of adhesion at the polyimide–metal interface *Mater. Res. Soc. Symp. Proc.* **108** 189
- [53] Siegenthaler H 1995 STM in electrochemistry *Scanning Tunnelling Microscopy II* ed R Wiesendanger and H-J Güntherodt (Berlin: Springer) ch 2
- [54] Bard A J and Fan F F 1993 Applications in electrochemistry *Scanning Tunnelling Microscopy and Spectroscopy* ed D A Bonnell (Weinheim: VCH) ch 9
- [55] Cataldi T R I, Blackham I G, Briggs G A D, Pethica J B and Hill H A O 1990 New insight for electrochemical electrode–surface investigations *J. Electroanal. Chem.* **290** 1

- [56] Christoph R, Siegenthaler H, Rohrer H and Wiese H 1989 *In situ* scanning tunnelling microscopy at potential controlled Ag(100) substrates *Electrochim. Acta* **34** 1011
- [57] Siegenthaler H 1995 Recent developments *Scanning Tunnelling Microscopy II* ed R Wiesendanger and H-J Güntherodt (Berlin: Springer) ch 9
- [58] Magnussen O M, Hotlos J, Beitel G, Kolb D M and Behm R J 1991 Atomic structure of ordered copper adlayers on single-crystalline gold electrodes *J. Vac. Sci. Technol. B* **9** 969
- [59] Chen S J, Sanz F, Ogletree D F, Hallmark V M, Devine T M and Salmeron M 1993 Selective dissolution of copper from Au-rich Cu–Au alloys: an electrochemical STM study *Surf. Sci.* **292** 289
- [60] Ogura K, Tsujigo M, Sakurai K and Yano J 1993 Electrochemical coloration of stainless steel and the scanning tunnelling microscopic study *J. Electrochem. Soc.* **140** 1311
- [61] Müller-Zülow B, Kipp S, Lacmann R and Schneeweiss M A 1994 Topological aspects of iron corrosion in alkaline solution by means of scanning force microscopy (SFM) *Surf. Sci.* **311** 153
- [62] Bard A J, Fan F F, Pierce D T, Unwin P R, Wipf D O and Zhou F 1991 Chemical imaging of surfaces with the scanning electrochemical microscope *Science* **254** 68
- [63] McIntyre B J, Salmeron M and Somorjai G A 1994 Nanocatalysis by the tip of a scanning tunnelling microscope operating inside a reactor cell *Science* **265** 1415
- [64] Stauer U 1995 Surface modification with a scanning proximity probe microscope *Scanning Tunnelling Microscopy II* ed R Wiesendanger and H-J Güntherodt (Berlin: Springer) ch 8
- [65] Eigler D M and Schweizer E K 1990 Positioning single atoms with a scanning tunnelling microscope *Nature* **344** 524
- [66] Crommie M F, Lutz C P and Eigler D M 1993 Confinement of electrons to quantum corrals on a metal surface *Science* **262** 218
- [67] Yokoyama T and Takayanagi K 1999 Size quantization of surface state electrons on the Si(001) surface *Phys. Rev. B* **59** 12 232
-

- [68] Salling C T and Lagally M G 1994 Fabrication of atomic-scale structures on Si(001) surfaces *Science* **265** 502
- [69] Kent A D, Shaw T M, Molnar S V and Awschalom D D 1993 Growth of high aspect ratio nanometer-scale magnets with chemical vapor deposition and scanning tunnelling microscopy *Science* **262** 1249
- [70] Rugar D and Hansma P K 1990 Atomic force microscopy *Physics Today* **43**(10) 23
- [71] Guenther K H, Wierer P G and Bennett J M 1984 Surface roughness measurements of low-scatter mirrors and roughness standards *Appl. Opt.* **23** 3820
- [72] Teague E C, Scire F E, Backer S M and Jensen S W 1982 Three-dimensional stylus profilometry *Wear* **83** 1
- [73] Binnig G, Quate C F and Gerber Ch 1986 Atomic force microscope *Phys. Rev. Lett.* **56** 930
- [74] Hansma P K, Elings V B, Marti O and Bracker C E 1988 Scanning tunnelling microscopy and atomic force microscopy: application to biology and technology *Science* **242** 209
- [75] Burnham N A and Colton R J 1993 Force microscopy *Scanning Tunnelling Microscopy and Spectroscopy* ed D A Bonnell (Weinheim: VCH) ch 7
- [76] Hartmann U 1991 van der Waals interactions between sharp probes and flat sample surfaces *Phys. Rev. B* **43** 2404
- [77] Wolter O, Bayer T and Greschner J 1991 Micromachined silicon sensors for scanning force microscopy *J. Vac. Sci. Technol. B* **9** 1353
- [78] Cross G, Schirmeisen A, Stalder A, Grütter P, Tschudy M and Dürig U 1998 Adhesion interaction between atomically defined tip and sample *Phys. Rev. Lett.* **80** 4685

- [79] Dai H, Hafner J H, Rinzler A G, Colbert D T and Smalley R E 1996 Nanotubes as nanoprobe in scanning probe microscopy *Nature* **384** 147
- [80] Meyer G and Amer N M 1988 Novel optical approach to atomic force microscopy *Appl. Phys. Lett.* **53** 1045
- [81] Sarid D, Pax P, Yi L, Howells S, Gallagher M, Chen T, Elings V and Bocek D 1992 Improved atomic force microscope using a laser diode interferometer *Rev. Sci. Instrum.* **63** 3905
Nonnenmacher M, Vaez-Iravani M and Wickramasinghe H K 1992 Attractive mode force microscopy using feedback-controlled fiber interferometer *Rev. Sci. Instrum.* **63** 5373
- [82] Giessibl F J 1995 Atomic resolution of the silicon (111)-(7 × 7) surface by atomic force microscopy *Science* **260** 67
- [83] Pethica J B 1986 Comment on interatomic forces in scanning tunnelling microscopy: giant corrugations of the graphite surface *Phys. Rev. Lett.* **57** 3235
- [84] Carpick R W, Agraït N, Ogletree D F and Salmeron M 1996 Measurement of interfacial shear (friction) with an ultrahigh vacuum atomic force microscope *J. Vac. Sci. Technol. B* **14** 1289
- [85] Wilson D L, Kump K S, Eppell S J and Marchant R E 1995 Morphological restoration of atomic force microscopy images *Langmuir* **11** 265
- [86] Sheiko S S, Moller M, Reuvekamp E M C M and Zandbergen H W 1993 Calibration and evaluation of scanning-force-microscopy probes *Phys. Rev. B* **48** 5675
- [87] Roberts C J, Williams P M, Davies J, Dawkes A C, Sefton J, Edwards J C, Haymes A G, Bestwick C, Davies M C and Tendler S J B 1995 Real-space differentiation of IgG and IgM antibodies deposited on microtiter wells by scanning force microscopy *Langmuir* **11** 1822
Shivji A P, Brown F, Davies M C, Jennings K H, Roberts C J, Tendler S J B, Wilkinson M J and Williams P M 1995 Scanning tunnelling microscopy studies of β -amyloid fibril structure and assembly *FEBS Lett.* **371** 25–8

- [88] Ducker W A, Senden T J and Pashley R M 1992 Measurement of forces in liquids using a force microscope *Langmuir* **8** 1831
Ducker W A, Senden T J and Pashley R M 1991 Direct measurement of colloidal forces using an atomic force microscope *Nature* **353** 239
- [89] Jarvis S P, Yamada H, Yamamoto S-I, Tokumoto H and Pethica J B 1996 Direct mechanical measurement of interatomic potentials *Nature* **384** 247
- [90] Gotsmann B, Anczykowski B, Seidel C and Fuchs H 1999 Determination of tip-sample interaction forces from measured dynamic force spectroscopy curves *Appl. Surf. Sci.* **140** 314
- [91] Tabor D and Winterton R H S 1969 The direct measurement of normal and retarded van der Waals forces *Proc. R. Soc. A* **312** 435
- [92] Thundat T, Zheng X-Y, Chen G Y, Sharp S L, Warmack R J and Schowalter L J 1993 Characterization of atomic force microscope tips by adhesion force measurements *Appl. Phys. Lett.* **63** 2150
- [93] Hutter J L and Bechhoefer J 1994 Measurement and manipulation of van der Waals forces in atomic-force microscopy *J. Vac. Sci. Technol. B* **12** 2251
- [94] Hansma H G, Vesenka J, Siegerist C, Kelderman G, Morrett H, Sinsheimer R L, Bustamante C, Elings V and Hansma P K 1992 Reproducible imaging and dissection of plasmid DNA under liquid with the atomic force microscope *Science* **256** 1180
- [95] Jarvis S P, Dürig U, Lantz M A, Yamada H and Tokumoto H 1998 Feedback stabilized force-sensors: a gateway to the direct measurement of interaction potentials *Appl. Phys. A* **66** S211
- [96] Jarvis S P, Yamamoto S-I, Yamada H, Tokumoto H and Pethica J B 1997 Tip-surface interactions studied using a force controlled atomic force microscope in ultrahigh vacuum *Appl. Phys. Lett.* **70** 2238
- [97] Giessibl F J 1997 Forces and frequency shifts in atomic-resolution dynamic-force microscopy *Phys. Rev. B* **56** 16 010
- [98] Mate C M, Lorenz M R and Novotny V J 1989 Atomic force microscopy of polymeric liquid films *J. Chem. Phys.* **90**

- [99] Weisenhorn A L and Hansma P K 1989 Forces in atomic force microscopy in air and water *Appl. Phys. Lett.* **54** 2651
- [100] Proceedings from the First International Conference on NC-AFM *Appl. Surf. Sci.* **140** 243
- [101] Albrecht T R, Grütter P, Horne D and Rugar D 1991 Frequency modulation detection using high-Q cantilevers for enhanced force microscope sensitivity *J. Appl. Phys.* **69** 668
- [102] Dürig U, Züger O and Stalder A 1992 Interaction force detection in scanning probe microscopy: methods and applications *J. Appl. Phys.* **72** 1778
- [103] Sasaki N and Tsukada M 1999 Theory for the effect of the tip–surface interaction potential on atomic resolution in forced vibration system of noncontact AFM *Appl. Surf. Sci.* **140** 339
- [104] Perez R, Payne M C, Stich I and Terakura K 1997 Role of covalent tip–surface interactions in noncontact atomic force microscopy on reactive surfaces *Phys. Rev. Lett.* **78** 678
- [105] Zong Q, Inniss D, Kjoller K and Elings V B 1993 Fractured polymer/silica fiber surface studied by tapping mode atomic force microscopy *Surf. Sci. Lett.* **290** L688
- [106] Hansma P K *et al* 1994 Tapping mode atomic force microscopy in liquids *Appl. Phys. Lett.* **64** 1738
- [107] Hobbs P, Abraham D and Wickramasinghe H 1989 Magnetic force microscopy with 25 nm resolution *Appl. Phys. Lett.* **55** 2357

- [108] Grütter P, Mamin H J and Rugar D 1995 Magnetic force microscopy (MFM) *Scanning Tunnelling Microscopy II* ed R Wiesendanger and H-J Güntherodt (Berlin: Springer) ch 5
- [109] Rugar D, Mamin H J, Guenther P, Lambert S E, Stern J E, McFadyen I and Yogi T 1990 Magnetic force microscopy: general principles and application to longitudinal recording media *J. Appl. Phys.* **68** 1169
- [110] Overney R and Meyer E 1993 Tribological investigations using friction force microscopy *MRS Bull.* **XVIII** 20
- [111] Mate C M, Erlandsson R, McClelland G M and Chiang S 1987 Atomic-scale friction of a tungsten tip on a graphite surface *Phys. Rev. Lett.* **59** 1942
- [112] Meyer G and Amer N M 1990 Simultaneous measurement of lateral and normal forces with an optical-beam-deflection atomic force microscope *Appl. Phys. Lett.* **57** 2089
- [113] Marti O, Colchero J and Mlynek J 1990 Combined scanning force and friction microscopy of mica *Nanotechnology* **1** 141
- [114] Burnham N A and Colton R J 1989 Measuring the nanomechanical properties and surface forces of materials using atomic force microscope *J. Vac. Sci. Technol. A* **7** 2906
- [115] van der Werf K O, Putman C A J, Groth B G and Greve J 1994 Adhesion force imaging in air and liquid by adhesion mode atomic force microscopy *Appl. Phys. Lett.* **65** 1195
Miyatani T, Horii M, Rosa A, Fujihira M and Marti O 1997 Mapping of electric double-layer force between tip and sample surfaces in water with pulsed-force-mode atomic force microscopy *Appl. Phys. Lett.* **71** 2632
- [116] Kolosov O and Yamanaka K 1993 Nonlinear detection of ultrasonic vibrations in an atomic force microscope *Japan. J. Appl. Phys.* **32** L1095
- [117] Erlandsson R, Olsson L and Martensson P 1996 Inequivalent atoms and imaging mechanisms in AC-mode atomic force microscopy of Si(111) 7×7 *Phys. Rev. B* **54** 8309
- [118] Nakagiri N, Suzuki M, Okiguchi K and Sugimura H 1997 Site discrimination of adatoms in Si(111)- 7×7 by noncontact atomic force microscopy *Surf. Sci.* **375** L329
- [119] Takayanagi K, Takashiro Y, Takahashi M and Takahashi S 1985 Structural analysis of Si(111)- 7×7 by UHV-transmission electron diffraction and microscopy *J. Vac. Sci. Technol. A* **3** 1502

- [120] Bammerlin M, Luthi R, Meyer E, Baratoff A, Lü J, Guggisberg M, Gerber Ch, Howald L and Güntherodt H-J 1997 True atomic resolution on the surface of an insulator via ultrahigh vacuum dynamic force microscopy *Probe Microsc.* **1** 3
- [121] Fukui K, Onishi H and Iwasawa Y 1997 Atom-resolved image of the TiO₂(110) surface by noncontact atomic force microscopy *Phys. Rev. Lett.* **79** 4202
- [122] Raza H, Pang C L, Haycock S A and Thornton G 1999 Non-contact atomic force microscopy imaging of TiO₂(100) surfaces *Appl. Surf. Sci.* **140** 271
- [123] Friedbacher G, Hansma P K, Ramli E and Stucky G D 1991 Imaging powders with the atomic force microscope: from biominerals to commercial materials *Science* **253** 1261
- [124] Erlandsson R, Hadziioannou G, Mate M, McClelland G and Chiang S 1988 Atomic scale friction between the muscovite mica cleavage plane and a tungsten tip *J. Chem. Phys.* **89** 5190
- [125] Heckl W, Ohnesorge F, Binnig G, Specht M and Hashmi M 1991 Ring structures on natural molybden disulfide investigated by scanning tunnelling and scanning force microscopy *J. Vac. Sci. Technol. B* **9** 1072

-54-

- [126] Hegenbart G and Müssig Th 1992 Atomic force microscopy studies of atomic structures on AgBr(111) surfaces *Surf. Sci. Lett.* **275** L655
- [127] Keita B, Nadjo L and Kjoller K 1991 Surface characterization of a single crystal of sodium decatungstocerate (IV) by the atomic force microscope *Surf. Sci. Lett.* **256** L613
- [128] Colchero J, Marti O, Mlynek J, Humbert A, Henry C R and Chapon C 1991 Palladium clusters on mica: a study by scanning force microscopy *J. Vac. Sci. Technol. B* **9** 794
- [129] Baski A A and Fuchs H 1994 Epitaxial growth of silver on mica as studied by AFM and STM *Surf. Sci.* **313** 275
- [130] van Duyne R P, Hulthen J C and Treichel D A 1993 Atomic force microscopy and surface-enhanced Raman spectroscopy. Ag island films and Ag film over polymer nanosphere surfaces supported on glass *J. Chem. Phys.* **99** 2101
- [131] Roark S E and Rowlen K L 1993 Atomic force microscopy of thin Ag films. Relationship between morphology and optical properties *Chem. Phys. Lett.* **212** 50
- [132] Overney R, Howald L, Frommer J, Meyer E and Güntherodt H 1991 Molecular surface structure of tetracene mapped by the atomic force microscope *J. Chem. Phys.* **94** 8441
- [133] Overney R, Howald L, Frommer J, Meyer E, Brodbeck D and Güntherodt H 1992 Molecular surface structure of organic crystals observed by atomic force microscopy *Ultramicroscopy* **42-44** 983
- [134] Schwartz D K, Viswanathan R and Zasadzinski J A N 1993 Commensurate defect superstructures in a Langmuir-Blodgett film *Phys. Rev. Lett.* **70** 1267
- [135] Bourdieu L, Ronsin O and Chatenay D 1993 Molecular positional order in Langmuir-Blodgett films by atomic force microscopy *Science* **259** 798
- [136] Liu G-Y and Salmeron M B 1994 Reversible displacements of chemisorbed n-alkanethiol molecules on Au(111) surface: an atomic force microscopy study *Langmuir* **10** 367
- [137] Salmeron M, Neubauer G, Folch A, Tomitori M, Ogletree D F and Sautet P 1993 Viscoelastic and electrical properties of self-assembled monolayers on Au(111) films *Langmuir* **9** 3600
- [138] Salmeron M, Liu G-Y and Ogletree D F 1995 Molecular arrangement and mechanical stability of self-assembled monolayers on Au(111) under applied load *Force in Scanning Probe Methods* ed H-J Güntherodt *et al* (Amsterdam: Kluwer)
- [139] Grim P C M, Brouwer H J, Seyger R M, Oostergetel G T, Bergsma-Schutter W G, Arnberg A C, Gütthner P, Dransfeld K and Hadziioannou G 1992 Investigation of polymer surfaces using scanning force microscopy (SFM) 'A new direct look on old polymer problems' *Makromol. Chem., Macromol. Symp.* **62** 141

- [140] Annis B K, Noid D W, Sumpter B G, Reffner J R and Wunderlich B 1992 Application of atomic force microscopy (AFM) to a block copolymer and an extended chain polyethylene *Makromol. Chem., Rapid. Commun.* **13** 169
 Annis B K, Schwark D W, Reffner J R, Thomas E L and Wunderlich B 1992 Determination of surface morphology of diblock copolymers of styrene and butadiene by atomic force microscopy *Makromol. Chem.* **193** 2589
- [141] Krausch G, Hipp M, Bölltau M, Mlynek J and Marti O 1995 High resolution imaging of polymer surfaces with chemical sensitivity *Macromolecules* **28** 260
- [142] Feldman K, Tervoort T, Smith P and Spencer N D 1998 Toward a force spectroscopy of polymer surfaces *Langmuir* **14** 372
- [143] Snéitivy D and Vancso G J 1994 Atomic force microscopy of polymer crystals: 7. Chain packing, disorder and imaging of methyl groups in oriented isotactic polypropylene *Polymer* **35** 461

-55-

- [144] Vansteenkiste S O, Davies M C, Roberts C J, Tendler S J B and Williams P M 1998 Scanning probe microscopy of biomedical interfaces *Prog. Surf. Sci.* **57** 95
- [145] Ikai A, Yoshimura K, Arisaka F, Ritani A and Imai K 1993 Atomic force microscopy of bacteriophage T4 and its tube-baseplate complex *FEBS Lett.* **326** 39
- [146] Wagner P 1998 Immobilization strategies for biological scanning probe microscopy *FEBS Lett.* **430** 112
- [147] Bustamente C, Vesenka J, Tang C L, Rees W, Guthold M and Keller R 1992 Circular DNA molecules imaged in air by scanning force microscopy *Biochemistry* **31** 22
- [148] Allen M J, Hud N V, Balooch M, Tench R J, Siekhaus W J and Balhorn R 1992 Tip-radius-induced artifacts in AFM images of protamine-complexed DNA fibers *Ultramicroscopy* **42-44** 1095
- [149] Keller D and Chou C C 1992 Imaging steep, high structures by scanning force microscopy with electron beam deposited tips *Surf. Sci.* **268** 333
- [150] Thundat T, Warmack R J, Allison D P, Bottomley L A, Lourenco A J and Ferrell T L 1992 Atomic force microscopy of deoxyribonucleic acid strands adsorbed on mica: the effect of humidity on apparent width and image contrast *J. Vac. Sci. Technol. A* **10** 630
- [151] Bustamente C, Keller D and Yang G 1993 Scanning force microscopy of nucleic acids and nucleoprotein assemblies *Curr. Opin. Struct. Biol.* **3** 363
- [152] Hansma H G, Vesenka J, Siegerist C, Kelderman G, Morrett H, Sinsheimer R L, Elings V, Bustamente C and Hansma P K 1992 Reproducible imaging and dissection of plasmid DNA under liquid with the atomic force microscope *Science* **256** 1180
- [153] Hansma H G, Sinsheimer R L, Li M-Q and Hansma P K 1992 Atomic force microscopy of single- and double-stranded DNA *Nucleic Acids Res.* **20** 3585
- [154] Rees W A, Keller R W, Vesenka J P, Yang G and Bustamente C 1993 Evidence of DNA bending in transcription complexes imaged by scanning force microscopy *Science* **260** 1646
- [155] Hansma H G, Laney D E, Bezanilla M, Sinsheimer R L and Hansma P K 1995 Applications for atomic-force microscopy of DNA *Biophys. J.* **68** 1672
- [156] Samori B, Siligardi G, Quagliariello C, Weisenhorn A L, Vesenka J and Bustamente C 1993 Chirality of DNA supercoiling assigned by scanning force microscopy *Proc. Natl Acad. Sci. USA* **90** 3598
- [157] Hansma H G, Sinsheimer R L, Groppe J, Bruice T C, Elings V, Gurley G, Bezanilla M, Mastrangelo I A, Hough P V C and Hansma P K 1993 Recent advances in atomic force microscopy of DNA *Scanning* **15** 296
- [158] Radmacher M, Fritz M, Hansma H G and Hansma P K 1994 Direct observation of enzyme activity with the atomic force microscope *Science* **265** 1577
- [159] Bottomley L A, Coury J E and First P N 1996 Scanning probe microscopy *Anal. Chem.* **68** 185R
 Shao Z and Yang J 1995 Progress in high-resolution atomic-force microscopy in biology *Qt. Rev. Biophys.* **28** 195
 Shao Z, Mou J, Czajkowsky D M, Yang J and Yuan J 1996 Biological atomic force microscopy: what is achieved and

what is needed *Adv. Phys.* **45** 1

Ikai A 1996 STM and AFM of bio/organic molecules and structures *Surf. Sci. Rep.* **26** 263

Lal R and John S A 1994 Biological applications of atomic-force microscopy *Am. J. Physiol.* **266** C1

Shao Z F, Yang J and Somlyo A P 1995 Biological atomic force microscopy: from microns to nanometers and beyond *Ann. Rev. Cell Dev. Biol.* **11** 241

Kasas S, Thompson N H, Smith B L, Hansma P K, Miklossy J and Hansma H G 1997 Biological applications of the AFM: from single molecules to organs *Int. J. Im. Syst. Technol.* **8** 151

-56-

- [160] Müller D J, Baumeister W and Engel A 1996 Conformational change of the hexagonally packed intermediate layer of *Deinococcus radiodurans* monitored by atomic force microscopy *J. Bacteriol.* **178** 3025
Müller D W, Fotiadis D and Engel A 1998 Mapping flexible protein domains at subnanometre resolution with the atomic force microscope *FEBS Lett.* **430** 105
- [161] Lal R, Kim H, Garavito R M and Arnsdorf M F 1993 Imaging of reconstituted biological channels at molecular resolution by atomic force microscopy *Am. J. Physiol.* **265** C851
- [162] Rief M, Gautel M, Oesterhelt F, Fernandez J M and Gaub H E 1997 Reversible unfolding of individual titin immunoglobulin domains by AFM *Science* **276** 1109
- [163] Henderson E, Haydon P G and Sakaguchi D S 1992 Actin filament dynamics in living glial cells imaged by atomic force microscopy *Science* **257** 1944
- [164] Brandow S L, Turner D C, Ratna B R and Gaber B P 1993 Modification of supported lipid membranes by atomic force microscopy *Biophys. J.* **64** 898
- [165] Parpura V, Haydon P G, Sakaguchi D S and Henderson E 1993 Atomic force microscopy and manipulation of living glial cells *J. Vac. Sci. Technol. A* **11** 773
- [166] Lee G U, Kidwell D A and Colton R J 1994 Sensing discrete streptavidin–biotin interactions with atomic force microscopy *Langmuir* **10** 354
- [167] Florin E-L, Moy V T and Gaub H E 1994 Adhesion forces between individual ligand-receptor pairs *Science* **264** 415
- [168] Lee G U, Chrisey L A and Colton R J 1994 Direct measurement of the forces between complementary strands of DNA *Science* **266** 771
- [169] Bongrand P 1999 Ligand–receptor interactions *Rep. Prog. Phys.* **62** 921
- [170] Ducker W A, Senden T J and Pashley R M 1991 Direct measurement of colloidal forces using an atomic force microscope *Nature* **353** 239
- [171] Li Y Q, Tao N J, Pan J, Garcia A A and Lindsay S M 1993 Direct measurement of interaction forces between colloidal particles using the scanning force microscope *Langmuir* **9** 637
- [172] Larson I, Drummond C J, Chan D Y C and Grieser F 1993 Direct force measurements between TiO₂ surfaces *J. Am. Chem. Soc.* **115** 11 885
- [173] Biggs S, Mulvaney P, Zukoski C F and Grieser F 1994 Study of anion adsorption at the gold–aqueous solution interface by atomic force microscopy *J. Am. Chem. Soc.* **116** 9150
- [174] Greenwood J A 1967 On the area of contact between rough surfaces and flats *J. Lub. Tech. (ASME)* **1** 81
- [175] Johnson K L, Kendall K and Roberts A D 1971 Surface energy and the contact of elastic solids *Proc. R. Soc. A* **324** 301
Sperling G 1964 *Dissertation* Karlsruhe Technische Hochschule
- [176] Hu J, Xiao X-D, Ogletree D F and Salmeron M 1995 Atomic scale friction and wear of mica *Surf. Sci.* **327** 358
- [177] Lüthi R, Meyer E, Haefke H, Howald L and Güntherodt H-J 1995 Nanotribology: an UHV-SFM study of thin films of AgBr(001) *Tribol. Lett.* **1** 23
- [178] Overney R M, Meyer E, Frommer J, Brodbeck D, Lüthi R, Howald L, Güntherodt H-J, Fujihara M, Takano H and Gotoh Y 1992 Friction measurements of phase separated thin films with a modified atomic force microscope *Nature*

-
- [179] Brager W R, Koleske D D, Feldman K, Krueger D and Colton R J 1996 Small change-big effect: SPM studies of two-component fatty-acid monolayers *ACS Polymer Preprints* **37** 606
- [180] Kim H I, Graupe M, Oloba O, Koini T, Imaduddin S, Lee T R and Perry S S 1999 Molecularly specific studies of the frictional properties of monolayer films: a systematic comparison of CF_3 -, $(\text{CH}_3)_2\text{CH}$ -, and CH_3 -terminated films *Langmuir* **15** 3179
- [181] Yoshizawa H, Chen Y-L and Israelachvili J N 1993 Fundamental mechanisms of interfacial friction. 1. Relation between adhesion and friction *J. Chem. Phys.* **97** 4128
Israelachvili J N, Chen Y-L and Yoshizawa H 1994 Relationship between adhesion and friction forces *J. Adhesion Sci. Technol.* **8** 1234
- [182] Haugstad G, Gladfelter W L, Weberg E B, Weberg R T and Jones R R 1995 Friction force microscopy as a probe of molecular relaxation on polymer surfaces *Tribol. Lett.* **1** 253
- [183] Burnham N A and Colton R J 1989 Measuring the nanomechanical properties and surface forces of materials using an atomic force microscope *J. Vac. Sci. Technol. A* **7** 2906
- [184] Salmeron M, Folch A, Neubauer G, Tomitori M and Ogletree D F 1992 Nanometer scale mechanical properties of Au (111) thin films *Langmuir* **8** 2832
- [185] Rivière J C 1990 *Practical Surface Analysis* 2nd edn, vol 1, ed D Briggs and M P Seah (Chichester: Wiley)
Briggs D 1990 *Practical Surface Analysis* 2nd edn, vol 2, ed D Briggs and M P Seah (Chichester: Wiley)
- [186] Akari S, Horn D, Keller H and Schrepp W 1995 Chemical imaging by scanning force microscopy *Adv. Mater.* **7** 549
- [187] Frisbie C D, Rozsnyai L F, Noy A, Wrighton M S and Lieber C M 1994 Functional group imaging by chemical force microscopy *Science* **265** 2071
- [188] McKendry R, Theoclitou M-E, Rayment T and Abell C 1998 Chiral discrimination by chemical force microscopy *Nature* **391** 566
- [189] Marti A, Hähner G and Spencer N D 1995 The sensitivity of frictional forces to pH on a nanometer scale: a lateral force microscopy study *Langmuir* **11** 4632
Hähner G, Marti A and Spencer N D 1997 The influence of pH on friction between oxide surfaces in electrolytes, studied with lateral force microscopy: application as a nanochemical imaging technique *Tribol. Lett.* **3** 359
- [190] Sittig C, Hähner G, Marti A, Textor M, Spencer N D and Hauert R 1999 The implant material, Ti6Al7Nb: surface microstructure, composition, and properties *J. Mater. Sci.* **10** 191
- [191] Bergasa F and Saenz J J 1992 Is it possible to observe biological macromolecules by electrostatic force microscopy? *Ultramicroscopy* **42-44** 1189
- [192] Berger C E H, van der Werf K O, Kooyman R P H, de Groot B G and Greve J 1995 Functional group imaging by adhesion AFM applied to lipid monolayers *Langmuir* **11** 4188
- [193] Pohl D W 1991 Scanning near-field optical microscopy (SNOM) *Advances in Optical and Electron Microscopy* vol 12, ed R Barer and V E Cosslett (London: Academic)
- [194] Abbé E 1873 *Archiv. Mikroskop. Anal.* **9** 413
- [195] Synge E H 1928 Extending microscopic resolution into the ultra-microscopic region *Phil. Mag.* **6** 356
- [196] Pohl D W, Denk W and Lanz M 1984 Optical stethoscopy: image recording with resolution $\lambda/20$ *Appl. Phys. Lett.* **44** 651
- [197] Ash E A and Nichols G 1972 Super-resolution aperture scanning microscope *Nature* **237** 510
-

- [198] Dürig U T, Pohl D W and Rohner F 1986 Near-field optical-scanning microscopy *J. Appl. Phys.* **59** 3318
Pohl D W 1991 Scanning near-field optical microscopy (SNOM) *Adv. Opt. Electron. Microsc.* **12** 243
- [199] Betzig E, Finn P L and Weiner J S 1992 Combined shear force and near-field scanning optical microscopy *Appl. Phys. Lett.* **60** 2484
- [200] Fischer U Ch 1985 Optical characteristics of 0.1 μm circular apertures in a metal film as light sources for scanning ultramicroscopy *J. Vac. Sci. Technol. B* **3** 386
Fischer U Ch, Dürig U T and Pohl D W 1988 Near-field scanning microscopy in reflection *Appl. Phys. Lett.* **52** 249
Cline J A, Barshatzky H and Isaacson M 1991 Scanned-tip reflection-mode near-field scanning optical microscopy *Ultramicroscopy* **38** 299
- [201] Betzig E and Chichester R J 1993 Single molecules observed by near-field scanning optical microscopy *Science* **262** 1422
Trautman J K, Macklin J J, Brus L E and Betzig E 1994 Near-field spectroscopy of single molecules at room temperature *Nature* **369** 40
- [202] Sharp S L, Warmack R J, Goudonnet J P, Lee I and Ferrell T L 1993 Spectroscopy and imaging using the photon scanning-tunnelling microscope *Acc. Chem. Res.* **26** 377
- [203] Tsai D P, Othonos A, Moskovits M and Uttamchandani D 1994 Raman spectroscopy using a fibre optic probe with subwavelength aperture *Appl. Phys. Lett.* **64** 1768
- [204] Smith D A, Webster S, Ayad M, Evans S D, Fogherty D and Batchelder D 1995 Development of a scanning near-field optical probe for localised Raman spectroscopy *Ultramicroscopy* **61** 247
- [205] Takahashi S, Futamata M and Kojima I 1999 Spectroscopy with scanning near-field optical microscopy using photon tunnelling mode *J. Microscopy* **194** 519
- [206] Emory S R and Nie S 1997 Near-field surface-enhanced Raman spectroscopy on single silver nanoparticles *Anal. Chem.* **69** 2631
- [207] Zeisel D, Deckert V, Zenobi R and Vo-Dinh T 1998 Near-field surface-enhanced Raman spectroscopy of dye molecules adsorbed on silver island films *Chem. Phys. Lett.* **283** 381
- [208] Guenther P, Fischer U Ch and Dransfeld K 1989 Scanning near-field acoustic microscopy *Appl. Phys. B* **48** 89
- [209] Williams C C and Wickramasinghe H K 1986 Scanning thermal profiler *Appl. Phys. Lett.* **49** 1587
- [210] Hammiche A, Hourston D J, Pollock H M, Reading M and Song M 1996 Scanning thermal microscopy: sub-surface imaging, thermal mapping of polymer blends, localised calorimetry *J. Vac. Sci. Technol. B* **14** 1486
- [211] Hansma P K, Drake B, Marti O, Gould S A C and Prater C B 1989 The scanning ion-conductance microscope *Science* **243** 641
- [212] Jarrell J A, King J G and Mills J W 1981 A scanning micropipette molecule microscope *Science* **211** 277
- [213] Binnig *et al* 1982 Surface studies by scanning tunnelling microscopy *Phys. Rev. Lett.* **49** 57
- [214] Hansma P K and Tersoff J 1987 Scanning tunneling microscopy *J. Appl. Phys.* **61** R1
- [215] Weisenhorn A L 1991 *PhD Thesis* University of California, Santa Barbara
-

- [216] Manne S, Hansma P K, Massie J, Elings V B and Gewirth A A 1991 Atomic-resolution electrochemistry with the atomic force microscope: copper deposition on gold *Science* **251** 183
- [217] Jarvis S P and Tokumoto H 1997 Measurement and interpretation of forces in the atomic force microscope *Probe Microscopy* **1** 65

B1.20 The surface forces apparatus

Manfred Heuberger

B1.20.1 INTRODUCTION

Compared with other direct force measurement techniques, a unique aspect of the surface forces apparatus (SFA) is to allow *quantitative measurement of surface forces and intermolecular potentials*. This is made possible by essentially three measures: (i) well defined contact geometry, (ii) high-resolution interferometric distance measurement and (iii) precise mechanics to control the separation between the surfaces.

It is remarkable that the roots of the SFA go back to the early 1960s [1]. Tabor and Winterton [2] and Israelachvili and Tabor [3] developed it to the current state of the art some 15 years before the invention of the more widely used atomic force microscope (AFM) (see [chapter B1.19](#)).

Although only a few dozen laboratories worldwide are actively practising the SFA technique, it has produced many notable findings and a fundamental understanding of surface forces [4]. These are applicable to various research fields such as polymer science [5, 6, 7, 8, 9 and 10], thin-film rheology [11, 12, 13 and 14], biology [15, 16, 17 and 19], liquid crystals [20, 21 and 22], food sciences [12, 23, 24], molecular tribology [9, 14, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34 and 35] and automotive tribology [36]—to name just a few. Numerous experimental extensions of the SFA technique have been developed to take advantage of the unique experimental set-up.

One of the most important extensions is the measurement of lateral forces (friction). Friction measurements have accompanied the SFA technique since its early beginnings in the Cavendish laboratory in Cambridge [37] and a variety of different lateral force measurements are practised throughout the SFA community.

B1.20.2 PRINCIPLES

B1.20.2.1 DIRECT FORCE MEASUREMENT

The measurement of surface forces calls for a rigid apparatus that exhibits a high force sensitivity as well as distance measurement and control on a subnanometre scale [38]. Most SFAs make use of an optical interference technique to measure distances and hence forces between surfaces. Alternative distance measurements have been developed in recent years—predominantly capacitive techniques, which allow for faster and simpler acquisition of an *averaged* distance [11, 39, 40] or even allow for simultaneous dielectric loss measurements at a confined interface.

The predominant method of measuring forces is to detect the proportional deflection of an elastic spring. The proportionality factor is commonly called the *spring constant* and, in the SFA, may range from some 10 N m^{-1} to some 100 kN m^{-1} . It is essential that the apparatus frame and surface compliance be at least 1000 times stiffer than the force-measuring spring. In a typical SFA, one of the two surfaces is attached to the force-measuring spring, while the other surface is rigidly mounted to the apparatus frame. In this set-up, shown

schematically in figure B1.20.1 a set of at least three parameters must be controlled or measured simultaneously for a direct force measurement.

Figure B1.20.1. Direct force measurement via deflection of an elastic spring—essential design features of a direct force measurement apparatus.

The accurate and absolute measurement of the distance, D , between the surfaces is central to the SFA technique. In a typical experiment, the SFA controls the base position, z_3 , of the spring and simultaneously measures D , while the *spring constant*, k , is a known quantity. Ideally, the simple relationship $\Delta F(D) = k\Delta(D - z_3)$ applies. Since surface forces are of limited range, one can set $F(D = \infty) = 0$ to obtain an absolute scale for the force. Furthermore, $\delta F(D = \infty)/\delta D \approx 0$ so that one can readily obtain a calibration of the distance control at large distances relying on an accurate measurement of D . Therefore, D and F are obtained at high accuracy to yield $F(D)$, the so-called *force versus distance curve*.

Most interferometric SFAs allow one to measure the distance, D , as a function of a selected lateral dimension, x' , which can be used to obtain information about the entire contact geometry. Knowledge of the contact geometry, together with the assumption that the underlying intermolecular forces are additive, allows the *intermolecular potential* $W(D)$ to be deduced from the measured $F(D)$ using the so-called *Derjaguin approximation* [4, 41]. For the idealized geometry of an undeformed sphere of radius R on a flat:

$$(B1.20.1)$$

where R becomes an effective radius $R' = (R, R_2)^{0.5}$ for the case of two cylinders (SFA) with radii R_1 and R_2 .

-3-

In accordance with [equation \(B1.20.1\)](#), one can plot the so-called *surface force parameter*, $P = F(D) / 2\pi R$, versus D . This allows comparison of different direct force measurements in terms of intermolecular potentials $W(D)$, i.e. independent of a particular contact geometry. Figure B1.20.2 shows an example of the attractive van der Waals force measured between two curved mica surfaces of radius $R \approx 10$ mm.

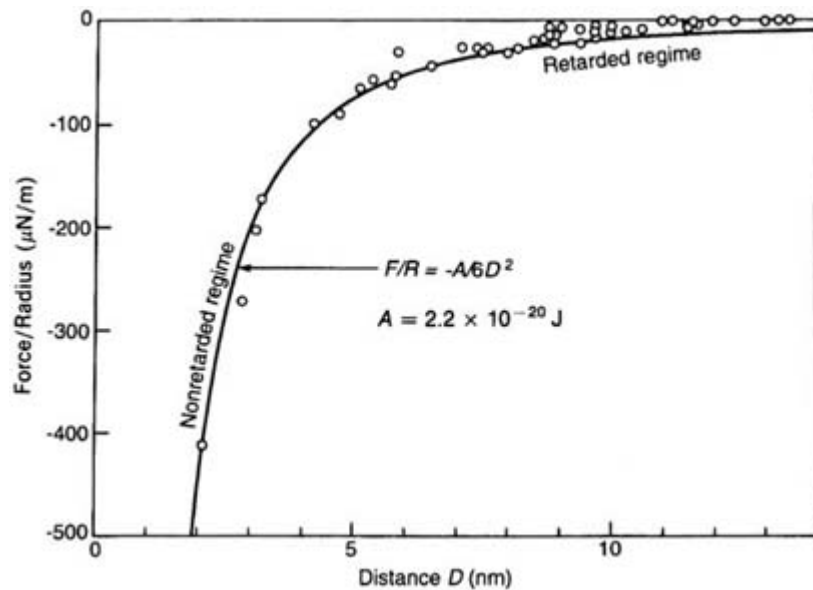


Figure B1.20.2. Attractive van der Waals potential between two curved mica surfaces measured with the SFA. (Reproduced with permission from [4], figure 11.6.)

An alternative way of measuring $F(D)$ is to control forces using magnetic fields instead of controlling

distances [42, 43]. An electronically controlled current through one or more coils creates a magnetic gradient, which exerts a force on the force-measuring spring. The surfaces are allowed to equilibrate at a distance, D , which is then measured. This alternative set-up allows one to measure $D(F)$ instead of $F(D)$ allowing for a true equilibrium force measurement.

B1.20.2.2 SAMPLE PREPARATION

Well defined contact geometry and absolute cleanliness are crucial factors for a successful SFA experiment. Therefore, two curved sheets of mica are brought into contact in crossed-cylinder geometry.

The sample preparation of these mica sheets is a delicate process that requires some experience and often takes 1–2 days prior to an SFA experiment. Through successive cleaving, one has to prepare 1–5 μm thick and uniform sheets of mica. Mica is a natural material that is available in different qualities [44].

Each newly cleaved mica surface is very clean. However, it is known that mica has a strong tendency to spontaneously adsorb particles [45] or organic contaminants [46], which may affect subsequent measurements. The mica sheets are cut into 10 mm \times 10 mm sized samples using a hot platinum wire, then laid down onto a thick and clean 100 mm \times 100 mm mica *backing sheet* for protection. On the backing sheet, the mica samples can be transferred into a vacuum chamber for thermal evaporation of typically 50–55 nm thick silver mirrors.

-4-

It was the idea of Winterton [2] to glue the otherwise fragile mica sheets onto polished *silica discs* to give them better mechanical stability, especially for friction experiments. The glue layer determines the final surface compliance of the silica/glue/mica stack which is typically around 4×10^{10} Pa. The use of mica samples from the same original sheet guarantees that the interferometer will be perfectly symmetrical (see section (b1.20.2.3)).

The silica discs that now hold the back-silvered mica samples are finally mounted into the SFA so that the cylinder axes are crossed and the clean mica surfaces are facing each other.

B1.20.2.3 MULTIPLE BEAM INTERFEROMETRY

The absolute measurement of the distance, D , between the surfaces is central to the SFA technique. In *interferometric* SFAs, it is realized through an optical method called multiple beam interferometry (MBI), which has been described by Tolansky [47].

A 50 nm film of metal (silver) is deposited onto the atomically smooth mica sheets. White light with a coherence length of some 10 μm is directed normally through the mica sheets to illuminate the contact zone. The mica–silver interfaces have a reflectivity of typically 97% and form an optical resonator. A constructive interference occurs for light that has a wavelength equal to half the optical distance between the mirrors. This resonance is called *interference fringe of chromatic order* $N = 1$. The larger the optical distance, the more this resonance shifts towards the red end of the spectrum. In analogy to an organ pipe, one also observes harmonic fringes at higher frequencies which are identified with integer numbers $N = 2, 3, \dots$. The emerging light from the silver/mica/mica/silver interferometer is focused onto the slit of an imaging spectrograph for further analysis. The light selected to enter the spectrograph entrance slit corresponds to a one-dimensional cut through the illuminated contact zone. The distance information along this (linear) dimension is maintained and the exit plane of the imaging spectrograph displays the interference pattern as a function of a position x' . Typically, one uses fringes in the visible spectrum ($15 < N < 35$) for distance measurements in the SFA, but other near-visible wavelengths can equally be employed. Generally, the distance resolution of MBI decreases linearly with increasing N and longer wavelengths used.

Because the mica surfaces are curved, the optical distance between the mirrors is a function of the lateral dimension, x' , i.e. $T = T(x')$. Therefore, the wavelength of a given fringe becomes itself a function of x' . Since the chromatic order, N , is invariant within a given fringe, these fringes are commonly called fringes of equal chromatic order (FECO) (see [figure B1.20.3.](#))

-5-

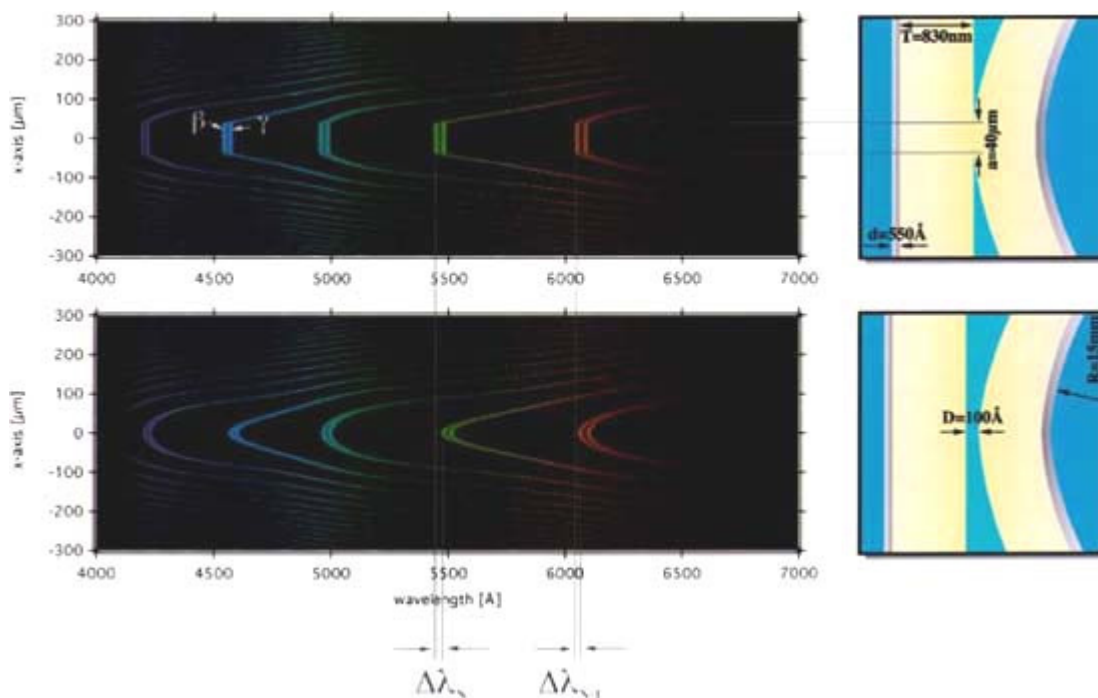


Figure B1.20.3. FECO allow optical distances in the SFA to be measured at subnanometre resolution. The FECOs depicted here belong to a symmetric, three-layer interferometer. The elastically deformed contact region appears as flattened fringes in the upper graph (vertical part). Once the surfaces are separated to a distance D in a medium of index n , the wavelength shifts are measured to calculate D and n . Since mica is birefringent, one can observe β and γ components for each fringe. To simplify data evaluation, birefringence can be suppressed using polarizing filters after the interferometer.

The problem of calculating surface separations from a given FECO pattern is in general far more complex than it may seem at first sight. The main reason is the fact that the refractive index is different for each of the different layers in the interferometer. For the simplest case of two mica surfaces in contact, surrounded by a medium, one has to solve for a one-layer interferometer inside the contact area and for a three-layer interferometer outside the contact area—that is where light travels partially through the medium. The analytical treatment of a three-layer interferometer is considerably simplified if the two mica sheets have exactly the same thickness, i.e. when the interferometer is *symmetric*. Based on the work of Hunter *et al* [48], Israelachvili [49] has derived the first useful analytical expressions and methodology for the symmetrical three-layer interferometer in the SFA. Clarkson [50] extended MBI with a numerical analysis of asymmetric three-layer interferometers by applying the multilayer matrix method [51]. Later, Horn *et al* [52] derived useful analytical expressions for the asymmetric interferometer.

Nonetheless, the symmetric interferometer remains very useful, because there, the wavelengths of fringes with *even* chromatic order, N , strongly depend on the refractive index, n_3 , of the central layer, whereas fringes with *odd* chromatic order are almost insensitive to n_3 . This lucky combination allows one to measure the thickness as well as the refractive index of a layer between the mica surfaces independently and simultaneously [49].

To simplify FECO evaluation, it is common practice to experimentally filter out one of the components by the use of a linear polarizer after the interferometer. Mica birefringence can, however, be useful to study thin films of birefringent molecules [49] between the surfaces. Rabinowitz [53] has presented an eigenvalue analysis of birefringence in the multiple beam interferometer.

Partial reflections at the inner optical interfaces of the interferometer lead to so-called *secondary* and *tertiary* fringe patterns as can be seen from figure B1.20.4. These additional FECO patterns become clearly visible if the reflectivity of the silver mirrors is reduced. Methods for analysis of such secondary and tertiary FECO patterns were developed to extract information about the topography of non-uniform substrates [54].

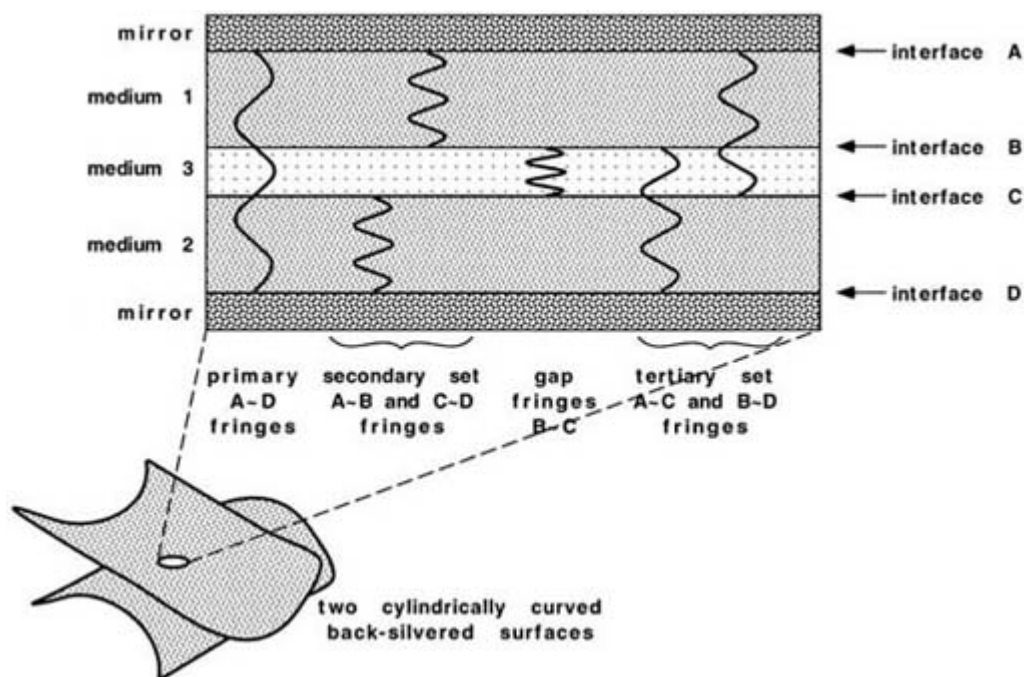


Figure B1.20.4. Cross-sectional sideview of a symmetric, three-layer interferometer illustrating the origin of primary, secondary, tertiary and gap FECOs. (Reproduced with permission from [54].)

In the *symmetric*, three-layer interferometer, only even-order fringes are sensitive to refractive index and it is possible to obtain spectral information of the confined film by comparison of the different intensities of odd- and even-order fringes. The absorption spectrum of thin dye layers between mica was investigated by Müller and Mächtle [55, 56] using this method.

Instead of an absorbing dye layer between the mica, Levins *et al* [57] used thin metallic films and developed a method for FECO analysis using an extended spectral range.

The preparation of the reflecting silver layers for MBI deserves special attention, since it affects the optical properties of the mirrors. Another important issue is the optical phase change [58] at the mica/silver interface, which is responsible for a wavelength-dependent shift of all FECOs. The phase change is a function of silver layer thickness, T , especially for $T < 40$ nm [54]. The roughness of the silver layers can also have an effect on the resolution of the distance measurement [59, 60].

Another interesting extension of the FECO technique, using a capillary droplet of mercury as the second mirror, was developed by Horn *et al* [61]. The light from this special interferometer is analysed in reflection,

which yields an inverted FECO pattern.

Every property of an interface that can be optically probed can, in principle, be measured with the SFA. This may include information obtainable from absorption spectroscopy [55], fluorescence, dichroism, birefringence, or nonlinear optics [43], some of which have already been realized.

B1.20.2.4 COMMON DESIGNS AND ATTACHMENTS

Israelachvili and Adams [62] designed one of the first SFAs, known as the Mk I, in the mid-1970s. Later, Israelachvili developed the Mk II [63, 64], which is based on the Mk I but has improved mechanics—in particular, a double cantilever spring to avoid surface tilt upon deflection (force), as well as a number of new attachments for a variety of different experimental set-ups. The Mk I and Mk II designs served as basis for further versions. More recent and improved versions include the Mk III developed by Israelachvili [65], as well as the circular steel Mk IV developed by Parker *et al* [66] and the circular glass SFA designed by Klein [67]. Some of the most common designs are schematically reproduced in [figure B1.20.5](#).

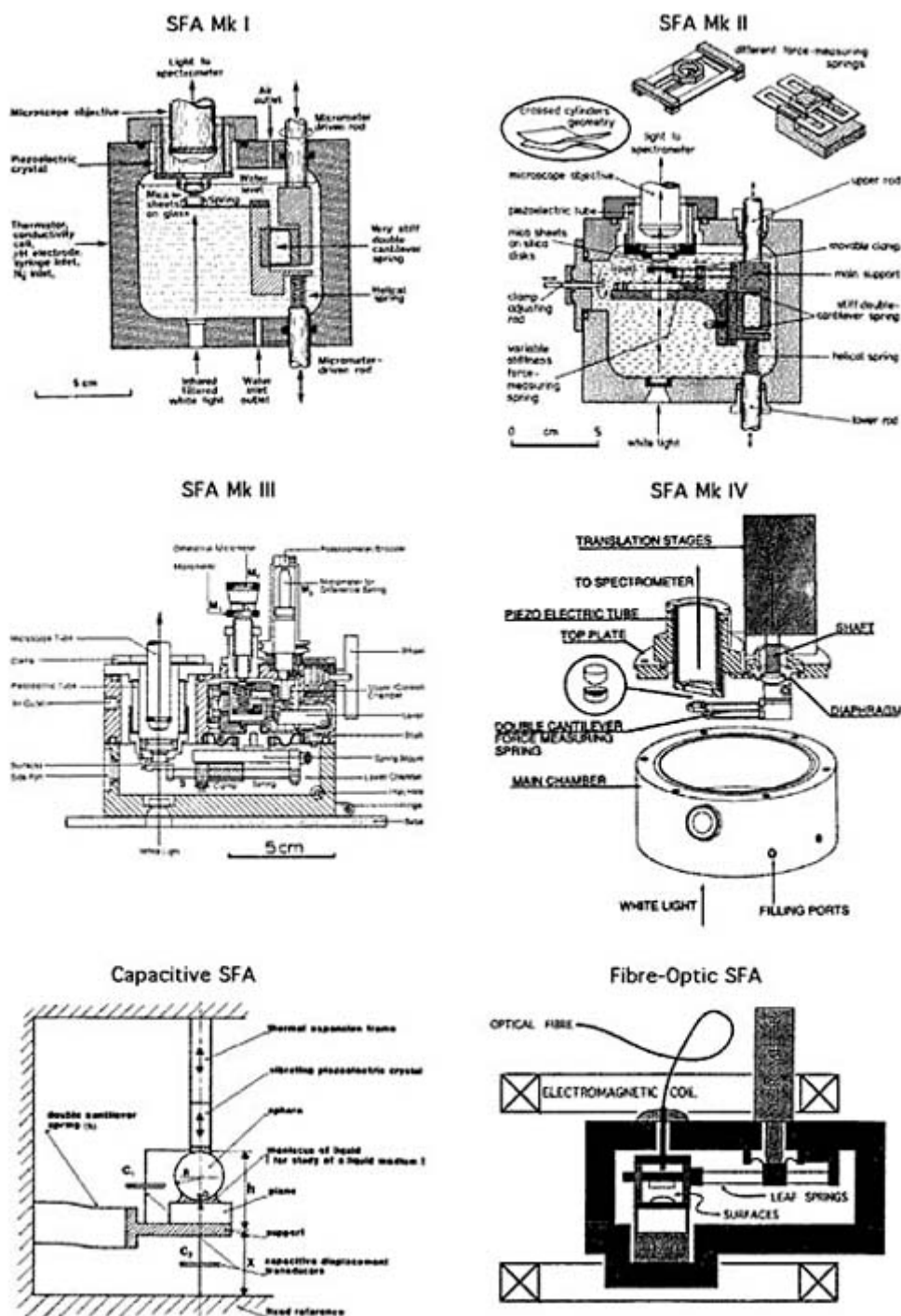


Figure B1.20.5. A selection of common SFA designs: SFA Mk I [62], SFA Mk II [64], SFA Mk III [65], SFA Mk IV [66], capacitive SFA [11], fibre-optic SFA [43]. Figures reproduced with permission from indicated references.

A considerable number of experimental extensions have been developed in recent years. Luckham *et al* [5] and Dan [68] review examples of dynamic measurements in the SFA. Studying the visco-elastic response of surfactant films [69] or adsorbed polymers [7, 9] promises to yield new insights into molecular mechanisms of frictional energy loss in boundary-lubricated systems [28, 70].

The measurement of lateral forces (friction and shear) in the SFA has recently been reviewed by Kumacheva

[32]. To measure friction and shear response, one has to laterally drive one surface and simultaneously measure the response of a lateral spring mount. A variety of versions have been devised. Lateral drives are often based on piezoelectric or bimorph deflection [13, 71] or DC motor drives, whereas the response can be measured via strain gauges, bimorphs, capacitive or optical detection.

Another promising extension uses x-rays to probe the structure of confined molecules [72].

In summary, the SFA is a versatile instrument that represents a unique platform for many present and future implementations. Unlike any other experimental technique, the SFA yields quantitative insight into molecular dimensions, structures and dynamics under confinement.

B1.20.3 APPLICATIONS

This section deals with some selected examples of typical SFA results, collected from various research areas. It is not meant to be a comprehensive review, rather a brief glance at the kind of questions that can be addressed with the SFA.

The earliest SFA experiments consisted of bringing the two mica sheets into contact in a controlled atmosphere (figure B1.20.6) or (confined) liquid medium [14, 27, 73, 74 and 75]. Later, a variety of surfactant layers [76, 77], polymer surfaces [5, 9, 10, 13, 68, 78], polyelectrolytes [79], novel materials [80] or biologically relevant molecular layers [15, 19, 81, 85 and 86] or model membranes [84, 87, 88] were prepared on the mica substrate. More recently, the SFA technique has also been extended to thick layers of other materials, such as silica [73, 89], polymer [10], as well as metals [59] and metal oxides [90, 91].

-10-

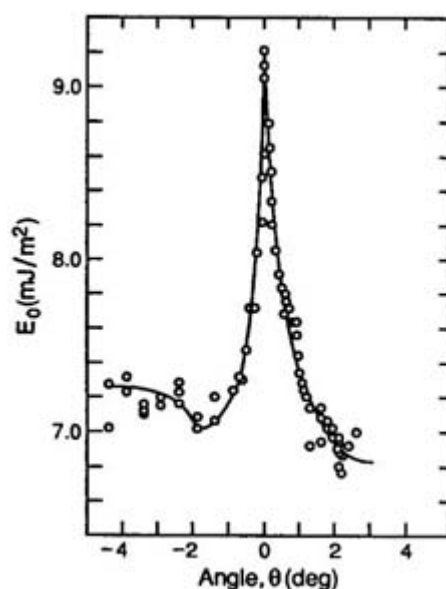


Figure B1.20.6. Short-range adhesion of a mica–mica contact as a function of the relative crystallographic orientation of the mica surfaces, measured in a dry nitrogen atmosphere. With permission from [94].

B1.20.3.1 MEASURING SHORT-RANGE SOLVATION AND HYDRATION FORCES

The measurement of surface forces out-of-plane (normal to the surfaces) represents a central field of use of the SFA technique. Besides the ubiquitous van der Waals dispersion interaction between two (mica) surfaces

in dry air ([figure B1.20.2](#)) and [figure B1.20.6](#), there is a wealth of other surface forces arising when the surfaces are brought into contact in a liquid medium. Many of these forces result from the specific properties of the liquid medium and originate from a characteristic ordering or reorientation of atoms or molecules—processes which are often entropy driven.

The well defined contact geometry and the ionic structure of the mica surface favours observation of *structural* and *solvation forces*. Besides a monotonic entropic repulsion one may observe superimposed periodic force modulations. It is commonly believed that these modulations are due to a metastable layering at surface separations below some 3–10 molecular diameters. These diffuse layers are very difficult to observe with other techniques [92]. The periodicity of these *oscillatory forces* is regularly found to correspond to the characteristic molecular diameter. [Figure B1.20.7](#) shows a typical measurement of solvation forces in the case of ethanol between mica.

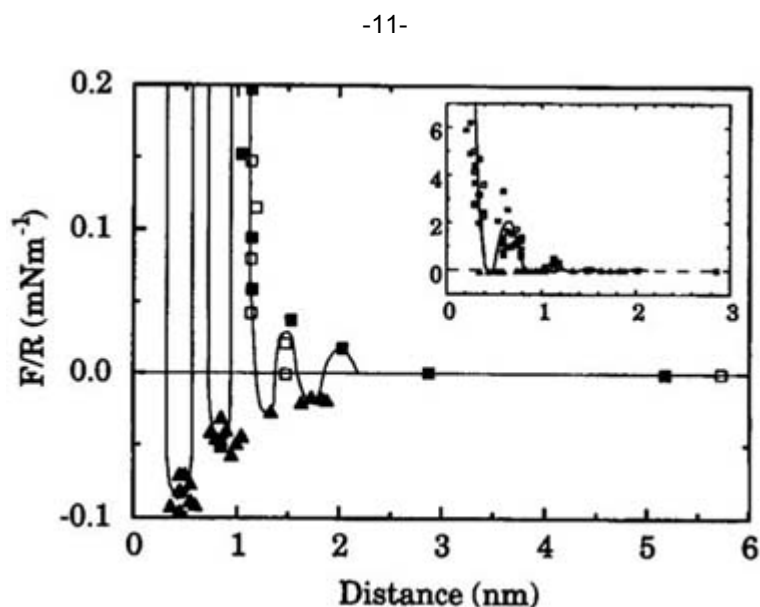


Figure B1.20.7. The solvation force of ethanol between mica surface. The inset shows the full scale of the experimental data. With permission from [75].

In the case of water, these forces are called *hydration forces* [93, 94]. The behaviour of water close to surfaces has attracted considerable attention due to its importance in the understanding of colloidal and biological interactions. Water seems to be a molecule of remarkable intermolecular interactions [95], mainly due to its capability to form hydrogen bonds. A number of aspects of water have been vigorously debated using results obtained with the SFA technique. These include the apparent viscosity in ultra-thin confined films [73] or the structure of water near surfaces, which is believed to give rise to *hydrophilic repulsion* or *hydrophobic attraction* [96, 97], or, indeed, the very origin of hydration forces and oscillatory forces [93].

B1.20.3.2 MEASURING LONG-RANGE DLVO-TYPE INTERACTIONS

In a given aqueous electrolyte there is a certain population of hydrated ions and counterions. In addition, many surfaces, including mica, exhibit a net charge in aqueous solution. Counterions are known to form a diffuse screening layer near a charged molecule, particle or surface with roughly exponentially decreasing concentration into the solution. The characteristic decay length of this *double layer* is called the *Debye length* and decreases with increasing ionic strength. This double layer gives rise to an entropically driven,

exponentially decreasing surface force, the so-called *double-layer repulsion*. Before the discovery of short-range forces (see above), it was commonly accepted that surface forces in liquid media could always be decomposed into an attractive dispersion component (van der Waals) and a repulsive double-layer force. These two forces are combined in the well known Derjaguin–Landau, Verwey–Overbeck (DLVO) theory and a number of systems have been measured in the SFA to confirm the predictions of this theory. Figure B1.20.8 shows results obtained in aqueous solutions of NaCl on silica surfaces. The ranges of the observed long-range repulsion forces are in good agreement with the DLVO theory. The inset nicely demonstrates the effect of the monotonic short-range forces described in [section \(b1.20.3.1\)](#). As a contrast to the results obtained on silica surfaces, [figure B1.20.9](#) schematically displays the measured DLVO-type forces between mica surfaces in aqueous electrolytes. The main differences are that the monotonic hydration (solvation) force is dependent on the salt concentration and that there are oscillatory forces superimposed. On the mica surface, the monotonic hydration force hence seems to be mainly the result of the presence of hydrated ions in the double-layer, whereas the silica surface is strongly hydrophilic and hence ‘intrinsically’ hydrated [94].

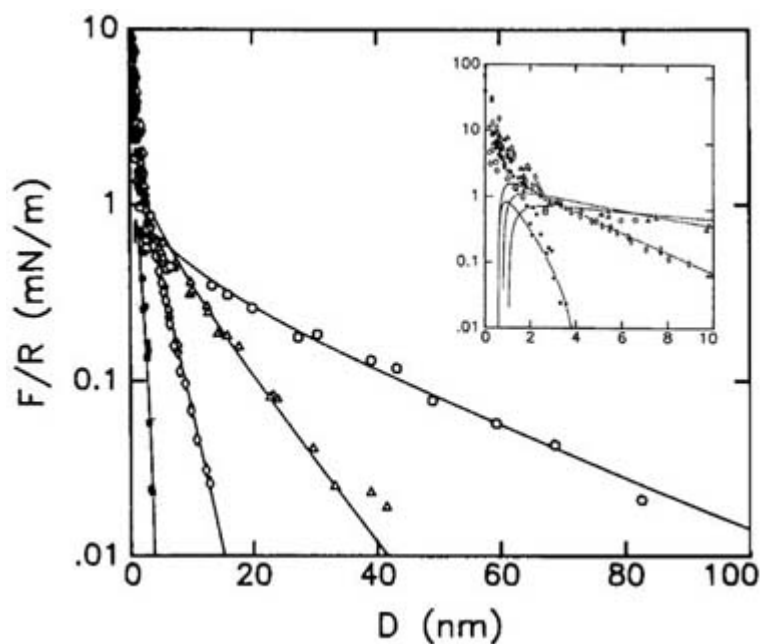


Figure B1.20.8. DLVO-type forces measured between two silica glass surfaces in aqueous solutions of NaCl at various concentrations. The inset shows the same data in the short-range regime up to $D = 10$ nm. The repulsive deviation at short range (< 2 nm) is due to a monotonic solvation force, which seems not to depend on the salt concentration. Oscillatory surface forces are not observed. With permission from [73].

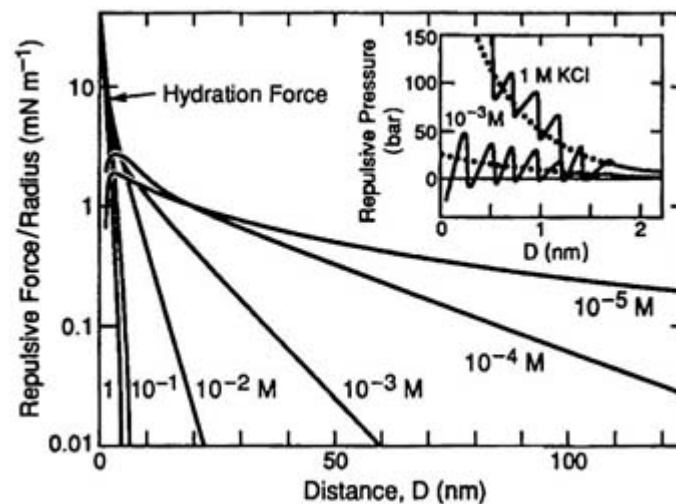


Figure B1.20.9. Schematic representation of DLVO-type forces measured between two mica surfaces in aqueous solutions of KNO_3 or KCl at various concentrations. The inset reveals the existence of oscillatory and monotonic structural forces, of which the latter clearly depend on the salt concentration. Reproduced with permission from [94].

B1.20.3.3 MEASURING BIOLOGICAL INTERACTIONS

Interactions between macromolecules (proteins, lipids, DNA, . . .) or biological structures (e.g. membranes) are considerably more complex than the interactions described in the two preceding paragraphs. The sum of all biological interactions at the molecular level is the basis of the complex mechanisms of *life*. In addition to computer simulations, direct force measurements [98], especially the surface forces apparatus, represent an invaluable tool to help understand the molecular interactions in biological systems.

Proteins can be physisorbed or covalently attached to mica. Another method is to immobilise and orient them by specific binding to receptor-functionalized planar lipid bilayers supported on the mica sheets [15]. These surfaces are then brought into contact in an aqueous electrolyte solution, while the pH and the ionic strength are varied. Corresponding variations in the force-versus-distance curve allow conclusions about protein conformation and interaction to be drawn [99]. The local electrostatic potential of protein-covered surfaces can hence be determined with an accuracy of ± 5 mV.

A typical force curve showing the specific avidin–biotin interaction is depicted in figure B1.20.10. The SFA revealed the strong influence of hydration forces and membrane undulation forces on the specific binding of proteins to membrane-bound receptors [81].

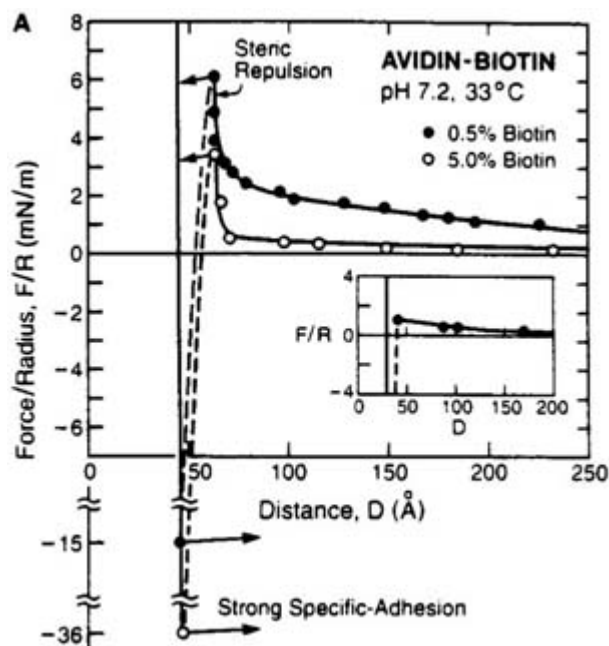


Figure B1.20.10. Typical force curve for a streptavidin surface interacting with a biotin surface in an aqueous electrolyte of controlled pH. This result demonstrates the power of specific protein interactions. Reproduced with permission from [81].

Direct measurement of the interaction potential between tethered ligand (biotin) and receptor (streptavidin) have been reported by Wong *et al* [16] and demonstrate the possibility of controlling range and dynamics of specific biologic interactions *via* a flexible PEG-tether.

The adhesion and fusion mechanisms between bilayers have also been studied with the SFA [88, 100]. Kuhl *et al* [17] found that solutions of short-chained polymers (PEG) could produce a short-range depletion attraction between lipid bilayers, which clearly depends on the polymer concentration (figure B1.20.11). This *depletion attraction* was found to induce a membrane fusion within 10 minutes that was observed, in real-time, using FECO fringes. There has been considerable progress in the preparation of fluid membranes to mimic natural conditions in the SFA [87], which promises even more exciting discoveries in biologically relevant areas.

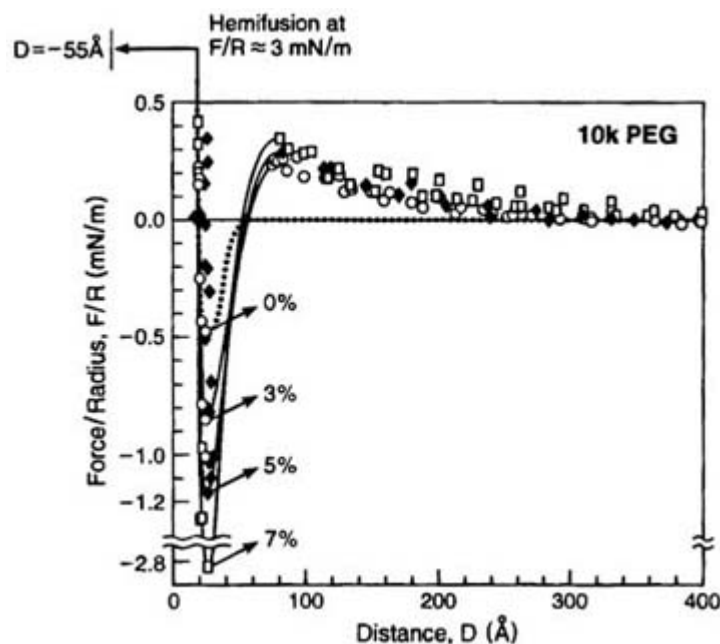


Figure B1.20.11. Force curves of DMPC/DPPE (dimyristoyl phosphatidylcholine and dipalmitoyl phosphatidylethanolamine) bilayers across a solution of PEG at different concentrations. Clearly visible is a concentration-dependent depletion attraction, with permission from [17].

B1.20.3.4 MEASUREMENTS IN MOLECULAR TRIBOLOGY

Using friction attachments (see section (b1.20.2.4)), many remarkable discoveries related to thin-film and boundary lubrication have been made with the SFA. The dynamic aspect of confined molecules at a sliding interface has been extensively investigated and the SFA had laid the foundation for *molecular tribology* long before the AFM technique was available.

The often-cited Amontons' law [101, 102] describes friction in terms of a friction coefficient, which is, *a priori*, a material constant, independent of contact area or dynamic parameters, such as sliding velocity, temperature or load. We know today that all of these parameters can have a significant influence on the magnitude of the measured friction force, especially in thin-film and boundary-lubricated systems.

Using the SFA technique, it could be demonstrated that there is an intimate relationship between adhesion hysteresis and friction [28, 29, 68, 77]. Both processes dissipate energy through non-equilibrium mechanisms at the interface [30]. Friction can be represented as a sum of two terms, one adhesion related and the other load-related. It was recently shown with the SFA that, in the absence of adhesion, the load related portion linearly depends on the load and not necessarily on the *real contact area*, as commonly believed [25]. [Figure B1.20.12](#) nicely illustrates this finding.

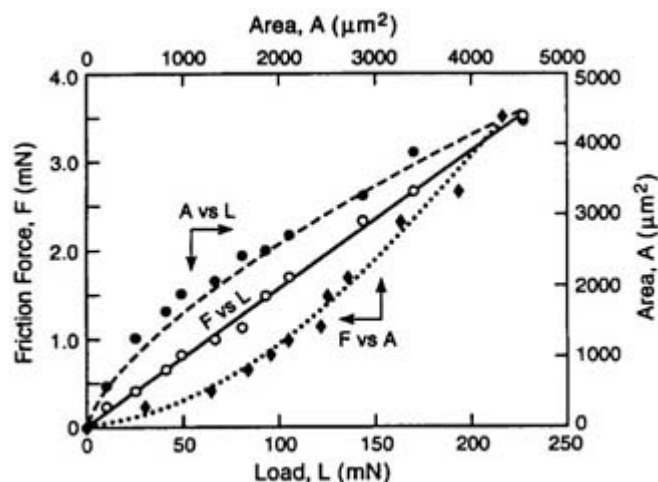


Figure B1.20.12. Measured friction force, F , and real contact area, A , against externally applied load, L , for two molecularly smooth mica surfaces sliding in 0.5 M KCl solution, i.e. in the absence of adhesion, with permission from [25].

A traditional subject of discussion is the phenomenon of intermittent friction, also called *stick – slip* friction [35]. It is thought that interfacial molecules can switch between different dynamic states, namely between a solid-like and a liquid-like state [103]. It was also found that liquids become oriented [7] and solid-like, when confined in a narrow gap. A diffuse layering of the trapped molecules may occur (see also section (B1.20.3.1)). The ordering mechanisms are particularly susceptible to the shape of the molecules and can spur substantial history and time effects, as illustrated in figure B1.20.13. Molecules in such ordered arrangements no longer behave as liquids and exhibit, for example, a finite yield stress [8]. When sliding above a critical speed, however, intermittent friction often disappears and the interface remains liquid-like at all times. This dynamic phenomenon, which occurs in the absence of hydrodynamic lubrication, is due to a time effect at the molecular level. Furthermore, on molecular layers of surfactants it was observed that, at even higher speeds, the system can enter a *superkinetic* regime with vanishingly small friction forces [70], as depicted in figure B1.20.14. In this high-speed regime, mechanisms of molecular entanglement are too slow to dissipate energy. New findings point out a more complex behaviour of such systems in terms of multiple relaxation mechanisms at the molecular interface [77]. It has also been shown experimentally [28] and by molecular dynamics simulation [104] that there is a potential to control interfacial dynamics with subnanometre out-of-plane excitations to achieve ultralow friction at arbitrarily slow sliding velocities.

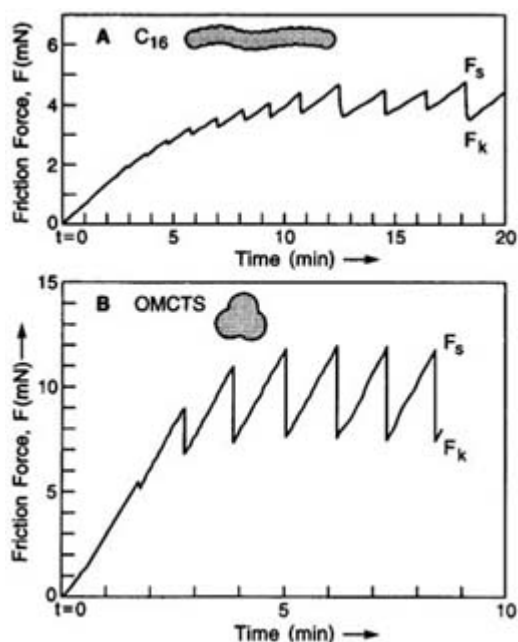


Figure B1.20.13. Temporal development of intermittent friction following commencement of sliding. The shape of the molecules has a great influence on history and time effects in the system. Reproduced with permission from [34].

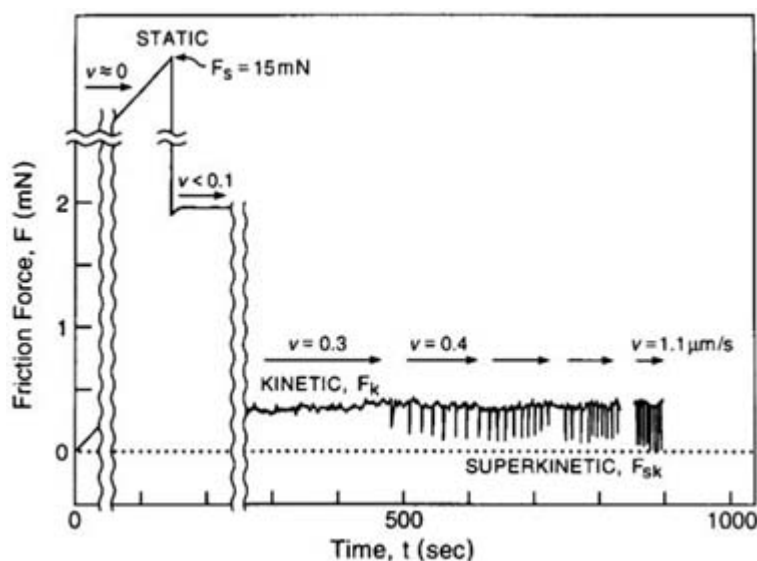


Figure B1.20.14. Dynamic friction at different velocities of DHDAA-coated mica (DHDAA = dihexadecyldimethylammonium acetate). Reproduced with permission from [70].

Polymer-bearing surfaces have also attracted considerable attention in view of their particular friction properties and underlying mechanisms. Klein *et al* [9] have demonstrated the possibility of achieving ultra-low friction on surfaces covered with polymer brushes. Figure B1.20.15 displays the strong lubricating effect, which is due to a very fluid layer at the interface between polymer and solvent—a layer that remains fluid even under high compression. Accompanying molecular dynamics simulations [105] suggest that a dynamic disentanglement of the polymer chains is responsible for the observed reduction in friction.

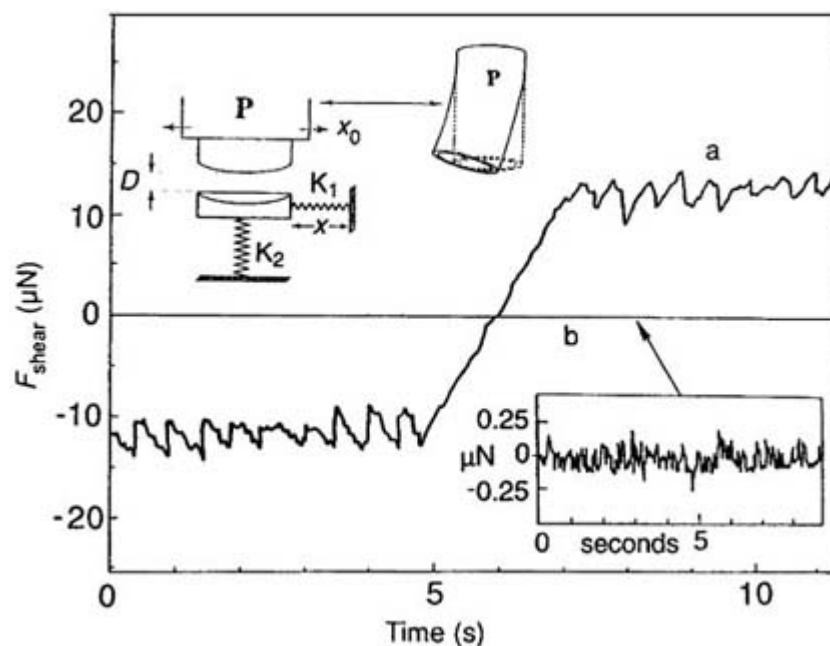


Figure B1.20.15. Shear force as a function of time for (a) bare mica in toluene and (b) polystyrene-covered mica in toluene. Reproduced with permission from [9].

REFERENCES

- [1] Bowden F P and Tabor D 1964 *Friction and Lubrication of Solids Part II* (Oxford: Oxford University Press)
- [2] Tabor D and Winterton R H S 1969 The direct measurement of normal and retarded van der Waals forces *Proc. R. Soc. London A* **312** 435–50
- [3] Israelachvili J N and Tabor D 1972 The measurement of van der Waals dispersion forces in the range 1.5 to 130 nm *Proc. R. Soc. London A* **331** 19–38
- [4] Israelachvili J N 1991 *Intermolecular and Surface Forces* 2nd edn (London: Academic)
- [5] Luckham P F and Manimaaran S 1997 Investigating adsorbed polymer layer behaviour using dynamic surface forces apparatuses—a review *Adv. Colloid Interface Sci.* **73** 1–46
- [6] Kelly T W *et al* 1998 Direct force measurements at polymer brush surfaces by atomic force microscopy *Macromolecules* **31** 4297–300
- [7] Dhinojwala A and Granick S 1997 Surface forces in the tapping mode: solvent permeability and hydrodynamic thickness of adsorbed polymer brushes *Macromolecules* **30** 1079–85
- [8] Luengo G, Israelachvili J N and Granick S 1996 Generalized effects in confined fluids: new friction map for boundary lubrication *Wear* **200** 328–35
- [9] Klein J *et al* 1994 Reduction of frictional forces between solid surfaces bearing polymer brushes *Nature* **370** 634–7
- [10] Mangipudi V S *et al* 1996 Measurement of interfacial adhesion between glassy polymers using the JKR method *Macromol. Symp.* **102** 131–43
- [11] Tonck A, Georges J M and Loubet J L 1988 Measurements of intermolecular forces and the rheology of dodecane between alumina surfaces *J. Colloid Interface Sci.* **126** 150–5
- [12] Borwankar R P and Case S E 1997 Rheology of emulsions, foams and gels *Curr. Opin. Colloid Interface Sci.* **2** 584–9

- [13] Luengo G *et al* 1997 Thin film rheology and tribology of confined polymer melts: contrasts with bulk properties *Macromolecules* **30** 2482–94
- [14] Demirel A L and Granick S 1996 Relaxations in molecularly thin liquid films *J. Phys. : Condens. Matter* **8** 9537–9
- [15] Leckband D 1995 The surface force apparatus—a tool for probing molecular protein interactions *Nature* **376** 617–18
- [16] Wong J Y *et al* 1997 Direct measurement of a tethered ligand–receptor interaction potential *Science* **275** 820–2
- [17] Kuhl T *et al* 1996 Direct measurement of polyethylene glycol induced depletion attraction between lipid bilayers *Langmuir* **12** 3003–14
- [18] Pincet F *et al* 1994 Long-range attraction between nucleosides with short-range specificity: direct measurements *Phys. Rev. Lett.* **73** 2780–3
- [19] Ionov R, De Coninck J and Angelova A 1996 On the origin of the long-range attraction between surface-confined DNA bases *Thin Solid Films* **284–285** 347–51
- [20] Richetti P *et al* 1996 Measurement of the interactions between two ordering surfaces under symmetric and asymmetric boundary conditions *Phys. Rev. E* **54** 1749–62
- [21] Idziak S H J *et al* 1996 Structure in a confined smectic liquid crystal with competing surface and sample elasticities *Phys. Rev. Lett.* **76** 1477–80
- [22] Idziak S H J *et al* 1996 Structure under confinement in a smectic-A and lyotropic surfactant hexagonal phase *Physica B* **221** 289–95
- [23] Giasson S, Israelachvili J N and Yoshizawa H 1997 Thin film morphology and tribology study of mayonnaise *J. Food Sci.* **62** 640–4
- [24] Luengo G *et al* 1997 Thin film rheology and tribology of chocolate *J. Food Sci.* **62** 767–72
- [25] Berman A, Drummond C and Israelachvili J N 1998 Amontons' law at the molecular level *Tribol. Lett.* **4** 95–101
- [26] Bhushan B, Israelachvili J N and Landman U 1995 Nanotribology: friction, wear and lubrication at the atomic scale *Nature* **374** 607–16
- [27] Granick S 1991 Motions and relaxations of confined liquids *Science* **253** 1374–9
- [28] Heuberger M, Drummond C and Israelachvili J N 1998 Coupling of normal and transverse motions during frictional sliding *J. Phys. Chem. B* **102** 5038–41
- [29] Israelachvili J N, Chen Y L and Yoshizawa H 1994 Relationship between adhesion and friction forces *J. Adhes. Sci. Technol.* **8** 1231–49
- [30] Israelachvili J N and Berman A 1995 Irreversibility, energy dissipation, and time effects in intermolecular and surface interactions *Israel J. Chem.* **35** 85–91

-20-

- [31] Krim J 1996 Friction at the atomic scale *Sci. Am.* **275** 74–80
- [32] Kumacheva E 1998 Interfacial friction measurement in surface force apparatus *Prog. Surf. Sci.* **58** 75–120
- [33] Reiter G, Demirel A L and Granick S 1994 From static to kinetic friction in confined liquid films *Science* **263** 1741–4
- [34] Yoshizawa H and Israelachvili J N 1993 Fundamental mechanisms of interfacial friction. 2. Stick–slip friction of spherical and chain molecules *J. Phys. Chem.* **97** 11 300–13
- [35] Yoshizawa H, McGuiggan P and Israelachvili J N 1993 Identification of a second dynamic state during stick-slip motion *Science* **259** 1305–8
- [36] Everson M P and Ohtani M 1998 New opportunities in automotive tribology *Tribol. Lett.* **5** 1–12
- [37] Tabor D 1992 *Fundamentals of Friction* ed I L Singer and H M Pollock (London: Kluwer)
- [38] Luesse C *et al* 1988 Drive mechanism for a surface force apparatus *Rev. Sci. Instrum.* **59** 811–2
- [39] Frantz P, Agrait N and Salmeron M 1996 Use of capacitance to measure surface forces. 1. Measuring distance of separation with enhanced spacial and time resolution *Langmuir* **12** 3289–94
- [40] Frantz P *et al* 1997 Use of capacitance to measure surface forces. 2. Application to the study of contact mechanics *Langmuir* **13** 5957–61
- [41] Derjaguin B V 1934 *Kolloid Zeitschrift* **69** 155–64

- [42] Stewart A M and Christenson H K 1990 Use of magnetic forces to control distance in a surface force apparatus *Meas. Sci. Technol* **12** 1301–3
- [43] Frantz P *et al* 1997 Design of surface forces apparatus for tribology studies combined with nonlinear optical spectroscopy *Rev. Sci. Instrum.* **68** 2499–2504
- [44] Ribbe P H (ed) 1984 *Micas Reviews in Mineralogy* vol 13 (Chelsea, MI: BookCrafters)
- [45] Ohnishi S *et al* 1999 Presence of particles on melt-cut mica sheets *Langmuir* **15** 3312–6
- [46] Frantz P and Salmeron M 1998 Preparation of mica surfaces for enhanced resolution and cleanliness in the surface forces apparatus *Tribol. Lett.* **5** 151–3
- [47] Tolansky S 1948 *Multiple Beam Interferometry of Surfaces and Films* (Oxford: Oxford University Press)
- [48] Hunter S C and Nabarro F R N 1952 The origin of Glauert's superposition fringes *Phil. Mag.* **43** 538–46
- [49] Israelachvili J N 1973 Thin film studies using multiple beam interferometry *J. Colloid Interface Sci.* **44** 259–71
- [50] Clarkson M T 1989 Multiple-beam interferometry with thin metal films and unsymmetrical systems *J. Phys. D: Appl. Phys.* **22** 475–82
- [51] Born M and Wolf E 1980 *Principles of Optics* 6th edn (Oxford: Pergamon)
- [52] Horn R G and Smith D T 1991 Analytical solution for the three-layer multiple beam interferometer *Appl. Opt.* **30** 59–65
- [53] Rabinowitz P 1995 Eigenvalue analysis of the surface forces apparatus interferometer *J. Opt. Soc. Am A* **12** 1593–601
- [54] Heuberger M, Luengo G and Israelachvili J N 1997 Topographic information from multiple beam interferometry in the surface forces apparatus *Langmuir* **13** 3839–48
- [55] Müller C, Mächtle P and Helm C A 1994 Enhanced absorption within a cavity. A study of thin dye layers with the surface forces apparatus *J. Phys. Chem.* **98** 11 119–25
- [56] Mächtle P, Müller C and Helm C A 1994 A thin absorbing layer at the center of a Fabry–Pérot interferometer *J. Physique II* **4** 481–500
- [57] Levins J M and Vanderlick T K 1994 Extended spectral analysis of multiple beam interferometry: a technique to study metallic films in the surface forces apparatus *Langmuir* **10** 2389–94

- [58] Farrell B, Bailey A I and Chapman D 1995 Experimental phase changes at the mica–silver interface illustrate the experimental accuracy of the central film thickness in a symmetrical three-layer interferometer *Appl. Opt.* **34** 2914–20
- [59] Levins J M and Vanderlick T K 1993 Impact of roughness of reflective films on the application of multiple beam interferometry *J. Colloid Interface Sci.* **158** 223–7
- [60] Levins J M and Vanderlick T K 1992 Reduction of the roughness of silver films by the controlled application of surface forces *J. Phys. Chem.* **96** 10 405–11
- [61] Horn R G *et al* 1996 The effect of surface and hydrodynamic forces on the shape of a fluid drop approaching a solid surface *J. Phys. : Condens. Matter* **8** 9483–90
- [62] Israelachvili J N and Adams G E 1976 Direct measurement of long range forces between two mica surfaces in aqueous KNO₃ solutions *Nature* **262** 774
- [63] Israelachvili J N 1987 Direct measurements of forces between surfaces in liquids at the molecular level *Proc. Nat. Acad. Sci. USA* **84** 4722–5
- [64] Israelachvili J N 1989 Techniques for direct measurement of forces between surfaces in liquids at the atomic scale *Chemtracts—Anal. Phys. Chem.* **1** 1–12
- [65] Israelachvili J N and McGuiggan P M 1990 Adhesion and short-range forces between surfaces. I. New apparatus for surface force measurements *J. Mater. Res.* **5** 2223–31
- [66] Parker J L, Christenson H K and Ninham B W 1989 Device for measuring the force and separation between two surfaces down to molecular separations *Rev. Sci. Instrum.* **60**
- [67] Klein J 1983 *J. Chem. Soc. Faraday Trans. I* **79** 99
- [68] Dan N 1996 Time-dependent effects in surface forces *Current Opinion Colloid Interface Sci.* **1** 48–52
- [69] Kutzner H B, Luckham P F and Rennie J 1996 Measurement of the viscoelastic properties of thin surfactant films

- [70] Yoshizawa H, Chen Y L and Israelachvili J N 1993 Recent advances in molecular level understanding of adhesion, friction and lubrication *Wear* **168** 161–6
- [71] Peachey J, van Alsten J and Granick S 1991 Design of an apparatus to measure the shear response of ultrathin liquid films *Rev. Sci. Instrum.* **62** 463–73
- [72] Idziak S H J *et al* 1994 The x-ray surface forces apparatus: structure of a thin smectic liquid crystal film under confinement *Science* **264** 1915–8
- [73] Horn R, Smith D T and Haller W 1989 Surface forces and viscosity of water measured between silica sheets *Chem. Phys. Lett.* **162** 404–8
- [74] Ruths M, Steinberg S and Israelachvili J N 1996 Effects of confinement and shear on the properties of thin films of thermotropic liquid crystal *Langmuir* **12** 6637–50
- [75] Wanless E J and Christenson H K 1994 Interaction between surfaces in ethanol: adsorption, capillary condensation, and solvation forces *J. Chem. Phys.* **101** 4260–7
- [76] Dedinaite A *et al* 1998 Interactions between modified mica surfaces in triglyceride media *Langmuir* **14** 5546–54
- [77] Yamada S and Israelachvili J N 1998 Friction and adhesion hysteresis of fluorocarbon surfactant monolayer-coated surfaces measured with the surface forces apparatus *J. Phys. Chem. B* **102** 234–44
- [78] Ruths M and Granick S 1998 Rate-dependent adhesion between opposed perfluoropoly (alkyl ether) layers: dependence on chain-end functionality and chain length *J. Phys. Chem. B* **102** 6056–63
- [79] Lowack K and Helm C A 1998 Molecular mechanisms controlling the self-assembly process of polyelectrolyte multilayers *Macromolecules* **31** 823–33
-

- [80] Luengo G *et al* 1997 Measurement of the adhesion and friction of smooth C-60 surfaces *Chem. Mater.* **9** 1166–71
- [81] Leckband D *et al* 1994 Direct force measurements of specific and nonspecific protein interactions *Biochemistry* **33** 4611–23
- [82] Chowdhury P B and Luckham P F 1995 Interaction forces between kappa-casein adsorbed on mica *Colloids Surfaces B* **4** 327–34
- [83] Holmberg M *et al* 1997 Surface force studies of Langmuir–Blodgett cellulose films *J. Colloid Interface Sci.* **186** 369–81
- [84] Kuhl T L *et al* 1994 Modulation of interaction forces between bilayers exposing short-chained ethylene oxide headgroups *Biophys. J.* **66** 1479–88
- [85] Nylander T and Wahlgren N M 1997 Forces between adsorbed layers of beta-casein *Langmuir* **13** 6219–25
- [86] Yu Z W, Calvert T L and Leckband D 1998 Molecular forces between membranes displaying neutral glycosphingolipids: evidence for carbohydrate attraction *Biochemistry* **37** 1540–50
- [87] Seitz M *et al* 1998 Formation of tethered supported bilayers via membrane-inserting reactive lipids *Thin Solid Films* **327–9** 767–71
- [88] Wolfe J *et al* 1991 The interaction and fusion of bilayers formed from unsaturated lipids *Eur. Biophys. J.* **19** 275–81
- [89] Rutland M W and Parker J L 1994 Surface forces between silica surfaces in cationic surfactant solutions: adsorption and bilayer formation at normal and high pH *Langmuir* **10** 1110–21
- [90] Xu Z H, Ducker W and Israelachvili J N 1996 Forces between crystalline alumina (sapphire) surfaces in aqueous sodium dodecyl sulfate surfactant solutions *Langmuir* **12** 2263–70
- [91] Horn R G, Clarke D R and Clarkson M T 1988 Direct measurement of surface forces between sapphire crystals in aqueous solutions *J. Mater. Res.* **3** 413–6
- [92] Cleveland J P, Schäffer T E and Hansma P K 1995 Probing oscillatory hydration potentials using thermal-mechanical noise in an atomic force microscope *Phys. Rev. B* **52** R8692–5
- [93] Israelachvili J N and Pashley R M 1983 Molecular layering of water at surfaces and origin of repulsive hydration forces *Nature* **306** 249–50
- [94] Israelachvili J N, McGuiggan P and Horn R 1992 Basic physics of interactions between surfaces in dry, humid, and aqueous environments *1st Int. Symp. on Semiconductor Wafer Bondings: Science, Technology and Applications* (Pennington, NJ: Electrochemical Society)

- [95] Stanley H E 1999 Unsolved mysteries of water in its liquid and glass states *MRS Bull.* May 22–30
- [96] Israelachvili J N 1996 Role of hydration and water structure in biological and colloidal interactions *Nature* **379** 219–25
- [97] Müller H J 1998 Extraordinarily thick water films on hydrophilic solids: a result of hydrophobic repulsion? *Langmuir* **14** 6789–92
- [98] Pierres A, Benoliel A M and Bongrand P 1996 Measuring bonds between surface-associated molecules *J. Immunol. Methods* **196** 105–20
- [99] Leckband D *et al* 1993 Measurements of conformational changes during adhesion of lipid and protein (polylysine and S-layer) surfaces *Biotech. Bioeng.* **42** 167–77
- [100] Helm C A, Israelachvili J N and McGuiggan P M 1992 Role of hydrophobic forces in bilayer adhesion and fusion *Biochemistry* **31** 1794–805
- [101] Amontons G 1699 De la résistance causé dans les machines *Mémoires de l'Académie Royale A* 275–82
- [102] Coulomb C A 1785 Théorie des machines simples *Mémoire de Mathématique et de Physique de l'Académie Royale* 161–342
- [103] Gee M L *et al* 1990 Liquid to solidlike transitions of molecularly thin films under shear *J. Chem. Phys.* **93** 1895–905
- [104] Gao J, Luedtke W and Landman U 1998 Friction control in thin-film lubrication *J. Phys. Chem. B* **102** 5033–7
- [105] Grest G S 1996 Interfacial sliding of polymer brushes: a molecular dynamics simulation *Phys. Rev. Lett.* **76** 4979–82
-

-23-

FURTHER READING

Derjaguin B V 1934 *Research in Surface Forces* (New York: Consultants Bureau)

An old classic: four volumes.

Israelachvili J N 1991 *Intermolecular and Surface Forces* (London: Academic)

The most often cited reference about surface forces and SFA.

Hutchings I M 1992 *Friction and Wear of Engineering Materials* (London: Arnold)

A good introduction to tribology.

Bhushan B 1999 *Handbook of Micro/Nano Tribology* (Boca Raton, FL: CRC)

A valuable reference to anyone involved with friction at small scales.

-1-

B1.21 Surface structural determination: diffraction methods

Michel A Van Hove

B1.21.1 INTRODUCTION

Diffraction methods have provided the large majority of solved atomic-scale structures for both the bulk materials and their surfaces, mainly in the crystalline state. Crystallography by diffraction tends to filter out defects and focus on the periodic part of a structure. By adding contributions from very many unit cells, diffraction gives results that are, in effect, averaged over space and time. This is excellent for investigating stable states of solid matter as they occur in well crystallized samples; some forms of disorder can also be analysed reasonably well. Diffraction, however, is much less appropriate for examining inhomogeneous and time-dependent events such as transition states and pathways in chemical reactions.

For bulk structural determination (see [chapter B1.9](#)), the main technique used has been x-ray diffraction (XRD). Several other techniques are also available for more specialized applications, including: electron diffraction (ED) for thin film structures and gas-phase molecules; neutron diffraction (ND) and nuclear magnetic resonance (NMR) for magnetic studies (see [chapter B1.12](#) and [chapter B1.13](#)); x-ray absorption fine structure (XAFS) for local structures in small or unstable samples and other spectroscopies to examine local structures in molecules. Electron microscopy also plays an important role, primarily through imaging (see [chapter B1.17](#)).

At surfaces, the primary challenge is to obtain the desired surface sensitivity. Ideally, one wishes to gain structural information about those atomic layers which differ in their properties from the underlying bulk material. This means in practice extracting the structure of the first few monolayers, i.e. atoms within about 5–10 Å (0.5–1 nm) of the vacuum above the surface. The above-mentioned bulk methods, if applied unchanged, do not easily provide sensitivity to this very thin slice of matter. The challenge becomes even greater when dealing with an interface between two materials, including solid/liquid and solid/gas interfaces. A number of mechanisms are available to obtain surface sensitivity on the required depth scale. We shall describe some of them in the next section, with emphasis on the solid/vacuum interface.

However, it is necessary to first discuss the meaning of ‘diffraction’, because this concept can be interpreted in several ways. After these fundamental aspects are dealt with, we will take a statistical and historical view of the field. It will be seen that many different diffraction methods are available for surface structural determination.

It will also be useful to introduce concepts of two-dimensional ordering and the corresponding nomenclature used to characterize specific structures. We can then describe how the surface diffraction pattern relates to the ordering and, thus, provides important two-dimensional structural information.

We will, in the latter part of this discussion, focus only on those few methods that have been the most productive, with low-energy electron diffraction (LEED) receiving the most attention. Indeed, LEED has been the most successful surface structural method in two quite distinct ways. First, LEED has become an almost universal characterization

technique for single-crystal surfaces: the diffraction pattern is easily imaged in real time and is very helpful in monitoring the state of the surface in terms of the ordering and, hence, also density, of adsorbed atoms and molecules. Second, LEED has been quite successful in determining the detailed atomic positions at a surface (e.g., interlayer distances, bond lengths and bond angles), especially for ordered structures. This relies primarily on simulating the intensity (current) of diffracted beams as a function of electron energy in order to fit assumed model structures to measured data. Because of multiple scattering, such simulation and fitting is a very different and much more difficult task than looking at a diffraction pattern. We will close with a

description of the state of the art and an outlook on the future of the field.

B1.21.2 FUNDAMENTALS OF SURFACE DIFFRACTION METHODS

B1.21.2.1 DIFFRACTION

(A) DIFFRACTION AND STRUCTURE

Diffraction is based on wave interference, whether the wave is an electromagnetic wave (optical, x-ray, etc), or a quantum mechanical wave associated with a particle (electron, neutron, atom, etc), or any other kind of wave. To obtain information about atomic positions, one exploits the interference between different scattering trajectories among atoms in a solid or at a surface, since this interference is very sensitive to differences in path lengths and hence to relative atomic positions (see [chapter B1.9](#)).

It is relatively straightforward to determine the *size and shape of the three- or two-dimensional unit cell* of a periodic bulk or surface structure, respectively. This information follows from the *exit directions* of diffracted beams relative to an incident beam, for a given crystal orientation: measuring those exit angles determines the unit cell quite easily. But no *relative positions* of atoms within the unit cell can be obtained in this manner. To achieve that, one must measure *intensities* of diffracted beams and then computationally analyse those intensities in terms of atomic positions.

With XRD applied to bulk materials, a detailed structural analysis of atomic positions is rather straightforward and routine for structures that can be quite complex (see [chapter B1.9](#)): *direct methods* in many cases give good results in a single step, while the resulting atomic positions may be refined by iterative fitting procedures based on simulation of the diffraction process.

With ED, by contrast, the task is more complicated due to *multiple scattering* of the electrons from atom to atom (see [chapter B1.17](#)). Such multiple scattering is especially strong at the relatively low energies employed to study surfaces. This dramatically restricts the application of direct methods and strongly increases the computational cost of simulating the diffraction process. As a result, an iterative *trial-and-error fitting* is the method of choice with ED, even though it can be a slow process when many trial structures have to be tested.

Also, the result of any diffraction-based trial-and-error fitting is not necessarily unique: it is always possible that there exists another untried structure that would give a better fit to experiment. Hence, a multi-technique approach that provides independent clues to the structure is very fruitful and common in surface science: such clues include chemical composition, vibrational analysis and position restrictions implied by other structural methods. This can greatly restrict the number of trial structures which must be investigated.

(B) NON-PERIODIC STRUCTURES

Diffraction is not limited to periodic structures [1]. Non-periodic imperfections such as defects or vibrations, as well as sample-size or domain effects, are inevitable in practice but do not cause much difficulty or can be taken into account when studying the ordered part of a structure. Some other forms of disorder can also be handled quite well in their own right, such as *lattice-gas* disorder in which a given site in the unit cell is randomly occupied with less than 100% probability. At surfaces, lattice-gas disorder is very common when atoms or molecules are adsorbed on a substrate. The local adsorption structure in the given site can be studied in detail.

(C) NON-PLANAR INITIAL WAVES

More fundamental is the distinction between planar and spherical initial waves. In XRD, for instance, the incident x-rays are well described by plane waves; this is generally true of probes that are aimed at the sample from macroscopic distances, as is the case also in most forms of ED and ND. However, there are techniques in which a wave is generated locally within the sample, for instance through emission of an x-ray (by fluorescence) or an electron (by photoemission) from a sample atom. In such *point-source emission*, the wave which performs the useful diffraction initially has a spherical rather than planar character; it is centred on the nucleus of an atom, with a rapidly decaying amplitude as it travels away from the emitting site. (Depending on the excitation mechanism, this initial wave need not be spherically symmetrical, but may also have an angular variation, as given by spherical harmonics, for instance, or combinations thereof.)

This spherical outgoing wave can diffract only from atoms that are near to the emitting atom, mainly those atoms within a distance of a few atomic diameters. In these circumstances, the crystallinity of the sample is of less importance: the diffracting wave sees primarily the *local* atomic-scale neighbourhood of the emitting atoms. As long as the same local neighbourhood predominates everywhere in the sampled part of the surface, information about the structure of that neighbourhood can be extracted. It also helps very much if the local neighbourhood has a constant orientation, so that the experiment does not average over a multitude of orientations, since these tend to average out diffraction effects and thus wash away structural information.

(D) VARIETY OF DIFFRACTION METHODS

From the above descriptions, it becomes apparent that one can include a wide variety of techniques under the label ‘diffraction methods’. [Table B1.21.1](#) lists many techniques used for surface structural determination, and specifies which can be considered diffraction methods due to their use of wave interference ([table B1.21.1](#) also explains many technique acronyms commonly used in surface science). The diffraction methods range from the classic case of XRD and the analogous case of LEED to much more subtle cases like XAFS (listed as both SEXAFS (surface extended XAFS) and NEXAFS (near-edge XAFS) in the table).

-4-

Table B1.21.1. Surface structural determination methods. The second column indicates whether a technique can be considered a diffraction method, in the sense of relying on wave interference. Also shown are statistics of surface structural determinations, extracted from the Surface Structure Database [[14](#)], up to 1997. Counted here are only ‘detailed’ and complete structural determinations, in which typically the experiment is simulated computationally and atomic positions are fitted to experiment. (Some structural determinations are performed by combining two or more methods: those are counted more than once in this table, so that the columns add up to more than the actual 1113 structural determinations included in the database.)

Surface structural determination method	Diffraction method?	Number of structural determinations	Percentage of structural determinations
LEED	yes	751	67.5
IS (including LEIS, MEIS and HEIS for low-, medium- and high-energy ion scattering)	no	102	9.2

PD (covers a variety of other acronyms, like ARPEFS, ARXPD, ARXPS, ARUPS, NPD, OPD, PED)	yes	88	7.9
SEXAFS	yes	67	6.0
XSW	yes	52	4.7
XRD (also GIXS, GIXD)	yes	40	3.6
TOF-SARS (time-of-flight scattering and recoiling spectrometry)	no	13	1.2
NEXAFS (also called XANES)	yes	11	1.0
RHEED	yes	10	0.9
LEPD	yes	5	0.4
HREELS	yes	4	0.4
MEED	yes	3	0.3
AED	yes	3	0.3
SEELFS (surface extended energy loss fine structure)	yes	2	0.2
TED	yes	1	0.1
AD	yes	1	0.1
STM	no	1	0.1

XAFS is a good example of less obvious diffraction [2, 3]. In XAFS, an electron is emitted by an x-ray locally within the sample. It propagates away as a spherical wave, which is allowed to back-scatter from neighbouring atoms to the emitter atom. The back-scattered electron wave interferes at the emitting atom with the emitted wave, thereby modulating the probability of the emitting process itself when the energy (wavelength) is varied: as one cycles through constructive and destructive interferences, the emission probability oscillates with a period that reflects the interatomic distances. This emission probability is, however, measured through yet another process (e.g., absorption of the incident x-rays, or emission of other x-rays or other electrons), which oscillates in synchrony with the interference. Thus, the structure-determining diffraction is in such a case buried relatively deeply in the overall process, and does not closely resemble the classic plane-wave diffraction of XRD.

B1.21.2.2 SURFACE SENSITIVITY

There are several approaches to gain the required surface sensitivity with diffraction methods. We review several of these here, emphasizing the case of solid/vacuum interfaces; some of these also apply to other interfaces.

(A) SHORT MEAN FREE PATH

One obvious method to obtain surface sensitivity is to choose probes and conditions that give shallow penetration. This can be achieved through a short *mean free path* λ , i.e. a short average distance until the probe (e.g., x-ray or electron) is absorbed by energy loss or is otherwise removed from the useful diffraction channels. For typical x-rays, λ is of the order of micrometres in many materials, which is too large compared to the desired surface thickness [4]. But for electrons of low kinetic energies, i.e. $E \approx 10\text{--}1000$ eV, the mean free path λ is of the order of 5–20 Å [5]. The mean free path has a minimum in the 100–200 eV range, with larger mean free paths existing both below and above this range.

Such ideal low mean free paths are the basis of LEED, the technique that has been used most for determining surface structures on the atomic scale. This is also the case of photoelectron diffraction (PD): here, the mean free path of the emitted electrons restricts sensitivity to a similar depth (actually double the depth of LEED, since the incident x-rays in PD are only weakly attenuated on this scale).

(B) GRAZING INCIDENCE AND/OR EMERGENCE

Another approach to limit the penetration of the probe into the surface region is to use *grazing incidence* and/or *grazing emergence*; this works for those probes that already have a reasonably small mean free path λ . A grazing angle θ (measured from the surface normal, i.e., θ close to 90°) then allows the probe to penetrate to a depth of only about $\lambda \cos(\theta)$. This approach is used primarily for higher-energy electrons above about 1000 eV in a technique called reflection high-energy electron diffraction (RHEED) [6].

With XRD, however, the mean free path is still too long to make this approach practical by itself [4]: as an example, to obtain even 100 Å penetration, one would typically need to use a grazing angle of about 0.05° , which is technically extremely demanding. The penetration depth is proportional to the grazing angle of incidence at such small angles, so that a ten times smaller penetration depth requires a further tenfold reduction in grazing angle. In addition, such small grazing angles require samples with a flatness that is essentially impossible to achieve, in order that the x-rays see a flat surface rather than a set of ridges that shadow much of the surface.

-6-

(C) TOTAL EXTERNAL REFLECTION

In XRD, surface sensitivity can, however, be achieved through another phenomenon [4]: *total external reflection*. This also occurs at grazing angles of incidence, giving rise to the technique acronym of GIXS for grazing-incidence x-ray scattering. At angles within approximately 0.5° of $\theta = 90^\circ$, x-rays cannot penetrate by refraction into materials: the laws of optics imply that the wave velocity of refracted waves in the material would have to be larger than the speed of light under those circumstances, which is impossible for propagating waves. Instead, the incident wave is totally reflected. However, this is accompanied by a shallow penetration of waves that decay exponentially into the bulk while propagating parallel to the surface. Under such conditions, the decay length into the surface is of the order of 10–30 Å, as desired. This penetration depth depends on the material and not on the wavelength of the x-rays. Note that total external reflection does not require vacuum: it can occur at various kinds of interfaces, depending on the relative optical constants of the phases in contact.

(D) HIGH-SURFACE AREA MATERIALS

None of the above methods is sufficient for neutrons, however. Neutrons penetrate matter so easily that the only effective approach is to use materials with a very high surface-to-volume ratio. This can be accomplished with small particles and exfoliated graphite, for instance, but the technique has essentially been abandoned in surface studies [7, 8].

(E) SUPERLATTICE DIFFRACTION

One further method for obtaining surface sensitivity in diffraction relies on the presence of two-dimensional *superlattices* on the surface. As we shall see further below, these correspond to periodicities that are different from those present in the bulk material. As a result, additional diffracted beams occur (often called fractional-order beams), which are uniquely created by and therefore sensitive to this kind of surface structure. XRD, in particular, makes frequent use of this property [4]. Transmission electron diffraction (TED) also has used this property, in conjunction with ultrathin samples to minimize bulk contributions [9].

(F) HYBRID METHODS

As we have seen, the electron is the easiest probe to make surface sensitive. For that reason, a number of hybrid techniques have been designed that combine the virtues of electrons and of other probes. In particular, electrons and photons (x-rays) have been used together in techniques like PD [10] and SEXAFS (or EXAFS, which is the high-energy limit of XAFS) [2, 11]. Both of these rely on diffraction by electrons, which have been excited by photons. In the case of PD, the electrons themselves are detected after emission out of the surface, limiting the depth of ‘sampling’ to that given by the electron mean free path.

(G) ELEMENTAL AND CHEMICAL-STATE RESOLUTION

With some techniques, another mechanism can give high surface sensitivity, namely *elemental resolution* through spectroscopic filtering of emitted electrons or x-rays. In this approach, one detects, by setting an energy window, only those electrons or x-rays that are emitted by a particular kind of atom, since each electronic level produces a line at a particular energy given by the level energy augmented by the excitation energy.

-7-

Thus, if a ‘foreign’ element is present only at the surface, one can detect a signal that only comes from that element and, therefore, only from the surface. Given sufficient energy resolution, one can even differentiate electrons coming from the same atoms in different bonding environments: e.g., in the case of a clean surface, atoms of the outermost layer *versus* bulk atoms [10]. This *chemical-state resolution* is due to the fact that electronic levels are shifted by bonding to other atoms, resulting in different emitted lines from atoms in different bonding situations.

Elemental and chemical-state resolution affords the possibility of detecting only a monolayer or even a fraction of a monolayer. This approach is prevalent in PD and in methods based on x-ray fluorescence.

It is also used in SEXAFS [11]: as we have seen, photoexcited electrons are back-reflected to the photoemitting atoms, thereby modulating the x-ray absorption cross-section through electron wave interference, after which a secondary electron or ion or fluorescent x-ray is ejected from the surface and finally detected. This latter ejection process provides surface sensitivity, through the electronic mean free path or the shallowness of ionic emission. However, elemental and chemical-state selection by energy filtering is essentially universal here, and again can give monolayer resolution with emission from foreign surface atoms different from the bulk atoms.

A similar device can be applied to a form of x-ray diffraction called the x-ray standing wave (XSW) method [12, 13], as detected by fluorescence. Here, x-ray waves reflected from bulk atomic planes form a standing wave pattern near the surface. The maxima and minima of this standing wave pattern can be arranged to fall at different locations on the atomic scale, by varying the energy and incidence angles. Thereby, the induced fluorescence varies with the location of those maxima and minima. Since the fluorescence is element specific, one can thus determine positions of foreign surface atoms relative to the extended bulk lattice (it remains

difficult, however, to locate those substrate atoms that are close to the fluorescing surface atoms, because they are drowned by the bulk signal).

B1.21.3 STATISTICS OF FULL STRUCTURAL DETERMINATIONS

Many methods have been developed to determine surface structure: we have mentioned several in the previous section and there are many more. To get an idea of their relative usage and importance, we here examine historical statistics. We also review the kinds of surface structure that have been studied to date, which gives a feeling for the kinds of surface structures that current methods and technology can most easily solve. This will provide an overview of the range of surfaces for which detailed surface structures are known, and those for which very little is known.

As source of information we use the Surface Structure Database [14], a critical compilation of surface structures solved in detail, covering the period to the end of 1997. It contains 1113 structural determinations with, on average, two determinations for each structure: thus there are approximately 550 distinct solved structures available.

In terms of individual techniques, [table B1.21.1](#) lists the breakdown totalled over time, counting from the inception of surface structural determination in the early 1970s. It is seen that LEED has contributed altogether about 67% of all structural determinations included in the database. The annual share of LEED was 100% until 1978, and has generally remained over 50% since then. In 1979 other methods started to produce structural determinations, especially PD, ion scattering (IS) and SEXAFS. XRD and then XSW started to contribute results in the period 1981–3.

-8-

As the table shows, a host of other techniques have contributed a dozen or fewer results each. It is seen that diffraction techniques have been very prominent in the field: the major diffraction methods have been LEED, PD, SEXAFS, XSW, XRD, while others have contributed less, such as NEXAFS, RHEED, low-energy position diffraction (LEPD), high-resolution electron energy loss spectroscopy (HREELS), medium-energy electron diffraction (MEED), Auger electron diffraction (AED), SEELFS, TED and atom diffraction (AD). The major non-diffraction method is IS, which is described in [chapter B1.23](#).

The database provides interesting perspectives on the evolution of surface structural determination since its inception around 1970. Not surprisingly, there is a clear temporal trend toward more complex and more diverse materials, such as compound substrates, alloyed bimetallic surfaces, complex adsorbate-induced relaxations and reconstructions, epitaxial and pseudomorphic growth, alkali adsorption on semiconductor and transition metal substrates, and molecular adsorbates as well as co-adsorbates on metal surfaces. The complexity of some solved structures has grown to about 100 times that of the earliest structures. The range of structure types can also be gauged, for instance, from the list of substrate lattice categories included in the SSD database: bcc, CdCl₂, CdI₂, corundum, CsCl, CuAu I, Cu₃Au, diamond, fcc, fluorite, graphite, hcp, hexagonal, NaCl, perovskite, rutile, spinel, wurtzite, zincblende, 2H–MoS₂, 2H–NbSe₂ and 6H–SiC.

Nonetheless, when counting all structures solved over time, one finds a strong predominance of studies in certain narrow categories, as exhibited by the following uneven statistics:

- fcc metals far outdistance any other substrate lattice type, with 60% of the total;
- the diamond lattice (C, Si and Ge) forms the next most numerous lattice category, about 10%, followed by the bcc (9%) and hcp (7%) lattices;
- elemental solids (with or without foreign adsorbates) form 85% of the substrates examined, the rest being metallic alloys (7%) or other compounds (8%);

- the surfaces of *non-reconstructed* elemental metal substrates (with or without adsorbates) constitute about 77% of the results; the remainder are *reconstructed*, i.e. have undergone a substantial structural change from the ideal termination of the bulk lattice, involving bond breaking and/or bond making;
- looking at electronic properties, metals again dominate heavily, with 81% of the total, followed by semiconductors (16%), insulators (3%) and semimetals (less than 1%);
- atomic overlayers comprise about 54% of all types of adsorption, as opposed to interstitial (1%) or substitutional (5%) underlayers, molecular overlayers (10%), multilayers (9%) or mixes of these adsorption modes.

There is much room for further study of various important categories of materials: one prominent example is oxides and other compounds (carbides, nitrides, . . .); another is all types of adsorption on oxides and other compounds.

However, recent advances in techniques will ensure further diversification and complexification of solved surface structures. The present maturity of techniques will thus increasingly allow the analysis of structures chosen for their practical interest rather than for their simplicity.

B1.21.4 TWO-DIMENSIONAL ORDERING AND NOMENCLATURE

In diffraction, the degree and kind of structural ordering is an important consideration, since the diffraction reflects those structural properties. As a result, diffraction methods are ideal for characterizing the degree and type of ordering that a surface exhibits. In particular, at surfaces, LEED has always been a favourite tool for ‘fingerprinting’ a particular state of ordering of a surface, enhancing experimental reproducibility. It is therefore useful to first briefly examine the forces that are responsible for the variety of ordering types that occur at surfaces. Then, we can introduce standard notation to succinctly describe specific forms of ordering that occur at surfaces.

B1.21.4.1 TWO-DIMENSIONAL ORDERING

A large number of ordered surface structures can be produced experimentally on single-crystal surfaces, especially with adsorbates [15]. There are also many disordered surfaces. Ordering is driven by the interactions between atoms, ions or molecules in the surface region. These forces can be of various types: covalent, ionic, van der Waals, etc; and there can be a mix of such types of interaction, not only within a given bond, but also from bond to bond in the same surface. A surface could, for instance, consist of a bulk material with one type of internal bonding (say, ionic). It may be covered with an overlayer of molecules with a different type of intramolecular bonding (typically covalent); and the molecules may be held to the substrate by yet another form of bond (e.g., van der Waals).

Strong adsorbate–substrate forces lead to *chemisorption*, in which a chemical bond is formed. By contrast, weak forces result in *physisorption*, as one calls non-chemical ‘physical’ adsorption.

The balance between these different types of bonds has a strong bearing on the resulting ordering or disordering of the surface. For adsorbates, the relative strength of adsorbate–substrate and adsorbate–adsorbate interactions is particularly important. When adsorbate–substrate interactions dominate, well ordered overlayer structures are induced that are arranged in a *superlattice*, i.e. a periodicity which is closely related to that of the substrate lattice: one then speaks of *commensurate* overlayers. This results from the tendency for each adsorbate to seek out the same type of *adsorption site* on the surface, which means that all adsorbates attempt to bond in the same manner to substrate atoms.

An example of commensurate overlayers is provided by atomic sulfur chemisorbed on a Ni (100) surface: all S atoms tend to adsorb in the *fourfold coordinated hollow sites*, i.e. each S atom tries to bond to four Ni atoms. At typical high coverages and moderate temperatures, this results in an ordered array of S atoms on the Ni (100) surface. However, high temperatures will disorder such overlayers; also, this layer may be kinetically disordered during its formation, as a result of gradual addition of sulfur atoms before they manage to order. The same is often true of molecular adsorption. Although intramolecular bonding can be strong enough to keep an adsorbed molecular species intact despite its bonding to the substrate, there is usually only a relatively weak mutual interaction among adsorbed molecular species.

Relatively strong adsorbate–adsorbate interactions have a different effect: the adsorbates attempt to first optimize the bonding between them, before trying to satisfy their bonding to the substrate. This typically results in close-packed overlayers with an internal periodicity that it is not matched, or at least is poorly matched, to the substrate lattice. One thus finds well ordered overlayers whose periodicity is generally not closely related to the substrate lattice: this leads

-10-

to so-called *incommensurate* overlayers. Such behaviour is best exemplified by very cohesive overlayers like graphite sheets or oxide thin films that adopt their own preferred lattice constant regardless of the substrate material on which they are adsorbed.

B1.21.4.2 COVERAGE AND MONOLAYER DEFINITIONS

It is useful to define the terms *coverage* and *monolayer* for adsorbed layers, since different conventions are used in the literature. The surface coverage measures the two-dimensional density of adsorbates. The most common definition of coverage sets it to be equal to one monolayer (1 ML) when each two-dimensional surface unit cell of the unreconstructed substrate is occupied by one adsorbate (the adsorbate may be an atom or a molecule). Thus, an overlayer with a coverage of 1 ML has as many atoms (or molecules) as does the outermost single atomic layer of the substrate.

However, many adsorbates cannot reach a coverage of 1 ML as defined in this way: this occurs most clearly when the adsorbate is too large to fit in one unit cell of the surface. For example, benzene molecules normally lie flat on a metal surface, but the size of the benzene molecule is much larger than typical unit cell areas on many metal surfaces. Thus, such an adsorbate will saturate the surface at a lower coverage than 1 ML; deposition beyond this coverage can only be achieved by starting the growth of a second layer on top of the first layer.

It is thus tempting to define the first saturated layer as being one monolayer, and this often done, causing some confusion. One therefore also often uses terms like *saturated monolayer* to indicate such a single adsorbate layer that has reached its maximal two-dimensional density. Sometimes, however, the word ‘saturated’ is omitted from this definition, resulting in a different notion of monolayer and coverage. One way to reduce possible confusion is to use, for contrast with the saturated monolayer, the term *fractional monolayer* for the term that refers to the substrate unit cell rather than the adsorbate size as the criterion for the monolayer density.

B1.21.4.3 TWO-DIMENSIONAL CRYSTALLOGRAPHIC NOMENCLATURE

(A) MILLER INDICES

Single-crystal surfaces are characterized by a set of Miller indices that indicate the particular crystallographic orientation of the surface plane relative to the bulk lattice [5]. Thus, surfaces are labelled in the same way that atomic planes are labelled in bulk x-ray crystallography. For example, a Ni (111) surface has a surface plane

that is parallel to the (111) crystallographic plane of bulk nickel. Thus, the Ni (111) surface exposes a hexagonally close-packed layer of atoms, given that nickel has a face-centred close-packed (fcc) cubic bulk lattice, see [figure B1.21.1 a](#)). Some authors use the more correct notation $\{111\}$ instead of (111), as is common in bulk crystallography to emphasize that the (111) plane is only one of several symmetrically equivalent plane orientations, like $(11\bar{1})$, $(\bar{1}11)$, etc. The $\{111\}$ notation implicitly includes all such equivalent planes.

-11-

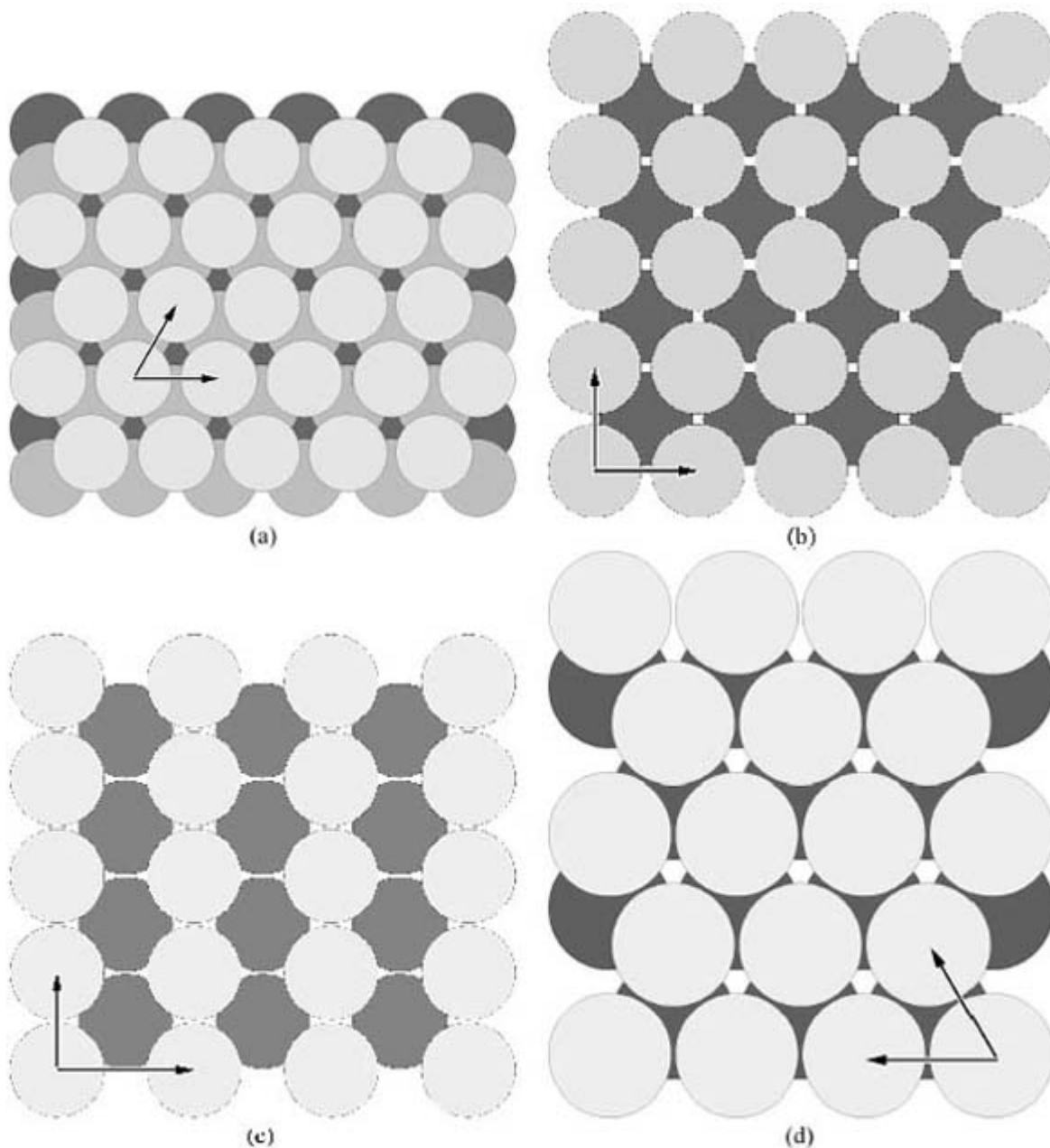


figure continued on next page.

-12-

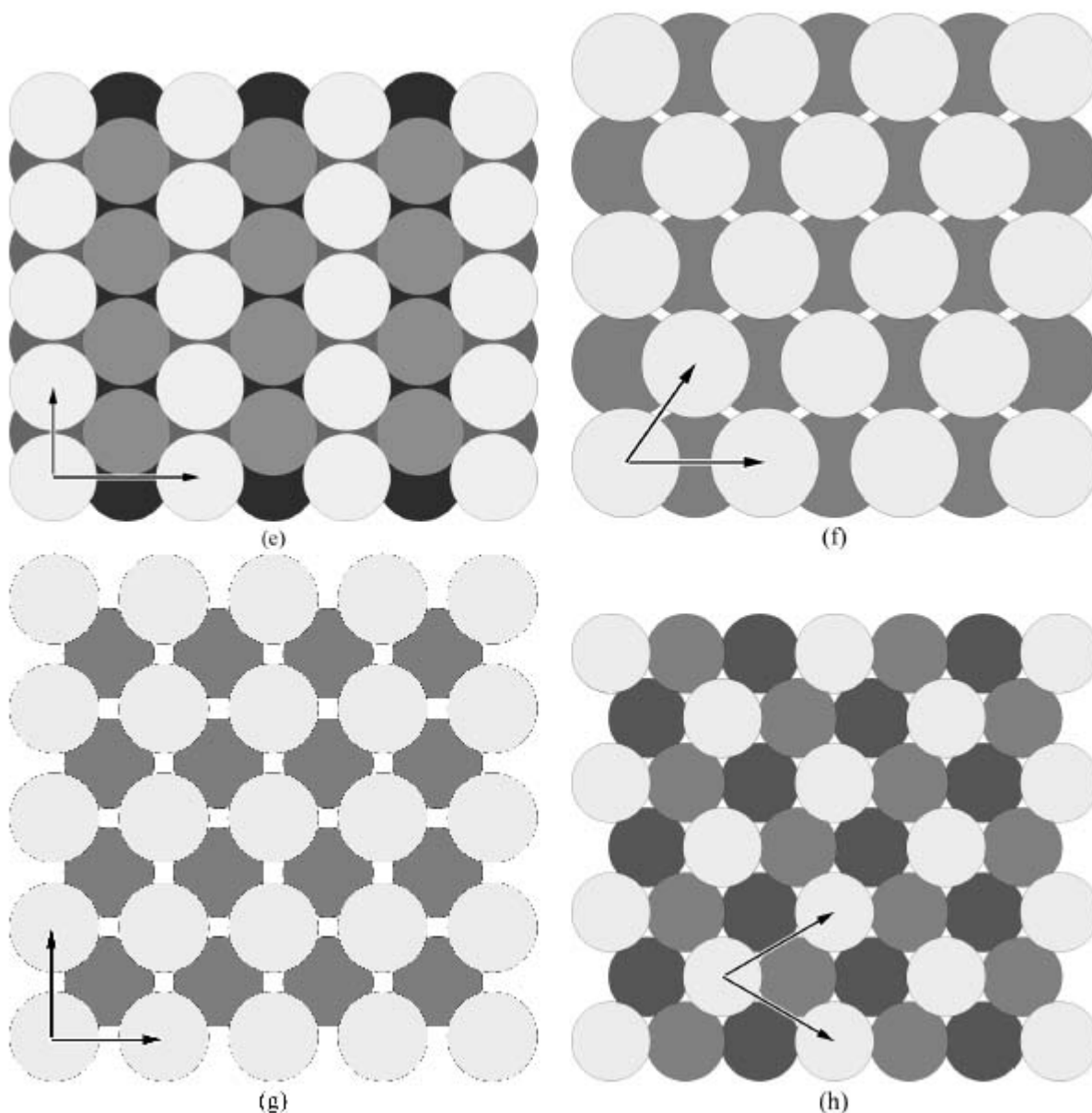


Figure B1.21.1. Atomic hard-ball models of low-Miller-index bulk-terminated surfaces of simple metals with face-centred close-packed (fcc), hexagonal close-packed (hcp) and body-centred cubic (bcc) lattices: (a) fcc (111)–(1 × 1); (b) fcc(100)–(1 × 1); (c) fcc(110)–(1 × 1); (d) hcp(0001)–(1 × 1); (e) hcp(10-10)–(1 × 1), usually written as hcp(10 $\bar{1}$ 0)–(1 × 1); (f) bcc(110)–(1 × 1); (g) bcc(100)–(1 × 1) and (h) bcc(111)–(1 × 1). The atomic spheres are drawn with radii that are smaller than touching-sphere radii, in order to give better depth views. The arrows are unit cell vectors. These figures were produced by the software program BALSAC [35].

Figure B1.21.1 shows a number of other clean unreconstructed low-Miller-index surfaces. Most surfaces studied in surface science have low Miller indices, like (111), (110) and (100). These planes correspond to relatively close-packed surfaces that are atomically rather smooth. With fcc materials, the (111) surface is the densest and smoothest, followed by the (100) surface; the (110) surface is somewhat more ‘open’, in the sense that an additional atom with the same or smaller diameter can bond directly to an atom in the *second* substrate layer. For the hexagonal close-packed (hcp) materials, the (0001) surface is very similar to the fcc (111) surface: the difference only occurs deeper into the surface, namely in the fashion of stacking of the hexagonal close-packed monolayers onto each other (ABABAB... *versus* ABCABC..., in the convenient layer-stacking notation). The hcp (10 $\bar{1}$ 0) surface resembles the fcc (110) surface to some extent, in that it also

presents open troughs between close-packed rows of atoms, exposing atoms in the second layer. With the body-centred cubic (bcc) materials, the (110) surface is the densest and smoothest, followed by the (100) surface; in this case, the (111) surface is rather more open and atomically ‘rough’.

(B) HIGH-MILLER-INDEX OR STEPPED SURFACES

The atomic structures of high-Miller-index surfaces are composed of *terraces*, separated by *steps*, which may have *kinks* in them [5]. Examples are shown in figure B1.21.2. Thus, the (755) surface of an fcc crystal consists of (111) terraces, six atoms deep (from one step to the next), separated by straight steps of (100) orientation and of single-atom height. The fcc (10,8,7) has ‘kinks’ in its step edges, i.e. the steps themselves are not straight. The steps and kinks provide a degree of roughness that can be very important as sites for chemical reactions or for nucleation of crystal growth.

The step notation [5, 16] compacts the terrace/step information into the general form $w(h_t k_t l_t) \times (h_s k_s l_s)$. Here $(h_t k_t l_t)$ and $(h_s k_s l_s)$ are the Miller indices of the terrace plane and the step plane, respectively, while w is the number of atoms that are counted in the width of the terrace, including the step-edge atom and the in-step atom. Thus, the fcc (755) surface can be denoted by $6(111) \times (100)$, since its terraces are six atoms in depth. A kinked surface, like fcc (10,8,7), can also be approximately expressed in this form: the step plane $(h_s k_s l_s)$ is a stepped surface itself, and thus has higher Miller indices than the terrace plane. However, the step notation does not exactly tell us the relative location of adjacent steps, and it is not entirely clear how the terrace width w should be counted. A more complete microfacet notation is available to describe kinked surfaces generally [5].

-14-

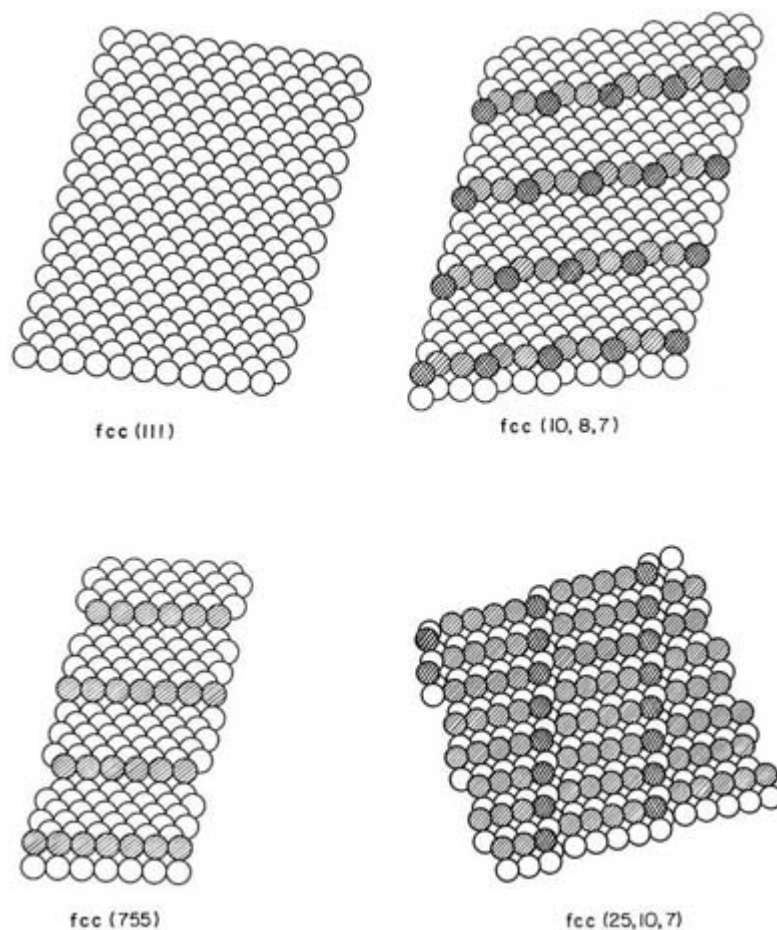


Figure B1.21.2. Atomic hard-ball models of ‘stepped’ and ‘kinked’ high-Miller-index bulk-terminated surfaces of simple metals with fcc lattices, compared with an fcc(111) surface: fcc(755) is stepped, while fcc

(10,8,7) and fcc(25,10,7) are 'kinked'. Step-edge atoms are shown singly hatched, while kink atoms are shown cross-hatched.

-15-

(C) SUPERLATTICES

Many surfaces exhibit a different periodicity than expected from the bulk lattice, as is most readily seen in the diffraction patterns of LEED: often additional diffraction features appear which are indicative of a *superlattice*. This corresponds to the formation of a new two-dimensional lattice on the surface, usually with some simple relationship to the expected 'ideal' lattice [5]. For instance, a layer of adsorbate atoms may occupy only every other equivalent adsorption site on the surface, in both surface dimensions. Such a lattice can be labelled (2×2) : in each surface dimension the repeat distance is doubled relative to the ideal substrate. In this example, the unit cell of the original bulk-like surface is magnified by a factor of two in both directions, so that the new surface unit cell has dimensions (2×2) relative to the original unit cell. For instance, an oxygen overlayer on Pt (111), at a quarter-monolayer coverage, is observed to adopt an ordered (2×2) superlattice: this can be denoted as Pt (111) + (2×2) -O, which provides a compact description of the main crystallographic characteristics of this surface. This particular notation is that of the Surface Structure Database [14]; other equivalent notations are also common in the literature, such as Pt (111)-(2 × 2)-O or Pt (111) 2×2 -O.

This (2×2) notation can be generalized. First, it can take on the form $(m \times n)$, where the numbers m and n are two independent stretch factors for the two unit cell vectors. These numbers are often integers, but need not be. In addition, this new stretched unit cell can be rotated by any angle about the surface normal: this is denoted as $(m \times n) R\alpha^\circ$, where α is the rotation angle in degrees [5, 17, 18 and 19]; the suffix $R\alpha^\circ$ is omitted when $\alpha = 0$, as is the case for Pt (111) + (2×2) -O. This *Wood notation* [5, 19] allows the original unit cell to be stretched and rotated; however, it conserves the angle between the two unit cell vectors in the plane of the surface, therefore not allowing 'sheared' unit cells.

As a particular case, a surface may be given the Wood notation (1×1) , as in Ni (111)-(1 × 1): this notation indicates that the two-dimensional unit cell of the surface has the same size as the two-dimensional unit cell of the bulk (111) layers. Thus, an ideally terminated bulk lattice without overlayers or reconstructions will carry the label (1×1) .

The Wood notation can be generalized somewhat further, by adding either the prefix 'c' for centred, or the prefix 'p' for primitive. For instance, one may have a c (2×2) unit cell or a p (2×2) unit cell, the latter often abbreviated to (2×2) because it is identical to it. In a centred unit cell, the centre of the cell is an exact copy of the corners of the cell; this makes the cell non-primitive, i.e. it is no longer the smallest cell that, when repeated periodically across the surface, generates the entire surface structure. Nonetheless, the centred notation is often used because it can be quite convenient, as the next example will illustrate.

-16-

The c (2×2) unit cell can also be written as $(\sqrt{2} \times \sqrt{2})R45^\circ$. Here, the original unit vectors of the (1×1) structure have both been stretched by factors $\sqrt{2}$ and then rotated by 45° . Thus, sulfur on Ni (100) forms an ordered half-monolayer structure that can be labelled as Ni (100) + c (2×2) -S or, equivalently, Ni (100) + $(\sqrt{2} \times \sqrt{2}) R45^\circ$ -S. The c (2×2) notation is clearly easier to write and also easier to convert into a geometrical model of the structure, and hence is the favoured designation.

A more general notation than Wood's is available for all kinds of unit cells, including those that are sheared, so that the superlattice unit cell can take on any shape, size and orientation. It is the *matrix notation*, defined

as follows [5]. We connect the unit cell vectors \mathbf{a}' and \mathbf{b}' of the superlattice to the unit cell vectors \mathbf{a} and \mathbf{b} of the substrate by the general relations

$$\mathbf{a}' = m_{11}\mathbf{a} + m_{12}\mathbf{b}$$

$$\mathbf{b}' = m_{21}\mathbf{a} + m_{22}\mathbf{b}.$$

The coefficients m_{11} , m_{12} , m_{21} and m_{22} define the matrix $\mathbf{M} = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}$, which serves to denote the superlattice. The (1×1) , (2×2) and $c(2 \times 2)$ lattices are then denoted respectively by the matrices $\mathbf{M} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\mathbf{M} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ and $\mathbf{M} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$. This allows the Ni (100) + $c(2 \times 2)$ -S structure to be also written as Ni(100)+ $\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$ -S. Clearly, this notation is not as intuitive and compact as the $c(2 \times 2)$ Wood notation. However, when the Wood notation is not capable of a clear and compact notation, use of the matrix notation is necessary. Thus, a structure characterized by a matrix like $\mathbf{M} = \begin{pmatrix} 4 & -3 \\ 2 & 5 \end{pmatrix}$ could not be described in the Wood notation.

In LEED experiments, the matrix \mathbf{M} is determined by visual inspection of the diffraction pattern, thereby defining the periodicity of the surface structure: the relationship between surface lattice and diffraction pattern will be described in more detail in the next section.

A superlattice is termed *commensurate* when all matrix elements m_{ij} are integers. If at least one matrix element m_{ij} is an irrational number (not a ratio of integers), then the superlattice is termed *incommensurate*. A superlattice can be incommensurate in one surface dimension, while commensurate in the other surface dimension, or it could be incommensurate in both surface dimensions.

A superlattice can be caused by adsorbates adopting a different periodicity than the substrate surface, or also by a reconstruction of the clean surface. In [figure B1.21.3](#) several superlattices that are commonly detected on low-Miller-index surfaces are shown with their Wood notation.

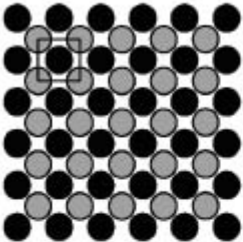
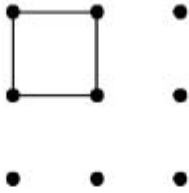
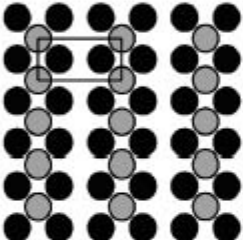
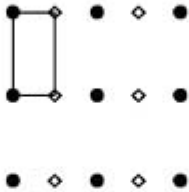
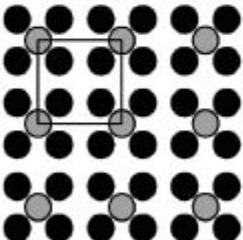
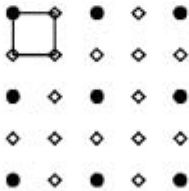
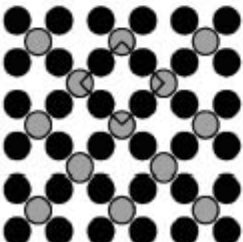
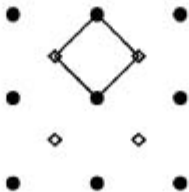
atomic structure	diffraction pattern	nomenclature
		fcc(100)-(1x1)
		fcc(100)-(2x1)
		fcc(100)-(2x2)
		fcc(100)-c(2x2)

figure continued on next page.

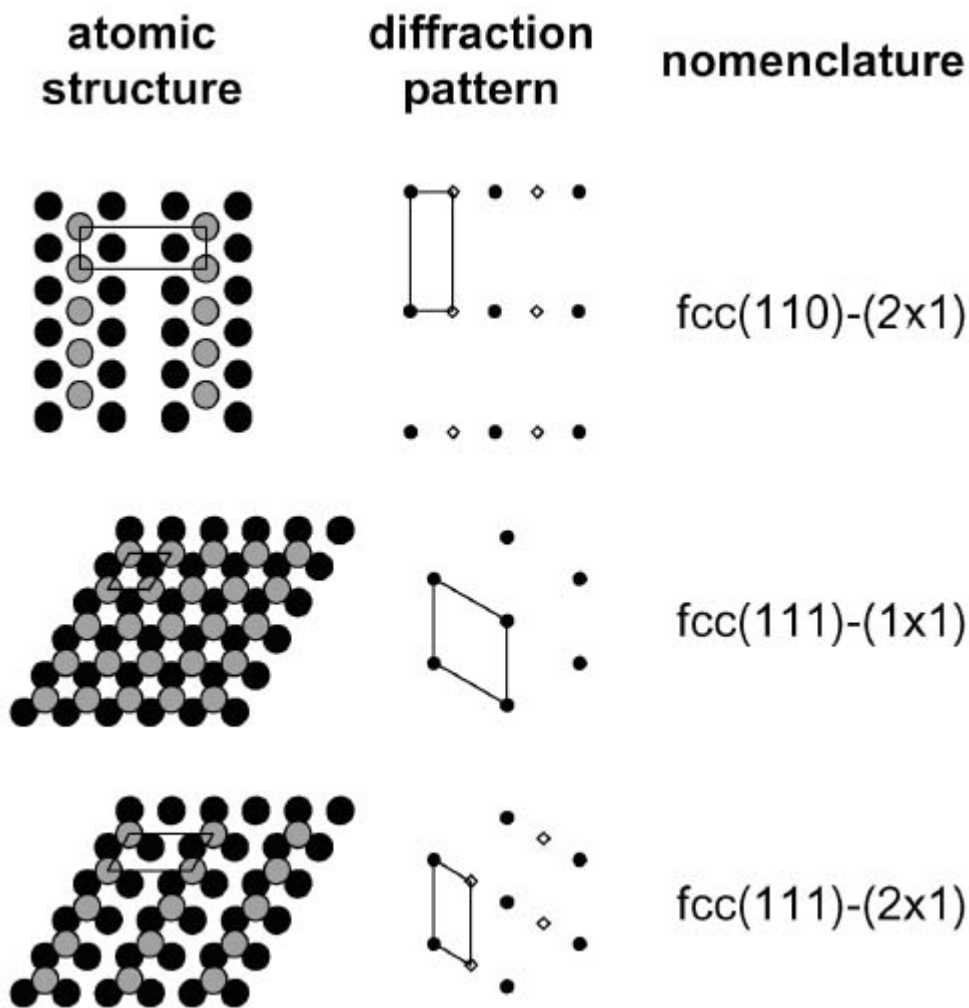


Figure B1.21.3. ‘Direct’ lattices (at left) and corresponding reciprocal lattices (at right) of a series of commonly occurring two-dimensional superlattices. Black circles correspond to the ideal (1 × 1) surface structure, while grey circles represent adatoms in the direct lattice (arbitrarily placed in ‘hollow’ positions) and open diamonds represent fractional-order beams in the reciprocal space. Unit cells in direct space and in reciprocal space are outlined.

B1.21.5 SURFACE DIFFRACTION PATTERN

The diffraction pattern observed in LEED is one of the most commonly used ‘fingerprints’ of a surface structure. With XRD or other non-electron diffraction methods, there is no convenient detector that images in real time the corresponding diffraction pattern. Point-source methods, like PD, do not produce a convenient spot pattern, but a diffuse diffraction pattern that does not simply reflect the long-range ordering.

So it is essential to relate the LEED pattern to the surface structure itself. As mentioned earlier, the diffraction pattern does not indicate relative atomic positions within the structural unit cell, but only the size and shape of that unit cell. However, since experiments are mostly performed on surfaces of materials with a known crystallographic bulk structure, it is often a good starting point to assume an ideally terminated bulk lattice; the actual surface structure will often be related to that ideal structure in a simple manner, e.g. through the creation of a superlattice that is directly related to the bulk lattice.

In this section, we concentrate on the relationship between diffraction pattern and surface lattice [5]. In direct analogy with the three-dimensional bulk case, the *surface lattice* is defined by two vectors \mathbf{a} and \mathbf{b} parallel to the surface (defined already above), subtended by an angle γ ; \mathbf{a} and \mathbf{b} together specify one unit cell, as illustrated in figure B1.21.4. Within that unit cell atoms are arranged according to a *basis*, which is the list of atomic coordinates within that unit cell; we need not know these positions for the purposes of this discussion. Note that this unit cell can be viewed as being infinitely deep in the third dimension (perpendicular to the surface), so as to include all atoms below the surface to arbitrary depth.

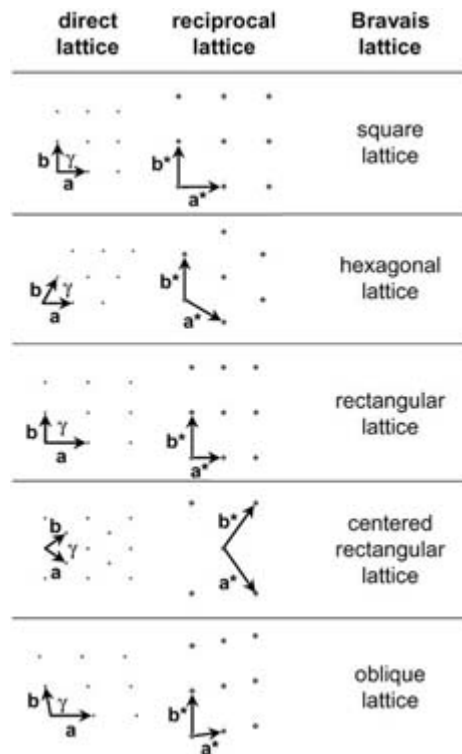


Figure B1.21.4. ‘Direct’ lattices (at left) and reciprocal lattices (middle) for the five two-dimensional Bravais lattices. The reciprocal lattice corresponds directly to the diffraction pattern observed on a standard LEED display. Note that other choices of unit cells are possible: e.g., for hexagonal lattices, one often chooses vectors \mathbf{a} and \mathbf{b} that are subtended by an angle γ of 120° rather than 60° . Then the reciprocal unit cell vectors also change: in the hexagonal case, the angle between \mathbf{a}^* and \mathbf{b}^* becomes 60° rather than 120° .

There are several special shapes of the surface lattice, forming the five two-dimensional Bravais lattices shown in [figure B1.21.4](#). The Bravais lattices form the complete list of possible lattices. They are characterized by unit cell vectors of equal length (in the case of the square and hexagonal lattices) and/or a subtended angle of 90° or 60° (for the square, rectangular and hexagonal lattices) or by completely general values (for the oblique lattice). The rectangular lattice comes in two varieties: primitive and centred. The centred lattice has the particularity that its atomic basis is duplicated: each atom is reproduced by displacement through the vector $1/2 (\mathbf{a} + \mathbf{b})$. The main value of the centred rectangular lattice is its convenience: it is easier to think in terms of the rectangle (with duplicated basis) than to think of the rhombus with arbitrary angle γ . One could also centre any of the other lattices, but one would only produce another instance of a square, rectangular or oblique lattice, i.e. nothing more convenient.

The diffraction of low-energy electrons (and any other particles, like x-rays and neutrons) is governed by the translational symmetry of the surface, i.e. the surface lattice. In particular, the directions of emergence of the diffracted beams are determined by conservation of the linear momentum parallel to the surface, $\hbar \mathbf{k}_\parallel$. Here \mathbf{k}

denotes the wavevector of the incident plane electron wave that represents the incoming electron beam. This conservation can occur in two ways. After the diffractive scattering, the parallel component of the momentum $\hbar\mathbf{k}'_{\parallel}$ can be equal to that of the incident electron beam, i.e. $\hbar\mathbf{k}'_{\parallel} = \hbar\mathbf{k}_{\parallel}$; this corresponds to *specular* (mirror-like) reflection, with equal polar angles of incidence and emergence with respect to the surface normal, and with a simple reversal of the perpendicular momentum $\hbar\mathbf{k}'_{\perp} = -\hbar\mathbf{k}_{\perp}$.

Alternatively, the electron can exchange parallel momentum with the lattice, but only in well defined amounts given by vectors $\hbar\mathbf{g}$ that belong to the *reciprocal lattice* of the surface. That is, the vector \mathbf{g} is a linear combination of two reciprocal lattice vectors \mathbf{a}^* and \mathbf{b}^* , with integer coefficients. Thus, $\mathbf{g} = h\mathbf{a}^* + k\mathbf{b}^*$, with arbitrary integers h and k (note that all the vectors \mathbf{a} , \mathbf{b} , \mathbf{a}^* , \mathbf{b}^* and \mathbf{g} are parallel to the surface). The reciprocal lattice vectors \mathbf{a}^* and \mathbf{b}^* are related to the ‘direct-space’ lattice vectors \mathbf{a} and \mathbf{b} through the following non-transparent definitions, which also use a vector \mathbf{n} that is perpendicular to the surface plane, as well as vectorial dot and cross products:

$$\mathbf{a}^* = 2\pi \left(\frac{\mathbf{b} \times \mathbf{n}}{\mathbf{a} \cdot (\mathbf{b} \times \mathbf{n})} \right) \quad \text{and} \quad \mathbf{b}^* = 2\pi \left(\frac{\mathbf{n} \times \mathbf{a}}{\mathbf{b} \cdot (\mathbf{n} \times \mathbf{a})} \right).$$

These two equations are a special case of the corresponding three-dimensional definition, common in XRD, with the surface normal \mathbf{n} replacing the third lattice vector \mathbf{c} .

[Figure B1.21.4](#) illustrates the ‘direct-space’ and reciprocal-space lattices for the five two-dimensional Bravais lattices allowed at surfaces. It is useful to realize that the vector \mathbf{a}^* is always perpendicular to the vector \mathbf{b} and that \mathbf{b}^* is always perpendicular to \mathbf{a} . It is also useful to notice that the length of \mathbf{a}^* is inversely proportional to the length of \mathbf{a} , and likewise for \mathbf{b}^* and \mathbf{b} . Thus, a large unit cell in direct space gives a small unit cell in reciprocal space, and a wide rectangular unit cell in direct space produces a tall rectangular unit cell in reciprocal space. Also, the hexagonal direct-space lattice gives rise to another hexagonal lattice in reciprocal space, but rotated by 90° with respect to the direct-space lattice.

-21-

The reciprocal lattices shown in [figure B1.21.3](#) and [figure B1.21.4](#) correspond directly to the diffraction patterns observed in LEED experiments: each reciprocal-lattice vector produces one and only one diffraction spot on the LEED display. It is very convenient that the hemispherical geometry of the typical LEED screen images the reciprocal lattice without distortion; for instance, for the square lattice one observes a simple square array of spots on the LEED display.

One of the spots in such a diffraction pattern represents the specularly reflected beam, usually labelled (00). Each other spot corresponds to another reciprocal-lattice vector $\mathbf{g} = h\mathbf{a}^* + k\mathbf{b}^*$ and is thus labelled (hk) , with integer h and k .

When a superlattice is present, additional spots arise in the diffraction pattern, as shown in [figure B1.21.3](#) in terms of the reciprocal lattice: again, each reciprocal lattice point corresponds to a spot in a diffraction pattern. This can be easily understood from the fact that a larger unit cell in direct space imposes a smaller unit cell in reciprocal space. For instance, a (2×1) superlattice has a unit cell doubled in length in one surface direction relative to the (1×1) lattice, i.e. \mathbf{a} is replaced by $2\mathbf{a}$. According to the above equations, this has no effect on \mathbf{b}^* , but halves \mathbf{a}^* . This is equivalent to allowing h to be a half-integer in $\mathbf{g} = h\mathbf{a}^* + k\mathbf{b}^*$, thus doubling the number of spots in the diffraction pattern. These additional spots are therefore often called half-order spots in the (2×1) case, or fractional-order spots in the general case.

With some practice, one can easily recognize specific superlattices from their LEED pattern. Otherwise, one can work through the above equations to connect particular superlattices to a given LEED pattern. A number

of examples are given and discussed in some detail in [5]. A discussion can also be found there of the special case of stepped and kinked surfaces.

B1.21.6 DIFFRACTION PATTERN OF DISORDERED SURFACES

Many forms of disorder in a surface structure can be recognized in the LEED pattern. The main manifestations of disorder are *broadening* and *streaking* of diffraction spots and *diffuse* intensity between spots [1].

Broadening of spots can result from thermal diffuse scattering and island formation, among other causes. The thermal effects arise from the disorder in atomic positions as they vibrate around their equilibrium sites; the sites themselves may be perfectly crystalline.

Islands occur particularly with adsorbates that aggregate into two-dimensional assemblies on a substrate, leaving bare substrate patches exposed between these islands. Diffraction spots, especially fractional-order spots if the adsorbate forms a superlattice within these islands, acquire a width that depends inversely on the average island diameter. If the islands are systematically anisotropic in size, with a long dimension primarily in one surface direction, the diffraction spots are also anisotropic, with a small width in that direction. Knowing the island size and shape gives valuable information regarding the mechanisms of phase transitions, which in turn permit one to learn about the adsorbate–adsorbate interactions.

-22-

Lattice-gas disorder, in which adatoms occupy a periodic lattice of equivalent sites with a random occupation probability, produces diffuse intensity distributions between diffraction spots. For complete disorder, one observes such diffuse intensity throughout the diffraction pattern. If there is order in one surface direction, but disorder in the other, one observes streaking in the diffraction pattern: the direction of the streaks corresponds to the direction in which disorder occurs. In principle, the diffuse intensity distribution can be converted into a direct-space distribution, including a pair-correlation function between occupied sites, e.g. by Fourier transformation. However, the diffuse intensity is too much affected by other diffraction effects (like multiple scattering) to be very useful in this manner. It nonetheless can be interpreted in terms of local structure, i.e. bond lengths and angles, by a procedure that is very similar to the multiple-scattering modelling for solving structures in full detail [20].

LEED has found a strong competitor for studying surface disorder: scanning tunnelling microscopy, STM (see [chapter B1.20](#)). Indeed, STM is the ideal tool for investigating irregularities in periodic surface structures. LEED (as any other diffraction method) averages its information content over macroscopic parts of the surface, giving only statistical information about disorder. By contrast, STM can provide a direct image of individual atoms or defects, enabling the observation of individual atomic behaviour. By observing a sufficiently large area, STM can also provide statistical information, if desired.

B1.21.7 FULL STRUCTURAL DETERMINATION

In the previous sections we have emphasized the two-dimensional information available through the diffraction pattern observed in LEED. But, as mentioned before, one can extract the detailed atomic positions as well, including interlayer spacings, bond lengths, bond angles, etc. Here we sketch how this more complete structural determination is accomplished. We focus on the case of LEED, since this method has produced by far the most structural determinations [5, 17, 18, 21]. The procedures employed to analyse PD data are in fact very similar to those for LEED, in many details. With XRD, the kinematic (single-scattering) nature of the

problem makes the analysis simpler, but still considerable for complex structures: there also, a trial-and-error search for the solution is common.

To obtain spacings between atomic layers and bond lengths or angles between atoms, it is necessary to measure and analyse the *intensity* of diffraction spots. This is analogous to measuring the intensity of XRD reflections.

The measurement of LEED spot intensities is nowadays mostly accomplished by digitizing the image recorded by a video camera that observes the diffraction pattern, which is visibly displayed on a fluorescent screen within an ultra-high vacuum system [22]. The digitized image is then processed by computer to give the integrated spot intensity, after removal of the background. This is repeated for different incident electron energies. Thereby, the intensity of each spot is obtained as a function of the incident electron energy, resulting in an *IV curve* (intensity–voltage curve) for each spot. Computer codes for this purpose are available, and are normally packaged together with the required hardware [23]. The resulting IV curves form the experimental database to which theory can fit the atomic structure. It typically takes between minutes and an hour to accumulate such a database, once the sample has been prepared.

Since ED by a surface is a complicated process, there is no routine method available to *directly* and accurately extract atomic positions from the experimental data. Direct holographic methods have been proposed [24], but have not yet

-23-

become routine methods, and in any case they yield only approximate atomic positions (with uncertainties on the scale of 0.2–0.5 Å) and work only for relatively simple structures; when they do work they have to be followed up by refinement using the same trial-and-error approach that we discuss next.

A detailed structural determination proceeds by modelling the full multiple scattering of the electrons that are diffracted through the surface structure. The multiple scattering means that an electron can bounce off a succession of atoms in an erratic path before emerging from the surface. Various theoretical and computational methods are available to treat this problem to any degree of precision: a compromise between precision and computing expense must be struck, with progress moving toward higher precision, even for more complex structures.

The modelling of the multiple scattering requires input of all atomic positions, so that the *trial-and-error approach* must be followed: one guesses reasonable models for the surface structure, and tests them one by one until satisfactory agreement with experiment is obtained. For simple structures and in cases where structural information is already known from other sources, this process is usually quite quick: only a few basic models may have to be checked, e.g. adsorption of an atomic layer in hollow, bridge or top sites at positions consistent with reasonable bond lengths. It is then relatively easy to refine the atomic positions within the best-fit model, resulting in a complete structural determination. The refinement is normally performed by some form of *automated steepest-descent optimization*, which allows many atomic positions to be adjusted simultaneously [21]. Computer codes are also available to accomplish this part of the analysis [25]. The trial-and-error search with refinement may take minutes to hours on current workstations or personal computers.

In more complex cases, and when little additional information is available, one must test a larger number of possible structural models. The computational time grows rapidly with complexity, so that it may take hours to check a single model. More time-consuming, however, is often the human factor in guessing what are reasonable models to test. This is a much more difficult problem, which is the issue of finding the ‘global optimum’, not just a ‘local optimum’. At present, several approaches to *global optimization* are being examined, such as *simulated annealing* [26] and *genetic algorithms* [27]. In any event, these will require larger amounts of computer time, since a wide variety of surface models must be tested in such a global

search.

B1.21.8 PRESENT CAPABILITIES AND OUTLOOK

Surface crystallography started in the late 1960s, with the simplest possible structures being solved by LEED [14]. Such structures were the clean Ni (111), Cu(111) and Al(111) surfaces, which are unreconstructed and essentially unrelaxed, i.e. very close to the ideal termination of the bulk shown in [figure B1.21.1 a](#)): typically, only one unknown structural parameter was fitted to experiment, namely the spacing between the two outermost atomic layers.

Progress in experiment, theory, computational methods and computer power has contributed to the capability to solve increasingly complex structures [28, 29]. [Figure B1.21.5](#) quantifies this progress with three measures of complexity, plotted logarithmically: the achievable two-dimensional unit cell size, the achievable number of fit parameters and the achievable number of atoms per unit cell per layer: all of these measures have grown from 1 for simple clean metal

-24-

surfaces, like Ni (111) (see [Figure B1.21.1 \(a\)](#)), to about 50–100 in the case of the reconstructed Si(111)–(7 × 7) surface, the most complicated structure examined to date [30] (note that the basic model which solved the Si(111)–(7 × 7) surface was mainly derived from another diffraction study, using TED [9]). All these measures thus exhibit a progression by about two orders of magnitude over less than 25 years.

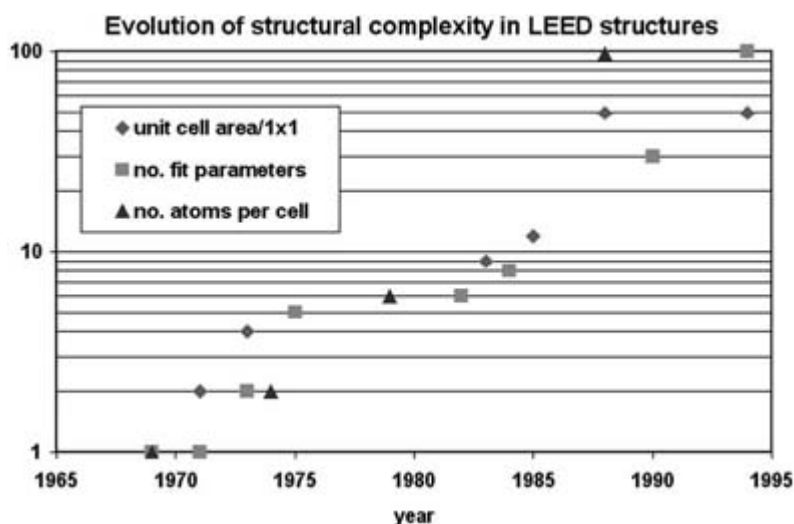


Figure B1.21.5. Evolution with time of the complexity of structural determination achievable with LEED. The unit cell area is measured relative to the unit cell area of the simple (1 × 1) structures studied in the early days: thus a ($n \times n$) superstructure (due to reconstruction and/or adsorption) has a unit cell size of n^2 . A (7 × 7) structure gives a complexity of 49 on this scale. The number of fit parameters measures the number of coordinates fitted to experiment in any given structure: a value of 1 was typical of many early determinations, when only one interlayer spacing was fitted to experiment. The Si(111)–(7 × 7) structure has over 100 fit parameters, if one allows only those structural changes in the top two double layers and the adatom layer that maintain the $p3m1$ symmetry of the substrate. The number of atoms per unit cell refers to so-called composite layers, which are groups of closely spaced layers: this number dramatically affects computation time in multiple-scattering methods. It has grown from 1 in the simplest structures to about 100 in the Si(111)×(7 × 7) structure.

Figure B1.21.6, figure B1.21.7, figure B1.21.8 and figure B1.21.9 show several of the more complex structures solved by LEED in recent years. They exhibit various effects observed at surfaces:

- clustering of adatoms in Re(0001) – $(2\sqrt{3} \times 2\sqrt{3})R30^\circ-6S$ [31], see figure B1.21.6
- hollow-site adsorption and adsorbate-induced relaxations of substrate atoms both in Re(0001)– $(2\sqrt{3} \times 2\sqrt{3})R30^\circ-6S$ [31] and in Mo (100) – $c(4 \times 2)-3S$ [32], see figure B1.21.7
- adsorbate-induced reconstruction as well as substitutional adsorption in Cu(100)– $(4 \times 4)-10Li$ [33], see figure B1.21.8 note that this is the most complex surface structural determination by LEED to date, involving far more adjustable structural parameters than were fitted in the Si(111) – (7×7) structure [30];
- compound ionic surface with a large bulk unit cell and very large surface relaxations in $Fe_3O_4(111)$ [34], see figure B1.21.9

-25-

Further progress towards solving more complex surface structures is possible. The biggest challenge on the computational and theoretical side is the identification of the globally optimum structure. Holographic and other methods have not yet provided a convenient way to accomplish this, and would actually fail with structures that have the complexity of Cu(100) – $(4 \times 4)-10Li$ and Si(111) – (7×7) . Global-search algorithms, like simulated annealing and genetic algorithms, may provide workable, if perhaps not cheap, solutions.

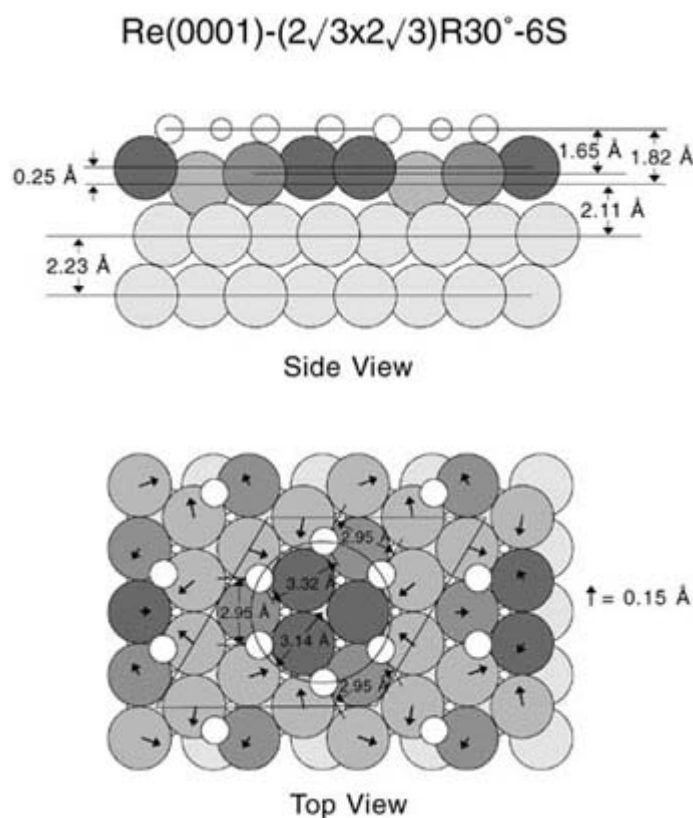


Figure B1.21.6. Side and top views of the best-fit structure of the Re(0001)– $(2\sqrt{3} \times 2\sqrt{3})R30^\circ-6S$ surface structure (with a half-monolayer coverage of sulfur), as determined by LEED [31]. A $(2\sqrt{3} \times 2\sqrt{3})R30^\circ$ unit cell is outlined in the top view. Sulfur atoms are drawn as small open circles, Re atoms as large grey circles. Sulfur–sulfur distances in a ring of six alternate between 2.95 and 3.32 Å, expanded from the unrelaxed distance between hollow sites of 2.75 Å. Arrows represent lateral relaxations in the topmost metal layer, with the scale of displacements indicated by the lone arrow on the right. The bulk interlayer spacing in Re(0001) is 2.23 Å. Shades of grey identify atoms that are equivalent by symmetry in the sulfur and outermost rhenium layers. The darkest-grey rhenium atoms forming a triangle within a sulfur ring are pulled out of the surface by the adsorbed sulfur, relative to the lighter-grey rhenium atoms in the same layer.

Mo(100)-c(4x2)-3S

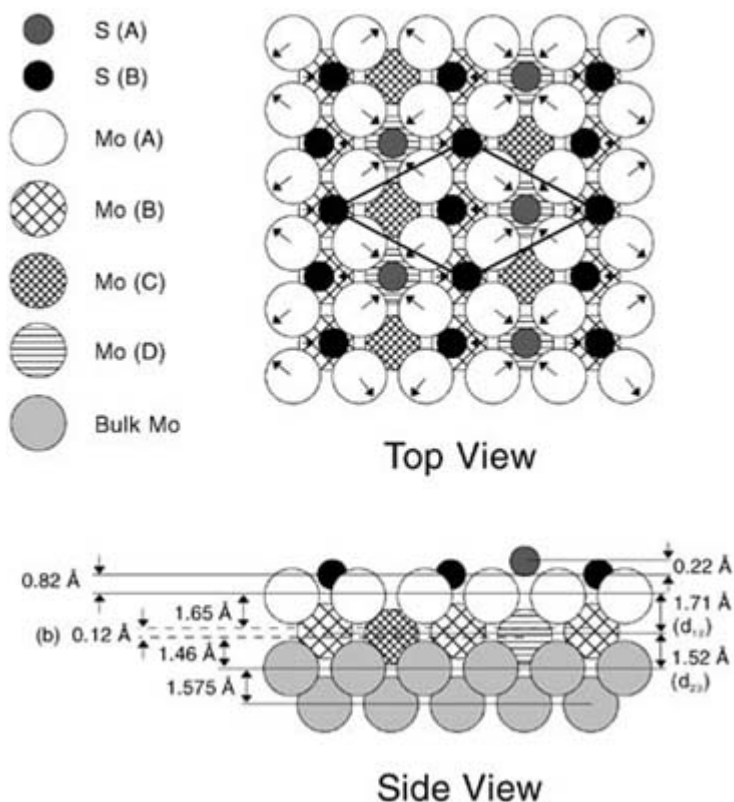


Figure B1.21.7. Top and side views of the best-fit structure of the Mo(100)-c(4 × 2)-3S surface structure (with a 3/4-monolayer coverage of sulfur), as determined by LEED [32]. A c(4 × 2) unit cell is outlined in the top view. The sulfur sizes (small black and dark grey circles) have been reduced from covalent for clarity, while the molybdenum atoms (large circles) are drawn with touching radii. The same cross-hatching has been assigned to molybdenum atoms that are equivalent by symmetry in the topmost two metal layers. Two-thirds of the sulfur atoms are displaced away from the centre of the hollow sites in which they are bonded: these displacements by 0.13 Å are drawn exaggerated. Arrows in the top view also indicate the directions and relative magnitudes of molybdenum atom displacements (these substrate atoms are drawn in their undisplaced positions, except for the buckling seen in the second molybdenum layer in the side view). The bulk interlayer spacing in Mo(100) is 1.575 Å.

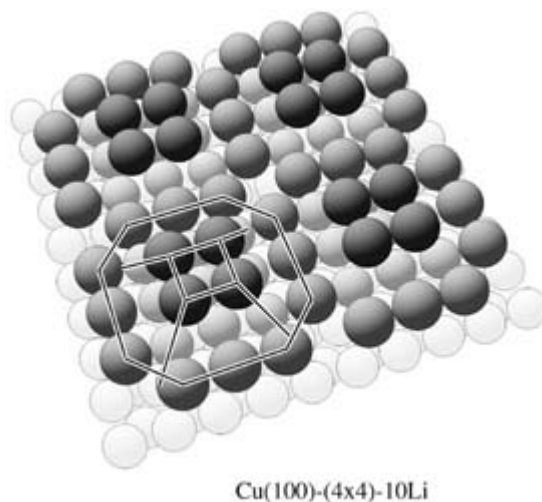
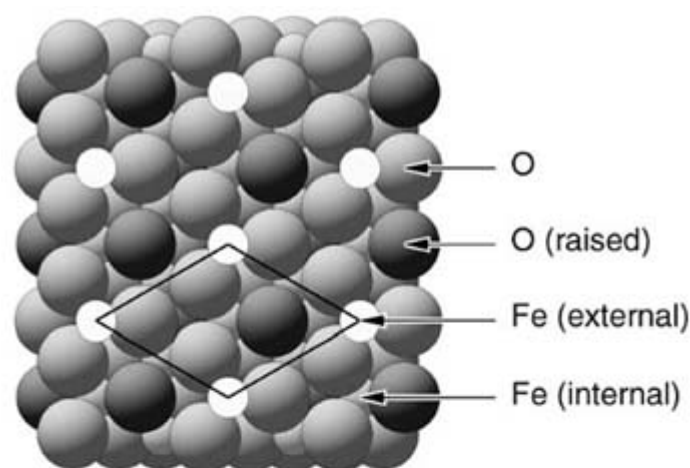
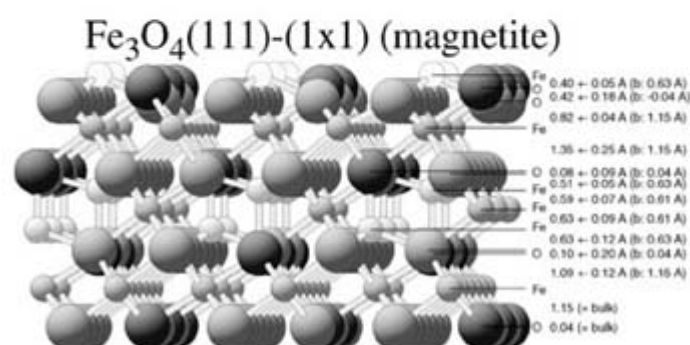


Figure B1.21.8. Perspective view of the structure of the Cu(100)–(4 × 4)–10Li surface structure (with a 10/16-monolayer coverage of lithium), as determined by LEED [33]. The atoms are drawn with radii that are reduced by about 15% from covalent radii. The surface fragment shown includes four (4 × 4) unit cells. Lithium atoms are shown as larger spheres. In each unit cell, four lithium atoms (dark grey) form a flat-topped pyramid (as outlined): the lithium atoms rest in hollow sites on a 3 × 3 base of nine Cu atoms (lighter grey). Around each pyramid 12 lithium atoms occupy substitutional sites, i.e. have taken the place of Cu atoms: these lithium atoms are shown linked by an octagon. Since the lithium atoms are about 15% larger than the copper atoms that they replace, fewer lithium atoms can fit in the troughs evacuated by the copper atoms; thus, they do not fill the troughs completely, and leave a hole at each intersection between troughs (e.g. at the exact centre of the fragment). The lightest-grey atoms underneath are the bulk Cu(100) termination: some small local distortions in the atomic positions are also detected by LEED there. This and the following figure were produced with the SARCH/LATUSE/PLOT3D/BALSAC software, available from the author.



Top view



Side view

$\text{Fe}_3\text{O}_4(111)-(1 \times 1)$ (magnetite)

Figure B1.21.9. Perspective side and top view of the best-fit structure of $\text{Fe}_3\text{O}_4(111)$, as determined by LEED [34]. A unit cell is outlined in the top view, in which all atoms are drawn with nearly touching radii, while smaller radii are used in the side view. This iron oxide was grown as an ultrathin film on a Pt(111) substrate, in order to prevent electrical charging of the surface. The free surface is at the top end of the side view, exposing 1/4 monolayer of ‘external’ iron ions (shown as small light-grey circles in both views). Large circles represent oxygen ions, forming hexagonally close-packed layers. In each such layer, one-fourth of the oxygen ions (drawn in darkest grey) is not coplanar with the others: in particular, in the outermost oxygen layer, these ions are raised outward by a large amount (0.42 Å, compared to 0.04 Å in the opposite direction in the bulk). Small circles below the surface represent iron ions in tetrahedral or octahedral interstitial positions between the O layers: the lightest-grey of these are in tetrahedral positions. Interlayer spacings as determined by LEED are given at the right with error bars and with corresponding bulk values in parentheses.

On the experimental side, a larger measured database is required than is commonly available to determine the large number of structural parameters to be fitted. For instance, LEED calculations for the $\text{Si}(111) - (7 \times 7)$ surface have been attempted to fit the many tens of unknown structural parameters; however, the amount of experimental data was insufficient for the task, resulting in a multitude of locally-optimum structures, without the ability to discriminate between them. Increasing the database size can be achieved by extending the energy range to higher energies, or by acquiring data at a number of different incidence directions: either way, the calculations become disproportionately more time-consuming, because the computing effort rises quickly

with energy and non-symmetrical off-normal incidence directions.

ACKNOWLEDGMENTS

Professor K Hermann is gratefully thanked for the images in [figure B1.21.1](#) which were produced with the program BALSAC [35].

REFERENCES

- [1] Henzler M 1997 Capabilities of LEED for defect analysis *Surf. Rev. Lett* **4** 489–500
 - [2] Rehr J J 1995 Multiple-scattering approach to surface EXAFS—theory versus experiment *Surf. Rev. Lett* **2** 63–9
 - [3] Stöhr J 1992 *NEXAFS Spectroscopy* (Heidelberg: Springer)
 - [4] Feidenhans'l R 1989 Surface structure determination by x-ray diffraction *Surf. Sci. Rep* **10** 105–88
 - [5] Van Hove M A, Weinberg W H and Chan C-M 1986 *LEED Experiment, Theory and Structural Determination* (Heidelberg: Springer)
 - [6] Ichimiya A, Ohno Y and Horio Y 1997 Structural analysis of crystal surfaces by reflection high energy electron diffraction *Surf. Rev. Lett* **4** 501–11
 - [7] Kjems J K, Passell L, Taub H, Dash J G and Novaco A D 1976 Neutron scattering study of nitrogen adsorbed on basal plane-oriented graphite *Phys. Rev. B* **13** 1446–62
 - [8] McTague J P, Nielsen M and Passell L 1979 Neutron scattering by adsorbed monolayers *Crit. Rev. Solid State Sci* **8** 125–56
 - [9] Takayanagi K 1990 Surface structure analysis by transmission electron diffraction—effects of the phases of structure factors *Acta. Crystallogr A* **46** 83–6
 - [10] Fadley C S et al 1997 Photoelectron diffraction: space, time and spin dependence of surface structures *Surf. Rev. Lett* **4** 421–40
-

- [11] Lee P A, Citrin P H, Eisenberger P and Kincaid B M 1981 Extended x-ray absorption fine structure—its strengths and limitations as a structural tool *Rev. Mod. Phys.* **53** 769–806
- [12] Cowan P L, Golovchenko J L and Robbins M F 1980 X-ray standing waves at crystal surfaces *Phys. Rev.L* **44** 1680–3
- [13] Woodruff D P, Cowie B C C and Ettem a A R H F 1994 Surface structure determination using x-ray standing waves: a simple view *J. Phys.. Condens.* **6** 10 633–45
- [14] Watson P R, Van Hove M A and Hermann K 1999 *NIST Surface Structure Database Ver. 3.0* (Gaithersburg, MD: NIST Standard Reference Data Program)
- [15] Ohtani H, Kao C-T, Van Hove M A and Somorjai G A 1986 A tabulation and classification of clean

solid surfaces and of adsorbed atomic and molecular monolayers as determined from low-energy electron diffraction *Prog. Surf. Sci* **23** 155–316

- [16] Lang B, Joyner R W and Somorjai G A 1972 LEED studies of high index crystal surfaces of platinum *Surf. Sci* **30** 440–53
- [17] Pendry J B 1974 *Low-Energy Electron Diffraction* (London: Academic)
- [18] Clarke L J 1985 *Surface Crystallography, An Introduction to LEED* (Chichester: Wiley)
- [19] Wood E A 1964 Vocabulary of surface crystallography *J. Appl. Phys.* **35** 1306–12
- [20] Heinz K 1994 Diffuse LEED and local surface structure *Phys. Status. Solidi A* **146** 195–204
- [21] Van Hove M A, Moritz W, Over H, Rous P J, Wander A, Barbieri A, Materer N, Starke U and Somorjai G A 1993 Automated determination of complex surface structures by LEED *Surf. Sci. Rep* **19** 191–229
- [22] Heinz K and Müller K 1982 LEED intensities—experimental progress and new possibilities of surface structure determination *Springer Tracts in Modern Physics* **91** 1–54 (Heidelberg: Springer)
- [23] e.g., SPECS GmbH, Voltastrasse 5, D-13355 Berlin, Tel. + 49 -30 -46 78 24 - 0, Fax. + 49 - 30 - 46 42 083, email: support@specs.de, <http://www.specs.de>
- [24] Saldin D K, Chen X, Vamvakas J A, Ott M, Wedler H, Reuter K, Heinz K and De Andres P L 1997 Holographic LEED: a review of recent progress *Surf. Rev. Lett* **4** 991–1001
- [25] Van Hove M A <http://electron.lbl.gov/software/software.html> Heinz K TLEED@fkp.physik.uni-erlangen.de
- [26] Rous P J 1993 A global approach to the search problem in surface crystallography by low-energy electron diffraction *Surf. Sci* **296** 358–73
- [27] Döll R and Van Hove M A 1996 Global optimization in LEED structure determination using genetic algorithms *Surf. Sci* **355** L393–8
- [28] Van Hove M A 1996 Complex surface structures from LEED *Surf. Rev. Lett* **3** 1271–84

- [29] Van Hove M A 1997 Determination of complex surface structures with LEED *Surf. Rev. Lett* **4** 479–87
- [30] Tong S Y, Huang H, Wei C M, Packard W E, Men F K, Glander G and Webb M B 1988 Low-energy electron diffraction analysis of the Si(111) - (7 × 7) structure *J. Vac. Sci. Technol A* **6** 615–24
- [31] Barbieri A, Jentz D, Materer N, Held G, Dunphy J, Ogletree D F, Sautet P, Salmeron M, Van Hove M A and Somorjai G A 1994 Surface crystallography of Re (0001) – (2 × 2) – S and Re (0001) – (2√3 × 2√3) r30° – 6S: a combined LEED and STM study *Surf. Sci* **312** 10–20
- [32] Jentz D, Rizzi S, Barbieri A, Kelly D, Van Hove M A and Somorjai G A 1995 Surface structures of sulfur and carbon overlayers on Mo(100): a detailed analysis by automated tensor LEED *Surf. Sci* **329** 14–31

- [33] Mizuno S, Tochiyama H, Barbieri A and Van Hove M A 1995 Completion of the structural determination of and rationalizing of the surface-structure sequence $(2 \times 1) \rightarrow (3 \times 3) \rightarrow (4 \times 4)$ formed on Cu(001) with increasing Li coverage *Phys. Rev. B* **52** 11 658–61
- [34] Barbieri A, Weiss W, Van Hove M A and Somorjai G A 1994 Magnetite Fe_2O_4 (111): surface structure by LEED crystallography and energetics *Surf. Sci* **302** 259–79
- [35] Hermann K <http://www.fhi-berlin.mpg.de/th/personal/hermann/bal pam.html>
-

FURTHER READING

Pendry J B 1974 *Low-energy Electron Diffraction* (London: Academic)

A full description of the principles of low-energy electron diffraction (LEED) as of 1974, containing all the basic physics still in use today.

Van Hove M A, Weinberg W H and Chan C-M 1986 *LEED Experiment, Theory and Structural Determination* (Heidelberg: Springer)

Covers in great detail the practical application of low-energy electron diffraction (LEED) for structural studies, excepting more recent techniques like tensor LEED and holography.

Van Hove M A 1997 Determination of complex surface structures with LEED *Surf. Rev. Lett* **4** 479–88

Describes the state of the art in surface structural determination by low-energy electron diffraction (LEED), focusing on complex structures.

Fadley C S *et al* 1997 Photoelectron diffraction: space, time and spin dependence of surface structures *Surf. Rev. Lett* **4** 421–40

Summarizes the state of the art of photoelectron diffraction as a structural tool.

Rehr J J 1995 Multiple-scattering approach to surface EXAFS—theory versus experiment *Surf. Rev. Lett* **2** 63–9

Addresses the need for advanced methods in surface extended x-ray absorption fine structure (SEXAFS) for accurate structural determination.

-32-

-1-

B1.22 Surface characterization and structural determination: optical methods

Francisco Zaera

B1.22.1 INTRODUCTION

As discussed in more detail elsewhere in this encyclopaedia, many optical spectroscopic methods have been developed over the last century for the characterization of bulk materials. In general, optical spectroscopies make use of the interaction of electromagnetic radiation with matter to extract molecular parameters from the substances being studied. The methods employed usually rely on the examination of the radiation absorbed,

emitted or scattered by a system, and may be based on simple linear optical processes, resonance transitions and/or nonlinear processes. Molecular spectroscopy probes energy transitions at all levels, from the excitation of spins in the radiofrequency range (NMR), to rotational (microwave), vibrational (infrared) and electronic valence (visible–UV) and core (x-rays) excitations. Additional diffraction- and polarization-based techniques provide structural information and laser-based pump–probe methods allow for the study of molecular dynamics down to the femtosecond time scale.

In spite of the wide range of applications of optical techniques for the study of bulk samples, however, they have so far found only limited use in the characterization of surfaces. One of the main reasons for this is the fact that it is quite difficult to discriminate optical signals originating from the surface from those arriving from the bulk of a given material. To illustrate this problem, imagine a typical metal sample consisting of a cube one centimetre long on each side. At a density of approximately 10 g cm^{-3} , this represents about 0.1 moles (for molybdenum, to pick an example), or approximately 6×10^{22} atoms, and of those about 1×10^{16} , that is, only one in six million atoms, are on the surface of the cube. This means that if one wants to selectively characterize a surface phenomenon, one would need to develop a technique with a large dynamic range (of at least seven orders of magnitude) and/or the ability to discriminate between signals from surface and bulk elements.

Moreover, with the advent of relatively cheap vacuum technologies over the past decades, physicists have been able to develop a large number of alternative particle-based (electrons, ions, atoms) techniques capable of selectively probing solid surfaces. Most particles interact strongly with matter, and therefore cannot penetrate deeply into the substance being probed. Consequently, whatever information can be obtained from the interactions of those particles with solid samples, it must be related to the properties of the surface. The same argument does not work as well with optical techniques, because photons penetrate through most substances to depths comparable to their wavelength, microns in the case of IR radiation. In order to overcome this difficulty, alternative ways have been devised to gain surface sensitivity with optical spectroscopies. Among them are the following:

- (1) Increasing the surface-to-bulk ratio of the sample to be studied. This is easily done in the case of highly porous materials, and has been exploited for the characterization of supported catalysts, zeolites, sol-gels and porous silicon, to mention a few.
- (2) Taking advantage of the intrinsic physical and chemical differences of surfaces introduced by the discontinuity of the bulk environment. Specifically, most solids display specific structural relaxations and reconstructions, surface

-2-

phonons and surface electronic states easy to discriminate from those of the bulk. A clearer surface specificity is introduced in the study of adsorbates by the uniqueness of the molecules present at the interface.

- (3) Taking advantage of the symmetry changes induced by the presence of a surface. Many nonlinear techniques rely on the fact that the surface breaks the centrosymmetrical nature of the bulk. The use of polarized light can also discriminate among dipole moments in different orientations.
- (4) Illuminating the sample at grazing angles. The penetration depth of photons depends on the cosine of the incidence angle and, therefore, can be reduced by this procedure. Although such an approach has limited use, it has been successfully employed in a few instances, such as for x-ray diffraction experiments.

The power of optical spectroscopies is that they are often much better developed than their electron-, ion- and atom-based counterparts, and therefore provide results that are easier to interpret. Furthermore, photon-based techniques are uniquely poised to help in the characterization of liquid–liquid, liquid–solid and even solid–solid interfaces generally inaccessible by other means. There has certainly been a renewed interest in the use of optical spectroscopies for the study of more ‘realistic’ systems such as catalysts, adsorbates, emulsions, surfactants, self-assembled layers, etc.

In this chapter we review some of the most important developments in recent years in connection with the use of optical techniques for the characterization of surfaces. We start with an overview of the different approaches available to the use of IR spectroscopy. Next, we briefly introduce some new optical characterization methods that rely on the use of lasers, including nonlinear spectroscopies. The following section addresses the use of x-rays for diffraction studies aimed at structural determinations. Lastly, passing reference is made to other optical techniques such as ellipsometry and NMR, and to spectroscopies that only partly depend on photons.

B1.22.2 IR SPECTROSCOPY

Perhaps the optical technique most used for surface characterization has been infrared (IR) spectroscopy. The reason for this may very well be because the vibrational modes identified by the interaction of IR radiation with matter are among the most specific and thus the most informative for chemical characterization. Not only can vibrational frequencies be easily identified with specific localized vibrational groups within a molecule (metal-adsorbate vibrations, O-H stretches, C-C-C deformation modes, etc), but they also depend strongly on the local environment in which the probed moiety is placed [1, 2]. The use of IR spectroscopy was greatly enhanced by the development of Fourier-transform (FTIR) spectrometers in the early 1970s, an event that brought about an enormous improvement in performance in terms of sensitivity, acquisition time, dynamic range and ease of data processing (spectra ratioing in particular) over the conventional scanning apparatus; this made the extension of IR spectroscopy to difficult systems quite feasible. The several experimental approaches pursued for the implementation of IR spectroscopy in surface studies include straight transmission, diffuse reflectance, reflection-absorption, attenuated total reflectance and emission. Each of these is discussed in some detail below.

-3-

B1.22.2.1 TRANSMISSION IR SPECTROSCOPY

The most common use of spectroscopy in general is in its transmission mode, and this was the first method employed for surface characterization as well. The pioneering work of Terenin *et al* on porous glasses [3] and of Eischens and others on chemisorption over supported metals [4] has already been reviewed in the past [5, 6]. Extensive studies have been carried out since on the characterization of catalysts upon chemisorption of many reactants, from simple molecules such as carbon and nitrogen oxides to hydrocarbons and other complex species [7, 8]. In a recent use of transmission IR absorption to surface problems, the reactivity of silicon towards water and other gases was addressed by first creating highly reproducible porous surfaces by the controlled etching of silicon single crystals [9]. Unfortunately, the general application of transmission IR spectroscopy to surface studies faces some significant limitations, in particular the need for high-surface-area solids (which usually have quite heterogeneous and ill characterized surfaces) and the restricted range of frequencies available away from the regions where the solid absorbs (above 1300 cm⁻¹ for silica, 1050 cm⁻¹ for alumina, 1200 cm⁻¹ for titania, 800 cm⁻¹ for magnesia).

B1.22.2.2 DIFFUSE-REFLECTANCE IR SPECTROSCOPY

Another useful technique for the IR characterization of surfaces in powders is diffuse-reflectance IR spectroscopy (DRIFTS) [10]. In the past, the challenge in using this approach has been in the development of efficient optics to collect the diffuse reflected radiation from the sample once illuminated with a focused IR beam, but nowadays this problem has been solved, and several cell designs are available commercially for this endeavour [11]. DRIFTS has, in theory, several advantages over conventional transmission arrangements. First, loose powders can be used without the need to press them into pellets, thus avoiding any sample distortions due to severe physical treatments, allowing for better exposures of the surface to adsorbates, and

avoiding losses in the high-frequency range due to light scattering. Second, band intensities in the DRIFTS mode can be as much as four times more intense than in the transmission mode, possibly because of the potential multiple internal reflection of the light in the vicinity of the surface before its emergence towards the detector. Lastly, DRIFTS is better for opaque samples than transmission IR spectroscopy, although the diffuse reflectance may still be low in spectral regions where the absorptivity of the substrate is high. On the negative side, there is a potential lack of reproducibility in the intensities of the DRIFTS bands because of variations in scattering coefficients with cell geometry and sample-loading procedure. Furthermore, diffuse-reflectance spectroscopy suffers from the same key limitation in transmission IR spectroscopy, namely, it requires high-surface-area samples, and therefore provides average spectra only from many types of surface local ensembles and adsorption sites.

The use of DRIFTS for the characterization of surfaces has to date been limited, but has recently been used for applications in fields as diverse as sensors development [12], soils science [13], forensic chemistry [14], corrosion [15], wood science [16] and art [17]. Given that there is in general no reason for preferring transmission over diffuse reflectance in the study of high-area powder systems, DRIFTS is likely to become much more popular in the near future.

B1.22.2.3 REFLECTION–ABSORPTION IR SPECTROSCOPY

The best way to perform IR spectroscopy studies on small samples is in the reflection–absorption (RAIRS) or attenuated total-reflectance (ATR) modes, which work best for opaque and transparent substrates, respectively. RAIRS has in fact become the method of choice for the study of adsorbates on well characterized metal samples, including single crystals. The first attempt to obtain spectra from adlayers on bulk metal samples was that of Pickering and Eckstrom, who in 1959 looked at the adsorption of carbon monoxide and hydrogen on metal films by using a multiple reflection technique with an incoming beam at close to normal incidence to the surface [18]. It soon became clear that better spectra could be obtained by using glancing incidence angles instead [19], and that the gain from using multiple reflection was not worth the complications connected with the required experimental set-up (the optimum number of reflections usually varies between 3 and 10, and results in signal intensity increases of only about 30–50% compared to those from single reflection) [20]. The theory for IR radiation reflection at metal surfaces was later developed by Greenler, who proved that only the p-polarized component of the incident beam is capable of strong interaction with adsorbates on metals, and that interference between that component of the incident and reflected rays sets an intense standing field at the surface which can yield an intensity enhancement of a factor of up to 25 compared to that from the perpendicularly polarized photons [21]. Many surface scientists have since taken advantage of these properties to perform reflection–absorption measurements of monolayers on solid metals [22, 23 and 24]. Even though the initial RAIRS experiments were carried out with molecules with large dynamic moments such as CO (in order to take advantage of their large absorption cross sections), recent FTIR developments have led to the possibility of detecting submonolayer quantities of species like hydrocarbons with much weaker signals on single crystals of less than 1 cm² area [25].

A recent example of the usefulness of RAIRS for the characterization of supported catalyst surfaces is given in [figure B1.22.1](#) which displays spectra obtained for a mixture of carbon and nitrogen monoxides coadsorbed on different palladium surfaces [26]. Both CO and NO stretching frequencies are quite sensitive to their adsorption sites, so they can be used to probe local surface sites by determining adsorption geometries in an analogous way as in organometallic discrete complexes. In this example, signals can be easily seen for the two-fold coordination of both CO and NO on Pd(100) surfaces and for three-fold, bridge and atop coordination of CO on Pd(111). The peaks from adsorption on single crystals are used as signatures for the different planes in palladium particles, so an estimate can be obtained on the relative abundance of (100) *versus* (111) sites available on the supported-metal system.

On metals in particular, the dependence of the radiation absorption by surface species on the orientation of the electrical vector can be fully exploited by using one of the several polarization techniques developed over the past few decades [27, 28, 29 and 30]. The idea behind all those approaches is to acquire the p-to-s polarized light intensity ratio during each single IR interferometer scan; since the adsorbate only absorbs the p-polarized component, that spectral ratio provides absorbance information for the surface species exclusively. Polarization-modulation methods provide the added advantage of being able to discriminate between the signals due to adsorbates and those from gas or liquid molecules. Thanks to this, RAIRS data on species chemisorbed on metals have been successfully acquired *in situ* under catalytic conditions [31], and even in electrochemical cells [32].

-5-

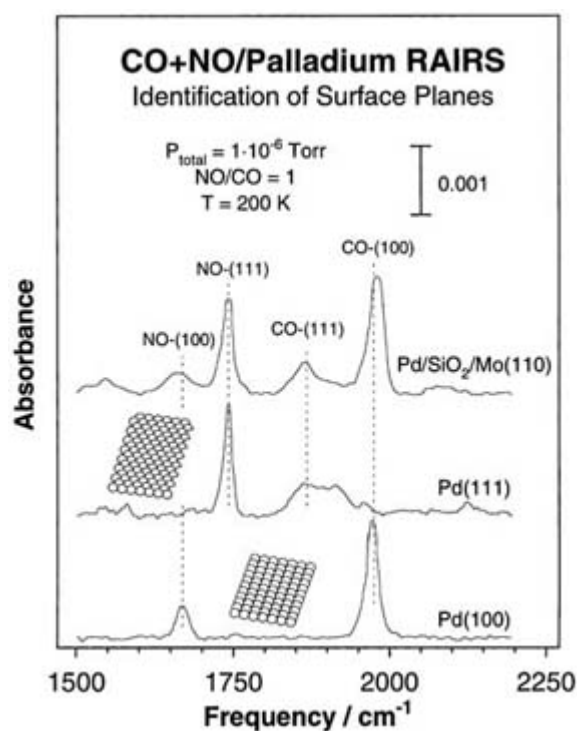


Figure B1.22.1. Reflection-absorption IR spectra (RAIRS) from palladium flat surfaces in the presence of a 1×10^{-6} Torr 1:1 NO:CO mixture at 200 K. Data are shown here for three different surfaces, namely, for Pd (100) (bottom) and Pd(111) (middle) single crystals and for palladium particles (about 500 Å in diameter) deposited on a 100 Å thick SiO₂ film grown on top of a Mo(110) single crystal. These experiments illustrate how RAIRS titration experiments can be used for the identification of specific surface sites in supported catalysts. On Pd(100) CO and NO each adsorbs on twofold sites, as indicated by their stretching bands at about 1970 and 1670 cm⁻¹, respectively. On Pd(111), on the other hand, the main IR peaks are seen around 1745 cm⁻¹ for NO (on-top adsorption) and about 1915 cm⁻¹ for CO (threefold coordination). Using those two spectra as references, the data from the supported Pd system can be analysed to obtain estimates of the relative fractions of (100) and (111) planes exposed in the metal particles [26].

The polarization dependence of the photon absorbance in metal surface systems also brings about the so-called surface selection rule, which states that only vibrational modes with dynamic moments having components perpendicular to the surface plane can be detected by RAIRS [22, 23 and 24]. This rule may in some instances limit the usefulness of the reflection technique for adsorbate identification because of the reduction in the number of modes visible in the IR spectra, but more often becomes an advantage thanks to the simplification of the data. Furthermore, the relative intensities of different vibrational modes can be used to estimate the orientation of the surface moieties. This has been particularly useful in the study of self-

assembled and Langmuir–Blodgett monolayers, where RAIRS data have been unique in providing information on the orientation of the hydrocarbon chains [33]. Figure B1.22.2 shows an example in which RAIRS was used to determine a collective change in adsorption geometry for alkyl halides on metal single crystals as the surface coverage is increased past the half-monolayer [34].

-6-

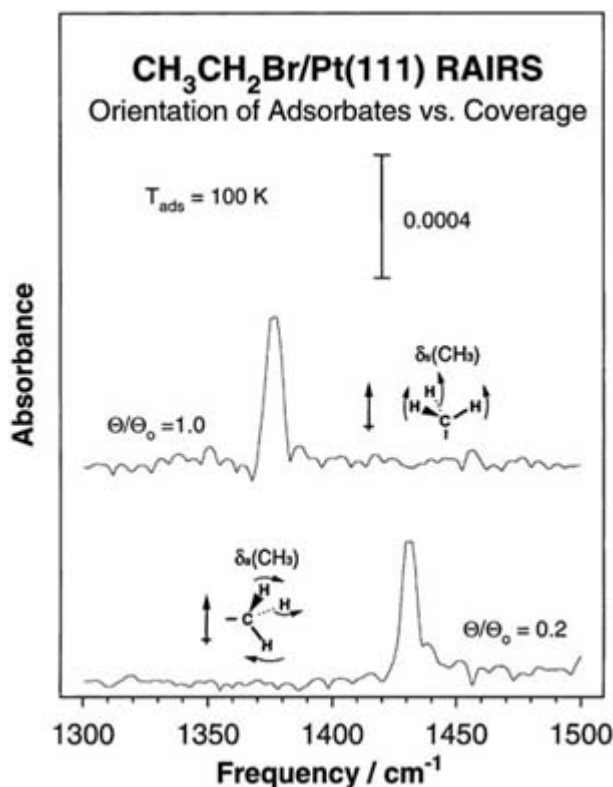


Figure B1.22.2. RAIRS data from molecular ethyl bromide adsorbed on a Pt(111) surface at 100 K. The two traces shown, which correspond to coverages of 20% and 100% saturation, illustrate the use of the RAIRS surface selection rule for the determination of adsorption geometries. Only one peak, but a different one, is observed in each case: while the signal detected at low coverages is due to the asymmetric deformation of the terminal methyl group (1431 cm^{-1}), the feature corresponding to the symmetric deformation (1375 cm^{-1}) is the one seen at high coverages instead. Given that only vibrations with dynamic dipole moments perpendicular to the surface are visible with this technique, it is concluded that a flat adsorption geometry prevails at low coverages but that a collective rearrangement of the adsorbates to a standing-up configuration takes place at about half-saturation [34].

The use of RAIRS has recently been extended from its regular mid-IR characterization of adsorbates on metals into other exciting and promising directions. For one, changes in optics and detectors have allowed for an extension of the spectral range towards the far-IR region in order to probe substrate–adsorbate vibrations [35]. The use of intense synchrotron sources in particular looks quite promising for the detection of such weak modes [36]. Thanks to the speed with which Fourier-transform spectrometers can acquire complete IR spectra, kinetic studies of surface reactions can be carried out as well. To date this has only been done in a few cases, usually for reactions that take seconds or more to occur [37], but the advent of step scanners promises the availability of time resolutions of 10^{-8} s or better in the near future [38]. In terms of the lifetime of the vibrational excitations themselves, this can in some instances be estimated from IR absorption line shapes. Because of the efficient coupling between the vibrations of adsorbate and phonons and other electronic surface states, the former are generally short-lived, and therefore yield IR absorption bands several wavenumbers wide. Nevertheless, bands as narrow as 0.7 cm^{-1} have been observed in some cases [39]. Finally, the use of RAIRS is not limited to metal surfaces. Although the surface selection rules change

significantly for non-metal surfaces, they can still be used to obtain orientational information for adsorbates on transparent substrates, as recently demonstrated in the elegant study by Hoffmann *et al* on the adsorption of long-chain hydrocarbons on silicon (figure B1.22.3) [40], and even for the analysis of air–liquid interfaces [41]. There are many clear new directions still unexplored for the use of RAIRS in surface characterization studies.

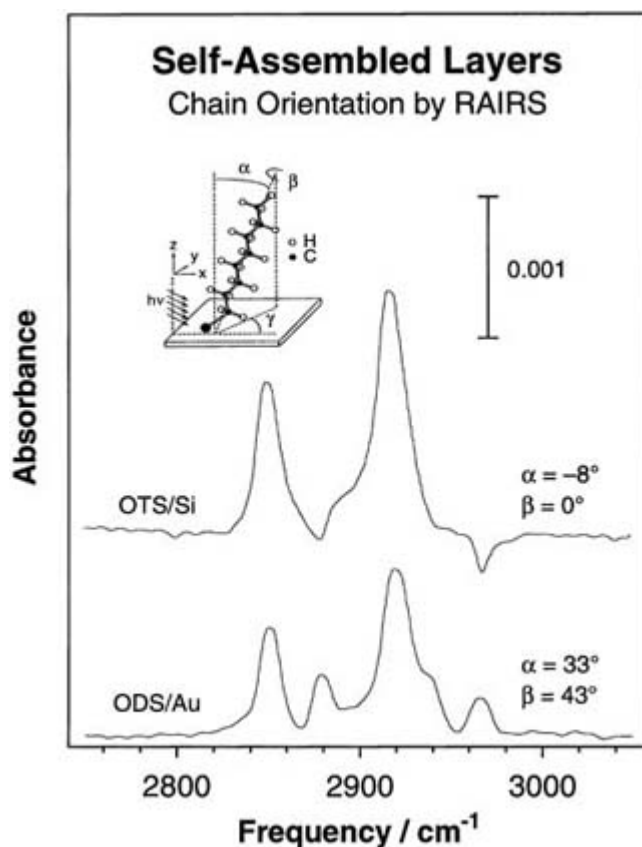


Figure B1.22.3. RAIRS data in the C–H stretching region from two different self-assembled monolayers, namely, from a monolayer of dioctadecylsulfide (ODS) on gold (bottom), and from a monolayer of octadecyltrichlorosilane (OTS) on silicon (top). Although the RAIRS surface selection rules for non-metallic substrates are more complex than those which apply to metals, they can still be used to determine adsorption geometries. The spectra shown here were, in fact, analysed to yield the tilt (α) and twist (β) angles of the molecular chains in each case with respect to the surface plane (the resulting values are also given in the figure) [40].

B1.22.2.4 ATTENUATED TOTAL REFLECTANCE IR SPECTROSCOPY

In 1960, Harrick demonstrated that, for transparent substrates, absorption spectra of adsorbed layers could be obtained using internal reflection [42]. By cutting the sample in a specific trapezoidal shape, the IR beam can be made to enter through one end, bounce internally a number of times from the flat parallel edges, and exit the other end without any losses, leading to high adsorption coefficients for the species adsorbed on the external surfaces of the plate (higher than in the case of external reflection) [24]. This is the basis for the ATR technique.

In recent years, ATR has been used primarily in connection with the characterization of semiconductor surfaces. For instance, ATR studies have led to the detailed mapping of the complex series of reconstructions that silicon surfaces follow upon thermal treatment and/or hydrogen exposures [43]. Surface electronic excitations have been studied with this technique as well; see, for instance, the pioneering work of McCombe *et al* on the characterization of inter-subband optical transitions in silicon MOS field-effect transistors (figure B1.22.4) [44]. One interesting additional extension of the use of multiple internal reflection to the characterization of non-transparent samples was discussed by Bermudez, who suggested that the sensitivity to adsorbates in IR-reflection spectroscopy can be enhanced by burying a metal layer beneath the surface of a dielectric material [45].

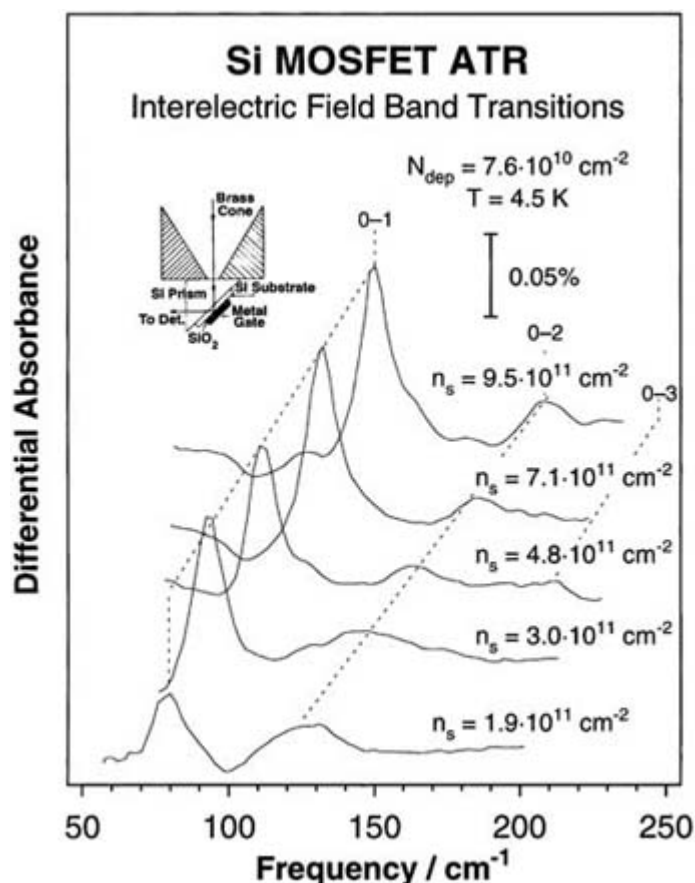


Figure B1.22.4. Differential IR absorption spectra from a metal–oxide silicon field-effect transistor (MOSFET) as a function of gate voltage (or inversion layer density, n_s , which is the parameter reported in the figure). Clear peaks are seen in these spectra for the 0–1, 0–2 and 0–3 inter-electric-field subband transitions that develop for charge carriers when confined to a narrow ($<100 \text{ \AA}$) region near the oxide–semiconductor interface. The inset shows a schematic representation of the attenuated total reflection (ATR) arrangement used in these experiments. These data provide an example of the use of ATR IR spectroscopy for the probing of electronic states in semiconductor surfaces [44].

B1.22.2.5 OTHER SURFACE IR SPECTROSCOPY ARRANGEMENTS

There have been a few other experimental set-ups developed for the IR characterization of surfaces. Photoacoustic (PAS), or, more generally, photothermal IR spectroscopy relies on temperature fluctuations caused by irradiating the sample with a modulated monochromatic beam: the acoustic pressure wave created in the gas layer adjacent to the solid by the adsorption of light is measured as a function of photon wavelength

in order to determine the absorption spectra [11]. It has sometimes been thought that PAS is more surface sensitive than DRIFTS, but in fact that depends on the specific optical and thermal properties of the material being studied. In emission spectrometry (EMS), the IR radiation emitted by the sample is directly collected and analysed. The detection of the (non-monochromatized) IR radiation from thin films has recently been combined with molecular beam techniques in order to perform differential microcalorimetric measurements on adsorption processes [46]. Finally, the sample itself can be used as the detector of IR radiation. None of these techniques have found much use in surface studies to date.

B1.22.3 LASER-BASED SPECTROSCOPIES

Although the development of a large variety of lasers with different spectral ranges, intensities and temporal resolutions has led to the surge of many new optical characterization techniques, most of those have yet to make a large impact in surface science. As discussed above, the signals from surfaces are often weak and hard to differentiate from those from the bulk, and this is particularly troublesome in nonlinear techniques which rely on the absorption of more than one photon. Furthermore, the increase of the laser power to levels where signal intensities are no longer an issue may lead to damaging of the substrate. In spite of these limitations, some laser-based methods have already been developed for surface-characterization studies.

B1.22.3.1 RAMAN SPECTROSCOPY

Perhaps the best known and most used optical spectroscopy which relies on the use of lasers is Raman spectroscopy. Because Raman spectroscopy is based on the inelastic scattering of photons, the signals are usually weak, and are often masked by fluorescence and/or Rayleigh scattering processes. The interest in using Raman for the vibrational characterization of surfaces arises from the fact that the technique can be used *in situ* under non-vacuum environments, and also because it follows selection rules that complement those of IR spectroscopy.

Regular Raman has been employed mainly for the characterization of high-surface-area solids [47]. Specifically, a good methodology has been developed for the determination of bond orders, bond lengths and local geometries in many metal oxides used for catalysis. [Figure B1.22.5](#) illustrates this point by displaying some examples where the Raman vibrational signals for metal–oxygen single and double bonds as well as for oxygen–metal–oxygen deformations were used to determine the structure of a number of supported and highly dispersed transition-metal oxides [48].

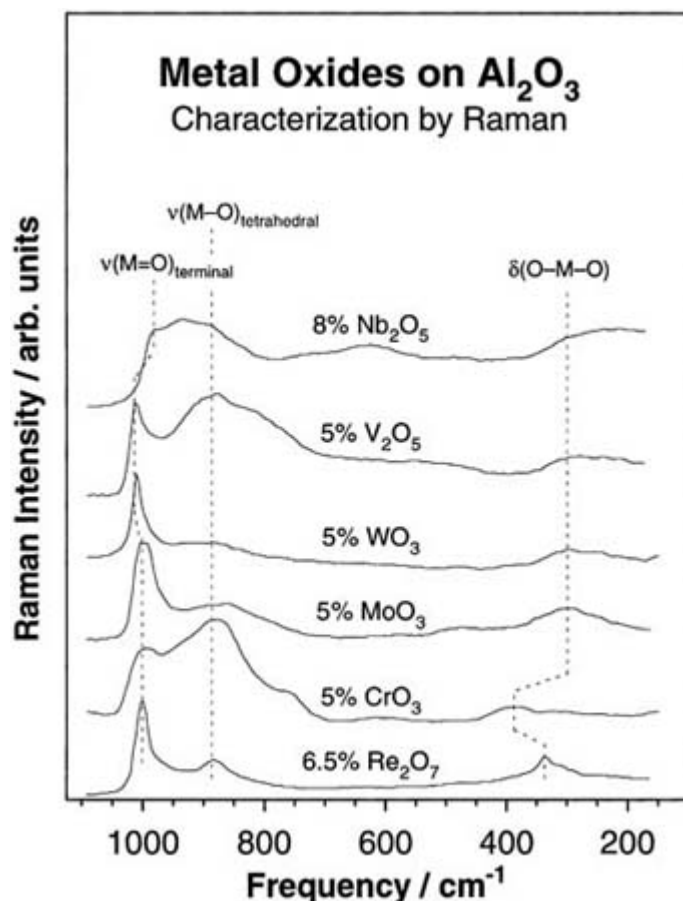


Figure B1.22.5. *In situ* Raman spectra from a family of transition-metal oxides dispersed on high-surface-area alumina substrates. Three distinct regions can be differentiated in these spectra, namely, the peaks around 1000 cm^{-1} , which are assigned to the stretching frequency of terminal metal–oxygen double bonds, the features about 900 cm^{-1} , corresponding to metal–oxygen stretches in tetrahedral coordination sites, and the low-frequency ($<400\text{ cm}^{-1}$) range associated with oxygen–metal–oxygen deformation modes. Data such as these can be used to determine the nature and geometry of supported oxides as a function of metal loading and subsequent treatment [48].

In an interesting development in Raman spectroscopy, Fleischmann *et al* noticed in 1974 that there is a significant enhancement in the Raman signal intensities from solid surfaces if the substrate is comprised of small silver particles [49]. The same phenomenon has since been observed with copper, silver, gold, lithium, sodium, potassium, indium, platinum and rhodium, and has become the basis for surface-enhanced Raman spectroscopy (SERS) [50, 51]. The reasons for this enhancement are still not completely clear, but have been recognized to be the result of a combination of effects, including a surface electromagnetic field enhancement (in particular when illuminating rough samples with photons of energies near those of localized plasmons) and a chemical enhancement due to the change of polarizability in molecules when interacting with surfaces [52].

Since its initial development, SERS has been used for the surface characterization of a good number of systems. One important extension to the use of SERS has been in the determination of surface geometries. Figure B1.22.6 shows

an example of the SERS C–H stretching frequency data used to determine the different chemisorption geometries of 2-butanol and 2-butanethiol on silver electrodes [53]. Notice that, being an optical technique, SERS works quite well in solid–liquid interfaces. On the other hand, the need for signal enhancement normally limits the use of SERS to a handful of metals and/or to samples with rough surfaces.

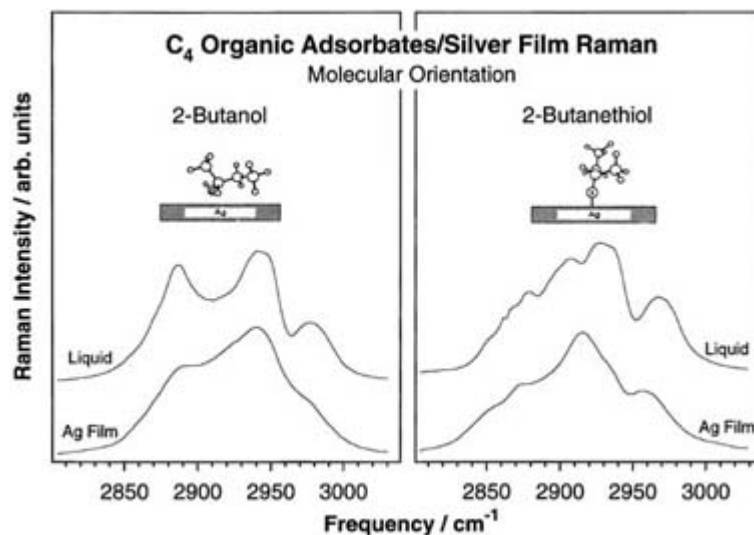


Figure B1.22.6. Raman spectra in the C–H stretching region from 2-butanol (left frame) and 2-butanethiol (right), each either as bulk liquid (top traces) or adsorbed on a rough silver electrode surface (bottom). An analysis of the relative intensities of the different vibrational modes led to the proposed adsorption structures depicted in the corresponding panels [53]. This example illustrates the usefulness of Raman spectroscopy for the determination of adsorption geometries, but also points to its main limitation, namely the need to use rough silver surfaces to achieve adequate signal-to-noise levels.

Another recent development in the use of Raman spectroscopy for the characterization of surfaces has been the employment of UV light for the initial excitation of the sample [54]. The advantage of UV over conventional Raman spectroscopy is twofold: (1) since the normal Raman scattering cross sections are proportional to the fourth power of the scattered light frequency, the use of higher-energy photons significantly increases the signal intensity and (2) by using UV light the spectral range is moved away from that where fluorescence de-excitation is observed. This allows for the Raman characterization of virtually any high-surface-area sample, including opaque solids such as black carbon. Unfortunately, UV-Raman spectroscopy is still not commercially available.

B1.22.3.2 OTHER NONLINEAR OPTICAL TECHNIQUES

Other nonlinear optical spectroscopies have gained much prominence in recent years. Two techniques in particular have become quite popular among surface scientists, namely, second harmonic (SHG) [55] and sum-frequency (SFG) [56] generation. The reason why both SHG and SFG can probe interfaces selectively without being overwhelmed by the signal from the bulk is that they rely on second-order processes that are electric-dipole forbidden in centrosymmetric media; by breaking the bulk symmetry, the surface places the molecular species in an environment where their second-order nonlinear susceptibility, the term responsible for the absorption of SHG and SFG signals, becomes non-zero.

In SHG the sample is illuminated with light of a single colour and the component at twice the initial frequency is filtered from the emitted light and analysed. Because these experiments usually involve near-IR or visible–UV photons, SHG most often probes electronic transitions. In fact, SHG has often been used as a way to measure changes in work function or localized electrostatic surface potentials. When using polarized light, SHG can also be used to determine the geometrical alignment of polar molecules at interfaces and, by sweeping the incident photon energy, spectroscopic information can be obtained on molecular orbital energies as well. Figure B1.22.7 shows an example of the latter for the case of rhodamine 6G [55]. This figure also shows a clever extension of the technique as a microscope to provide spatial information on adsorbates.

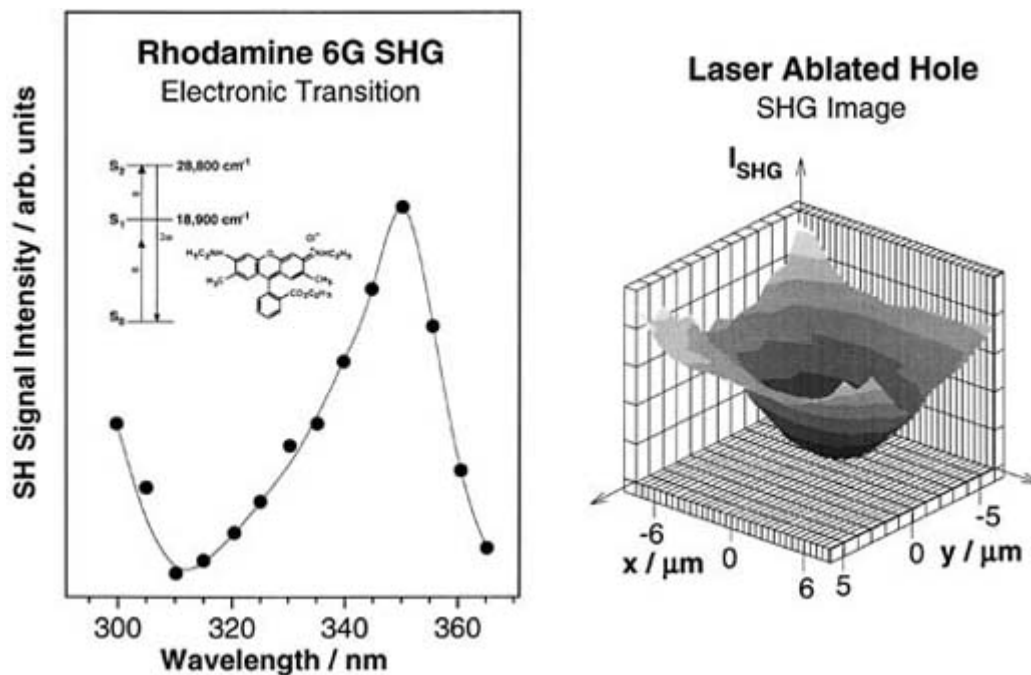


Figure B1.22.7. Left: resonant second-harmonic generation (SHG) spectrum from rhodamine 6G. The inset displays the resonant electronic transition induced by the two-photon absorption process at a wavelength of approximately 350 nm. Right: spatially resolved image of a laser-ablated hole in a rhodamine 6G dye monolayer on fused quartz, mapped by recording the SHG signal as a function of position in the film [55]. SHG can be used not only for the characterization of electronic transitions within a given substance, but also as a microscopy tool.

By combining two beams on the surface, one of visible or IR fixed frequency and a second, of variable energy in the IR region, resonance absorption can be measured by detecting the intensity of the outgoing light resulting from the addition of the two incident beams as a function of the photon energy of the variable laser. The net effect of this SFG is the acquisition of vibrational absorption spectra for surface species where signals are seen only for the modes active in both IR and Raman. Vibrational information can be obtained with SFG for almost any interface as long as lasers are available to cover the frequency range of interest and the bulk materials are transparent to the laser light. Also, as with many of the other techniques described above, orientational information can be obtained with SFG as well. Figure B1.22.8 displays data demonstrating that an increase in the concentration of acetonitrile dissolved in water leads to a collective molecular orientation change at the air/water interface [57].

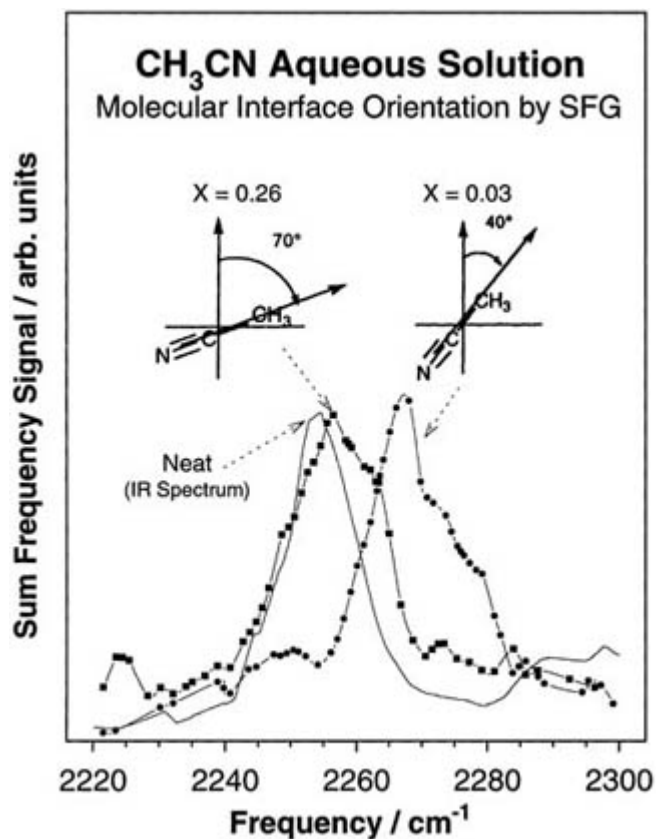


Figure B1.22.8. Sum-frequency generation (SFG) spectra in the $\text{C}\equiv\text{N}$ stretching region from the air/aqueous acetonitrile interfaces of two solutions with different concentrations. The solid curve is the IR transmission spectrum of neat bulk CH_3CN , provided here for reference. The polar acetonitrile molecules adopt a specific orientation in the air/water interface with a tilt angle that changes with changing concentration, from 40° from the surface normal in dilute solutions (molar fractions less than 0.07) to 70° at higher concentrations. This change is manifested here by the shift in the $\text{C}\equiv\text{N}$ stretching frequency seen by SFG [57]. SFG is one of the very few techniques capable of probing liquid/gas, liquid/liquid, and even liquid/solid interfaces.

There are a few other surface-sensitive characterization techniques that also rely on the use of lasers. For instance surface-plasmon resonance (SPR) measurements have been used to follow changes in surface optical properties as a function of time as the sample is modified by, for instance, adsorption processes [58]. SPR has proven useful to image adsorption patterns on surfaces as well [59].

B1.22.3.3 TIME-RESOLVED PUMP-PROBE EXPERIMENTS

The dynamics of fast processes such as electron and energy transfers and vibrational and electronic de-excitations can be probed by using short-pulsed lasers. The experimental developments that have made possible the direct probing of molecular dissociation steps and other ultrafast processes in real time (in the femtosecond time range) have, in a few cases, been extended to the study of surface phenomena. For instance, two-photon photoemission has been used to study the dynamics of electrons at interfaces [60]. Vibrational relaxation times have also been measured for a number of modes such as the O–H stretching in silica and the C–O stretching in carbon monoxide adsorbed on transition metals [61]. Pump-probe laser experiments such as these are difficult, but the field is still in its infancy, and much is expected in this direction in the near future.

B1.22.4 X-RAY DIFFRACTION AND X-RAY ABSORPTION

Because x-rays are particularly penetrating, they are very useful in probing solids, but are not as well suited for the analysis of surfaces. X-ray diffraction (XRD) methods are nevertheless used routinely in the characterization of powders and of supported catalysts to extract information about the degree of crystallinity and the nature and crystallographic phases of oxides, nitrides and carbides [62, 63]. Particle size and dispersion data are often acquired with XRD as well.

One way to obtain surface sensitivity in XRD experiments with crystalline samples is to illuminate the substrate at glancing angles. Under normal conditions, x-rays projected onto the sample at incident angles of less than 10° still penetrate to a depth of 10 μm or more but, beyond a critical angle, x-ray photons are completely reflected from the surface and light propagation into the solid is *via* a rapidly attenuating evanescent wave, and this renders the x-ray probe quite surface sensitive [64]. There are several inherent difficulties in implementing grazing-angle XRD experiments, which require focusing high-intensity x-rays onto surfaces at angles of the order of 0.1° , but recent experiments have proved the usefulness of this technique in providing interesting information on the structure of robust surfaces such as oxides, nitrides, silicides and other thin films.

A related technique that also relies on the interference of x-rays for solid characterization is extended x-ray absorption fine structure (EXAFS) [65, 66]. Because the basis for EXAFS is the interference of outgoing photoelectrons with their scattered waves from nearby atoms, it does not require long-range order to work (as opposed to diffraction techniques), and provides information about the local geometry around specific atomic centres. Unfortunately, EXAFS requires the high-intensity and tunable photon sources typically available only at synchrotron facilities. Further limitations to the development of surface-sensitive EXAFS (SEXAFS) have come from the fact that it requires technology entirely different from that of regular EXAFS, involving in many cases ultrahigh-vacuum environments and/or photoelectron detection. One interesting advance in SEXAFS came with the design by Stöhr *et al* of fluorescence detectors for the x-rays absorbed by the surface species of small samples; that allows for the characterization of well defined systems such as single crystals under non-vacuum conditions [67]. Figure B1.22.9 shows the S K-edge x-ray absorption data obtained for a $(2 \times 2)\text{S-Ni}(100)$ overlayer using their original experimental set-up. This approach has since been extended to the analysis of lighter atoms (C, O, F) on many different substrates and under atmospheric pressures [68].

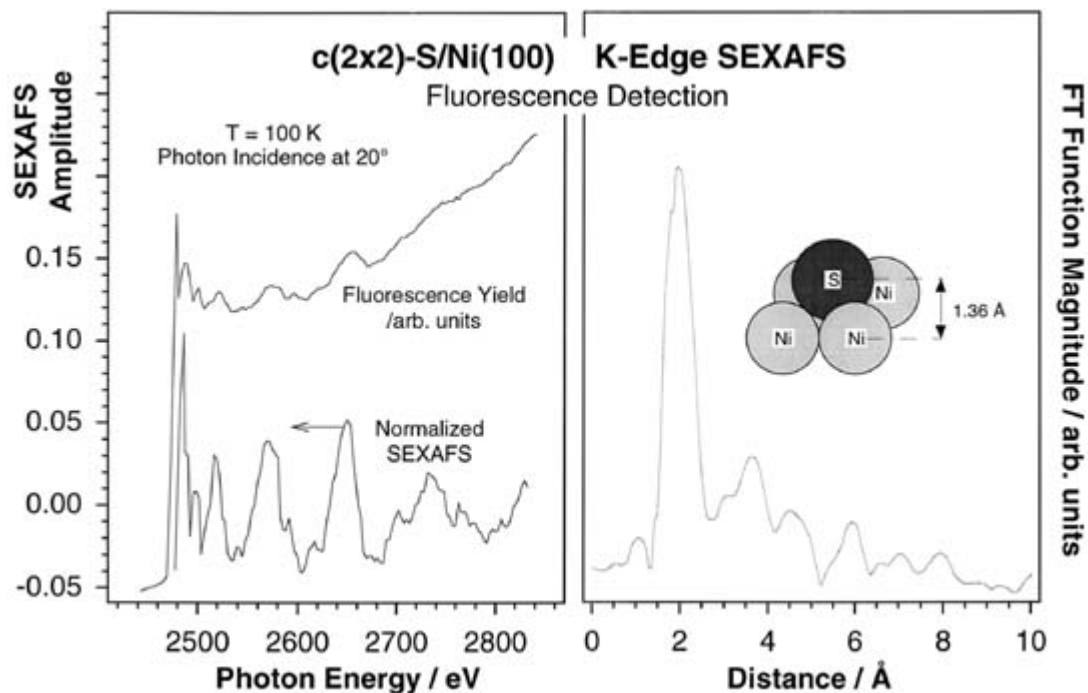


Figure B1.22.9. Fluorescence-yield, surface-extended x-ray absorption fine structure (FY-SEXAFS) spectra for the S K-edge of a $c(2 \times 2)$ ordered monolayer of sulfur atoms adsorbed on a Ni(100) surface. The upper trace in the left panel corresponds to the raw data obtained at an incident photon angle of 20° from the surface plane, while the bottom trace displays the background-subtracted and normalized SEXAFS oscillations calculated from the original spectrum. The right panel, which corresponds to the Fourier-transformed SEXAFS data, provides information on both the Ni–S bond length (2.22 \AA) and the Ni near-neighbour coordination number around each sulfur atom (4) [67]. Because EXAFS relies on the interference of an outgoing photoelectron with its own scattering from nearby atoms, it provides local geometry information without requiring long-range order. This example also illustrates the high sensitivity of the technique (these experiments were carried out with a 1 cm^2 area single crystal). The use of fluorescence detection allows for the extension of this type of study to samples in non-vacuum environments [68].

Soon after the development of EXAFS it was recognized that the signal near the x-ray absorption edge is quite complex and provides information on electronic transitions from atomic core levels to valence bands and/or molecular orbitals [69]. The analysis of that signal constitutes the basis for a technique named near-edge x-ray absorption fine structure (NEXAFS, or XANES). The shape of the x-ray absorption spectra near the absorption edge has long been used as an empirical fingerprint for the local chemical environment of oxides and other supported catalysts, but newer developments allow for the extraction of a more detailed picture of the nature and geometrical arrangement of adsorbates from those data. This is possible thanks in great part to the combination of the polarized nature of synchrotron radiation and the simplicity of the electronic transition dipoles for absorption from core levels [70]. [Figure B1.22.10](#) displays an example where the geometry of vinyl moieties adsorbed on Ni(100) surfaces was determined by using NEXAFS [71].

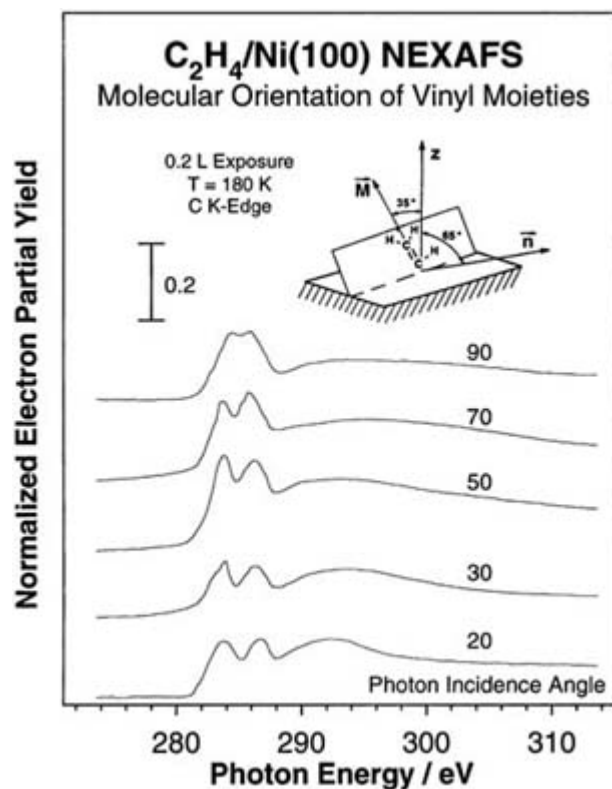


Figure B1.22.10. Carbon K-edge near-edge x-ray absorption (NEXAFS) spectra as a function of photon incidence angle from a submonolayer of vinyl moieties adsorbed on Ni(100) (prepared by dosing 0.2 l of ethylene on that surface at 180 K). Several electronic transitions are identified in these spectra, to both the pi (284 and 286 eV) and the sigma (>292 eV) unoccupied levels of the molecule. The relative variations in the intensities of those peaks with incidence angle can be easily converted into adsorption geometry data; the vinyl plane was found in this case to be at a tilt angle of about 65° from the surface [71]. Similar geometrical determinations using NEXAFS have been carried out for a number of simple adsorbate systems over the past few decades.

B1.22.5 OTHER OPTICAL TECHNIQUES

A few additional optical techniques need to be mentioned in this review. As discussed above, these are by and large well known spectroscopies for the study of bulk samples; it is only their extension to the study of surfaces what has not been realized to its fullest potential yet.

B1.22.5.1 OTHER UV-VISIBLE OPTICAL TECHNIQUES

Spectroscopies such as UV-visible absorption and phosphorescence and fluorescence detection are routinely used to probe electronic transitions in bulk materials, but they are seldom used to look at the properties of surfaces [72]. As with other optical techniques, one of the main problems here is the lack of surface discrimination, a problem that has sometime been bypassed by either using thin films of the materials of interest [73, 74], or by using a reflection detection scheme. Modulation of a parameter, such as electric or magnetic fields, stress, or temperature, which affects the optical properties of the sample and detection of the AC component of the signal induced by such periodic changes, can also be used to achieve good surface sensitivity [75]. This latter approach is the basis for techniques such as surface reflectance spectroscopy, reflectance difference spectroscopy/reflectance anisotropy spectroscopy, surface photoadsorption

spectroscopy and surface differential reflectivity [76, 77 and 78]. Early optical characterization studies of solid surfaces were instrumental in the detection and characterization of intrinsic surface electronic states due to the uniqueness of the interface environments. Ellipsometry is also a mature technique often used to obtain film thickness and other optical properties [79].

One interesting new field in the area of optical spectroscopy is near-field scanning optical microscopy, a technique that allows for the imaging of surfaces down to sub-micron resolution and for the detection and characterization of single molecules [80, 81]. When applied to the study of surfaces, this approach is capable of identifying individual adsorbates, as in the case of oxazine molecules dispersed on a polymer film, illustrated in figure B1.22.11 [82]. Absorption and emission spectra of individual molecules can be obtained with this technique as well, and time-dependent measurements can be used to follow the dynamics of surface processes.

B1.22.5.2 MAGNETIC RESONANCE

NMR has developed into a powerful analytical technique in the past decades, and has been used extensively in the characterization of a great number of chemical systems. Its extension to the study of surfaces, however, has been hampered by the need of large samples because of its poor sensitivity. On the other hand, the development of magic-angle-spinning NMR (MAS-NMR) and the extension of NMR to many nuclei besides hydrogen have opened the doors for the use of that technique to many solids. For instance, MAS ^1H NMR has been quite useful in the research on Brønsted acidity in oxides [83]. Also, the study of zeolites with NMR is now practically routine: the chemical shifts in ^{29}Si NMR data are easily interpreted in terms of the number of aluminium atoms next to a given silicon centre, and the position of the ^{27}Al peaks provides information on the coordination number and geometry of the aluminium atoms [84]. ^{129}Xe NMR has been used to probe both the local environment inside porous materials [83] and heterogeneities in adsorbates [85]. Dynamic studies on the thermal conversion of adsorbates on transition metal catalysts have been performed as well [86].

Other magnetic measurements of catalysts include electron paramagnetic resonance and magnetic susceptibility. Although those are not as common as NMR, they can be used to look at the properties of paramagnetic and ferromagnetic samples. Examples of these applications can be found in the literature [87, 88].

-18-

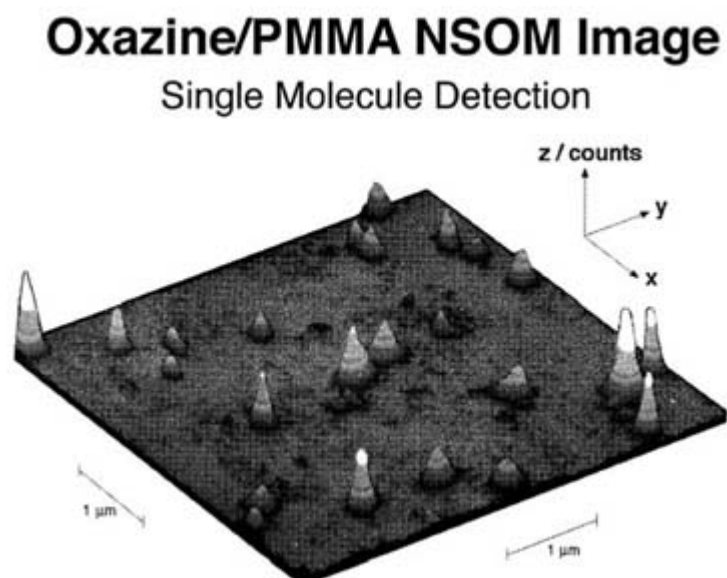


Figure B1.22.11. Near-field scanning optical microscopy fluorescence image of oxazine molecules dispersed on a PMMA film surface. Each protuberance in this three-dimensional plot corresponds to the detection of a single molecule, the different intensities of those features being due to different orientations of the molecules. Sub-diffraction resolution, in this case on the order of a fraction of a micron, can be achieved by the near-field scanning arrangement. Spectroscopic characterization of each molecule is also possible. (Reprinted with permission from [82]. Copyright 1996 American Chemical Society.)

B1.22.5.3 SPECTROSCOPIES WHICH RELY ONLY IN PART ON PHOTONS

A number of surface-sensitive spectroscopies rely only in part on photons. On the one hand, there are techniques where the sample is excited by electromagnetic radiation but where other particles ejected from the sample are used for the characterization of the surface (photons in; electrons, ions or neutral atoms or moieties out). These include photoelectron spectroscopies (both x-ray- and UV-based) [89, 90 and 91], photon stimulated desorption [92], and others. At the other end, a number of methods are based on a particles-in/photons-out set-up. These include inverse photoemission and ion- and electron-stimulated fluorescence [93, 94]. All these techniques are discussed elsewhere in this encyclopaedia.

REFERENCES

- [1] Nakamoto K 1978 *Infrared and Raman Spectra of Inorganic and Coordination Compounds* 3rd edn (New York: Wiley-Interscience)
- [2] Socrates G 1994 *Infrared Characteristic Group Frequencies: Tables and Charts* 2nd edn (Chichester: Wiley)
- [3] Yaroslavskii N G and Terenin A N 1949 Infrared absorption spectra of adsorbed molecules *Dokl. Akad. Nauk* **66** 885–8
- [4] Eischens R P, Pliskin W A and Francis S A 1954 Infrared spectra of chemisorbed CO *J. Chem. Phys.* **22** 1986–7
- [5] Little L H 1966 *Infrared Spectra of Adsorbed Species* (New York: Academic)
- [6] Hair M L 1967 *Infrared Spectroscopy in Surface Chemistry* (New York: Marcel Dekker)
- [7] Sheppard N and De La Cruz C 1996 Vibrational spectra of hydrocarbons adsorbed on metals. Part I. Introductory principles, ethylene, and higher acyclic alkenes *Adv. Catal.* **41** 1–112
- [8] Sheppard N and De La Cruz C 1998 Vibrational spectra of hydrocarbons adsorbed on metals. Part II. Adsorbed acyclic alkynes and alkanes, cyclic hydrocarbons including aromatics and surface hydrocarbon groups derived from the decomposition of alkyl halides, etc *Adv. Catal.* **42** 181–313
- [9] Gupta P, Dillon A C, Bracker A S and George S M 1991 FTIR studies of H₂O and D₂O decomposition on porous silicon surfaces *Surf. Sci.* **245** 360–72
- [10] Willey R R 1976 Fourier transform infrared spectrophotometer for transmittance and diffuse reflectance measurements *Appl. Spectrosc.* **30** 593–601
- [11] Griffiths P R and de Haseth J A 1986 *Fourier Transform Infrared Spectrometry* (New York: Wiley)
- [12] Benitez J J, Centeno M A, Merdrignac O M, Guyader J, Laurent Y and Odriozola J A 1995 DRIFTS chamber for *in situ* and simultaneous study of infrared and electrical response of sensors *Appl. Spectrosc.* **49** 1094–6
- [13] Vreugdenhil A J and Butler I S 1998 Investigation of MMT adsorption on soils by diffuse reflectance infrared spectroscopy DRIFTS and headspace analysis gas-phase infrared spectroscopy HAGIS *Appl. Organomet. Chem.*

- [14] Lennard C J, Mazzella W D and Margot P A 1993 Some applications of diffuse reflectance infrared Fourier transform spectroscopy DRIFTS in forensic science *Analysis* **21** M34–7
- [15] Job A, Somuah S K, Siddiqui A H and Abbas N M 1990 Diffuse reflectance infrared DRIFTS studies of corrosion products *Anal. Lett.* **23** 1537–52
- [16] Kazayawoko M, Balatinez J J and Woodhams R T 1997 Diffuse reflectance Fourier transform infrared spectra of wood fibers treated with maleated polypropylenes *J. Appl. Polymer Sci.* **66** 1163–73
- [17] Zeine C and Grobe J 1997 Diffuse reflectance infrared Fourier transform DRIFT spectroscopy in the preservation of historical monuments: studies on salt migration *Mikrochim. Acta* **125** 279–82
- [18] Pickering H L and Eckstrom H C 1959 Studies by infrared absorption *J. Phys. Chem.* **63** 512–17
- [19] Francis S A and Ellison A H 1959 Infrared spectra of monolayers on metal mirrors *J. Opt. Soc. Am.* **49** 131–8
- [20] Hollins P and Pritchard J 1985 Infrared studies of chemisorbed layers on single crystals *Prog. Surf. Sci.* **19** 275–350
- [21] Greenler R G 1966 Infrared study of adsorbed molecules on metal surfaces by reflection techniques *J. Chem. Phys.* **44** 310–15
- [22] Bradshaw A M 1982 Vibrational spectroscopy of adsorbed atoms and molecules *Appl. Surf. Sci.* **11/12** 712–29
- [23] Hoffmann F M 1983 Infrared reflection–absorption spectroscopy of adsorbed molecules *Surf. Sci. Rep.* **3** 107–92
- [24] Chabal Y J 1988 Surface infrared spectroscopy *Surf. Sci. Rep.* **8** 211–357

- [25] Zaera F and Hoffmann H 1991 Detection of chemisorbed methyl and methylene groups: surface chemistry of methyl iodide on Pt(111) *J. Phys. Chem.* **95** 6297–303
- [26] Xu X, Chen P and Goodman D W 1994 A comparative study of the coadsorption of CO and NO on Pd(100), Pd(111), and silica-supported palladium particles with infrared reflection–absorption spectroscopy *J. Phys. Chem.* **98** 9242–6
- [27] Dowrey A E and Marcott C 1982 A double-modulation Fourier transform infrared approach to studying adsorbates on metal surfaces *Appl. Spectrosc.* **36** 414–16
- [28] Ishida H, Ishino Y, Buijs H, Tripp C and Dignam M J 1987 Polarization-modulation FT-IR reflection spectroscopy using a polarizing Michelson interferometer *Appl. Spectrosc.* **41** 1288–94
- [29] Hoffmann H, Wright N A, Zaera F and Griffiths P R 1989 Differential-polarization dual-beam FT-IR spectrometer for surface analysis *Talanta* **36** 125–31
- [30] Barner B J, Green M J, Sáez E I and Corn R M 1991 Polarization modulation Fourier transform infrared reflectance measurements of thin films and monolayers at metal surfaces utilizing real-time sampling electronics *Anal. Chem.* **63** 55–60
- [31] Hoffmann F M and Weisel M D 1993 Fourier transform infrared reflection absorption spectroscopy studies of adsorbates and surface reactions—bridging the pressure gap between surface science and catalysis *J. Vac. Sci. Technol. A* **11** 1957–63
- [32] Bewick A and Pons S 1985 Infrared spectroscopy of the electrode–electrolyte solution interface *Advances in Infrared and Raman Spectroscopy* ed R J H Clark and R E Hester (New York: Wiley Heyden) **12** 1–63
- [33] Parikh A N and Allara D L 1992 Quantitative determination of molecular structure in multilayered thin films of biaxial and lower symmetry from photon spectroscopies. I. Reflection infrared vibrational spectroscopy *J. Chem. Phys.* **96** 927–45

- [34] Zaera F, Hoffmann H and Griffiths P R 1990 Determination of molecular chemisorption geometries using reflection-absorption infrared spectroscopy: alkyl halides on Pt(111) *J. Electron. Spectrosc. Relat. Phenom.* **54/55** 705–15
- [35] Malik I J and Trenary M 1989 Infrared reflection-absorption study of the adsorbate-substrate stretch of CO on Pt(111) *Surf. Sci.* **214** L237–45
- [36] Dumas P, Suhren M, Chabal Y J, Hirschmugl C J and Williams G P 1997 Adsorption and reactivity of NO on Cu(111): a synchrotron infrared reflection absorption spectroscopic study *Surf. Sci.* **371** 200–12
- [37] Janssens T V W, Stone D, Hemminger J C and Zaera F 1998 Kinetics and mechanism for the H/D exchange between ethylene and deuterium over Pt(111) *J. Catal.* **177** 284–95
- [38] Johnson T J, Simon A, Weil J M and Harris G W 1993 Applications of time-resolved step-scan and rapid-scan FT-IR spectroscopy: dynamics from ten seconds to ten nanoseconds *Appl. Spectrosc.* **47** 1376–81
- [39] Agrawal V K and Trenary M 1989 Infrared spectrum from 400 to 1000 cm^{-1} of PF_3 chemisorbed on the Pt(111) surface *J. Vac. Sci. Technol. A* **7** 2235–7
- [40] Hoffmann H, Mayer U and Krischanitz A 1995 Structure of alkylsiloxane monolayers on silicon surfaces investigated by external reflection infrared spectroscopy *Langmuir* **11** 1304–12
- [41] Mendelsohn R, Brauner J W and Gericke A 1995 External infrared reflection absorption spectrometry of monolayer films at the air-water interface *Ann. Rev. Phys. Chem.* **46** 305–34
- [42] Harrick N J 1960 Physics and chemistry of surfaces from frustrated total internal reflections *Phys. Rev. Lett.* **4** 224–6
-

- [43] Chabal Y J 1986 High-resolution infrared spectroscopy of adsorbates on semiconductor surfaces: hydrogen on silicon(100) and germanium(100) *Surf. Sci.* **168** 594–608
- [44] McCombe B D, Holm R T and Schafer D E 1979 Frequency domain studies of intersubband optical transitions in Si inversion layers *Solid State Commun* **32** 603–8
- [45] Bermudez V M 1992 Infrared optical properties of dielectric/metal layer structures of relevance to reflection absorption spectroscopy *J. Vac. Sci. Technol. A* **10** 152–7
- [46] Borroni-Bird C E and King D A 1991 An ultrahigh vacuum single crystal adsorption microcalorimeter *Rev. Sci. Instrum* **62** 2177–85
- [47] Stencel J M 1990 *Raman Spectroscopy for Catalysis* (New York: Van Nostrand Reinhold)
- [48] Vuurman M A and Wachs I E 1992 *In situ* Raman spectroscopy of alumina-supported metal oxide catalysts *J. Phys. Chem.* **96** 5008–16
- [49] Fleischmann M, Hendra P J and McQuillan A J 1974 Raman spectra of pyridine adsorbed at a silver electrode *Chem. Phys. Lett.* **26** 163–6
- [50] Moskovits M 1985 Surface-enhanced spectroscopy *Rev. Mod. Phys.* **57** 783–826
- [51] Creighton J A 1988 The selection rules for surface-enhanced Raman spectroscopy *Spectroscopy of Surfaces* ed R J H Clark and R E Hester (Chichester: Wiley) pp 37–89
- [52] Kambhampati P, Child C M, Foster M C and Champion A 1998 On the chemical mechanism of surface enhanced Raman scattering: experiment and theory *J. Chem. Phys.* **108** 5013–26
- [53] Pemberton J E, Bryant M A, Sobocinski R L and Joa S L 1992 A simple method for determination of orientation of adsorbed organics of low symmetry using surface-enhanced Raman scattering *J. Phys. Chem.* **96** 3776–82

- [54] Stair P C and Li C 1997 Ultraviolet Raman spectroscopy of catalysts and other solids *J. Vac. Sci. Technol. A* **15** 1679–84
- [55] Shen Y R 1989 Optical second harmonic generation at interfaces *Ann. Rev. Phys. Chem.* **40** 327–50
- [56] Eisenthal K B 1996 Liquid interfaces probed by second-harmonic and sum-frequency spectroscopy *Chem. Rev.* **96** 1343–60
- [57] Wang H, Borguet E, Yan E C Y, Zhang D, Gutow J and Eisenthal K B 1998 Molecules at liquid and solid surfaces *Langmuir* **14** 1472–7
- [58] Caruso F, Jory M J, Bradberry G W, Sambles J R and Furlong D N 1998 Acousto-optic surface-plasmon resonance measurements of thin films on gold *J. Appl. Phys.* **83** 1023–8
- [59] Jordan C E and Corn R M 1997 Surface plasmon resonance imaging measurements of electrostatic biopolymer adsorption onto chemically modified gold surfaces *Anal. Chem.* **69** 1449–56
- [60] Harris C B, Ge N-H, Lingle R L Jr, McNeill J D and Wong C M 1997 Femtosecond dynamics of electrons on surfaces and at interfaces *Ann. Rev. Phys. Chem.* **48** 711–44
- [61] Heilweil E J, Casassa M P, Cavanagh R R and Stephenson J C 1989 Picosecond vibrational energy transfer studies of surface adsorbates *Ann. Rev. Phys. Chem.* **40** 143–71
- [62] Peiser H S 1960 *X-Ray Diffraction by Polycrystalline Materials* (London: Chapman and Hall)
- [63] Cohen J B and Schwartz L H 1977 *Diffraction from Materials* (New York: Academic)
-

-22-

- [64] Brunel M 1995 Glancing angle x-ray diffraction *Encyclopedia of Analytical Science* vol 8, ed A Townshend (London: Academic) **8** 4922–30
- [65] Kongingsberger D C and Prins R (ed) 1988 *X-Ray Absorption: Principles, Applications, Techniques of EXAFS, SEXAFS and XANES* (New York: Wiley)
- [66] Iwasawa Y (ed) 1996 *X-Ray Absorption Fine Structure for Catalysis and Surfaces* (Singapore: World Scientific)
- [67] Stöhr J, Kollin E B, Fischer D A, Hastings J B, Zaera F and Sette F 1985 Surface extended x-ray-absorption fine structure of low-Z adsorbates studied with fluorescence detection *Phys. Rev. Lett.* **55** 1468–71
- [68] Zaera F, Fischer D A, Shen S and Gland J L 1998 Fluorescence yield near-edge X-ray absorption spectroscopy under atmospheric conditions: CO and H₂ coadsorption on Ni(100) at pressures between 10⁻⁹ and 0.1 Torr *Surf. Sci.* **194** 205–16
- [69] Stöhr J 1992 *NEXAFS Spectroscopy* (Berlin: Springer)
- [70] Stöhr J and Jaeger R 1982 Adsorption-edge resonances, core-hole screening and orientation of chemisorbed molecules: CO, NO and N₂ on Ni(100) *Phys. Rev. B* **26** 4111–31
- [71] Zaera F, Fischer D A, Carr R G and Gland J L 1988 Determination of chemisorption geometries for complex molecules by using near-edge X-ray absorption fine structure: ethylene on Ni(100) *J. Chem. Phys.* **89** 5335–41
- [72] Sharma A and Khatri R K 1995 Surface analysis: optical spectroscopy *Encyclopedia of Analytical Science* ed A Townshend (London: Academic) **8** 4958–65
- [73] Shanthi E, Dutta V, Banerjee A and Chopra K L 1980 Electrical and optical properties of undoped and antimony-doped tin oxide films *J. Appl. Phys.* **51** 6243–51
- [74] Wruck D A and Rubin M 1993 Structure and electronic properties of electrochromic NiO films *J. Electrochem. Soc.* **140** 1097–104

- [75] Chiarotti G 1994 Electronic surface states investigated by optical spectroscopy *Surf. Sci.* **299/300** 541–50
- [76] McGlip J F 1990 Epioptics: linear and non-linear optical spectroscopy of surfaces and interfaces *J. Phys.: Condens Matter* **2** 7985–8006
- [77] Aspens D E and Dietz N 1998 Optical approaches for controlling epitaxial growth *Appl. Surf. Sci.* **130–132** 367–76
- [78] Frederick B G, Power J R, Cole R J, Perry C C, Chen Q, Haq S, Bertrams T, Richardson N V and Weightman P 1998 Adsorbate azimuthal orientation from reflectance anisotropy spectroscopy *Phys. Rev. Lett.* **80** 4490–3
- [79] Tompkins H G 1993 *A User's Guide to Ellipsometry* (Boston: Academic)
- [80] Emedocles S A, Norris D J and Bawendi M G 1996 Photoluminescence spectroscopy of single CdSe nanocrystallite quantum dots *Phys. Rev. Lett.* **77** 3873–6
- [81] Moerner W E 1996 High-resolution optical spectroscopy of single molecules in solids *Acc. Chem. Res.* **29** 563–71
- [82] Xie X S 1996 Single-molecule spectroscopy and dynamics at room temperature *Acc. Chem. Res.* **29** 598–606
- [83] Haw J F 1992 Nuclear magnetic resonance spectroscopy *Anal. Chem.* **64** R243–54
- [84] Engelhardt G and Günter 1987 *High-Resolution Solid-State NMR of Silicates and Zeolites* (Chichester: Wiley)

-23-

- [85] Chmelka B F, Pearson J G, Liu S B, Ryoo R, de Menorval L C and Pines A 1991 NMR study of the distribution of aromatic molecules in NaY zeolite *J. Phys. Chem.* **95** 303–10
- [86] Wang P-K, Slichter C P and Sinfelt J H 1984 NMR study of the structure of simple molecules adsorbed on metal surfaces: C₂H₂ on Pt *Phys. Rev. Lett.* **53** 82–5
- [87] Brey W S 1983 Applications of magnetic resonance in catalytic research *Heterogeneous Catalysis: Selected American Stories* ed B H Davis and W P Hettinger Jr (Washington: American Chemical Society)
- [88] Deviney M L and Gland J L (eds) 1985 *Catalyst Characterization Science: Surface and Solid State Chemistry* (Washington, DC: American Chemical Society)
- [89] Briggs D (ed) 1978 *Handbook of X-ray and Ultraviolet Photoelectron Spectroscopy* (London: Heyden)
- [90] Ertl G and Küppers J 1985 *Low Energy Electrons and Surface Chemistry* (Weinheim: VCH)
- [91] Woodruff D P and Delchar T A 1988 *Modern Techniques of Surface Science* (Cambridge: Cambridge University Press)
- [92] Avouris P, Bozso F and Walkup R E 1987 Desorption via electronic transitions: fundamental mechanisms and applications *Nucl. Instrum. Methods Phys. Res. B* **27** 136–46
- [93] Czanderna A W and Hercules D M (ed) 1991 *Ion Spectroscopies for Surface Analysis* (New York: Plenum)
- [94] Zangwill A 1988 *Physics at Surfaces* (Cambridge: Cambridge University Press)

-1-

B1.23 Surface structural determination: particle scattering methods

J Wayne Rabalais

B1.23.1 INTRODUCTION

The origin of scattering experiments has its roots in the development of modern atomic theory at the beginning of this century. As a result of both the Rutherford experiment on the scattering of alpha particles (He nuclei) by thin metallic foils and the Bohr theory of atomic structure, a consistent model of the atom as a small massive nucleus surrounded by a large swarm of light electrons was confirmed. Following these developments, it was realized that the inverse process, namely, analysis of the scattering pattern of ions from crystals, could provide information on composition and structure. This analysis is straightforward because the kinematics of energetic atomic collisions is accurately described by classical mechanics. Such scattering occurs as a result of the mutual Coulomb repulsion between the colliding atomic cores, that is, the nucleus plus core electrons. The scattered primary atom loses some of its energy to the target atom. The latter, in turn, recoils into a forward direction. The final energies of the scattered and recoiled atoms and the directions of their trajectories are determined by the masses of the pair of atoms involved and the closeness of the collision. By analysis of these final energies and angular distributions of the scattered and recoiled atoms, the elemental composition and structure of the surface can be deciphered.

Low-energy (1–10 keV) ion scattering spectrometry (ISS) had its beginning as a modern surface analysis technique with the 1967 work of Smith [1], which demonstrated both surface elemental and structural analysis. Over the next twenty years, it was clearly demonstrated [2, 3, 4, 5 and 6] that direct surface structural information could be obtained from ISS. Most of the early workers used electrostatic analysers to measure the kinetic energies of the scattered ions. There are two problems with this technique. (i) It analyses only the scattered ions; these are typically only a very small fraction (<5%) of the total scattered flux. Thus, high primary ion doses are required for spectral acquisition which are potentially damaging to the surface and adsorbate structures. (ii) Neutralization probabilities are a function of the ion beam incidence angle α with the surface and the azimuthal angle δ along which the ion beam is directed. This is not a simple behaviour since the probabilities depend on the distances of the ion to specific atoms. As a result, it is difficult to separate scattering intensity changes due to neutralization effects from those due to structural effects. The use of alkali primary ions [7] which have low neutralization probabilities leads to higher scattering intensities and pronounced focusing effects, however, the contamination of the sample surface by the reactive alkali ions is a potential problem with this method. Buck and co-workers [8], who had been developing time-of-flight (TOF) methods for ion scattering since the mid 1970s, used TOF methods for surface structure analysis in 1984 and demonstrated the capabilities and high sensitivity of the technique when both neutrals and ions are detected simultaneously. A TOF spectrometer system with a long flight path for separation of the scattered and recoiled particles and continuous variation of the scattering θ and recoiling ϕ angles was developed in 1990 [9]. This coupling of TOF methods with detection of both scattered and recoiled particles led to the development of TOF scattering and recoiling spectrometry (TOF-SARS) as a tool for structural analysis [10]. A large, time-gated, position-sensitive microchannel plate detector was used in 1997 to obtain images of the scattered and recoiled particles, leading to the development of scattering and recoiling imaging spectrometry (SARIS) [11]. Several research groups [12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23 and 24] throughout the world are now engaged in surface structure determinations using some form of low-keV ISS.

This section will concentrate on TOF-SARS and SARIS, for it is felt that these are the techniques that will be most important in future applications. Also, emphasis will be placed on surface structure determinations rather than surface elemental analysis, because TOF-SARS and SARIS are capable of making unique contributions in the area of structure determination. The use of TOF methods and large-area position-sensitive detectors has led to a surface crystallography that is sensitive to all elements, including the ability to directly detect hydrogen adsorption sites. TOF detection of both neutrals and ions provides the high sensitivity necessary for non-destructive analysis. Detection of atoms scattered and recoiled from surfaces in simple collision sequences, together with calculations of shadowing and blocking cones, can now be used to make direct

measurements of interatomic spacings and adsorption sites within an accuracy of less than 0.1 Å.

B1.23.2 BASIC PHYSICS UNDERLYING KEV ION SCATTERING AND RECOILING

There are two basic physical phenomena which govern atomic collisions in the keV range. First, repulsive interatomic interactions, described by the laws of classical mechanics, control the scattering and recoiling trajectories. Second, electronic transition probabilities, described by the laws of quantum mechanics, control the ion–surface charge exchange process.

B1.23.2.1 KINEMATICS OF ION–SURFACE COLLISIONS AND ELEMENTAL ANALYSIS

The dynamics of ion surface scattering at energies exceeding several hundred electronvolts can be described by a series of binary collision approximations (BCAs) in which only the interaction of one energetic particle with a solid atom is considered at a time [25]. This model is reasonable because the interaction time for the collision is short compared with the period of phonon frequencies in solids, and the interaction distance is shorter than the interatomic distances in solids. The BCA simplifies the many-body interactions between a projectile and solid atoms to a series of two-body collisions of the projectile and individual solid atoms. This can be described with results from the well known two-body central force problem [26].

Within the BCA, the trajectories of energetic particles on the surface become a series of linear motion segments between neighbouring atoms. Both the scattered and recoiled atoms have high, discrete kinetic energy distributions. The simplest case of ion–surface scattering phenomena is quasi-single scattering (QSS), which represents the case of one large-angle deflection that is preceded and/or followed by a few small deflections. [Figure B1.23.1](#) shows an example of QSS. This typically produces a sharp scattering peak whose energy is near that of the theoretical single-collision energy. The energies of scattered and recoiled particles in single scattering (SS) can be derived from the laws of conservation of energy and momentum. The energy E_s of a projectile scattered from a stationary target is given as

-3-

$$E_s = E_0 \{ (\cos \theta \pm (A^2 - \sin^2 \theta)^{1/2})^2 / (1 + A)^2 \} \quad (\text{B1.23.1})$$

where $A = M_t/M_p$, and E_0 , M_t , and M_p are the initial energy of the projectile and the mass of the target and the projectile, respectively. If the mass of the impinging particle is less than or equal to that of the target atom then $A \geq 1$, and only the positive sign is used. If the scattering angle is chosen as 90° , equation (B1.23.1) can be simplified to

$$\frac{E_s}{E_0} = \frac{A - 1}{A + 1} \quad (\text{B1.23.2})$$

Solving for M_t yields

$$M_t = M_p \frac{1 + B}{1 - B} \quad (\text{B1.23.3})$$

where $B = E_s/E_0$. If the mass of the impinging particle is greater than that of the target atom, $M_p > M_t$, and both signs are used in the equation. The energy of the scattered particle is then found to be a double-valued function of the scattering angle θ , i.e. there are two E_s for each θ . For the case of $A < 1$, the maximum SS scattering angle is

$$\theta_{\max} = \sin^{-1} A. \quad (\text{B1.23.4})$$

For angles greater than θ_{\max} , only multiple scattering can occur.

The energy of scattered or recoiled ions can be measured directly by means of an electrostatic energy analyser. If the TOF method is used, the relation between scattering energy E_s and TOF t_s is expressed as

$$E_s = \frac{1}{2} M_s v_s^2 = \frac{1}{2} M_s \left(\frac{d_{\text{tof}}}{t_s} \right)^2 \quad (\text{B1.23.5})$$

and

$$t_s = d_{\text{tof}} \sqrt{\frac{M_s}{2E_s}} \quad (\text{B1.23.6})$$

where d_{tof} is the flight distance of the scattered atom.

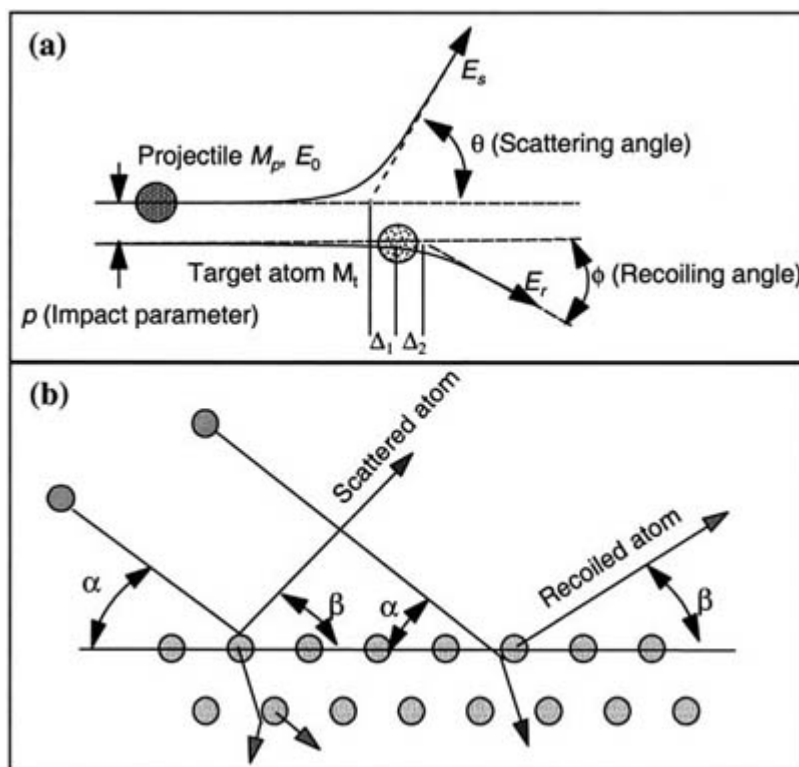


Figure B1.23.1. (a) Two-body collision of a projectile of mass M_p and kinetic energy E_0 approaching a stationary target of mass M_t with an impact parameter p . (b) Quasi-single scattering and direct recoiling with

incident angle α and exit angle β , based on the BCA.

The intensity of SS I_i from an element i in the solid angle $\Delta\Omega$ is proportional to the initial beam intensity I_0 , the concentration of the scattering element N_i , the neutralization probability P_i , the differential scattering cross section $d\sigma(\theta)/d\Omega$, the shadowing coefficient $f_{si}(\alpha, \delta_{in})$ and the blocking coefficient $f_{bi}(\alpha, \delta_{out})$ for the i th component on the surface:

$$I_i \approx I_0 N_i P_i f_{si} f_{bi} \frac{d\sigma}{d\Omega} \Delta\Omega. \quad (\text{B1.23.7})$$

Similar to QSS, direct recoil (DR) of surface atoms produces energetic atoms that have a relatively narrow velocity distribution. DR particles are those species which are recoiled from the surface layers as a result of a direct collision of the primary ion. They escape from the surface with little energy loss through collisions with neighbouring atoms. The energy E_r of a DR surface atom can be expressed as

$$E_r = E_0 \{4A \cos^2 \phi\} / (1 + A)^2. \quad (\text{B1.23.8})$$

-5-

From geometry considerations, DR is observed only in the forward-scattering direction for which $\phi < 90^\circ$. A similar expression as [equation \(B1.23.7\)](#) is applicable for recoiling intensity evaluation. All elements, including hydrogen, can be analysed by either scattering, recoiling, or both techniques. TOF peak identification of QSS and DR is straightforward using the equations above.

B1.23.2.2 SHADOW CONES, BLOCKING CONES, AND STRUCTURAL ANALYSIS

A simple interpretation based on the BCA yields some important concepts for ion scattering and recoiling: shadow cones and blocking cones. As shown in [figure B1.23.2\(a\)](#) scattering of ions by a target atom produces a region in which no ion can penetrate behind the target atom. This region is called a shadow cone. The cone dimensions can be evaluated from known interatomic distances in experiments. [figure B1.23.2\(b\)](#) shows the normalized ion flux density across the shadow cone. There is zero flux density inside of the shadow cone and unit flux density far outside of the cone. Highly focused ion flux density appears at the boundary of the cone. This anisotropic distribution of ion flux after interaction with a target atom is the basis of ISS structural determinations. If a neighbouring atom lies inside the shadow cone (A in (a)), it cannot be scattered or recoiled. If it is well away from the cone (C), the cone has no effect on the intensity. When it lies in the focusing region (B), enhanced intensity scattering and/or recoiling intensity is observed. TOF-SARS measures the intensity change due to the shadow cone effect on neighbouring atoms as a function of incident beam direction.

Trajectories of ions interacting with an additional target atom (blocking atom) after scattering from the initial target atom (scattering atom) produce a hollow region behind the blocking atom called a blocking cone ([figure B1.23.2\(c\)](#)). It can be regarded as an interaction of a target atom (blocking atom) with ions emitted from an adjacent point ion source (scattering atom). A shadow cone is different because it originates from the interaction with ions from an infinitely distant point ion source (ion source in the ion beam line). Unlike a shadow cone, a blocking cone diverges with a measurable blocking cone angle ξ . The closer the blocking atom is to the scattering atom, the larger is the angle of divergence. In traditional ISS, the variation of the interactions of both shadowing and blocking cones are measured. It is possible to minimize the effect of blocking cones by a judicious choice of scattering geometry. Blocking cones are, however, inevitable, especially for scattering trajectories from deep subsurface layers.

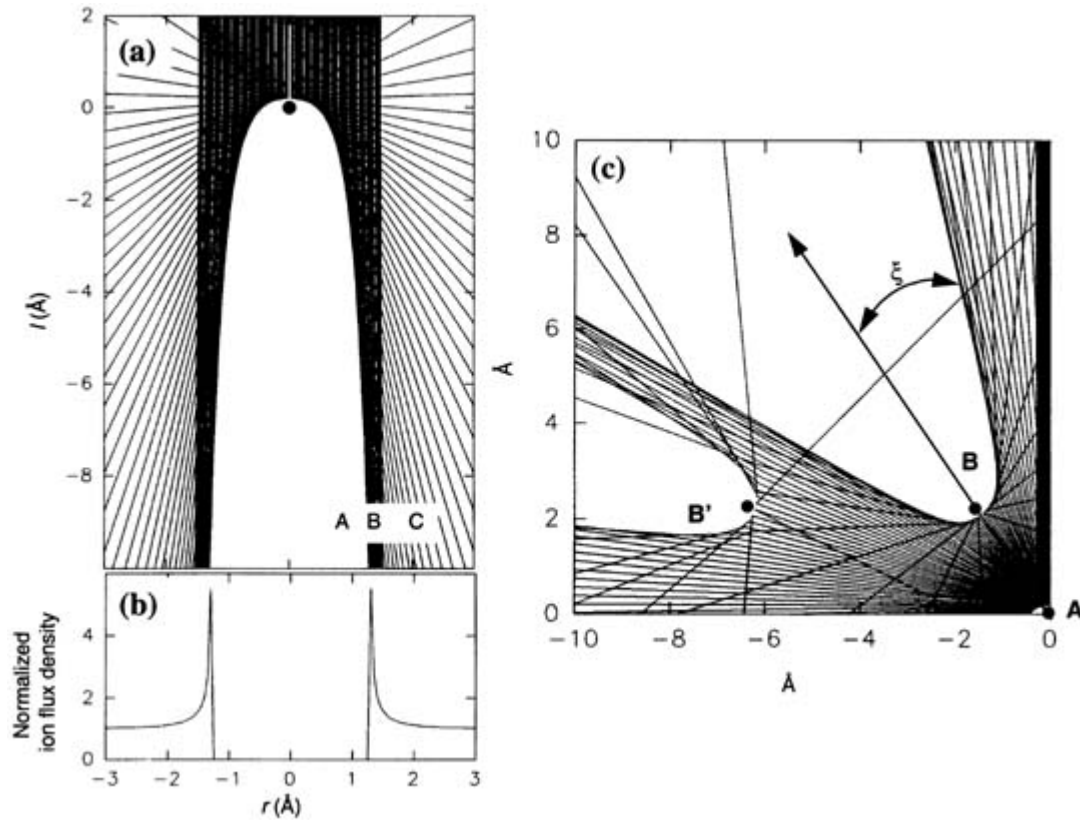


Figure B1.23.2. (a) Shadow cone of a stationary Pt atom in a 4 keV Ne^+ ion beam, appearing with the overlapping of ion trajectories as a function of the impact parameter. The initial position of the target atom that recoils in the collision is indicated by a solid circle. (b) Plot of the normalized ion flux distribution density across the shadow cone in (a). The flux density changes from 0 inside the shadow cone, to much greater than ~ 1 in the focusing region, converging to 1 away from the shadow cone edge. (c) Blocking cones of Pt atoms (B, B') cast by 4 keV Ne^+ ions scattered from another Pt atom (A). Note the different blocking angles of the two blocking atoms, which is due to the differences in the interatomic spacings between the scattering and blocking atoms.

Considering a large number of ions with parallel trajectories impinging on a target atom, the ion trajectories are bent by the repulsive potential such that there is an excluded volume, called the shadow cone, in the shape of a paraboloid formed behind the target atom as shown in [figure B1.23.3\(a\)](#). Ion trajectories do not penetrate into the shadow cone, but instead are concentrated at its edges much as rain pours off an umbrella. Atoms located inside the cone behind the target atom are shielded from the impinging ions. Similarly, if the scattered ion or recoiling atom trajectory is directed towards a neighbouring atom, that trajectory will be blocked. For a large number of scattering or recoiling trajectories, a blocking cone will be formed behind the neighbouring atom into which no particles can penetrate, as shown in [figure B1.23.3\(b\)](#). The dimensions of the shadowing and blocking cones can be determined experimentally from scattering measurements along crystal azimuths for which the interatomic spacings are accurately known.

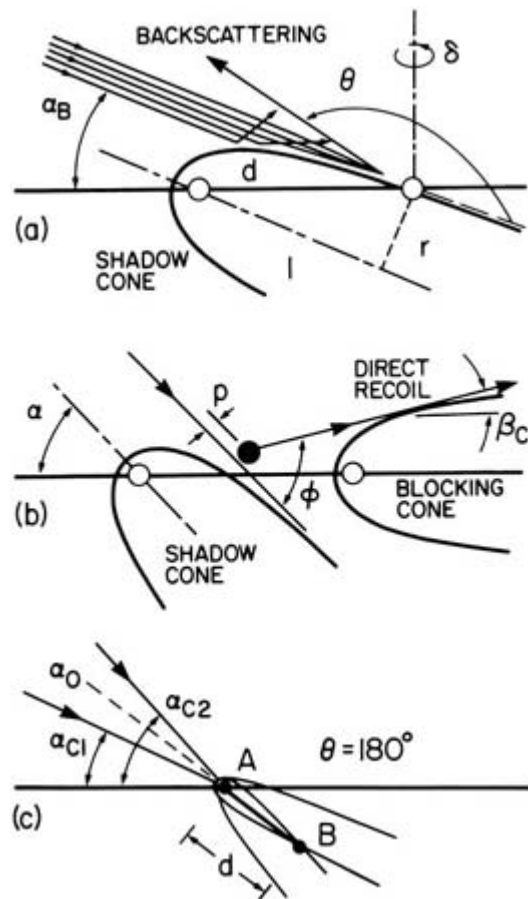


Figure B1.23.3. Schematic illustrations of backscattering with shadowing and direct recoiling with shadowing and blocking.

B1.23.2.3 SCATTERING AND RECOILING ANISOTROPY CAUSED BY SHADOWING AND BLOCKING CONES

When an isotropic ion fluence impinges on a crystal surface at a specific incident angle α , the scattered and recoiled atom flux is anisotropic. This anisotropy is a result of the incoming ion's eye view of the surface, which depends on the specific arrangement of atoms and the shadowing and blocking cones. The arrangement of atoms controls the atomic density along the azimuths and the ability of ions to channel, that is, to penetrate into empty spaces between atomic rows. The cones determine which nuclei are screened from the impinging ion flux and which exit trajectories are blocked, as depicted in figure B1.23.3. By measuring the ion and atom flux at specific scattering and recoiling angles as a function of ion beam incident α and azimuthal δ angles to the surface, structures are observed which can be interpreted in terms of the interatomic spacings and shadow cones from the ion's eye view.

(A) TIME OF FLIGHT SCATTERING AND RECOILING SPECTROMETRY (TOF-SARS)—SHADOW CONE BASED EXPERIMENT

In TOF-SARS [9], a low-keV, monoenergetic, mass-selected, pulsed noble gas ion beam is focused onto a sample surface. The velocity distributions of scattered and recoiled particles are measured by standard TOF methods. A channel electron multiplier is used to detect fast (>800 eV) neutrals and ions. This type of detector has a small acceptance solid angle. A fixed angle is used between the pulsed ion beam and detector directions with respect to the sample as shown in figure B1.23.4. The sample has to be rotated to measure ion scattering

and recoiling anisotropy as a function of the incident angle (α) and azimuthal angle (δ) of the incident beam. Since the sample rotation changes both the incident beam direction and the detector direction, the spectra are affected by both shadow cones and blocking cones. In order to reduce blocking cone effects for simpler interpretations, high-angle scattering is preferred. Elemental analyses are achieved by converting the velocity distributions into energy distributions and relating those to the masses of the target atoms through the kinematic relationship that describes classical scattering and recoiling (equation (B1.23.1) and equation (B1.23.8)). Structural analyses are achieved by monitoring the scattered and recoiled particles as a function of both beam incident angle α and crystal azimuthal angle δ . The anisotropic features in these α - and δ -scans are interpreted by means of shadow cone and blocking cone analyses. It requires several hours to collect data needed to construct a contour map of intensities as a function of both α and δ . Moreover, the experimental geometry is restricted to fixed scattering angles and in-plane scattering and recoiling trajectories.

(B) SCATTERING AND RECOILING IMAGING SPECTROMETRY (SARIS)—BLOCKING CONE BASED EXPERIMENT

In an ideal SARIS system [11], it would be desirable to measure the velocity distributions of all the energetic particles scattered and recoiled from a sample surface in a short time period. This concept requires a hemispherical, time-resolving, position-sensitive detector which covers all of the solid angle space above the sample surface. Implementation of this concept is not currently feasible. If data collection time is unimportant, a point detector as described above, such as a channel electron multiplier mounted on a flexible goniometer which allows movement of the detector over a large solid angle, can be used to collect TOF spectra. In order to compromise the size of detector with data collection time, a hybrid configuration can be used. This is a large, time-resolving, position-sensitive microchannel plate detector mounted on a triple-axis UHV goniometer. This instrument makes it possible to capture scattering and recoiling intensity distributions without changing the incident beam direction. It gives a great advantage in comparison of experimental results with those of computer simulations. The large-area detector provides the intensity distribution over a limited solid angle. Since the optimum detector position and flight distance for a specific experiment are variable, the detector has to be moved around to cover a large solid angle, and to compromise TOF resolution with the detector acceptance solid angle.

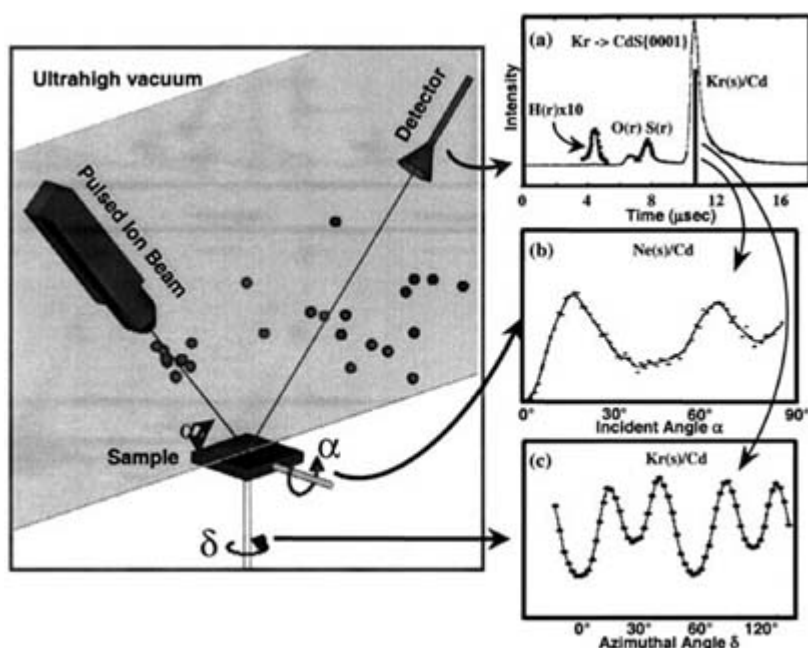


Figure B1.23.4. Schematic diagram of TOF scattering and recoiling spectrometry (TOF-SARS) illustrating the plane of scattering formed by the ion beam, sample and detector. TOF spectra (a) are collected with fixed

positions of the ion beam, sample and detector. In order to measure the incident angle (α) or azimuthal angle (δ) intensity variation of a peak in a TOF spectrum, the sample is rotated about an axis that goes through its normal or through its plane, respectively. Such α and δ intensity variations of a peak are shown in (b) and (c).

B1.23.3 INSTRUMENTATION

The basic requirements [9] for low-energy ion scattering are an ion source, a sample mounted on a precision manipulator, an energy or velocity analyser and a detector as shown in [figure B1.23.5](#). The sample is housed in an ultra-high vacuum (UHV) chamber in order to prepare and maintain well defined clean surfaces. The UHV prerequisite necessitates the use of differentially pumped ion sources. Ion scattering is typically done in a UHV chamber which houses other surface analysis techniques such as low-energy electron diffraction (LEED), x-ray photoelectron spectroscopy (XPS) and Auger electron spectroscopy (AES). The design of an instrument for ion scattering is based on the type of analyser to be used. An electrostatic analyser (ESA) measures the kinetic energies of ions while a TOF analyser measures the velocities of both ions and fast neutrals.

-10-

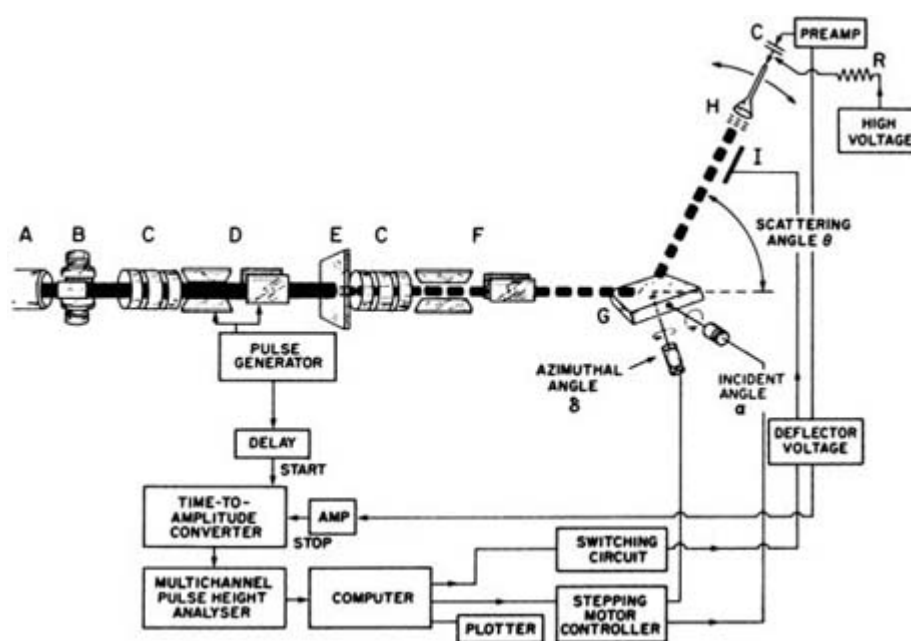


Figure B1.23.5. Schematic illustration of the TOF-SARS spectrometer system. A = ion gun, B = Wien filter, C = Einzel lens, D = pulsing plates, E = pulsing aperture, F = deflector plates, G = sample, H = electron multiplier detector with energy prefilter grid and I = electrostatic deflector.

B1.23.3.1 ION SOURCE AND BEAM LINE

The critical requirements for the ion source are that the ions have a small energy spread, there are no fast neutrals in the beam and the available energy is 1–10 keV. Both noble gas and alkali ion sources are common. For TOF experiments, it is necessary to pulse the ion beam by deflecting it past an aperture. A beam line for such experiments is shown in [figure B1.23.5](#) it is capable of producing ion pulse widths of ≈ 15 ns.

B1.23.3.2 ANALYSERS

An ESA provides energy analysis of the ions with high resolution. A TOF analyser provides velocity analysis

of both fast neutrals and ions with moderate resolution. In an ESA the energy separation is made by spatial dispersion of the charged particle trajectories in a known electrical field. ESAs were the first analysers used for ISS; their advantage is high-energy resolution and their disadvantages are that they analyse only ions and have poor collection efficiency due to the necessity for scanning the analyser. A TOF analyser is simply a long field-free drift region. It has the advantage of high efficiency since it collects both ions and fast neutrals simultaneously in a multichannel mode; its disadvantage is only moderate resolution.

B1.23.3.3 DETECTORS

The most common detectors used for TOF-SARS are continuous dynode channel electron multipliers which are capable of multiplying the signal pulses by 10^6 – 10^7 . They are sensitive to both ions and fast neutrals. Neutrals with

-11-

velocities $\gtrsim 10^6$ cm s⁻¹ are detected with the same efficiency as ions. Since the cones of these detectors are usually less than 1 cm² and the TOF flight paths are of the order of ~1 m, the acceptance solid angles of such detectors are very small. Incident and azimuthal angle scans are made by rotating the sample with the detector at a fixed position.

SARIS overcomes the limitations of small-area detectors by using a large, time-resolving, position-sensitive microchannel plate (MCP) detector and TOF methods to capture images of both ions and fast neutrals that are scattered and recoiled from a surface. Due to the large solid angle subtended by the MCP, atoms that are scattered and recoiled in both planar and non-planar directions are detected simultaneously. For example, with a 75 × 95 mm MCP situated at a distance of 16 cm from the sample, it spans a solid angle of ~0.3 sr corresponding to an azimuthal range of ~26°. Using a beam current of ~0.1 nA cm⁻², the four images required to make up a 90° azimuthal range can be collected in ~2 min with a total ion dose of ~10¹¹ ions cm⁻². The time gating of the MCP provides resolution of the scattered and recoiled atoms into time frames as short as 10 ns, thereby providing element-specific spatial-distribution images. These SARIS images contain features that are sharply focused into well defined patterns as a function of both space and time by the crystal structure of the target sample. If the MCP is mounted on a goniometer that provides both horizontal and vertical rotation and translation away from the sample, it is possible to change the solid angle of collection and the flight path length.

B1.23.4 COMPUTER SIMULATION METHODS

It is extremely helpful to use classical ion trajectory simulations in order to visualize the ion trajectories to improve the understanding of ion behaviour in the surface region and to provide a systematic method for surface structure determination. Such simulations are based on the BCA in which the trajectories of the energetic particles are assumed to be a series of straight lines corresponding to the asymptotes of the scattering trajectories due to sequential binary collisions. The BCA has been proven to be valid for the keV range of energies.

B1.23.4.1 BINARY COLLISION APPROXIMATION

Atom–surface interactions are intrinsically many-body problems which are known to have no analytical solutions. Due to the shorter de Broglie wavelength of an energetic ion than solid interatomic spacings, the energetic atom–surface interaction problem can be treated by classical mechanics. In the classical mechanical

framework, the problem becomes a set of Newtonian equations of motion [26] for i th particle in an N -body problem.

$$\mathbf{F}_i = m_i \frac{d^2 \mathbf{r}_i}{dt^2} = -\nabla \Phi(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N). \quad (\text{B1.23.9})$$

The summation of pair-wise potentials is a good approximation for molecular dynamics calculations for simple classical many-body problems [27]. It has been widely used to simulate hyperthermal energy (>1 eV) atom–surface scattering:

$$\mathbf{F}_i \approx - \sum_j^{N, i \neq j} \nabla \Phi(\mathbf{r}_i, \mathbf{r}_j). \quad (\text{B1.23.10})$$

-12-

As the kinetic energy involved in the system goes higher, the interaction of energetic particles is more and more localized near the nuclei. When the interaction distance is much smaller than interatomic distances in the system, the BCA is valid:

$$\mathbf{F}_i \approx -\nabla \Phi(\mathbf{r}_i, \mathbf{r}_j) \quad j = \text{the nearest neighbour of } i. \quad (\text{B1.23.11})$$

In the BCA, each collision process is regarded as an isolated event. Ion–solid interactions are approximated by a series of two-body interactions which are reduced to one-body problems in the centre-of-mass (CM) coordinates. The projectile is assumed to converge to the asymptote after a collision before interacting with the next collision partner. In the central force one-body problem, evaluation of scattering integrals and time integrals replaces the time-consuming numerical integration of a set of differential equations.

B1.23.4.2 SCATTERING INTEGRAL AND TIME INTEGRAL

Once atom–surface scattering is reduced to the BCA, one can calculate the energy relationship between two particles involved in scattering with a known scattering angle and the laws of conservation of energy and momentum as shown in section B1.23.2.1. The scattering angle as a function of impact parameter necessitates evaluation of the scattering integrals. The scattering angle, χ in the CM coordinate system, and the CM energy E , are given by

$$\chi = \pi - 2 \int_{r_0}^{\infty} \frac{p \, dr}{r^2} \left(1 - \frac{V(r)}{E} - \frac{p^2}{r^2} \right)^{-1/2} \quad (\text{B1.23.12})$$

where

$$E = \frac{A}{1+A} E_0. \quad (\text{B1.23.13})$$

$A = M_t/M_p$, p is impact parameter and r_0 is the distance of closest approach (apsis) of the collision pair. The transformations from the CM coordinates (scattering angle χ) to the laboratory coordinates with the scattering angle θ for the primary particle and ϕ for the recoiled surface atoms is given by

$$\tan \theta = \frac{A \sin \chi}{1 + A \cos \chi} \quad (\text{B1.23.14})$$

and

$$\tan \phi = \frac{\sin \theta}{1 - \cos \chi}. \quad (\text{B1.23.15})$$

-13-

For accurate ion trajectory calculation in the solid, it is necessary to evaluate the exact positions of the intersections of the asymptotes (Δ_1, Δ_2) of the incoming trajectory and that of the outgoing trajectories of both the scattered and recoiled particles in a collision. The evaluation of these values requires time integrals and the following transformation equations:

$$t = \sqrt{r_0^2 - p^2} - \int_{r_0}^{\infty} \left\{ \left(1 - \frac{V(r)}{E_r} - \frac{p^2}{r^2} \right)^{-1/2} \left(1 - \left(\frac{p^2}{r^2} \right) \right)^{-1/2} \right\} dr \quad (\text{B1.23.16})$$

$$\Delta_1 = \frac{t + (A - 1)p \tan \frac{1}{2}\chi}{1 + A} \quad (\text{B1.23.17})$$

$$\Delta_2 = \frac{p}{\tan \phi} - \Delta_1. \quad (\text{B1.23.18})$$

Numerical integration methods are widely used to solve these integrals. The Gauss–Muhler method [28] is employed in all of the calculations used here. This method is a Gaussian quadrature [29] which gives exact answers for Coulomb scattering.

B1.23.4.3 POTENTIAL FUNCTION

One of the most important issues in simulation of energetic atom–surface scattering is the determination of the interaction potential between the colliding atoms. In the low-keV energy region, electrons have a screening effect on the Coulomb interaction of nuclei so that the actual nuclear charges affecting the trajectories are less than the atomic numbers (Z) of the atoms involved. This screening effect decreases the potential $V(r)$ by an amount which is expressed as a screening function $\Phi(r)$. The form of the potential function is

$$V(r) = \frac{Z_1 Z_2 e^2}{r} \Phi(r). \quad (\text{B1.23.19})$$

The $\Phi(r)$ can be expressed in various forms [30], e.g. the Bohr, Born–Mayer, Thomas–Fermi–Firsov and Moliere models, as well as the ‘universal potential’ of Ziegler, Biersack and Littmark known as the ZBL potential [31]. The ZBL potential function is expressed as

$$\Phi(r) = 0.1818 e^{-3.2/ar} + 0.5099 e^{0.9423/ar} + 0.2802 e^{0.4029/ar} + 0.02817 e^{0.2016/ar} \quad (\text{B1.23.20})$$

where the screening length $a = (0.8853 C_F a_0 / (Z_1^{0.23} + Z_2^{0.23}))$, a_0 is Bohr radius (0.53 Å), Z_1, Z_2 are the atomic numbers of the atoms involved and C_F is a screening constant for adjusting the screening length to calibrate the potential to experimental scattering data. The ZBL potential provides good agreement between simulated

and experimental results.

-14-

B1.23.4.4 GENERAL DESCRIPTION OF SIMULATION PROGRAM

Classical ion trajectory computer simulations based on the BCA are a series of evaluations of two-body collisions. The parameters involved in each collision are the type of atoms of the projectile and the target atom, the kinetic energy of the projectile and the impact parameter. The general procedure for implementation of such computer simulations is as follows. All of the parameters involved in the calculation are defined: the surface structure in terms of the types of the constituent atoms, their positions in the surface and their thermal vibration amplitude; the projectile in terms of the type of ion to be used, the incident beam direction and the initial kinetic energy; the detector in terms of the position, size and detection efficiency; the type of potential functions for possible collision pairs.

After defining the input parameters, the calculation of the trajectories of an incident ion begins with a randomly chosen initial entrance point on the surface. The next step is to find the first collision partner. Taking advantage of the symmetry of the crystal structure, one can list the positions of surface atoms within a certain distance from the projectile. The atoms are sorted in ascending order of the scalar product of the interatomic vector from the atom to the projectile with the unit velocity vector of the projectile. If the collision partner has larger impact parameter than a predefined maximum impact parameter (p_{\max}), it is discarded. If a partner has a shorter impact parameter than p_{\max} , the evaluation of the collision is initiated by converting three-dimensional information, such as the positions of the projectile and the target atom and the velocity of the projectile, into the parameters necessary to calculate the expressions for the impact parameter and the relative energy. After the equations are solved with these parameters, the values are converted back to three-dimensional information to search for a new collision partner with a new set of parameters calculated in the previous collision. This procedure is repeated until the kinetic energy of the projectile falls below a predefined cutoff energy or it ejects from the surface. If it is necessary to follow recoiled particles, they are regarded as new projectiles in subsequent collisions. After finishing a trajectory calculation, a new calculation starts with a new randomly chosen entrance point. Millions of trajectory calculations with different initial impact parameters are carried out in order to compare the results with those of experiments. In order to increase the speed of the calculation, a precalculated table of scattering angle and scattered energy as a function of impact parameter and kinetic energy is used. A two-dimensional spline method is used to interpolate a scattering angle and energy from the table.

(A) CRYSTAL MODELS WITH THERMAL VIBRATION INCLUSION

The lattice atoms in the simulation are assumed to vibrate independently of one another. The displacements from the equilibrium positions of the lattice atoms are taken as a Gaussian distribution, such as

$$\rho(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\Delta x^2/2\sigma^2} \quad (\text{B1.23.21})$$

where σ^2 and Δx are the variance and the displacement from the lattice equilibrium position, respectively. The variance of the distribution can be expressed as [32]

$$\sigma^2 = \overline{\Delta x^2} = \frac{2h^2T}{4\pi^2mk\Theta_D^2} \quad (\text{B1.23.22})$$

where Θ_D is the Debye temperature. The Box–Muller method [29] is used for generating random deviates with a Gaussian distribution.

(B) COMPARISON OF SIMULATED AND EXPERIMENTAL DATA

A systematic comparison of two sets of data requires a numerical evaluation of their likeliness. TOF-SARS and SARIS produce one- and two-dimensional data plots, respectively. Comparison of simulated and experimental data is accomplished by calculating a one- or two-dimensional reliability (R) factor [33], respectively, based on the R -factors developed for LEED [34]. The R -factor between the experimental and simulated data is minimized by means of a multiparameter simplex method [33].

B1.23.5 ELEMENTAL ANALYSIS FROM SCATTERING AND RECOILING

TOF-SARS and SARIS are capable of detecting all elements by either scattering, recoiling or both techniques. TOF peak identification is straightforward by converting equation (B1.23.1) and equation (B1.23.8) to the flight times of the scattered t_s and recoiled t_r particles as

$$t_s = L(M_1 + M_2)/(2M_1 E_0)^{1/2} \{ \cos + [(M_2/M_1)^2 - \sin^2 \theta]^{1/2} \} \quad (\text{B1.23.23})$$

and

$$t_r = L(M_1 + M_2)/(8M_1 E_0)^{1/2} \cos \phi \quad (\text{B1.23.24})$$

where L is the flight distance, that is, the distance from target to detector. Collection of neutrals plus ions results in scattering and recoiling intensities that are determined by elemental concentrations, shadowing and blocking effects and classical cross sections. The main advantage of TOF-SARS for surface compositional analyses is its extreme surface sensitivity as compared to the other surface spectrometries, i.e. mainly XPS and AES. Indeed, with a correct orientation and aperture of the shadow cone, the first monolayer can be probed selectively. At selected incident angles, it is possible to delineate signals from specific subsurface layers. Detection of the particles independently of their charge state eliminates ion neutralization effects. Also, the multichannel detection requires primary ion doses of only $\approx 10^{11}$ ions cm^{-2} or $\approx 10^{-4}$ ions/surface atom for spectral acquisition; this ensures true static conditions during analyses.

Examples of typical TOF spectra obtained from 4 keV Ar^+ impinging on a $\text{Si}\{100\}$ surface with chemisorbed H_2O and H_2 are shown in figure B1.23.6 [35]. Peaks due to Ar scattering from Si and recoiled H, O and Si are observed. The intensities necessary for structural analysis are obtained by integrating the areas of fixed time windows under these peaks.

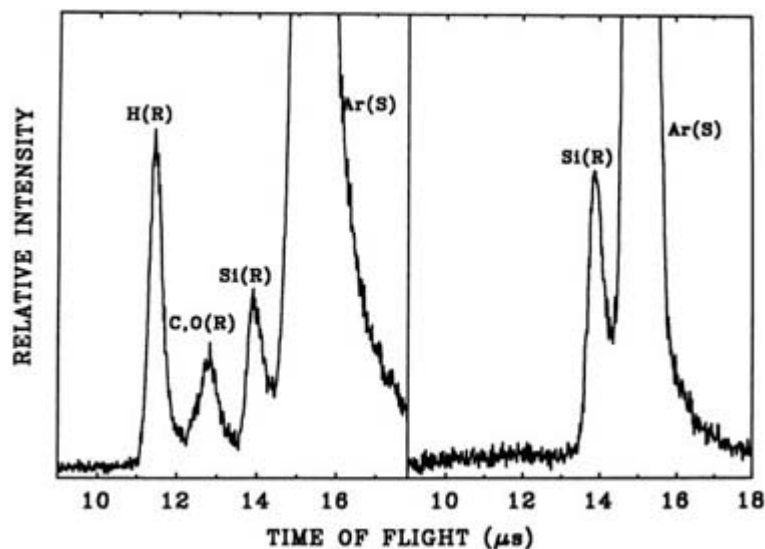


Figure B1.23.6. TOF spectra of a Si{100} surface with chemisorbed H₂O (left) and clean Si (right). Peaks due to scattered Ar and recoiled H, O, and Si are observed. Conditions: 4 keV Ar⁺, scattering angle $\theta = 28^\circ$, incident angle $\alpha = 8^\circ$.

While qualitative identification of scattering and recoiling peaks is straightforward, quantitative analysis requires relating the scattered or recoiled flux to the surface atom concentration. The flux of scattered or recoiled atoms is dependent on several parameters as detailed in [equation \(B1.23.7\)](#). Compositional analyses by TOF-SARS and ISS have been applied in different areas of surface science, mainly in situations where the knowledge of the uppermost surface composition (first monolayer) is crucial. Some of these areas are as follows: gas adsorption, surface segregation, compounds and polymer blends, surface composition of real supported catalysts, surface modifications due to preferential sputtering by ion beams, diffusion, thin film growth and adhesion.

B1.23.6 STRUCTURAL ANALYSIS FROM TOF-SARS

The atomic structure of a surface is usually not a simple termination of the bulk structure. A classification exists based on the relation of surface to bulk structure. A *bulk truncated* surface has a structure identical to that of the bulk. A *relaxed* surface has the symmetry of the bulk structure but different interatomic spacings. With respect to the first and second layers, *lateral relaxation* refers to shifts in layer registry and *vertical relaxation* refers to shifts in layer spacings. A *reconstructed* surface has a symmetry different from that of the bulk symmetry. The methods of structural analysis will be delineated below.

B1.23.6.1 SCATTERING VERSUS INCIDENT ANGLE SCANS

When an ion beam is incident on an atomically flat surface at grazing angles, each surface atom is shadowed by its neighbouring atom such that only forwardscattering (FS) is possible; these are large impact parameter (p) collisions.

As α increases, a critical value $\alpha_{c,sh}^i$ is reached each time the i th layer of target atoms moves out of the shadow cone allowing for large-angle backscattering (BS) or small- p collisions as shown in [figure B1.23.3](#). If the BS intensity I_{RC} is monitored as a function of α , steep rises [36] with well defined maxima are observed when the

focused trajectories at the edge of the shadow cone pass close to the centre of neighbouring atoms. This is illustrated for scattering of Ne^+ from a Pt(110) surface in figure B1.23.7. From the shape of the shadow cone, i.e. the radius (R) as a function of distance (l) behind the target atom (figure B1.23.3)), the interatomic spacing (δ) can be directly determined from the I_{BS} versus α plots. For example, by measuring $\alpha_{c,sh}^1$ along directions for which specific crystal azimuths are aligned with the projectile direction and using $d = r/\sin \alpha_{c,sh}^1$, one can determine interatomic spacings in the first atomic layer. The first–second layer spacing can be obtained in a similar manner from $\alpha_{c,sh}^2$ measured along directions for which the first- and second-layer atoms are aligned, providing a measure of the vertical relaxation in the outermost layers.

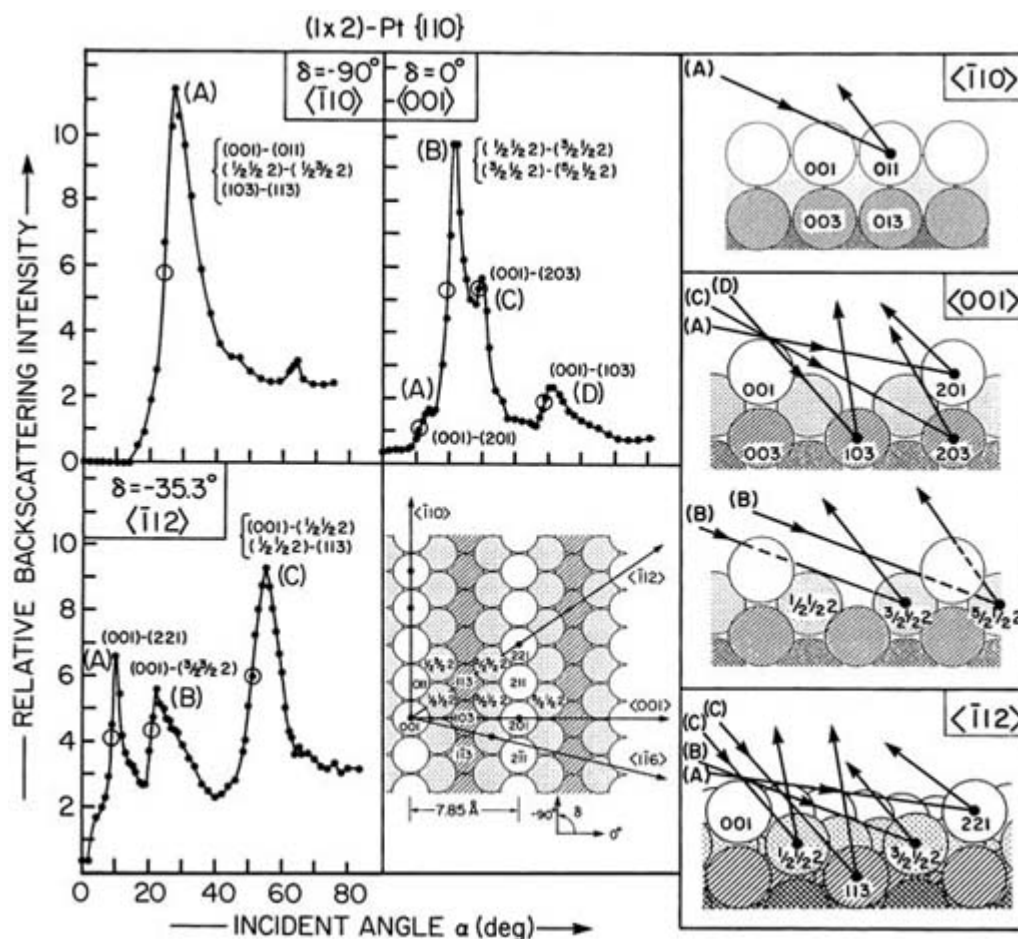


Figure B1.23.7. Scattering intensity versus incident angle α scans for 2 keV Ne^+ incident on $(1 \times 2)\text{-Pt}\{110\}$ at $\theta = 149^\circ$ along the $\langle \bar{1}10 \rangle$, $\langle 001 \rangle$ and $\langle \bar{1}12 \rangle$ azimuths. A top view of the (1×2) missing-row Pt $\{110\}$ surface along with atomic labels is shown. Cross-section diagrams along the three azimuths illustrating scattering trajectories for the peaks observed in the scans are shown on the right.

B1.23.6.2 SCATTERING VERSUS AZIMUTHAL ANGLE Δ SCANS

Fixing the incident beam angle and rotating the crystal about the surface normal while monitoring the backscattering intensity provides a scan of the crystal azimuthal angles δ [37]. Such scans reveal the periodicity of the crystal structure. For example, one can obtain the azimuthal alignment and symmetry of the outermost layer by using a low α value such that scattering occurs from only the first atomic layer. With higher α values, similar information can be obtained for the second atomic layer. Shifts in the first–second layer registry can be detected by carefully monitoring the $\alpha_{c,sh}^2$ values for second-layer scattering along directions near those azimuths for which the second-layer atoms are expected, from the bulk structure, to be

directly aligned with the first-layer atoms. The $\alpha_{c,sh}^2$ values will be maximum for those δ values where the first- and second-layer neighbouring atoms are aligned.

When the scattering angle θ is decreased to a forward angle ($<90^\circ$), both shadowing effects along the incoming trajectory and blocking effects along the outgoing trajectory contribute to the patterns. The blocking effects arise because the exit angle $\beta = \theta - \alpha$ is small at high α values. Surface periodicity can be read directly from these features [37], as shown in figure B1.23.8 for Pt{110}. Minima are observed at the δ positions corresponding to alignment of the beam along specific azimuths. These minima are a result of shadowing and blocking along the close-packed directions, thus providing a direct reading of the surface periodicity.

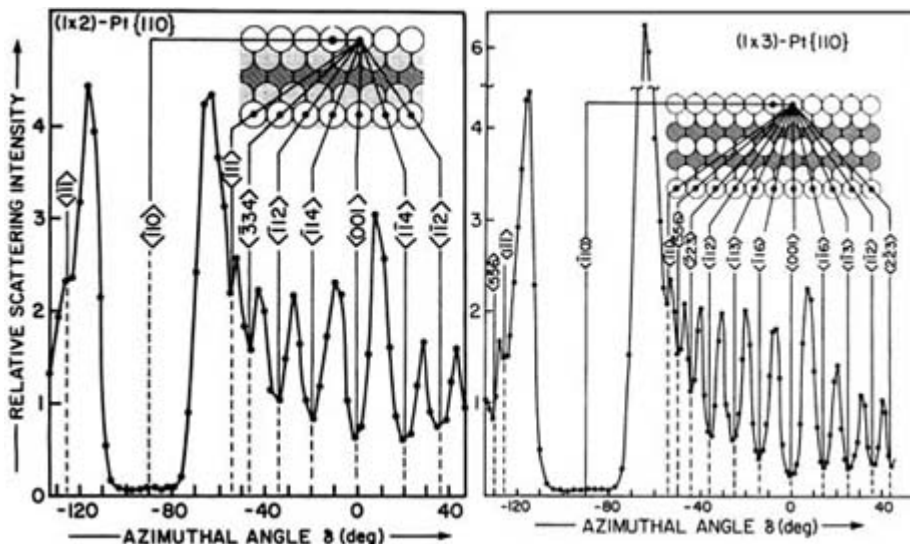


Figure B1.23.8. Scattering intensity of 2 keV Ne^+ versus azimuthal angle δ scans for Pt{110} in the (1×2) and (1×3) reconstructed phases. Scattering angle $\theta = 28^\circ$ and incident angle $\alpha = 6^\circ$.

Azimuthal scans obtained for three surface phases of Ni{110} are shown in figure B1.23.9 [38]. The minima observed for the clean and hydrogen-covered surfaces are due only to Ni atoms shadowing neighbouring Ni atoms, whereas for the oxygen-covered surface minima are observed due to both O and Ni atoms shadowing neighbouring Ni atoms. Shadowing by H atoms is not observed because the maximum deflection in the Ne^+ trajectories caused by H atoms is less than 2.8° .

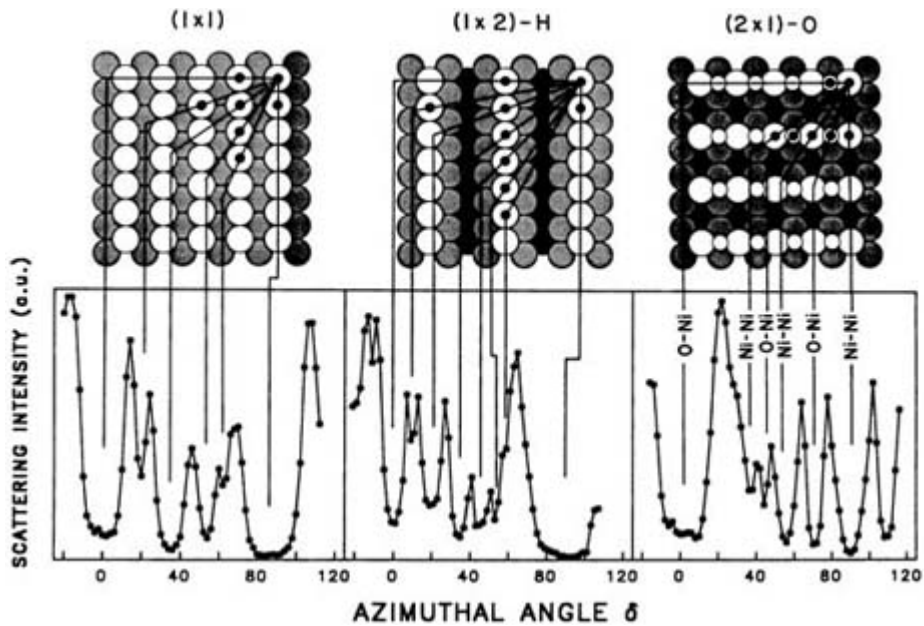


Figure B1.23.9. Scattering intensity of 4 keV Ne^+ versus azimuthal angle δ for a $\text{Ni}\{110\}$ surface in the clean (1×1) , (1×2) -H missing row, and (2×1) -O missing row phases. The hydrogen atoms are not shown. The oxygen atoms are shown as small open circles. O–Ni and Ni–Ni denote the directions along which O and Ni atoms, respectively, shadow the Ni scattering centre.

B1.23.7 STRUCTURAL ANALYSIS FROM SARIS

An example of the SARIS experimental arrangement is shown in [figure B1.23.10](#) [39, 40 and 41]. The velocity distributions of scattered and recoiled ions plus fast neutrals are measured by analysing the positions of the particles on the detector along with their correlated TOF from sample to detector. The detector is gated so that it can be activated in windows of several microseconds duration, which are appropriate for TOF collection of specific scattered or recoiled particles. These windows are divided into 255 time frames with the time duration of 16.7 ns for each frame. Good statistics are obtained in a total acquisition time of ~ 1 min. The image ordinate represents particle exit angles (β) and the abscissa represents the crystal azimuthal angles (δ), i.e. an image in (β, δ) -space.

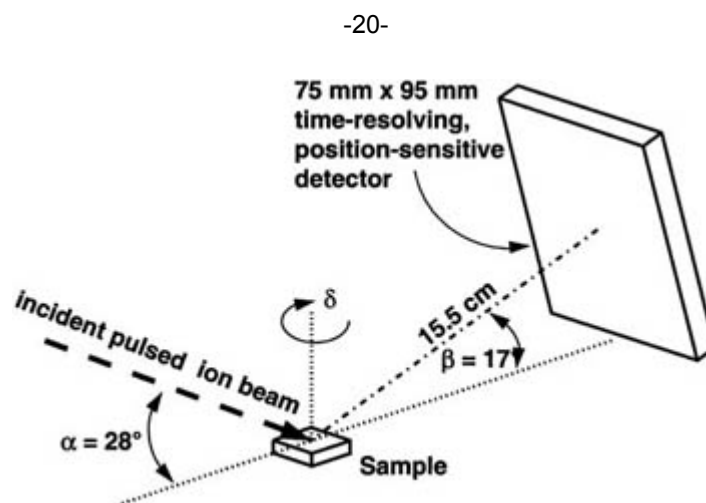


Figure B1.23.10. Schematic diagram of a scattering and recoiling imaging spectrometer (SARIS). A large-area ($95 \times 75 \text{ mm}^2$), time-resolving, position-sensitive microchannel plate (MCP) detector captures a large

solid angle of the scattering and recoiling particles. A triple-axis UHV goniometer moves the MCP inside the vacuum chamber in order to vary the scattering angle, the distance from detector to sample, the TOF resolution and the acceptance solid angle of the detector.

B1.23.7.1 INTERACTION OF 4 KEV AR WITH PT{111}

(A) AR SCATTERING

The time-resolved images of Ar scattering [33] from Pt{111} of figure B1.23.11 correspond to selected frames of scattered Ar atoms with the azimuthal angle of the incident beam aligned along $\langle\bar{1}\bar{1}2\rangle$. The overall scattering intensity is maximal at 1.17 μs (3.54 keV) for the scattered Ar atoms, corresponding to SS as predicted by the BCA. The two intense spots at 1.17 μs result from the scattering from a first-layer Pt atom and focusing of the scattered beam by an ‘atomic lens’ formed by neighbouring first-layer Pt atoms (2, 3, 4 in figure B1.23.11). The intense spots are at small β since most of the Ar atoms are scattered and focused by first-layer Pt atoms. Focused high- β scattering usually arises from subsurface collisions.

-21-

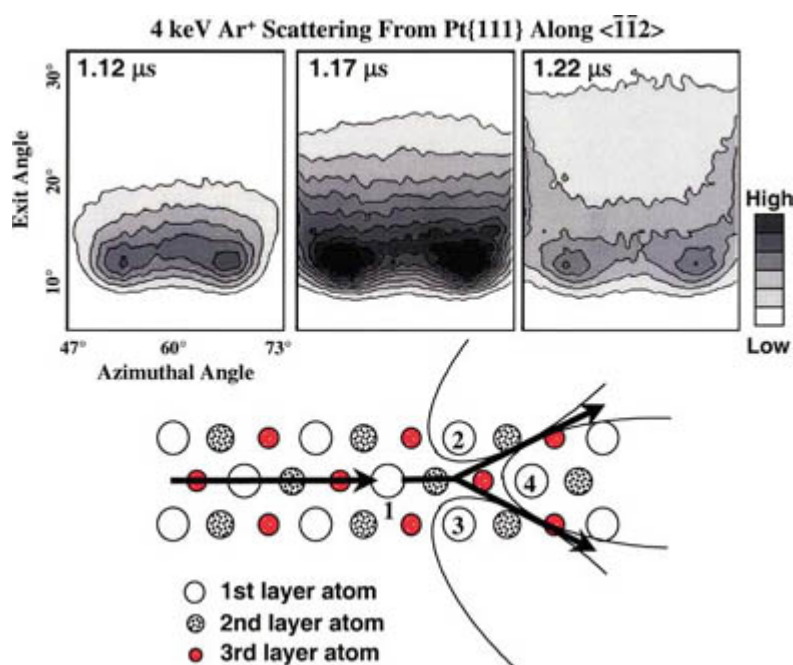


Figure B1.23.11. Above: selected time-resolved SARIS images of 4 keV Ar^+ scattering from Pt{111} along $\langle\bar{1}\bar{1}2\rangle$. Below: view of Pt{111} surface along $\langle\bar{1}\bar{1}2\rangle$ showing Ar^+ scattering from a first-layer Pt atom (1) and splitting into two focused beams by an ‘atomic lens’ formed by neighbouring first-layer Pt atoms (2, 3, 4).

(B) PT RECOILING

The images of recoiled Pt atoms [33] by 4 keV Ar^+ are shown in figure B1.23.12. With increasing TOF, the recoil Pt images change from diffuse, to a focused recoil spot at $\beta \sim 25^\circ$ and, finally, to movement of this spot to a higher β that is partially off the MCP. This focused recoil is observed along the $0^\circ \langle\bar{2}11\rangle$ and $60^\circ \langle\bar{1}\bar{1}2\rangle$ azimuths but not along the $30^\circ \langle\bar{1}01\rangle$ and $90^\circ \langle 0\bar{1}1\rangle$ azimuths. The diffuse images at short TOF, e.g. 3.77 μs , correspond to recoil of Pt atoms from the first layer. These recoils have more isotropic distributions and higher energies because there are no atoms above them by which they can be blocked. At longer TOF, e.g. 4.57 μs , the focused recoil is due to Pt atoms from the second layer which have been focused by an ‘atomic lens’ created by first-layer atoms (figure B1.23.12)). The second-layer Pt atom can either be recoiled directly into the atomic lens or it can scatter from the neighbouring aligned third-layer atom into the atomic lens.

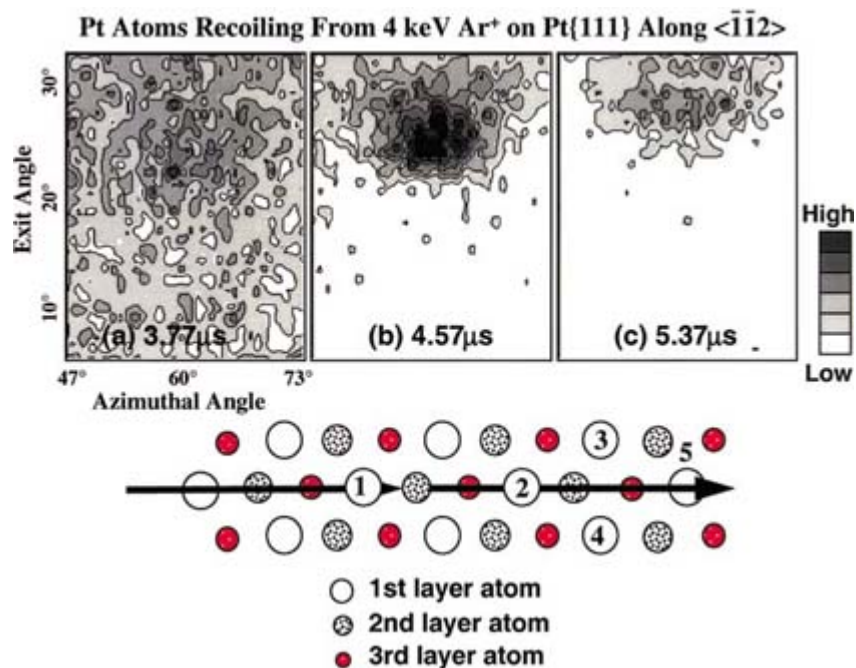


Figure B1.23.12. Above: selected time-resolved SARIS images of 4 keV Ar⁺ recoiling Pt atoms from Pt {111} along $\langle\bar{1}\bar{1}2\rangle$. Below: view of Pt{111} surface along $\langle\bar{1}\bar{1}2\rangle$ showing a focused second-layer Pt recoil trajectory (atoms 1–4 form a focusing ‘atomic lens’).

B1.23.7.2 INTERACTION OF 4 KEV HE WITH PT{111}

A series of time-resolved He scattering images [33] taken as a function of azimuthal angle is shown in [figure B1.23.13](#). The crystal was rotated about its surface normal by 3° for each image. Each image is taken from a 16.7 ns frame corresponding to the QSS TOF. The same intensity scale was used for all of the frames. The observed images are rich in features which change in position and intensity as a function of azimuthal angle. The regions of low intensity correspond to the positions of the centres of the blocking cones; these regions have mainly circular or oval shapes with distortions caused by other overlapping blocking cones. The regions of high intensity correspond to the positions of intersection or near-overlap of blocking cones; atom trajectories are highly focused along the edges of the cones.

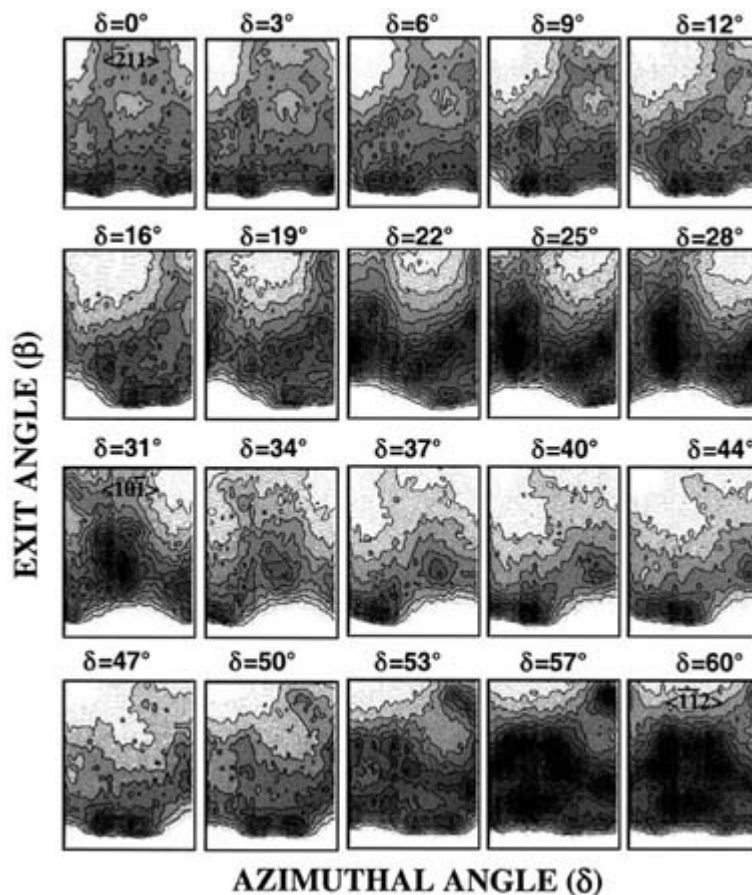


Figure B1.23.13. A series of 20 time-resolved SARIS frames for 4 keV He^+ scattering from $\text{Pt}\{111\}-(1 \times 2)$ taken every 3° of rotation about the azimuthal angle δ , starting with $\delta = 0^\circ$ as the $\langle 211 \rangle$ azimuth and 60° as the $\langle \bar{1}\bar{1}2 \rangle$ azimuth. Each frame represents a 16.7 ns window centred at the TOF corresponding to QSS as predicted by the BCA. The abscissa is the crystal azimuthal angle (δ) and the ordinate is the particle exit angle (β).

The images at $\delta = 0^\circ$ and 60° along the $\langle 211 \rangle$ and $\langle \bar{1}\bar{1}2 \rangle$ azimuths, respectively, are symmetrical about a vertical line through the centre of the frame, as is the crystal structure along these azimuths as shown in [figure B1.23.14](#). The shifts in the positions and sizes of the blocking cones can be monitored as the azimuthal angle δ is rotated away from the symmetrical 0° or 60° directions. There are large variations in the intensities as a function of δ , with the highest intensities being observed along the directions $\delta = 22\text{--}32^\circ$ and $56\text{--}60^\circ$. These high-intensity features result from focusing of ions onto second-layer atoms by the shadow cones of first-layer atoms. The first-layer atoms are symmetrical; however, the second-layer atoms are in sites which are asymmetrical with respect to the first layer, resulting in non-planar scattering trajectories. Very intense features in asymmetrical positions are observed at higher exit angles. These intense features correspond to semichanneling in asymmetrical channels. Semichannels are ‘valleys’ in surfaces through which scattered ions are guided. Along $\langle \bar{1}01 \rangle$ the first-layer atoms form the ‘walls’ and the second-layer atoms form the ‘floor’ of the semichannel. However, the second-layer rows are not centred in the bottom of the channel, resulting in an asymmetrical channel. As a result, the scattered atom trajectories are bent and focused along directions determined by the asymmetry of the channel.

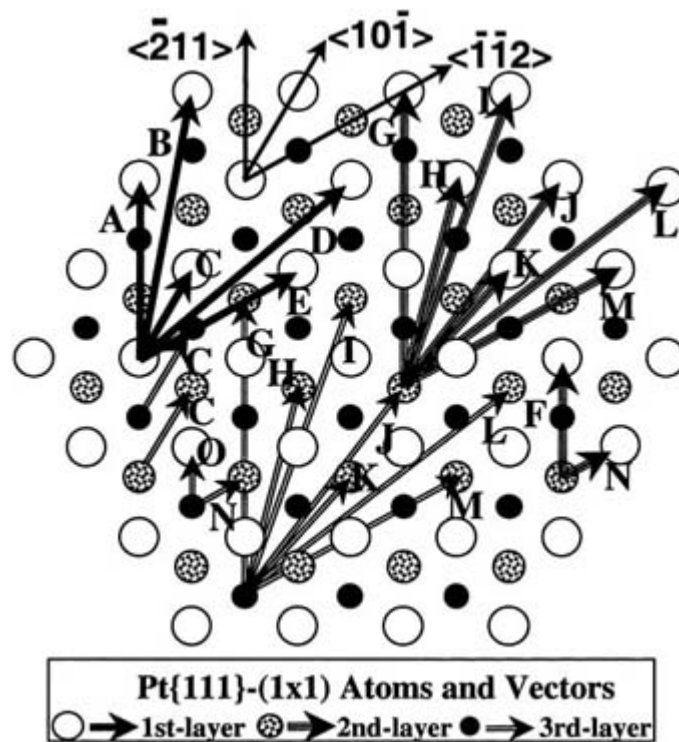


Figure B1.23.14. Schematic illustration of the Pt{111}-(1 × 1) surface. Arrows are drawn to indicate the nearest-neighbour first–first-, second–first-, and third–first-layer interatomic vectors.

The frames along the 0° $\langle \bar{2}11 \rangle$ and 60° $\langle \bar{1}\bar{1}2 \rangle$ azimuths in figure B1.23.13 were selected to compare with those of blocking cone analyses and classical ion trajectory simulations. The arrangement of the first-layer atoms is identical along both of these azimuths; however, the second- and third-layer atoms have a different arrangement with respect to the first-layer atoms. The atoms scattered from second- and third-layer atoms experience a different arrangement of blocking cones on their exit from the surface. The positions of the blocking cones were calculated [39] from the interatomic vectors of figure B1.23.14 and the critical blocking angles or sizes of the cones were calculated with the method described in section B1.23.4. The results are shown in figure B1.23.15. The blocking of scattering trajectories from n th-layer atoms by their neighbouring n th-layer atoms are observed at low β since these atoms are all in the same plane. This first/first-layer atom scattering contributes most of the intensity at low β . The arcs corresponding to the edges of the blocking cones (figure B1.23.15) resulting from the vectors A , B , D and E in figure B1.23.14 occur at $\beta \sim 10^\circ$. The features at higher β correspond to scattering trajectories from second- and third-layer atoms that are blocked and focused by first-layer atoms. The cones resulting from the vectors F , G and O along $\langle \bar{2}11 \rangle$ and M and N along $\langle \bar{1}\bar{1}2 \rangle$ are due to scattering trajectories from second- and third-layer atoms that are blocked by first-layer atoms along these symmetrical directions. These are centred along the azimuths and are directed to higher β values for shorter interatomic spacings. Blocking cones due to the vectors H , I , K and L result from second- and third-layer scattering and are observed at δ values off of the 0° and 60° directions due to non-planar scattering trajectories.

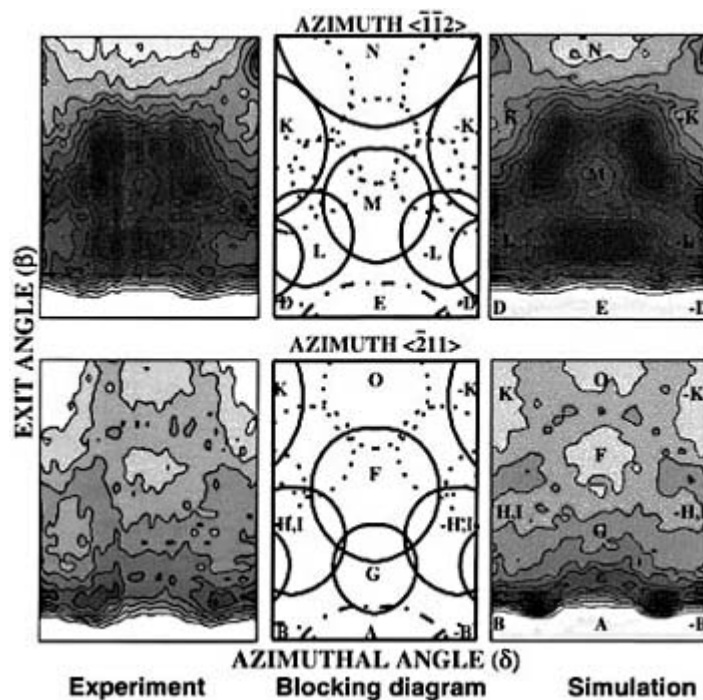


Figure B1.23.15. Experimental images (left), simulated images (right) and blocking cone analyses (centre) for He^+ scattering along the $\langle \bar{2}11 \rangle$ and $\langle \bar{1}\bar{1}2 \rangle$ azimuths. For the calculated blocking cones, first–first, first–second, and first–third layer interactions are identified by dash-dot, solid, and dotted lines, respectively. The scattering parameters are: scattering angle (with respect to incident beam direction) $\theta = 45^\circ$; beam incident angle (with respect to surface plane) $\alpha = 28^\circ$; exit angle of scattered particles along detector normal (relative to surface plane) $\beta = 17^\circ$; flight path to detector (along the detector normal)= 15.5 cm. Angular space subtended by MCP is 27° of crystal azimuthal angle δ and 33° of particle exit angle β .

(A) QUANTITATIVE ANALYSIS

Quantitative analyses can be achieved by using the scattering and recoiling imaging code (SARIC) simulation and minimization of the R -factor [33] (section B1.23.4.4) between the experimental and simulated images as a function of the structural parameters. The SARIC was used to generate simulated images of 4 keV He^+ scattering from bulk-terminated $\text{Pt}\{111\}$ as a function of the first–second interlayer spacing d . Anisotropic thermal vibrations with an amplitude of 0.1 Å were included in the model. A two-dimensional reliability, or R , factor, based on the differences between the experimental and simulated patterns, was calculated as a function of the deviation d of the first–second interlayer spacing from the bulk value. The plots shown in figure B1.23.16 exhibit minima at $d_{\min} = -0.005$ and $+0.005$ Å for the $\langle \bar{1}\bar{1}2 \rangle$ and $\langle \bar{2}11 \rangle$ azimuths, respectively. The optimized simulated images corresponding to d_{\min} are shown in figure B1.23.16, rightmost frames; there is good agreement between these simulated and experimental images. The R -factors are sensitive to changes in the interlayer spacing at the level of 0.01 Å. Based on these data, we conclude that the $\text{Pt}\{111\}$ surface is bulk terminated with the first–second layer spacing within ± 0.01 Å, or 0.4%, of the 2.265 Å bulk spacing. This sensitivity is less than the uncertainty due to the thermal vibrations because SARIS samples the average positions of lattice atoms.

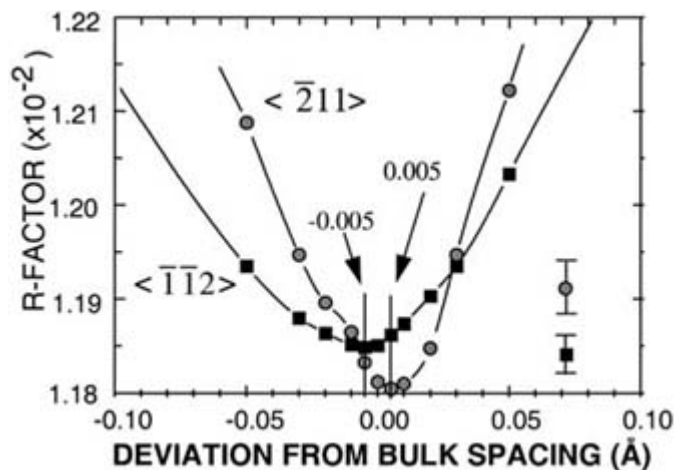


Figure B1.23.16. Plots of the two-dimensional R -factors as a function of the deviation (d) of the first–second interlayer spacing from the bulk value. The experimental and simulated images along the $\langle \bar{2}11 \rangle$ and $\langle \bar{1}\bar{1}2 \rangle$ azimuths of [figure B1.23.15](#) were used in the comparison.

B1.23.8 ION–SURFACE ELECTRON EXCHANGE

One of the unsolved problems in the interaction of low-energy ions with surfaces is the mechanism of charge transfer and prediction of the charge composition of the flux of scattered, recoiled and sputtered atoms. The ability to collect spectra of neutrals plus ions and only neutrals provides a direct measure of scattered and recoiled ion fractions. SARIS images can provide electronic transition probability contour maps which are related to surface electron density and reactivity along the various azimuths.

Ion–surface electron transition probabilities are determined by electron tunnelling between the valence bands of the surface and the atomic orbitals of the ion [42]. Such transition probabilities are highest for close distances of approach. Since TOF-SARS is capable of directly measuring the scattered and recoiled ion fractions, it provides an excellent method for studying ion–surface charge exchange. For simplicity, electron exchange [43] between ions or atoms and surfaces can be discussed in terms of two regions: (i) along the incoming and outgoing trajectories where the particle is within Ångstroms of the surface and (ii) in the close atomic encounter where the core electron orbitals of the collision partners overlap. In region (i), the dominating processes are resonant and Auger electron tunnelling transitions, both of which are fast ($\tau < 10^{-15}$ s). Since the work functions of most solids are lower than the ionization potentials of most gaseous atoms, keV scattered and recoiled species are predominately neutrals as a result of electron capture from the solid. In region (ii), as the interatomic distance R decreases, the atomic orbitals (AOs) of the separate atoms of atomic number Z_1 and Z_2 evolve into molecular orbitals (MOs) of a quasimolecule and finally into the AO of the ‘united’ atom of atomic number ($Z_1 + Z_2$). As R decreases, a critical distance is reached where electrons are promoted into higher-energy MOs because of electronic repulsion and the Pauli exclusion principle. This can result in collisional reionization of neutral species. The fraction of species scattered and recoiled as ions is sensitive to atomic structure through changes in electron density along the trajectories. A direct method for measuring the spatial dependence of charge transfer probabilities with atomic-scale resolution has been developed using the method of DR ion fractions [43]. The data demonstrate the need for an improved understanding of how atomic energy levels shift and broaden near surfaces.

These types of measurements, combined with theoretical modelling, can provide a detailed microscopic map of the local reactivity of the surface as well as electron tunnelling rates within the surface unit cell. This information is of crucial importance for the understanding of various impurity-induced promotion and

poisoning phenomena in catalysis and electron-density maps from scanning tunnelling microscopy.

B1.23.9 ROLE OF SCATTERING AND RECOILING AMONG SURFACE SCIENCE TECHNIQUES

Scattering and recoiling contribute to our knowledge of surface science through (i) elemental analysis, (ii) structural analysis and (iii) analysis of electron exchange probabilities. We will consider the merits of each of these three areas.

B1.23.9.1 ELEMENTAL ANALYSIS

There are two unique features of scattering and recoiling spectrometry: (1) sensitivity to the outermost atomic layer of a surface and (2) sensitivity to surface hydrogen. Using an ESA, it is possible to resolve ions scattered from all elements of mass greater than carbon. The TOF technique is sensitive to all elements, including hydrogen, although it has limited mass resolution. For general qualitative and quantitative surface elemental analyses, XPS and AES remain the techniques of choice.

B1.23.9.2 STRUCTURAL ANALYSIS

The major role of TOF-SARS and SARIS is as surface structure analysis techniques which are capable of probing the positions of all elements with an accuracy of $\lesssim 0.1$ Å. They are sensitive to short-range order, i.e. individual interatomic spacings that are < 10 Å. They provide a direct measure of the interatomic distances in the first and subsurface layers and a measure of surface periodicity in real space. *One of its most important applications is the direct determination of hydrogen adsorption sites by recoiling spectrometry [12, 41].* Most other surface structure techniques do not detect hydrogen, with the possible exception of He atom scattering and vibrational spectroscopy.

TOF-SARS and SARIS are complementary to LEED, which probes long-range order, minimum domain size of 100–200 Å and provides a measure of surface and adsorbate symmetry in reciprocal space. Coupling TOF-SARS, SARIS and LEED provides a powerful combination for surface structure investigations. The techniques of medium- and high- (Rutherford backscattering) energy ion scattering only sample subsurface and bulk structure and are not as surface sensitive as TOF-SARS.

B1.23.10 LOW-ENERGY SCATTERING OF LIGHT ATOMS

Atomic and molecular beams of light atoms such as He, H and H₂ formed from supersonic nozzle beam sources typically have kinetic energies of 20 to 100 meV [44]. Scattering of such low-energy light atoms from surfaces is predominantly elastic. Coherently scattered waves from regularly spaced surface atoms can interfere with each other, giving rise to well known diffraction phenomena. Such hyperthermal atoms have classical turning points that are

typically about 3 Å above the centres of the outermost atomic layer of the surface. The diffraction data probe the outer regions of the atom–surface potential. This information is usually expressed as a corrugation function. Structural information such as bond angles and lengths is obtained by calculating the potential or corrugation function from assumed geometries [44].

The basic components of an apparatus for such atom scattering consist of a UHV scattering chamber equipped with a supersonic atomic and molecular beam source, a sample manipulator and a rotating mass spectrometer. Cryogenic sample temperatures are usually used in order to reduce the vibrational amplitudes of the surface atoms. Data are obtained from the in-plane and out-of-plane diffraction intensity distributions as a function of scattering angle. A set of corrugation parameters is derived from this data. These parameters can be calculated from a first principles approach based on a proportionality relation between the atom–surface interaction potential and the surface charge density. Such diffraction experiments of thermal He atoms coupled with theoretical simulations of the data have been shown to be a very useful structural tool for studying adsorption on surfaces. The method is extremely surface-sensitive, is capable of providing adsorption site information and is one of the few techniques that can detect surface hydrogen.

B1.23.11 SUMMARY

Emphasis in this chapter has been placed on the physical concepts and structural applications of TOF-SARS and SARIS. These techniques are now established as surface structural analysis methods that will have a significant impact in areas as diverse as thin-film growth, catalysis, hydrogen embrittlement and penetration of materials, surface reaction dynamics and analysis of interfaces. Surface crystallography is evolving from the classical concept of a static surface and the question of ‘Where do atoms sit?’ to the concept of a dynamically changing surface. The development of large-area detectors with rapid acquisition of scattering and recoiling structural images, as described in B1.23.7, will provide a technique for capturing time-resolved snapshots of such dynamically changing surfaces.

REFERENCES

- [1] Smith D P 1967 Scattering of low-energy noble gas ions from metal surfaces *J. Appl. Phys.* **38** 340–7
- [2] Heiland W and Taglauer E 1977 The backscattering of low energy ions and surface structure *Surf. Sci.* **68** 96–107
- Heiland W and Taglauer E 1975 Low energy ion scattering and Auger electron spectroscopy studies of clean nickel surfaces and adsorbed layers *Surf. Sci.* **47** 234–43
- Heiland W and Taglauer E 1973 Bombardment induced surface damage in a nickel single crystal observed by ion scattering and LEED *Rad. Effects.* **19** 1–6
- Heiland W, Iberl F, Taglauer E and Menzel D 1975 Oxygen adsorption on (110) silver *Surf. Sci.* **53** 383–92
-

- [3] Brongersma H H and Theeten J B 1976 The structure of oxygen adsorbed on Ni(0001) as determined by ion scattering spectroscopy *Surf. Sci.* **54** 519–24
- Brongersma H H and Mul P 1973 Analysis of the outermost atomic layer of a surface by low-energy ion scattering *Surf. Sci.* **35** 393–412
- [4] Suurmijer E P Th M and Boers A L 1973 Low-energy ion reflection from metal surfaces *Surf. Sci.* **43** 309–52
- [5] DeWit A G J, Bronckers R P N and Fluit J M 1979 Oxygen adsorption on Cu(110): determination of atom positions with low energy ion scattering *Surf. Sci.* **82** 177–94

- [6] Aono M, Hou Y, Souda R, Oshima C, Otani S, Ishizawa Y, Matsuda K and Shimizu R 1982 Interaction potential between He⁺ and Ti in a keV range as revealed by a specialized technique in ion scattering spectroscopy *Japan. J. Appl. Phys. Lett.* **21** L670–2
- Aono M, Hou Y, Oshima C and Ishizawa Y 1982 Low-energy ion scattering from the Si(001) surface *Phys. Rev. Lett.* **49** 567–70
- Aono M and Souda R 1985 Quantitative surface atomic structure analysis by low energy ion scattering spectroscopy *Japan. J. Appl. Phys. Part 1* **24** 1249–62
- [7] Niehus H 1984 Analysis of the Pt(110) × (1 × 2) surface reconstruction *Surf. Sci.* **145** 407–18
- Niehus H and Comsa G 1984 Determination of surface reconstruction with impact-collision alkali ion scattering *Surf. Sci.* **140** 18–30
- [8] Marchut L, Buck T M, Wheatley G H and McMahon C J Jr 1984 Surface structure analysis using low energy ion scattering *Surf. Sci.* **141** 549–66
- [9] Grizzi O, Shi M, Bu H, and Rabalais J W 1990 Time-of-flight scattering and recoiling spectrometer (TOF-SARS) for surface analysis *Rev. Sci. Instrum.* **61** 740–52
- [10] Rabalais J W 1990 Scattering and recoiling spectrometry: an ion's eye view *Science* **250** 521–7
- [11] Kim C, Hofner C, Al-Bayati A and Rabalais J W 1998 Scattering and recoiling imaging spectrometer (SARIS) *Rev. Sci. Instrum.* **69** 1676–84
- [12] Aono M, Katayama M and Nomura E 1992 Exploring surface structures by coaxial impact collision ion scattering spectroscopy (CAICISS) *Nucl. Instrum. Methods B* **64** 29–37
- [13] Ghrayeb R, Purushotham M, Hou M and Bauer E 1987 Estimate of repulsive interatomic pair potentials by low-energy alkalimetal-ion scattering and computer simulation *Phys. Rev. B* **36** 7364–70
- [14] Chester M and Gustafsson T 1991 Geometric structure of the Si(111)-(√3 × √3)R30°-Au surface *Surf. Sci.* **256** 135–46
- [15] Hetterich W, Höfner C and Heiland W 1991 An ion scattering study of the surface structure and thermal vibrations on Ir(110) *Surf. Sci.* **251/252** 731–6
- [16] O'Connor D J, King B V, MacDonald R J, Shen Y G and Chen X 1990 The study of surfaces using ion beams *Aust. J. Phys.* **43** 601
- [17] Dodonoy A I, Mashkova E S and Molchanov V A 1989 Medium-energy ion scattering by solid surfaces. III: ejection of fast recoil atoms from solids under ion bombardment *Rad. Eff. Def. Sol.* **110** 227–341
- [18] Mintz M H, Atzmony U and Shamir N 1987 Initial adsorption kinetics of oxygen on polycrystalline copper *Surf. Sci.* **185** 413–30

- [19] van de Riet E and Niehus A 1991 Application of low energy neutral ionization spectroscopy for surface structure analysis *Surf. Sci.* **243** 43–8
- [20] Niehus H, Spitzl R, Besocke K and Comsa G 1991 N-induced (2 × 3) reconstruction of Cu(110): evidence for long-range, highly directional interaction between Cu–N–Cu bonds *Phys. Rev. B* **43** 12 619–25
- [21] Shoji F, Kashihara K, Sumitomo K and Oura K 1991 Low-energy recoil-ion spectroscopy studies of hydrogen adsorption on Si(100)-2 × 1 surfaces *Surf. Sci.* **242** 422–7
- [22] Overbury S H, Mullins D R, Paffett M T and Koel B E 1991 Surface structure determination of Sn deposited on Pt(111) by low energy alkali ion scattering *Surf. Sci.* **254** 45–57
- [23] Taglauer E, Beckschulte M, Margraf R and Mehl D 1988 Recent developments in the applications of ion scattering spectroscopy *Nucl. Instrum. Methods B* **35** 404–9
- [24] Bracco G, Canepa M, Catini P, Fossa F, Mattered L, Terreni S and Truffelli D 1992 Impact-collision ion scattering study of Ag(110) *Surf. Sci.* **269/270** 61–7

- [25] Mashkova E S and Molchanov V A 1985 *Medium-Energy Ion Reflection From Solids* (Amsterdam: North-Holland)
- [26] Goldstein H 1980 *Classical Mechanics* 2nd edn (Reading, MA: Addison-Wesley)
- [27] Gibson J B, Goland A N, Milgram M and Vineyard G H 1960 Dynamics of radiation damage *Phys. Rev. Series 2* **120** 1229–53
- [28] Johnson L W and Reiss R D 1982 *Numerical Analysis* (Berlin: Springer)
- [29] Press W H, Flannery B P, Teukolsky S A and Vetterling W T 1988 *Numerical Recipe in C* (Cambridge: Cambridge University Press) p 131
- [30] Parilis E S *et al* 1993 *Atomic Collisions on Solids* (New York: North-Holland)
- [31] Zeigler J F, Biersack J P and Littmark U 1985 *The Stopping and Range of Ions in Solids* (New York: Pergamon)
- [32] Zangwill A 1989 *Physics at Surfaces* (Cambridge: Cambridge University Press) p 117
- [33] Kim C, Hofner C and Rabalais J W 1997 Surface structure determination from ion scattering images *Surf. Sci.* **388** L1085–91
- [34] Xu M L and Tong S Y 1985 Multilayer relaxation for the clean Ni(110) surface *Phys. Rev. B* **31** 6332–6
- [35] Shi M, Wang Y and Rabalais J W 1993 Structure of the Si{100} surface in the clean (2 × 1), (2 × 1)-H monohydride, (1 × 1)-H dihydride, and c(4 × 4)-H phases *Phys. Rev. B* **48** 1678–88
- [36] Masson F and Rabalais J W 1991 Time-of-flight scattering and recoiling spectrometry (TOF-SARS) analysis of Pt{110}. I. Quantitative structure study of the clean (1 × 2) surface *Surf. Sci.* **253** 245–57
- Masson F and Rabalais J W 1991 Time-of-flight scattering and recoiling spectrometry (TOF-SARS) analysis of Pt{110}. II. The (1 × 2)-to-(1 × 3) interconversion and characterization of the (1 × 3) phase **253** 258–69
- [37] Masson F and Rabalais J W 1991 Surface periodicity exposed through shadowing and blocking effects *Chem. Phys. Lett.* **179** 63–7
- [38] Roux C D, Bu H and Rabalais J W 1991 Structure of the hydrogen induced Ni{110}-p(1 × 2)-H reconstructed surface *Surf. Sci.* **259** 253–65
- Roux C D, Bu H and Rabalais J W 1992 Hydrogen adsorption site on the Ni{110}-p(1 × 2)-H surface from time-of-flight scattering and recoiling spectrometry (TOF-SARS) *Surf. Sci.* **271** 68–80

- [39] Kim C and Rabalais J W 1997 Projections of atoms in terms of interatomic vectors *Surf. Sci.* **385** L938–44
- [40] Kim C, Höfner C, Bykov V and Rabalais J W 1997 Element-, time-, and spatially-resolved images of scattered and recoiled atoms *Nucl. Instrum. Methods B* **125** 315–22
- Kim C and Rabalais J W 1998 Focusing of He⁺ ions on semichannel planes in the Pt{111} surface *Surf. Sci.* **395** 239–47
- [41] Hofner C, Bykov V and Rabalais J W 1997 Three-dimensional focusing patterns of He⁺ ions scattering from a Au{110} surface *Surf. Sci.* **393** 184–93
- Kim C, Ahn J, Bykov V and Rabalais J W 1998 Element-, velocity-, and spatially-resolved images of Kr⁺ scattering and recoiling from a CdS surface *Int. J. Mass Spectrom. Ion Phys.* **174** 305–15
- [42] Hsu C C and Rabalais J W 1991 Structure sensitivity of scattered Ne⁺ ion fractions from a Ni{100} surface *Surf. Sci.* **256** 77–86
- [43] Hsu C C, Bu H, Bousetta A, Rabalais J W and Nordlander P 1992 Angular dependence of charge

transfer probabilities between O- and a Ni{100}-c(2 × 2)-O surface *Phys. Rev. Lett.* **69** 188–91

- [44] Engel T and Rieder K H 1982 Structural studies of surfaces with atomic and molecular beam diffraction *Structural Studies of Surfaces With Atomic and Molecular Beam Scattering (Springer Tracts in Modern Physics vol 91)* (Berlin: Springer) pp 55–180
-

FURTHER READING

Rabalais J W 1994 Low energy ion scattering and recoiling *Surf. Sci.* **299/300** 219–32

A review of the ion scattering literature starting in the 1960s.

Rabalais J W 1992 Surface crystallography from ion scattering *Chemistry in Britain* **28** 37–71

A simple survey of ion scattering for students.

Niehus H, Heiland W and Taglauer E 1993 Low-energy ion scattering at surfaces *Surf. Sci. Rep.* **17** 213–304

An excellent review of ion scattering.

Rabalais J W (ed) 1994 *Low Energy Ion–Surface Interactions* (Chichester: Wiley)

A volume with contributions from several authors that treats ion–surface interactions at different energies.

Taglauer E 1997 Low-energy ion scattering and Rutherford backscattering *Surface Analysis; The Principal Techniques* ed J C Vickerman (Chichester: Wiley) pp 215–66

A textbook that treats the principal techniques of surface science.

-1-

B1.24 Rutherford backscattering, resonance scattering, PIXE and forward (recoil) scattering

C C Theron, V M Prozesky and J W Mayer

B1.24.1 INTRODUCTION

The use of million electron volt (MeV) ion beams for materials analysis was instigated by the revolution in integrated circuit technology. Thin planar structures were formed in silicon by energetic ion implantation of dopants to create electrical active regions and thin metal films were deposited to make interconnections between the active regions. Ion implantation was a new technique in the early 1960s and interactions between metal films and silicon required analysis. For example, the number of ions implanted per square centimetre (ion dose) and thicknesses of metal layers required careful control to meet the specifications of integrated circuit technology. Rutherford backscattering spectrometry (RBS) and MeV ion beam analysis were developed in response to the needs of the integrated circuit technology. In turn integrated circuit technology provided the electronic sophistication used in the instrumentation in ion beam analysis. It was a synergistic development of analytical tools and the fabrication of integrated circuits.

Rutherford backscattering spectrometry is the measurement of the energies of ions scattered back from the surface and the outer microns (1 micron = 1 μm) of a sample. Typically, helium ions with energies around 2 MeV are used and the sample is a metal coated silicon wafer that has been ion implanted with about a

monolayer (10^{15} ions cm^{-2}) of electrically active dopants such as arsenic. Only moderate vacuum levels (about 10^{-4} Pa) are required so that sample exchange is rapid allowing the analysis of the number of implanted atoms per square centimetre and their distribution in depth to be carried out in periods of about 15 minutes or less. The sample is not damaged structurally during the analysis and therefore Rutherford backscattering spectrometry is considered non-destructive. This is in contrast to surface sensitive techniques such as Auger electron spectroscopy where surface erosion by sputtering is required for depth analysis. One of the strong features of Rutherford backscattering spectrometry is that the scattering cross sections are well known so that the analysis is quantitative. In other analytical techniques, such as secondary ion mass spectrometry (SIMS), the cross sections are not well defined. The relative ion yields can vary over three orders of magnitude depending on the nature of the surface. Rutherford backscattering has been a convenient way to calibrate secondary ion mass spectrometry, which in turn is more sensitive to the detection of trace elements than Rutherford backscattering. The two techniques are thus complementary.

Ion beam analysis grew out of nuclear physics research on cross sections and reaction products involved in atomic collisions. In this work million volt accelerators were developed and used extensively. As the energies of the incident particles increased, the lower energy accelerators became available for use in solid state applications. The early nuclear physics research used magnetic spectrometers to measure the energies of the particles. This analytical procedure was time consuming and the advent of the semiconductor nuclear particle detector allowed simultaneous detection of all particle energies. It was an energy dispersive spectrometer. The semiconductor detector is a Schottky barrier (typically a gold film on silicon) or shallow diffused p-n junction with the active region defined by the high electric field in the depletion layer. The active region extends tens of microns below the surface of the detector so that in almost every application the penetration of the energetic particles is confined within the active region. The response of the detector is linear with the energy of the particles providing a true particle energy spectrometer.

-2-

Analysis of ion implanted layers and metal-silicon interactions was carried out with Rutherford backscattering at 2.0 MeV energies and with semiconductor nuclear particle detectors for several years. Rutherford backscattering became well established and was utilized in materials analysis in industrial and university laboratories across the world. The importance of hydrogen and its influence in solid state chemistry led to the development of forward scattering in which one measures the energy of the recoiling hydrogen atom. The helium ion is heavier than that of hydrogen so that by tilting the sample it is possible to measure the recoil energy of the emerging hydrogen, again with a nuclear particle detector. In other words, the modification to the Rutherford backscattering spectrometry target chamber geometry was only to tilt the target and to move the detector. These forward recoil techniques have of course become more sophisticated with use of heavy incident ions and detectors which measure both the energy and the mass of the recoiling particles ($\Delta E - E$ or time of flight detector).

Analysis of silicon is an almost ideal experimental situation because the masses of most implanted atoms and metal layers exceed that of silicon. In Rutherford backscattering the mass of the atom must be greater than that of the silicon target to separate the energy signals of the target atom from those of the silicon spectrum. Oxygen is an exception. It is lighter than silicon and also is ubiquitous in surface and interface layers. The analysis of oxygen, and also carbon and nitrogen, are carried out in the same experimental chamber as used in Rutherford backscattering, but the energy of the incident helium ions is increased to energies where there are resonances in the backscattering cross sections. These resonances increase the yield of the scattered particle by nearly two orders of magnitude and provide high sensitivity to the analysis of oxygen and carbon in silicon. The use of these high energies, 3.04 MeV for the helium-oxygen resonance, is called resonance scattering and the word Rutherford is inappropriate for a descriptor.

By inserting a semiconductor x-ray detector into the analysis chamber, one can measure particle induced x-rays. The cross section for particle induced x-ray emission (PIXE) is much greater than that for Rutherford backscattering and PIXE is a fast and convenient method for measuring the identity of atomic species within

the outer microns of the sample surface. The energy resolution in helium ion Rutherford backscattering spectrometry does not allow discrimination between the signals from high atomic number (high Z) elements close to each other in the periodic table. With conventional semiconductor detectors one cannot distinguish between gold and tungsten, for example, whereas the ion induced x-ray energies are easily distinguished for the two high Z elements. PIXE, then, becomes another tool in the MeV ion analysis chamber and only requires the addition of a x-ray detector system.

The dimensions of the incident ion beam are typically 1 mm across the width of the incident beam impinging on the sample surface. This dimension can be easily obtained using slits in the beam handling system. The beam diameter can be reduced by orders of magnitude by using quadrupole or electrostatic lenses to focus the ion beam to diameters of about one micron on the sample surface. The beam is then rastered across the surface to provide a visual image of the surface with micron resolution. In this work the large cross sections for PIXE are important, because sample analysis can be performed without sample damage caused by the high current density of incident ions.

This overview covers the major techniques used in materials analysis with MeV ion beams: Rutherford backscattering, channelling, resonance scattering, forward recoil scattering, PIXE and microbeams. We have not covered nuclear reaction analysis (NRA), because it applies to special incident-ion–target-atom combinations and is a topic of its own [1, 2].

B1.24.2 RUTHERFORD BACKSCATTERING SPECTROMETRY (RBS)

The discussion of Rutherford backscattering spectrometry starts with an overview of the experimental target chamber, proceeds to the particle kinematics that determine mass identification and depth resolution, and then provides an example of the analysis of a silicide.

B1.24.2.1 TARGET CHAMBER

Figure B1.24.1 shows the placement of the sample and detectors in the target chamber. The sample is located so that its surface is on an axis of rotation of a goniometer so that the beam position does not shift across the sample as the sample is tilted with respect to the incident ion beam. The backscattering detector is mounted as close to the incident beam as possible so that the average backscattering angle, θ , is close to 180° , typically 170° , with a detector solid angle of about 3–5 milliradians (msr). In some cases annular detectors are used with the incident beam passing through the centre of the detector aperture in order to provide larger analysis solid angles. The sample is rotated to glancing angle geometries when the forward scattering detector is used. Typically a thin foil is placed in front of the detector to block the helium ions while allowing the hydrogen ions to pass through with only minimal energy loss. The stopping power (energy loss) of MeV helium ions is ten times that of the recoiling hydrogen ions. As shown in [section B1.24.8](#) below, the forward scattering detector system can be augmented to include a $\Delta E - E$ detector to allow identification of the ion mass as well as energy. The x-ray detector is placed so that the active region is in full view of the sample surface bombarded with the incident ions.

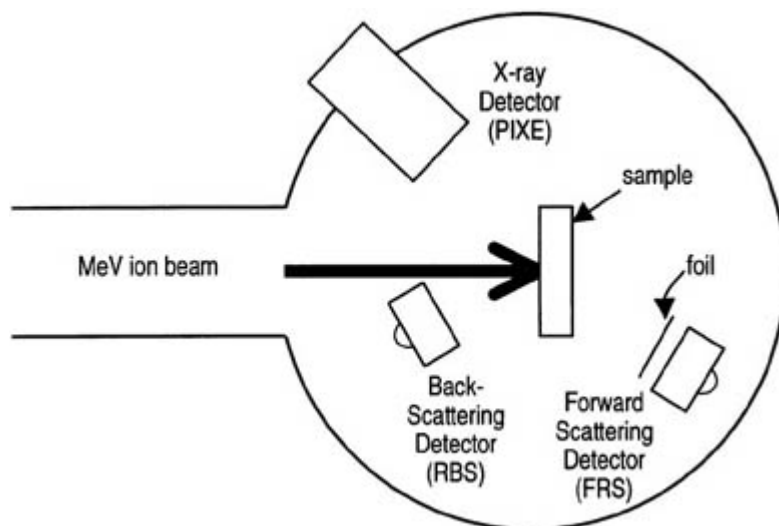


Figure B1.24.1. Schematic diagram of the target chamber and detectors used in ion beam analysis. The backscattering detector is mounted close to the incident beam and the forward scattering detector is mounted so that, when the target is tilted, hydrogen recoils can be detected at angles of about 30° from the beam direction. The x-ray detector faces the sample and receives x-rays emitted from the sample.

-4-

For conventional backscattering spectrometry with helium ions, the energy resolution of the semiconductor particle detector is typically 15 kiloelectron volts (keV). This resolution can be improved to 10 keV with special detectors and detector cooling. The output signal, which is typically millivolts in pulse height, is processed by silicon integrated circuit electronics and provides an energy spectrum in terms of number of particles versus energy. It is often displayed as particles versus channel number as the energy scale is divided into channels which must be calibrated to give the energy scale. The calibration between the measured particle energy and the channel number is independent of the ion energy and sample analysed and only depends on the semiconductor detector and associated electronics response to the energy of the ion beams.

The vacuum requirements in the target chamber are relatively modest (10^{-4} Pa) and are comparable to those in the accelerator beam lines. All that is required is that the ion beam does not lose energy on its path to the sample and that there is minimal deposition of contaminants and hydrocarbons on the surface during analysis.

B1.24.2.2 KINEMATICS

In ion beam analysis the incident particle penetrates into the silicon undergoing inelastic collisions, predominantly with target electrons, and loses energy as it penetrates to the end of its range. The range of 2.5 MeV helium ions is about 10 microns in silicon; the range of comparable energy protons is about ten times that of the helium ions (the range of 3 MeV hydrogen is about 100 microns in silicon). During the penetration of the helium ions, a small fraction undergo elastic collisions with the target atom to give the backscattering signal.

Figure B1.24.2 is a schematic representation of the geometry of an elastic collision between a projectile of mass M_1 and energy E_0 with a target atom of mass M_2 initially at rest. After collision the incident ion is scattered back through an angle θ and emerges from the sample with an energy E_1 . The target atom after collision has a recoil energy E_2 . There is no change in target mass, because nuclear reactions are not involved and energies are non-relativistic.

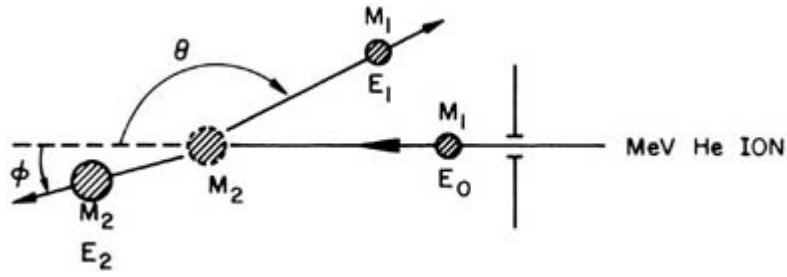


Figure B1.24.2. A schematic representation of an elastic collision between a particle of mass M_1 and energy E_0 and a target atom of mass M_2 . After the collision the projectile and target atoms have energies of E_1 and E_2 respectively. The angles θ and ϕ are positive as shown. All quantities refer to the laboratory frame of reference.

The ratio of the projectile energies for $M_1 < M_2$ is given by

$$K = \frac{E_1}{E_0} = \left[\frac{(M_2^2 - M_1^2 \sin^2 \theta)^{1/2} + M_1 \cos \theta}{M_2 + M_1} \right]^2. \quad (\text{B1.24.1})$$

-5-

The energy ratio, called the kinematic factor $K = E_1/E_0$, shows that the energy after scattering is determined by the masses of the incident particle and target atoms and the scattering angle. For a direct backscattering through 180° the energy ratio has its lowest value given by

$$\frac{E_1}{E_0} = \left(\frac{M_2 - M_1}{M_2 + M_1} \right)^2. \quad (\text{B1.24.2})$$

For incident helium ions ($M_1 = 4$) at $E_0 = 2.0$ MeV the energy E_1 of the backscattered particle for silicon ($M_2 = 28$) is 1.12 MeV and for palladium ($M_2 = 106$) the energy is 1.72 MeV.

The energy E_2 transferred to the target atom has a general relation given by

$$\frac{E_2}{E_0} = \frac{4M_1M_2}{(M_1 + M_2)^2} \cos^2 \phi \quad (\text{B1.24.3})$$

and at $\theta = 180^\circ$ the energy E_2 transferred to the target atom has its maximum value given by

$$\frac{E_2}{E_0} = \frac{4M_1M_2}{(M_1 + M_2)^2}. \quad (\text{B1.24.4})$$

In collisions where $M_1 = M_2$ at $\theta = 180^\circ$ the incident particle is at rest after the collision, with all the energy transferred to the target atom. For 2.0 MeV helium ions colliding with silicon the recoil energy E_2 is 0.88 MeV and from palladium is 0.28 MeV.

The ability to identify different mass species depends on the energy resolution of the detector which is typically 15 keV full width at half maximum (FWHM). For example, silver has a mass $M_2 = 108$ and tin has a mass $M_2 = 119$. The difference between $K_{\text{Ag}} = 0.862$ and $K_{\text{Sn}} = 0.874$ is 0.012. For 2 MeV helium ions the

difference in backscattering energy is 24 keV which is well outside the range of the detector resolution, indicating that signals from Ag and Sn on the surface can be resolved. The difference between gold and tungsten K values is 0.005, and at 2 MeV energies one would not resolve the signals between gold and tungsten. With Rutherford backscattering and conventional detectors with energy resolution of 15 keV one can resolve the signals from and identify the elements of masses up to 100. One can resolve isotopes up to a mass of around 60. For example, all the silicon isotopes can be identified.

B1.24.2.3 SCATTERING CROSS SECTION

The identity of target elements is established by the energy of the scattered particles after an elastic collision. The number of atoms per unit area, N_S , is found from the number Q_D of detected particles (called the yield, Y) for a given number Q of particles incident on the target. The connection is given by the scattering cross section $\sigma(\theta)$ by

$$Y = Q_D = \sigma(\theta)\Omega Q N_S. \quad (\text{B1.24.5})$$

-6-

This is shown schematically in figure B1.24.3. In the simplest approximation the scattering cross section σ is given by

$$\sigma(\theta) = \left(\frac{Z_1 Z_2 e^2}{4E} \right)^2 \frac{1}{\sin^4 \theta/2}, \quad (\text{B1.24.6})$$

the scattering cross section originally derived by Rutherford. For 2 MeV helium ions incident on silver, $Z_2 = 47$ at 180° , the cross section is $2.89 \times 10^{-24} \text{ cm}^2$ or 2.89 barns where the barn = 10^{-24} cm^2 . The distance of closest approach is about $7 \times 10^{-4} \text{ \AA}$ which is smaller than the K-shell radius of silver (10^{-2} \AA). This means that the incident helium ion penetrates well within the innermost radius of the electrons so that one can use an unscreened Coulomb potential for the scattering. The distance of closest approach is sufficiently large that penetration into the nuclear core does not occur and one neglects nuclear reactions.

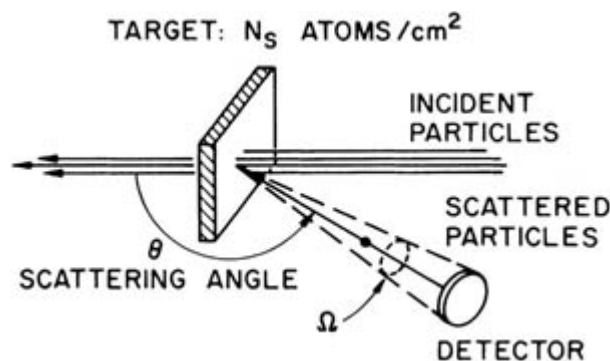


Figure B1.24.3. Layout of a scattering experiment. Only primary particles that are scattered within the solid angle Ω spanned by the solid state detector are counted.

The cross sections are sufficiently large that one can detect sub-monolayers of most heavy mass elements on silicon. For example, the yield of 2.0 MeV helium ions from 10^{14} cm^{-2} silver atoms (approximately one-tenth of a monolayer) is 800 counts for a current of 100 nanoamperes for 15 minutes and detector area of 5 msr.

This represents a large signal for a small number of atoms on the surface. With care, 10^{-4} monolayers of gold on silicon can be detected.

B1.24.2.4 DEPTH SCALE

Light ions such as helium lose energy through inelastic collision with atomic electrons. In backscattering spectrometry, where the elastic collision takes place at depth t below the surface, one considers the energy loss along the inward path and on the outward path as shown in [figure B1.24.4](#). The energy loss on the way in is weighted by the kinematic factor and the total is

$$\Delta E = \Delta t \left(K \frac{dE}{dx} \Big|_{\text{in}} + \frac{1}{|\cos \theta|} \frac{dE}{dx} \Big|_{\text{out}} \right) = \Delta t [S] \quad (\text{B1.24.7})$$

-7-

where dE/dx is the rate of energy loss with distance and $[S]$ is the energy loss factor. An example illustrating the influence of depth on analysis is given in [figure B1.24.5](#) which shows two thin gold layers on the front and back of a nickel film. The scattering from gold at the surface is clearly separated from gold at the back layer. The energy width between the gold signals is closely equal to that of the energy width of the nickel signal. This signal is nearly square shaped because nickel exists from the front to the back surface. The depth scales are determined from energy loss values, which are given in tabular form as a function of energy [1, 2]. It is often expressed as a stopping cross section in terms of $(1/N) dE/dx$, which gives values in eV cm^2 . The depth resolution is given by dividing the detector resolution by the energy loss factor. For 2 MeV helium in silicon one might expect a depth resolution of about 200 Å for 180° scattering geometries. This can be reduced to values of about 50 Å for glancing incident and exit angles. These values of depth resolution degrade as the particle penetrates into the sample and energy straggling becomes a factor.

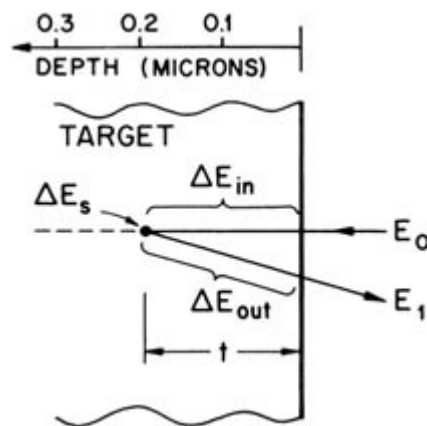


Figure B1.24.4. Energy loss components for a projectile that scatters from depth t . The particle loses energy ΔE_{in} via inelastic collisions with electrons along the inward path. There is energy loss ΔE_S in the elastic scattering process at depth t . There is energy lost to inelastic collisions ΔE_{out} along the outward path. For an incident energy E_0 the energy of the exiting particle is $E_1 = E_0 - \Delta E_{\text{in}} - \Delta E_S - \Delta E_{\text{out}}$.

-8-

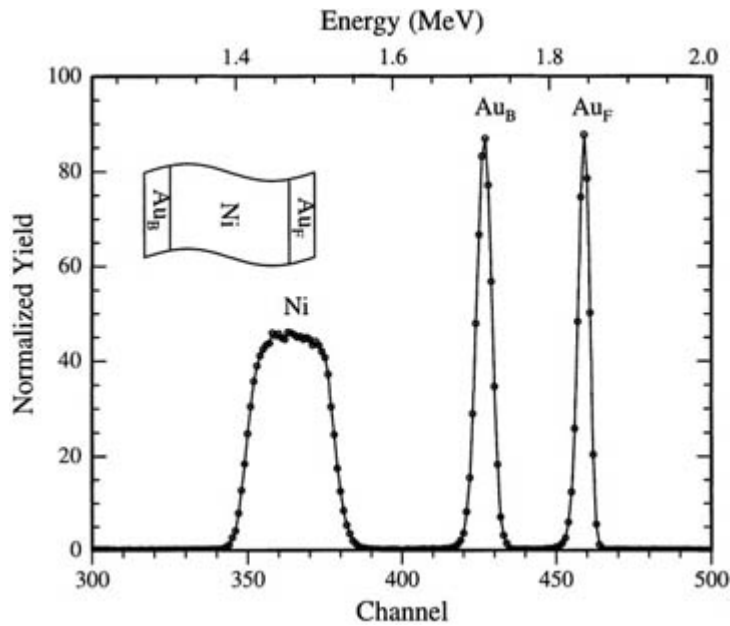


Figure B1.24.5. Backscattering spectrum of a thin Ni film (950 Å) with near monolayers ($\approx 30 \times 10^{15}$ at cm^{-2}) of Au on the front and back surfaces of the Ni film. The signals from the front and back layers of Au are shown and are separated in energy from each other by nearly the same energy width as the Ni signal.

B1.24.2.5 SIMULATION

Rutherford backscattering spectra can be analysed by use of some of the available analysis programs. Programs such as RUMP and GISA provide a layer-by-layer signal for multielement targets [3, 4, 5]. These programs include detector resolution, energy straggling and individual isotopes, and can also be applied to forward recoil spectrometry for detection of light elements. These programs also include provisions for enhanced cross section for light elements such as carbon and oxygen.

B1.24.2.6 SILICIDE FORMATION

An example of Rutherford backscattering spectrometry of the formation of PtSi is shown in figure B1.24.6 for the case where the original Pt layer has reacted to form Pt₂Si. The backscattering signals at the high energy end near 1.8 MeV represent Pt at the surface of the sample. The plateau extends downward in energy to 1.7 MeV where there is a step down to the signals from Pt in Pt₂Si. In the Pt signal the contribution from Pt₂Si is shown shaded. In the Si portion of the spectrum the signal steps upward around 1.0 MeV and represents the silicon in the Pt₂Si. The second step represents the Si signal from the Si substrate. In this case, the signals from the unreacted Pt, the Pt and Si in Pt₂Si and the Si in the substrate are clearly identified and can be used to specify the thickness and composition of the silicide layer.

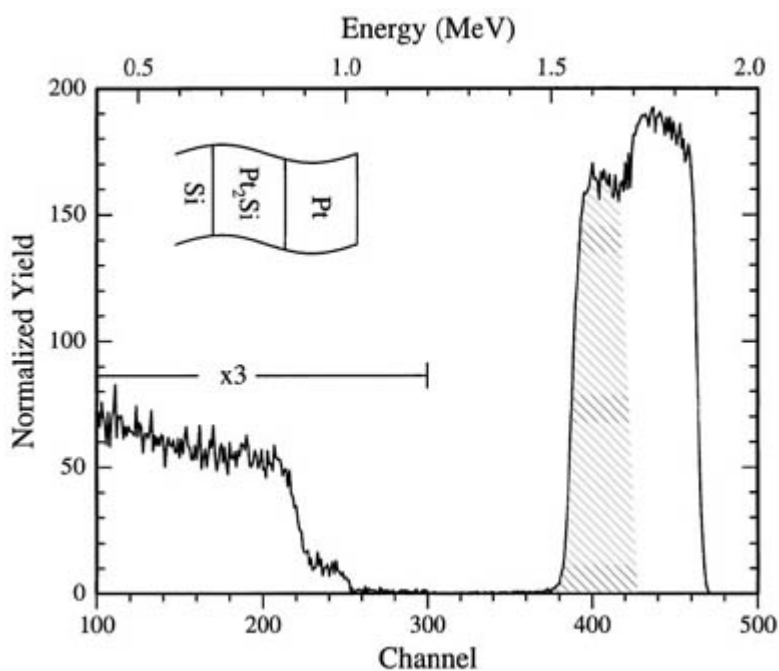


Figure B1.24.6. Backscattering spectrum of a layer of Pt on Si that has been thermally heated so that approximately half the Pt has been consumed in the formation of Pt_2Si , the first stage in the reaction of Pt with Si. In the spectrum the signals from Si have been multiplied by a factor of three for visibility, because the atomic number of Si (14) is much less than that of Pt (78). In the spectrum the signals from Pt are in the region of 1.6–1.8 MeV. The higher energy corresponds to scattering from unreacted Pt and the step around 1.7 MeV corresponds to the transition from Pt to Pt_2Si . The signal at 1.6 MeV corresponds to the back interface of the Pt_2Si in contact with silicon. The silicon signal in the energy range from about 0.9–1.0 MeV corresponds to the Si in the Pt_2Si . At lower energies the spectrum represents signals from the Si substrate.

B1.24.3 IN SITU REAL-TIME RBS

The essentially non-destructive nature of Rutherford backscattering spectrometry, combined with the its ability to provide both compositional and depth information, makes it an ideal analysis tool to study thin-film, solid-state reactions. In particular, the non-destructive nature allows one to perform *in situ* RBS, thereby characterizing both the composition and thickness of formed layers, without damaging the sample. Since only about two minutes of irradiation is needed to acquire a Rutherford backscattering spectrum, this may be done continuously to provide a real-time analysis of the reaction [6].

There are two main applications for such real-time analysis. The first is the determination of the chemical reaction kinetics. When the sample temperature is ramped linearly with time, the data of thickness of formed phase together with ramped temperature allows calculation of the complete reaction kinetics (that is, both the activation energy and the pre-exponential factor) from a single sample [6], instead of having to perform many different temperature ramps as is the usual case in differential thermal analysis [7, 8, 9, 10 and 11]. The second application is in determining the

contribution that each of the elements in the reaction couple makes to the overall atomic transport across the forming layer. For this purpose, thin, inert markers (analogous to the thin wires used by Kirkendall [12, 13]) are inserted into the layers to establish a reference frame within which to measure the contribution of each element's flux to the overall growth. Without the use of a real-time analysis technique, one must rely on the

use of different samples, which, although nominally identical, do not necessarily behave identically since many of these reactions depend critically on the exact conditions at the interfaces between the layers. On the other hand, if analysis can be performed on a single sample, small changes in the position of the marker can then confidently be interpreted. Examples of these two applications are presented below.

B1.24.3.1 PT-SI

When a thin (about 3000 Å) layer of Pt is deposited onto a Si wafer and then heated, the first phase that forms at the interface between Pt and Si is Pt₂Si. After all the Pt has been consumed, the newly formed Pt₂Si layer reacts with the Si to form PtSi, which is stable in contact with excess Si. No further reaction is observed.

Figure B1.24.7 shows the progress of this reaction as the temperature is ramped linearly at a rate of 1°C min⁻¹. At time zero, the signal between 1.5 and 1.9 MeV is from the unreacted Pt layer, whereas the signal from the Si wafer appears below about 0.9 MeV. The signal from the Si has been magnified by a factor of three to compensate for the differences in cross sections between Pt (Z = 78) and Si (Z = 14).

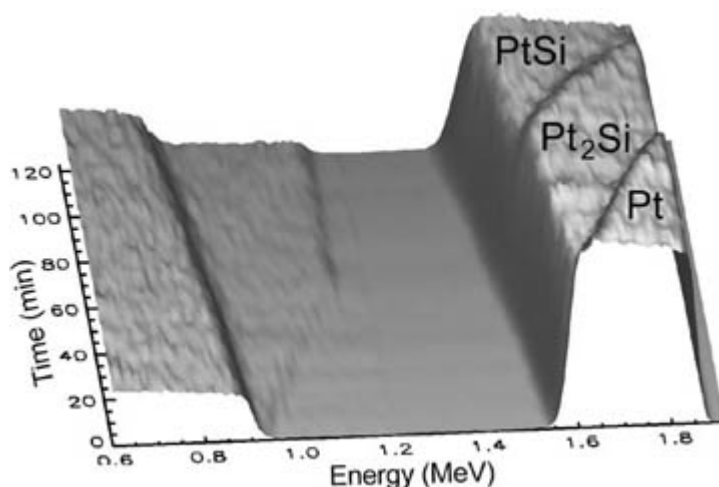


Figure B1.24.7. A three-dimensional plot of backscattering signal versus time for a Pt film deposited onto Si and heated at a rate of 1 °C min⁻¹. At time zero, the Pt signal shows a square-topped energy distribution. As time progresses and the sample is heated, a step appears in the Pt signal, indicating the formation of the first phase Pt₂Si. At longer times a second step appears, indicating the formation of the second phase PtSi after all the Pt has been consumed. The energy widths of the Pt signals give the thickness of the formed phases. The heights of the Pt signals, relative to those from Si, give the composition of the phases.

After 60 minutes of annealing, all the Pt has reacted to form Pt₂Si. Almost immediately thereafter the reaction between Pt₂Si and Si to form PtSi starts and after a further 60 minutes all the Pt₂Si has reacted, resulting in a stable PtSi film on Si. The data of silicide thickness versus ramped temperature can be plotted in reduced form in an Arrhenius-like plot to give the activation energy [6, 14].

B1.24.3.2 PD₂SI ON CRSI₂

When a thin film structure of Si(100)/Pd/Cr (see figure B1.24.8(a) is heated to 300 °C, the Pd quickly reacts with the Si to form Pd₂Si (b). Upon further heating the Cr reacts with the Si to form CrSi₂ on top of the Pd₂Si. The required silicon can either be supplied directly by the diffusion of Si atoms from the crystalline substrate (c) or by Pd₂Si dissociation followed by Pd diffusion (d). The motion of a thin Ta marker embedded in the Pd₂Si layer is used to distinguish between these two mechanisms. In (c) there is no movement of the marker relative to the Pd₂Si layer, while in (d) the marker moves towards the Pd₂Si / CrSi₂ interface.

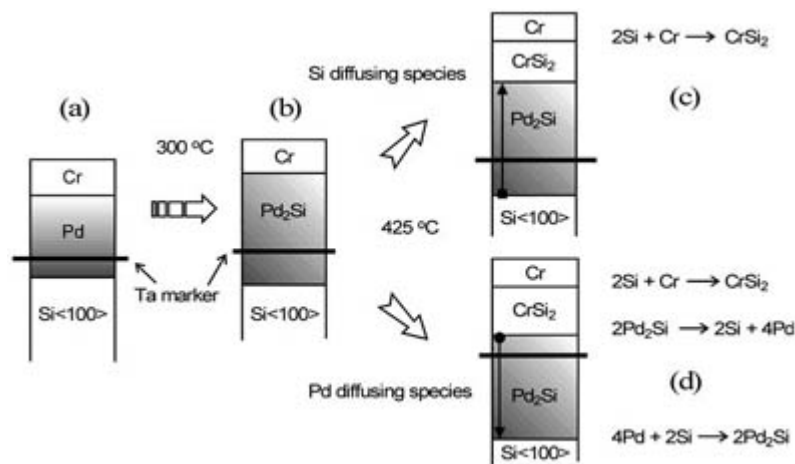


Figure B1.24.8. Schematic diagram of the reaction of Pd/Cr layers on (100) Si with a Ta marker placed inside the Pd layer. When the sample (a) is heated to 300°C, the Pd reacts with the Si to form Pd₂Si (b). Upon further heating Cr reacts with Si to form CrSi₂ on top of Pd₂Si. The required Si can either be supplied directly by the diffusion of Si atoms from the crystalline substrate (c) or by Pd₂Si dissociation followed by Pd diffusion (d). The motion of a thin Ta marker, embedded in the Pd₂Si layer, is used to distinguish between these two mechanisms. In (c) there is no movement of the marker relative to the Pd₂Si layer, while in (d) the marker moves towards the Pd₂Si/CrSi₂ interface.

In figure B1.24.9 the *in situ*, real-time, RBS spectrum of the formation of CrSi₂ on Pd₂Si at 425°C is shown. The Ta marker embedded in the Pd₂Si layer shifts to lower energies during CrSi₂ formation in agreement with the prediction for the case of Si diffusion (c). In the figure, the element from which backscattering has taken place has been underlined [14].

-12-

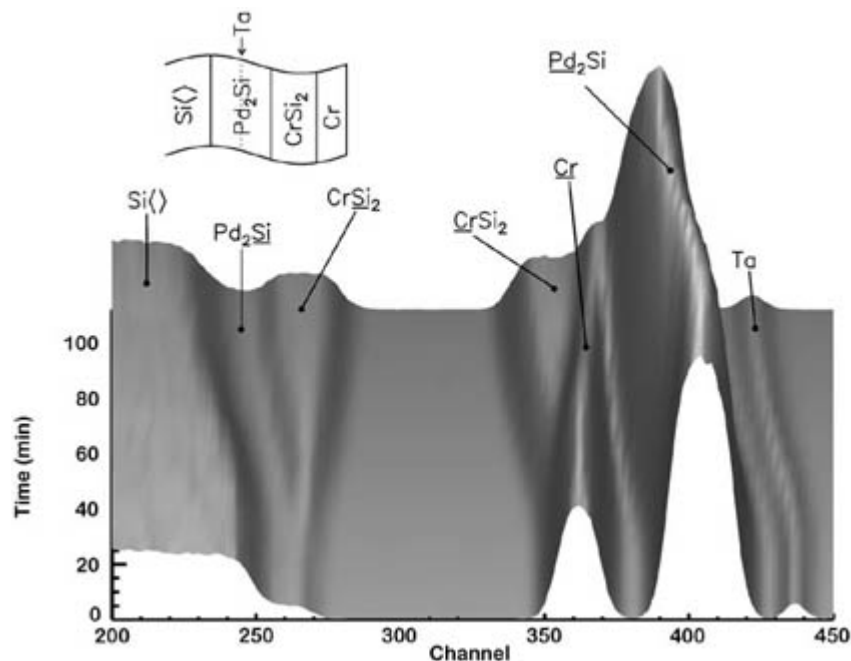


Figure B1.24.9. *In situ*, real-time, backscattering spectrum of the formation of CrSi₂ on Pd₂Si at 425°C. The structure is shown in the inset. The elements underlined represent the origin of the signal; the Si signal in CrSi₂ is around channel 260 and the Cr signal from CrSi₂ appears around channel 350. The Ta marker embedded in the Pd₂Si layer shifts to lower energy during CrSi₂ formation indicating that Si diffusion through

the Pd₂Si layer has occurred as indicated in diagram (c) of [figure B1.24.8](#).

B1.24.4 CHANNELLING

If the sample is mounted on a goniometer so that the crystal axis or planes of a single crystal sample, such as silicon, are aligned within about 0.1 or 0.5 degrees, the crystal lattice can steer the trajectories of incident ions penetrating the crystal [15, 16]. This steering of the incident energetic beam is known as ‘channelling’ as the atomic rows and planes are guides that steer the energetic ions along the channels between rows and planes. The channelled ions do not closely approach the lattice atoms with the result that the backscattering yield can be reduced 100-fold (an aligned spectrum compared to that when the incident ions are misaligned from the lattice atoms gives a random spectrum). Channelling measurements can determine the amount of lattice disorder in which displaced atoms are located within the channels and hence accessible to backscattering collisions with the channelled ions. Channelling can also be used to measure the number of impurity atoms located sufficiently far from substitutional lattice sites that they are accessible to backscattering from the channelled ions.

Channelling phenomena were studied before Rutherford backscattering was developed as a routine analytical tool. Channelling phenomena are also important in ion implantation, where the incident ions can be steered along the lattice planes and rows. Channelling leads to a deep penetration of the incident ions to depths below that found in the normal, near Gaussian, depth distributions characterized by non-channelled energetic ions. Even today, implanted channelled

-13-

ions are of concern when one attempts to form shallow junctions in ion implantation of integrated circuit structures. Channelling effects can be overcome if the silicon crystal is amorphized by a prior implantation of silicon or germanium atoms.

Figure B1.24.10 shows schematically a random and aligned spectrum for MeV helium ions incident on silicon. The aligned spectrum is characterized by a peak at the high energy end of the spectrum. The peak represents ions scattered from the outermost layer of atoms directly exposed to the incident beam. This peak is called the ‘surface peak’. Behind the surface peak, at lower energies, the aligned spectrum drops to a value of 1/40th of the silicon random spectrum indicating that nearly 98% of the incident ions are channelled and do not make close impact collisions with the lattice atoms. The rise in the aligned spectrum at lower energies represent the ions that are deflected from the steering by the lattice atoms and can then collide in close impact collisions with the lattice atoms and hence directly contribute to the backscattering spectra.

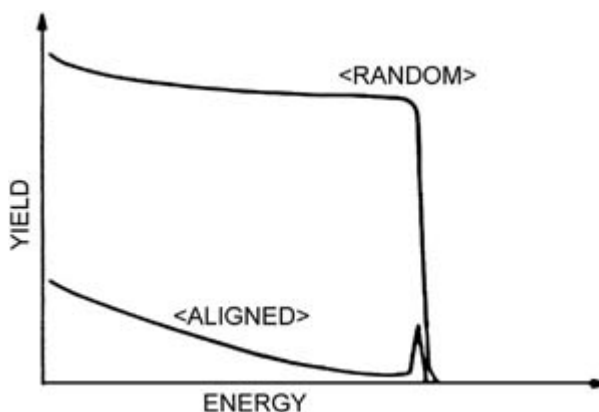


Figure B1.24.10. Random and aligned (channelled) backscattering spectrum from a single crystal sample of silicon. The aligned spectrum has a peak at the high energy end of the Si signal. This peak represents helium

ions scattered from the outer layers of Si that are exposed to the incident beam. The yield behind the peak is 1/40 th of the random yield because the Si atoms are shielded from close encounter elastic collisions from the ion beam that is channelled along the axial rows of the Si crystal.

The application of channelling to Rutherford backscattering spectrometry is used to determine the amount of damage in ion-implanted crystal and the lattice location of ion-implanted dopant atoms. One of the important contributions of channelling to integrated circuit technology is the analysis of amorphous layer formation during ion implantation and its subsequent reanneal at temperatures near 550°C, approximately half the melting temperature of silicon (1414°C). [Figure B1.24.11](#) shows a channelling spectrum in a silicon sample, where the outer 4200 Å of the silicon were converted into an amorphous layer by implantation of silicon atoms at liquid nitrogen temperatures [17]. In the spectrum of the as-implanted sample, marked '0 minutes', the yield of the silicon spectra matches that of the random spectra at energies of around 1 to 1.1 MeV. This shows that the implanted amorphous layer has atoms that are displaced from the underlying single crystal silicon. The silicon signal at 0 minutes shows a decrease at around 0.9 MeV. This decrease represents the fraction of channelled ions in the silicon lattice. The yield does not drop to the non-implanted level because the incident helium atoms suffer multiple collisions penetrating through the amorphous layer and their angular distribution is broadened well beyond the critical angle for channelling. The critical angle for channelling of 1 MeV helium ions along the <100> axis of silicon at room temperature is 0.63 degrees. As the sample is

-14-

thermally annealed at temperatures above 500°C, the amorphous layer reorders epitaxially on the silicon substrate. The rear edge of the amorphous spectrum moves towards the surface such that after 30 minutes half of the layer has recrystallized. The yield from the single-crystal silicon behind the implanted layer decreases since fewer of the incident ions suffer multiple collisions sufficient to make their angular distribution exceed that of the critical angle. Finally, after 60 minutes annealing, almost all the implanted layer is recrystallized and one is left with a surface peak slightly greater than that in the non-implanted case.

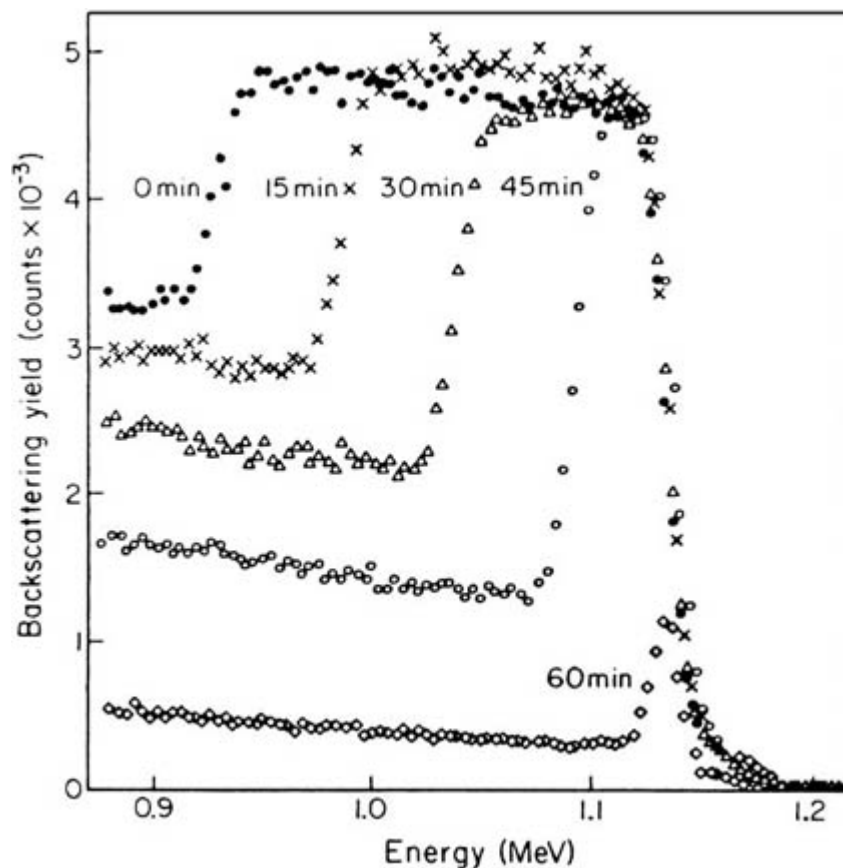


Figure B1.24.11. The backscattering yield from an Si sample that has been implanted with Si atoms to form an amorphous layer. Upon annealing this amorphous layer recrystallizes epitaxially leading to a shift in the amorphous/single-crystal interface towards the surface. The aligned spectra have a step between the amorphous and crystal substrate which shifts towards the surface as the amorphous layer epitaxially recrystallizes on the Si.

Channelling only requires a goniometer to include the effect in the battery of MeV ion beam analysis techniques. It is not as commonly used as the conventional backscattering measurements because the lattice location of implanted atoms and the annealing characteristics of ion implanted materials is now reasonably well established [18]. Channelling is used to analyse epitaxial layers, but even then transmission electron microscopy is used to characterize the defects.

B1.24.5 RESONANCES

At 2 MeV energies the incident helium ion does not penetrate through the barrier around the nucleus. At higher energies and for lighter target atoms such as carbon, nitrogen and oxygen, the helium ion can penetrate and resonances in the cross sections lead to enhanced backscattering yields. This allows one to investigate these target atoms within silicon and even higher mass substrates.

An example of the oxygen resonance cross section is shown in figure B1.24.12 which displays the cross section *versus* energy [19]. The resonance that occurs at 3.04 MeV shows a strong peak. This results in a peak in the backscattering spectra as shown in figure B1.24.13 for 3.05 MeV He⁴ incident on an LaCaMnO₃ film on an LaAlO₃ substrate. In the analysis one increases the energy of the beam to move the resonance to increasing depths.

Carbon also has a resonance in its cross section leading to a 100-fold increase in the backscattering signal. This resonance has been very convenient for analysing 1% carbon in silicon-germanium films. The resonance for nitrogen is not as pronounced and has not been used extensively.

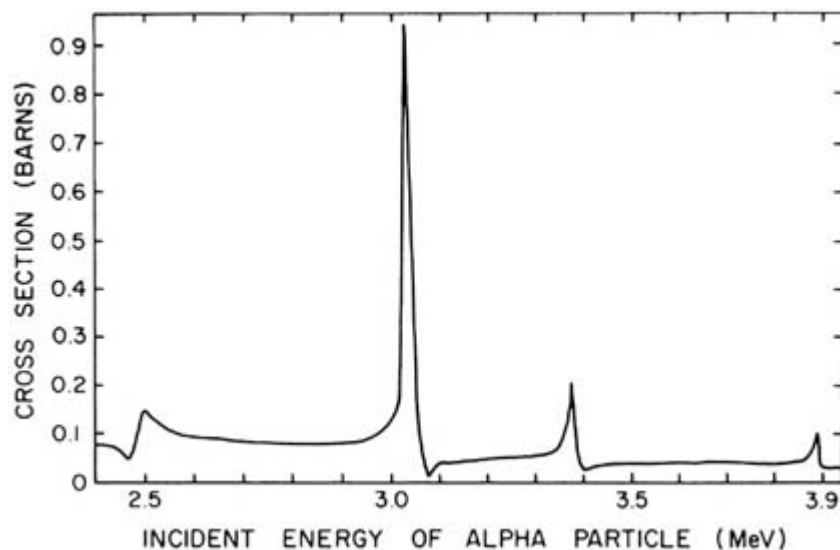


Figure B1.24.12. Elastic cross section of helium ions scattered from oxygen atoms. The pronounced peak in the spectrum around 3.04 MeV represents the resonance scattering cross section that is often used in detection

of oxygen.

-16-

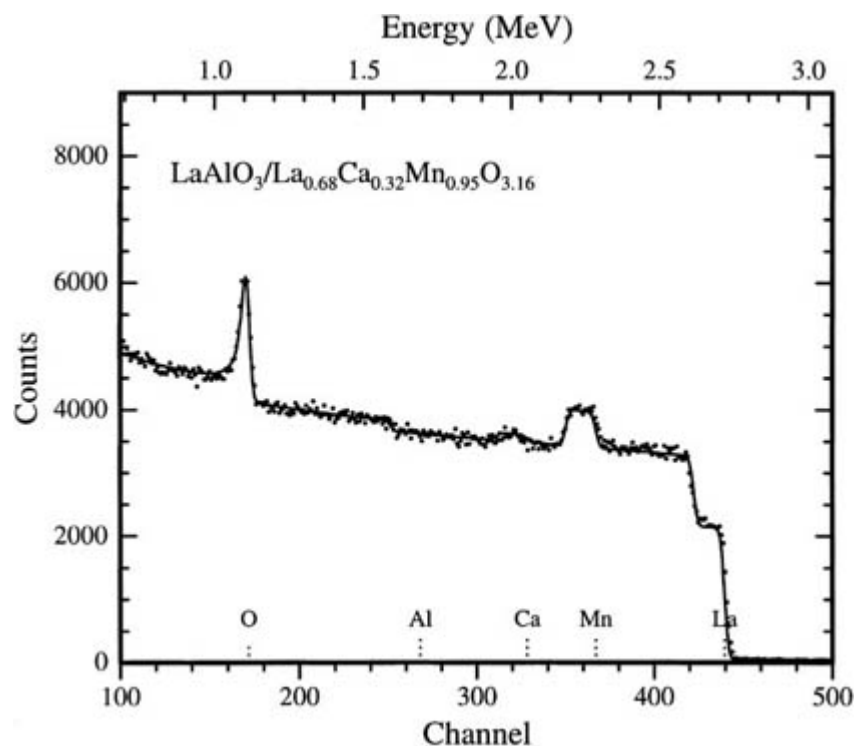


Figure B1.24.13. A thin film of LaCaMnO_3 on an LaAlO_3 substrate is characterized for oxygen content with 3.05 MeV helium ions. The sharp peak in the backscattering signal at channel 160 is due to the resonance in the scattering cross section for oxygen. The solid line is a simulation that includes the resonance scattering cross section and was obtained with RUMP [3]. Data from E B Nyeanchi, National Accelerator Centre, Faure, South Africa.

B1.24.6 PARTICLE-INDUCED X-RAY EMISSION (PIXE)

The PIXE method [20] is based on the spectrometry of the radiation released during the filling of vacancies of inner atomic levels. These vacancies are produced by bombarding a sample with energetic (a few MeV) ions that are normally derived from a high-voltage accelerator. The binding energies of the electrons in the outer layers of the electron shell of an atom are of the order of eV and radiation produced from the rearrangement of electrons in these levels will be in the region of visible wavelength. On the other hand, the binding energies of the inner levels are of the order of keV and radiation produced from processes involving these levels will be in the x-ray region. More importantly, as the electron energy levels of each element are quantized and unique, the measurement of the x-ray energy offers the possibility of determining the presence of a specific element in the sample. Furthermore, the x-ray intensity of a specific energy is proportional to the concentration of the corresponding element in the sample.

The relative simplicity of the method and the penetrative nature of the x-rays, yield a technique that is sensitive to elements with $Z > 10$ down to a few parts per million (ppm) and can be performed quantitatively from first principles. The databases for PIXE analysis programs [21, 22 and 23] are typically so well developed as to include accurate fundamental parameters, allowing the absolute precision of the technique to be around 3% for major elements and 10–20% for trace elements. A major factor in applying the PIXE technique is that the bombarding energy of the

projectiles is a few orders of magnitude more than that of the binding energies of the electrons in the atom and, as the x-rays are produced from the innermost levels, no chemical information is obtained in the process. The advantage of this is that the technique is also not matrix dependent and offers quantitative information regardless of the chemical states of the atoms in the sample. The major application of the technique is the determination of trace element concentrations and, due to the accuracy and non-destructive nature of the technique [24], there are few other techniques that can compete.

The PIXE technique is described schematically in figure B1.24.14. A beam of energetic ions (normally protons of around 3 MeV) is used to eject inner-shell electrons from atoms in a sample. This unstable condition of the atom cannot be maintained and these vacancies are filled by outer-shell electrons. This means that the electrons make a transition in energy in moving from one level to another, and this energy can be released in the form of characteristic x-rays, the energies of which identify the atom. In a competing process, called Auger electron spectroscopy, this energy can also be transferred to another electron that is ejected from the atom and can be detected by an electron detector. Therefore, the step from ionization to x-ray production is not 100% efficient. The x-ray production efficiency is called the fluorescence yield and must be included in the database for quantitative measurements. The x-rays that are emitted from the sample are measured using an energy dispersive detector that has a typical energy resolution around 2.5% (150 eV at 6 keV).

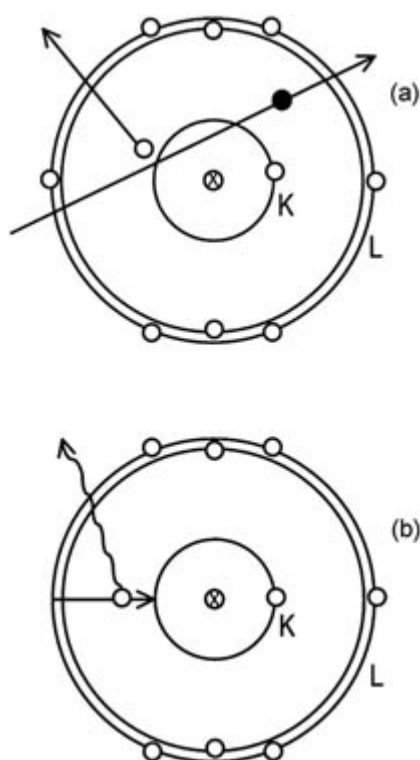


Figure B1.24.14. A schematic diagram of x-ray generation by energetic particle excitation. (a) A beam of energetic ions is used to eject inner-shell electrons from atoms in a sample. (b) These vacancies are filled by outer-shell electrons and the electrons make a transition in energy in moving from one level to another; this energy is released in the form of characteristic x-rays, the energy of which identifies that particular atom. The x-rays that are emitted from the sample are measured with an energy dispersive detector.

By convention, the transitions filling vacancies in the innermost shell are called K x-rays, those filling the next shell are L x-rays, etc. The energies of L x-rays are normally much lower than those of K x-rays, and

similarly M x-rays have much lower energies than L x-rays. Due to the structure of the electron shells, there are naturally more possible transitions yielding L x-rays and even more possibilities of yielding M x-rays; therefore it becomes more complex to measure the higher-order x-rays. Typically, the analytical method is limited to K, L and M x-rays. The limitation of detecting elements with $Z > 10$ is due to the low energies of x-rays from the light elements that are absorbed before reaching the detector. The high yield of low-energy x-rays that originate from the major elements of a sample can be eliminated by a filter in front of the detector. Although the stopping of the bombarding ion is depth dependent, the measured x-ray energy gives no direct indication of the depth at which it was produced, and therefore the technique does not provide depth distribution information.

Typically, PIXE measurements are performed in a vacuum of around 10^{-4} Pa, although they can be performed in air with some limitations. Ion currents needed are typically a few nanoamperes and current is normally not a limiting factor in applying the technique with a particle accelerator. This beam current also normally leads to no significant damage to samples in the process of analysis, offering a non-destructive analytical method sensitive to trace element concentration levels.

An example of a PIXE spectrum is shown in [figure B1.24.15](#) this spectrum was obtained from the analysis of a piece of ivory to establish whether its source could be determined from trace element concentrations [25]. The spectrum shows the contribution from the different elements, also showing the high Ca yield originating from the Ca-rich matrix of the ivory. In this case an 80 μm Al filter was used to filter most of the x-rays from Ca, as they tend to dominate the spectrum. As most interest was focused on the higher-energy part of the spectrum (the higher-energy x-rays are typically not absorbed as much as those of low energy through the same filter) this enabled better sensitivities for the heavier elements to be obtained. To maximize the sensitivity and statistical accuracy, the yields from all the K- or L-shell x-rays from an element are used together to determine the concentration for each element [21]. As can be seen in the figure, the x-ray peaks are situated on a continuous background due to *bremsstrahlung* of the projectiles and secondary electrons and, typically, PIXE software programs perform non-linear iterative procedures to obtain accurate information on peaks and this background [26].

-19-

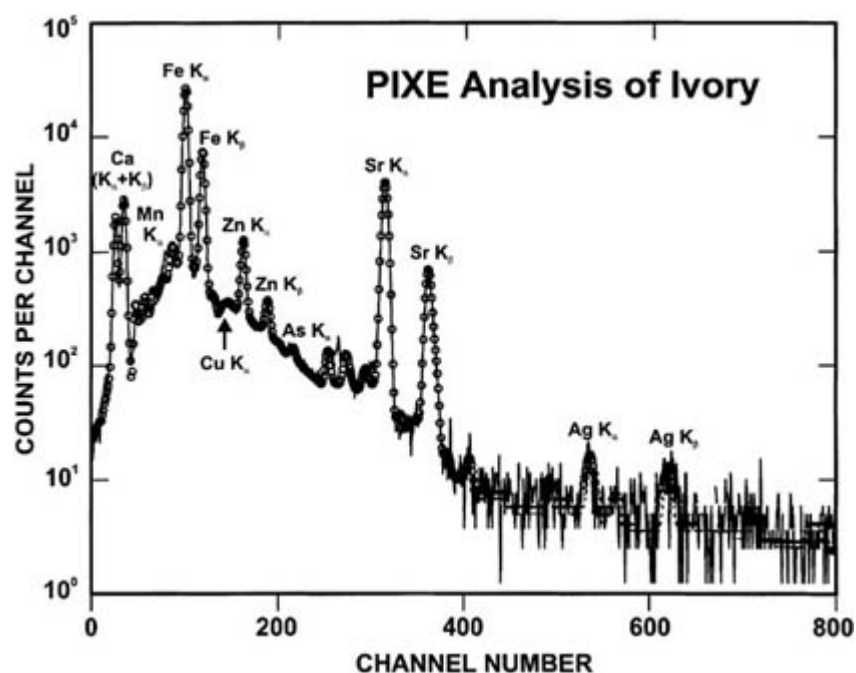


Figure B1.24.15. An example of a PIXE spectrum. This spectrum was obtained from the analysis of a piece of ivory to establish whether the origin of the ivory could be determined from trace element concentrations.

The spectrum shows the contribution from the different elements, also showing the Ca yield originating from the Ca-rich matrix of the ivory. In this case an 80 μm Al filter was used to filter most of the x-rays from Ca, as they tend to dominate the spectrum. As interest was focused on the higher-energy part of the spectrum (the higher-energy x-rays are typically not absorbed as much as those of low energy through the same filter), this enabled better sensitivities for the heavier elements to be obtained. The x-ray peaks are situated on a continuous background of *bremstrahlung* from the projectiles and secondary electrons and, typically, PIXE software programs perform non-linear iterative procedures to obtain accurate information on peaks and this background.

B1.24.7 NUCLEAR MICROPROBE (NMP)

A microprobe employs a focused beams of energetic ions, to provide information on the spatial distribution of elements at concentration levels that range from major elements to a few parts per million [27]. The range of techniques available that allowed depth information plus elemental composition to be obtained could all be used in exactly the same way; it simply became possible to obtain lateral information simultaneously.

-20-

The basis of the nuclear microprobe (NMP) is a source of energetic ions from a particle accelerator. These ions are then focused using either magnetic or electrostatic lenses to a minimum spot size less than $1 \mu\text{m}^2$. The technology of focusing ions is still at the stage where the resolution of the NMP does not compete with the electron microprobe. A set of deflection plates allows movement or scanning of the ion beam and, using computer control, the position of the bombarding beam is known. At each point of irradiation, analytical data are acquired. In this way, the analytical information obtained can be presented as an image. Naturally, the imaging capabilities of the NMP is limited by the sensitivity of each technique used since an entire spectrum must be collected at each 'pixel' in the image. For example, a 128×128 pixel image is equivalent to 16 384 single-point analyses. For this reason the analytical techniques used in imaging are mostly limited to PIXE in the case of trace elements, RBS and forward recoil spectrometry (FRS) for depth information and light elements and nuclear reaction analysis (NRA) to the detection of elements at high concentrations. Naturally the NMP can also be used in a point analysis mode, as for example in the case of geological applications [28]. Here the grain sizes that need to be analysed are often of the order of a few microns and a reasonably small bombarding beam is necessary to limit the analysis to a specific grain.

An example of the application of the PIXE technique using the NMP in the imaging mode [29] is shown in [figure B1.24.16](#). The figures show images of the cross section through a root of the *Phaseolus vulgaris L.* plant. In this case the material was sectioned, freeze-dried and mounted in vacuum for analysis. The scales on the right hand sides of the figures indicate the concentrations of the elements presented in ppm by weight. From the figures it is clear that the transports of the elements through the root are very different not only in the cases of the major elements Ca and K, but also in the case of the trace element Zn. These observations offer a wealth of information that is useful to a botanist studying dynamic processes in plant material.

The quantitative imaging capability of the NMP is one of the major strengths of the technique. The advanced state of the databases available for PIXE [21, 22 and 23] allows also for the analysis of layered samples as, for example, in studying non-destructively the elemental composition of fluid inclusions in geological samples.

The application of RBS is mostly limited to materials applications, where concentrations of elements are fairly high. RBS is specifically well suited to the study of thin film structures. The NMP is useful in studying lateral inhomogeneities in these layers [30] as, for example, in cases where the solid state reaction of elements in the surface layers occur at specific locations on the surfaces. Other aspects, such as lateral diffusion, can also be studied in three-dimensions.

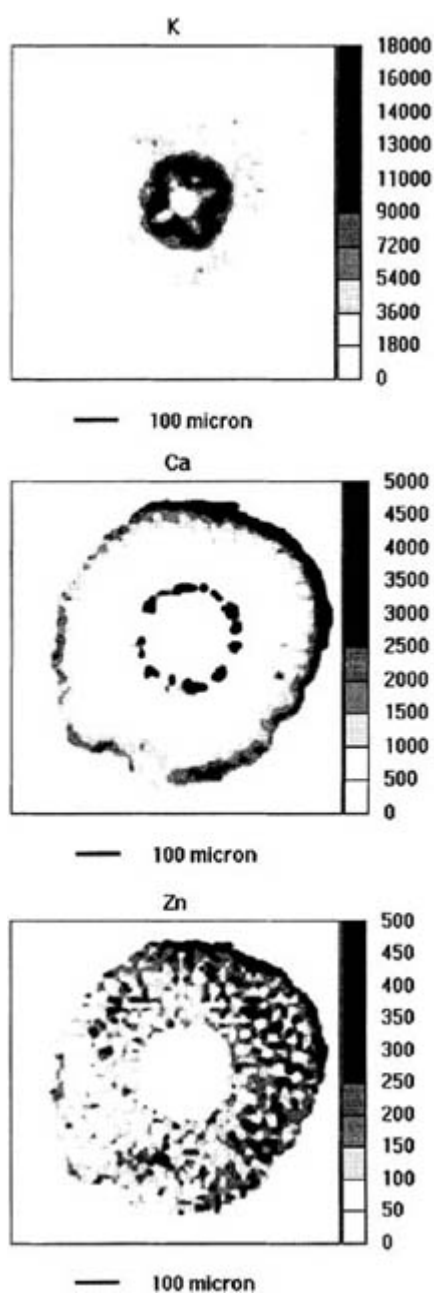


Figure B1.24.16. An example of the application of the PIXE technique using the NMP in the imaging mode. The figures show images of the cross section through a root of the *Phaseolus vulgaris L.* plant. In this case the material was sectioned, freeze-dried and mounted in vacuum for analysis. The scales on the right of the figures indicate the concentrations of the elements in ppm by weight. It is clear that the transports of the elements through the root are very different, not only in the cases of the major elements Ca and K, but also in the case of the trace element Zn.

There are some special techniques that can be used with the NMP, specifically scanning transmission ion microscopy (STIM). In this case the bombarding ion beam is allowed to penetrate through a thin sample and the energies of the transmitted ions are measured. As the energy loss of the ions through the sample is directly proportional to the amount of material traversed, the sample 'thickness' can be imaged with very high

efficiency. The technique is so efficient because every ion is counted. Beam currents of only a few fA are needed, thus permitting an imaging resolution of about 100 nm. The technique is well suited for the study of biological material where cell structure is easily identifiable due to the thickness differences in different parts of cells. An example is shown of STIM measurements of human oral cancer cells in figure B1.24.17. The different images indicate areas of different thickness, starting from thin to thick areas. The technique offers a ‘thickness’ scan through the sample and, in this case, the cell walls of one specific cell can be seen in the areas dominated by thicker structures. There is relatively little material in the inner areas of the cell.

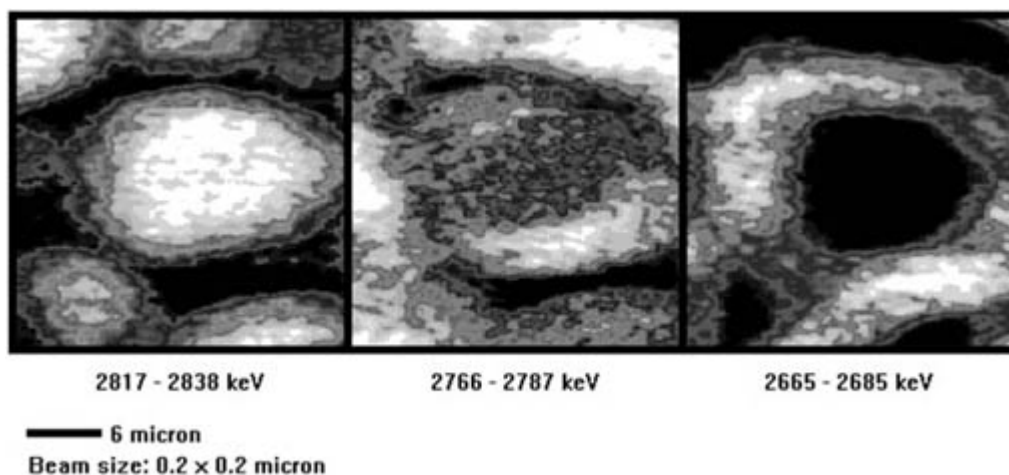


Figure B1.24.17. An example of scanning transmission ion microscopy (STIM) measurements of a human oral cancer cell. The different images indicate different windows in the energy of transmitted helium ions as indicated in the figure. White indicate areas of high counts. The technique offers a ‘thickness’ scan through the sample, and, in this case, the cell walls of one specific cell can be seen in the areas dominated by thicker structures (data from C A Pineda, National Accelerator Centre, Faure, South Africa).

Another special application of the NMP is the measurement of single event upset (SEU) in memory structures of computer chips [31]. In this technique, a single ion is directed onto a part of the memory structure, with a subsequent measurement of whether the memory bit was changed by the ion impact. In this way, the radiation hardness of different parts of the memory can be imaged. This information is valuable for production of components for space applications, where devices are subjected to high fluxes of ionizing radiation. A modern trend is also to study SEUs in living biological material to detect structures susceptible to radiation damage. A similar technique is the study of ion-beam-induced charge (IBIC) collection [32] from p–n junctions in semiconductor material. In this case, the ion beam is directed onto a p–n junction and the current flowing through the junction is measured. By rastering the beam an image can be obtained of the quality of these junctions.

B1.24.8 FORWARD RECOIL SPECTROMETRY (FRS)

Forward recoil spectrometry (FRS) [33], also known as elastic recoil detection analysis (ERDA), is fundamentally the same as RBS with the incident ion hitting the nucleus of one of the atoms in the sample in an elastic collision. In this case, however, the recoiling nucleus is detected, not the scattered incident ion. RBS and FRS are near-perfect complementary techniques, with RBS sensitive to high-Z elements, especially in the presence of low-Z elements. In contrast, FRS is sensitive to light elements and is used routinely in the detection of H at sensitivities not attainable with other techniques [34]. As the technique is also based on an incoming ion that is slowed down on its inward path and an outgoing nucleus that is slowed down in a similar fashion, depth information is obtained for the elements detected.

The analytically important parameters in FRS are exactly those of RBS. Naturally, the target nucleus can only recoil in the forward direction and, therefore, thick targets must be bombarded at an oblique angle to allow detection of the recoil. Thin targets allow the recoiling nucleus to be transmitted through the target. In the case of thick targets, the incident ion also has a high probability of scattering from the target into the detector. It is common that a filter is applied in front of the detector to remove scattered projectiles. This is possible because the projectile has a higher Z and lower energy and can be stopped while allowing the recoils through to the detector. Because of the kinematics as well as straggling, the depth resolution is somewhat worse than that obtainable with RBS. A simple schematic of the experimental setup for FRS is shown in figure B1.24.18.

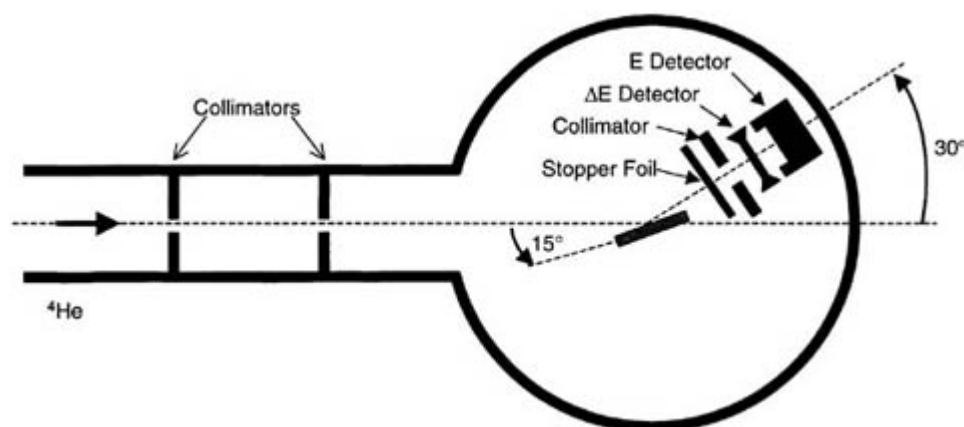


Figure B1.24.18. A simple schematic diagram of the experimental setup for FRS. The most common use of the technique is that of hydrogen detection using ^4He of a few MeV with the recoils being detected at 30° with respect to the beam direction and using a stopper foil to keep ^4He from hitting the detector. This set-up can be generalized to include an energy loss (ΔE) detector in front of the detector thus allowing, in one experiment, the separation of signals due to hydrogen, deuterium and tritium.

The most common use of FRS is the detection of H using ^4He of a few MeV, with the recoils being detected at 30° with respect to the beam direction and a stopper foil to keep ^4He from hitting the detector. This set-up can be generalized to include an energy loss (ΔE) detector in front of the detector, thus allowing the separation of signals from H, D and T in one experiment. The result of such an experiment [35] is shown in figure B1.24.19 where a sample was analysed for hydrogen, deuterium and tritium content using a 3.8 MeV ^4He beam, and detecting the recoils at 30° with a $13.6 \mu\text{m}$ ΔE detector. The three-dimensional graph shows a plot of the counts obtained versus ΔE on the one axis and the energy measured (with a surface barrier detector) after passing through the ΔE detector on the other axis. The traces due to the three isotopes of hydrogen can clearly be seen, with the edges at high E corresponding to the surface of the sample. The shape of the traces from the surface to lower energies are indicative of the depth distribution of isotopes in the sample.

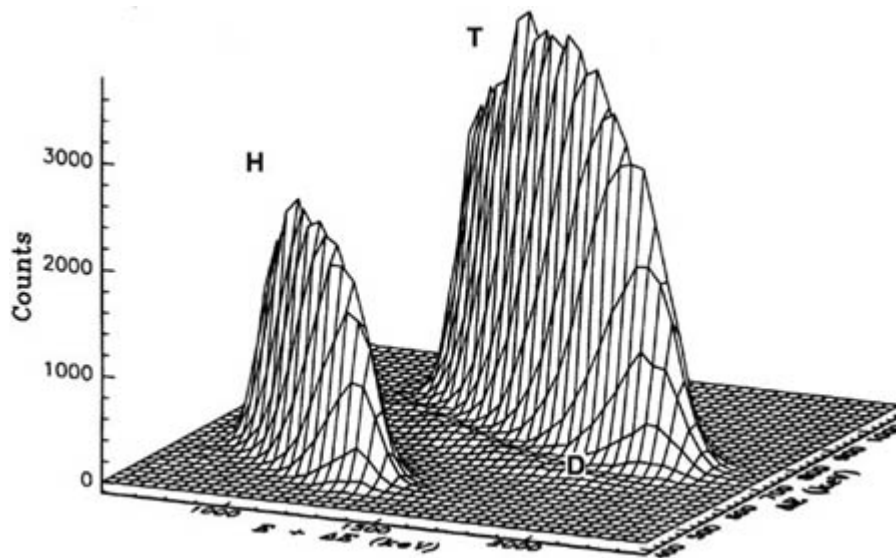


Figure B1.24.19. The FRS result of an experiment where a sample was analysed for hydrogen, deuterium and tritium content using a 3.8 MeV ^4He beam, detecting the recoils at 30° with a $13.6\ \mu\text{m}$ ΔE detector. The three-dimensional graph shows a plot of the counts obtained versus ΔE on the one axis and the total energy ($E + \Delta E$) on the other axis. The traces due to the three isotopes of hydrogen can clearly be seen, with the edges at high E corresponding to the surface of the sample. The shape of the traces from the surface to lower energies are indicative of the depth distribution of these isotopes in the sample.

The most advanced applications of the FRS technique employ high-energy (some tens of MeV) heavy ions, such as Cl and I [36]. In this case a number of nuclei lighter than the projectile can be detected. A detector that can separate different nuclei is required. A two-stage detector is used in which either the time of flight or the energy loss ΔE is determined together with the energy, thus allowing the separation of nuclei with different mass (and Z in the case of ΔE).

An example of such a measurement is shown in [figure B1.24.20](#) where a $\Delta E-E$ detector telescope was used to discriminate between different elements. When using heavy ions as incident particles in the analysis of surface layers, care must be taken not to damage the surface.

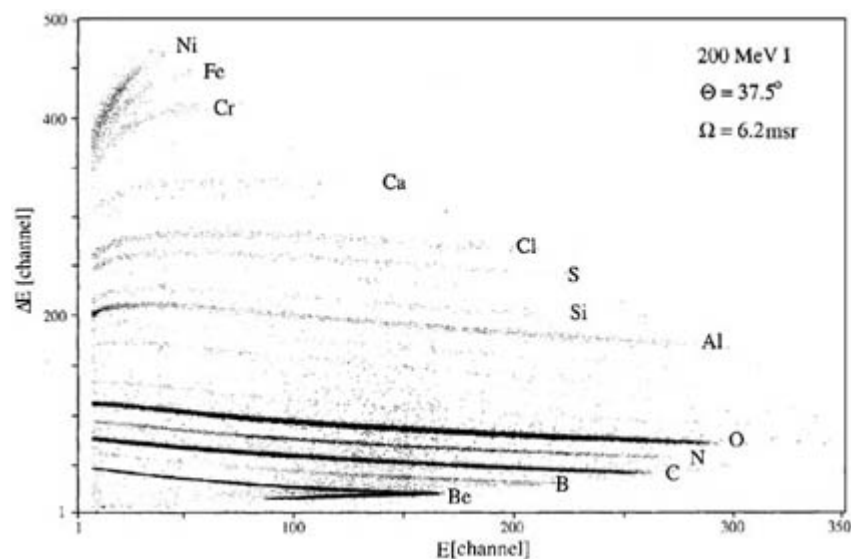


Figure B1.24.20. An example of a heavy-ion (iodine) FRS measurement where a large $\Delta E-E$ detector

telescope was used to enable the discrimination of different elements. The plot shows counts (as intensity plot) versus ΔE and E . The sample analysed was graphite introduced in an experimental nuclear fusion device (tokamak). In this device, a plasma causes different elements to be deposited on the surface of the sample. These elements were determined using a 200 MeV I beam with the detector telescope at 37.5° with respect to the incident beam. It is clear that all the elements from Ni down to Be can be detected in one experiment. The starting points of the traces at high energies indicate concentration of the elements at or near the surface. The trend of the lines as a function of E indicates the concentrations of the elements as a function of depth.

REFERENCES

- [1] Feldman L C and Mayer J W 1986 *Fundamentals of Surface and Thin Film Analysis* (Amsterdam: Elsevier)
- [2] Tesmer J R and Nastasi M (eds) 1995 *Handbook of Modern Ion Beam Materials Analysis* (Pittsburgh, PA: Materials Research Society)
- [3] Doolittle L R 1986 A semiautomatic algorithm for Rutherford backscattering analysis *Nucl. Instrum. Methods B* **15** 227
- [4] Saarihahti J and Rauhala E 1992 Interactive personal-computer data analysis of ion backscattering spectra *Nucl. Instrum. Methods B* **64** 734
- [5] Leavitt J A, McIntyre L C Jr and Weller M R 1995 Backscattering spectrometry *Handbook of Modern Ion Beam Materials Analysis* ed J R Tesmer and M Nastasi (Pittsburgh, PA: Materials Research Society) ch 4
- [6] Theron C C 1997 In situ, real-time characterization of solid-state reaction in thin films *PhD Thesis* University of Stellenbosch
- [7] Kissinger H E 1957 Reaction kinetics in differential thermal analysis *Anal. Chem.* **29** 1702
- [8] Mittemeijer E J, Cheng L, der Schaaf P J V, Brakman C M and Korevaar B M 1988 Analysis of nonisothermal transformation kinetics; tempering of iron-carbon and iron-nitrogen martensites *Metall. Trans. A* **19** 925

- [9] Mittemeijer E J, Gent A V and der Schaaf P J V 1986 Analysis of transformation kinetics by nonisothermal dilatometry *Metall. Trans. A* **17** 1441
- [10] Colgan E G 1995 Activation energy for Pt₂Si and PtSi formation measured over a wide range of ramp rates *J. Mater. Res.* **10** 1953
- [11] Colgan E G 1996 Activation energy for Ni₂Si and NiSi formation measured over a wide range of ramp rates *Thin Solid Films* **279** 193
- [12] Kirkendall E O 1942 Diffusion of zinc in α -brass *Trans. Metall. Soc. AIME* **147** 104
- [13] Smigelskas A D and Kirkendall E O 1947 Zn diffusion in α -brass *Trans. Metall. Soc. AIME* **171** 130
- [14] Theron C C, Mars J A, Churms C L, Farmer J and Pretorius R 1998 In situ, real-time RBS measurement of solid state reaction in thin films *Nucl. Instrum. Methods B* **139** 213
- [15] Feldman L C, Mayer J W and Picraux S T 1982 *Materials Analysis by Ion Channelling* (New York: Academic)
- [16] Swanson M L 1995 Channelling *Handbook of Modern Ion Beam Materials Analysis* ed J R Tesmer and M Nastasi (Pittsburgh, PA: Materials Research Society) ch 10, p 231
- [17] Csepregi L, Kennedy E F, Gallagher T J, Mayer J W and Sigmon T W 1978 Substrate orientation dependence of the epitaxial regrowth rate from Si-implanted amorphous Si *J. Appl. Phys.* **49** 3906

- [18] Rimini E 1995 *Ion Implantation: Basics to Device Fabrication* (Boston, MA: Kluwer)
- [19] Cox R P, Leavitt J A and McIntyre L C Jr 1995 Non-Rutherford elastic backscattering cross sections *Handbook of Modern Ion Beam Materials Analysis* ed J R Tesmer and M Nastasi (Pittsburgh, PA: Materials Research Society) ch A7, p 481
- [20] Johannson T B, Akselsson K R and Johannson S A E 1970 X-ray analysis: elemental trace analysis at the 10^{-12} g level *Nucl. Instrum. Methods* **84** 141
- [21] Ryan C G, Cousins D R, Sie S H, Griffin W L, Suter G F and Clayton E 1990 Quantitative PIXE microanalysis of geological material using the CSIRO proton microprobe *Nucl. Instrum. Methods B* **47** 55
- [22] Ryan C G and Jamieson D N 1993 Dynamic analysis—online quantitative PIXE microanalysis and its use in overlap-resolved elemental mapping *Nucl. Instrum. Methods B* **77** 203
- [23] Maxwell J A, Campbell J L and Teesdale W J 1989 The Guelph PIXE software package *Nucl. Instrum. Methods B* **43** 218
- [24] Campbell J L, Russell S B, Faiq S, Sculte C W, Ollerhead R W and Gingerich R R 1981 Optimization of PIXE sensitivity for biomedical applications *Nucl. Instrum. Methods* **181** 285
- [25] Prozesky V M, Raubenheimer E R, van Heerden W F P, Grotepass W P, Przybyłowicz W J, Pineda C A and Swart R 1995 Trace element concentration and distribution in ivory *Nucl. Instrum. Methods B* **104** 638
- [26] Vekemans B, Janssens K, Vincze L, Adams F and van Espen F 1995 Comparison of several background compensation methods useful for evaluation of energy-dispersive x-ray fluorescence spectra *Spectrochim. Acta B* **50** 149
- [27] Watt F, Grime G W and Hilger A *Principles and Applications of High-Energy Ion Microbeams* (Bristol: Institute of Physics)
- [28] Ryan C G 1995 The nuclear microprobe as a probe of earth structure and geological processes *Nucl. Instrum. Methods B* **104** 69

-27-

- [29] van As J A, Jooste J H, Mesjasz-Przybyłowicz J and Przybyłowicz W J 1995 Nuclear microprobe studies of the mechanisms of Zn uptake *Phaseolus vulgaris* L., *Microsc. Soc. of Southern Africa—Proceedings* **25** 34
- [30] de Waal H S, Pretorius R, Prozesky V M and Churms C L 1997 The study of voids in the Au–Al thin-film system using the nuclear microprobe *Nucl. Instrum. Methods B* **130** 722
- [31] Breese M B H, Jamieson D N and King P J C 1996 *Materials Analysis with a Nuclear Microprobe* (New York: Wiley)
- [32] Jamieson D N 1998 Structural and electrical characterization of semiconductor materials using a nuclear microprobe *Nucl. Instrum. Methods B* **136–138** 1
- [33] Tirira J, Serruys Y and Trocellier P 1996 *Forward Recoil Spectrometry—Applications to Hydrogen Determination in Solids* (New York: Plenum)
- [34] Sweeney R J, Prozesky V M, Churms C L, Padayachee J and Springhorn K 1998 Application of a ΔE – E telescope for sensitive ERDA measurement of hydrogen *Nucl. Instrum. Methods B* **136–138** 685
- [35] Prozesky V M, Churms C L, Pilcher J V and Springhorn K A 1994 ERDA measurement of hydrogen isotopes with a ΔE – E telescope *Nucl. Instrum. Methods B* **84** 373
- [36] Behrisch R, Prozesky V M, Huber H and Assmann W 1996 Hydrogen desorption induced by heavy-ions during surface analysis with ERDA *Nucl. Instrum. Methods B* **118** 262
-

FURTHER READING

Chu W K, Mayer J W and Nicolet M-A 1978 *Backscattering Spectrometry* (New York: Academic)

Comprehensive and detailed coverage of Rutherford backscattering and channelling.

Feldman L C, Mayer J W and Picraux S T 1982 *Materials Analysis by Ion Channelling* (New York: Academic)

Fundamental treatment suitable for both graduate students and researchers.

Feldman L C and Mayer J W 1986 *Fundamentals of Surface and Thin Film Analysis* (New York: Elsevier)

General coverage of analytical techniques suitable for undergraduates, graduate students and researchers.

Tesmer J R and Nastasi M (ed) 1995 *Handbook of Modern Ion Beam Materials Analysis* (Pittsburgh, PA: Materials Research Society)

This comprehensive handbook covers all aspects of ion beam analysis from energy loss to radiological safety. It is valuable for the researcher.

Johanssen S A E and Campbell J L 1988 *PIXE: A Novel Technique for Elemental Analysis* (Chichester: Wiley)

Covers PIXE in detail and is a good reference for graduate students and researchers.

Tirira J, Serruys Y and Trocellier P 1996 *Forward Recoil Spectrometry* (New York: Plenum)

A clear description of applications to hydrogen determination in solids for students and researchers.

-28-

-1-

B1.25 Surface chemical characterization

L Coulier and J W Niemantsverdriet

B1.25.1 INTRODUCTION

Chemical characterization of surfaces plays an important role in various fields of physics and chemistry, e.g. catalysis, polymers, metallurgy and organic chemistry. This section briefly describes the concepts and a few examples of the techniques that are most frequently used for chemical surface characterization, which are x-ray photoelectron spectroscopy (XPS), Auger electron spectroscopy (AES), ultraviolet photoelectron spectroscopy (UPS), secondary ion mass spectrometry (SIMS), temperature programmed desorption (TPD) and electron energy loss spectroscopy (EELS), respectively. We have tried to give examples in a broad range of fields. References to more extensive treatments of these techniques and others are given at the end of the section, see '[Further reading](#)'.

Although the techniques described undoubtedly provide valuable results on various materials, the most useful information almost always comes from a combination of several (chemical and physical) surface characterization techniques. Table B1.25.1 gives a short overview of the techniques described in this chapter.

B1.25.2 ELECTRON SPECTROSCOPY (XPS, AES, UPS)

B1.25.2.1 X-RAY PHOTOELECTRON SPECTROSCOPY (XPS)

X-ray photoelectron spectroscopy (XPS) is among the most frequently used surface chemical characterization techniques. Several excellent books on XPS are available [1, 2, 3, 4, 5, 6 and 7]. XPS is based on the photoelectric effect: an atom absorbs a photon of energy $h\nu$ from an x-ray source; next, a core or valence electron with binding energy E_b is ejected with kinetic energy (figure B1.25.1):

$$E_k = h\nu - E_b - \varphi \quad (\text{B1.25.1})$$

where E_k is the kinetic energy of the photoelectron, h is Planck's constant, ν is the frequency of the exciting radiation, E_b is the binding energy of the photoelectron with respect to the Fermi level of the sample and φ is the work function of the spectrometer. Routinely used x-ray sources are **Mg K α** ($h\nu = 1253.6$ eV) and **Al K α** ($h\nu = 1486.3$ eV). In XPS one measures the intensity of photoelectrons $N(E)$ as a function of their kinetic energy E_k . The XPS spectrum is a plot of $N(E)$ versus E_b ($= h\nu - E_k - \varphi$).

-2-

Table B1.25.1 Overview of the surface characterization techniques described in this chapter.

Acronym	Full name	Principle of measurement	Key information
XPS	X-ray photoelectron spectroscopy	Absorption of a photon by an atom, followed by the ejection of a core or valence electron with a characteristic binding energy.	Composition, oxidation state, dispersion
AES	Auger electron spectroscopy	After the ejection of an electron by absorption of a photon, an atom stays behind as an unstable ion, which relaxes by filling the hole with an electron from a higher shell. The energy released by this transition is taken up by another electron, the Auger electron, which leaves the sample with an element-specific kinetic energy.	Surface composition, depth profiles
UPS	UV photoelectron spectroscopy	Absorption of UV light by an atom, after which a valence electron is ejected.	Chemical bonding, work function
SIMS	Secondary ion mass spectroscopy	A beam of low-energy ions impinges on a surface, penetrates the sample and loses energy in a series of inelastic collisions with the target atoms leading to emission of secondary ions.	Surface composition, reaction mechanism, depth profiles
TPD	Temperature programmed desorption	After pre-adsorption of gases on a surface, the desorption and/or reaction products are measured while the temperature increases linearly with time.	Coverages, kinetic parameters, reaction mechanism
EELS	Electron energy loss spectroscopy	The loss of energy of low-energy electrons due to excitation of lattice vibrations.	Molecular vibrations, reaction mechanism

Photoelectron peaks are labelled according to the quantum numbers of the level from which the electron originates. An electron coming from an orbital with main quantum number n , orbital momentum l (0, 1, 2, 3, ... indicated as s, p, d, f, ...) and spin momentum s (+1/2 or -1/2) is indicated as nl_{l+s} . For every orbital momentum $l > 0$ there are two values of the total momentum: $j = l + 1/2$ and $j = l - 1/2$, each state filled with $2j + 1$ electrons. Hence, most XPS peaks come in doublets and the intensity ratio of the components is $(l + 1)/l$. When the doublet splitting is too small to be observed, the subscript $l + s$ is omitted.

Figure B1.25.2 shows the XPS spectra of two organoplatinum complexes which contain different amounts of chlorine. The spectrum shows the peaks of all elements expected from the compounds, the Pt 4f and 4d doublets (the 4f doublet is unresolved due to the low energy resolution employed for broad energy range scans), Cl 2p and Cl 2s, N 1s and C 1s. However, the C 1s cannot be taken as characteristic for the complex only. All surfaces that have not been cleaned by sputtering or oxidation in the XPS spectrometer contain carbon. The reason is that adsorbed hydrocarbons from the atmosphere give the optimum lowering of the surface free energy and hence, all surfaces are covered by hydrocarbon fragments [9].

-3-

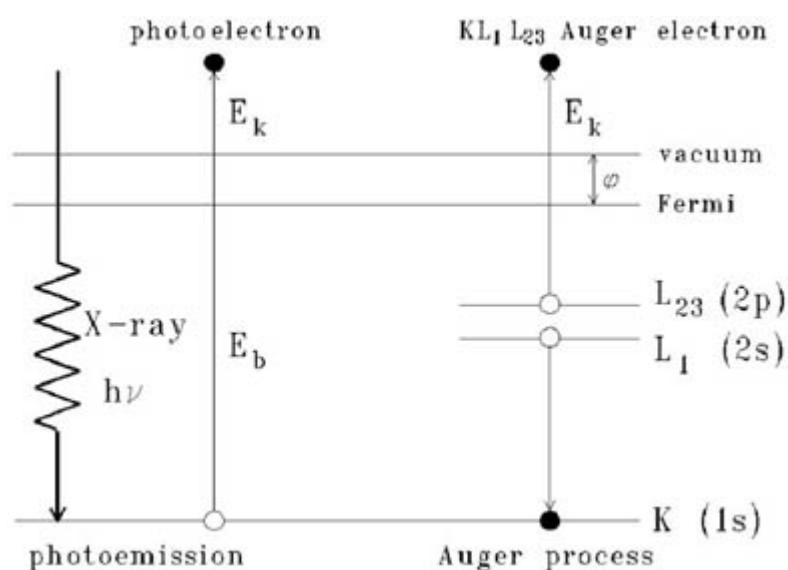


Figure B1.25.1. Photoemission and Auger decay: an atom absorbs an incident x-ray photon with energy $h\nu$ and emits a photoelectron with kinetic energy $E_k = h\nu - E_b$. The excited ion decays either by the indicated Auger process or by x-ray fluorescence.

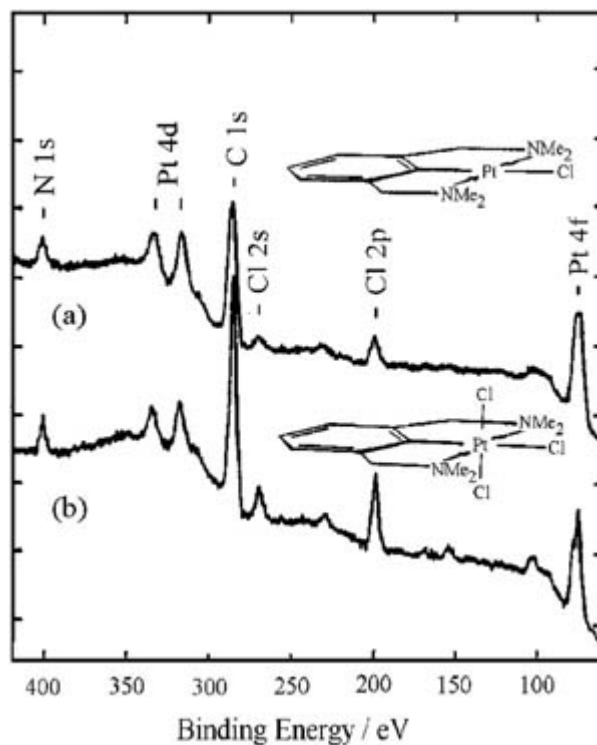


Figure B1.25.2. XPS scans between 0 and 450 eV of two organoplatinum complexes showing peaks due to Pt, Cl, N and C. The C 1s signal represents not only carbon in the compound but also contaminant hydrocarbon fragments, as in any sample. The abbreviation ‘Me’ in the structures stands for CH₃ (courtesy of J C Muijsers, Eindhoven).

Because a set of binding energies is characteristic for an element, XPS can analyse chemical composition. Almost all photoelectrons used in laboratory XPS have kinetic energies in the range of 0.2 to 1.5 keV, and probe the outer layers of the sample. The mean free path of electrons in elemental solids depends on the kinetic energy. Optimum surface sensitivity is achieved with electrons at kinetic energies of 50–250 eV, where about 50% of the electrons come from the outermost layer.

Binding energies are not only element specific but contain chemical information as well: the energy levels of core electrons depend on the chemical state of the atom. Chemical shifts are typically in the range 0–3 eV. In general, the binding energy increases with increasing oxidation state and, for a fixed oxidation state, with the electronegativity of the ligands. Figure B1.25.3 illustrates the sensitivity of XPS binding energy to oxidation states for platinum in metal, and in the two organoplatinum complexes of [Figure B1.25.2](#). The Pt 4f_{7/2} peak of the metal comes at 71.0 eV, that of the complex where Pt has one Cl ligand at 72.0 eV, characteristic of Pt²⁺, while the binding energy of the Pt⁴⁺ in the complex with three Cl ligands on platinum is again 2 eV higher, 74.4 eV [9]. The binding energy goes up with the oxidation state of the platinum. The reason is that the 74 electrons in the Pt⁴⁺ ion (lower curve) feel a higher attractive force from the nucleus with a positive charge of 78⁺ than the 76 electrons in Pt²⁺ (middle curve) or the 78 in the neutral Pt atom (upper curve).

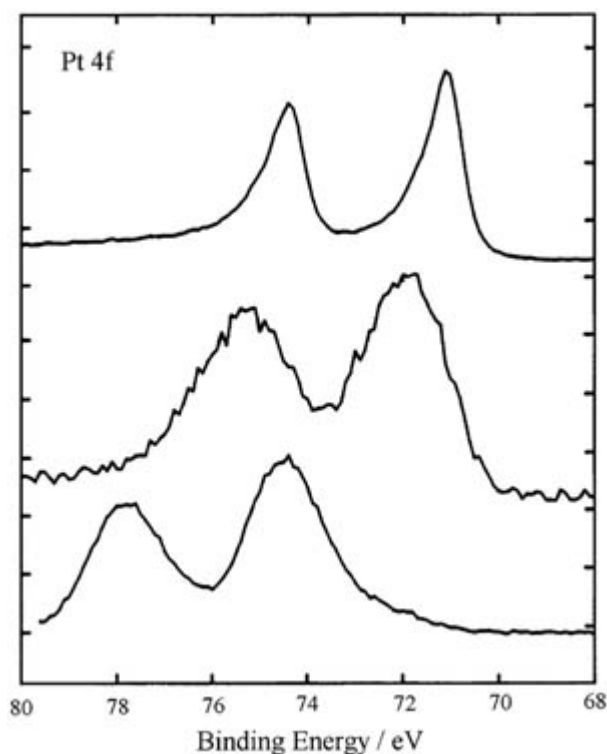


Figure B1.25.3. Pt 4f XPS spectra of platinum metal (top) and of the two organoplatinum compounds (a) and (b), middle and bottom respectively, shown in figure B1.25.2, illustrating that the Pt 4f binding energy reflects the oxidation state of platinum (from [9]).

-5-

Note that XPS measures binding energies. These are not necessarily equal to the energy of the orbitals from which the photoelectron is emitted. The difference is caused by reorganization of the remaining electrons when an electron is removed from an inner shell. Thus, the binding energy of the photoelectron contains both information on the state of the atom before photoionization (the initial state) and on the core-ionized atom left behind after the emission of an electron (the final state) [6]. Fortunately, it is often correct to interpret binding energy shifts in terms of initial state effects.

Determining compositions is possible if the distribution of elements over the outer layers of the sample and the surface morphology is known. Two limiting cases are considered, namely a homogeneous composition throughout the outer layers and an arrangement in which one element covers the other.

For homogeneous mixed samples it is relatively easy to determine the relative concentrations of the various constituents. For two elements one has approximately:

$$n_1/n_2 = (I_1/S_1)/(I_2/S_2) \quad (\text{B1.25.2})$$

where n_1/n_2 is the ratio of elements 1 and 2, I_1, I_2 are the intensities of the peaks of elements 1 and 2 (i.e. the area of the peaks) and S_1, S_2 are atomic sensitivity factors which are tabulated [8].

A more accurate calculation will account for differences in the energy dependent mean free paths of the elements and for the transmission characteristics of the electron analyser (see [7]).

An example in which XPS is used for studying surface compositions and oxidation states is illustrated in

[figure B1.25.2](#) and [figure B1.25.3](#) for two organoplatinum complexes. The samples were prepared for XPS by letting a solution of the complexes in dichloromethane dry on a stainless steel sample stub. The sample should thus be homogeneous and the use of [\(B1.25.2\)](#) permitted. [Figure B1.25.2](#) shows the wide scan up to a binding energy of 450 eV [9]. The figure shows immediately that the Cl peaks in the spectrum of the trichloride complex are about three times as intense as in the spectrum of the compound with one Cl. If we apply [\(B1.25.2\)](#) for the elements Pt, N and Cl, we obtain Pt:N:Cl = 1:1.9:4 for the trichloride complex, close to the true stoichiometry of 1:2:3.

In the case of metal particles distributed on a support material (e.g. supported catalysts), XPS yields information on the dispersion. A higher metal/support intensity ratio (at the same metal content) indicates a better dispersion [3].

Another good example of the application of XPS in a different field of chemistry is shown in [figure B1.25.4](#). This figure shows the C 1s spectrum of a polymer [11]. Four different carbon species can be distinguished. Reference tables indicate that the highest binding energy peak is due to carbon-fluorine species [7, 8]. The other three peaks are attributed (from high to low binding energy) to an ester species, an ether species and hydrocarbon/benzene fragments, respectively [8]. Hence, the carbon XPS spectrum nicely reflects the structure of the polymer. This example is also a nice illustration of the influence of the electronegativity of the ligands on the binding energy of carbon: the binding energy of carbon increases with the electronegativity of the ligands.

-6-

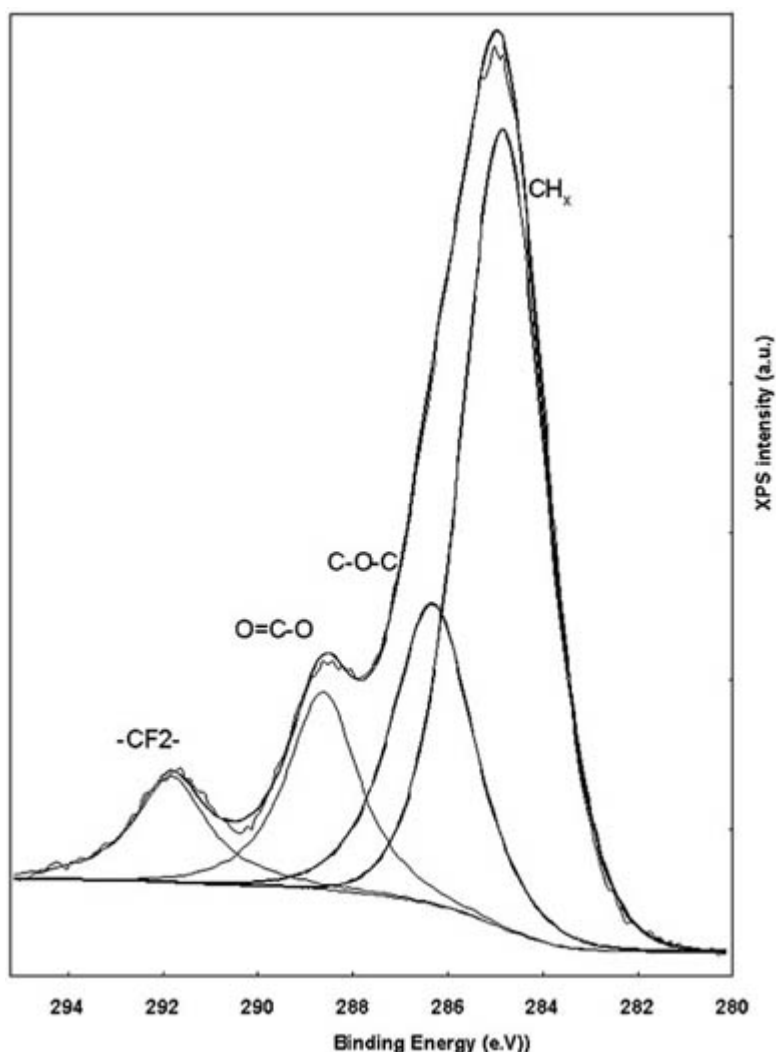


Figure B1.25.4. C 1s XPS spectrum of a polymer, illustrating that the C 1s binding energy is influenced by the chemical environment of the carbon. The spectrum clearly shows four different kinds of carbon, which corresponds well with the structure of the polymer (courtesy of M W G M Verhoeven, Eindhoven).

Owing to the limited escape depth of photoelectrons, the surface sensitivity of XPS can be enhanced by placing the analyser at an angle to the surface normal (the so-called take-off angle of the photoelectrons). This can be used to determine the thickness of homogeneous overlayers on a substrate.

This is demonstrated by the XPS spectra in [figure B1.25.5\(a\)](#) which show the Si 2p spectra of a silicon crystal with a thin (native) oxide layer, measured under take-off angles of 0° and 60° [12]. When the take-off angle is high, relatively more photoelectrons from the oxide surface region reach the analyser and the Si^{4+}/Si intensity ratio increases with increasing angle. [Figure B1.25.5\(b\)](#) shows the intensity ratio as a function of take-off angle, the line being a fit corresponding to a flat, homogeneous oxide layer with a uniform thickness of 2 nm.

-7-

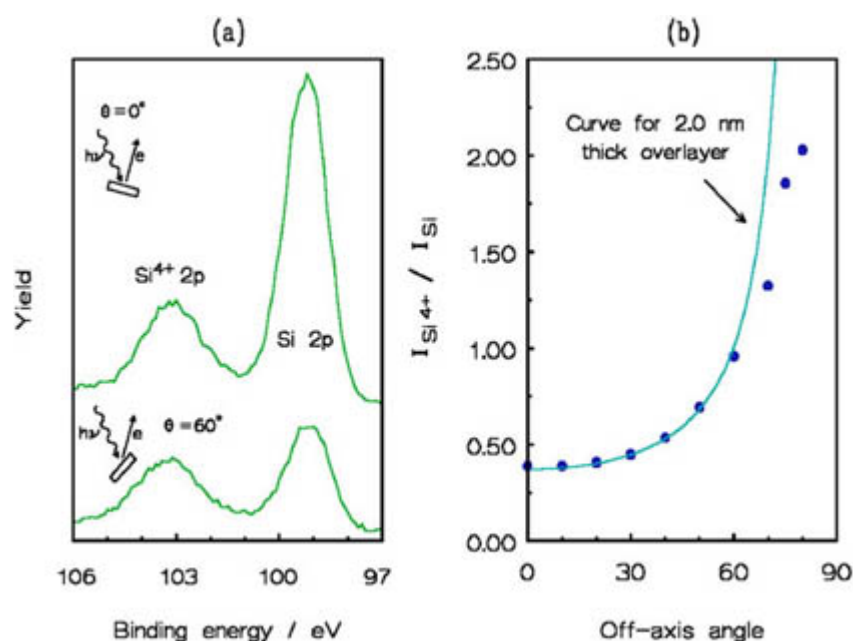


Figure B1.25.5. (a) XPS spectra at take-off angles of 0° and 60° as measured from the surface normal from a silicon crystal with a thin layer of SiO_2 on top. The relative intensity of the oxide signal increases significantly at higher take-off angles, illustrating that the surface sensitivity of XPS increases. (b) Plot of Si^{4+}/Si 2p peak areas as a function of take-off angle. The solid line is a fit which corresponds to an oxide thickness of 2.0 nm (from [12]).

An experimental problem in XPS is that electrically insulating samples may charge up during measurement, due to photoelectrons leaving the sample. Since the sample thereby acquires a positive charge, all XPS peaks in the spectrum shift by the same amount to higher binding energies. More serious than the shift itself is that different parts of the sample may acquire slightly different amounts of charge. This phenomenon, called differential charging, gives rise to broadening of the peaks and degrades the resolution. Correction for charging-induced shifts is made by using the binding energy of a known compound (in most cases one uses the C 1s binding energy of 284.6 eV). Alternatively, in certain circumstances, a low energy electron beam can be used to neutralize the charged surface and eliminate the shift.

Sensitive materials, such as metal salts or organometallic compounds, may decompose during XPS analysis, particularly when a standard x-ray source is used. Apart from the x-rays themselves, heat and electrons from the source may cause damage to the samples. In such cases, a monochromated x-ray source can offer a

solution [9]. Damage is in particular an issue in imaging XPS, where the x-ray intensity is focused in a narrow spot. In this mode, a small hole in front of the analyser entrance enables one to select electrons from an area of a few micrometres, such that an image of the surface composition can be made.

B1.25.2.2 AUGER ELECTRON SPECTROSCOPY (AES)

Auger electron spectroscopy is a powerful technique in the fields of materials and surface science [2, 3 and 4, 7]. In AES, core holes are created by exciting the sample with a beam of electrons. The excited ion relaxes by filling the core hole with an electron from a higher shell. The energy released by this transition is taken up by another electron, the Auger electron, which leaves the sample with an element-specific kinetic energy (figure B1.25.1). The Auger electrons appear as small peaks on a high background of secondary electrons, scattered by the sample from the incident beam. To enhance the visibility of the Auger peaks, spectra are usually presented in the derivative (dN/dE) mode, see figure B1.25.6.

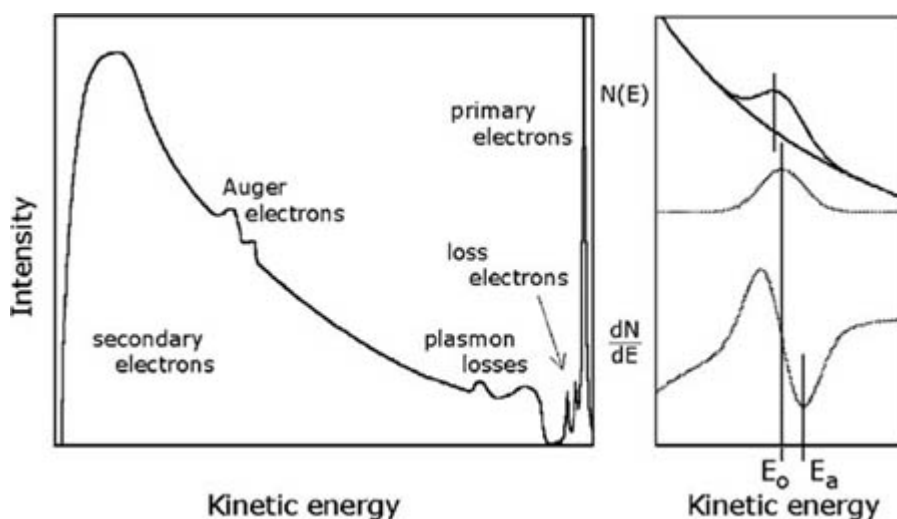


Figure B1.25.6. Energy spectrum of electrons coming off a surface irradiated with a primary electron beam. Electrons have lost energy to vibrations and electronic transitions (loss electrons), to collective excitations of the electron sea (plasmons) and to all kinds of inelastic process (secondary electrons). The element-specific Auger electrons appear as small peaks on an intense background and are more visible in a derivative spectrum.

Auger peaks are labelled according to the x-ray level nomenclature. For example, KL_1L_2 stands for a transition in which the initial core hole in the K shell is filled from the L_1 shell, while the Auger electron is emitted from the L_2 shell. Valence levels are indicated by ‘V’ as in the KVV transitions of carbon or oxygen. The energy of an Auger electron formed in a KLM transition is to a good approximation given by

$$E_{KLM} = E_K - E_L - E_M - \delta E - \varphi \quad (B1.25.3)$$

where E_{KLM} is the kinetic energy of the Auger electron, E_i is the binding energy of an electron in the i shell ($i = K, L, M, \dots$), δE is the energy shift caused by relaxation effects and φ is the work function of the spectrometer. The δE term accounts for the relaxation effect involved in the decay process, which leads to a final state consisting of a heavily excited, doubly ionized atom.

Auger peaks also appear in XPS spectra. In this case, the x-ray ionized atom relaxes by emitting an electron with a specific kinetic energy E_k . One should bear in mind that in XPS the intensity is plotted against the binding energy, so one uses (B1.25.1) to convert to kinetic energy.

The strong point of AES is that it provides a quick measurement of elements in the surface region of conducting samples. For elements having Auger electrons with energies in the range of 100–300 eV where the mean free path of the electrons is close to its minimum, AES is considerably more surface sensitive than XPS.

Auger electron spectroscopy allows for three types of measurement. First, it provides the elemental surface composition of a sample. If the Auger decay process involves valence electrons, one often obtains information on the oxidation state as well, although XPS is certainly the better technique for this purpose. Second, owing to the short data collection times, AES can be combined with sputtering to measure concentrations as a function of depth, see figure B1.25.7. Third, as electron beams are easily collimated and deflected electrostatically, AES can be used to image the composition into a chemical map of the surface (scanning Auger spectroscopy). The best obtained resolution is now around 25 nm [7].

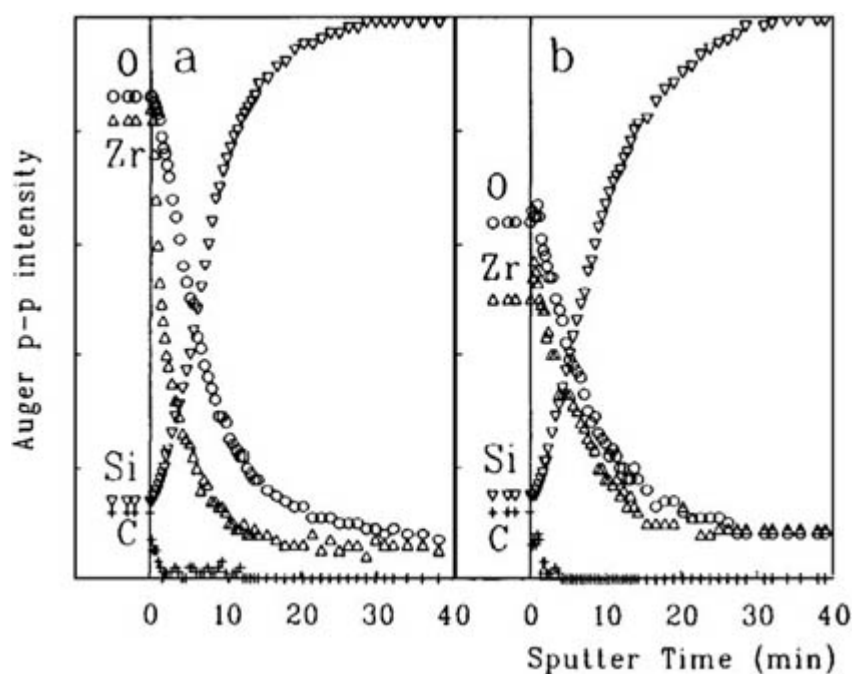


Figure B1.25.7. Auger sputter depth profile of a layered $ZrO_2/SiO_2/Si$ model catalyst. While the sample is continuously bombarded with argon ions that remove the outer layers of the sample, the Auger signals of Zr, O, Si and C are measured as a function of time. The depth profile is a plot of Auger peak intensities against sputter time. The profile indicates that the outer layer of the model catalyst contains carbon. Next Zr and O are sputtered away, but note that oxygen is also present in deeper layers where Zr is absent. The left-hand pattern is characteristic for a layered structure, and confirms that the zirconium is present in a well dispersed layer over the silicon oxide. The right-hand pattern is consistent with the presence of zirconium oxide in larger particles (from [27]).

A disadvantage of AES is that the intense electron beam easily causes damage to sensitive materials (polymers, insulators, adsorbate layers). Charging of insulating samples also causes serious problems.

B1.25.2.3 ULTRAVIOLET PHOTOELECTRON SPECTROSCOPY (UPS)

Ultraviolet photoelectron spectroscopy (UPS) [2, 3 and 4, 6] differs from XPS in that UV light (He I, 21.2 eV; He II, 40.8 eV) is used instead of x-rays. At these low exciting energies, photoemission is limited to valence electrons.

Hence, UPS spectra contain important chemical information. At low binding energies, UPS probes the density of states (DOS) of the valence band (but images it in a distorted way in a convolution with the unoccupied states). At slightly higher binding energies (5–15 eV), occupied molecular orbitals of adsorbed gases may become detectable. UPS also provides a quick measure of the macroscopic work function, ϕ , the energy separation between the Fermi and the vacuum level: $\phi = h\nu - W$, where W is the width of the spectrum. UPS is a surface science technique typically applied to single crystals, the main reason being that all elements contribute peaks to the valence band region. As a result, the UPS spectra of compounds which contain more than two elements are rather complicated.

B1.25.3 SECONDARY ION MASS SPECTROMETRY (SIMS)

Secondary ion mass spectrometry (SIMS) is by far the most sensitive surface technique, but also the most difficult one to quantify. SIMS is very popular in materials research for making concentration depth profiles and chemical maps of the surface. For a more extensive treatment of SIMS the reader is referred to [3] and [14, 15 and 16]. The principle of SIMS is conceptually simple: When a surface is exposed to a beam of ions (Ar^+ , Cs^+ , Ga^+ or other elements with energies between 0.5 and 10 keV), energy is deposited in the surface region of the sample by a collisional cascade. Some of the energy will return to the surface and stimulate the ejection of atoms, ions and multi-atomic clusters (figure B1.25.8). In SIMS, secondary ions (positive or negative) are detected directly with a mass spectrometer.

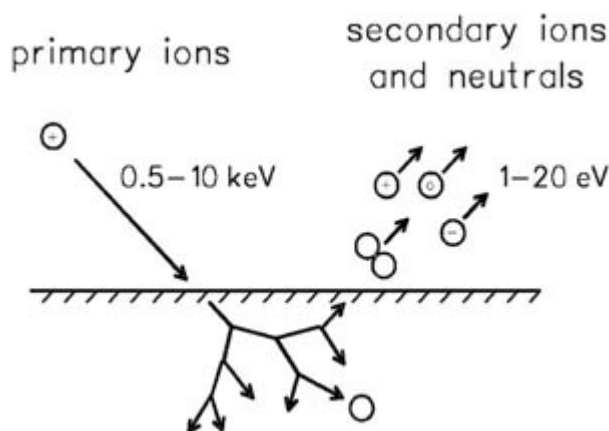


Figure B1.25.8. The principle of SIMS: Primary ions with an energy between 0.5 and 10 keV cause a collisional cascade below the surface of the sample. Some of the branches end at the surface and stimulate the emission of neutrals and ions. In SIMS, the secondary ions are detected directly with a mass spectrometer.

SIMS is, strictly speaking, a destructive technique, but not necessarily a damaging one. In the dynamic mode, used for making concentration depth profiles, several tens of monolayers are removed per minute. In static SIMS, however, the rate of removal corresponds to one monolayer per several hours, implying that the surface structure does not change during the measurement (between seconds and minutes). In this case one can be sure that the molecular ion fragments are truly indicative of the chemical structure on the surface.

The advantages of SIMS are its high sensitivity (ppm detection limit for certain elements), its ability to detect hydrogen and the emission of molecular fragments which often bear tractable relationships with the parent

structure on the surface. A disadvantage is that secondary ion formation is a poorly understood phenomenon and that quantitation is usually difficult. A major drawback is the matrix effect: Secondary ion yields of one element can vary tremendously with its chemical environment. This matrix effect and the elemental sensitivity variation of five orders of magnitude across the periodic table make quantitative interpretation of SIMS spectra of many compounds extremely difficult.

Figure B1.25.9(a) shows the positive SIMS spectrum of a silica-supported zirconium oxide catalyst precursor, freshly prepared by a condensation reaction between zirconium ethoxide and the hydroxyl groups of the support [17]. Note the simultaneous occurrence of single ions (H^+ , Si^{4+} , Zr^+) and molecular ions (SiO^+ , SiOH^+ , ZrO^+ , ZrOH^+ , ZrO_2^+). Also, the isotope pattern of zirconium is clearly visible. Isotopes are important in the identification of peaks, because all peak intensity ratios must agree with the natural abundance. In addition to the peaks expected from zirconia on silica mounted on an indium foil, the spectrum in figure B1.25.9(a) also contains peaks from the contaminants, Na^+ , K^+ and Ca^+ . This is typical for SIMS: sensitivities vary over several orders of magnitude and elements such as the alkalis are already detected when present in trace amounts.

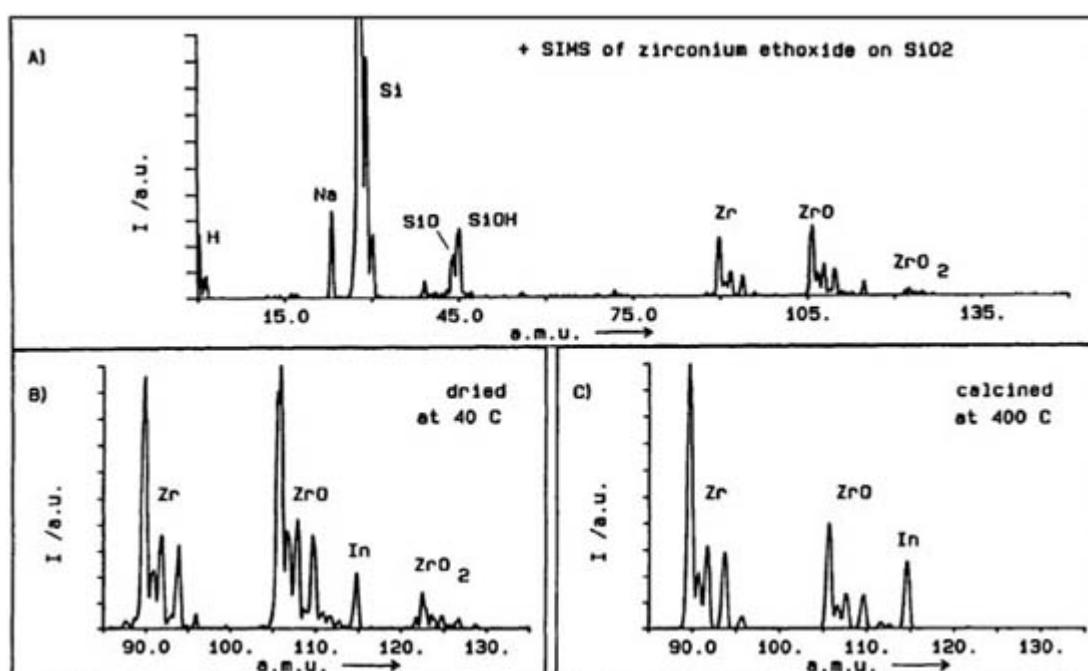


Figure B1.25.9. Positive SIMS spectra of a $\text{ZrO}_2/\text{SiO}_2$ catalyst, (a) after preparation from $\text{Zr}(\text{OC}_2\text{H}_5)_4$, (b) after drying at 40°C and (c) after calcination in air at 400°C (from [17]).

-12-

The most useful information is in the relative intensities of the Zr^+ , ZrO^+ , ZrOH^+ and ZrO_2^+ ions. This is illustrated in figure B1.25.9(b) and figure B1.25.9(c) which show the isotope patterns of these ions of a freshly dried and a calcined catalyst, respectively [17]. Note that the SIMS spectrum of the fresh catalyst contains small but significant contributions from ZrOH^+ ions (107 amu, $^{90}\text{ZrOH}^+$, and 111 amu, $^{94}\text{ZrOH}^+$). ZrOH^+ is most probably a fragment ion from zirconium ethoxide. In the spectrum of the catalyst which was oxidized at 400°C , the isotope pattern in the ZrO range resembles that of Zr , indicating that ZrOH species are absent. Spectrum figure B1.25.9(b) of the zirconium ethoxide (O:Zr = 4:1) shows higher intensities of the ZrO_2^+ and ZrO^+ signals than the calcined ZrO_2 (O:Zr = 2:1) does. The way to interpret this information is to compare the spectra of the catalysts with reference spectra of ZrO_2 and zirconium ethoxide [17].

For single crystals, matrix effects are largely ruled out and excellent quantization has been achieved by

calibrating SIMS yields by means of other techniques such as EELS and TPD (see further) [18]. Here SIMS offers the challenging perspective to monitor reactants, intermediates and products of catalytic reactions in real time while the reaction is in progress.

A good example of monitoring adsorbed species on surfaces with SIMS is shown in figure B1.25.10. This figure shows positive SIMS spectra of the interaction of NO with the Rh(111) surface [19]. The lower curve shows the adsorption of molecular NO (peak at 236 amu) on Rh(111) at 120 K. The middle curve shows the situation after heating the sample to 400 K. The presence of the peaks at 220 amu (Rh_2N^+) and 222 amu (Rh_2O^+) and the absence of the Rh_2NO^+ (236 amu) indicate that NO has dissociated. Heating the sample at 400 K in H_2 causes the removal of atomic oxygen and, thus, the disappearance of the Rh_2O^+ at 222 amu, as can be seen in the upper curve.

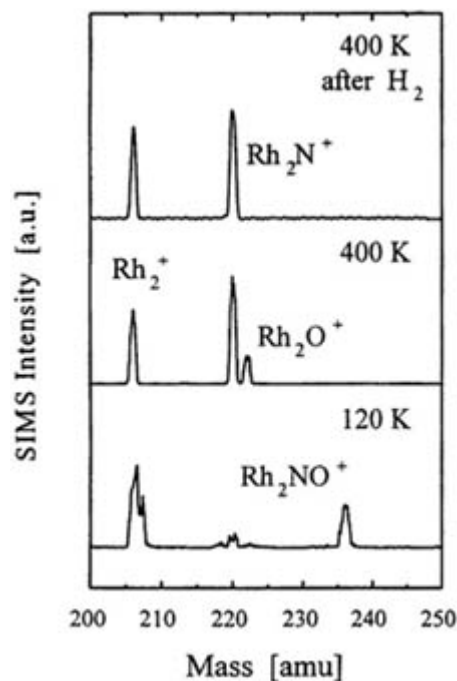


Figure B1.25.10. SIMS spectra of the Rh(111) surface after adsorption of 0.12 ML NO at 120 K (bottom), after heating to 400 K (middle) and after reaction with H_2 at 400 K (top) (from [19]).

As in Auger spectroscopy, SIMS can be used to make concentration depth profiles and, by rastering the ion beam over the surface, to make chemical maps of certain elements. More recently, SIMS has become very popular in the characterization of polymer surfaces [14, 15 and 16].

B1.25.4 TEMPERATURE PROGRAMMED DESORPTION (TPD)

Thermal desorption spectroscopy (TDS) or temperature programmed desorption (TPD), as it is also called, is a simple and very popular technique in surface science. A sample covered with one or more adsorbate(s) is heated at a constant rate and the desorbing gases are detected with a mass spectrometer. If a reaction takes place during the temperature ramp, one speaks of temperature programmed reaction spectroscopy (TPRS).

TPD is frequently used to determine (relative) surface coverages. The area below a TPD spectrum of a certain species is proportional to the total amount that desorbs. In this way one can determine uptake curves that correlate gas exposure to surface coverage. If the pumping rate of the UHV system is sufficiently high, the mass spectrometer signal for a particular desorption product is linearly proportional to the desorption rate of the adsorbate [20, 21]:

$$r = -d\theta/dt = k_{\text{des}}\theta^n = \nu(\theta)\theta^n \exp(-E_{\text{des}}(\theta)/RT) \quad (\text{B1.25.4})$$

$$T = T_0 + \beta t$$

where r is the rate of desorption, E_{des} is the activation energy of desorption, θ is the coverage in monolayers, R is the gas constant, t is the time, T is the temperature, k_{des} is the reaction rate constant for desorption, T_0 is the temperature at the start, n is the order of desorption, β is the heating rate, equal to dT/dt and ν is the preexponential factor of desorption.

With the aid of (B1.25.4), it is possible to determine the activation energy of desorption (usually equal to the adsorption energy) and the preexponential factor of desorption [21, 24]. Attractive or repulsive interactions between the adsorbate molecules make the desorption parameters E_{des} and ν dependent on coverage [22]. In the case of TPRS one obtains information on surface reactions if the latter is rate determining for the desorption.

Figure B1.25.11 shows the temperature programmed reaction between 0.15 ML of CO and 0.24 ML of NO adsorbed at 150 K on a Rh(111) single crystal [23]. The spectra show the desorption of species with masses 28, 29, 30, 44 and 45, corresponding to N_2 , ^{13}CO , NO, N_2O and $^{13}\text{CO}_2$ respectively, as functions of the temperature. N_2 , ^{13}CO and $^{13}\text{CO}_2$ are the only desorption products, indicating that NO is totally dissociated and all N atoms are converted to N_2 . CO is not decomposed at all; part of it desorbs as CO and part of it reacts with atomic oxygen to CO_2 . At higher NO coverages (not shown) the TPD spectrum has also a mass 30 signal, which is due to desorption of NO [28]. In this case, the coverage of adsorbed CO is too low to convert all NO to N_2 and CO_2 .

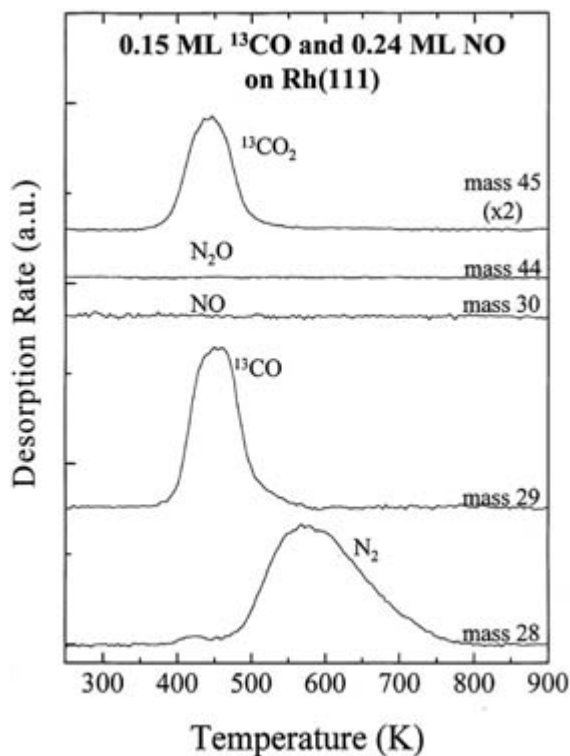


Figure B1.25.11. Temperature programmed reactions of 0.15 ML of ^{13}CO coadsorbed with 0.24 ML of NO. Adsorption was done at 150 K and the heating rate was 5 K s^{-1} (from [23]).

The disadvantage of TPD is that, in order to derive the kinetic parameters, rather involved computations are necessary [21, 24]. As an alternative to the complete desorption analysis, many authors rely on simplified methods. The analysis of spectra using simplified analysis should be made with care, as simplified analysis methods may easily give erroneous results [21].

B1.25.5 ELECTRON ENERGY LOSS SPECTROSCOPY (EELS)

Molecules possess discrete levels of vibrational energy. Vibrations in molecules can be excited by interaction with waves and with particles. In electron-energy loss spectroscopy (EELS, sometimes HREELS for high resolution EELS), a beam of monochromatic, low energy electrons falls on the surface, where it excites lattice vibrations of the substrate, molecular vibrations of adsorbed species and even electronic transitions. An energy spectrum of the scattered electrons reveals how much energy the electrons have lost to vibrations, according to the formula

$$E = E_0 - h\nu \quad (\text{B1.25.5})$$

where E is the energy of the scattered electron, E_0 is the energy of the incident electrons, h is Planck's constant and ν is the frequency of the excited vibration. The use of electrons requires that experiments are done in high vacuum and preferably on the flat surfaces of single crystals or foils (making ultrahigh vacuum conditions desirable).

While infrared and Raman spectroscopy are limited to vibrations in which a dipole moment or the molecular polarizability changes, EELS detects all vibrations. Two excitation mechanisms play a role in EELS: dipole

and impact scattering [4].

In dipole scattering we are dealing with the wave character of the electron. Close to the surface, the electron sets up an electric field with its image charge in the metal. This oscillating field is perpendicular to the surface and excites only those vibrations in which a dipole moment changes in a direction normal to the surface, similarly as in reflection absorption infrared spectroscopy. The outgoing electron wave has lost an amount of energy equal to $h\nu$, see (B1.25.5), and travels mainly in the specular direction.

The second excitation mechanism, impact scattering, involves a short range interaction between the electron and the molecule (put simply, a collision) which scatters the electrons over a wide range of angles. The useful feature of impact scattering is that all vibrations may be excited and not only the dipole active ones. As in Raman spectroscopy, the electron may also take an amount of energy $h\nu$ away from excited molecules and leave the surface with an energy equal to $E_0 + h\nu$.

Figure B1.25.12 illustrates the two scattering modes for a hypothetical adsorption system consisting of an atom on a metal [3]. The stretch vibration of the atom perpendicular to the surface is accompanied by a change in dipole moment; the bending mode parallel to the surface is not. As explained above, the EELS spectrum of electrons scattered in the specular direction detects only the dipole-active vibration. The more isotropically scattered electrons, however, undergo impact scattering and excite both vibrational modes. Note that the comparison of EELS spectra recorded in specular and off-specular direction yields information about the orientation of an adsorbed molecule.

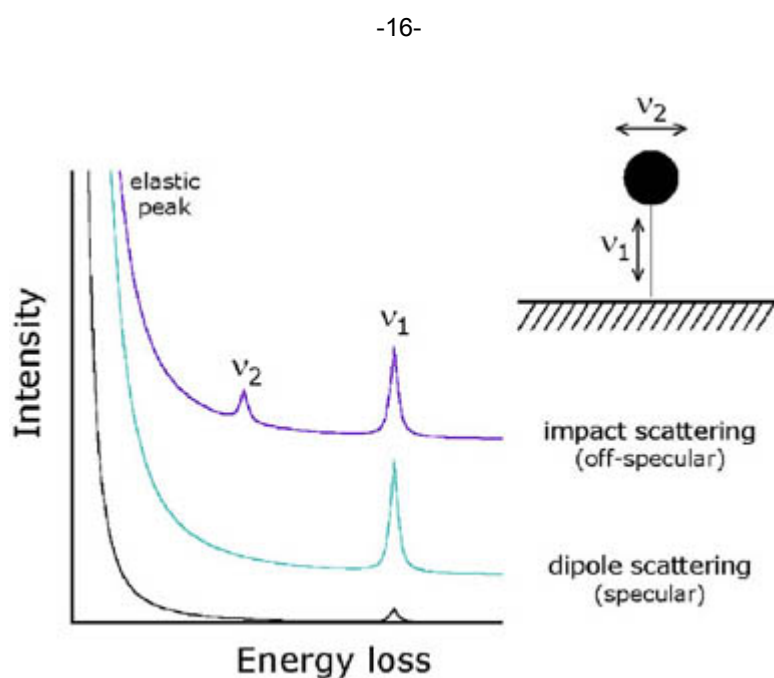


Figure B1.25.12. Excitation mechanisms in electron energy loss spectroscopy for a simple adsorbate system: Dipole scattering excites only the vibration perpendicular to the surface (ν_1) in which a dipole moment normal to the surface changes; the electron wave is reflected by the surface into the specular direction. Impact scattering excites also the bending mode ν_2 in which the atom moves parallel to the surface; electrons are scattered over a wide range of angles. The EELS spectra show the highly intense elastic peak and the relatively weak loss peaks. Off-specular loss peaks are in general one to two orders of magnitude weaker than specular loss peaks.

A strong point of EELS is that it detects losses in a very broad energy range, which comprises the entire infrared regime and extends even to electronic transitions at several electron volts. EELS spectrometers have to satisfy a number of stringent requirements. First, the primary electrons should be monochromatic. Second,

the energy of the scattered electrons should be measured with a high accuracy. Third, the low energy electrons must effectively be shielded from magnetic fields [25].

Figure B1.25.13 shows an HREELS spectrum of CO adsorption on a Rh(111) surface [26]. In the experiment 3 L of CO was adsorbed at a pressure of 1×10^{-8} mbar and $T = 200$ K. At zero energy loss one observes the highly intense elastic peak. The other peaks in the spectrum are loss peaks. At high energy, loss peaks due to dipole scattering are visible. In this case they are caused by CO vibration perpendicular to the surface. The peak at 2070 cm^{-1} is attributed to on-top adsorption of CO on Rh, while the peak at 1861 cm^{-1} corresponds to CO adsorption on a threefold Rh site. The loss peaks at low energy loss are due to metal-adsorbate vibrations. In this case it is the Rh-CO bond. The peak at 434 cm^{-1} is due to the Rh-CO vibration of the CO adsorbed on top, that at 390 cm^{-1} to the Rh-CO vibration in threefold CO. The CO molecules order in a $(2 \times 2) - 3\text{CO}$ structure, with one linear and two threefold CO molecules per (2×2) unit cell.

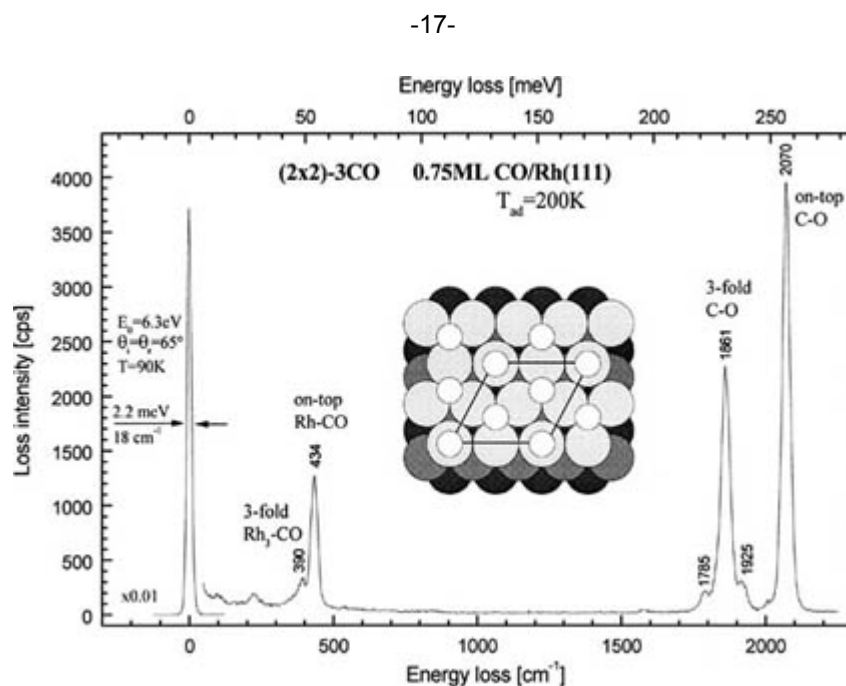


Figure B1.25.13. HREELS spectrum of CO adsorbed on Rh(111) at $T = 200$ K. Visible are the C-O vibration peaks at energy losses around $1800\text{--}2100 \text{ cm}^{-1}$ and the Rh-CO signals at energy losses $300\text{--}500 \text{ cm}^{-1}$ (courtesy of R Linke [26]).

REFERENCES

- [1] Delgass W N, Haller G L, Kellerman R and Lunsford J H 1979 *Spectroscopy in Heterogeneous Catalysis* (New York: Academic)
- [2] Feldman L C and Mayer J W 1986 *Fundamentals of Surface and Thin Film Analysis* (Amsterdam: North-Holland)
- [3] Niemantsverdriet J W 1993 *Spectroscopy in Catalysis, an Introduction* (Weinheim: VCH)
- [4] Ertl G and Kuppers J 1985 *Low Energy Electrons and Surface Chemistry* (Weinheim: VCH)
- [5] Ghosh P K 1983 *Introduction to Photoelectron Spectroscopy* (New York: Wiley)
- [6] Feuerbacher B, Fitton B and Willis R F (eds) 1978 *Photoemission and the Electronic Properties of*

Surfaces (New York: Wiley)

- [7] Briggs D and Seah M P (eds) 1983 *Practical Surface Analysis by Auger and X-ray Photoelectron Spectroscopy* (New York: Wiley)
- [8] Wagner C D, Riggs W M, Davis L E, Moulder J F and Muilenburg G E 1979 *Handbook of X-ray Photoelectron Spectroscopy* (Eden Prairie, MN: Perkin Elmer)
-

-18-

- [9] Muijsers J C, Niemantsverdriet J W, Wehman-Ooyevaar I C M, Grove D M and van Koten G 1992 *Inorg. Chem.* **31** 2655
- [10] Somorjai G A 1981 *Chemistry in Two Dimensions, Surfaces* (Ithaca, NY: Cornell University Press)
- [11] Verhoeven M W G M and Niemantsverdriet J W, unpublished results
- [12] Gunter P L J, de Jong A M, Niemantsverdriet J W and Rheiter H J H 1992 *Surf. Interface Anal.* **19** 161
- [13] Haas T W, Grant J T and Dooley G J 1972 *J. Appl. Phys.* **43** 1853
- [14] Benninghoven A, Rudenauer F G and Werner H W 1987 *Secondary Ion Mass Spectrometry, Basic Concepts, Instrumental Aspects, Applications and Trends* (New York: Wiley)
- [15] Vickerman J C, Brown A and Reed N M (eds) 1989 *Secondary Ion Mass Spectrometry, Principles and Applications* (Oxford: Clarendon)
- [16] Briggs D, Brown A and Vickerman J C 1989 *Handbook of Static Secondary Ion Mass Spectrometry* (Chichester: Wiley)
- [17] Meijers A C Q M, de Jong A M, van Gruijthuisen L M P and Niemantsverdriet J W 1991 *Appl. Catal.* **70** 53
- [18] Borg H J and Niemantsverdriet J W 1994 *Catalysis, Specialist Periodical Report* vol 11, ed J J Spivey and S K Agarwal (Cambridge: Royal Society of Chemistry) ch 11
- [19] van Hardeveld R M, Borg H J and Niemantsverdriet J W 1998 *J. Mol. Catal. A* **131** 199
- [20] King D A 1975 *Surf. Sci.* **47** 384
- [21] de Jong A M and Niemantsverdriet J W 1990 *Surf. Sci.* **233** 355
- [22] Cassuto A and King D A 1981 *Surf. Sci.* **102** 388
- [23] Hopstaken M J P, van Gennip W J H and Niemantsverdriet J W 1999 *Surf. Sci.* 433–435
- [24] Falconer J L and Schwarz J A 1983 *Catal. Rev. Sci. Eng.* **25** 141
- [25] Ibach H 1990 *Electron Energy Loss Spectrometers: the Technology of High Performance* (Berlin: Springer)
- [26] Linke R and Niemantsverdriet J W, to be published
- [27] Eshelman L M, de Jong A M and Niemantsverdriet J W 1991 *Catal. Lett.* **10** 201
-

FURTHER READING

Ertl G and Kuppers J 1985 *Low Energy Electrons and Surface Chemistry* (Weinheim: VCH)

Feldman L C and Mayer J W 1986 *Fundamentals of Surface and Thin Film Analysis* (Amsterdam: North-Holland)

-19-

Woodruff D P and Delchar T A 1986 *Modern Techniques of Surface Science* (Cambridge: Cambridge University Press)

Niemantsverdriet J W 1993 *Spectroscopy in Catalysis, an Introduction* (Weinheim: VCH)

-1-

B1.26 Surface physical characterization

W T Tysoe and Gefei Wu

B1.26.1 INTRODUCTION

The physical structure of a surface, its area, morphology and texture and the sizes of orifices and pores are often crucial determinants of its properties. For example, catalytic reactions take place at surfaces. Simple statistical mechanical estimates suggest that a surface-mediated reaction should proceed about 10^{12} times faster than the corresponding gas-phase reaction for identical activation energies [1]. The catalyst operates by lowering the activation energy of the reaction to accelerate the rate. The reaction rate, however, also increases in proportion to the exposed surface area of the active component of the catalyst so that maximizing its area also strongly affects its activity. Catalysts often have complicated morphologies, consisting of exposed regions, and small micro- and meso-pores. A traditional method for measuring these areas, which is still the workhorse for the catalytic chemists, is to titrate the surface with molecules of known 'areas' and to measure the amount that just covers it. This is done by pressurizing the sample using probe gases and gauging when a single layer of adsorbate forms. This relies on developing robust theoretical methods for determining the equilibrium between the gas phase and the surface. This was done in 1938 by Brunauer, Emmett and Teller. Brunauer and Emmett were catalytic chemists and Teller a theoretical physicist who was persuaded to undertake the theoretical task of developing an adsorption isotherm [2]. This he apparently did in one day and the Brunauer–Emmett–Teller isotherm was born. This, with minor modifications, is the isotherm still used today.

On planar systems, morphologies can be generally measured by directly imaging them using optical ([section B1.19](#) and [section B1.21](#)) or electron ([B1.18](#)) microscopies or using scanning probes (see [section B1.20](#)). Coarse morphologies can be measured using crude probes such as profilometers [3] and, more recently, at the atomic level, using atomic force microscopy ([section B1.20](#)). The measurement of the thickness and properties of thin films deposited onto planar surfaces is more of a challenge. Electron-based spectroscopic probes can measure the nature of the outer selvedge of planar films (see [section B1.6](#), [section B1.7](#), [section B1.21](#)). For example, x-ray photoelectron spectroscopy is useful for measuring films of few Ångströms thick using the electron escape depth. The film itself can be probed using optical spectroscopic techniques such as infrared ([B1.2](#)) and Raman ([B1.3](#)) spectroscopies. Ellipsometry, the change in polarization of linearly polarized light as it passes through the film, is used to measure film thickness non-destructively over a wide range. It is particularly useful for probing surface coatings such as anti-reflection and protective films and has been used very effectively to probe overlayers in ultrahigh vacuum and the formation kinetics of self-

assembled monolayers on gold.

The final technique addressed in this chapter is the measurement of the surface work function, the energy required to remove an electron from a solid. This is one of the oldest surface characterization methods, and certainly the oldest carried out *in vacuo* since it was first measured by Millikan using the photoelectric effect [4]. The observation of this effect led to the proposal of the Einstein equation:

$$E_k = h\nu - e\Phi \quad (\text{B1.26.1})$$

-2-

where ν is the light frequency, h is Planck's constant, E_k the kinetic energy of the emitted electron, e the charge on an electron and Φ the material work function. The resulting notion of wave-particle duality led directly to the development of quantum mechanics. This is not strictly a physical probe since the work function of a clean sample depends on its electronic structure. This is strongly affected by the presence of adsorbates, electronegative adsorbates leading to an increase in work function, and electropositive adsorbates to a decrease. The observations have technological implications, so that filaments used today as electron sources in cathode ray (television) and vacuum tubes (valves) are coated with electropositive alkaline earth compounds that lower the work function and enhance the thermionically emitted current. This allows the filaments to operate effectively at lower temperatures and thereby increases their lifetimes. The main experimental utility of this method is to measure, in a simple and direct way, the coverage of an adsorbate on a surface (see [section A1.7](#)).

B1.26.2 THE BRUNAUER-EMMETT-TELLER (BET) METHOD

B1.26.2.1 PRINCIPLES

(A) MEASUREMENTS OF SURFACE AREA BY GAS ADSORPTION

The central idea underlying measurements of the area of powders with high surface areas is relatively simple. Adsorb a close-packed monolayer on the surface and measure the number N of these molecules adsorbed per unit mass of the material (usually per gram). If the specific area occupied by each molecule is A_m , then the total surface area S_A of the sample is simply given by:

$$S_A = NA_m. \quad (\text{B1.26.2})$$

The saturation coverage during chemisorption on a clean transition-metal surface is controlled by the formation of a chemical bond at a specific site [5] and not necessarily by the area of the molecule. In addition, in this case, the heat of chemisorption of the first monolayer is substantially higher than for the second and subsequent layers where adsorption is *via* weaker van der Waals interactions. Chemisorption is often useful for measuring the area of a specific component of a multi-component surface, for example, the area of small metal particles adsorbed onto a high-surface-area support [6], but not for measuring the *total* area of the sample. Surface areas measured using this method are specific to the molecule that chemisorbs on the surface. Carbon monoxide titration is therefore often used to define the number of 'sites' available on a supported metal catalyst. In order to measure the total surface area, adsorbates must be selected that interact relatively weakly with the substrate so that the area occupied by each adsorbent is dominated by intermolecular interactions and the area occupied by each molecule is approximately defined by van der Waals radii. This

generally necessitates experiments being carried out at low temperatures such that $kT \ll \Delta H_{(\text{ads})}$, the heat of adsorption. Since now both the interaction of the first and subsequent adsorbate layers is dominated by van der Waals

-3-

forces, this leads not simply to the formation of a single monolayer, but also to the growth of second, third and subsequent layers. This distinction is shown in figure B1.26.1 which plots coverage versus pressure at constant temperature (an isotherm) for a molecule (hydrogen) which chemisorbs at the surface where the saturation of the monolayer is clearly evident from the appearance of a plateau [7]. In the case of physisorption, as demonstrated in figure B1.26.2, subsequent layers can grow so that the number of molecules adsorbed in the first monolayer is much more difficult to identify [8]. It is clear, in this case, that the first monolayer forms somewhere near the first 'knee' of the isotherm, labelled point 'B'. The importance of this point was first emphasized by Emmett [9]. In order to usefully measure the total surface area, the shape of the adsorption isotherm must be analysed to more clearly distinguish between monolayer (point B) and multilayer adsorption. In 1985, IUPAC introduced a classification of six different types of adsorption isotherm [11] (figure B1.26.3) exhibited by real surfaces. Types I–V were originally classified by Brunauer, Denning, Denning and Teller (BDDT) [10]. Type I represents the Langmuir isotherm [12] for monolayer coverage and is most often exhibited for chemisorption where the heat of adsorption in the first layer is much greater than that in subsequent layers, but also corresponds to physisorption by microporous adsorbents (pore width < 2 nm) within the solid. Type II are monolayer–multilayer isotherms and represent non-porous or macroporous adsorbents (pore width > 50 nm). Industrial adsorbents and catalysts which possess mesoporous (pore width 2–50 nm) structures often exhibit type IV behaviour. The shapes of types III and V isotherms are analogous to type II and IV respectively, but with weak gas–solid interactions. The stepwise type VI isotherms can be obtained with well defined, uniform solids. The origin of the shapes of some of these isotherms will be discussed in greater detail below.

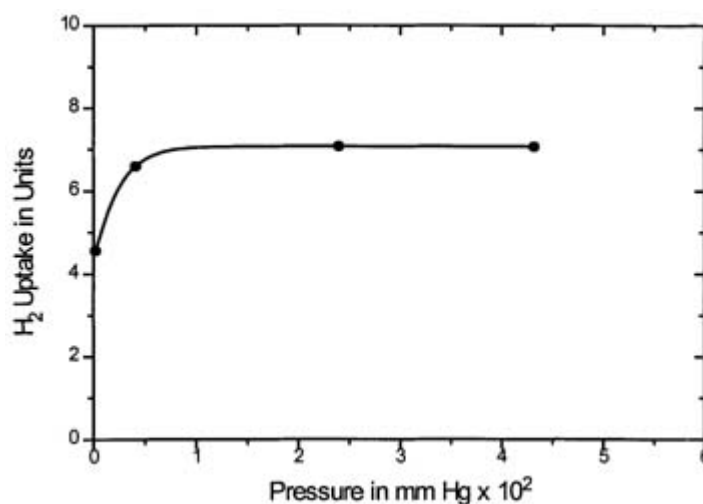


Figure B1.26.1. Sorption isotherm for chemisorption of hydrogen on palladium film at 273 K (Stephens S J 1959 *J. Phys. Chem.* **63** 188–94).

-4-

(B) TYPICAL ISOTHERMS FOR THE PHYSICAL ADSORPTION OF GASES ON SURFACES

As noted above, an isotherm plots the number of molecules adsorbed on the surface at some temperature in equilibrium with the gas at some pressure. Adsorption gives rise to a change in the free energy which, of

course, depends on the number of molecules already adsorbed on the surface (defined as the coverage, θ , see [section A1.7](#)). If it is assumed for simplicity that the structure of the adsorbed phase is similar to that of a solid (so that, at equilibrium, the chemical potential of the bulk phase equals the chemical potential for the gas phase in equilibrium with it) then:

$$\mu_{\text{bulk}} = \mu^0 + RT \ln P_0 \quad (\text{B1.26.3})$$

where P_0 is the equilibrium vapour pressure. The chemical potential of the adsorbed phase in equilibrium with the gas at some temperature T can similarly be written as:

$$\mu_{\text{ads}} = \mu^0 + RT \ln P. \quad (\text{B1.26.4})$$

We assume for simplicity that the adsorbed phase has the same entropy as the solid so that only an energy change is associated with the transfer of material from the bulk to the adsorbed phase, then:

$$\Delta F = \mu_{\text{ads}} - \mu_{\text{bulk}} = N \Delta E \quad (\text{B1.26.5})$$

where ΔE is the change in energy per adsorbed atom or molecule in going from the solid to the adsorbed phase. Combining equation (B1.26.3), equation (B1.26.4) and equation (B1.26.5) yields:

$$\ln \frac{P}{P_0} = \frac{\Delta E}{kT}. \quad (\text{B1.26.6})$$

If we knew the variation in ΔE as a function of coverage θ , this would be the equation for the isotherm. Typically the energy for physical adsorption in the first layer, $-\Delta E_1$, when adsorption is predominantly through van der Waals interactions, is of the order of $10kT$ where T is the temperature and k the Boltzmann constant, so that, according to equation (B1.26.6), the first layer condenses at a pressure given by $P/P_0 \sim 10^{-3}$. This accounts for the rapid initial rise in the isotherm for low values of P/P_0 shown in [figure B1.26.2](#). In the case of chemisorption, where the interaction is even stronger, the first monolayer saturates at even lower values of P/P_0 . It is initially assumed that adsorption into the second layer is dominated by van der Waals interactions with the sample. The attractive van der Waals energy ϵ_a between two molecules in the gas-phase separated by a distance r is $\epsilon_a \sim -K_a/r^6$ where K_a is a constant. When combined with an r^{-12} repulsive potential, this yields the Lennard-Jones 6–12 equation. An adsorbed species interacts with atoms in the truncated bulk of the sample, and the number of these increases as $\sim r^3$ so that the net van der Waals interaction with the surface varies as r^{-3} . This indicates that the energy of adsorption in the second layer ΔE_2 , assuming this to be dominated by van der Waals interactions with the surface, is given by: $\Delta E_2 = \Delta E_1/2^3$.

From [equation \(B1.26.6\)](#), this yields $\ln(P/P_0) \sim -1.3$ so that the pressure at which this layer is complete should be $P/P_0 \sim 0.3$, a value significantly higher than that required to saturate the first layer. Similarly, the energy required to saturate the third layer will be reduced by a factor 3^3 , the fourth layer 4^3 and so on. Therefore, if a value of ΔE is assumed for the first layer, the pressures at which the second and subsequent layer saturate can be calculated from this value. Writing this in terms of the coverage θ yields a simple form for the variation of ΔE as a function of coverage as $\Delta E_1/\theta^3$. This yields an isotherm for physisorption dominated by van der Waals interactions with the surface as:

$$\ln \frac{P}{P_0} = \frac{\Delta E_1}{kT\theta^3}. \quad (\text{B1.26.7})$$

Such isotherms are shown in [figure B1.26.4](#) for the physical adsorption of krypton and argon on graphitized carbon black at 77 K [13] and are examples of type VI isotherms ([figure B1.26.3](#)). Equation (B1.26.7) further predicts that a plot of $\ln(P/P_0)$ versus $1/\theta^3$ should be linear: such a plot is displayed in [figure B1.26.5](#) [13] for the adsorption of argon on graphitized carbon black at 77 K and yields a good straight line.

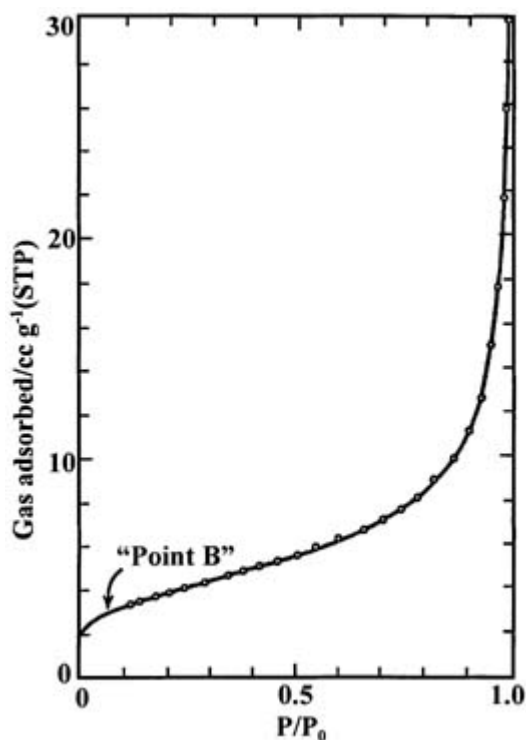


Figure B1.26.2. Adsorption isotherm for nitrogen on anatase at 77 K showing ‘point B’ (Harkins W D 1952 *The Physical Chemistry of Surface Films* (New York: Reinhold)).

-6-

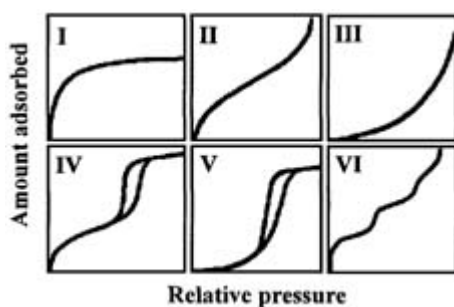


Figure B1.26.3. The IUPAC classification of adsorption isotherms for gas–solid equilibria (Sing K S W, Everett D H, Haul R A W, Mosoul L, Pierotti R A, Rouguerol J and Siemieniwska T 1985 *Pure. Appl. Chem.* 57 603–19).

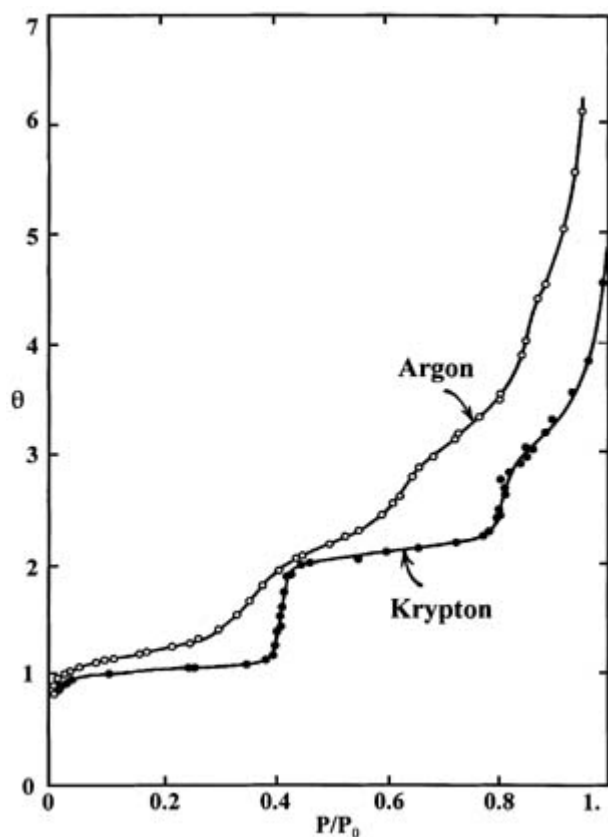


Figure B1.26.4. The adsorption of argon and krypton on graphitized carbon black at 77 K (Eggers D F Jr, Gregory N W, Halsey G D Jr and Rabinovitch B S 1964 *Physical Chemistry* (New York: Wiley) ch 18).

-7-

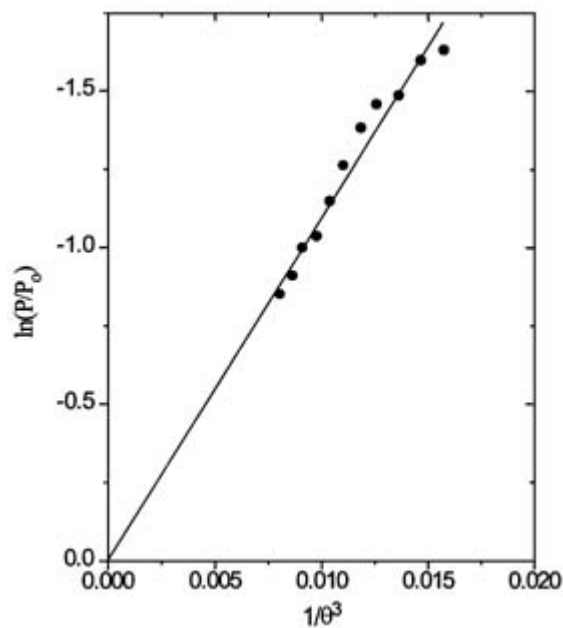


Figure B1.26.5. Plot of $\ln(P/P_0)$ versus $1/\theta^3$ for argon on graphitized carbon black at 77 K (from the argon data in [figure B1.26.4](#)) (Eggers D F Jr, Gregory N W, Halsey G D Jr and Rabinovitch B S 1964 *Physical Chemistry* (New York: Wiley) ch 18).

It is clear, however, that not all isotherms display the stepwise behaviour shown in [figure B1.26.4](#). For

example, the isotherm for the adsorption of nitrogen on anatase [8] (figure B1.26.2) has a rapid increase in coverage at low pressures, corresponding to the saturation of the first monolayer, but varies much more smoothly with coverage at higher pressures. This effect is also seen in the data for argon on carbon black [13] (figure B1.26.4) where the steps become much less pronounced for multilayer adsorption. Part of the reason for this discrepancy is that, as the layer becomes thicker, intermolecular van der Waals interactions within the layer become more important. This has two effects. First, the difference in energy between layers becomes less pronounced, leading to a smoothing out of the curve. In addition, this decrease in energy difference for each layer means that subsequent layers start to grow even before previous monolayers have saturated. Finally, in the case of high-surface-area samples, the surface tends to become more heterogeneous, leading to a further smoothing out of the steps. The limiting case is an isotherm calculated on a different basis to that used for equation (B1.26.8). Here it is assumed that adsorption in the first layer is dominated by surface van der Waals interactions, but that adsorption into second and subsequent layers is dominated by intermolecular interactions between adsorbates. This clearly no longer results in a strong variation in ΔE with each layer and allows multilayer films to be formed in which another layer can start before the previous layer has been completed. This smooths out the isotherm resulting in a variation of coverage with P that more closely resembles that shown in figure B1.26.2. Such an isotherm has the advantage that it often more closely mimics the behaviour of nitrogen physisorbed on high-surface-area materials and, as such, is more useful in reproducibly identifying ‘point B’. This forms the basis of a reproducible method for measuring surface areas using the BET isotherm. The calculation of the BET isotherm assumes that:

-8-

1. There are B equivalent sites available for adsorption in the first layer.
2. Each molecule adsorbed in the first layer is considered to be a possible adsorption site for molecules adsorbing into a second layer, and each molecule adsorbed in the second layer is considered to be a ‘site’ for adsorption into the third layer, and so on.
3. All molecules in the second and subsequent layers are assumed to behave similarly to a liquid, in particular to have the same partition function. This is assumed to be different to the partition function (A2.2) of molecules adsorbed into the first layer.
4. Intermolecular interactions are ignored for all layers.

Detailed derivations of the isotherm can be found in many textbooks and exploit either statistical thermodynamic methods [1] or independently consider the kinetics of adsorption and desorption in each layer and set these equal to define the equilibrium coverage as a function of pressure [14]. The most common form of BET isotherm is written as a linear equation and given by:

$$\frac{P}{V(P - P_0)} = \frac{1}{V_m C} + \frac{P(C - 1)}{V_m C P_0}. \quad (\text{B1.26.8})$$

Here V_m is the volume of gas required to saturate the monolayer, V the total volume of gas adsorbed, P the sample pressure, P_0 the saturation vapour pressure and C a constant related to the enthalpy of adsorption. The resulting shape of the isotherm is shown plotted in figure B1.26.6 for $C = 500$. A plot of $P/V(P - P_0)$ against P/P_0 should give a straight line having a slope $(C - 1)/V_m C$ and an intercept $1/V_m C$. The BET surface area is then calculated using the following equation:

$$S_A = V_m N A_m \quad (\text{B1.26.9})$$

where S_A is the required surface area of the sample, V_m the volume of the adsorbed monolayer, N Avogadro's number and A_m the cross-sectional area of the adsorbed molecule. In the BET method, where nitrogen is generally used, the value of A_m is taken to be 16.2 \AA^2 per nitrogen molecule. Classically, this BET equation is used for only for systems that exhibit type II and IV isotherms.

-9-

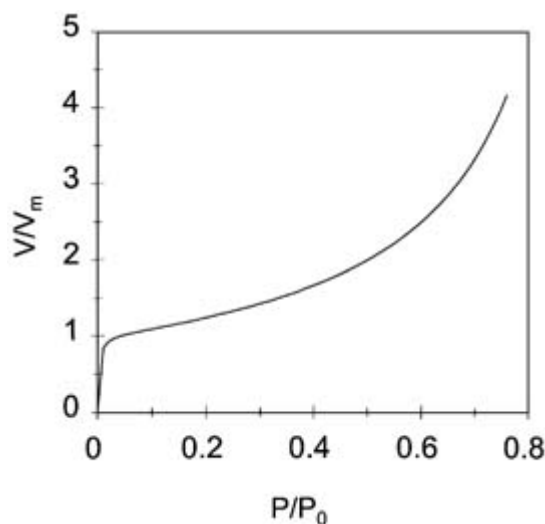


Figure B1.26.6. BET isotherm plotted from [equation \(B1.26.8\)](#) using a value of $C = 500$.

(C) MEASUREMENT OF BET ISOTHERMS

Practically, using the BET method to measure surface area involves three steps: (1) obtaining a full adsorption isotherm, (2) evaluating the monolayer capacity and (3) the calculation of surface area using [equation \(B1.26.9\)](#). It should be emphasized that the BET surface area represents a 'standard' method for measuring the area of a high-surface-area material that allows samples from different laboratories to be compared. Note that, due to the simplifying assumptions of the derivation, the BET method does not work well for type III and V isotherms where the weak interaction between gas and solid makes it hard to discern the formation of the first layer. Attempts to modify the BET equation to take account of these situations have proven unreliable and impractical. Beyond the BET method, approaches such as the Gibbs adsorption isotherm [15], immersion calorimetry [16] or adsorption from solution [17] have been used but the BET method continues as a standard procedure for the determination of surface areas [18]. Generally the measured value of BET-area can be regarded as an effective area unless the material is ultramicroporous. In the case of porous materials, it is important to know the pore sizes and their distributions. This can be calculated for type IV adsorbents where the mesopore size distribution can be obtained using the Kelvin equation:

$$\ln \left(\frac{P}{P_0} \right) = -\frac{2\gamma V}{rRT} \quad (\text{B1.26.10})$$

This equation describes the additional amount of gas adsorbed into the pores due to capillary action. In this case, V is the molar volume of the gas, γ its surface tension, R the gas constant, T absolute temperature and r the Kelvin radius. The distribution in the sizes of micropores may be determined using the Horvath–Kawazoe method [19]. If the sample has both micropores and mesopores, then the T -plot calculation may be used [20]. The T -plot is obtained by plotting the volume adsorbed against the statistical thickness of adsorbate. This thickness is derived from the surface area of a non-porous sample, and the volume of the liquified gas.

B1.26.2.2 INSTRUMENTATION

Two parameters must be measured to apply the BET equation, the pressure at the sample and the amount adsorbed at this pressure. There are three common methods for measuring the amount of gas adsorbed, called the volumetric method, the gravimetric method and the dynamic method, of which the volumetric method is the commonest [21].

(A) VOLUMETRIC METHOD

This method essentially consists of admitting successive charges of gas (generally nitrogen) to the adsorbent using some form of volumetric measuring device such as a gas burette or pipette with the sample held at liquid nitrogen temperature (77 K). Nowadays, the amount of gas admitted to the sample can most conveniently be measured using a mass-flow controller. When equilibrium has been attained, the gas pressure in the dead space surrounding the sample is read using a manometer, and the quantity of gas remaining unadsorbed is then calculated with the aid of the gas law, assuming a perfect gas. The volume of the dead space of the apparatus must, of course, be accurately calibrated. A precision manometer should be employed. The quantity of gas adsorbed onto the surface can be calculated by subtracting the amount remaining unadsorbed from the total amount which has been admitted. This type of apparatus can be simply constructed from glass. Alternatively, commercial BET measuring apparatuses are also available using computers to collect the data and to calculate the resulting surface area [22]. Shown in figure B1.26.7 is a schematic diagram of a typical apparatus for measuring BET isotherms [23]. Helium, which does not adsorb at liquid nitrogen temperatures, is used to calibrate the volume of the dead space and the furnace is used to outgas the sample prior to gas adsorption. A mass-flow controller is often used instead of a burette or pipette to measure total amount of nitrogen that has been admitted.

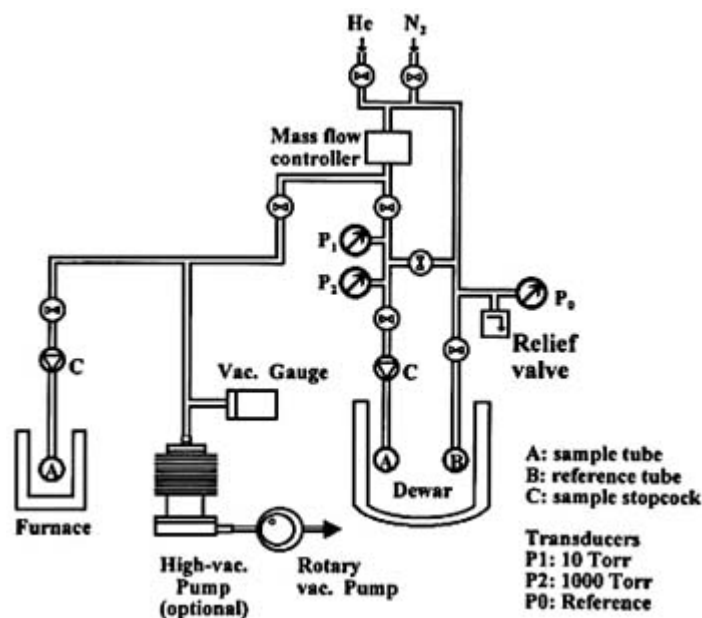


Figure B1.26.7. Diagram of a BET apparatus (representing an OMNISORB 100) (Beckman Coulter 1991 *OMNISORB Manual*).

(B) GRAVIMETRIC METHOD

This method is simple but experimentally more cumbersome than the volumetric method and involves the use of a vacuum microbalance or beam balance [22]. The solid is suspended from one arm of a balance and its increase in weight when adsorption occurs is measured directly. The ‘dead space’ calculation is thereby avoided entirely but a buoyancy correction is required to obtain accurate data. Nowadays this method is rarely used.

(C) DYNAMIC METHOD

This method has been developed using gas chromatographic techniques. The most popular way of implementing this method is by using a continuous nitrogen flow as first described by Nelsen and Eggertsen [24]. A known mixture of nitrogen and helium is passed through a bed of solid sample at ambient temperature (~ 300 K) where the exit gas is measured using a gas chromatographic detector. When the gas composition equilibrates as indicated by a constant base line on the GC recorder chart, the sample tube is immersed in a liquid nitrogen bath. The adsorption of nitrogen by the solid is then indicated by a negative excursion on the recorder chart corresponding to a loss of nitrogen from the system due to adsorption. After equilibrium is established at the particular partial pressure with the sample held at 77 K, the baseline attains its original level. The sample tube is then allowed to warm up to room temperature (300 K) by removal of the liquid nitrogen bath so that a positive excursion appears due to nitrogen desorption (see figure B1.26.8) [24]. The areas under these two curves should be equal and constitute a measure of the amount of nitrogen adsorbed. This method has drawn considerable interest due to its simplicity and speed and since it does not require a vacuum system. One of the main problems is that of deciding on the most appropriate conditions for ‘outgassing’ which can considerably affect the precision.

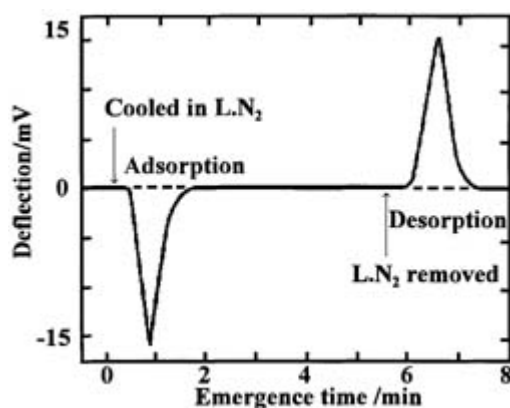


Figure B1.26.8. Adsorption/desorption peaks for nitrogen obtained with the continuous flow method (Nelsen F M and Eggertsen F T 1958 *Anal. Chem.* **30** 1387–90).

B1.26.2.3 EXPERIMENTAL NOTES

Nitrogen is the most widely used adsorbent (at 77 K) for the BET method and has been employed almost universally. Argon is more suited to the measurement of microporous zeolites. Krypton may be used for the measurement of very low-surface-area (less than $3 \text{ m}^2\text{g}^{-1}$) samples because it has a very low saturation vapour pressure (~ 2.45 Torr) at liquid nitrogen temperature so that the absolute pressure range is from 0 to approximately 0.75 Torr for a ratio of P/P_0 from 0 to 0.3. The absolute pressure range used for the measurement depends upon the type of data reduction required and the adsorbent's properties. The optimum

pressure range needed for BET surface area determinations may be taken to be for P/P_0 up to 0.3 [23].

B1.26.3 ELLIPSOMETRY

B1.26.3.1 PRINCIPLES

The term ellipsometry was first coined by Rothen in 1945 [25] to refer to the measurement of thin films of materials by monitoring the light reflected from them at some incident angle θ . The method is illustrated in [figure B1.26.9](#). The technique was used extensively before this [26]. In the simplest case, the film is transparent with refractive index n , such that light can be reflected or transmitted at the first interface. The transmitted beam propagates through the material and is reflected from the substrate. The reflected portion of the beam can either subsequently reflect from the film/air interface or reflect once again into the film and undergo further reflections. The multiple beams are eventually emitted together so that, in general they are attenuated and one of the parameters that can most simply be measured is the reflectivity of the film. In addition, because of the phase shifts that occur because of path length differences as the beam passes through the film, there is also a change of the phase of this beam which depends on the film thickness d and the wavelength of the light, λ . The way in which this phase change can be measured will be described below. Each of these values, that is, the reflectivity and the phase shift, depend on the polarization of the radiation (see below). When the light is polarized parallel to the surface it is said to be s polarized and when it is polarized perpendicularly to the surface, p polarized. The corresponding reflectivities are denoted R^s and R^p which, since there is a phase change on reflection, are generally complex numbers. The ratio of these values is defined as ρ which is given by:

$$\rho = \frac{R^p}{R^s}. \quad (\text{B1.26.11})$$

Since this is a complex number, it can be separated into an amplitude and a phase and written as:

$$\rho = \frac{|R^p|}{|R^s|} \exp(i\Delta) \quad (\text{B1.26.12})$$

where Δ is the difference between the phase shifts for p- and s-polarized light: $\Delta = \delta_s - \delta_p$. By convention, the ratio of the moduli of the reflectivities of the p- and s-polarized light $|R^p|/|R^s|$ is written in terms of another parameter Ψ where $\tan\Psi = |R^p|/|R^s|$. Thus, the parameters that are measured in ellipsometry are Ψ and Δ which can be related to the refractive index and thickness of the film.

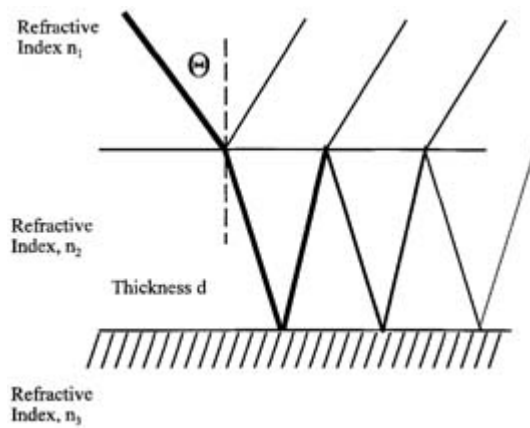


Figure B1.26.9. Schematic diagram showing the reflection of light incident at an angle Θ from a medium with refractive index n_1 through a film of thickness d with refractive index n_2 .

(A) THE NATURE OF ELECTROMAGNETIC RADIATION

As shown by Maxwell, light consists of an oscillating electromagnetic field propagating in vacuum at the speed of light [27, 28, 29, 30, 31, 32 and 33]. Both the electric and magnetic fields are oriented perpendicularly to the direction of propagation of the light and perpendicularly to each other. By convention, the direction in which the electric field points defines the polarization direction so that plane-polarized light has an electric field component only in one direction. A common way to obtain plane-polarized light is to use a polarizer which only transmits light of one polarization and absorbs the other. Polarized spectacles have lenses made from such polarizing material. Maxwell was also able to demonstrate that the velocity v with which light propagates in a material is given by $v = 1/\sqrt{\mu\epsilon}$ where ϵ is the permittivity of the material, ϵ_0 that of vacuum, μ the permeability of the material and μ_0 the value in vacuum. When light propagates in vacuum, this reduces to $v = 1/\sqrt{\mu_0\epsilon_0}$ and yields the speed of light, c . This correspondence provided confirmation that light and electromagnetic radiation were one and the same. The refractive index of a material, n , is defined as $n = c/v$, and is therefore given by $n = \sqrt{\mu\epsilon}$. Since most materials under investigation are not magnetic, $\mu = 1$, and the equation for refractive index simplifies to $n = \sqrt{\epsilon}$.

As electromagnetic radiation propagates through space, the electric field converts into a magnetic field during the oscillation cycle. Thus, the energy present in the electric field converts to energy in a magnetic field. The magnetic field oscillates at the same frequency but 90° out of phase with the electric field so that, when the electric field is a maximum, the magnetic field is a minimum, and *vice versa*. This idea allows us to calculate the relative electric and magnetic field amplitudes in an electromagnetic wave. Let the electric field amplitude be E_0 and the corresponding magnetic field amplitude H_0 . The energy in a magnetic field is proportional to μH_0^2 and that in an electric field is proportional to ϵE_0^2 . Since the electric and magnetic fields interconvert in an electromagnetic wave, conservation of energy requires that these be equal, so that $H_0 = \sqrt{(\epsilon/\mu)}E_0$. Again, assuming that $\mu = 1$ for a non-magnetic material and using the equation for the refractive index above yields: $H_0 = nE_0$.

Non-polarized electromagnetic radiation, of course, comprises two perpendicular polarizations, which can change both in amplitude and in phase with respect to each other. If the two polarizations are in phase with each other, the resultant is just another linearly polarized beam, with the resultant polarization direction given by a simple vector addition of the

two electric field components. When the electric fields of the two polarizations are out of phase with each other, the resulting electric field precesses as it propagates through space. For example, if the two electric

fields are of equal magnitude but 90° out of phase with each other the electric field spirals as it moves through space. Depending on the relative phases, this rotation can be either clockwise or anti-clockwise. In this case, if we were able to look end on at the electric field vector, this would describe a circle and is therefore referred to as circularly polarized light. This particular combination of electric fields also carries with it angular momentum, and is responsible for the angular momentum selection rules in spectroscopy; this is really just the law of conservation of angular momentum. Different relative phases or amplitudes generally lead to elliptically polarized light, so that the phase shift Δ between the reflected and transmitted light measured in ellipsometry is manifested as elliptically polarized light. The different types of polarized light found as the phase shifts between 0 and 360° are shown in figure B1.26.10.

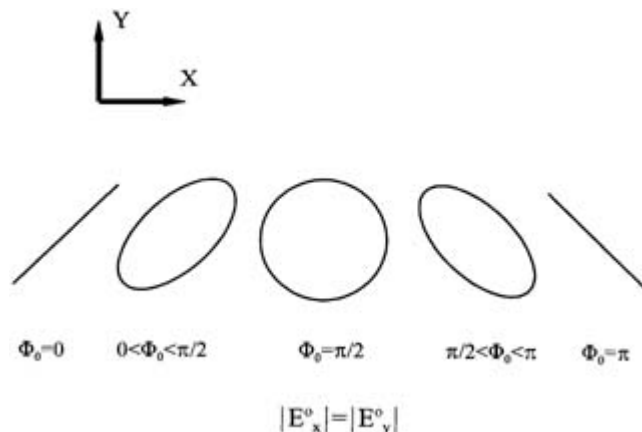


Figure B1.26.10. Various polarization configurations corresponding to different values of the phase shift, Φ

Light can also be absorbed by a material through which it passes. This leads to an attenuation in intensity of the light as it passes through the material, which decays exponentially as a function of distance through the material and is described mathematically by the Beer–Lambert law [34]:

$$I = I_0 \exp(-\epsilon cl) \quad (\text{B1.26.13})$$

where c is the concentration of the absorbant, l the path length and ϵ the extinction coefficient. This is represented in the mathematical description of the propagation of an electromagnetic wave by modifying the dielectric constant to add an imaginary part, k , which is generally written as: $n - ik$ where $i = \sqrt{-1}$.

(B) REFLECTION OF LIGHT FROM A DIELECTRIC SURFACE

Before discussing multiple reflections at a thin film, we will first examine the reflection of light from one material (with real refractive index n_1) into another (with refractive index n_2) [27, 32, 33]. It is assumed that the material does not absorb light and this situation is depicted in [figure B1.26.11](#). Since electromagnetic waves propagate in both materials, the behaviour at the interface is dictated by the boundary conditions for electric and magnetic fields. The components of \mathbf{E} and \mathbf{H} parallel to the surface (in the x direction in [figure B1.26.11](#)) and the components of \mathbf{D} and \mathbf{B}

perpendicular to the surface (in the z direction in [figure B1.26.11](#)) are continuous across the boundary. Thus the reflected and transmitted wave amplitudes can be calculated by simply applying these equations. We will consider the reflection of p-polarized radiation from the surface. The calculation of the equations for s-polarized radiation is essentially identical. The beam is incident at an angle θ_i , and a portion is reflected at an

angle θ_r . Of course, θ_i and θ_r are equal. The remaining light is transmitted at an angle θ_t which will be different to θ_i , due to refraction at the surface. The electric field amplitudes are taken to be E_i in the incident beam, E_r in the reflected beam and E_t in the transmitted beam. The corresponding values for magnetic field amplitudes are H_i , H_r and H_t respectively. Application of the above boundary conditions gives:

$$E_i^0 \cos \theta_i - E_r^0 \cos \theta_r = E_t^0 \cos \theta_t \quad (\text{B1.26.14})$$

$$H_i^0 + H_r^0 = H_t^0. \quad (\text{B1.26.15})$$

Using the above relationship between the electric and magnetic field amplitudes from equation (B1.26.15):

$$n_1 E_i^0 + n_1 E_r^0 = n_2 E_t^0. \quad (\text{B1.26.16})$$

In order to calculate the reflected amplitude, E_r^0 can be eliminated from equation (B1.26.15) and equation (B1.26.16) to yield:

$$\frac{(E_i^0 - E_r^0) \cos \theta_i}{\cos \theta_t} = \frac{n_1(E_i^0 - E_r^0)}{n_2} \quad (\text{B1.26.17})$$

where we have used $\theta_i = \theta_r$. Writing the reflection coefficient for p-polarized radiation as r^p (which equals E_r^0/E_i^0) yields:

$$r^p = \frac{n_2 \cos \theta_i - n_1 \cos \theta_t}{n_2 \cos \theta_i + n_1 \cos \theta_t} \quad (\text{B1.26.18})$$

which is the Fresnel equation for p-polarized radiation. A similar analysis can be carried out for s-polarized radiation using the boundary conditions in a similar way to yield:

$$r^s = \frac{n_1 \cos \theta_i - n_2 \cos \theta_t}{n_1 \cos \theta_i + n_2 \cos \theta_t}. \quad (\text{B1.26.19})$$

Corresponding equations can also be written for the transmitted portion of the beam. For a non-absorbing sample, the refractive index is real, so that $\exp(i\phi)$ is real where ϕ is the phase shift on reflection. This means that ϕ is either 0 or 180°. If, however, the sample absorbs light, the refractive index can become complex, resulting in a phase change of the light. This, for example, must be taken into account when reflecting light from a metallic (mirror) surface.

The reflection coefficients r^p and r^s give the electric field in the reflected beam for each polarization. Since the intensity of light is proportional to the square of the electric field, the reflectances for s- and p-polarized light can be written as $R^p = |r^p|^2$ and $R^s = |r^s|^2$, respectively. These are plotted in [figure B1.26.12](#) for a light beam incident from air ($n = 1$) onto a material with refractive index $n = 3$. It is evident that p-polarized radiation is reflected to a much lesser extent than s-polarized radiation, and exclusively s-polarized radiation is reflected at the polarizing angle. This effect is exploited in polarized sunglasses (as mentioned above) to minimize the reflective glare from surfaces by only allowing p-polarized light to be transmitted. At normal incidence, these equations reduce to:

$$R^s = R^p = \left(\frac{n_2 - 1}{n_2 + 1} \right)^2$$

which for glass with $n \sim 1.5$ gives about 4% reflectivity. The polarizing angle shown in [figure B1.26.12](#) is given by:

$$\tan \theta_i = \frac{n_2}{n_1}$$

and is known as the Brewster angle. p-polarized radiation is perfectly transmitted at this angle and Brewster windows (oriented at this angle) are used in lasers to minimize the loss of radiation in the laser cavity. This often results in laser light being polarized. This effect is exploited in polarimeters (see below). Note finally that the phase shift on reflection from a dielectric is 0° below the Brewster angle and 180° above.

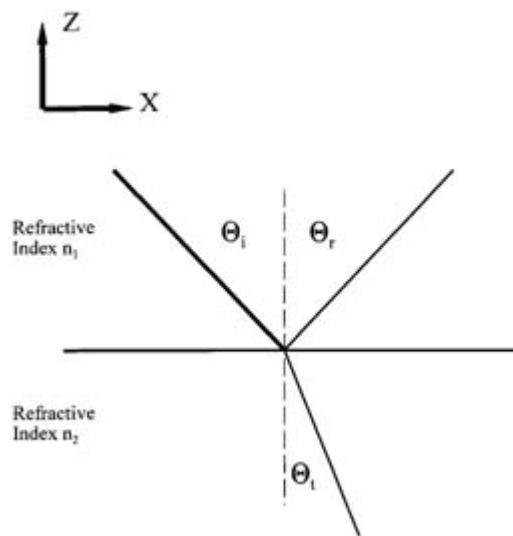


Figure B1.26.11. Diagram showing light impinging from a material of refractive index n_1 at an angle θ_i onto a material with refractive index n_2 and reflected at an angle θ_r and transmitted at an angle θ_t .

-17-

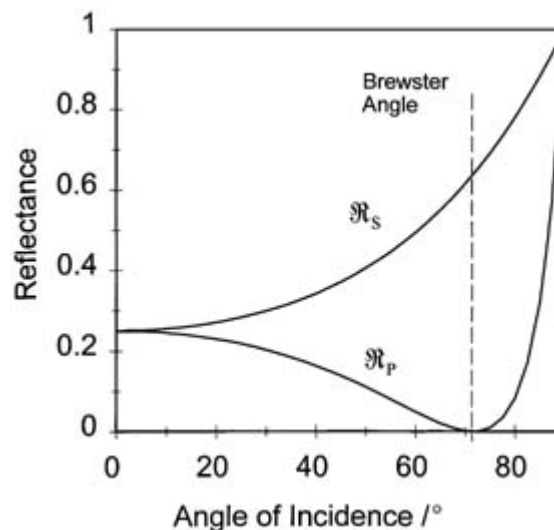


Figure B1.26.12. Plot of the reflectivity of s- and p-polarized light from a material with refractive index $n = 3$.

(C) REFLECTION AT MULTIPLE INTERFACES

We are now in a position to calculate the reflections from multiple interfaces using the simple example of a thin film of material of thickness d with refractive index n_2 sandwiched between a material of refractive index n_1 (where this is generally air with $n = 1$) deposited onto a substrate of refractive index n_3 [35, 36]. This is depicted in [figure B1.26.9](#). The resulting reflectivities for p- and s-polarized light respectively are given by:

$$R^p = \frac{r_{12}^p + r_{23}^p \exp(-i2\beta)}{1 + r_{12}^p r_{23}^p \exp(-i2\beta)} \quad (\text{B1.26.20})$$

and

$$R^s = \frac{r_{12}^s + r_{23}^s \exp(-i2\beta)}{1 + r_{12}^s r_{23}^s \exp(-i2\beta)} \quad (\text{B1.26.21})$$

where r_{ij}^s and r_{ij}^p are the reflection coefficients for s- and p-polarized radiation, respectively, at the interface between material i and j (equation (B1.26.18) and equation (B1.26.19)). The path length difference due to the film results in phase differences between different emerging beams giving rise to complex reflection coefficients and hence phase shifts. As noted above, this produces elliptically polarized light which can be analysed to yield the amplitude and phase shift and ultimately β . This parameter depends on the film thickness and the wavelength of light and is given by:

$$\beta = 2\pi \left(\frac{d}{\lambda} \right) n_2 \cos \theta_2. \quad (\text{B1.26.22})$$

-18-

These equations are generally too complex to be solved analytically even for relatively simpler systems and are therefore solved numerically. The way in which this is done will be described below. The measurement of Δ and Ψ clearly depends on the wavelength of light directly through this equation. However, both the real and imaginary parts of the refractive indices also depend on the wavelength of light. Now the reflectivity of the surface for s- and p-polarized light is $\mathfrak{R}^p = |r^p|^2$ and $\mathfrak{R}^s = |r^s|^2$ respectively. This dependence is also exploited in ellipsometry by measuring Δ and Ψ and a function of light wavelength in a technique known as spectroscopic ellipsometry [37].

B1.26.3.2 APPLICATIONS

(A) MEASUREMENT OF THE OPTICAL CONSTANTS OF MATERIALS USING ELLIPSOMETRY

In this case, the Fresnel (equation (B1.26.18) and equation (B1.26.19)) for reflection at a single interface are used. The phase shift is zero or 180° for a dielectric with a real refractive index, which can be measured directly from the reflectivities ([figure B1.26.12](#)). Intermediate phase shifts are found for absorbant materials with complex refractive indices which can also be measured from Δ and Ψ . This is generally done by numerically solving the Fresnel equation (B1.26.18) and equation (B1.26.19). Many ellipsometers include software to calculate these values directly and Fortran programs are also available to calculate these values. The variation in Ψ plotted as a function of the imaginary part of the refractive index is shown in [figure B1.26.13](#). This varies between 0 and 45° over the range of K values. The variation in Δ as a function of K with

$n = 2$ and an incidence angle of 70° is displayed in [figure B1.26.14](#). Again, the value of Δ varies as the imaginary part of the refractive index changes and can vary between 0 and 180° . It is important to emphasize that, unless a material is extremely pure, its optical constants can vary so that it is generally important to measure these parameters for a particular sample.

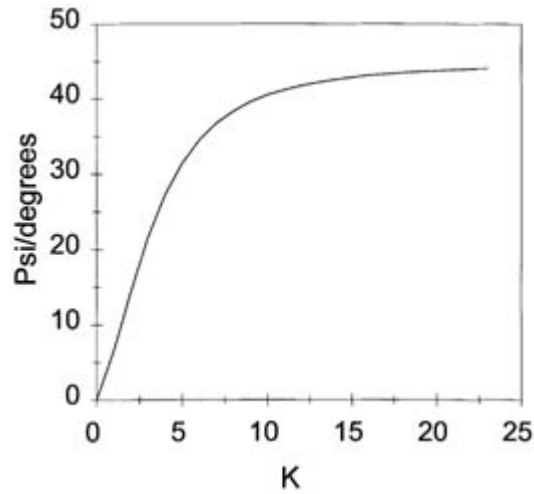


Figure B1.26.13. Plot of Ψ versus K , the imaginary part of the refractive index.

-19-

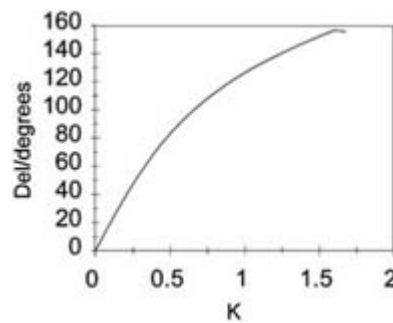


Figure B1.26.14. Plot of Δ versus K , the imaginary part of the refractive index.

(B) MEASUREMENT OF FILM THICKNESS AND OPTICAL PROPERTIES

The way in which ellipsometry can be used to measure film thickness will be illustrated for the simple case of a material with thickness d and complex refractive index n_2 deposited onto a substrate with complex refractive index n_3 with light incident from a material of refractive index n_1 . Since light is generally incident from air, n_1 is usually taken to be unity. This is done by measuring Δ and Ψ and using [equation \(B1.26.20\)](#) and [equation \(B1.26.21\)](#) to calculate the optical properties of the film and its thickness. Since, in addition to the film thickness, there are potentially six other variables in the system (the real and imaginary parts of each of the refractive indices) whereas only two parameters Δ and Ψ are measured, several of these must be determined independently. The substrate parameters can, for example, be measured prior to film deposition as described in the previous section and the refractive index of air is known. The variation in Δ and Ψ with the thickness d of a film of silicon dioxide (with refractive index 1.46) deposited onto a silicon substrate (with refractive index $3.872 - i0.037$) is shown plotted in [figure B1.26.15](#) [37]. This yields a trajectory of the allowed values of Δ and Ψ as the film thickness. Since the film thickness is measured from the interference between the light reflected from the boundary between air and the film and that reflected from the interface between the film and the substrate, the maximum film thickness d_{\max} that can be measured before the

trajectory shown in [figure B1.26.15](#) retraces itself is given by:

$$2d_{\max} \cos \theta_t = \frac{\lambda}{n_2} \quad (\text{B1.26.23})$$

where n_2 is included to take account of the change in wavelength as the light passes through the film with this refractive index. The angle of the transmitted light (θ_t) and the incidence angle (θ_i) are related through Snell's law so that:

$$\cos \theta_t = \sqrt{1 - \frac{\sin^2 \theta_i}{n_2^2}} \quad (\text{B1.26.24})$$

assuming that the light is incident from air ($n = 1$). Substitution into equation (B1.26.23) yields a value of d_{\max} as:

-20-

$$d_{\max} = \frac{\lambda}{2\sqrt{n_2^2 - \sin^2 \theta_2}} \quad (\text{B1.26.25})$$

This shows that the values of Δ and Ψ are identical for films of thickness nd_{\max} where n is an integer. Thus the trajectory shown in [figure B1.26.15](#) retraces itself for multiples of this maximum thickness. For the silicon dioxide film deposited onto silicon, this yields a value of $d_{\max} = 2832 \text{ \AA}$. Thus, in [figure B1.26.15](#) if values Δ and Ψ were measured to be 90 and 25° respectively, this would correspond to a film of $600 + n \times 2832 \text{ \AA}$ thick where $n = 0, 1, 2$ etc.

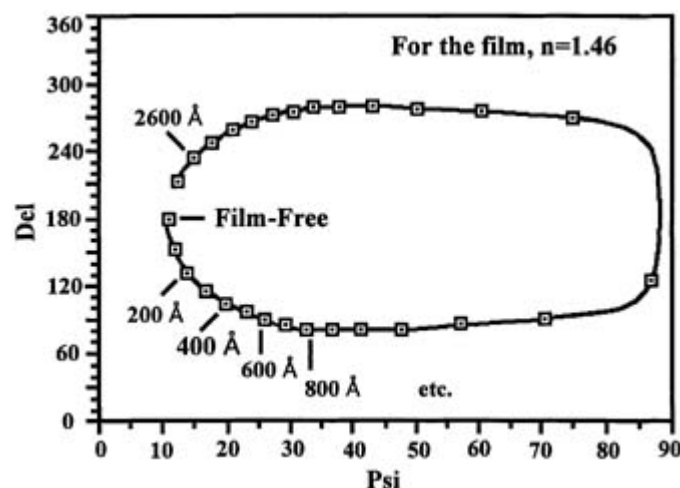


Figure B1.26.15. The Del/Psi trajectory for silicon dioxide on silicon with angle of incidence $\phi_1 = 70^\circ$ and wavelength $\lambda = 6328 \text{ \AA}$ (Tompkins H G 1993 *A Users Guide to Ellipsometry* (San Diego, CA: Academic)).

B1.26.3.3 INSTRUMENTATION

The following components make up an ellipsometer:

- (1) *Monochromatic light source.* This is generally a small laser, usually a helium–neon laser emitting red light at 6328 Å. The plasma tube of these lasers is usually terminated by a Brewster window (see [section B1.26.3.1](#)) to minimize losses in the laser cavity so that the light emitted by the laser is linearly polarized. If an unpolarized source is used, a polarizer is placed after the light source to produce linearly polarized light.
- (2) *An element that converts linearly polarized light into elliptically polarized light.* This component is made of a birefringent material where the refractive index of light that passes through the material depends on the light polarization with respect to its crystal lattice [33] where there are generally two perpendicular axes with different refractive indices for light polarized along each of these directions. Since, as shown above, the velocity of light through a medium depends on the refractive index, where the velocity of light v in the medium is given by $v = c/n$, this implies that light travels at different velocities depending on the polarization with respect to each of these directions. When the light is polarized along the direction of the smallest refractive index, the light travels the fastest and this is known as the ‘fast’ axis. Since light polarized along the other, higher-refractive-index axis travels more slowly, this is known as the ‘slow’ axis. The refractive indices of the fast and slow axis are

-21-

designated n_f and n_s , respectively, where it follows from the definitions that $n_s > n_f$. If we now imagine linearly polarized light incident on this material with the electric field vector oriented at 45° to each of these axes, the electric field of the incident light along both the slow and fast axes will be equal at $E_0 \cos 45^\circ$, where E_0 is the electric field amplitude of the polarized light. They will also be in phase. The frequency of the light, ν_0 , will be the same for both components within the material, but they will travel with different velocities. This means that after traversing a plate of this material of thickness d , the two components will no longer be in phase with each other so that, according to the discussion in [section B1.26.3.1](#), the light will, in general, be elliptically polarized. In order to calculate this, we first calculate the time required for light polarized along the fast and slow axes to traverse a disc of the birefringent material of thickness d . This is given by $t_s = d/v_s$ and $t_f = d/v_f$ for the slow and fast axes respectively. The difference in transit time for the two beams, $\Delta t = t_s - t_f$. If the period of oscillation of the light is τ ($= 1/\nu_0$), then if $\Delta t = \tau$, the slow beam is one period behind the fast beam and the phase difference is 2π radians. The phase difference for intermediate values of Δt designated $\Delta\phi$ is given by $(2\pi\Delta t/\tau)$. Combining these equations yields:

$$\Delta\phi = \frac{2\pi\nu_0 d}{c}(n_s - n_f). \quad (\text{B1.26.26})$$

If the wavelength of the incident radiation *in vacuo* is λ_0 , equation B1.26.26 becomes:

$$\Delta\phi = \frac{2\pi d}{\lambda_0}(n_s - n_f) \quad (\text{B1.26.27})$$

so the value of $\Delta\phi$ can be selected to be any desired value merely by varying the value of d for a material with particular values of n_s , n_f and λ . It is common to select $\Delta\phi = 90^\circ = (2\pi)/4$ radians, and this is known as a quarter-wave plate for a particular wavelength and produces circularly polarized light from linearly polarized light if the polarization direction of the incident beam is oriented at 45° to the fast and slow axes. Orientating the incident polarization to intermediate angles with respect to the fast and slow axes yields elliptically polarized light. This effect is exploited in the ellipsometer.

- (3) A polarizer which transmits only one polarization of radiation is required to define the state of polarization of the reflected beam.
- (4) The intensity of the reflected light must also be measured. Historically, this was done using the eye. Since, in general, a null (a measurement of the point at which the light decreases to zero) is required, this can be relatively sensitive. However, nowadays, the light intensity is generally measured using a photomultiplier tube.

- (5) These components must be mounted so that the incident and detection angles can be varied and kept equal and a place must be provided to mount the sample.

There are several possible configurations used to construct an ellipsometer [44]. We will describe one example of the one of the most common arrangements shown in [figure B1.26.16](#). The linearly polarized light emerging from the laser passes through the quarter-wave plate, which can be rotated to yield elliptically polarized light. When this reflects from the sample this also produces elliptically polarized light. The quarter-wave plate is rotated to find the condition such that, when the elliptically polarized light interacts with the sample, the phase change produced on reflection exactly compensates for the elliptical polarization of the incident light to produce linearly polarized light. The angle of the resulting linearly polarized light can be accurately determined using the polarizer placed before the detector (known as the analyser) by rotating it so that no light reaches the detector (the null condition). Being able to achieve this depends,

-22-

of course, on the quarter-wave plate being correctly oriented so as to exactly compensate for the effect of the sample and the analyser being oriented at exactly 90° to the direction of the resulting linearly polarized light. The experiment then consists of rotating both the quarter-wave plate and the analyser so that no light reaches the detector. This is often done automatically, and these resulting quarter-wave plate and analyser angles can be simply converted into the parameters Δ and Ψ [37, 44].

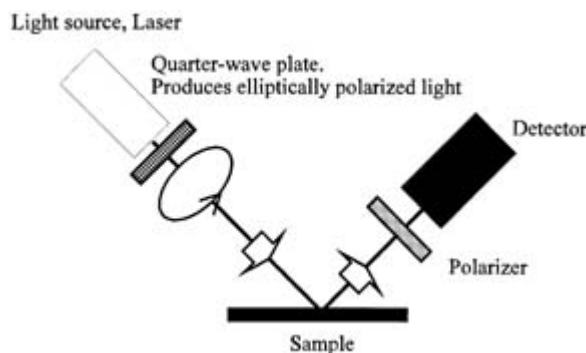


Figure B1.26.16. Schematic diagram of an ellipsometer.

B1.26.3.4 EXAMPLES

(A) BASIC STUDIES

Dielectric constants of metals, semiconductors and insulators can be determined from ellipsometry measurements [38, 39]. Since the dielectric constant can vary depending on the way in which a film is grown, the measurement of accurate film thicknesses relies on having accurate values of the dielectric constant. One common procedure for determining dielectric constants is by using a Kramers–Kronig analysis of spectroscopic reflectance data [39]. This method suffers from the series-termination error as well as the difficulty of making corrections for the presence of overlayer contaminants. The ellipsometry method is for the most part free of both these sources of error and thus yields the most accurate values to date [39].

(B) CHARACTERIZATION OF THIN FILMS AND MULTILAYER STRUCTURES

Ellipsometry measurements can provide information about the thickness, microroughness and dielectric function of thin films. It can also provide information on the depth profile of multilayer structures non-destructively, including the thickness, the composition and the degree of crystallinity of each layer [39]. The measurement of the various components of a complex multilayered film is illustrated in [figure B1.26.17](#) [40].

This also illustrates the use of different wavelengths of light to obtain much more information on the nature of the film. Here Δ and Ψ are plotted versus the wavelength of light (\bullet) and the line drawn through these data represents a fit calculated for the various films of yttrium oxide deposited on silica as shown at the bottom of the figure [40].

-23-

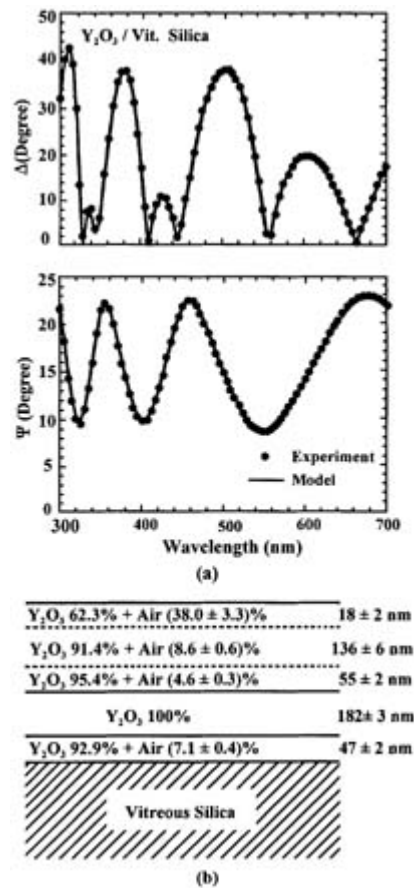


Figure B1.26.17. (a) Observed and calculated ellipsometric $[\Delta(\lambda), \Psi(\lambda)]$ spectra for the Y_2O_3 film on vitreous silica. Angle of incidence 75° . (b) Best-fit model of the Y_2O_3 film on vitreous silica (Chindaudom P and Vedam K 1994 *Physics of Thin Films* vol 19, ed K Vedam (New York: Academic) p 191).

(C) REAL-TIME STUDIES

With the development of multichannel spectroscopic ellipsometry, it is possible now to use real-time spectroscopic ellipsometers, for example, to establish the optimum substrate temperature in a film growth process [41, 42].

B1.26.4 WORK-FUNCTION MEASUREMENTS

B1.26.4.1 PRINCIPLES

The work function (Φ) is defined as the minimum work that has to be done to remove an electron from the bulk of the material to a sufficient distance outside the surface such that it no longer experiences an interaction with the surface electrostatic field [43, 44 and 45]. In other words, it is the minimum energy required to remove an electron from the

highest occupied level (the ‘Fermi level’) of a solid, through the surface, to the so-called vacuum reference level (figure B1.26.18). Thus it is influenced by two factors. The first is associated with the bulk electronic properties of the solid: work function increases with increasing binding energy. The second is associated with penetrating the surface dipole layer: work function changes with surface contamination and structure.

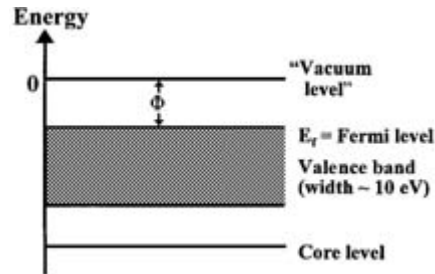


Figure B1.26.18. Schematic diagram of the energy levels in a solid.

In order to understand the tendency to form a dipole layer at the surface, imagine a solid that has been cleaved to expose a surface. If the truncated electron distribution originally present within the sample does not relax, this produces a steplike change in the electron density at the newly created surface (figure B1.26.19(A)).

Since the electron density $\rho(x) \propto |\psi(x)|^2$, where $\psi(x)$ is the electron wavefunction, this implies that the electron wavefunction varies in a similarly step-wise fashion at the interface. This indicates that $d^2\psi/dx^2|_s$, where s indicates that the derivative is evaluated at the surface, becomes infinite. Since the electron kinetic energy $E_K = (-\hbar^2/2m) (d^2\psi/dx^2)$, this creates an infinite-energy surface. This energy can be decreased by reducing $d^2\psi/dx^2$ and by allowing the wavefunction to become ‘smoother’ at the interface as shown in figure B1.26.19(B). This means that electron density previously within the sample extends outside the sample, producing a negative charge. Since the sample was originally electrically neutral, the excess charge outside the sample is balanced by a corresponding positive charge within it, resulting in an electric dipole moment at the surface. The work required to separate these charges increases the potential energy at the same time as the kinetic energy decreases. The equilibrium surface dipole moment corresponds to the minimum in this energy and this has been discussed in detail by Smoluchowski [46, 47]. In general, the greater the electron density of the sample, the larger will be the surface dipole. Thus, for the same metal, close-packed surfaces generally have the highest work functions; for example, in the case of copper, the work functions of the various surfaces are Cu(111): $\Phi = 4.94$ eV, Cu(100): $\Phi = 4.59$ eV, Cu(110): $\Phi = 4.48$ eV [45]. It is the presence of this surface dipole that renders the work function sensitive to changes in surface properties. For example, the surface dipole layer may change as a result of adsorption. Adsorbed species can be viewed as having a discrete dipole moment that tends to modify the total dipole layer at the surface by charge transfer and consequently change the work function. Thus, measurement of the work function change, $\Delta\Phi = \Phi_{\text{adsorbatecovered}} - \Phi_{\text{clean}}$ yields important information on the degree of charge reorganization upon adsorption and the surface coverage of the adsorbate. For example, adsorption of an electronegative species (for example, chlorine, hydrogen etc) will tend to increase the surface dipole moment causing an increase in work function ($\Delta\Phi$ positive).

Correspondingly, an electropositive adsorbate (for example, an alkali metal) decreases the surface dipole moment causing a decrease in the work function ($\Delta\Phi$ negative). These changes can be used to measure the adsorbate coverage as illustrated in figure B1.26.20. Here it is assumed that there are N adsorbates per unit area each having an effective charge q . This is balanced by an equal and opposite ‘image charge’ in the substrate a distance d away so that each adsorbate possesses a dipole moment $\mu = qd$. The separated layers of positive and negative charge can be thought of as forming a parallel-plate capacitor, where the potential difference between the capacitor plates corresponds to the change in work function of the sample. If the total charge per unit area on each of the plates is Q , then:

$$Q = C \Delta \Phi \tag{B1.26.28}$$

where C is the capacitance per unit area and is ϵ_0/d . Since $Q = Nq$, equation (B1.26.28) becomes:

$$Nq = \frac{\Delta \phi \epsilon_0}{d} \tag{B1.26.29}$$

which, remembering that $\mu = qd$, yields the Helmholtz equation:

$$\Delta \Phi = \frac{N\mu}{\epsilon_0} \tag{B1.26.30}$$

and shows that, for this simple case, the change in work function varies linearly with the adsorbate coverage N and the surface dipole moment of the adsorbate μ . A simplifying assumption in these equations is that either the coverage or the surface dipole moment or both are sufficiently small that the dipoles do not interact. If the distance between the dipoles decreases (as the coverage becomes larger) and/or if the dipole moment is large, the electric field created by one dipole can polarize adjacent dipoles to reduce their dipole moments. This effect is known as depolarization. This situation has been described by Topping [48] and Miller [49] and the resulting change in work function taking these effects into account is given by:

$$\Delta \Phi = \frac{N\mu(1 + 9\alpha N^{3/2})}{\epsilon_0} \tag{B1.26.31}$$

where α is the polarizability of the adsorbate. This reduces to equation (B1.26.30) for low coverages.

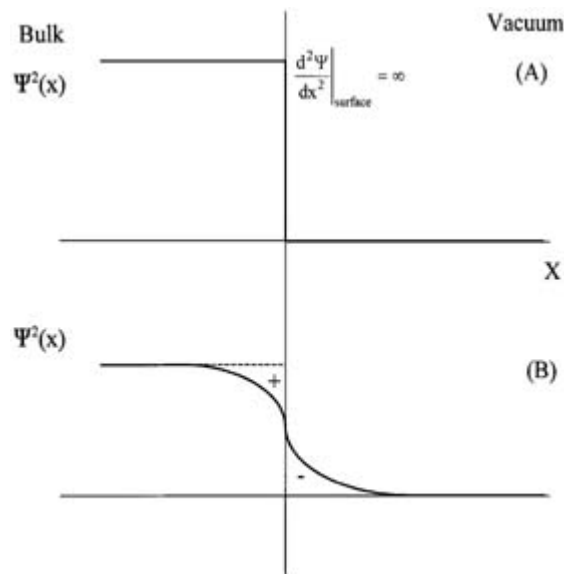


Figure B1.26.19. The variation of the electron density (A) from an unrelaxed surface and (B) showing the smoothing of the electron density to lower the kinetic density.

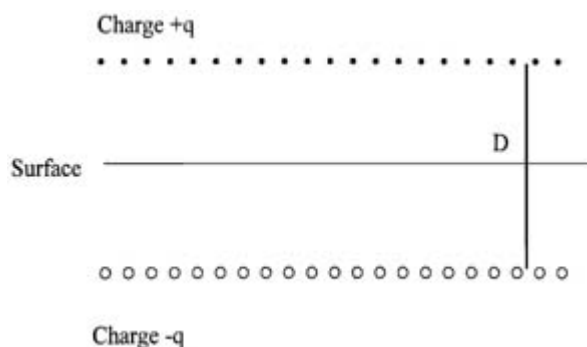


Figure B1.26.20. Diagram showing the dipole layer created on a surface by an electropositive adsorbate.

B1.26.4.2 INSTRUMENTATION

There are many ways to measure the work function or work function change which can be generally classified as electron emission methods (thermionic emission, field emission and photoelectron emission), low-energy electron beam (retarding-potential) methods and capacitance methods [50, 51 and 52]. Absolute work function values can be measured using emission methods while the other techniques measure only work-function changes.

-27-

The probes for measuring surface work functions are generally incorporated into an ultrahigh vacuum apparatus and supplement the existing vacuum-compatible, surface-sensitive probes (see for example [section B1.7](#), [section B1.9](#), [section B1.20](#), [section B1.21](#) and [section B1.25](#)). Rather than measuring the absolute value of the work function, it is often more interesting to measure the change in work function caused by some change to the surface. It can either be measured by modifying equipment already present in the vacuum system, for example using the electron gun of low-energy electron diffraction optics for the electron beam method or from the high-binding-energy cut-off in ultraviolet photoelectron spectroscopy. Specific probes for rapidly and conveniently measuring work-function changes can also be introduced separately into the chamber. The most common of these is the Kelvin probe or vibrating capacitor.

(A) THERMIONIC EMISSION METHOD

When a metal sample is heated, electrons are emitted from the surface when the thermal energy of the electrons, kT , becomes sufficient to overcome the work function Φ [53]. The probability of this electron emission depends on work function Φ and temperature T as expressed in the Richardson–Dushman equation:

$$J = A(1 - r)T^2 \exp(-e\Phi/kT) \quad (\text{B1.26.32})$$

where J is the thermionic emission current density, $A = 120 \text{ A cm}^{-2} \text{ deg}^{-2}$ and r is the reflection coefficient for electrons arriving at the work function barrier. Thus, plotting $\ln(J/T^2)$ against $1/kT$ yields a straight line with slope equal to $e\Phi$. The method is not generally suitable for monitoring adsorbates since the sample has to be heated to emit electrons.

(B) FIELD EMISSION METHOD

The barrier to electron removal from a surface can be reduced substantially by the presence of a strong

electric field. The physical situation involved in this process is shown in [figure B1.26.21](#) where, in the presence of an applied field, the work function is less than that at zero field. This increases the probability for electrons to ‘tunnel’ out of the sample [54]. This is a quantum mechanical phenomenon which can be formulated mathematically by considering a Fermi sea of electrons within the metal impinging on a potential barrier at the surface. The result is given by the Fowler–Nordheim equation as:

$$J = 6 \times 10^6 \left(\frac{\mu}{\Phi(\mu + \Phi)} \right) \left(\frac{E}{\alpha} \right)^2 \exp \left(\frac{-(6.8 \times 10^7 \alpha \Phi^{3/2})}{E} \right) \quad (\text{B1.26.33})$$

where J is the current arising from field emission, E is the electric field strength and α is a tabulated function equal to 0.95 ± 0.009 over the range of current densities normally encountered. A plot of $\ln(J/E^2)$ against $(1/E)$ can be used to determine Φ .

-28-

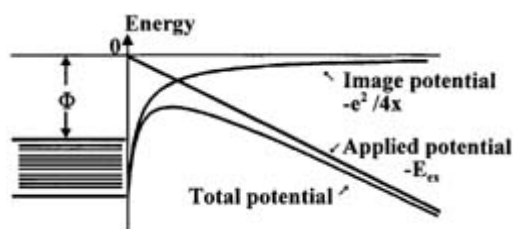


Figure B1.26.21. Potential energy curves for an electron near a metal surface. ‘Image potential’ curve: no applied field. ‘Total potential’ curve: applied external field = $-E_{\text{ex}}$.

This method needs a highly specialized sample configuration and the sample has to be a very sharp point (radius $\sim 10^{-5}$ cm) so that sufficiently high fields can be maintained for emission measurements. This requirement often precludes the use of other surface analysis techniques. However, for specialized applications, the field emission method may be the only one or the most convenient one available. It has been shown to be able to directly measure the work function change induced by the adsorption of a single atom on a tungsten plane [55]. Combined with field emission microscopy, the work function of different crystal planes can be detected. Much of the early understanding of adsorption was obtained with this device and a large number of data on single-crystal work functions have been produced by this technique [52].

(C) PHOTOELECTRON EMISSION METHOD

When photons of sufficiently high frequency ν are directed onto a metal surface, electrons are emitted in a process known as photoelectron emission [56]. The threshold frequency ν_0 is related to the work function by the expression

$$\Phi = h\nu_0/e. \quad (\text{B1.26.34})$$

The total electron current generated in this process is given by the Fowler equation:

$$J = BT^2 f \left[\frac{h(\nu - \nu_0)}{kT} \right] \quad (\text{B1.26.35})$$

where B is a parameter that depends on the material involved. The photoemitted current is measured as a function of photon energy; extrapolation allows the determination of ν_0 , and thus of Φ , to be made.

This technique requires a photon source (a light source with monochromator or filters) of calibrated spectral intensity and variable energy in the range around ν_0 , and an electron collector. Both the work function and the work-function change may be determined conveniently from the cut-off in inelastic electrons in a photoelectron spectrum [47]. As demonstrated by figure B1.26.18 electrons with the minimum kinetic energy barely surmount the work function, while electrons from the Fermi level (E_F , the highest occupied level) will have the maximum kinetic energy, $h\nu - \Phi$.

The work function can be calculated from the relationship

$$\Phi = h\nu - W \tag{B1.26.36}$$

where W is the energy width of the whole photoelectron spectrum. Adsorption changes the work function which in turn changes the width (W) of the UP spectrum. Φ and W are inversely related. The work-function change upon adsorption manifests itself in a shift in the low-kinetic-energy ‘secondary tail’, as shown in the inset in figure B1.26.22 [47].

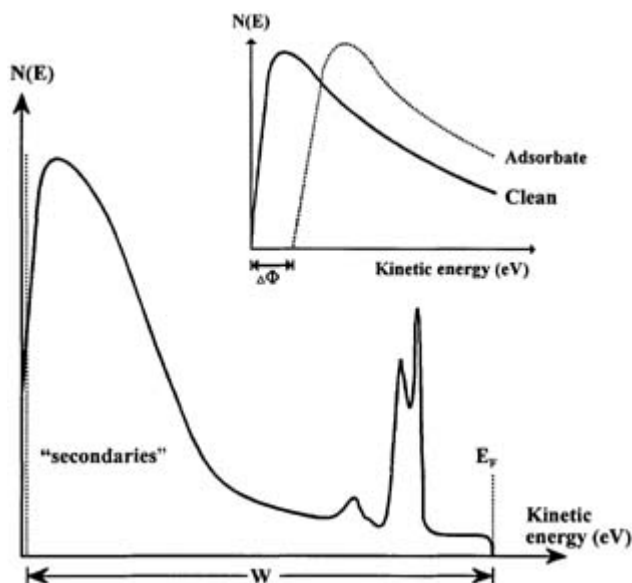


Figure B1.26.22. The energy width W of an ultraviolet photoelectron spectrum from a solid may be used to determine the work function. Changes in work function may be obtained from changes in the ‘cut-off’ of the secondary electron peak (inset) (Attard G and Barnes C 1988 *Surfaces* (Oxford: Oxford University Press)).

(D) LOW-ENERGY ELECTRON BEAM (RETARDING-POTENTIAL) METHODS

In this method, the sample is designed as an anode. Electrons emitted from the cathode by thermionic emission normally impinge on the anode sample on which a retarding potential is applied. Basically this can be arranged as a diode (known as the diode method) [57] or a triode (called the Shelton triode method) [58]. Different circuit arrangements give rise to different $I-V$ relationships and the difference in the work functions of the two electrode surfaces is measured. When a low-energy electron diffraction (LEED) apparatus is available (see section B1.9), this method can be implemented using the low-energy electrons from the LEED electron gun. In this case, electrons of known, low energy are incident on the sample. As the potential across

the sample is slowly made more negative, the current is measured and the relationship between sample and retarding voltage is shown in [figure B1.26.23](#) [50]. At low retarding voltage, most of the impinging electrons are collected. At some larger retarding voltage, the impinging electrons do not possess sufficient energy to reach the sample and so are reflected. This voltage is determined by the work functions of the sample and the electron gun filament, and by the accelerating voltage of the electron gun.

-30-

Changes in the work function of the sample, when the filament work function and the electron gun accelerating voltage are held constant, are manifested by a change in the cut-off retarding voltage (see [figure B1.26.23](#)).

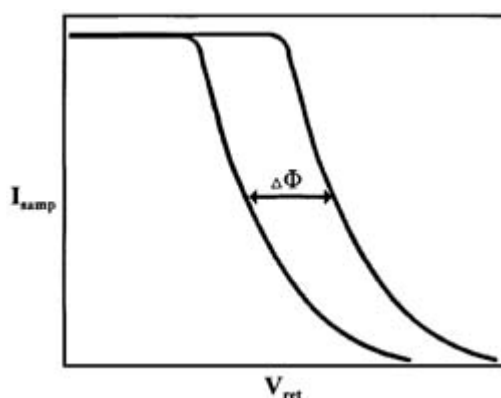


Figure B1.26.23. Current–voltage curves observed in the retarding potential difference method of work-function measurement (Hudson J B 1992 *Surface Science* (Stoneham, MA: Butterworth–Heinemann)).

A low-energy electron beam can also be obtained using a field emission tip and used in the field emission retarding-potential method. This combination provides an absolute measure of the sample work function and the resolution is excellent [52].

(E) CAPACITANCE METHODS

When an electrical connection is made between two metal surfaces, a contact potential difference arises from the transfer of electrons from the metal of lower work function to the second metal until their Fermi levels line up. The difference in contact potential between the two metals is just equal to the difference in their respective work functions. In the absence of an applied emf, there is electric field between two parallel metal plates arranged as a capacitor. If a potential is applied, the field can be eliminated and at this point the potential equals the contact potential difference of the two metal plates. If one plate of known work function is used as a reference electrode, the work function of the second plate can be determined by measuring this applied potential between the plates [52]. One can determine the zero-electric-field condition between the two parallel plates by measuring directly the tendency for charge to flow through the external circuit. This is called the static capacitor method [59].

Historically, the first and most important capacitance method is the vibrating capacitor approach implemented by Lord Kelvin in 1897. In this technique (now called the Kelvin probe), the reference plate moves relative to the sample surface at some constant frequency and the capacitance changes as the interelectrode separation changes. An AC current thus flows in the external circuit. Upon reduction of the electric field to zero, the AC current is also reduced to zero. Originally, Kelvin detected the zero point manually using his quadrant electrometer. Nowadays, there are many elegant and sensitive versions of this technique. A piezoceramic foil can be used to vibrate the reference plate. To minimize noise and maximize sensitivity, a phase-locked

detection circuit is used and a feedback loop may automatically null the electric field. The whole process can be carried out electronically to provide automatic recording of the contact potential and any changes which occur.

-31-

This technique does not involve heating the sample to high temperature or exposing it to high electric fields. Nor is there a need for a hot filament which cannot be used at high pressures. When mounted on a linear motion manipulator, the reference plate assembly can be moved away from the sample leaving no interference with other techniques. One major shortcoming of this technique is that the work function of the reference plate, or electrode, must be precisely known for an absolute determination of the sample work function. Moreover, when relative changes are examined, the work function of the reference electrode must be stable, either unaffected by adsorption or cleanable before each experiment. This disadvantage may be compensated for by using an inert metal such as gold. In addition, the reference electrode must be well shielded to reduce the effects of external electric and magnetic fields on the experiment [52].

B1.26.4.3 APPLICATIONS

(A) SURFACE CHARACTERIZATION

Measurement of the work function of a surface is an important part of overall surface characterization. Surface electron charge density can be described in terms of the work function and the surface dipole moment can be calculated from it (equation (B1.26.30) and equation (B1.26.31)). Likewise, changes in the chemical or physical state of the surface, such as adsorption or geometric reconstruction, can be observed through a work-function modification. For studies related to cathodes, the work function may be the most important surface parameter to be determined [52].

(B) MEASUREMENT OF ADSORPTION ISOTHERMS

Almost all adsorbates cause work function changes ($\Delta\Phi$). Plotting $\Delta\Phi$ versus pressure at various temperatures produces an adsorption isotherm which can be used to determine heats of adsorption and the surface coverage [60]. Much early understanding of adsorption was gained by this method. In particular alkali and alkali earth adsorption has been rather extensively studied in this way. This is illustrated in figure B1.26.24 for the adsorption of alkali metals on W(100) [61]. The change in work function is depicted as solid lines which initially decrease with increasing coverage as expected for an electropositive adsorbate. The dots connected by the dashed line represent the work functions of each of the bulk alkali metals and each of the work-function curves tends asymptotically to these values for large coverages. These curves are not linear as a function of coverage because of depolarization effects (equation (B1.26.31)) and so reach a minimum before attaining their bulk values.

-32-

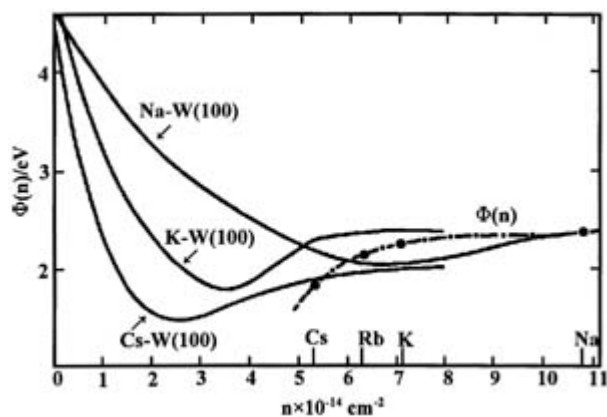


Figure B1.26.24. The change of work function of the (100) plane of tungsten covered by Na, K and Cs, and work function of alkali metals (dashed–dotted line) versus adatom concentration n (Kiejna A and Wojciechowski 1981 *Prog. Surf. Sci.* **11** 293–338).

REFERENCES

- [1] Clarke A 1970 *The Theory of Adsorption and Catalysis* (London: Academic)
- [2] Brunauer S, Emmett P H and Teller E 1938 Adsorption of gases in multimolecular layers *J. Am. Chem. Soc.* **60** 309–19
- [3] Halling J 1975 *Principles of Tribology* (New York: Macmillan) ch 1
- [4] Millikan R A 1916 A direct photoelectric determination of Planck's 'h' *Phys. Rev.* **7** 355–88
- [5] Yates J T and Garland C 1961 Infrared studies of carbon monoxide chemisorbed on nickel surfaces *J. Catal.* **65** 617–24
- [6] Scholten J J and van Montfoort A 1962 The determination of the free-metal surface area of palladium catalysts *J. Catal.* **1** 85–92
- [7] Stephens S J 1959 Surface reactions on evaporated palladium films *J. Phys. Chem.* **63** 188–94
- [8] Harkins W D 1952 *The Physical Chemistry of Surface Films* (New York: Reinhold)
- [9] Emmett P H and Brunauer S 1937 The use of low temperature van der Waals adsorption isotherms in determining the surface area of iron synthetic ammonia catalysts *J. Am. Chem. Soc.* **59** 1553–64
- [10] Brunauer S, Deming L S, Deming W S and Teller E A 1940 Theory of the van der Waals adsorption of gases *J. Am. Chem. Soc.* **62** 1723–32
- [11] Sing K S W, Everett D H, Haul R A W, Mosouf L, Pierotti R A, Rouguerol J and Siemieniewska T 1985 Reporting physisorption data to the determination of surface area and porosity *Pure Appl. Chem.* **57** 603–19
- [12] Langmuir I 1916 The constitution and fundamental properties of solids and liquids *J. Am. Chem. Soc.* **38** 2221–95

- [13] Eggers D F Jr, Gregory N W, Halsey G D Jr and Rabinovitch B S 1964 *Physical Chemistry* (New York: Wiley) ch 18

- [14] Castellan G W 1983 *Physical Chemistry* (Reading, MA: Addison-Wesley) ch 18
- [15] Donohue M D and Aranovich G L 1998 Classification of Gibbs adsorption isotherms *Adv. Colloid Interface Sci.* **76/77** 137–52
- [16] Harkins W D and Jura G 1944 An absolute method for the determination of the area of a finely divided crystalline solid *J. Am. Chem. Soc.* **66** 1362–6
- [17] Giles C H, Smith D and Huitson A 1974 General treatment and classification of the solute adsorption isotherm *J. Colloid Interface Sci.* **47** 755–65
- [18] Sing K S W 1998 Adsorption methods for the characterization of porous materials *Adv. Colloid Interface Sci.* **76/77** 3–11
- [19] Horvath G and Kawazoe K 1983 Method for calculation of effective pore size distribution in molecular sieve carbon *J. Chem. Eng. Japan* **16** 470–5
- [20] Lippens B C and deBoer J H 1965 Studies on pore systems in catalysts V. The *T* method *J. Catal.* **4** 319–23
- [21] Gregg S J and Sing K S W 1967 *Adsorption, Surface Area and Porosity* (London, UK: Academic) ch 8; 1982 2nd edn, ch 6
- [22] 'OMNISORP' Series, Beckman Coulter, USA; 'ASAP 2400', 'FlowSorb 2300', Micromeritics; 'Autosorb-1', 'nova 1000', Quantachrome, USA; 'Grimm 100', Labortechnik, Germany; 'Sorpty 1750', 'Sorptomatic', Carlo Erba, Italy
- [23] Beckman Coulter 1991 *OMNISORP Manual*
- [24] Nelsen F M and Eggertsen F T 1958 Determination of surface area: adsorption measurements by a continuous flow method *Anal. Chem.* **30** 1387–90
- [25] Rothen A 1945 The ellipsometer, an apparatus to measure thicknesses of thin surface films *Rev. Sci. Instrum.* **16** 26–30
- [26] Drude P 1901 *Theory of Optics* (New York: Longmans Green)
- [27] Jenkins F A and White H E 1957 *Fundamentals of Optics* (New York: McGraw-Hill)
- [28] Longhurst R S 1967 *Geometrical and Physical Optics* (New York: Wiley)
- [29] Klein M V 1970 *Optics* (New York: Wiley)
- [30] Welford W T 1988 *Optics* (Oxford: Oxford University Press)
- [31] Fincham W H A and Freeman M H 1980 *Optics* (London: Butterworths)
- [32] Born M and Wolf E 1969 *Principles of Optics* (New York: Pergamon)
- [33] Banerjee P P and Poon T C 1991 *Principles of Applied Optics* (Boston, MA: Asken)
- [34] Atkins P W 1998 *Physical Chemistry* (New York: Freeman) ch 19
- [35] Azzam R M A and Bashara N M 1977 *Ellipsometry and Polarized Light* (Amsterdam: North-Holland)
- [36] Hevens O S 1965 *Optical Properties of Thin Solid Films* (New York: Dover)
- [37] Tompkins H G 1993 *A Users Guide to Ellipsometry* (San Diego, CA: Academic)

- [38] Aspnes D E and Studna A A 1983 Dielectric functions and optical parameters of silicon and germanium *Phys. Rev. B* **27** 985–1009

- [39] Vedam K 1998 Spectroscopic ellipsometry: a historical overview *Thin Solid Films* **313/314** 1–9
- [40] Chindaudom P and Vedam K 1994 *Physics of Thin Films* vol 19, ed K Vedam (New York: Academic) p 191
- [41] Muller R H and Farmer J C 1984 Fast self-compensating spectral-scanning ellipsometer *Rev. Sci. Instrum.* **55** 371–4
- [42] Collins R W, An I, Fujiwara H, Lee J, Lu Y, Kol J and Rovira P I 1998 Advances in multichannel spectroscopic *Thin Solid Films* **313/314** 18–32
- [43] Riviere J C 1969 *Solid State Surface Science* vol I, ed M Green (New York: Dekker)
- [44] Wedler G 1976 *Chemisorption: an Experimental Approach* (London: Butterworth)
- [45] Smoluchowski R 1941 Volume magnetostriction of nickel *Phys. Rev.* **60** 249–51
- [46] Smoluchowski R 1941 The theory of volume magnetostriction *Phys. Rev.* **59** 309–17
- [47] Attard G and Barnes C 1998 *Surfaces* (Oxford: Oxford University Press)
- [48] Topping J 1927 Form and energy potential of atoms on surfaces *Proc. R. Soc. A* **114** 67–76
- [49] Miller A R 1946 The variation of the dipole moment of adsorbed particles with the fraction of the surface covered *Proc. Camb.Phil. Soc.* **42** 292–303
- [50] Hudson J B 1992 *Surface Science* (Stoneham, MA: Butterworth-Heinemann)
- [51] Woodruff D P and Delchar T A 1994 *Modern Technologies of Surface Science* 2nd edn (Cambridge: Cambridge University Press)
- [52] Swanson L W and Davis P R 1985 Work function measurements *Solid State Physics: Surfaces(Methods of Experimental Physics 22)* cd R L Park and M G Lagally (New York: Academic) ch1
- [53] Reimann A L 1934 *Thermionic Emission* (London: Chapman and Hall)

- [54] Gomer R 1994 Field emission, field ionization, and field desorption *Surf. Sci* **299/300** 129-52
- [55] Todd C J and Rhodin T N 1974 Adsorption of single alkali atoms on tungsten using field emission and field desorption *Surf. Sci.* **42** 109-21
- [56] Fowler R H 1931 The analysis of photoelectric sensitivity curves for clean metals and various temperatures *Phys. Rev.* **38** 45-56
- [57] Haas G A and Thomas R E 1972 Thermionic emission and work function *Tech. Met. Res.* **6** 91-262
- [58] Shelton H 1957 Thermionic emission from a planar tantalum crystal *Phys. Rev.* **107** 1553-7
- [59] Delchar T, Eberhagen A and Tompkins F C 1963 A static capacitor method for the measurement of the surface potential of gases on evaporated metal films *J. Sci. Instrum.* **40** 105-7
- [60] Conrad H, Ertl G and Latta E E 1974 Adsorption of hydrogen on palladium single crystal surfaces *Surf. Sci* **41** 435-46
- [61] Kiena A and Wojciechowski 1981 Work function of metals: relation between theory and experiment *Prog. Surf. Sci.* **11** 293-338

B1.27 Calorimetry

Kenneth N Marsh

B1.27.1 INTRODUCTION

Calorimetry is the basic experimental method employed in thermochemistry and thermal physics which enables the measurement of the difference in the energy U or enthalpy H of a system as a result of some process being done on the system. The instrument that is used to measure this energy or enthalpy difference (ΔU or ΔH) is called a calorimeter. In the first section the relationships between the thermodynamic functions and calorimetry are established. The second section gives a general classification of calorimeters in terms of the principle of operation. The third section describes selected calorimeters used to measure thermodynamic properties such as heat capacity, enthalpies of phase change, reaction, solution and adsorption.

B1.27.2 RELATIONSHIP BETWEEN THERMODYNAMIC FUNCTIONS AND CALORIMETRY

The first law of thermodynamics relates the energy change in a system at constant volume to the work done on the system w and the heat added to the system q ,

$$\Delta U = w + q. \quad (\text{B1.27.1})$$

Both heat and work are a flow of energy, heat being a flow resulting from a difference in temperature between the system and the surroundings and work being an energy flow caused by a difference in pressure or the application of other electromechanical forces such as electrical energy. For the case where the system is thermally isolated from its surrounding, termed *adiabatically enclosed*, there is no heat flow to or from the surrounds, i.e. $q = 0$ so that

$$\Delta U = w_{\text{adiabatic}}. \quad (\text{B1.27.2})$$

A calorimeter is a device used to measure the work w that would have to be done under adiabatic conditions to bring about a change from state 1 to state 2 for which we wish to measure $\Delta U = U_2 - U_1$. This work w is generally done by passing a known constant electric current \mathfrak{I} for a known time t through a known resistance R embedded in the calorimeter, and is denoted by w_{elec} where

$$w_{\text{elec}} = \mathfrak{I}^2 R t. \quad (\text{B1.27.3})$$

In general it is difficult to construct a calorimeter that is truly adiabatic so there will be unavoidable heat leaks q . It is also possible that non-deliberate work is done on the calorimeter such as that resulting from a change in volume against a non-zero external pressure $p_{\text{ext}}(-\int p_{\text{ext}} dV)$, often called pV work. Additional work w' may be done on the system by energy introduced from stirring or from energy dissipated due to self-heating of the device used to measure the temperature. The basic equation for the energy change in a calorimeter is

$$\Delta U = w_{\text{elec}} - \int p_{\text{ext}} dV + w' + q. \quad (\text{B1.27.4})$$

The pV work term is not normally measured. It can be eliminated by suspending the calorimeter in an evacuated space ($p = 0$) or by holding the volume of the calorimeter constant ($dV = 0$) to give

$$\Delta U = w_{\text{elec}} + w' + q. \quad (\text{B1.27.5})$$

This is the working equation for a constant volume calorimeter. Alternatively, a calorimeter can be maintained at constant pressure p equal to the external pressure p_{ext} in which case

$$- \int_{V_1}^{V_2} p_{\text{ext}} dV = p(V_2 - V_1) \quad (\text{B1.27.6})$$

and

$$\Delta U = U_2 - U_1 = w_{\text{elec}} - pV_1 - pV_2 + w' + q \quad (\text{B1.27.7})$$

hence

$$\Delta(U + pV) = (U + pV)_2 - (U + pV)_1 = w_{\text{elec}} + w' + q. \quad (\text{B1.27.8})$$

The quantity $U + pV$ is termed the enthalpy H , hence

$$\Delta H = H_2 - H_1 = w_{\text{elec}} + w' + q. \quad (\text{B1.27.9})$$

is the working equation for a constant pressure calorimeter.

The heat capacity at constant volume C_V is defined from the relations

-3-

$$\Delta U = \int_{T_1}^{T_2} C_V dT \quad (\text{B1.27.10})$$

and

$$C_V \stackrel{\text{def}}{=} (\partial U / \partial T)_V \stackrel{\text{def}}{=} \lim_{T_2 \rightarrow T_1} \{ [U(T_2, V_2) - U(T_1, V_1)] / (T_2 - T_1) \}. \quad (\text{B1.27.11})$$

Values of $C_V(T)$ can be derived from a constant volume calorimeter by measuring ΔU for small values of $(T_2 - T_1)$ and evaluating $\Delta U / (T_2 - T_1)$ as a function of temperature. The energy change ΔU can be derived from a knowledge of the amount of electrical energy required to change the temperature of the sample + container

from T_1 to T_2 , w_{elec} (sample + container), and the energy required to change the temperature of the container only from T_1 to T_2 , w_{elec} (container). If the volume of the sample is kept constant, and the calorimeter is adiabatic ($q = 0$) or the heat leak is independent of the amount of sample and no other work is done then:

$$\Delta U = U(T_2, V) - U(T_1, V) = w_{\text{elec}}(\text{sample + container}) - w_{\text{elec}}(\text{container}). \quad (\text{B1.27.12})$$

Except for gases, it is very difficult to determine C_V . For a solid or liquid the pressure developed in keeping the volume constant when the temperature is changed by a significant amount would require a vessel so massive that most of the total heat capacity would be that of the container. It is much easier to measure the difference

$$\Delta H = H(T_2, p) - H(T_1, p) = w_{\text{elec}}(\text{sample + container}) - w_{\text{elec}}(\text{container}) \quad (\text{B1.27.13})$$

between the enthalpies of the initial and final states when the pressure is kept constant and derive the heat capacity at constant pressure C_p defined by

$$C_p \stackrel{\text{def}}{=} (\partial H / \partial T)_p \stackrel{\text{def}}{=} \lim_{T_2 \rightarrow T_1} [(H(T_2, V_2) - H(T_1, V_1)) / (T_2 - T_1)]. \quad (\text{B1.27.14})$$

The enthalpy change ΔH for a temperature change from T_1 to T_2 can be obtained by integration of the constant pressure heat capacity

$$\Delta H = \int_{T_1}^{T_2} C_p \, dT. \quad (\text{B1.27.15})$$

The entropy change ΔS for a temperature change from T_1 to T_2 can be obtained from the following integration

$$\Delta S = \int_{T_1}^{T_2} (C_p / T) \, dT. \quad (\text{B1.27.16})$$

B1.27.3 OPERATING PRINCIPLE OF A CALORIMETER

All calorimeters consist of the calorimeter proper and its surround. This surround, which may be a jacket or a bath, is used to control the temperature of the calorimeter and the rate of heat leak to the environment. For temperatures not too far removed from room temperature, the jacket or bath usually contains a stirred liquid at a controlled temperature. For measurements at extreme temperatures, the jacket usually consists of a metal block containing a heater to control the temperature. With non-isothermal calorimeters (calorimeters where the temperature either increases or decreases as the reaction proceeds), if the jacket is kept at a constant temperature there will be some heat leak to the jacket when the temperature of the calorimeter changes. Hence, it is necessary to correct the temperature change observed to the value it would have been if there was no leak. This is achieved by measuring the temperature of the calorimeter for a time period both before and after the process and applying Newton's law of cooling. This correction can be reduced by using the technique of adiabatic calorimetry, where the temperature of the jacket is kept at the same temperature as the calorimeter as a temperature change occurs. This technique requires more elaborate temperature control and it is primarily used in accurate heat capacity measurements at low temperatures.

With most non-isothermal calorimeters, it is necessary to relate the temperature rise to the quantity of energy released in the process by determining the calorimeter constant, which is the amount of energy required to increase the temperature of the calorimeter by one degree. This value can be determined by electrical calibration using a resistance heater or by measurements on well-defined reference materials [1]. For example, in bomb calorimetry, the calorimeter constant is often determined from the temperature rise that occurs when a known mass of a highly pure standard sample of, for example, benzoic acid is burnt in oxygen.

B1.27.4 CLASSIFICATION OF CALORIMETERS

B1.27.4.1 CLASSIFICATION BY PRINCIPLE OF OPERATION

ISOTHERMAL CALORIMETERS (MORE PRECISELY, QUASI-ISOTHERMAL)

These include calorimeters referred to as calorimeters with phase transitions. The temperatures of the calorimeter (i.e. the vessel) $T(c)$ and the jacket $T(s)$ in such calorimeters remain constant throughout the experiment. For calorimeters with phase transition, the calorimetric medium is usually a pure solid (its stable modification) in equilibrium with the liquid phase of the same substance, for example, ice and water. The reaction chamber is placed inside a vessel inside the layer of this substance. The jacket also contains an equilibrium mixture of two phases of the same substance. For an exothermic process, part of the solid substance melts in the vessel, and the volume change of the liquid is precisely measured. Another calorimeter of this type is an isothermal titration calorimeter. One of the reactants is added at such a rate that, for an endothermic system, the enthalpy or energy change is balanced by the simultaneous addition of electrical energy so the calorimeter remains isothermal. The energy added is a direct measure of the energy or enthalpy change. For an exothermic system electrical energy that is added at a constant rate is counterbalanced by the removal of energy at a constant rate (by, for example, a thermoelectric cooling device) to maintain the calorimeter isothermal. The reactant is then added at such a rate that the calorimeter remains isothermal when the addition of electrical energy is discontinued.

-5-

ADIABATICALLY-JACKETED CALORIMETERS

The energy released when the process under study takes place makes the calorimeter temperature $T(c)$ change. In an adiabatically jacketed calorimeter, $T(s)$ is also changed so that the difference between $T(c)$ and $T(s)$ remains minimal during the course of the experiment; that is, in the best case, no energy exchange occurs between the calorimeter (unit) and the jacket. The thermal conductivity of the space between the calorimeter and jacket must be as small as possible, which can be achieved by evacuation or by the addition of a gas of low thermal conductivity, such as argon.

HEAT-FLOW CALORIMETERS

These calorimeters are enclosed in a thermostat ('heat sink') which has a much greater heat capacity than that of the calorimeter vessel proper. The energy released in the calorimeter is negligibly small compared with the heat capacity of the thermostat, and hence the thermostat temperature $T(s)$ does not change. The outer surface of the calorimeter (vessel) is in direct thermal contact with the inner surface of the thermostat, and the energy flow occurs through a series of thermopiles, which consist of a large number of thermocouples connected in series. The flow of energy through the thermocouples gives rise to a voltage. The thermopiles are designed so that the majority of energy that flows from the calorimeter to the thermostat flows through them as rapidly as possible and the area under the curve of the voltage produced against time is a measure of the overall quantity of the energy released (or taken up) in the process occurring in the calorimeter.

ISOPERIBOLE CALORIMETERS

This type of calorimeter is normally enclosed in a thermostatted-jacket having a constant temperature $T(s)$ and the calorimeter (vessel) temperature $T(c)$ changes through the energy released as the process under study proceeds. The thermal conductivity of the intermediate space must be as small as possible. Most combustion calorimeters fall into this group.

B1.27.4.2 CLASSIFICATION BY DESIGN

LIQUID CALORIMETERS

A liquid serves as the calorimetric medium in which the reaction vessel is placed and facilitates the transfer of energy from the reaction. The liquid is part of the calorimeter (vessel) proper. The vessel may be isolated from the jacket (isoperibole or adiabatic), or may be in good thermal contact (heat-flow type) depending upon the principle of operation used in the calorimeter design.

ANEROID (LIQUIDLESS) CALORIMETERS

The reaction vessel is situated inside a metal of high thermal conductivity having a cylindrical, spherical, or other shape which serves as the calorimetric medium. Silver is the most suitable material because of its high thermal conductivity, but copper is most frequently used.

-6-

COMBINED CALORIMETERS

These are a combination of the liquid and aneroid types.

B1.27.4.3 SELECTION OF METHOD OF MEASUREMENT

In designing a calorimeter, consideration must be given to the combination of the principle of its operation with the type of design and this depends on the ultimate goal. Thus, isoperibole calorimeters may be a liquid, aneroid, or a combined calorimeter type with a static or rotating vessel (dynamic). For adiabatically-jacketed calorimeters, one normally uses a combined or aneroid design that enables the creation of a vacuum between the unit and the jacket, which is essential for proper thermal isolation of the unit. Isothermal calorimeters require special consideration depending on their application. To characterize a calorimeter design, it is necessary to specify its type by all the classifications, for example, adiabatic aneroid static or isoperibole liquid dynamic calorimeter.

The selection of the operating principle and the design of the calorimeter depends upon the nature of the process to be studied and on the experimental procedures required. However, the type of calorimeter necessary to study a particular process is not unique and can depend upon subjective factors such as technical restrictions, resources, traditions of the laboratory and the inclinations of the researcher.

B1.27.5 CALORIMETERS FOR SPECIFIC APPLICATIONS

Various books and chapters in books are devoted to calorimeter design and specific applications of calorimetry. For several decades the Commission on Thermodynamics of the International Union of Pure and

Applied Chemistry (IUPAC) has been responsible for a series of volumes on experimental thermodynamics and thermochemistry. *Experimental Thermochemistry*, volume I, published in 1956, edited by F D Rossini [2], dealt primarily with combustion calorimetry. Volume II published in 1962, edited by H A Skinner [3], primarily documented advances in combustion calorimetry since the first volume. In 1979 an update of much of the material covered in *Experimental Thermodynamics*, volumes I and II, was published under the title *Combustion Calorimetry* with editors S Sunner and M Månsson [4]. The first volume in the series *Experimental Thermodynamics, Calorimetry of Non-Reacting Systems*, edited by J P McCullough and D W Scott [5] was published in 1968. This volume covered the general principle of calorimeter design for non-reacting systems. It included a detailed discussion of adiabatic and drop calorimeters for the measurements of heat capacity, calorimeters for measurement of enthalpies of fusion and vaporisation, and calorimeters for the measurement of heat capacities of liquids and solutions close to room temperature. The second volume, *Experimental Thermodynamics of Non-Reacting Systems*, edited by B LeNeindre and B Vodar [6], published in 1975, was concerned with the measurement of a broader class of thermodynamic and transport properties over a wide range of temperature and pressure. A number of the techniques covered, such as density of a fluid as a function of temperature and pressure and speed of sound, allow the calculation of energy differences by non-calorimetric methods. Volume III, *Measurement of Transport Properties of Fluids*, edited by W A Wakeham, A Nagashima and J V Sengers [7], published in 1991, was concerned primarily with the measurement of the transport properties of fluids. Volume IV, *Solution Calorimetry*, edited by K N Marsh and P A G O'Hare [8] was published in 1994. This book covered calorimetric techniques for the measurement of enthalpies of reaction of organic substances, heat capacity and excess enthalpy of mixtures of organic compounds in both the liquid and gas phase, calorimetry of electrolyte solutions at high

-7-

temperature and pressure, microcalorimetric application in biological systems, titration calorimetry, and the calorimetric determination of pressure effects. *IUPAC Chemical Data Series No 32, Enthalpies of Vaporization of Organic Compounds* by V Majer and V Svoboda [9], contains a detailed review of calorimeters used to measure enthalpy of vaporization. Other monographs dealing extensively with calorimetric techniques have been published. These include *Specialist Periodical Reports, Chemical Thermodynamics*, volume 1 [10], which covered combustion and reaction calorimetry, heat capacity of organic compounds, vapour-flow calorimetry and calorimetric methods at high temperature. *Physical Methods of Chemistry*, Volume VI, *Determination of Thermodynamic Properties* [11, 12] contains a chapter on calorimetry and a chapter devoted to differential thermal methods including differential thermal calorimetry.

B1.27.5.1 MEASUREMENT OF HEAT CAPACITY

The most important thermodynamic property of a substance is the standard Gibbs energy of formation as a function of temperature as this information allows equilibrium constants for chemical reactions to be calculated. The standard Gibbs energy of formation $\Delta_f G^\circ$ at 298.15 K can be derived from the enthalpy of formation $\Delta_f H^\circ$ at 298.15 K and the standard entropy ΔS° at 298.15 K from

$$\Delta_f G^\circ = \Delta_f H^\circ - T \Delta S^\circ. \quad (\text{B1.27.17})$$

The enthalpy of formation is obtained from enthalpies of combustion, usually made at 298.15 K while the standard entropy at 298.15 K is derived by integration of the heat capacity as a function of temperature from $T = 0$ K to 298.15 K according to [equation \(B1.27.16\)](#). The Gibbs–Helmholtz relation gives the variation of the Gibbs energy with temperature

$$\left\{ \partial(G/T) / \partial T \right\}_p = -H/T^2. \quad (\text{B1.27.18})$$

Hence it is necessary to measure the heat capacity of a substance from near 0 K to the temperature required for equilibrium calculations to derive the enthalpy as a function of temperature according to [equation \(B1.27.15\)](#).

-8-

LOW TEMPERATURE HEAT CAPACITY

For solids and non-volatile liquids accurate heat capacity measurements are generally made in an adiabatic calorimeter. A typical low temperature aneroid-type adiabatic calorimeter used to make measurements between 4 K and about 300 K is shown in [figure B1.27.1](#). The primary function of the complex assembly is to maintain the calorimeter proper at any desired temperature between 4 K and 300 K. The only energy gain should be from the addition of electrical energy during a measurement. The upper part of the calorimeter contains vessels for holding liquid nitrogen and helium that provide low temperature heat sinks. Construction materials are generally those having high thermal conductivity (e.g. copper) plated with reflectant material (e.g. chromium) to reduce radiant energy transfer. The calorimeter proper and its surrounding adiabatic shield are suspended by silk lines and can be raised to bring them into good thermal contact with the lower tank, thereby cooling the calorimeter. When the calorimeter proper has reached its desired temperature, thermal contact is broken by lowering the calorimeter and the adiabatic shield. Adiabatic conditions are maintained by keeping the temperature of the adiabatic shield at the temperature of the calorimeter and heat conduction is minimized by maintaining a high vacuum (10^{-3} Pa) inside the cryostat. The temperature is normally measured with high precision using a calibrated platinum resistance thermometer. A major source of heat leak is through the electrical leads. This can be minimized by tempering the leads as they pass through the nitrogen and helium tanks and then bringing them to the calorimeter temperature with an electrical heater on the floating ring. In operation, a known amount of electrical energy is added through the heater and the temperature rise (usually of the order of 5 K) is measured to within 10^{-3} K. The temperature of the adiabatic shield is automatically controlled to follow the temperature of the calorimeter.

-9-

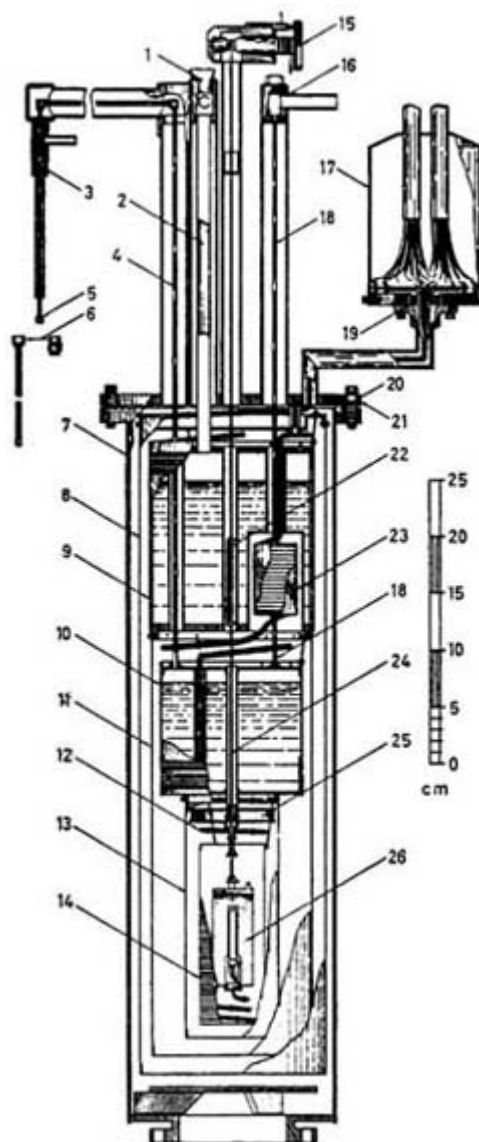


Figure B1.27.1. Aneroid-type cryostat for low-temperature adiabatic calorimeter: 1, 2, liquid nitrogen transfer; 3, 4, 5, 6, liquid helium transfer parts; 7, brass vacuum jacket; 8, outer floating radiation shield; 9, liquid nitrogen tank; 10, liquid helium tank; 11, nitrogen radiation shield; 12, lead wire; 13, helium radiation shield; 14, adiabatic shield; 15, windlass; 16, helium exit connector; 17, copper shield for terminal block; 18, helium exit tube; 19, vacuum seal; 20, O-ring gasket; 21, cover plate; 22, coil spring; 23, helium vapour exchanger; 24, supporting braided silk line; 25, floating ring; 26, calorimeter assembly. (Reprinted with permission from 1968 *Experimental Thermodynamics* vol I (Butterworth).)

HIGH TEMPERATURE HEAT CAPACITY

Adiabatic and drop calorimetry are the primary methods used to make measurements of heat capacity above room temperature. In drop calorimetry, a known mass of a sample at a known high temperature is dropped into a calorimeter vessel, usually close to room temperature and its temperature rise is measured. This method gives enthalpy differences, which are usually represented as a power series in the temperature. The equation can be differentiated to give the heat capacity. This method can be used to very high temperatures with moderate accuracy but it gives poor results when the sample undergoes phase transitions during the cooling process, since there may not be a complete transformation in the calorimeter. For systems with known phase

transitions adiabatic calorimetry is widely used. This technique is similar to that used in low temperature calorimetry except that no cooling is required. An example high temperature adiabatic calorimeter is shown in figure B1.27.2. An adiabatic shield that in the figure is the outer silver cup surrounds the calorimeter proper. Its temperature is controlled to be as close as possible to that of the calorimeter proper. The shield is surrounded by an inner and outer guard, which consist of multiple layers of thin aluminium. The inner guard is usually heated to a temperature close to the calorimeter temperature. Since the main heat loss mechanism at high temperature is radiation, the volume surrounding the calorimeter does not need to be evacuated.

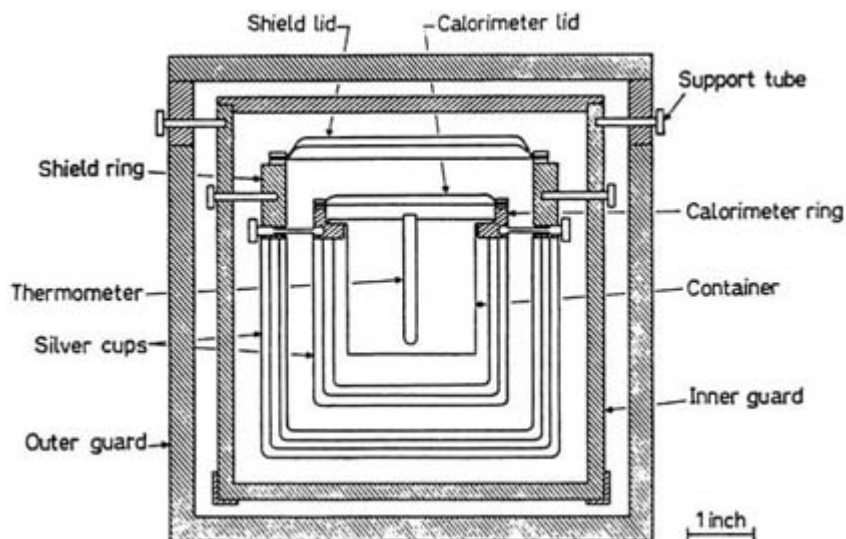


Figure B1.27.2. Schematic vertical section of a high-temperature adiabatic calorimeter and associated thermostat (Reprinted with permission from 1968 *Experimental Thermodynamics* vol I (Butterworth).)

Two methods are generally used when operating an adiabatic calorimeter. In the continuous-heating method, the calorimeter and shield are heated at a constant rate such that the temperature difference between the calorimeter and shield are minimal. At predetermined temperatures the power and time are recorded allowing the average heat capacity to be determined. This method allows for rapid measurement and the control of the shield temperature is less demanding. In the intermittent-heating method a known amount of power is added for a known time and the temperature change measured. The control of the adiabatic shield is more difficult because of the sudden changes in the rate of change of the temperature of the calorimeter at the beginning and end of the heating period. However, it is essential to use the intermittent method when a transition such as a solid–solid, liquid–solid, or an annealing process takes place.

HEAT CAPACITY OF GASES

The heat capacity of a gas at constant pressure C_p is normally determined in a flow calorimeter. The temperature rise is determined for a known power supplied to a gas flowing at a known rate. For gases at pressures greater than about 5 MPa Magee *et al* [13] have recently described a twin-bomb adiabatic calorimeter to measure C_p .

B1.27.5.2 COMBUSTION CALORIMETRY

Combustion or bomb calorimetry is used primary to derive enthalpy of formation values and measurements are usually made at 298.15 K. Bomb calorimeters can be subdivided into three types: (1) static, where the bomb or entire calorimeter (together with the bomb) remains motionless during the experiment; (2) rotating-

bomb calorimeters, where provision is made to rotate the bomb in the calorimetric media and (3) entirely rotating calorimeters, called dynamic. It is not necessary to use a rotating-bomb calorimeter for burning conventional organic compounds (containing only C, H, O and N). A stainless steel bomb without a corrosion-proof metal lining is suitable.

For burning organic substances containing heteroatoms of non-metals and metals, dynamic calorimeters of the combined or aneroid types are used. Liquid rotating-bomb calorimeters can also be used. For burning compounds containing halogens and sulphur, a bomb made of corrosion-resistant metal or lined with such a metal is generally used. The most resistant metal for the protection of the inner surface of a bomb used in combustion of chlorine-, sulphur- or bromine-containing organic compounds is tantalum, since it is very little affected by the products of combustion of these substances. Platinum can also be used as a protective layer, even though it is prone to react with the reaction products (e.g. $\text{Cl}_2 + \text{HCl} + \text{H}_2\text{O}$); the correction required to account for the enthalpy of such a reaction can be made by the analysis of the quantity of platinum dissolved. To study comparatively slow bomb processes, an adiabatically jacketed calorimeter designed as a combined or aneroid type or an isothermal calorimeter can be used only if the reaction can be conducted under static conditions.

STATIC BOMB CALORIMETER

An example of a static bomb calorimeter used to measure energies of combustion in oxygen is shown in [figure B1.27.3](#). The bomb is typically a heavy walled vessel capable of withstanding pressures of 20 MPa. A precisely known mass of the material to be burnt is held in a small platinum cup and oxygen is added to a pressure of about 3 MPa. The bomb is either immersed in a known mass of water or suspended in an evacuated vessel (aneroid type). The material is ignited by passing a current through a thin platinum wire stretched between the two metal posts, which causes an attached cotton or polythene fuse to burn. A small amount of water is added to the calorimeter to ensure that any solution that is formed is sufficiently dilute to allow the small corrections associated with the various solution processes to be

calculated. The amount of material in the vessel is chosen to give a temperature rise of from 1 K to 3 K, for a typical bomb immersed in about 3 kg H_2O . For an organic compound this corresponds to a mass of from 0.5 g to 1.6 g. A static bomb calorimeter is used for substances containing only carbon, hydrogen, oxygen and nitrogen giving only carbon dioxide, water and N_2 and possibly small amounts of HNO_2 and HNO_3 . The temperature rise is typically measured to between 10^{-4} K and 10^{-5} K and is measured with either a platinum resistance or a quartz thermometer. In order to relate the temperature rise to the energy of combustion, the calorimeter constant (the amount of energy required to increase the temperature of the calorimeter by 1 K) must be known. This can be obtained either by direct electrical calibration or by burning a certified reference material whose energy of combustion has been determined in specifically designed calorimeters. Direct electrical calibration is not simple as it involves the installation of a high power electrical heater within the bomb. Measurements on reference materials are usually made at a National Standards Laboratory. Reference materials suitable for the calibration of various calorimeters have been recommended by the International Union of Pure and Applied Chemistry (IUPAC) [1]. Benzoic acid is the most used reference material and its energy of combustion is known to about 1 part in 15,000. Involatile liquids or solids can usually be burnt directly. Volatile materials must be encapsulated in either plastic bags or glass ampoules. Combustion of these samples requires the addition of a known mass of an auxiliary material, which is usually an involatile oil whose energy of combustion is known.

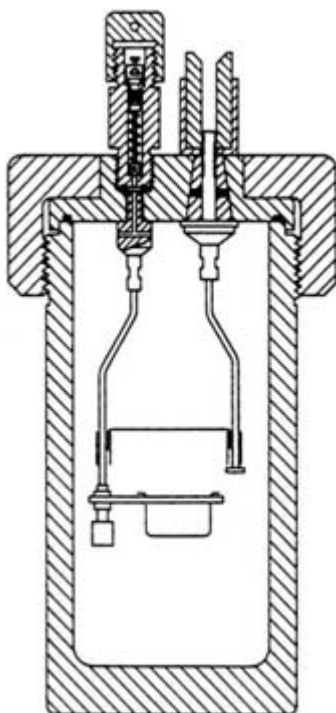


Figure B1.27.3. Typical static combustion bomb. (Reproduced with permission from A Gallencamp & Co. Ltd.)

ROTATING BOMB CALORIMETER

For substances containing elements additional to C, H, O and N a rotating bomb calorimeter is generally used. A typical rotating bomb calorimeter system is shown in [figure B1.27.4](#). With this calorimeter considerably more water is added to the combustion bomb and the continuous rotation of the bomb both about the cylindrical axis and end over

-13-

end ensures that the final solution is homogeneous and in equilibrium with the gaseous products. At the completion of the experiment this solution is withdrawn and analysed. Substances containing S, Si, P and the halogens can be studied in such calorimeters. To ensure that the products are in a known oxidation state it is generally necessary to add small quantities of reducing agents or other materials.

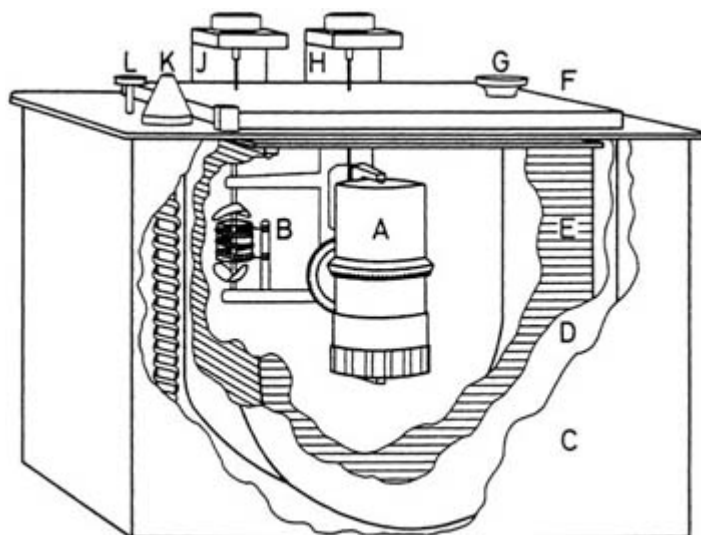


Figure B1.27.4. Rotating bomb isoperibole calorimeter. A, stainless steel bomb, platinum lined; B, heater; C, thermostat can; D, thermostat inner wall; E, thermostat water; G, sleeve for temperature sensor; H, motor for bomb rotation; J, motor for calorimeter stirrer; K, connection to cooling or heating unit for thermostat; L, circulation pump.

Precision combustion measurements are primarily made to determine enthalpies of formation. Since the combustion occurs at constant volume, the value determined is the energy change $\Delta_c U$. The enthalpy of combustion $\Delta_c H$ can be calculated from $\Delta_c U$, provided that the change in the pressure within the calorimeter is known. This change can be calculated from the change in the number of moles in the gas phase and assuming ideal gas behaviour. Enthalpies of formation of compounds that do not readily burn in oxygen can often be determined by combusting in fluorine and the enthalpy of formation of volatile substances can be determined using flame calorimetry. For compounds that only combust at an appreciable rate at high temperature, such as zirconium in chlorine, the technique of hot-zone calorimetry is used. In this method one heats the sample only very rapidly with a known amount of energy until it reaches a temperature where combustion will occur. Alternatively, a well characterized material such as benzoic acid can be used as an auxiliary material which, when it burns, raises the temperature sufficiently for the material to combust. These methods have been discussed in detail [2, 3 and 4].

B1.27.5.3 ENTHALPIES OF PHASE CHANGE

Accurate enthalpies of solid–solid transitions and solid–liquid transitions (fusion) are usually determined in an adiabatic heat capacity calorimeter. Measurements of lower precision can be made with a differential scanning calorimeter (see later). Enthalpies of vaporization are usually determined by the measurement of the amount of energy required to vaporize a known mass of sample. The various measurement methods have been critically reviewed by Majer and Svoboda [9]. The actual technique used depends on the vapour pressure of the material. Methods based on

-14-

vaporization into a vacuum are best suited for pressures from about 25 kPa down to 10^{-4} Pa. This method has been extensively developed by Sunner's group in Lund [14]. The most recent design allows measurement on samples down to 5 mg over a temperature range from 300 K to 423 K with an accuracy of about 1%. Methods based on vaporization into a steady stream of carrier gas, useful in the range 0.05 Pa to 25 kPa, have also been developed in Lund under Wadsö [15] and gas flow cells based on their designs are commercially available. The method is accurate to between 0.2 and 0.5%. Methods based on vaporization into a closed system, useful in the range 5 kPa to 3 MPa fall into two types; recycle and controlled withdrawal. Both types can give an accuracy approaching 0.1%. Methods based on recycle often contain a second calorimeter to determine the heat capacity of the flowing gas.

B1.27.5.4 SOLUTION CALORIMETRY

Solution calorimetry covers the measurement of the energy changes that occur when a compound or a mixture (solid, liquid or gas) is mixed, dissolved or adsorbed in a solvent or a solution. In addition it includes the measurement of the heat capacity of the resultant solution. Solution calorimeters are usually subdivided by the method in which the components are mixed, namely, batch, titration and flow.

BATCH CALORIMETERS

Batch calorimeters are instruments where there is no flow of matter in or out of the calorimeter during the time the energy change is being measured. Batch calorimeters differ in the way the reactants are mixed and in the method used to determine the enthalpy change. Enthalpy changes can be measured by the various methods

outlined above; isothermal, adiabatic, heat flow or isoperibole. It is necessary to have the reactants separated in the calorimeter. The most common method is to maintain one of the reactants in an ampoule that is broken to release its contents, which initiates the reaction. Initially, thin walled glass ampoules were used but these usually required the narrow neck to be flame-sealed after the contents were added. In recent years there have been significant improvements in ampoule design. An ampoule particularly suited for solids consists of a stainless steel cylinder with replaceable thin glass windows at each end. The cylinder can be taken apart with one half forming a cup into which the solid can be added and weighed. Wadsö and coworkers have developed a variety of ampoules that attach to the stirrer. The glass window is broken by depressing the stirrer so as to impinge against an ampoule-breaking pin. This technique ensures good mixing of the reactants. A typical solution calorimeter is shown in [figure B1.27.5](#)

Ampoules are satisfactory when the presence of a vapour space is not important. When volatile organic compounds are mixed in the presence of a vapour space there can be a considerable contribution to the measured heat effect from the vaporization. This results from the change in the vapour composition that occurs so as to maintain vapour-liquid equilibrium with the liquid mixture. For a small enthalpy of mixing, this correction can be greater than the enthalpy of mixing itself. A batch method suitable for the measurement of enthalpies of mixing in the absence of a vapour space is shown in [figure B1.27.6](#). Known masses of liquids A and B are separately confined over mercury and are mixed by rotation of the entire calorimeter. When liquids are mixed at constant pressure there is a volume change on mixing. The side arm C, which is partially filled with mercury, allows for the expansion or contraction of the mixture against the air space D. The calorimeter operates in the isoperibole mode.

-15-

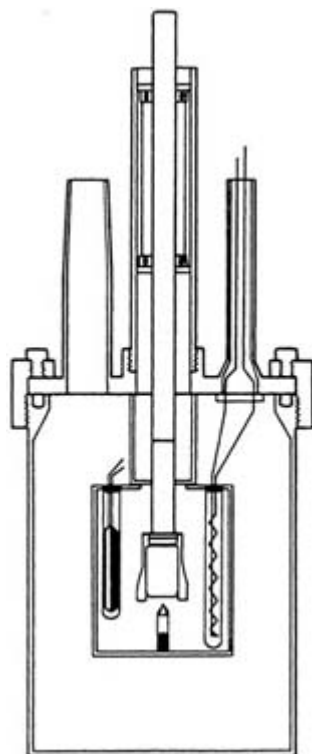


Figure B1.27.5. A typical solution calorimeter with thermometer, heater and an ampoule on the base of the stirrer which is broken by depressing it against the ampoule breaker. (Reproduced with permission from Sunner S and Wadsö I 1959 *Acta. Chem. Scand.* **13** 97.)

Reviews of batch calorimeters for a variety of applications are published in the volume on *Solution Calorimetry* [8]: cryogenic conditions by Zollweg [22], high temperature molten metals and alloys by Colinet and Pasturel [19], enthalpies of reaction of inorganic substances by Cordfunke and Ouweltjes [16], electrolyte

-16-

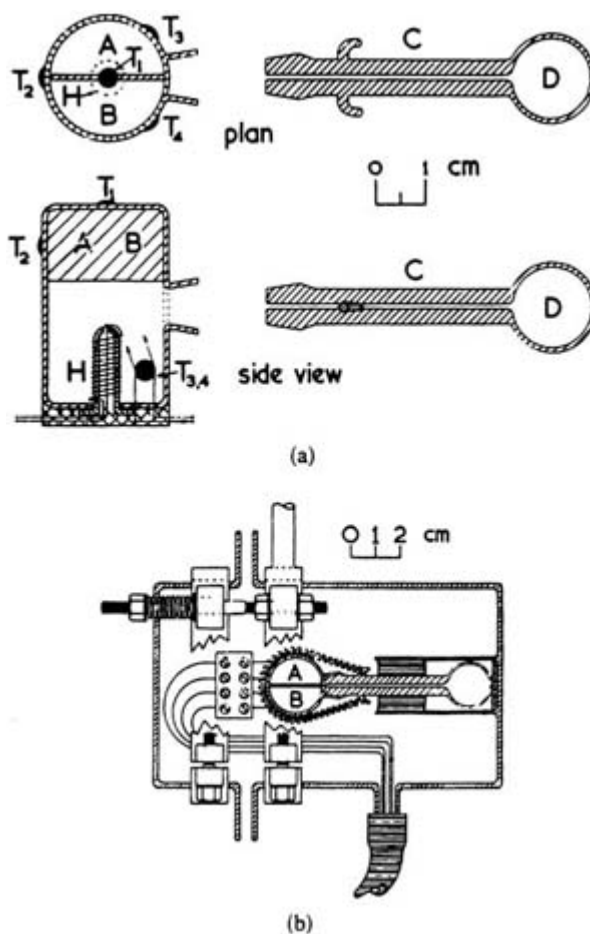


Figure B1.27.6. A calorimeter for enthalpies of mixing in the absence of a vapour space. (Reproduced with permission from Larkin J A and McGlashan M L 1961 *J. Chem. Soc.* 3245.)

-17-

TITRATION OR DILUTION CALORIMETERS

In a titration or dilution calorimetry one fluid is added from a burette, usually at a well-defined rate, into a calorimeter containing the second fluid. One titration is equivalent to many batch experiments. Calorimeters are typically operated in either the isoperibole or isothermal mode. The method has been extensively developed by Izatt, Christensen and co-workers at Brigham Young University for the measurement of formation constants and enthalpies of reaction for a variety of organic and inorganic compounds. This technique is described in detail by Oscarson *et al* [23]. A titration calorimeter typically has a vapour space, which for large enthalpies of reaction or solution in a solvent with a low vapour pressure does not give rise to significant errors. The vapour space can be eliminated by filling the cell completely and, as the titration proceeds, the excess liquid flows out of the calorimeter into a reservoir. The calculations of the enthalpy change for such a procedure is complex. To measure enthalpies of mixing of organic liquids, Stokes and Marsh [20] have used an alternative method that eliminates both the vapour space and the effect of volume changes on mixing. Their isothermal dilution calorimeter is shown in figure B1.27.7. The calorimeter proper, made from either stainless steel or glass, contains a stirrer, a sealed heater, a thermistor to measure the temperature, and a silver rod connected to a Peltier device. A known volume of mercury is added from a

pipette and one component is added to completely fill the calorimeter. The calorimeter is then brought to isothermal conditions within 10^{-3} K. For endothermic reactions the Peltier device removes energy at a rate sufficient to counterbalance the power introduced from the stirrer. The second component is then injected into the calorimeter from a burette, displacing the mercury. Electrical power is added to maintain the calorimeter approximately isothermal. Usually the rate of addition of the second component is adjusted so that the calorimeter remains isothermal to within 10^{-3} K during the addition. The injection is stopped at selected intervals and the calorimeter brought back to isothermal conditions by either the addition of additional electrical power or additional liquid from the burette. The volume of the mercury pipette is such that one run, which comprises about 20 individual measurement, covers over half the composition range. The components are then interchanged to determine the second half of the curve. The two runs should give results overlapping to within 1%. For exothermic systems the Peltier device is run at high power and the energy removed is counterbalanced by the addition of electrical energy to maintain the calorimeter isothermal. When the second component is added, the power is turned off for known periods of time. This calorimeter has been modified to measure enthalpies of solution of gases in liquids and Stokes [20] has described a version of this calorimeter that uses the overflow technique. Titration and dilution calorimeters have the disadvantage that they are difficult to operate at high pressures or at temperatures considerably removed from ambient. Flow calorimetry does not suffer these disadvantages.

-18-

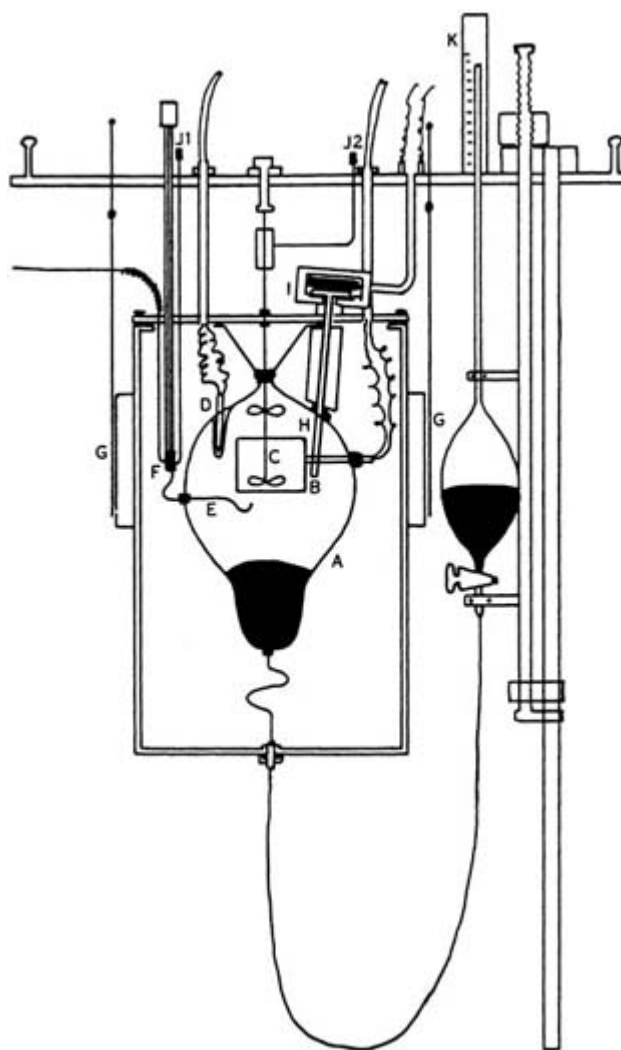


Figure B1.27.7. Schematic diagram of isothermal displacement calorimeter: A, glass calorimeter cell; B, sealed heater; C, stainless steel stirrer; D, thermistor; E, inlet tube; F, valve; G, window shutters; H, silver rod;

I, thermoelectric cooler; J, small ball valves; K, levelling device. (reproduced by permission from Costigan M J, Hodges L J, Marsh K N, Stokes R H and Tuxford C W 1980 *Aust. J. Chem.* **33** 2103.)

FLOW CALORIMETERS

In a flow calorimeter one or more streams flow in and out of the calorimeter. Flow calorimeters are suited for measurements over a wide range of temperatures and pressures on both liquids and gases. A wide pressure range is possible because measurement are usually made in a small bore tube, and a wide temperature range is feasible because the in-flowing material can be readily brought to the calorimeter temperature by heat exchange with the out-flowing fluid. In a flow experiment there is generally no vapour space and changes in volume on mixing are inconsequential. Flow methods are not suitable for measurement involving solids, and usually large volumes of materials are required.

-19-

Various flow calorimeters are available commercially. Flow calorimeters have been used to measure heat capacities, enthalpies of mixing of liquids, enthalpy of solution of gases in liquids and reaction enthalpies. Detailed descriptions of a variety of flow calorimeters are given in *Solution Calorimetry* by Grolier [17], by Albert and Archer [18], by Ott and Wormald [21], by Simonson and Mesmer [24] and by Wadsö [25].

A flow calorimeter developed by Picker suitable for the measurement of heat capacity of a liquid is shown in figure B1.27.8. The method measures the difference in heat capacity between the fluid under study and some reference fluid. The apparatus contains two thermistors T_1 and T_2 used to measure the temperature change that occurs when the flowing fluid is heated by two identical heaters Z_1 and Z_2 . The standard procedure is to flow the reference material from A through both cells. With the same fluid, same power and same flow rate the temperature change ΔT should be the same. The temperature difference observed on flowing the sample material from B through cell C_1 while the reference is still flowing through C_2 is a measure of the heat capacity difference between the two liquids. The flow method has been extensively developed for measurement on biological systems and on liquid mixtures at high temperatures and pressures. The apparatus constructed by Christensen and Izatt, shown in figure B1.27.9 can be used to measure positive and negative enthalpy changes at pressures up to 40.5 MPa and temperatures up to 673 K. Two high pressure pumps were used for the fluid flow. Mixing occurs in the top half of the isothermal cylinder where the fluid from the two pumps meet. A control heater encircles the cylinder, which is attached by three heat-leak rods to a base plate maintained 1 K below the cylinder temperature. A pulsed electrical current is passed through the control heater to maintain the temperature of the cylinder the same as that of the walls of the oven. During mixing the frequency of the pulses are either increased or decreased depending on the size of the enthalpy of mixing. Commercial calorimeters are available based on both the Picker and Christensen *et al* designs.

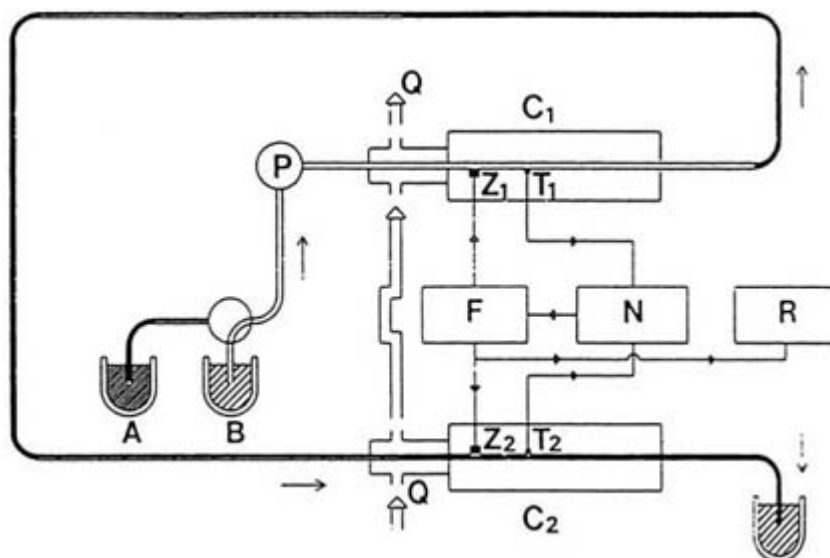


Figure B1.27.8. Schematic view of Picker's flow microcalorimeter. A, reference liquid; B, liquid under study; P, constant flow circulating pump; Z_1 and Z_2 , Zener diodes acting as heaters; T_1 and T_2 , thermistors acting as temperature sensing devices; F, feedback control; N, null detector; R, recorder; Q, thermostat. In the above A is the reference liquid and C_2 is the reference cell. When B circulates in cell C_1 this cell is the working cell. (Reproduced by permission from Picker P, Leduc P-A, Philip P R and Desnoyers J E 1971 *J. Chem. Thermo. B* 41.)

-20-

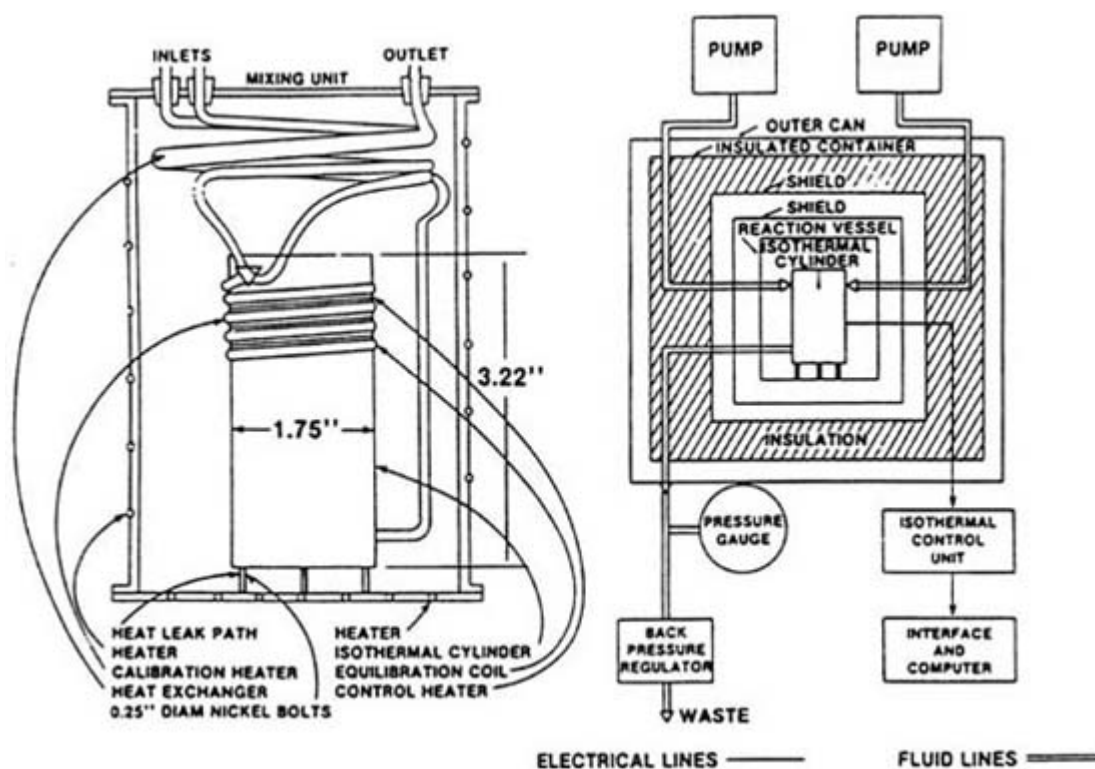


Figure B1.27.9. High-temperature heat-leak calorimeter. (Reproduced by permission from Christensen J J and Izatt R M 1984 An isothermal flow calorimeter designed for high-temperature, high-pressure operation *Thermochim. Acta* 73 117-29.)

B1.27.6 DIFFERENTIAL SCANNING CALORIMETRY

Boerio-Goats and Callanan [12] have recently reviewed different thermal methods, including differential scanning calorimetry. A differential scanning calorimeter (DSC) consists of two similar cells containing a sample and a reference material. In one type of DSC both cells are subjected to a controlled temperature change by applying power to separate heaters and the temperature difference between the sample and reference cells is observed. When an endothermic change occurs in the sample, for example, melting, the sample temperature lags behind that of the reference. In a power compensated DSC the power to that cell is increased to keep the heating rate of the sample and reference cells the same. A schematic diagram of a typical DSC is shown in [figure B1.27.10](#). Another type of calorimeter, developed initially by Tian and Calvet, is also considered a differential scanning calorimeter but is called a heat-flux or heat-conduction calorimeter. A schematic diagram of this type of calorimeter is shown in [figure B1.27.11](#). In this calorimeter two sets of thermopiles, consisting of multiple junction thermocouples, connect both cells to a large block enclosing the sample. The output of the two thermopiles, when connected in opposition gives a measure of the difference in energy flows between the sample and reference when both are heated at the same rate. Both types of DSC can be used to measure heat capacities, enthalpies of phase change, adsorption, dehydration, reaction and polymerization. The major advantage of DSC is the rapidity of the measurements, the small sample requirement, and

-21-

the ready availability of commercially available equipment capable of operating from liquid nitrogen temperatures to well above 1000 K by unskilled personnel. With the majority of DSC instruments it is possible to obtain heat capacities with an accuracy approaching 2–3%, provided one uses the optimum sample size and scan rate along with careful calibration of the temperature scale and the calorimetric response. Reference materials are used to calibrate a DSC and to check the correct operating conditions. Differential scanning calorimeters have been developed for specific applications. A very precise calorimeter, developed by Privilov and coworkers [17] and now available commercially, has been used to measure the heat capacity of very small amounts of biological materials in aqueous solutions.

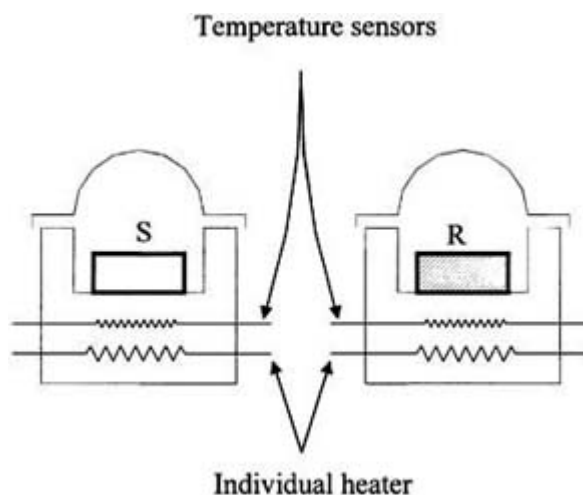


Figure B1.27.10. Schematic diagram of a power-compensated DSC.

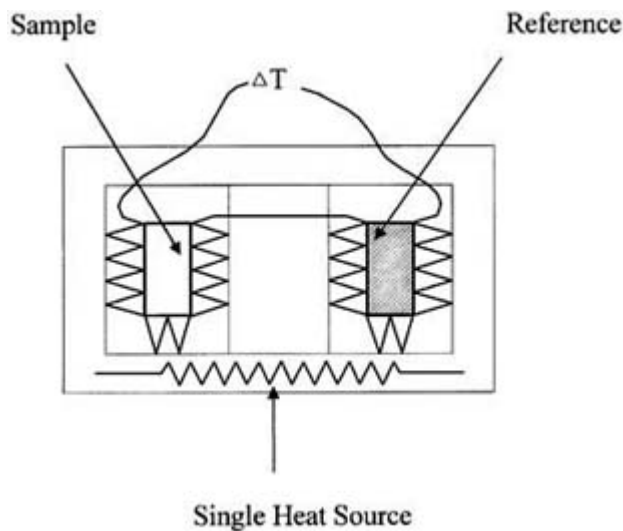


Figure B1.27.11. Schematic diagram of a Tian–Calvet heat-flux or heat-conduction calorimeter.

B1.27.7 ACCELERATING RATE CALORIMETRY

Special calorimeters have been developed to make thermal hazard evaluations. In an exothermic chemical reaction there is the possibility of a runaway reaction occurring where the energy released from the reaction increases the temperature with a consequent increase in the reaction rate, thus increasing the release of energy. If there are insufficient resources to remove the generated energy, hazardous temperature and pressure regimes can be encountered. An accelerating rate calorimeter or ARC, initially developed at Dow Chemical, is available commercially to study such thermal hazards. A schematic diagram is shown in figure B1.27.12. The reaction vessel consists of a spherical bomb that can withstand pressures greater than 20 MPa and temperatures to 770 K. The calorimeter operates in an adiabatic mode under computer control in a heat-wait-see mode. After the reaction comes to thermal equilibrium the rate of temperature rise due to the reaction is determined. If this is less than a preset value the calorimeter temperature is increased in steps and the process repeated until the reaction rate is sufficient to give the preset temperature rise. The chemical reaction then proceeds at its own rate and the temperature and pressure recorded. From these measurements the kinetic parameters are determined and used to establish the conditions that could lead to a runaway reaction. A problem with this calorimeter is that the massive vessel required to withstand the pressure has a heat capacity well in excess of the heat capacity of the reactants. This problem can be overcome by having a thin-walled vessel within the bomb and the pressure in the space between the reaction vessel and the bomb is automatically controlled to the pressure in the calorimeter.

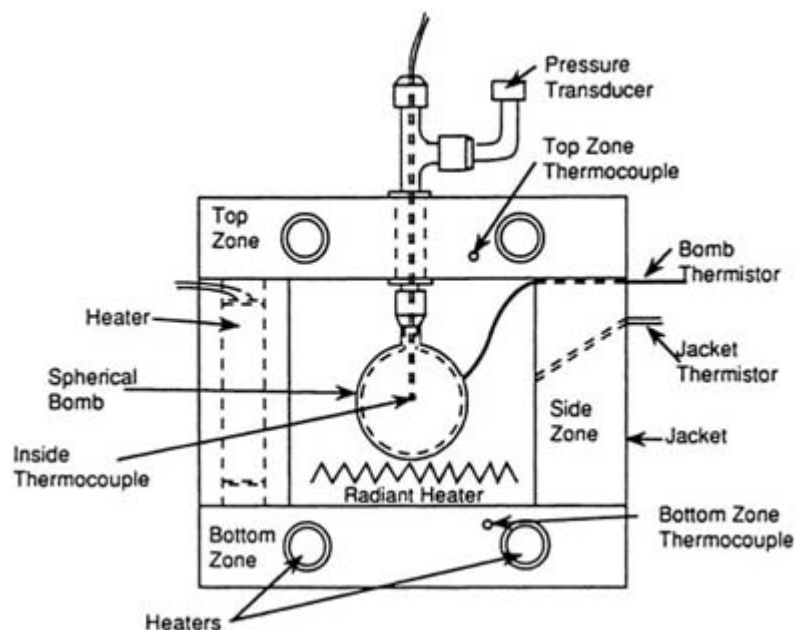


Figure B1.27.12. Schematic diagram of an accelerating rate calorimeter (ARC).

B1.27.8 SPECIALIZED CALORIMETERS

Ultra sensitive Calvet type microcalorimeters are available commercially to measure the deterioration of materials over relatively long periods. For example, the lifetime of a battery can be estimated from the energy released when it is placed in such a calorimeter on open circuit [27]. Similarly the shelf life of drugs and other digestible products can be evaluated from the very small calorimetric response that results from decomposition reactions [28]. Calorimeters have also been developed to measure the heat effects associated with the uptake of oxygen and carbon dioxide in living plants. Such measurements have been used to identify species that exhibit high rates of metabolism [29]. Calorimeters are also used to measure the rate of enzyme reaction and as a clinical tool to identify micro-organisms and test the effect of drugs in inhibiting the growth of such micro-organisms.

B1.27.9 RECENT DEVELOPMENTS

Recent developments in calorimetry have focused primarily on the calorimetry of biochemical systems, with the study of complex systems such as micelles, proteins and lipids using microcalorimeters. Over the last 20 years microcalorimeters of various types including flow, titration, dilution, perfusion calorimeters and calorimeters used for the study of the dissolution of gases, liquids and solids have been developed. A more recent development is pressure-controlled scanning calorimetry [26] where the thermal effects resulting from varying the pressure on a system either step-wise or continuously is studied.

REFERENCES

- [1] Head A J and Sabbah R 1987 *Enthalpy Recommended Reference Materials for the Realization of Physicochemical Properties* ed K N Marsh (Oxford: Blackwell)
- [2] Rossini F D (ed) 1956 *Experimental Thermochemistry* vol I (New York: Interscience)
- [3] Skinner H A (ed) 1962 *Experimental Thermochemistry* vol II (New York: Interscience)
- [4] Sunner S and Månsson M (eds) 1979 *Combustion Calorimetry* (Oxford: Pergamon)
- [5] McCullough J P and Scott D W (eds) 1968 *Experimental Thermodynamics Calorimetry of Non-Reacting Systems* vol I (London: Butterworths)
- [6] LeNeindre B and Vodar B (eds) 1975 *Experimental Thermodynamics Experimental Thermodynamics of Non-Reacting Systems* vol II (London: Butterworths)
- [7] Wakeham W A, Nagashima A and Sengers J V (eds) 1991 *Experimental Thermodynamics Measurement of Transport Properties of Fluids* vol III (Oxford: Blackwell)
- [8] Marsh K N and O'Hare P A G (eds) 1994 *Solution Calorimetry, Experimental Thermodynamics* vol IV (Oxford: Blackwell)
-

-24-

- [9] Majer V and Svoboda V 1985 (IUPAC Chemical Data Series No 32) *Enthalpies of Vaporization of Organic Compounds* (Oxford: Blackwell)
- [10] McGlashan M L (ed) 1973 *Specialist Periodical Reports, Chemical Thermodynamics* vol 1 (London: The Chemical Society)
- [11] Oscarson J L and Izatt R M 1992 *Calorimetry Physical Methods of Chemistry Determination of Thermodynamic Properties* 2nd edn, vol VI, ed B W Rossiter and R C Baetzold (New York: Wiley)
- [12] Boerio-Goates J and Callanan J E 1992 *Differential thermal methods Physical Methods of Chemistry Determination of Thermodynamic Properties* 2nd edn, vol VI, ed B W Rossiter and R C Baetzold (New York: Wiley)
- [13] Magee J W, Blanco J C and Deal R J 1998 High-temperature adiabatic calorimeter for constant-volume heat capacity of compressed gases and liquids *J. Res. Natl Inst. Stand. Technol.* **103** 63
- [14] Morawetz E 1972 Enthalpies of vaporization of *n*-alkanes from C₁₂ to C₂₀ *J. Chem. Thermodyn.* **4** 139
- [15] Wadsö I 1968 Heats of vaporization of organic compounds: II. Chlorides, bromides and iodides *Acta Chem. Scand.* **22** 2438
- [16] Cordfunke E H P and Ouweltjes W 1994 Solution calorimetry for the determination of enthalpies of reaction of inorganic substances at 298.15 K *Solution Calorimetry, Experimental Thermodynamics* vol IV, ed K N Marsh and P A G O'Hare (Oxford: Blackwell)
- [17] Grolier J-P E 1994 Heat capacity of organic liquids *Solution Calorimetry, Experimental Thermodynamics* vol IV, ed K N Marsh and P A G O'Hare (Oxford: Blackwell)
- [18] Albert H J and Archer D G 1994 Mass-flow isoperibole calorimeters *Solution Calorimetry, Experimental Thermodynamics* vol IV, ed K N Marsh and P A G O'Hare (Oxford: Blackwell)
- [19] Colinet C and Pasturel A 1994 High temperature solution calorimetry *Solution Calorimetry, Experimental Thermodynamics* vol IV, ed K N Marsh and P A G O'Hare (Oxford: Blackwell)
- [20] Stokes R H 1994 Isothermal displacement calorimeters *Solution Calorimetry, Experimental Thermodynamics* vol IV, ed K N Marsh and P A G O'Hare (Oxford: Blackwell)
- [21] Ott J B and Wormald C J 1994 Excess enthalpy by flow calorimetry *Solution Calorimetry, Experimental Thermodynamics* vol IV, ed K N Marsh and P A G O'Hare (Oxford: Blackwell)
- [22] Zollweg J A 1994 Mixing calorimetry at cryogenic conditions *Solution Calorimetry, Experimental Thermodynamics*

vol IV, ed K N Marsh and P A G O'Hare (Oxford: Blackwell)

- [23] Oscarson J L, Izatt R M, Hill J O and Brown P R 1994 Continuous titration calorimetry *Solution Calorimetry, Experimental Thermodynamics* vol IV, ed K N Marsh and P A G O'Hare (Oxford: Blackwell)
- [24] Simonson J M and Mesmer R E 1994 Electrolyte solutions at high temperatures and pressures *Solution Calorimetry, Experimental Thermodynamics* vol IV, ed K N Marsh and P A G O'Hare (Oxford: Blackwell)
- [25] Wadsö I 1994 Microcalorimetry of aqueous and biological systems *Solution Calorimetry, Experimental Thermodynamics* vol IV, ed K N Marsh and P A G O'Hare (Oxford: Blackwell)
- [26] Randzio S L 1994 Calorimetric determination of pressure effects *Solution Calorimetry, Experimental Thermodynamics* vol IV, ed K N Marsh and P A G O'Hare (Oxford: Blackwell)
- [27] Hansen L D and Hart R M 1978 Shelf-life prediction from induction period *J. Electrochem. Soc.* **125** 842

-25-

- [28] Hansen L D, Eatough D J, Lewis E A and Bergstrom R G 1990 Calorimetric measurements on materials undergoing autocatalytic decomposition *Can. J. Chem.* **68** 2111–14
- [29] Hansen L D, Hopkins M S and Criddle R S 1997 Plant calorimetry: a window to plant physiology and ecology *Thermochim. Acta* **300** 183–97

FURTHER READING

Oscarson J L and Izatt R M 1992 Calorimetry *Determination of Thermodynamic Properties, Physical Methods of Chemistry* 2nd edn vol VI ed B W Rossiter and R C Baetzold (New York: Wiley)

Boerio-Goates J and Callanan J E 1992 Differential thermal methods *Determination of Thermodynamic Properties, Physical Methods of Chemistry* 2nd edn vol VI ed B W Rossiter and R C Baetzold (New York: Wiley)

Rossini F D (ed) 1956 *Experimental Thermochemistry* vol I (New York: Interscience)

Skinner H A (ed) 1962 *Experimental Thermochemistry* vol II (New York: Interscience)

Sunner S and Månsson M (eds) 1979 *Combustion Calorimetry* (Oxford: Pergamon)

McCullough J P and Scott D W (eds) 1968 *Calorimetry of Non-Reacting Systems, Experimental Thermodynamics* vol I (London: Butterworths)

Marsh K N and O'Hare P A G (eds) 1994 *Solution Calorimetry, Experimental Thermodynamics* vol IV (Oxford: Blackwell)

Head A J and Sabbah R 1987 Enthalpy *Recommended Reference Materials for the Realization of Physicochemical Properties* ed K N Marsh (Oxford: Blackwell)

Majer V and Svoboda V 1985 *Enthalpies of Vaporization of Organic Compounds* (Oxford: Blackwell)

-26-

- [27] Hansen L D and Hart R M 1978 Shelf-life prediction from induction period *J. Electrochem. Soc.* **125** 842
- [28] Hansen L D, Eatough D J, Lewis E A and Bergstrom R G 1990 Calorimetric measurements on materials undergoing autocatalytic decomposition *Can. J. Chem.* **68** 2111–14
- [29] Hansen L D, Hopkins M S and Criddle R S 1997 Plant calorimetry: a window to plant physiology and ecology *Thermochim. Acta* **300** 183–97
-

FURTHER READING

Oscarson J L and Izatt R M 1992 Calorimetry *Determination of Thermodynamic Properties, Physical Methods of Chemistry* 2nd edn vol VI ed B W Rossiter and R C Baetzold (New York: Wiley)

Boerio-Goates J and Callanan J E 1992 Differential thermal methods *Determination of Thermodynamic Properties, Physical Methods of Chemistry* 2nd edn vol VI ed B W Rossiter and R C Baetzold (New York: Wiley)

Rossini F D (ed) 1956 *Experimental Thermochemistry* vol I (New York: Interscience)

Skinner H A (ed) 1962 *Experimental Thermochemistry* vol II (New York: Interscience)

Sunner S and Månsson M (eds) 1979 *Combustion Calorimetry* (Oxford: Pergamon)

McCullough J P and Scott D W (eds) 1968 *Calorimetry of Non-Reacting Systems, Experimental Thermodynamics* vol I (London: Butterworths)

Marsh K N and O'Hare P A G (eds) 1994 *Solution Calorimetry, Experimental Thermodynamics* vol IV (Oxford: Blackwell)

Head A J and Sabbah R 1987 *Enthalpy Recommended Reference Materials for the Realization of Physicochemical Properties* ed K N Marsh (Oxford: Blackwell)

Majer V and Svoboda V 1985 *Enthalpies of Vaporization of Organic Compounds* (Oxford: Blackwell)

-1-

B1.28 Electrochemical methods

Alexia W E Hodgson

B1.28.1 INTRODUCTION

Electrochemical methods may be classified into two broad classes, namely potentiometric methods and voltammetric methods. The former involves the measurement of the potential of a working electrode immersed in a solution containing a redox species of interest with respect to a reference electrode. These are equilibrium experiments involving no current flow and provide thermodynamic information only. The potential of the working electrode responds in a Nernstian manner to the activity of the redox species, whilst that of the reference electrode remains constant. In contrast, in voltammetric methods the system is perturbed

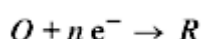
and involves the control of the electrode potential or the current as the independent variable, and measurement of the resulting current or potential.

The latter may be further subdivided into *transient experiments*, in which the current and potential vary with time in a non-repetitive fashion; *steady-state experiments*, in which a unique interrelation between current and potential is generated, a relation that does not involve time or frequency and in which the steady-state current achieved is independent of the method adopted and *periodic experiments*, in which current and potential vary periodically with time at some imposed frequency.

In this chapter, transient techniques, steady-state techniques, electrochemical impedance, photoelectrochemistry and spectroelectrochemistry are discussed.

B1.28.2 INTRODUCTION TO ELECTRODE REACTIONS

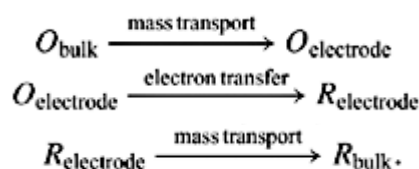
Electrode processes are a class of heterogeneous chemical reaction that involves the transfer of charge across the interface between a solid and an adjacent solution phase, either in equilibrium or under partial or total kinetic control. A simple type of electrode reaction involves electron transfer between an inert metal electrode and an ion or molecule in solution. Oxidation of an electroactive species corresponds to the transfer of electrons from the solution phase to the electrode (anodic), whereas electron transfer in the opposite direction results in the reduction of the species (cathodic). Electron transfer is only possible when the electroactive material is within molecular distances of the electrode surface; thus for a simple electrode reaction involving solution species of the form



in which species O is reduced at the electrode surface to species R by the transfer of n electrons, the overall conversion

-2-

may be divided into three steps [1, 3]:



The scheme involves the transport of the electroactive species from the bulk solution to the electrode surface, where it can undergo electron transfer, thus forming the reduced species R at the electrode surface. Finally, the reduced species is transported from the electrode surface back to the bulk solution. The overall reaction rate will be limited by the slowest step, therefore a particular reaction might be controlled by either the kinetics of electron transfer or by the rate at which material is brought to or from the electrode surface. The rate of electron transfer can be experimentally controlled through the electrode potential imposed and can vary by several orders of magnitude in a small potential interval. For the steps involving the transport of species to and from the electrode surface there are three distinct modes of mass transport regime which can occur: diffusion, migration and convection.

The nature of electrode processes can, of course, be more complex and also involve phase formation, homogeneous chemical reactions, adsorption or multiple electron transfer [1, 2, 3 and 4].

B1.28.2.1 ELECTRON TRANSFER

For a simple electron transfer reaction containing low concentrations of a redox couple in an excess of electrolyte, the potential established at an inert electrode under equilibrium conditions will be governed by the Nernst equation and the electrode will take up the equilibrium potential E_e for the couple O/R . In terms of current density, the dynamic situation at the electrode surface is expressed by $j = \bar{j} + \bar{j} = 0$, the sum of the partial cathodic and partial anodic current densities, which have opposite signs and the magnitude of which at equilibrium potential is defined as $j_0 = -\bar{j} = \bar{j}$. The exchange current density, j_0 , is a measure of the amount of electron transfer activity at the equilibrium potential. On applying a potential to the electrode, the system will seek to move towards a new equilibrium where the concentrations of the electroactive species are those demanded by the Nernst equation for the applied potential, and an associated current of reduction or oxidation will flow. The rate of electron transfer can be described by classical kinetics, and hence expressed by the product of a rate constant with the concentration of the reactant at the electrode surface. The rate of the heterogeneous electron transfer will depend on the potential gradient at the interface driving the transfer of electrons between the electrode and the solution phases and in general will take the form of $\bar{k} = k_0 \exp(-\alpha_c n F / RT) E$ and $\bar{k} = k_0 \exp(-\alpha_a n F / RT) E$ for a reduction and oxidation process respectively. α_c and α_a are the cathodic and anodic transfer coefficients, F is the Faraday constant and k_0 the standard rate constant [1, 2, 3 and 4]. By substituting and defining the overpotential, $\eta = E - E_e$, as the deviation of the potential from the equilibrium

value, the Butler–Volmer equation for current density may be derived:

$$j = j_0 \left\{ \exp\left(\frac{\alpha_a n F}{RT} \eta\right) - \exp\left(\frac{-\alpha_c n F}{RT} \eta\right) \right\}$$

which represents the fundamental equation of electrode kinetics. The equation may be simplified for the limiting cases in which very high positive or very high negative overpotentials are applied, leading to the Tafel equations, which provide a simple method for determining exchange current density and transfer coefficient (figure B1.28.1). For very low overpotentials the equation simplifies to $j = j_0 (nF/RT)\eta$, indicating that very close to the equilibrium potential, the current density varies linearly with overpotential.

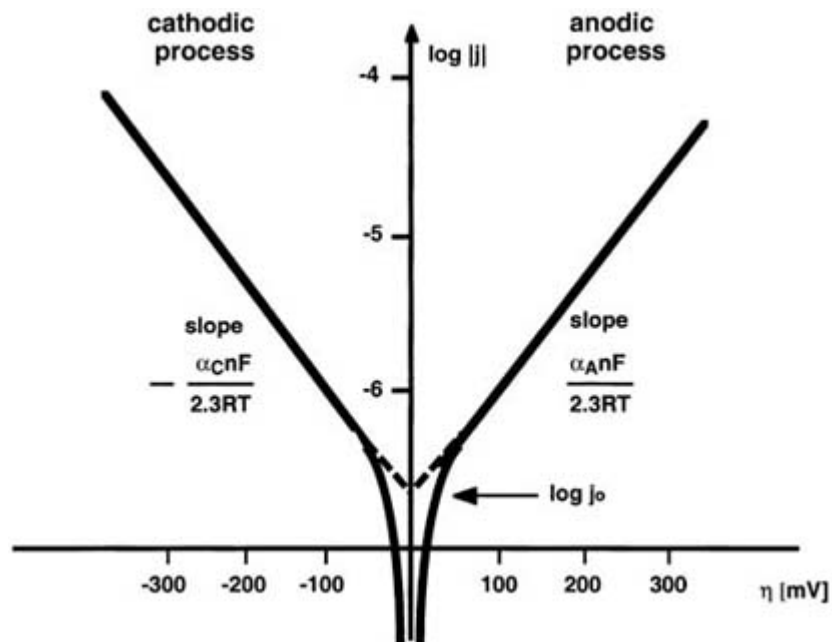


Figure B1.28.1. Schematic Tafel plot for the experimental determination of j_0 and α .

B1.28.2.2 MASS TRANSPORT

Diffusion, convection and migration are the forms of mass transport that contribute to the essential supply and removal of material to and from the electrode surface [1, 2, 3 and 4].

Diffusion may be defined as the movement of a species due to a concentration gradient, which seeks to maximize entropy by overcoming inhomogeneities within a system. The rate of diffusion of a species, the flux, at a given point in solution is dependent upon the concentration gradient at that particular point and was first described by Fick in 1855, who considered the simple case of linear diffusion to a planar surface:

-4-

$$\text{Flux} = -D \frac{dc_i(x)}{dx}$$

where dc_i/dx is the concentration gradient and D is the diffusion coefficient (figure B1.28.2). The flux of species to and from the electrode surface must also be accompanied by the conversion of reactant to product and by the flux of electrons. The flux of material crossing the electrode boundary can therefore be converted to current density by equating the two fluxes:

$$\frac{j}{nF} = -D \frac{dc_i(x)}{dx}$$

The second of Fick's laws expresses the change in concentration of a species at a point as a function of time due to diffusion (figure B1.28.2). Hence, the one-dimensional variation in concentration of material within a volume element bounded by two planes x and $x + dx$ during a time interval dt is expressed by $\partial c_i(x,t)/\partial t = D (\partial^2 c_i(x,t)/\partial x^2)$. Fick's second law of diffusion enables predictions of concentration changes of electroactive material close to the electrode surface and solutions, with initial and boundary conditions appropriate to a particular experiment, provide the basis of the theory of instrumental methods such as, for example, potential-step and cyclic voltammetry.

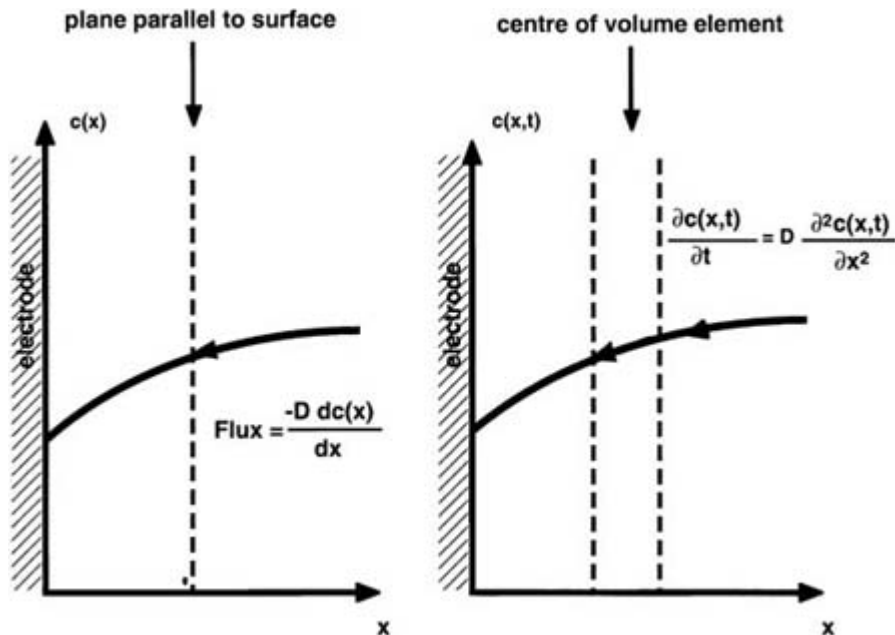


Figure B1.28.2. Fick's laws of diffusion. (a) Fick's first law, (b) Fick's second law.

Convection is the movement of a species due to external mechanical forces. This can be of two types: natural convection, which arises from thermal gradients or density differences within the solution, and forced convection, which can take the form of gas bubbling, pumping or stirring. The former is undesirable and can occur in any solution

at normal sized electrodes on the time scale of ten or more seconds. In contrast, the latter has the function of overcoming contributions from natural convection and of increasing the rate of mass transport and hence facilitates the study of the kinetics of electrode reactions. Forced convection usually possesses well defined hydrodynamic behaviour, thus enabling the quantitative description of the flow in the solution and the prediction of the pattern of mass transport to the electrode.

Migration is the movement of ions due to a potential gradient. In an electrochemical cell the external electric field at the electrode/solution interface due to the drop in electrical potential between the two phases exerts an electrostatic force on the charged species present in the interfacial region, thus inducing movement of ions to or from the electrode. The magnitude is proportional to the concentration of the ion, the electric field and the ionic mobility.

Most electrochemical experiments are designed so that one of the mass transport regimes dominates over the others, thus simplifying the theoretical treatment, and allowing experimental responses to be compared with theoretical predictions. Normally, specific conditions are selected where the mass-transport regime results only from diffusion or convection. Such regimes allow mass transport to be described by a set of mathematical equations, which have analytical solutions. A common experimental practice to render the migration of reactants and products negligible is to add an excess of inert supporting electrolyte, thus ensuring that any migration is dominated by the ions of the electrolyte. Electro-neutrality is also thus maintained, ensuring that electric fields do not build up in the solution. Furthermore, the addition of a high concentration of electrolyte increases the solution conductivity, compresses the double-layer region to dimensions of 10–20 Å, and ensures a constant ionic strength during the electrochemical experiment. As a consequence, the activities of the electroactive species and thus the applied potentials, as predicted by the Nernst equation and by the rate of electron transfer, remain constant throughout the experiment.

B1.28.3 TRANSIENT TECHNIQUES

Voltammetry relies on the registering of current–potential profiles, whether by controlling the potential of the working electrode and recording the resulting current or by measuring the potential response as a function of an applied current. The electrochemical cell, as well as a conducting medium, must also contain at least one other electrode. In a two-electrode configuration, the second electrode is a reference electrode that serves both as a standard against which the working potential is measured and as the necessary current-carrying electrode where the rate of charge transfer must be equal and opposite to that of the working electrode. Commonly, these two functions are separated in a three-electrode configuration, in which a secondary or counter electrode is employed as the current-carrying electrode and a separate reference electrode reports the potential of the working electrode. This prevents any undesirable polarization of the reference electrode, since only small currents flow in the reference electrode loop. Placement of the reference electrode close to the working electrode enables the exclusion of the majority of the solution IR_u drop, which is often achieved by the use of a Luggin capillary [1, 2].

The measurement of the current for a redox process as a function of an applied potential yields a voltammogram characteristic of the analyte of interest. The particular features, such as peak potentials, half-wave potentials, relative peak/wave height of a voltammogram give qualitative information about the analyte electrochemistry within the sample being studied, whilst quantitative data can also be determined. There is a wealth of voltammetric techniques, which are linked to the form of potential program and mode of current measurement adopted. Potential-step and potential-sweep

-6-

techniques are carried out under conditions where diffusion is the only mode of mass transport and the experiment is designed such that diffusion may be described by linear diffusion to a plane electrode and changes in concentration occur perpendicular to the surface [1, 2, 3, 4 and 5].

B1.28.3.1 LINEAR-SWEEP AND CYCLIC VOLTAMMETRY

Linear-sweep and cyclic voltammetry were first reported in 1938 and described theoretically in 1948 by Randles and Sevcik [1, 2, 3, 4, 5 and 6]. The techniques consist of scanning the potential between two chosen limits at a known sweep rate, v , and measuring the current response arising from any electron transfer process. In linear-sweep voltammetry the scan terminates at the chosen end potential, E_f , whereas in cyclic voltammetry, the potential is reversed back at E_f toward the starting potential E_i , or another chosen potential limit. The potential limits define the electrode reactions that take place so that the potential scan is normally chosen to start at a potential value where no electrode reaction occurs and swept towards positive or negative potentials to investigate oxidation or reduction processes, respectively. The current–potential curves for a simple reversible electrode reaction are characterized by unsymmetrical peaks with the current density increasing as the sweep rate is raised. On the forward sweep, the current begins to rise as the potential reaches the vicinity of the reversible formal potential E^0 , then passes through a maximum before decreasing again as the potential is sufficiently driven to produce a diffusion-limited current. The surface concentration of an electroactive species, R , decreases as the potential is made more positive and the rate of oxidation increases, until it becomes effectively zero, at which point the reaction is diffusion controlled. In terms of concentration profiles, the flux of species to the electrode surface increases with potential (hence time) and continues to increase until the surface concentration reaches zero, at which point the flux to the surface starts to decrease, since the surface concentration remains at zero, yielding the peak-shaped response (figure B1.28.3). On the reverse sweep, anodic current continues to flow until the potential is still sufficiently negative to cause the oxidation of R . When the potential, however, reaches the vicinity of E^0 , the oxidized species produced can diffuse back to the electrode surface to be reduced, the current becomes cathodic and a similar peak-shaped

response is obtained as the reaction becomes diffusion controlled.

The scan rate, $\nu = |dE/dt|$, plays a very important role in sweep voltammetry as it defines the time scale of the experiment and is typically in the range 5 mV s^{-1} to 100 V s^{-1} for normal macroelectrodes, although sweep rates of 10^6 V s^{-1} are possible with microelectrodes (see later). The short time scales in which the experiments are carried out are the cause for the prevalence of non-steady-state diffusion and the peak-shaped response. When the scan rate is slow enough to maintain steady-state diffusion, the concentration profiles with time are linear within the Nernst diffusion layer which is fixed by natural convection, and the current–potential response reaches a plateau steady-state current. On reducing the time scale, the diffusion layer cannot relax to its equilibrium state, the diffusion layer is thinner and hence the currents in the non-steady-state will be higher.

-7-

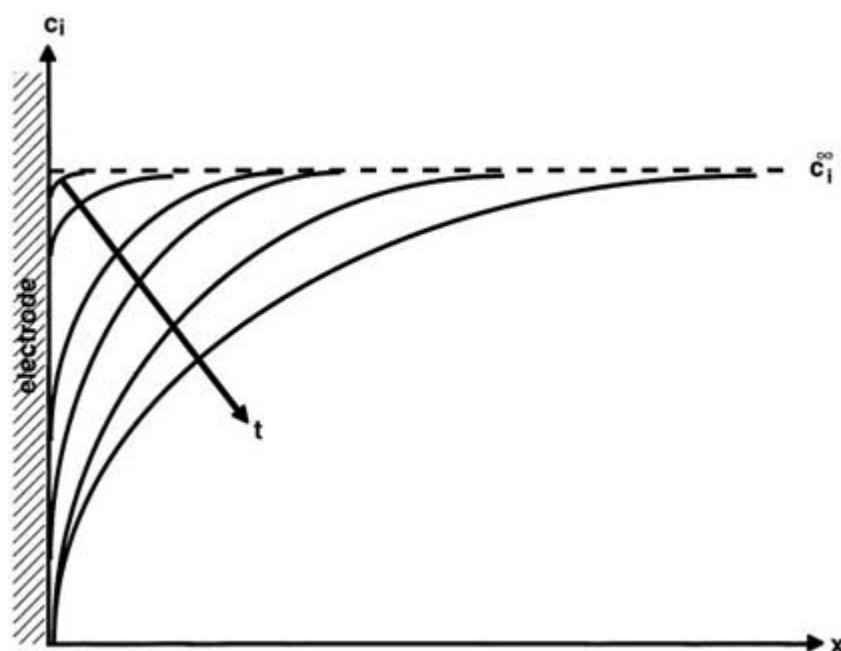


Figure B1.28.3. Concentration profiles of an electroactive species with distance from the electrode surface during a linear sweep voltammogram.

Cyclic voltammetry provides a simple method for investigating the reversibility of an electrode reaction ([table B1.28.1](#)). The reversibility of a reaction closely depends upon the rate of electron transfer being sufficiently high to maintain the surface concentrations close to those demanded by the electrode potential through the Nernst equation. Therefore, when the scan rate is increased, a reversible reaction may be transformed to an irreversible one if the rate of electron transfer is slow. For a reversible reaction at a planar electrode, the peak current density, j_p , is given by

$$j_p = 2.69 \times 10^5 n^{3/2} D^{1/2} c_i^\infty \nu^{1/2}$$

where n is the number of electrons, D is the diffusion coefficient, c_i^∞ the concentration of the electroactive species in the bulk and ν the sweep rate. Of particular importance is the proportionality of the peak current to the square root of the scan rate. In addition, for a reversible couple, the cathodic and anodic peak potentials are separated by $59/n$ mV, the reversible half-wave potential is situated midway between the peaks, the peak potential is independent of scan rate and the peak current ratio equals 1 ([figure B1.28.4](#)). As the response becomes less reversible, the separation between the peaks increases, as an overpotential is necessary to drive

reduction and oxidation reactions, and the shape of the peaks will become more drawn out. Beyond the peak, however, the electrode reaction remains diffusion controlled. For totally irreversible systems the reverse peak disappears completely and the peak current density is expressed by

$$j_p = 2.99 \times 10^5 n(\alpha n_\alpha)^{1/2} c_i^\infty D^{1/2} \nu^{1/2}$$

-8-

where α is the transfer coefficient and n_α the number of electrons transferred up to and including the rate-determining step. The majority of redox couples fall between the two extremes and exhibit quasi-reversible behaviour. When investigating an electrode reaction for reversibility it is essential to obtain results over a sweep-rate range of at least two orders of magnitude, in order not to reach erroneous conclusions. In addition, although subsequent cyclic voltammograms enable valuable mechanistic information to be deduced, the first sweep cycle only should be considered for accurate analysis of kinetic data.

If adsorbed electroactive species are present on the electrode surface, the shape of the cyclic voltammogram changes, since the species do not need to diffuse to the electrode surface. In this case the peaks are symmetrical with coincident peak potentials provided the kinetics are fast.

Table B1.28.1 Diagnostic tests for reversibility of electrode processes in cyclic voltammetry at 293 K.

Reversible process	Irreversible process	Quasi-reversible process
$\Delta E_p = E_p^A - E_p^C = 59/n$ (mV)	Absence of reverse peak	$\Delta E_p > 59/n$ (mV), increases with ν
$ E_p - E_{p/2} = 59/n$ (mV)	$ E_p - E_{p/2} = 48/\alpha_C n_\alpha$ (mV)	$ I_p^A/I_p^C = 1$, if $\alpha_C = \alpha_A = 0.5$
$ I_p^A/I_p^C = 1$	$I_p^C \propto \nu^{1/2}$	I_p increases with $\nu^{1/2}$; I_p not $\propto \nu^{1/2}$
$I_p \propto \nu^{1/2}$	E_p^C dependent of ν	E_p^C shifts with increasing ν
E_p independent of ν		
$E > E_p, I \propto t^{-1/2}$		

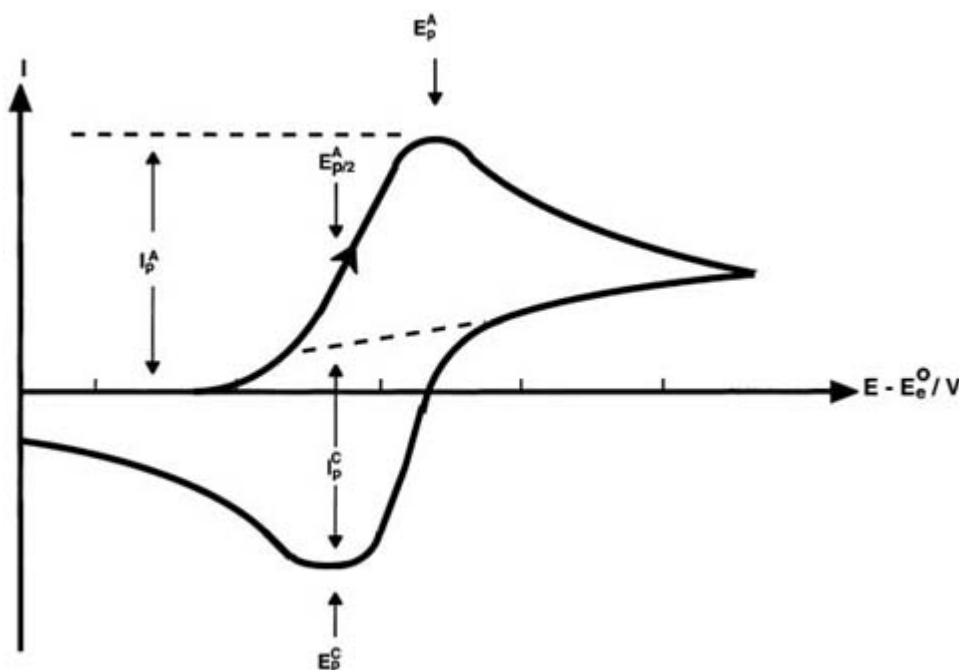


Figure B1.28.4. Cyclic voltammogram for a simple reversible electrode reaction in a solution containing only oxidized species.

On investigating a new system, cyclic voltammetry is often the technique of choice, since a number of qualitative experiments can be carried out in a short space of time to gain a feeling for the processes involved. It essentially permits an electrochemical spectrum, indicating potentials at which processes occur. In particular, it is a powerful method for the investigation of coupled chemical reactions in the initial identification of mechanisms and of intermediates formed. Theoretical treatment for the application of this technique extends to many types of coupled mechanisms.

B1.28.3.2 POTENTIAL-STEP TECHNIQUES

In a potential-step experiment, the potential of the working electrode is instantaneously stepped from a value where no reaction occurs to a value where the electrode reaction under investigation takes place and the current *versus* time (chronoamperometry) or the charge *versus* time (chronocoulometry) response is recorded. The transient obtained depends upon the potential applied and whether it is stepped into a diffusion control, in an electron transfer control or in a mixed control region. Under diffusion control the transient may be described by the Cottrell equation obtained by solving Fick's second law with the appropriate initial and boundary conditions [1, 2, 3, 4, 5 and 6]:

$$|j| = \frac{nFD^{1/2}c_i^\infty}{\pi^{1/2}t^{1/2}}.$$

Immediately after the imposition of a large negative overpotential in a solution containing oxidized species, O , a large current is detected, which decays steadily with time. The change in potential from E_c will initiate the very rapid reduction of all the oxidized species at the electrode surface and consequently of all the electroactive species diffusing to the surface. It is effectively an instruction to the electrode to instantaneously change the concentration of O at its surface from the bulk value to zero. The chemical change will lead to concentration gradients, which will decrease with time, ultimately to zero, as the diffusion-layer thickness increases. At time $t = 0$, on the other hand, $\partial c_i / \partial x)_{x=0}$ will tend to infinity. The linearity of a plot of j *versus* $t^{-1/2}$ confirms whether the reaction is under diffusion control and can be used to estimate values for the diffusion coefficient. It is a good technique for determining exact kinetic parameters when a mechanism is fully understood. Under mixed control, where the rates of diffusion and electron transfer are comparable, the current decays less steeply: at short times it will be controlled by electron transfer but, as the surface concentration is depleted, mass transport will become the rate-limiting step.

When analysing the data, it is important to consider a wide time range to ensure the reliability of the data, since at short times, <1 ms, it will be determined by the charging time of the double layer, and at longer times, >10 s, by the effects of natural convection.

Potential-step techniques can be used to study a variety of types of coupled chemical reactions. In these cases the experiment is performed under diffusion control, and each system is solved with the appropriate initial and boundary conditions.

Double potential steps are useful to investigate the kinetics of homogeneous chemical reactions following electron transfer. In this case, after the first step—raising to a potential where the reduction of O to R occurs under diffusion control—the potential is stepped back after a period τ , to a value where the reduction of O is mass-transport controlled. The two transients can then be compared and the kinetic information obtained by looking at the ratio of

the currents, which are a function of both τ and the homogeneous rate constant, k . This is a good method for obtaining exact information, provided that the mechanism is already understood.

B1.28.3.3 PULSE VOLTAMMETRY

Pulse techniques were originally devised to provide enhanced sensitivity in classical polarography for analytical applications [7, 8 and 9]. Sub-nanomolar detection limits can, in fact, be achieved with mercury electrodes where charging and background faradic currents are minimal. At solid electrodes (C, Pt, Au), charging currents and background currents arising from electrode surface reactions limit the level of analyte detection. However, pulse techniques remain particularly useful when looking at analyte concentrations of 10^{-5} M and lower, where voltammetric techniques such as linear-sweep voltammetry and cyclic voltammetry become limited by the difficulty of measuring faradic currents in the presence of background currents. Many step techniques have been devised, based on the succession of potential steps of varying height and in forward and reverse directions [2, 6, 7]. They find wide application in digitally based potentiostats, the electronics of which are suited to their exploitation. The current is normally sampled toward the end of the potential pulse, after the capacitative current has decayed, and the pulse widths are adjusted to fit between this limit and the onset of natural convection. Normal-pulse voltammetry (NPV), differential-pulse voltammetry (DPV) and square-wave voltammetry (SWV) are perhaps the most widely used of a variety of pulse techniques that have been developed.

In NPV [2, 7, 10, 11], short pulses of increasing height are superimposed on a constant base potential, E_b , where no reaction occurs (figure B1.28.5(a)). At the end of the pulse of width, t_p (typically 50–60 ms), the potential is returned to E_b , where it is held for another fixed period of a few seconds, before being pulsed again with a height increase determined by the scan rate. The current is sampled at the end of each pulse and the values are plotted against the potential to give a voltammetric profile similar to a steady-state voltammogram. The maximum current is given by the Cottrell equation, $j = nFDc_i^\infty / \pi^{1/2} t_s^{1/2}$, where t_s is the time at which the current is measured. NPV is therefore a good technique for determining diffusion coefficients.

In DPV [2, 7, 11] the pulse height is kept constant and the base potential is either swept constantly or is incremented in a staircase (figure B1.28.5(b)). The current is sampled just before the end of the pulse and just before pulse application, and the difference between the two measurements is plotted as a function of the potential. The resulting voltammogram is peak-shaped since it essentially is the differential of a steady-state shaped response, and for this feature the technique particularly lends itself to analytical purposes, enabling lower detection limits to be achieved. For a reversible system, the peak is symmetric with $E_p = E_{1/2} - \Delta E/2$, where ΔE is the pulse amplitude. In general, DPV is better at eliminating capacitative contributions and the peak-shaped response is useful for distinguishing two waves with close half-wave potentials.

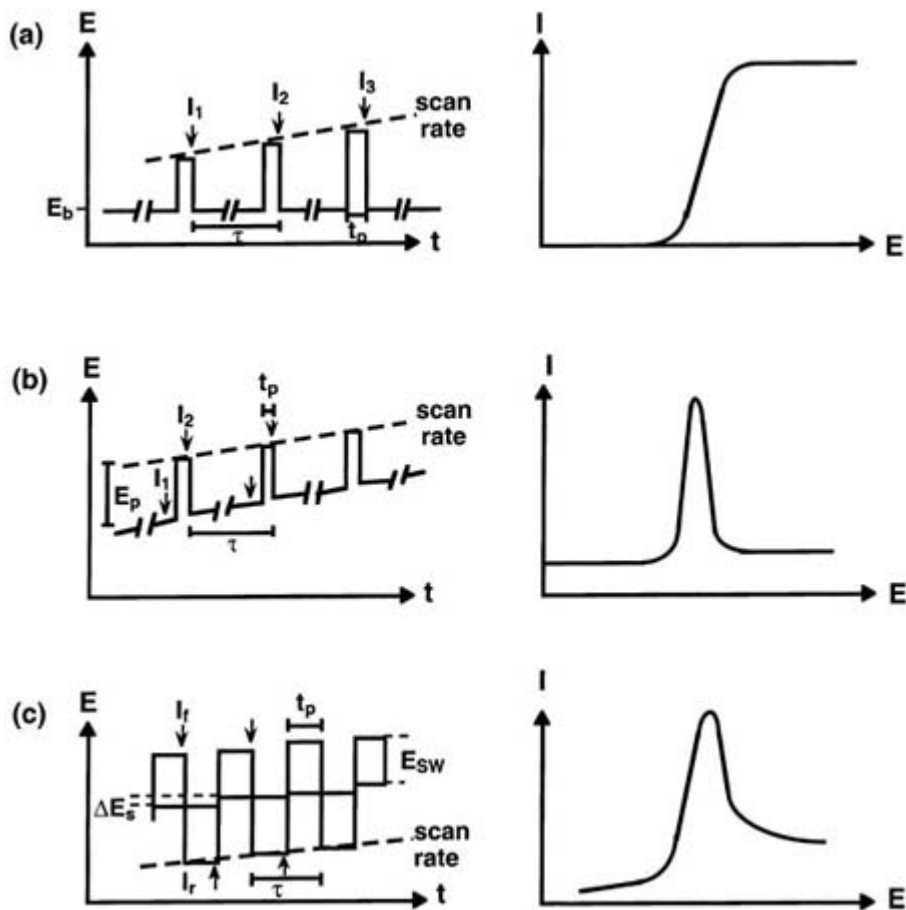


Figure B1.28.5. Applied potential–time waveforms for (a) normal pulse voltammetry (NPV), (b) differential pulse voltammetry (DPV), and (c) square-wave voltammetry (SWV), along with typical voltammograms obtained for each method.

SWV is an alternative voltammetric technique, first reported in 1952 by Barker and Jenkins [12] and subsequently developed into the form known today by Osteryoung *et al* [13, 14, 15 and 16]. The potential–time waveform is composed of a sequence of symmetrical square-wave pulses superimposed on an underlying ramp (figure B1.28.5(c)). The critical parameters are the step height of the underlying potential scan, ΔE_s ; the height of the square-wave pulse, E_{sw} ; the pulse width, t_p and the time at which the current is sampled on the forward and reverse pulses, t_s . Current measurements are made near the end of the pulse in each square-wave cycle: once at the end of the forward pulse and once at the end of the reverse pulse. However, capacitive contributions can be discriminated against before they decay, since over a small potential range between forward and reverse pulses, the capacity is constant and is thus annulled by subtraction. Consequently, shorter pulses than in DPV and NPV can be applied, enabling higher frequencies to be employed and much faster analysis to be carried out. The difference between the two currents, the net current, is plotted *versus* the base staircase potential, yielding a peak-shaped response. Since the square-wave modulation amplitude is large, the reverse pulses cause the reverse reaction to occur and, thus, the net current is larger

than either the forward or the reverse components. This, coupled with the effective discrimination against charging currents, enables a more sensitive analysis. The resulting peak-shaped voltammograms are symmetrical with characteristic position, width and height: the peak potential, E_p , coincides with the half-wave potential of a redox couple, the peak width indicates the effective number of electrons transferred and the peak current is proportional to the analyte concentration. In addition, the peak shape and position have been found to be largely independent of the size and geometry of the electrode. The net current is generally

compared with theoretical predictions of a dimensionless current Ψ , which are related by the Cottrell equation for the characteristic time:

$$j = \left(n F c \left(\frac{D}{\pi t_p} \right)^{1/2} \right) \Psi$$

where t_p is the pulse width. As well as for analysis, SWV has been found to be well suited to kinetic investigations.

B1.28.3.4 STRIPPING VOLTAMMETRY

Stripping voltammetry involves the pre-concentration of the analyte species at the electrode surface prior to the voltammetric scan. The pre-concentration step is carried out under fixed potential control for a predetermined time, where the species of interest is accumulated at the surface of the working electrode at a rate dependent on the applied potential. The determination step leads to a current peak, the height and area of which is proportional to the concentration of the accumulated species and hence to the concentration in the bulk solution. The stripping step can involve a variety of potential waveforms, from linear-potential scan to differential pulse or square-wave scan. Different types of stripping voltammetries exist, all of which commonly use mercury electrodes (dropping mercury electrodes (DMEs) or mercury film electrodes) [7, 17].

Anodic-stripping voltammetry (ASV) is used for the analysis of cations in solution, particularly to determine trace heavy metals. It involves pre-concentrating the metals at the electrode surface by reducing the dissolved metal species in the sample to the zero oxidation state, where they tend to form amalgams with Hg. Subsequently, the potential is swept anodically resulting in the dissolution of the metal species back into solution at their respective formal potential values. The determination step often utilizes a square-wave scan (SWASV), since it increases the rapidity of the analysis, avoiding interference from oxygen in solution, and improves the sensitivity. This technique has been shown to enable the simultaneous determination of four to six trace metals at concentrations down to fractional parts per billion and has found widespread use in seawater analysis.

Cathodic stripping voltammetry follows a similar sequence of events, except that trace anionic species are reduced in the form of insoluble salts with metal constituents on the electrode surface, e.g. Ag and Hg, during application of a short, relatively positive deposition potential. The applied potential is then swept linearly or pulsed from the deposition potential in the cathodic direction resulting in the selective desorption of the anionic species according to the respective formal potential values. Cathodic stripping voltammetry can be used to determine organic and inorganic compounds that form an insoluble film at the electrode surface. Various inorganic analytes such as halide ions, sulphide ions and oxo-anions are capable of forming insoluble Hg salts which can be pre-concentrated on the Hg electrode surface and be measured.

Adsorptive stripping analysis involves pre-concentration of the analyte, or a derivative of it, by adsorption onto the working electrode, followed by voltammetric measurement of the surface species. Many species with surface-active properties are measurable at Hg electrodes down to nanomolar levels and below, with detection limits comparable to those for trace metal determination with ASV.

Improved sensitivities can be attained by the use of longer collection times, more efficient mass transport or pulsed waveforms to eliminate charging currents from the small faradic currents. Major problems with these methods are the toxicity of mercury, which makes the analysis less attractive from an environmental point of view, and surface fouling, which commonly occurs during the analysis of a complex solution matrix. Several methods have been reported for the improvement of the pre-concentration step [17, 18]. The latter is, in fact,

strongly influenced by the choice of solvent, electrode material, pH, electrode potential and temperature. A constant mass-transport rate leads to better reproducibility and hence stirring is often used with static mercury drop electrodes and stationary electrodes. Hydrodynamic electrodes are also employed in order to increase the sensitivity and decrease the detection limits.

Recent years have witnessed the exploitation of stripping voltammetry in chemical sensors. Complex, fixed-site ASV analysers are used to determine a wide range of metals, such as Cr, Ni, Cu, V, Sn, As and Cd, in the effluents from mining, mineral processing, metal-finishing and related industries. The portable instrumentation and low power demands of stripping analysis satisfy many of the requirements for on-site *in situ* measurements. The development of remotely deployed submersible stripping probes, easy-to-use microfabricated metal sensor strips and micromachined, hand-held total stripping analysers have been reported to move the measurement of trace metals to the field and to perform them more rapidly, reliably and inexpensively [17, 18, 19 and 20].

B1.28.4 STEADY-STATE TECHNIQUES

In the study of electrode reactions, the rates of electron transfer are very often high compared to mass transport, rendering the extrapolation of mechanistic and kinetic data unfeasible. It is therefore essential for the study of electrode reactions and the extrapolation of kinetic information to disrupt the equilibrium by increasing the rate of mass transport and forcing the process into a mixed-control region where the rate of electron transfer is comparable to that of mass transport. There are several methods available for increasing and varying the rate of mass transport in a controlled way, amongst which are hydrodynamic electrodes and microelectrodes [1, 2, 3 and 4]. In both cases, the regime may be described by solvable systems that may be used to predict the rate of mass transport and in the interpretation of experimental data. In hydrodynamic electrodes, the increased rate of mass transport of species is brought about by external mechanical forces, which can arise from the movement of the electrode, agitation of the solution or flowing of the solution past the electrode surface. The resulting forced convection leads to the thinning of the Nernst diffusion layer with a consequent increase in the linear concentration gradient that exists across it and hence to current densities as large as 100 times greater than the steady-state diffusion-limited value. By measuring the current–potential response as a function of mass transport it is thus possible to extrapolate kinetic information regarding an electrode reaction, provided it is under mixed control. There are a number of electrode designs that fall into the category of hydrodynamic electrodes, which include the rotating-disc electrode (RDE), the rotating ring–disc electrode (RRDE), the wall-jet electrode, the wall-pipe electrode, the tube electrode and the channel electrode [21, 22, 23, 24, 25, 26 and 27]. The RDE and RRDE are perhaps the most commonly employed in kinetic and mechanistic studies, and these will be further discussed together with the channel electrode. Microelectrodes, scanning electrochemical microscopy (SECM) and sonoelectrochemistry are also discussed.

B1.28.4.1 ROTATING-DISC ELECTRODES

A rotating-disc electrode (RDE) consists of a disc of electrode material embedded into a larger insulating sheath, and attached to the rotor spindle *via* a suitable electrical contact. The disc and sheath are rotated about a vertical axis. Upon rotation, a pump-action flow is initiated, which brings solution perpendicularly to the electrode surface and throws it out in a radial direction on meeting disc and sheath (figure B1.28.6). A more quantitative description of the flow patterns can be made by the use of cylindrical polar coordinates, by looking at the variation of the solution-flow velocity components V_x , V_r and V_θ as a function of x , the distance perpendicular to the surface of the electrode. The change in concentration of an electroactive species with time due to convection and diffusion may be written as [1, 2, 4, 5]

$$\frac{\partial c_i}{\partial t} = D \left[\frac{\partial^2 c_i}{\partial x^2} + \frac{\partial^2 c_i}{\partial r^2} + \frac{1}{r} \frac{\partial c_i}{\partial r} + \frac{1}{r^2} \frac{\partial^2 c_i}{\partial \theta^2} \right] - \left[V_x \frac{\partial c_i}{\partial x} + V_r \frac{\partial c_i}{\partial r} + \frac{V_\theta}{r} \frac{\partial c_i}{\partial \theta} \right].$$

diffusion convection

However, the equation can be simplified, since the system is symmetrical and the radius of the disc is normally small compared to the insulating sheath. The access of the solution to the electrode surface may be regarded as uniform and the flux may be described as a one-dimensional system, where the movement of species to the electrode surface occurs in one direction only, namely that perpendicular to the electrode surface:

$$\frac{\partial c_i}{\partial t} = D \frac{\partial^2 c_i}{\partial x^2} - V_x \frac{\partial c_i}{\partial x} \quad \text{where} \quad V_x = -0.51 \omega^{3/2} \nu^{-1/2} x^2.$$

The importance of convection in the system increases as the square of the distance from the electrode surface, and close to the surface it is not a dominant form of mass transport. Hence concentration changes will arise due to both diffusion and convection. In the Nernst diffusion model, this trend is exaggerated, and for the mass transport behaviour at an RDE, a plot of the concentration of electroactive species, c_i , versus the distance from the electrode surface, x , is divided into two distinct zones (figure B1.28.6). At the electrode surface, i.e. $x = 0$, the concentration of the electroactive species will be c_i^0 and up to a distance δ away from the electrode, there is a stagnant layer, in which diffusion is the only form of mass transport (the Nernst diffusion layer). Outside this layer, mass transport is dominated by strong convection and the concentration is maintained at the bulk value, c_i^∞ . The diffusion-layer thickness is determined by the rotation rate of the disc, the layer becoming thinner with increasing rotation rate. In this model, the values of c_i^0 and δ will depend on the applied potential and the electrode rotation rate, respectively. In a linear-sweep experiment at a given rotation rate, the concentration profiles, (dc_i/dx) , within the diffusion layer will vary linearly as the applied potential at the RDE is swept from a value where no electron transfer occurs towards values positive to E_c^0 . As the experiment is driven further, the surface concentration of the electroactive species eventually reaches zero, at which point the current response reaches its limiting plateau value. The limiting current density is expressed by [1, 2, 4, 5]

$$j_L = \frac{nFDc_i^\infty}{\delta} = nFk_m c_i^\infty$$

-15-

where k_m is the mass-transport coefficient and the diffusion-layer thickness is $\delta = 1.61 \nu^{1/6} c_i^\infty \omega^{-1/2}$, where ω is the rotational speed in rad s^{-1} , and ν is the kinematic viscosity of the solution. Substituting for δ leads to the Levich equation

$$j_L = 0.621 n F D^{2/3} \nu^{-1/6} c_i^\infty \omega^{1/2}.$$

The expression for the mass-transport-limiting current density may be employed together with the Nernst equation to deduce the complete current–potential response in a solution containing only oxidized or reduced species

$$E = E_c^0 + \frac{2.3RT}{nF} \log \frac{I_L - I}{I}$$

in which I are current values from the rising portion of the curve (under mixed control). This equation, of

course, only holds for fast electron-transfer reactions, where the surface concentrations are related through the Nernst equation. For a reversible electrode reaction, therefore, a plot of potential *versus* the logarithmic quotient should have a slope of $(59/n)$ mV (at 298 K) and the formal potential for the couple will coincide with the half-wave potential, $E_{1/2}$, of the curve (potential corresponding to $1/2 I_L$) (figure B1.28.7). On reversing the scan, the current–potential curve will exactly retrace the forward scan, as the electroactive species will continue to be reduced or oxidized at the same potentials as in the forward sweep. The product formed at the electrode surface during both scans quickly disappears into the bulk solution through convection and is not available for the reverse electron transfer during the back-sweep. In addition to analysing the shape of a current–potential curve at a single rotation rate, the relationship between limiting current densities and mass transport can also be investigated. A common treatment for RDE data is to plot the limiting current density *versus* the square root of rotation speed, with a linear plot confirming conditions of mass-transport control, and the slope being used to determine the other parameters.

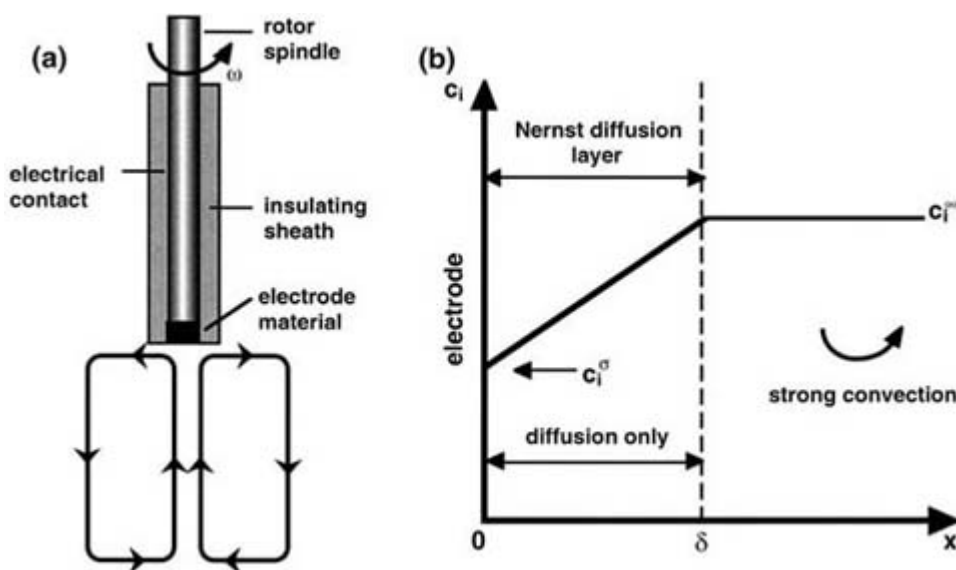


Figure B1.28.6. (a) Convection within the electrolyte solution, due to rotation of the electrode; (b) Nernst diffusion model for steady state.

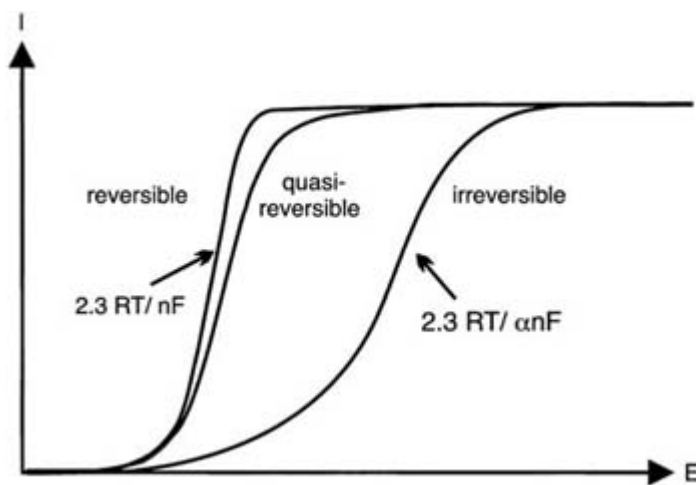


Figure B1.28.7. Schematic shape of steady-state voltammograms for reversible, quasi-reversible and irreversible electrode reactions.

In the case of an irreversible electrode reaction, the current-potential curve will display a similar shape, with

j_L still proportional to $\omega^{1/2}$, but the curve is drawn out along the potential axis. The current-potential curve may be described by [1, 2, 4, 5]

-17-

$$E = E_{1/2} + \frac{2.3RT}{\alpha nF} \log \frac{I_L - I}{I}$$

in which the log plot remains linear but with a slope of $2.3RT/\alpha nF$ (figure B1.28.7). The most obvious feature of the irreversible voltammogram is that the half-wave potential no longer falls near the reversible formal potential, reflecting the sluggish electron transfer kinetics. An activation overpotential is required to drive the reaction.

The rotating-disc is also well suited to the study of coupled chemical reactions [2, 4].

It is essential for the rotating-disc that the flow remain laminar and, hence, the upper rotational speed of the disc will depend on the Reynolds number and experimental design, which typically is 1000 s^{-1} or 10,000 rpm. On the lower limit, 10 s^{-1} or 100 rpm must be applied in order for the thickness of the boundary layer to be comparable to that of the radius of the disc.

The great advantage of the RDE over other techniques, such as cyclic voltammetry or potential-step, is the possibility of varying the rate of mass transport to the electrode surface over a large range and in a controlled way, without the need for rapid changes in electrode potential, which lead to double-layer charging current contributions.

B1.28.4.2 ROTATING RING-DISC ELECTRODES

The rotating ring-disc electrode (RRDE) consists of a central disc separated from a concentric ring electrode by a thin, non-conducting gap. It was first developed by Frumkin and Nekrasov to detect unstable intermediates in electrochemical reactions [1, 2, 22]. As with the RDE, on rotation of the disc, solution is pulled towards the centre of the disc and then thrown out radially across the surface of the structure. The ring is effectively situated downstream to the disc. This permits the intermediates formed on the disc, as the result of an oxidation or reduction process, to be detected at the ring following their mass transport across the insulating gap between the electrodes. Hence, information on intermediates can be obtained before they reach the bulk solution or react further with the electrolyte solution. The ring and the disc are independent from one another and can hence be potentiostatted independently.

In order to employ the RRDE for quantitative studies, it is necessary to describe the transport of species from disc to ring. In the absence of homogeneous chemical reactions, the electrogenerated species at the disc reaction is transported to the ring by diffusion across the stagnant layer at the electrode surface, by convection across the gap and diffusion across the stagnant layer at the ring electrode. The collection efficiency, N_0 , is defined as the ratio of the mass-transport-controlled current for the electrode reactions at ring and disc, $N_0 = -i_{\text{ring}}/i_{\text{disc}}$, where the minus sign arises because the reactions at the ring and at the disc occur in the opposite direction. The collection efficiency thus represents the fraction of material produced at the disc that is detected at the ring. Analytical solutions of the convective-diffusion transport at the ring-disc enables the collection efficiency for specific disc and ring dimensions to be calculated:

$$N_0 = 1 - F\left(\frac{\alpha}{\beta}\right) + \beta^{2/3} [1 - F(\alpha)] - (1 + \alpha + \beta)^{2/3} \left\{ 1 - F\left[\left(\frac{\alpha}{\beta}\right)(1 + \alpha + \beta)\right] \right\}$$

where α , β and F are defined as

$$\alpha = \left(\frac{r_2}{r_1}\right)^3 - 1$$

$$\beta = \left(\frac{r_3}{r_1}\right)^3 - \left(\frac{r_2}{r_1}\right)^3$$

$$F(\theta) = \frac{3^{1/2}}{4\pi} \ln \left\{ \frac{(1 + \theta^{1/3})^3}{1 + \theta} \right\} + \frac{3}{2\pi} \arctan \left(\frac{2\theta^{1/3} - 1}{3^{1/2}} \right) + \frac{1}{4}$$

and r_1 , r_2 and r_3 are the radius of the disc, the radius of the disc surrounded by the insulating sheath and the radius of the disc surrounded by sheath and ring, respectively. The collection efficiency is a function of r_1 , r_2 and r_3 and does not depend on the rotation speed or the nature of the redox species. Since access of material to the ring is not uniform, some of the material will be transported back into solution and collection efficiencies are typically around 0.2–0.3 and strongly depend on the geometry of the electrodes and the distance between them. The rotation rate will affect the time taken for the intermediates to be transported from the disc to the ring, short-lived intermediates requiring higher rotation rates and the construction of RRDEs with thin inter-electrode gaps. The rotation rate will not, however, affect the efficiency, since the currents at both the generator and collector electrodes will be enhanced.

A number of different types of experiment can be designed, in which disc and ring can either be swept to investigate the potential region at which the electron transfer reactions occur, or held at constant potential (under mass-transport control), depending on the information sought.

The RRDE is very useful for the detection of short-lived intermediates, in the investigation of reaction mechanisms, but also in the distinction of free and adsorbed intermediates, as the latter are not transported to the ring.

B1.28.4.3 CHANNEL-FLOW ELECTRODES

Forced convection can also arise from the movement of electrolyte solution over a stationary working electrode. In a channel electrode, the electrode is embedded smoothly in one wall of a thin, rectangular duct through which electrolyte is mechanically pumped [3, 6, 26, 27]. The design of the flow cell consists of two plates sealed together, with typical dimensions of 30–50 mm in length, <10 mm in width and a distance between the plates of less than 1 mm (the cell height). The electrode is embedded either at the centre of the base plate or attached to the centre of the cover plate by means of an adhesive.

The solution flow is normally maintained under laminar conditions and the velocity profile across the channel is therefore parabolic with a maximum velocity occurring at the channel centre. Thanks to the well defined hydrodynamic flow regime and to the accurately determinable dimensions of the cell, the system lends itself well to theoretical modelling. The convective–diffusion equation for mass transport within the rectangular duct may be described by

$$\frac{\partial c_i}{\partial t} = D \left(\frac{\partial^2 c_i}{\partial x^2} + \frac{\partial^2 c_i}{\partial y^2} + \frac{\partial^2 c_i}{\partial z^2} \right) c_i - \left(v_x \frac{\partial c_i}{\partial x} + v_y \frac{\partial c_i}{\partial y} + v_z \frac{\partial c_i}{\partial z} \right)$$

where v_x , v_y , v_z are the solution-velocity profiles in the directions x , y , z . By convention, the direction of flow is designated as the x -direction and the y -direction is that normal to the electrode. The equation may be considerably simplified by removal of the time dependence under steady-state conditions and by neglecting axial diffusion to the macroelectrode, since convection is considerably faster. The diffusion layer is situated very close to the electrode surface and is small compared to the cell depth, decreasing as the flow rate is increased. The Lévêque approximation further simplifies the system by approximating the parabolic flow to a linear flow near the electrode surface, provided that the electrode is less wide than the channel (for edge effects to be neglected) and the height, h , of the channel is much greater than the width, d . In an analogous fashion as for the RDE, solution for a simple mass-transport-limited electrode reaction leads to the Levich equation

$$I_{\text{lim}} = 0.925nF c_i^{\infty} D^{2/3} v_f^{1/3} (h^2 d)^{-1/3} w x_e^{2/3}$$

where v_f is the solution volume flow rate, x_e the length of the electrode, and d and w the height and width of the cell, respectively. Analytical solutions for the channel electrode also extend to more complicated electrode reactions involving coupled homogeneous reactions.

Amongst the greatest advantages of channel-flow electrodes is the possibility of controlling the rate of mass transport over a range of three orders of magnitude, from 10^{-4} to $10^{-1} \text{ cm}^3 \text{ s}^{-1}$, and of varying the mass-transport coefficient by altering the cell depth and the electrode length. These rates are, in fact, not attainable at other hydrodynamic electrodes, such as the RDE and the wall-jet electrode. In addition, there is no risk of a build-up of a stagnant zone since the spent solution flows to waste.

The channel-flow electrode has often been employed for analytical or detection purposes as it can easily be inserted in a flow cell, but it has also found use in the investigation of the kinetics of complex electrode reactions. In addition, channel-flow cells are immediately compatible with spectroelectrochemical methods, such as UV/VIS and ESR spectroscopy, permitting detection of intermediates and products of electrolytic reactions. UV–VIS and infrared measurements have, for example, been made possible by constructing the cell from optically transparent materials.

B1.28.4.4 MICROELECTRODES

A microelectrode is an electrode with at least one dimension small enough that its properties are a function of size, typically with at least one dimension smaller than $50 \mu\text{m}$ [28, 29, 30, 31, 32 and 33]. If compared with electrodes employed in industrial-scale electrosynthesis or in laboratory-scale synthesis, where the characteristic dimensions can be of the order of metres and centimetres, respectively, or electrodes for voltammetry with millimetre dimension, it is clear that the size of the electrodes can vary dramatically. This enormous difference in size gives microelectrodes their unique properties of increased rate of mass transport, faster response and decreased reliance on the presence of a conducting medium. Over the past 15 years, microelectrodes have made a tremendous impact in electrochemistry. They have, for example, been used to improve the sensitivity of ASV in environmental analysis, to investigate rapid

The *increased rate of mass transport* is one of the most attractive and advantageous properties of microelectrodes over conventional electrodes, as the increased transport of the reactant to the electrode surface allows it to reach steady-state regimes rapidly. The diffusion rates increase with decreasing electrode size beyond that obtained with other steady-state techniques. For example, with a 10 μm diameter disc, the steady-state, mass-transfer coefficient, k_m , is comparable to that of a rotating disc revolving at an experimentally impossible 250,000 rpm. The *discrimination against charging currents* is another very important property. In fact, the magnitude of the charging current depends on the area of the capacitor, and for a microelectrode it decreases with electrode area. Thus, a microelectrode has a very reduced interface capacitance, and the charging current decays much more quickly than with conventional electrodes and faster response times may be achieved. Another property of microelectrodes is the *decreased distortion from IR_u* , the potential drop between working and reference electrodes generated by the passage of current through a solution and expressed in terms of the product of the solution resistance and the current flowing in the circuit. With conventional electrodes, it is usual to add supporting electrolyte to minimize the solution resistance, but with microelectrodes, the current passing through the cell is low, often of the order of 10^{-9} A, and hence problems with IR_u drop are greatly reduced. This proves to be an advantage to experiments with either a large current, I , or a large resistance, R , as for example in experiments with solvents of very low dielectric constant, in media with very low ionic strength, or in studies of solutions with high concentrations of electroactive species. Electrochemical measurements can be therefore made in new and unique chemical environments, which are not amenable at larger electrodes, and experiments have been reported in frozen acetonitrile, low-temperature glasses, ionically conductive polymers, oil-based lubricants and milk. In addition, the use of electrolyte-free organic media can greatly extend the electrochemical potential window, thus allowing studies of species with high redox potentials. Furthermore, such dimensions offer obvious analytical advantages, including the exploration of microscopic domains, measurement of local concentration profiles, detection in micro-flow systems and analysis of very small sample volumes [28].

Microelectrodes with several geometries are reported in the literature, from spherical to disc to line electrodes; each geometry has its own critical characteristic dimension and diffusion field in the steady state. The diffusional flux to a spherical microelectrode surface may be regarded as planar at short times, therefore displaying a transient behaviour, but spherical at long times, displaying a steady-state behaviour [28, 34]. If a potential is applied so that the reaction $O + ne^- \rightarrow R$, becomes diffusion controlled, the current density at a microsphere electrode can be expressed by

$$j = \frac{nFD^{1/2}c_1}{\pi^{1/2}t^{1/2}} + \frac{nFDC_1}{r}$$

transient term steady-state term

This expression is the sum of a transient term and a steady-state term, where r is the radius of the sphere. At short times after the application of the potential step, the transient term dominates over the steady-state term, and the electrode is analogous to a plane, as the depletion layer is thin compared with the disc radius, and the current varies with time according to the Cottrell equation. At long times, the transient current will decrease to a negligible value, the depletion layer is comparable to the electrode radius, spherical diffusion controls the transport of reactant, and the current density reaches a steady-state value. At times intermediate to the limiting conditions of Cottrell behaviour or diffusion control, both transient and steady-state terms need to be considered and thus the full expression must be used. However, many experiments involving microelectrodes are designed such that one of the simpler current expressions is valid.

Of course, in order to vary the mass transport of the reactant to the electrode surface, the radius of the electrode must be varied, and this implies the need for microelectrodes of different sizes. Spherical electrodes are difficult to construct, and therefore other geometries are often employed. Microdiscs are commonly used in the laboratory, as they are easily constructed by sealing very fine wires into glass epoxy resins, cutting

perpendicular to the axis of the wire and polishing the front face of the disc that is created [30]. Because of its planar geometry, the diffusion field over the surface of a microdisc is non-uniform and the flux only approximates that of a hemisphere. The rate of diffusion to the edge of the disc will be higher than to the centre. Therefore, the rates of diffusion to the disc are estimated as space-averaged quantities and a factor of $4/\pi$ is required to adjust the equation for a spherical microelectrode to describe the diffusion of reactant to the surface of a microdisc electrode, which becomes $j = 4nFDc_1^\infty/\pi r$, where r is now the radius of the disc.

Similarly to the response at hydrodynamic electrodes, linear and cyclic potential sweeps for simple electrode reactions will yield steady-state voltammograms with forward and reverse scans retracing one another, provided the scan rate is slow enough to maintain the steady state [28, 35, 36, 37 and 38]. The limiting current will be determined by the slowest step in the overall process, but if the kinetics are fast, then the current will be under diffusion control and hence obey the above equation for a disc. The slope of the wave in the absence of IR_u drop will, once again, depend on the degree of reversibility of the electrode process.

All types of voltammetry may be applied to microelectrodes, including normal, reverse pulse and square-wave voltammetry. Pulse voltammetry and potential-step program at microelectrodes discriminate against charging currents and the boundary conditions are set much faster between pulses, since the electrode responds in a much more rapid fashion to a potential change [11, 15, 16]. Cyclic voltammetry measurements can be made on a much more rapid time scale than with electrodes of conventional size and be operated in the range of tens of nanoseconds, without important distortion by IR_u drop and concerns regarding charging currents. This renders the characterization of rates and mechanisms of very fast chemical reactions as well as determination of trace quantities of transient species possible. At high sweep rates, however, only linear diffusion needs to be considered [28].

The advantages of microelectrodes for low-volume detection and spatially and temporally resolved measurements have been largely exploited in biology and medicine [39, 40 and 41]. One of the most active and longer-standing fields is neuroscience, where the development of the electroanalysis of brain extracellular fluid has been remarkable, since it can be relatively non-invasive due to the small size and low currents flowing. An example is the monitoring of the release of neurotransmitters with carbon microelectrodes in either amperometric mode or using fast cyclic voltammetry [42, 43]. Carbon materials are the most common starting materials and have also been applied to other electroactive compounds such as histamine, anticancer drugs and ascorbic acid. Extension of the investigation of cellular systems to non-electroactive neurochemicals has led to the development of enzyme-modified microelectrodes for measurements of glutamate, glucose, and choline and acetylcholine. Voltammetric measurements have also been reported in single cells, although the living cells are separated from the parent organism. Microelectrodes have also found widespread use in sensor technology and environmental analysis. Due to the high rate of steady-state diffusion at a microelectrode, their response is independent of convection, thus enabling their use for the analysis of flowing systems. In order to enhance the current response at microelectrodes, a number of approaches have been described. Amongst these are random arrays of microdisc electrodes [44, 45 and 46] and interdigitated arrays of microband electrodes [47, 48]. Arrays of microelectrodes enable the enhancement of the current response, whilst retaining the properties of a single microelectrode, and have been used as highly sensitive detectors in flow-injection analysis and in liquid chromatography.

B1.28.4.5 SCANNING ELECTROCHEMICAL MICROSCOPY

SECM is a scanning-probe technique introduced by Bard *et al* in 1989 [49, 50 and 51] based on previous studies by the same group on *in situ* STM [52] and simultaneous work by Engstrom *et al* [53 and 54], who were the first to show that an amperometric microelectrode could be used as a local probe to map the concentration profile of a larger active electrode. SECM may be envisaged as a 'chemical' microscope based on faradic current changes as a microelectrode is moved across a surface of a sample. It has proved useful for

obtaining topographical and chemical information on a wide range of sample surfaces, including electrodes, minerals, polymers and biological materials.

The apparatus consists of a tip-position controller, an electrochemical cell with tip, substrate, counter and reference electrodes, a bipotentiostat and a data-acquisition system. The microelectrode tip is held on a piezoelectric pusher, which is mounted on an inchworm-translator-driven x - y - z three-axis stage. This assembly enables the positioning of the tip electrode above the substrate by movement of the inchworm translator or by application of a high voltage to the pusher *via* an amplifier. The substrate is attached to the bottom of the electrochemical cell, which is mounted on a vibration-free table [55, 56, 57 and 58]. A number of different size and shape tips have been reported. The most common are disc shaped with diameters of 0.6–25 μm formed by sealing a Pt, Au wire or carbon fibre of the required radius in a glass capillary and polishing the sealed end. The glass wall surrounding the disc is sharpened to a conical shape to decrease the possibility of contact between glass and substrate as the tip is moved close to the latter. For most studies, the ratio of the diameter of the entire tip end, including the insulator, to that of the electrode itself should typically be ≈ 10 . Metal electrodes down to the nanometre scale have also been fabricated by sealing an etched Pt or Pt–Ir wire in a suitable insulating material, leaving the etched end exposed. Commercial SECM instruments have only recently appeared on the market.

With SECM, almost any kind of electrochemical measurement may be carried out, whether voltammetric or potentiometric, and the addition of spatial resolution greatly increases the possibilities for the characterization of interfaces and kinetic measurements [55, 56, 57, 58 and 59]. It may be employed as an electrochemical tool for the investigation of heterogeneous and homogeneous reactions, as an imaging device, or for microfabrication, making use of different modes of operation. In amperometric *feedback mode* a three- or four-electrode configuration is employed, in which a microelectrode tip serves as the working electrode, the potential is controlled *versus* the reference electrode and the current flows between tip and counter-electrodes. The potential of the sample may also be controlled and it may thus serve as a second working electrode. The electrolyte solution contains a redox mediator, e.g. a reducible species O , such that when a suitably negative potential is applied to the tip, its reduction takes place at a rate governed by diffusion of the electroactive species to the electrode. If the tip is more than several tip diameters away from the surface, the steady-state current is given by $I_{T_{\infty}} = 4nFDc_i^{\infty}r$, for a disc-shaped tip, where r is the radius of the tip. However, when the tip is brought within a few tip radii to a conductive substrate, the reduced species formed at the tip diffuses to the substrate where it is re-oxidized. As a consequence, an additional flux of O to the tip is produced which leads to an increase in the tip current, known as positive feedback. The smaller the tip–substrate distance the larger is the effect. In contrast, if the substrate is an electrical insulator, the reducible species cannot be regenerated and, since the diffusion of O from the bulk is hindered at small distances to the substrate, the tip current will be smaller than $I_{T_{\infty}}$, i.e. negative feedback. Therefore, by scanning over the surface of a substrate, the variation in current can be related to changes in the distance and hence to the topography of the substrate. Besides feedback mode, several other modes exist, such as *generation/collection mode*, where species generated at one working electrode are detected at the second, *penetration mode*, in which a small tip is used to penetrate a microstructure and extract spatially resolved information about concentrations, kinetic and mass-transport parameters, and *ion-transfer feedback mode*, recently developed and useful for studies of ion-transfer reactions at liquid/liquid and liquid/membrane interfaces. The SECM

methodologies are based on quantitative theory, which has been developed for a variety of systems involving heterogeneous and homogeneous processes and different tip and substrate geometries. In many cases, analytical approximations allow the generation of theoretical dependences and an analysis of experimental data [60, 61, 62, 63 and 64].

The high rate of mass transfer in SECM enables the study of fast reactions under steady-state conditions and allows the mechanism and physical localization of the interfacial reaction to be probed. It combines the useful

features of microelectrodes and thin-layer cells in dimensions not easily attainable in larger electrochemical cells. The mass-transfer rate in SECM is a function of the tip–substrate distance. At large distances, d , $k_m \approx D/r$, whereas for small distances ($d < r$), $k_m \approx D/d$. The large effective k_m obtainable enables fast heterogeneous reaction rates to be measured under steady-state conditions. Zhou and Bard measured a rate constant of 6×10^7 Ms for the electro-hydrodimerization of acrylonitrile (AN) and observed the short-lived intermediate AN^- for this process [65].

Heterogeneous reactions at a substrate can also be probed without the need for an external voltage, if for example the mediator regeneration is chemical in nature rather than electrochemical. This has opened the possibility of studying dissolution of ionic single crystals and locating individual sites of reactivity in multiphase systems. It can be used to map the local surface reactivity in either feedback or collection mode. Feedback mode has been employed, for example, to probe the surface reactivity of a titanium substrate covered with TiO_2 and to individuate precursor sites for pitting corrosion, whereas collection mode has been used to image fluxes of species produced or consumed at the substrate, such as iontophoretic fluxes of electroactive species through porous membranes [56, 57, and 58].

Among the systems most studied by SECM are heterogeneous electron transfer reactions at the metal/ solution interface. Nonetheless, the diversity of interfaces and processes that can be studied with SECM has grown to include liquid/liquid and liquid/gas interfaces [66], and materials of biological significance [67]. It is a promising technique for the mapping of biochemical activity, for example transport in tissues and immobilized enzyme kinetics. Detection of single molecules has also recently been reported [68].

B1.28.4.6 SONOELECTROCHEMISTRY

The technique of applying ultrasound during electrochemical measurements and reactions is known as sonoelectrochemistry or sonovoltammetry and is a field that has grown rapidly in recent years [69, 70, 71 and 72]. The dominant ultrasonic effects are the enhanced mass transport of electroactive substrate to the electrode and the activation of the electrode surface through a cavitation cleaning action. The latter are violent collapses of oscillating bubbles, which can cause effects such as depassivation and erosion. The huge effects on the rate of mass transport to the electrode surface may be envisaged as an extremely thinned diffusion layer of uniform accessibility, partly induced by acoustic streaming, when ultrasonic horn transducers are employed, and by cavitation collapse of micro-bubbles at the solid–liquid interface.

Two major sources of ultrasound are employed, namely ultrasonic baths and ultrasonic immersion horn probes [70, 71]. The former consists of fixed-frequency transducers beneath the exterior of the bath unit filled with water in which the electrochemical cell is then fixed. Alternatively, the metal bath is coated and directly employed as electrochemical cell, but in both cases the results strongly depend on the position and design of the set-up. The ultrasonic horn transducer, on the other hand, is a transducer provided with an electrically conducting tip (often Ti6Al4V), which is immersed in a three-electrode thermostatted cell to a depth of 1–2 cm directly facing the electrode surface.

The ultrasound intensity and the distance between the horn and the electrode may be varied at a fixed frequency, typically of 20 kHz. This cell set-up enables reproducible results to be obtained due to the formation of a macroscopic jet of liquid, known as acoustic streaming, which is the main physical factor in determining the magnitude of the observed current.

The effects of ultrasound-enhanced mass transport have been investigated by several authors [73, 74, 75 and 76]. Empirically, it was found that, in the presence of ultrasound, the limiting current for a simple reversible electrode reaction exhibits quasi-steady-state characteristics with intensities considerably higher in magnitude compared to the peak current of the response obtained under silent conditions. The current density can be

described by $j_{\text{lim}} = nFDc_i^\infty/\delta$, where the diffusion layer δ depends on the distance between the horn and the electrode and on the ultrasound intensity. Superimposed on the faradic current a fluctuation or noise is also detected consistent with the turbulent nature of the macroscopic jet of liquid and the presence of oscillating and cavitating bubbles.

In an alternative design, the actual tip of the ultrasonic horn may be used as the working electrode after insertion of an isolated metal disc [77, 78 and 79]. With this electrode, known as the sonotrode, very high limiting currents are obtained at comparatively low ultrasound intensities, and diffusion layers of less than 1 μm have been reported. Furthermore, the magnitude of the limiting currents has been found to be proportional to $D^{2/3}$, enabling a parallel to be drawn with hydrodynamic electrodes.

The cleaning or depassivation effect is of great importance in sonoelectrochemistry, as it can be employed to wash off surface-adsorbed species and reduce blocking of the electrode by adsorption of reaction products. This effect has been reported, for example, for the depassivation of iron electrodes and for the removal of deposits and in the presence of polymer films on the electrode surface. However, damage of the electrode surface, especially for materials of low hardness such as lead or copper, can also occur under harsh experimental conditions and applied intensities [70, 71, 80].

Sonoelectrochemistry has been employed in a number of fields such as in electroplating for the achievement of deposits and films of higher density and superior quality, in the deposition of conducting polymers, in the generation of highly active metal particles and in electroanalysis. Furthermore, the sonolysis of water to produce hydroxyl radicals can be exploited to initiate radical reactions in aqueous solutions coupled to electrode reactions.

B1.28.5 ELECTROCHEMICAL IMPEDANCE SPECTROSCOPY

In contrast to transient techniques, which involve the perturbation of a system and studying its relaxation with time, when the perturbation is sinusoidal, the analysis is performed in the frequency domain, which is obtained by applying Laplace transforms to the time-domain information. Alternating-current impedance techniques employ the ratio of an imposed sinusoidal voltage and the resulting sinusoidal current to define the impedance, which is a function of the frequency of the signal. When a steady-state system is perturbed by an applied AC voltage, it relaxes to a new steady state and the time taken for this relaxation is known as τ . $\tau = RC$, where R is the resistance and C the capacitance of the system. Analysis of this relaxation process provides information about the system. In the frequency domain, fast processes, low τ , occur at high frequencies, while slow processes, with high τ , occur at low frequencies. Thus, dipolar properties may be studied at high frequencies, bulk properties at intermediate frequencies and surface properties at low frequencies.

Methods for measuring the impedance can be divided into controlled current and controlled potential [2, 4, 81]. Under controlled potential conditions, the potential of the electrode is sinusoidal at a given frequency with the amplitude being chosen to be sufficiently small to assure that the response of the system can be considered linear. The ratio of the response to the perturbation is the transfer function, or impedance, Z , when considering the response of an AC current to an AC voltage imposition and is defined as $E = IZ$, where E and I are the waveform amplitudes for the potential and the current respectively. Impedance may also be envisaged as the resistance to the flow of an alternating current.

Two different components contribute to impedance: the resistive or real component due to resistors and the reactive or imaginary component from AC circuitry elements, such as capacitors, inductors, etc. Unlike the resistive component, the reactive impedance affects not only the magnitude of the AC wave but also its time-

dependent characteristic, the phase. For example, when an alternating voltage wave is applied to a capacitor, the resulting current waveform will lead the applied voltage by 90°. Due to this reason the introduction of complex notation is convenient. Thus, when a system is perturbed by a sinusoidal potential, varying with time according to

$$E(t) = E_0 \exp(i\omega t)$$

the response can be expressed in terms of

$$I(t) = I_0 \exp(i\omega t - \varphi)$$

where i is the complex number, $E(t)$ and $I(t)$ are the instantaneous values, E_0 and I_0 the peak amplitude of the potential and the current respectively, φ the phase angle difference and ω the angular frequency in radians ($\omega = 2\pi f$).

Introducing the complex notation enables the impedance relationships to be presented as Argand diagrams in both Cartesian and polar co-ordinates (r, φ). The former leads to the Nyquist impedance spectrum, where the real impedance is plotted against the imaginary and the latter to the Bode spectrum, where both the modulus of impedance, r , and the phase angle are plotted as a function of the frequency. In AC impedance the cell is essentially replaced by a suitable model system in which the properties of the interface and the electrolyte are represented by appropriate electrical analogues and the impedance of the cell is then measured over a wide frequency range, usually between 10^4 and 10^{-3} Hz. By comparing the measured results with values calculated from the model system, the suitability of the model and the values of the parameters can be evaluated. In fact, one of the advantages of EIS is that impedance functions frequently display many of the features exhibited by passive electrical circuits. The most important elements employed in equivalent circuits are the resistor R , which represents the resistance that charge carriers encounter in a specific medium, the capacitor C , which represents the accumulation of charged species, and the inductance L , which represents the deposition of surface layers (figure B1.28.8). An analogy, however, is not always feasible, due to the active nature of electrochemical interfaces and the chemical nature of charge transfer processes, as well as the non-ideal electric behaviour of real electrochemical systems. Furthermore, problems can arise in selecting a correct equivalent circuit out of a large number of possibilities, because of the uncertainty connected with the impedance at low frequency and because of the large number of possible combinations of mechanistic reactions that can produce the same impedance shape within error limits. Various methods have been reported to discriminate for the correct equivalent circuit and to obtain values for the elements in the circuit [81, 82 and 83]. Spectra displaying one time constant are simple to interpret and may be easily resolved graphically directly from the spectra; however, more complex methods such as deconvolution and complex nonlinear least-square methods, or a combination of these are required for more complicated spectra. A number of software packages are based on these different types of methods.

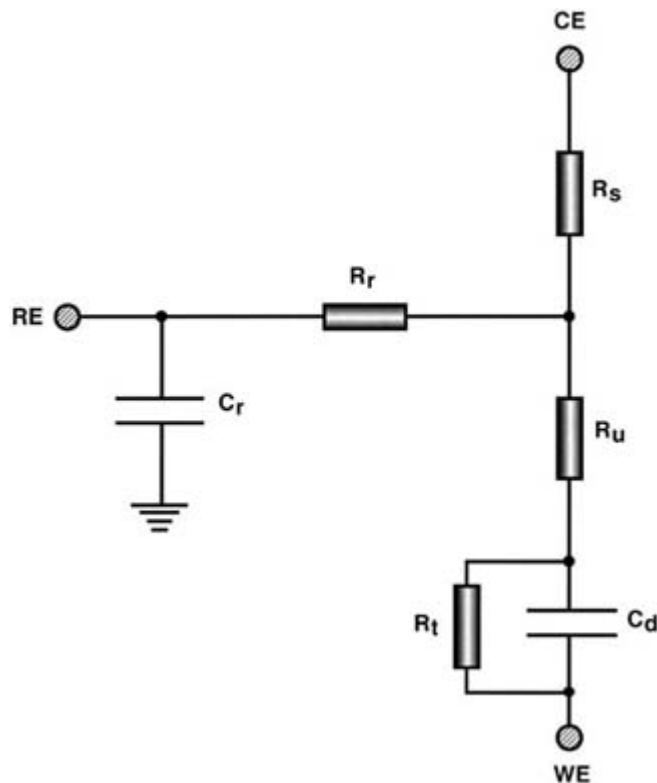


Figure B1.28.8. Equivalent circuit for a three-electrode electrochemical cell. WE, CE and RE represent the working, counter and reference electrodes; R_s is the solution resistance, R_u the uncompensated resistance, R_t the charge-transfer resistance, R_r the resistance of the reference electrode, C_d the double-layer capacitance and C_r the parasitic loss to the ground.

AC impedance spectroscopy is widely employed for the investigation of both solid- and liquid-phase phenomena. In particular, it has developed into a powerful tool in corrosion technology and in the study of porous electrodes for batteries [84, 85, 86 and 87]. Its usage has grown to include applications ranging from fundamental studies of corrosion mechanisms and material properties to very applied studies of quality control and routine corrosion engineering. In corrosion, EIS enables one to obtain instantaneous corrosion-rate information, polarization resistance and information on the kinetics and mechanisms of charge transfer processes such as oxide growth and metal dissolution. The technique is frequently employed in the monitoring of polymer-coated metals to investigate the corrosion protection, the dielectric properties, the onset of defect formation and the processes of coating degradation [84].

B1.28.6 PHOTOELECTROCHEMISTRY

The combination of electrochemistry and photochemistry is a form of dual-activation process. Evidence for a photochemical effect in addition to an electrochemical one is normally seen in the form of photocurrent, which is extra current that flows in the presence of light [88, 89 and 90]. In photoelectrochemistry, light is absorbed into the electrode (typically a semiconductor) and this can induce changes in the electrode's conduction properties, thus altering its electrochemical activity. Alternatively, the light is absorbed in solution by electroactive molecules or their reduced/oxidized products inducing photochemical reactions or modifications of the electrode reaction. In the latter case electrochemical cells (RDE or channel-flow cells) are constructed to allow irradiation of the electrode area with UV/VIS light to excite species involved in electrochemical processes and thus promote further reactions.

Conduction in semiconductors requires that electrons in the valence band be excited into the conduction band either by thermal or photochemical excitation. Upon excitation, an unoccupied vacancy (a hole) is left in the valence band. The hole and the excited electrons can move in response to an applied electric field and so permit the passage of current. Semiconduction can be controlled via doping of small quantities of material, which can be either electron donating or electron accepting, leading to n-type and p-type semiconductors [91, 92, 93 and 94]. In a solution containing a redox couple, electron transfer will occur until the electrochemical potentials of the semiconductor and the solution are equal (figure B1.28.9). The semiconductor will have a net positive or negative charge, which is situated near the surface of the solid, known as the space-charge layer (2–500 nm thickness). The bands may be envisaged as bent: band-bending downwards indicates excess of negative charge at the surface, whereas band-bending upwards indicates an excess of positive holes. Potentiostatic control of a semiconductor can change the energy of the conduction and valence bands with consequent changes in the band bending, leading to the supply and removal of charge carriers within the space layer and enabling electrolysis to occur at the solid/liquid interface. For the n-type semiconductor TiO_2 , for example, the conduction and valence bands are bent upward at the surface, provided that a very negative potential is not applied to the electrode. An applied potential leading to no band-bending—i.e. to an absence of the space-charge layer—is called the flat-band potential of the semiconductor. Upon irradiation with light of an equal or greater energy than the band gap, the photochemically promoted electrons will be swept into the bulk of the material by the electric field present in the space-charge layer, whilst the holes in the valence band will migrate to the surface of the solid (figure B1.28.10). As a result, photo-oxidation processes will be promoted to occur at the solid–liquid interface.

-28-

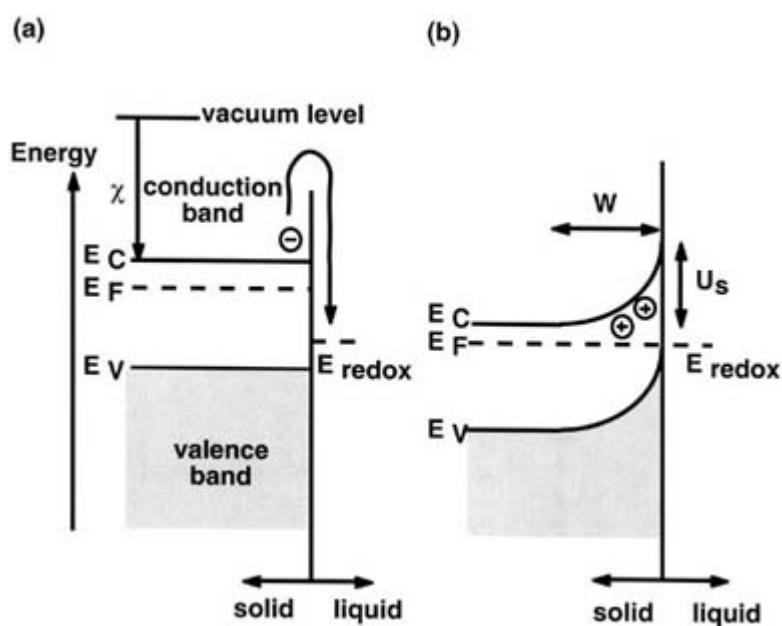


Figure B1.28.9. Energetic situation for an n-type semiconductor (a) before and (b) after contact with an electrolyte solution. The electrochemical potentials of the two systems reach equilibrium by electron exchange at the interface. Transfer of electrons from the semiconductor to the electrolyte leads to a positive space charge layer, W . U_s is the potential drop in the space-charge layer.

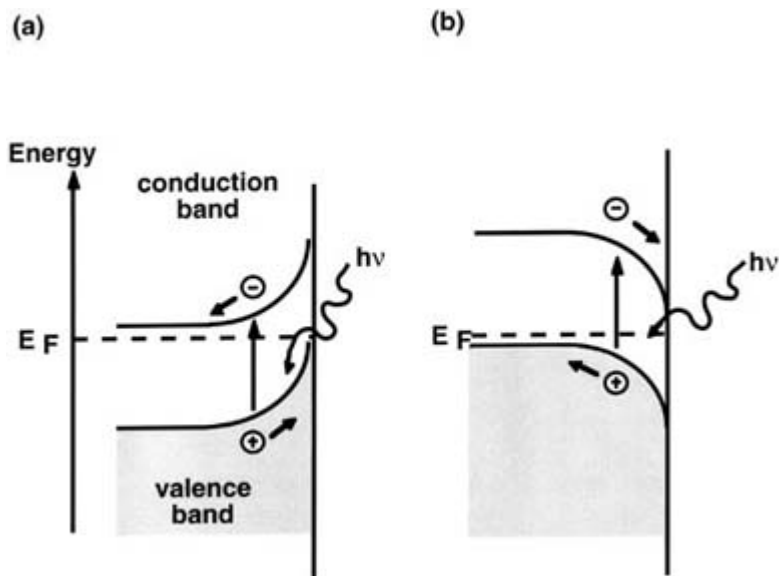


Figure B1.28.10. Schematic representation of an illuminated (a) n-type and (b) p-type semiconductor in the presence of a depletion layer formed at the semiconductor–electrolyte interface.

-29-

In the last 30 or more years, research in the field of photoelectrochemistry and photocatalysis has greatly expanded, with advances being made in the fundamental understanding of the faradic processes that control charge transfer at semiconductor/liquid junctions, and with the development of stable efficient and inexpensive photoelectrochemical cells [95, 96 and 97]. Semiconductor photoelectrochemistry has had an impact in a number of fields. Photocorrosion has been exploited to prepare technologically useful structures such as lenses that are integrated with light-emitting diodes and mated optical fibres. It has also led to the formation of porous Si electrodes, which have received much attention due to their interesting optoelectronic properties. Photoelectrochemical surface preparation of semiconductor materials has been used to clean solids and to evaluate etch pit densities.

Photoelectrochemistry may be used as an *in situ* technique for the characterization of surface films formed on metal electrodes during corrosion. Analysis of the spectra allows the identification of semiconductor surface phases and the characterization of their thickness and electronic properties.

Furthermore, semiconductor powders can be employed for the catalytic generation of useful products such as H₂ and O₂ that can be used for the destruction of pollutants [98, 99 and 100].

B1.28.7 SPECTROELECTROCHEMISTRY

In addition to the mechanism of electrode reactions, readily deduced using voltammetric techniques, the electrochemist seeks knowledge of the chemical composition and properties of electro-generated intermediates and films formed on electrode surfaces. Spectroelectrochemistry allows the simultaneous acquisition of electrochemical and spectroscopic data, which offer additional information to the investigation of a wide range of complex surface and homogeneous processes occurring in electrochemical systems [101, 102, 103 and 104]. Advances made in instrumentation over the past three decades have enabled the adaptation of spectroscopic methods to *in situ* application in an electrochemical cell and the development of new techniques, which have found widespread use in structure characterization of electrode surfaces, in identification of homogeneous phase molecules, and in studies of species adsorbed at the electrode/electrolyte interface.

One of the first *in situ* combined electrochemical/spectroscopic techniques to be investigated employed UV/VIS detection, where solution-phase spectra of organic radicals generated at an electrode could be recorded. Typically, organic intermediates or products possess additional absorption bands not observed in the parent molecule, which can be used to fingerprint the electro-generated species. In addition to spectra, useful information may be obtained by monitoring the absorbance as a function of time during a potential-step experiment. When the potential is stepped from a value where no electrode transfer takes place to one where an electro-generated species is formed, the spectroscopic intensity–time response of the products may be analysed and, from the shape and size, estimates of the lifetime of the electrode intermediates can be extrapolated.

In UV–VIS spectroelectrochemistry, optically transparent electrodes (OTEs) are utilized. A beam of monochromatic UV–VIS light is directed perpendicularly through the OTE, then through the diffusion layer next to the electrode and the bulk solution, before passing out of the electrochemical cell through an exit window and being detected. The beam is attenuated by the presence of absorbing species in the solution, therefore enabling spectral and temporal information about the concentrations of such species in the diffusion layer at the electrode to be obtained. An alternative design is the optically transparent thin layer electrode, in which a minigrad electrode is employed [2, 103, 105].

-30-

Infrared spectroscopy has also been widely employed in electrochemistry [105, 106, and 107]. Spectra aid the identification of reactants, of products and of long-lived intermediates and allow changes in the interfacial solvent to be tracked. A variety of spectral sampling and data acquisition methods have been developed to approach *in situ* detection of species. In external reflection sampling methods, the infrared beam is directed through a polarizer onto the front surface of a highly polished disc-shaped working electrode with high reflectivity in the infrared spectral region, such as Pt, Au or Ag. A special, thin-layer electrochemical cell is used that permits the infrared beam to enter and strike the disc, where it is reflected out of the cell and detected. In contrast, in attenuated total internal reflection sampling methods, the working electrode is a thin film of metal deposited on one surface of an ATR crystal. The metal film must be sufficiently thin to allow penetration of the IR evanescent wave beyond the metal solution interface. The ATR crystal forms the bottom of a chamber that holds the electrolyte solution and the counter and reference electrodes and the crystal is positioned so that the metal film is inside the chamber. This method has not been widely used in electrochemistry partly due to the difficulty in the preparation of the thin metal film working electrodes. Nonetheless, the latter design overcomes molecular transport limitations imposed by external reflection methods, where a thin solution layer of the order 1–5 μm between the front face of the working electrode and the infrared transparent window is required to minimize absorption of infrared radiation by the solvent. In fact, diffusion of species into and out of the thin-layer region is restricted and can lead to reactant depletion or product accumulation.

Luminescence has been used in conjunction with flow cells to detect electro-generated intermediates downstream of the electrode. The technique lends itself especially to the investigation of photoelectrochemical processes, since it can yield information about excited states of reactive species and their lifetimes. It has become an attractive detection method for various organic and inorganic compounds, and highly sensitive assays for several clinically important analytes such as oxalate, NADH, amino acids and various aliphatic and cyclic amines have been developed. It has also found use in microelectrode fundamental studies in low-dielectric-constant organic solvents.

One of the most important advances in electrochemistry in the last decade was the application of STM and AFM to structural problems at the electrified solid/liquid interface [108, 109]. Sonnenfield and Hansma [110] were the first to use STM to study a surface immersed in a liquid, thus extending STM beyond the gas/solid interfaces without a significant loss in resolution. *In situ* local-probe investigations at solid/liquid interfaces can be performed under electrochemical conditions if both phases are electronic and ionic conducting and this

offers a great advantage since the Fermi levels of both substrate and tip can be precisely adjusted independently of each other. This opens the possibility of correlating structural to physical properties, since charge transfer central to electrochemical reactivity occurs within a few atomic diameters of the electrode surface, in the inner Helmholtz plane, and the detailed arrangement of atoms and molecules at this interface strongly controls the corresponding electrochemical reactivity. Since its introduction in electrochemistry, STM investigations have focused on studies of metal electrodes, i.e. Au, Pt, Pd and Rh, their surface charges in the double-layer potential region, and the surface changes caused by the formation of surface oxides. In addition, it has been employed in reconstruction and restructuring studies of metal surfaces, in studies of underpotential deposition of metals, in the investigation of adsorption/desorption processes, as well as in the understanding of processes controlling deposition and corrosion at semiconductor electrodes.

Amongst other spectroscopic techniques which have successfully been employed *in situ* in electrochemical investigations are ESR, which is used to investigate electrochemical processes involving paramagnetic molecules, Raman spectroscopy and ellipsometry.

REFERENCES

- [1] Pletcher D 1991 *A First Course in Electrode Processes* (Romsey: The Electrochemical Consultancy)
- [2] Greef R, Peat R, Peter L M, Pletcher D and Robinson J 1993 *Instrumental Methods in Electrochemistry* (Chichester: Southampton Electrochemistry Group/Ellis Horwood)
- [3] Fisher A C 1996 *Electrode Dynamics (Oxford Chemistry Primers)* (New York: Oxford University Press)
- [4] Bard A J and Faulkner L R 1980 *Electrochemical Methods—Fundamentals and Applications* (New York: Wiley)
- [5] Evans D H 1991 Review of voltammetric methods for the study of electrode reactions *Microelectrodes: Theory and Applications (Nato ASI Series E vol 197)* ed M I Montenegro, M A Queirós and J L Daschbach (Dordrecht: Kluwer)
- [6] Brett C M A and Brett A M O 1998 *Electroanalysis (Oxford Chemistry Primers)* (New York: Oxford University Press)
- [7] Wang J 1994 *Analytical Electrochemistry* (Weinheim: VCH)
- [8] Bond A 1980 *Modern Polarographic Methods in Analytical Chemistry* (New York: Dekker)
- [9] Galus Z 1994 *Fundamentals of Electrochemical Analysis* (Chichester: Ellis Horwood/Polish Scientific Publishers PWN)
- [10] Osteryoung J 1983 Pulse voltammetry *J. Chem. Educ.* **60** 296
- [11] Osteryoung J and Murphy M M 1991 Normal and reverse pulse voltammetry at small electrodes *Microelectrodes: Theory and Applications (Nato ASI Series E vol 197)* ed M I Montenegro, M A Queirós and J L Daschbach (Dordrecht: Kluwer)
- [12] Barker G C and Jenkins I L 1952 *Analyst* **77** 685
- [13] Osteryoung J and O'Dea J J 1986 Square wave voltammetry *Electroanalytical Chemistry* ed A J Bard (New York: Dekker)
- [14] O'Dea J J, Osteryoung J and Osteryoung R A 1981 *Anal. Chem.* **53** 695
- [15] O'Dea J, Wojciechowski M and Osteryoung J 1985 Square wave voltammetry at electrodes having a small dimension *Anal. Chem.* **57** 954
- [16] Osteryoung J 1991 Square-wave and staircase voltammetry at small electrodes *Microelectrodes: Theory and Applications (Nato ASI Series)* ed M I Montenegro, M A Queirós and J L Daschbach (Dordrecht: Kluwer)

- [17] Wang J 1985 *Stripping Analysis: Principles, Instrumentation and Applications* (Weinheim: VCH)
- [18] Wang J 1987 *Anal. Proc.* **24** 325
- [19] Wang J, Tian B, Wang J, Lu J, Olsen C, Yarnitzky C, Olsen K, Hammerstrom D and Bennett W 1999 Stripping analysis into the 21st century: faster, smaller, cheaper, simpler and better *Anal. Chim. Acta* **385** 429
- [20] Economu A, Fielden P R and Packham A J 1994 *Analyst* **119** 279
- [21] Albery W J and Bruckenstein S 1983 Uniformly accessible electrodes *J. Electroanal. Chem.* **144** 105
-

-32-

- [22] Albery W J and Hitchman M L 1971 *Ring-Disc Electrodes* (Oxford: Clarendon)
- [23] Albery W J and Brett C M A 1983 The wall-jet ring disc electrode. 1. Theory *J. Electroanal. Chem.* **148** 201
- [24] Albery W J and Brett C M A 1983 The wall-jet ring disc electrode. 2. Collection efficiency, titration curves and anodic stripping voltammetry *J. Electroanal. Chem.* **148** 201
- [25] Albery W J 1985 The current distribution on a wall-jet electrode *J. Electroanal. Chem.* **191** 1
- [26] Unwin P R and Compton R G 1989 *Comprehensive Chemical Kinetics* vol 29, ed R G Compton (Lausanne: Elsevier)
- [27] Cooper J A and Compton R G 1998 Channel electrodes—a review *Electroanalysis* **10** 141
- [28] Montenegro M I, Queirós M A and Daschbach J L (eds) 1991 *Microelectrodes: Theory and Applications (Nato ASI Series E vol 197)* (Dordrecht: Kluwer)
- [29] Fleischmann M *et al* (eds) 1987 *Ultramicroelectrodes* (Morgantown: Datatech Systems Inc. Science Publishing)
- [30] Denuault G 1996 Microelectrodes *Chemistry and Industry* **18** 678
- [31] Pons S and Fleischmann M 1987 The behavior of microelectrodes *Anal. Chem.* **59** 1391A
- [32] Cassidy J F and Foley M B 1993 Microelectrodes—potential invaders *Chem. Br.* **29** 764
- [33] Forster R J 1994 Microelectrodes—new dimensions in electrochemistry *Chem. Soc. Rev.* **23** 289
- [34] Aoki K 1993 Theory of ultramicroelectrodes *Electroanalysis* **5** 627
- [35] Wightman R M and Wipf D O 1989 Voltammetry at ultramicroelectrodes *Electroanal. Chem.* **15** 267
- [36] Montenegro M I 1994 *Research in Chemical Kinetics* vol 2, ed R G Compton and G Hancock (Amsterdam: Elsevier)
- [37] Oldham K B 1991 Steady-state microelectrode voltammetry as a route to homogeneous kinetics *J. Electroanal. Chem.* **313** 3
- [38] de Carvalho R M, Kubota L T and Rohwedder J J 1999 *Quim. Nova* **22** 591
- [39] Koudelka-Hep M and Van der Wal P D 2000 Microelectrode sensors for biomedical and environmental applications *Electrochim. Acta* **45** 2437
- [40] Armstrong F A and Wilson G S 2000 Recent developments in faradaic bioelectrochemistry *Electrochim. Acta* **45** 2623
- [41] Tanaka K and Tokuda K 1996 *In vivo* electrochemistry with microelectrodes *Experimental Techniques in Bioelectrochemistry* ed V Brabec, D Walz and G Milazzo (Basel: Birkhäuser)
- [42] Stamford J A and Justice J B Jr 1996 Probing brain chemistry *Anal. Chem.* **68** 359A

- [43] Stamford J A, Palij P, Davidson C and Trout S J 1995 Fast cyclic voltammetry: neurotransmitter measurement in 'real time' and 'real space' *Bioelectrochem. Bioenerg.* **38** 289
- [44] Fletcher S 1991 Random assemblies of microdisk electrodes (RAM electrodes) for nucleation studies—a tutorial review *Microelectrodes: Theory and Applications (Nato ASI Series)* ed M I Montenegro, M A Queirós and J L Daschbach (Dordrecht: Kluwer)
-

-33-

- [45] Fletcher S and Horne M D 1999 Random assemblies of microelectrodes (RAM™ electrodes) for electrochemical studies *Electrochem. Commun.* **1** 502
- [46] Fungaro D A and Brett C M A 1999 Microelectrode arrays: application in batch-injection analysis *Anal. Chim. Acta* **385** 257
- [47] Schwarz J, Kaden H and Enseleit U 2000 Voltammetric examinations of ferrocene on microelectrodes and microarrayelectrodes *Electrochem. Commun.* **2** 606
- [48] Morita M, Niwa O and Horiuchi T 1997 Interdigitated array microelectrodes as electrochemical sensors *Electrochim. Acta* **42** 3177–83
- [49] Bard A J, Fan F-R F, Kwak J and Lev O 1989 Scanning electrochemical microscopy—introduction and principles *Anal. Chem.* **61** 132
- [50] Kwak J and Bard A J 1989 Scanning electrochemical microscopy—theory of the feedback mode *Anal. Chem.* **61** 1221
- [51] Kwak J and Bard A J 1989 Scanning electrochemical microscopy—apparatus and two-dimensional scans of conductive and insulating substrates *Anal. Chem.* **61** 1794
- [52] Liu H Y, Fan F-R F, Lin C W and Bard A J 1986 Scanning electrochemical and tunnelling ultramicroelectrode microscope for high-resolution examination of electrode surfaces in solution *J. Am. Chem. Soc.* **108** 3838
- [53] Engstrom R C, Webber M, Wunder D J, Burgess R and Winquist S 1986 Measurements within the diffusion layer using a microelectrode probe *Anal. Chem.* **58** 844
- [54] Winquist, Engstrom R C, Meaney T, Tople R and Wightman R M 1987 Spatiotemporal description of the diffusion layer with a microelectrode probe *Anal. Chem.* **59** 2005
- [55] Bard A J, Fan F-R F and Mirkin M V 1994 Scanning electrochemical microscopy *Electroanalytical Chemistry* vol 18, ed A J Bard (New York: Dekker)
- [56] Mirkin M V 1996 Recent advances in scanning electrochemical microscopy, analytical chemistry *Anal. Chem.* **68** 177A
- [57] Mirkin M V 1999 High resolution studies of heterogeneous processes with the scanning electrochemical microscope *Mikrochim. Acta* **30** 127
- [58] Mirkin M V and Horrocks B R 2000 Electroanalytical measurements using the scanning electrochemical microscope *Anal. Chim. Acta* **406** 119
- [59] Bard A J, Fan F-R F, Pierce D T, Unwin P R, Wipf D O and Zhou F 1991 Chemical imaging of surfaces with the scanning electrochemical microscope *Science* **254** 68
- [60] Unwin P R and Bard A J 1991 Scanning electrochemical microscopy—theory and application of the feedback mode to the measurement of following chemical-reaction rates in electrode processes *J. Phys. Chem.* **95** 7814
- [61] Unwin P R 1998 Dynamic electrochemistry as a quantitative probe of interfacial physicochemical processes *J. Chem. Soc., Faraday Trans.* **94** 3183
- [62] Fulian Q, Fisher A C and Denuault G 1999 Applications of the boundary element method in electrochemistry: scanning electrochemical microscopy *J. Phys. Chem. B* **103** 4387

- [63] Fulian Q, Fisher A C and Denuault G 1999 Applications of the boundary element method in electrochemistry: scanning electrochemical microscopy, part 2 *J. Phys. Chem. B* **103** 4393
-

-34-

- [64] Selzer Y and Manler D 2000 Scanning electrochemical microscopy. Theory of the feedback mode for hemispherical ultramicroelectrodes: steady-state and transient behavior *Anal. Chem.* **72** 2383
- [65] Zhou F and Bard A J 1994 Detection of the electrohydrodimerization intermediate acrylonitrile radical-anion by scanning electrochemical microscopy *J. Am. Chem. Soc.* **116** 393
- [66] Barker A L, Gonsalves M, Macpherson J V, Slevin C J and Unwin P R 1999 Scanning electrochemical microscopy: beyond the solid/liquid interface *Anal. Chim. Acta* **385** 223
- [67] Kranz C, Wittstock G, Wohlschläger H and Schumann W 1997 Imaging of microstructured biochemically active surfaces by means of scanning electrochemical microscope *Electrochim. Acta* **42** 3105
- [68] Bard A J and Fan F-R F 1996 Electrochemical detection of single molecules *Acc. Chem. Res.* **29** 572
- [69] Mason T J and Lorimer J P 1998 *Sonochemistry: Theory, Applications and Uses of Ultrasound in Chemistry* (Chichester: Ellis Horwood)
- [70] Compton R G, Eklund J C and Marken F 1997 Sonoelectrochemical processes: a review *Electroanalysis* **9** 509
- [71] Compton R G, Eklund J C, Marken F, Rebbitt T O, Akkermans R P and Waller D N 1997 Dual activation, coupling ultrasound to electrochemistry—an overview *Electrochim. Acta* **42** 2912
- [72] Akkermans R P, Wu M, Bain C D, Fidel-Suárez M and Compton R G 1998 Electroanalysis of ascorbic acid: a comparative study of laser ablation voltammetry and sonovoltammetry *Electroanalysis* **10** 613
- [73] Compton R G, Eklund J C, Page S D, Mason T J and Walton D J 1996 Voltammetry in the presence of ultrasound: mass transport effects *J. Appl. Electrochem.* **26** 775
- [74] Walton D J, Phull S S, Chyla A, Lorimer J P, Mason T J, Burke L D, Murphy M, Compton R G, Eklund J C and Page S D 1995 Sonovoltammetry at platinum electrodes: surface phenomena and mass transport processes *J. Appl. Electrochem.* **25** 1083
- [75] Birkin P R and SilvaMartinez S 1995 The effect of ultrasound on mass-transport to a microelectrode *J. Chem. Soc., Chem. Commun.* **17** 1807
- [76] Birkin P R and SilvaMartinez S 1997 A study on the effects of ultrasound on electrochemical phenomena *Ultrasonics Sonochemistry* **4** 121
- [77] Reisse J, Francois H, Vandercammen J, Fabre O, Kirschdemesmäker A, Märschalk C and Delplancke J L 1994 *Electrochim. Acta* **39** 37
- [78] Eklund J C, Marken F, Waller D N and Compton R G 1996 Voltammetry in the presence of ultrasound, a novel sono-electrode geometry *Electrochim. Acta* **41** 1541
- [79] Compton R G, Eklund J C, Marken F and Waller D N 1996 Electrode processes at the surfaces of sonotrodes *Electrochim. Acta* **41** 315
- [80] Birkin P R, O'Connor R, Rapple C and SilvaMartinez S 1998 Electrochemical measurement of erosion from individual cavitation events generated from continuous ultrasound *J. Chem. Soc., Faraday Trans.* **94** 3365
- [81] MacDonald J R 1987 *Impedance Spectroscopy* (New York: Wiley)
- [82] Scully J R, Silverman D C and Kendig M W (eds) 1993 *Electrochemical Impedance—Analysis and Interpretation* (Philadelphia: ASTM)
-

- [83] Urquindi-Macdonald M and Egan P C 1997 Validation and extrapolation of electrochemical impedance spectroscopy data *Corr. Rev.* **15** 169
- [84] Amirudin A and Thierry D 1995 Application of electrochemical impedance spectroscopy to the study and degradation of polymer-coated metals *Prog. Org. Coat.* **26** 1
- [85] Grundmeier G, Schmidt W and Stratmann M 2000 Corrosion protection by organic coatings: electrochemical mechanism and novel methods of investigation *Electrochim. Acta* **45** 2515
- [86] Gomes W P and VanMaelkelbergh D 1996 Impedance spectroscopy at semiconductor electrodes: review and recent developments *Electrochim. Acta* **41** 967
- [87] Swarup J and Sharma P C 1996 Electrochemical techniques for the monitoring of corrosion of reinforcement in concrete structures *Bull. Electrochem.* **12** 103
- [88] Gerischer H 1970 *Physical Chemistry* vol 9, ed H Eyring, D Henderson and W Jost (New York: Academic)
- [89] Morrison S R 1977 *The Chemical Physics of Surfaces* (New York: Plenum)
- [90] Pleskov Y V and Gurevich Y Y 1986 *Semiconductor Photoelectrochemistry* (New York: Plenum)
- [91] Stimming U 1986 Photoelectrochemical studies of passive films *Electrochim. Acta* **31** 415
- [92] Gerischer H 1990 On the interpretation of photoelectrochemical experiments with passive layers on metals *Corr. Sci.* **31** 81
- [93] Kamat P V 1993 Photochemistry on non-reactive and reactive (semiconductor) surfaces *Chem. Rev.* **93** 267
- [94] Gerischer H 1990 The impact of semiconductors on the concepts of electrochemistry *Electrochim. Acta* **35** 1677
- [95] Chandra S 1985 *Photoelectrochemical Solar Cells* (New York: Gordon and Breach)
- [96] Tryk D A, Fujishima A and Honda K 2000 Recent topics in photoelectrochemistry: achievements and future prospects *Electrochim. Acta* **45** 2363
- [97] Lewis N S 1996 Photoelectrochemistry—energy conversion using semiconductor electrodes *ECS Interface* Autumn, 28
- [98] Rajeshwar K 1995 Photoelectrochemistry and the environment *J. Appl. Electrochem.* **25** 1067
- [99] Pleskov Y V 1994 Semiconductor photoelectrochemistry for a cleaner environment: utilisation of solar energy *Environmental Oriented Electrochemistry (Studies in Environmental Science 59)* ed C A C Sequeira (Amsterdam: Elsevier)
- [100] Haram S K and Santhanam K S V 1994 Prospective usage of photoelectrochemistry for environmental control *Environmental Oriented Electrochemistry (Studies in Environmental Science 59)* ed C A C Sequeira (Amsterdam: Elsevier)
- [101] Gale R G (ed) 1988 *Spectroelectrochemistry, Theory and Practice* (New York: Plenum)
- [102] Gutiérrez C and Melendres C 1990 *Spectroscopic and Diffraction Techniques in Interfacial Electrochemistry (NATO ASI Series C vol 320)* (Dordrecht: Kluwer)
- [103] Christensen P A and Hamnett A 1994 *Techniques and Mechanisms in Electrochemistry* (Glasgow: Blackie) (an imprint of Chapman and Hall)

- [104] Plieth W, Wilson G S and de la Fe C 1998 Spectroelectrochemistry: a survey of *in situ* spectroscopic techniques *Pure Appl. Chem.* **70** 1395
- [105] Beden B 1995 On the use of '*in situ*' UV-visible and infrared spectroscopic techniques for studying corrosion products and corrosion inhibitors *Mater. Sci. Forum* **192-4** 277
- [106] Korzeniewski C 1997 Infrared spectroscopy in electrochemistry: new methods and connections to UHV surface science *Crit. Rev. Anal. Chem.* **27** 81
- [107] Christensen P and Hamnett A 2000 *In situ* techniques in electrochemistry—ellipsometry and FTIR *Electrochim. Acta* **45** 2443
- [108] Gewirth A A and Niece B K 1997 Electrochemical applications of *in situ* scanning probe microscopy *Chem. Rev.* **97** 1129
- [109] Lillehei P T and Bottomley L A 2000 Scanning probe microscopy *Anal. Chem.* **72** 189R
- [110] Sonnenfield R and Hansma P K 1986 Atomic-resolution microscopy in water *Science* **232** 211
-

FURTHER READING

Pletcher D 1991 *A First Course in Electrode Processes* (Romsey: The Electrochemical Consultancy)

Fisher A C 1996 *Electrode Dynamics (Oxford Chemistry Primers)* (New York: Oxford University Press)

Brett C M A and Brett A M O 1998 *Electroanalysis (Oxford Chemistry Primers)* (New York: Oxford University Press)

These books provide an excellent introduction to the subject.

Bard A J and Faulkner L R 1980 *Electrochemical Methods-Fundamentals and Applications* (New York: Wiley)

For in-depth coverage of electrochemical methods including mathematical derivations.

Montenegro M I, Queirós M A and Daschbach J L 1991 *Microelectrodes: Theory and Applications (Nato ASI Series)* (Dordrecht: Kluwer)

An essential introduction to the field of microelectrodes.

MacDonald J R 1987 *Impedance Spectroscopy* (New York: Wiley)

For in-depth theory of impedance.

Sawyer D T, Sobkowiak A and Roberts J L Jr 1995 *Electrochemistry for Chemists* (New York: Wiley)

Brabec V, Walz D and Milazzo G (eds) 1996 *Experimental Techniques in Bioelectrochemistry* (Basel: Birkhäuser)

These are a good introduction of specific electrochemical techniques for organic chemists and biologists.

Christensen P A and Hamnett A 1994 *Techniques and Mechanisms in Electrochemistry* (Glasgow: Blackie) (an imprint of Chapman and Hall)

This contains a good overview of spectroelectrochemical techniques.

B1.29 High-pressure studies

Malcolm F Nicol

B1.29.1 INTRODUCTION

This chapter introduces the physical chemistry of materials under high pressures. Space limitations permit only a broad-brush introductory survey. High-pressure studies range from designing equipment to generate, to confine and to measure high pressures to spectroscopic studies from 10^5 Hz to beyond 10^{19} Hz at temperatures from below 1 K to 10^5 K and beyond for all sorts of elements, compounds, solutions and mixtures. To say that these are extreme ranges of conditions is an understatement.

To gain a sense of the range of behaviours, consider what happens to one element familiar to every chemist and physicist: oxygen. At ambient temperature, oxygen, O_2 , exists as the canonical odourless, colourless gas of elementary school and as a purple, orange, red, blue or black solid depending on the pressure and the direction from which you look at the crystals. It becomes a metal and, at low temperatures, an antiferromagnet or a superconductor. In the solid phase stable above 10 GPa, O_2 has a strong infrared vibrational absorption band in the stretching region. Then, of course, there is the O_3 isomer which has not been studied at high pressures.

Similarly ‘strange’ things happen to other materials. Above 5 GPa, CO spontaneously polymerizes; the

structure of the product is $(O=C-C=O)_n$. Indeed, almost every carbon compound with unsaturated bonds becomes unstable with respect to reactions that produce saturated compounds at this or slightly higher pressures. Somewhat above 100 GPa, CsI and Xe also are metals. By 100 TPa and 10^5 K—yes, experiments have been done to these conditions—the density of Al exceeds 12 g cm^{-3} , or about five times greater than at ambient pressure. All but the 1s and possibly 2s electrons remain localized on an Al nucleus.

Books are available on many of these subjects. The objective here, therefore, is to introduce several fundamental issues and point to additional information by citing key references and suggesting further reading. We begin by briefly delimiting what we mean by *high pressure*. Then, we discuss how high pressures are achieved and measured before describing the behaviours of a few familiar materials at high pressures.

B1.29.2 WHAT IS PRESSURE?

Almost everyone has a concept of ‘pressure’ from weather reports of the pressure of the atmosphere around us. In this context, ‘high pressure’ is a sign of good weather while very low pressures occur at the ‘eyes’ of cyclones and hurricanes. In elementary discussions of mechanics, hydrostatics of fluids and the gas laws, most scientists learn to compute pressures in static systems as force per unit area, often treated as a scalar quantity. They also learn that unbalanced pressures cause fluids to flow. Winds are the flow of the atmosphere from regions of high to low

pressures. However, high and low pressures in the atmosphere rarely deviate by as much as 10% from the local mean pressure, about 0.1 megapascal at ‘sea level’. The pascal (Pa) is the SI unit of pressure, $1 \text{ Pa} = 1 \text{ N m}^{-2} = 10^{-5} \text{ bar}$. One standard atmosphere is about $1.013 \times 10^5 \text{ Pa}$. Local fluctuations in the pressure of the atmosphere are, however, much smaller than the difference between the average pressure at ‘sea level’ and at the peaks of high mountains. The average pressure near the top of Mount Everest is less than one-quarter of atmospheric pressure at ‘sea level’.

This example of high and low pressure also shows the ambiguities of these terms in science. All these pressures are essentially *constant* in terms of the range of pressures encountered in nature. From negative pressures in solids under tension (e.g., on the wall of flask confining a fluid), pressure in nature increases through the very low-pressure vacuum of interplanetary space (less than 10^{-13} Pa) to well in excess of 10^{20} Pa at the centres of neutron stars! In these terms, *high pressure* and *low pressure* are relative terms with different meanings in different areas of chemistry and physics. Test this by searching an electronic database for ‘high pressure’. A discussion of high-pressure studies, therefore, must decide what pressure is and what high means; just how high is high.

Relationships from thermodynamics provide other views of pressure as a macroscopic state variable. Pressure, temperature, volume and/or composition often are the controllable independent variables used to constrain equilibrium states of chemical or physical systems. For fluids that do not support shears, the pressure, P , at any point in the system is the same in all directions and, when gravity or other accelerations can be neglected, is constant throughout the system. That is, the equilibrium state of the system is subject to a hydrostatic pressure. The fundamental differential equations of thermodynamics:

$$dU = -PdV + TdS$$

$$dA = -PdV - SdT$$

identify P through the Maxwell relations:

$$P = -(\partial U/\partial V)_S = -(\partial A/\partial V)_T.$$

Two other Maxwell relations define the direction systems change to achieve equilibrium:

$$V = \rho^{-1} = (\partial H/\partial P)_S = (\partial G/\partial P)_T.$$

In both mechanical (constant S , minimize H) and thermal (constant T , minimize G) contexts, pressure drives a system to become smaller or denser.

The situation is more complex for rigid media (solids and glasses) and more complex fluids: that is, for most materials. These materials have finite yield strengths, support shears and may be anisotropic. As samples, they usually do not relax to hydrostatic equilibrium during an experiment, even when surrounded by a hydrostatic pressure medium. For these materials, P should be replaced by a stress tensor, σ_{ij} , and the appropriate thermodynamic equations are more complex.

The take-home lesson is that the vast majority of high-pressure studies are on solids or other rigid media and are not done under hydrostatic conditions. The stresses and stress-related properties may vary throughout the sample. Unless the probes are very local and focus on a small region of the sample, measurements are averages over a range of, often uncharacterized, conditions.

As well as macroscopic equations of state relating free energies, enthalpy, entropy, density, composition, temperatures and pressure, high-pressure science also concerns how changes of pressure and other macroscopic constraints affect the microscopic molecular and electronic structures of matter. At low pressures, the chemistry of most materials is described in terms of electrons tightly bound to specific atoms, molecules or ions and relatively weaker intermolecular van der Waals or ionic forces. The itinerant conduction electrons of metals are an exception; they are delocalized throughout the solid. The highly local nature of most electrons reflects the drive to minimize their potential energies.

To a rough approximation, the kinetic and potential energies of electrons in simple systems vary with density as $\rho^{2/3}$ and $-\rho^{1/3}$, respectively. This means that kinetic energy considerations should dominate at very high densities. Localized electrons should, therefore, eventually delocalize at very high pressures, converting ionic and molecular materials to more closely packed extended network structures and to metals at high pressures. Again, of course, the question occurs: how high is very high? Experiments provide the answer, or at least a lower limit. Where the answer is missing, experimenters are driven to try to attain even higher pressures.

Many experiments support the pressure-delocalization principle. Most unsaturated organic molecules including CO and C₂N₂ polymerize at pressures of the order of 10 GPa [1, 2 and 3]. Layered covalent solids like graphite and hexagonal boron nitride (h-BN) transform to dense, three-dimensional network materials, diamond and cubic boron nitride (c-BN) [4, 5 and 6]. Solid oxygen, solid and fluid iodine, fluid hydrogen and nitrogen, xenon, and cesium iodide are examples of materials developing metallic behaviour at pressures of the order of 100 GPa [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21 and 22]. Later sections provide more details about some of these transformations. At intermediate pressures, the energies of atomic and molecular orbitals change with pressure. By measuring electronic spectra at high pressures, differences of these energy changes can be determined for various different orbitals. In some cases, spectral features change by as much as 0.1 eV GPa⁻¹. Drickamer named this phenomenon pressure-tuning spectroscopy and has written extensively about observations for many systems [23, 24 and 25].

B1.29.3 WHAT PRESSURES ARE HIGH?

What then are high pressures? The answer to this question involves the bias of personal experience. I often remark in an off-hand manner that 'In my laboratory, we consider 5 kbar a low pressure'. We have several reasons for setting the low-high boundary around 1 GPa. This and slightly higher pressures can conveniently be achieved by use of commercial autoclaves several litres in size, mechanical compressors and fluid or even compressed-gas pressure media. Many commercial processes run at these and lower pressures. These include the Haber synthesis of ammonia, a method for producing high-density polyethylene, and recently developed methods for producing vaccines by denaturing viruses or for sterilizing (pascalizing) strawberry jam and other foods at ambient temperature which preserve their flavours better than pasteurizing at higher temperatures.

The rates of several chemical reactions accelerate by factors of 10⁴ or more between 0.1 and 100 MPa at ambient temperature, so much interesting chemistry occurs at these lower pressures. At such 'low' pressures, Bridgman [26] even showed how to cook eggs at 'room' temperature.

At ambient temperature, however, few materials remain fluid at pressures much higher than 1 GPa. Fluids also are much more difficult to confine at higher pressures. Absolute pressures have been measured with dead-weight testers only to about 2.5 GPa, that is to the pressure of a solid-solid phase transition in elemental bismuth at 298 K [27]. The importance of non-hydrostatic stresses and changes to the technology of high-pressure studies above 1 GPa suggest the rough dividing line which I have adopted for this essay.

The energies of chemical changes provide a third criterion for defining high pressures. Many unsaturated

organic compounds dimerize or polymerize at high pressures because the products are denser by about $10^{-5} \text{ m}^3 \text{ mol}^{-1}$ ($10 \text{ cm}^3 \text{ mol}^{-1}$). At a pressure of 1 GPa, the corresponding decrease of the energy, enthalpy and free energy is 10 kJ mol^{-1} or relatively modest compared with chemical bond energies. At 10 GPa, for the same difference of molar volume, the energies decrease by 100 kJ mol^{-1} , an amount comparable to bond energies. That is, chemical change can be anticipated at pressures somewhat above 1 GPa.

B1.29.4 HOW ARE HIGH PRESSURES ACHIEVED?

Laboratory high-pressure studies follow many approaches. The pressure may remain constant (so-called static experiments) or be transient (so-called dynamic experiments where shock waves generated by an explosion, impact or laser ablation compress a sample for a few microseconds or shorter times). Many devices have been used for static experiments to about 20 GPa; Jayaraman described many of these in three articles [28, 29 and 30]. Many static high-pressure cells are variants of the piston–cylinder apparatus frequently used to illustrate compression in elementary discussions of the thermodynamics of gases. An external force on a piston free to move within a cylinder applies pressure to a sample that is confined as long as the seal between the piston and cylinder remains leaktight. Bridgman’s anvils [31] represent a different concept. The concept confines the sample between two pistons made of a hard material shaped as truncated cones and a crushable cylindrical gasket. The classical Bridgman design used cemented tungsten carbide anvils and a lava (pyrophyllite) gasket. External force applied to the anvils crush the cylinder, preventing the sample from ‘blowing out’, while applying pressure to the confined sample. Many dual-piston, tetrahedral and cubic cells elaborate on one or both of these concepts of compressing a sample within a confined, sealed volume.

All static studies at pressures beyond 25 GPa are done with diamond-anvil cells conceived independently by Jamieson [32] and by Weir *et al* [33]. In these variants of Bridgman’s design, the anvils are single-crystal gem-quality diamonds, the hardest known material, truncated with small flat faces (culets) usually less than 0.5 mm in diameter. Diamond anvils with 50 μm diameter or smaller culets can generate pressures to about 500 GPa, the highest static laboratory pressures equivalent to the pressure at the centre of the Earth.

Dynamic experiments with conventional (chemical) explosives or projectiles accelerated in gas guns have achieved 1 TPa in favourable cases. Laser-driven shocks have produced higher shock pressures [34], and measurements to 75 TPa have been reported for shock waves generated during underground tests of nuclear explosives (for a recent discussion see [35]). Sample volumes in static experiments range from litres at pressures up to 10 GPa to 0.1 nl at 500 GPa. Samples for commercial dynamic high-pressure production of diamond powder was done on the 100 kl scale. Most samples for shock wave studies are smaller; laser-driven shock wave experiments often use microlitre samples.

In static experiments, the temperatures of samples can be controlled from less than 1 to more than 5000 K and can be measured with reasonable accuracy. For low temperatures, entire pressure vessels are thermostatted by mounting them in cryostats or surrounding them with heaters or furnaces. With these techniques, the temperatures are uniform throughout the sample. The strengths of the materials used to construct the vessel, however, limit the temperatures and pressures that can be achieved by such external heating methods. For higher temperatures, internal heating is used: that is, the sample is heated while the confining pressure vessel is kept at a lower temperature to maintain its mechanical strength. This can be done by surrounding the sample with heating elements mounted inside the pressure vessel around the sample or by irradiating the sample with an intense infrared or visible laser. Internal-heating methods may involve very large temperature gradients, up to 10^8 K m^{-1} ($100 \text{ K } \mu\text{m}^{-1}$), and challenge thermometry. The irreversible nature of the work done when shock waves compress a sample necessarily increases the sample’s temperature; however, the temperature of the shocked state is often impossible to characterize. Indeed, measuring temperatures achieved during dynamic experiments is one of the biggest unsolved problems of high-pressure research.

B1.29.5 HOW ARE HIGH PRESSURES MEASURED?

Absolute pressure measurements by dead-weight piston–cylinder methods have been made only to 2.5 GPa, although Getting recently developed a cell which may extend absolute measurements to 5 GPa [27]. Pressures achieved during shock experiments are computed with the Rankine–Hugoniot equations which assume that the shocked, high-temperature state is one of thermodynamic equilibrium and mass, momentum, and energy are conserved [35]. Several procedures have been developed to relate densities and other properties of states achieved during shock experiments to values of the same properties on ambient or zero-Kelvin isotherms. Pressures in static experiments above 2.5 GPa are often determined by measuring the density of a convenient material like NaCl or Au confined with the sample being studied by x-ray diffraction or by a secondary probe that has been calibrated in terms of x-ray densities. Typical secondary probes include luminescence spectra of ruby (dilute Cr^{3+} in Al_2O_3) [36] or Sm:YAG [37] or Raman spectra of nitrogen [38] or diamond [39]. In each secondary probe, a spectral feature—a narrow emission line or vibrational band—whose energy can be measured precisely changes energy with pressure in an established, ideally simple manner.

B1.29.6 HIGH-PRESSURE FORMS OF FAMILIAR OR USEFUL MATERIALS: DIAMOND, FLUID METALLIC HYDROGEN, METALLIC OXYGEN, IONIC CARBON DIOXIDE, GALLIUM NITRIDE

The most important commercial products of high-pressure science are the extremely hard materials, synthetic diamond [40, 41 and 42] and (c-BN) [43]. At ambient pressure, diamond is less stable than the less dense allotrope, graphite. c-BN also may be less stable than the less dense h-BN isomer with its graphitic structure of rings of alternating B and N atoms. Diamond and c-BN with four equivalent sp^3 bonds per atom are denser than graphite and h-BN with three sp^2 bonds, each shorter than an sp^3 bond and one very long intermolecular van der Waals bond per atom. The volume change, ΔV , for the transformations from graphite to diamond is negative. Thus, the $\Delta(PV) = P\Delta V$ contribution to the change of enthalpy or Gibbs' free energy is negative and becomes even more negative at higher pressures, so the denser forms of each material become more stable at higher pressures. Rearranging the bonds around each atom involves high-energy barriers that separate the low and high density forms (for a historical review of these processes see [44]). Although Yagi and Utsumi [45] showed that the graphite converted to the hexagonal form of diamond at

ambient temperature above 10 GPa, complete conversion and recovery to ambient pressure was not possible unless the product was heated under pressure to more than 1100 K. That is, high temperatures and high pressures are used to overcome the thermodynamic and kinetic barriers to making the desirable dense hard materials that can be recovered metastably. The high barrier also impedes reversion of the recovered diamond and c-BN to the stable graphitic forms at low temperatures. Besides the high-pressure route, diamond can be synthesized at low pressures, even less than 0.1 MPa, by kinetically controlled gas–surface reactions.

Large quantities of both diamond and c-BN are produced by static or shock methods for industrial cutting applications. Most of the synthetic material is finely powdered and can be bound or compacted to make tools. Manufacturing costs for large crystals are too high for the commercial gem market; however, large diamonds of exceptionally high quality are made for special applications: e.g. x-ray monochromators for high-intensity synchrotrons. For cutting steels, c-BN is particularly valuable, because diamond tools tend to react with the steel, forming iron carbide.

Metallic hydrogen has been a holy grail of high-pressure research since Wigner and Huntington suggested that

it might be stable above about 10 GPa [46]. Many claims to the contrary notwithstanding, metallic solid hydrogen has not been found at ambient temperatures to 342 GPa [47]. Weir *et al*, however, found that fluid hydrogen becomes highly conductive under shock compression at lower pressures, 140 GPa, and higher temperatures [11]. The fact that homonuclear diatomic molecular fluids become conductive at lower pressures—and necessarily higher temperatures—than solid phases of the same systems is evident in nitrogen and iodine, and may be a general phenomenon. Further careful experiments must be done to confirm this conjecture. Detailed studies of the conductivities of supercritical Cs and Hg show that the transition from low to high electrical conductivity has neither the characteristics of a thermodynamic phase transition nor a general relationship to the vapour–liquid critical point [48, 49, 50 and 51].

Oxygen is the low-*Z* diatomic which is known to transform to a metal and, at about 1 K, a superconductor at high pressures. The transition pressure is slightly greater than 100 GPa [7]. The conductive phase consists of O₂ molecules; that is, it is not an atomic phase. Optical, infrared and visual spectral, and x-ray diffraction data show that the relevant ϵ phase of oxygen is very anisotropic, and it is reasonable to conjecture that the electrical conductivity also depends upon crystallographic orientation. The other group VI elements also have metallic, superconductive phases at high pressures and low temperatures.

Recent work on the carbon dioxide system shows another unusual high-pressure behaviour. Raman spectra of carbon dioxide show that CO₂ molecules remain the basis of the phases to more than 40 GPa at temperatures below a few hundred Kelvin [52]. These results, however, do not mean that the molecular crystals are the stable phases; indeed, recent studies of the combustion of carbon at high pressures by Yoo *et al* [53] reach another conclusion. They initiated combustion of a mixture of carbon and oxygen at pressures between 7 and 13 GPa by heating the carbon with a Nd:YAG laser, quenching the products to ambient temperature under pressure and recording their Raman spectra. As well as features of unreacted O₂ and CO₂ in some samples, they found vibron bands characteristic of the carbonate ion near 734 and 1079 cm⁻¹, a band assigned to CO⁺⁺ near 2243 cm⁻¹, and several lattice modes between 100 and 350 cm⁻¹. These features, including the shape of the lattice-mode spectrum, closely match the spectra of ϵ , the ionic dimer of NO₂, reported by Agnew *et al* [54], apart from minor shifts because of different pressures, force constants and reduced masses. At higher pressures, heating carbon to ignition was more difficult because diamond formed, which greatly reduced the absorption of the Nd:YAG radiation. The could not be quenched to ambient pressure at ambient temperature; it transformed to CO₂ below 2 GPa. When the was compressed above 15 GPa at ambient temperature, the sharp band near 1100 cm⁻¹ either disappeared or broadened

-7-

above. This change was reversible with pressure and was attributed to either amorphization of CO⁺⁺CO₃⁻ or transitions to larger multimers.

An ionic dimer of carbon dioxide has a critical implication for understanding detonation chemistry of energetic organic molecules. This dimerization provides a reasonable explanation for the kink in shock compression Hugoniot of CO₂ [55], which, if correct, implies that the dimer forms on the time scale of shock loading and detonation. Because water and some nitrogen oxides also become more ionic at these high pressures, interactions and chemical reactivities of these ionic H–C–N–O species will differ from those of the neutral species assumed, in most models, to be the major detonation products over a wide range of high pressures and temperatures [56]. Furthermore, because ionic species are implicated in planetary ‘ices’ like H₂O, CH₄ and NH₃ [57, 58, 59 and 60], the ionic dimer should have some bearing in understanding the internal structure and magnetism of the Jovian planets [61].

Interest in AlN, GaN, InN and their alloys for device applications as blue light-emitting diodes and blue lasers has recently opened up new areas of high-pressure synthesis. Near atmospheric pressure, GaN and InN are unstable with respect to decomposition to the elements far below the temperatures where they might melt. Thus, large boules of these materials typically used to make semiconductor devices cannot be grown from the

melt or annealed at temperatures approaching their melting points. Devices have been grown by heteroepitaxial methods, depositing GaN on Al₂O₃ or SiC substrates, although high defect concentrations because of mismatched lattice constants and thermal expansivities are a serious problem. Grzegory *et al* [62] showed how to overcome this limitation for GaN by growing large crystals from N₂ dissolved in liquid Ga at pressures up to 2 GPa. AlN but not InN also have been grown this way [63]. These relatively slow processes produce excellent materials.

Wallace *et al* explored another approach, metathesis reactions. By igniting a mixture of GaI₃ and Li₃N confined at pressures of the order of 4 GPa, they produced fine-crystalline GaN [63]. The thermodynamic driving force for the process is the very negative enthalpy of forming LiI. With appropriate mixtures and pressures, they produced CrN, Cr₂N, TaN and other nitrides by this metathesis route. The metathesis method, like the direct reactions between elements that Yoo *et al* used to make c-BN, β-Si₃N₄, B₂O₃ and other materials [64, 65], yields more crystalline products at higher confining pressure. Recently, Wallace [66] devised combinations of reagents, chemical diluents and confinement so that GaN and InN crystals of similar quality can be made at as low as ambient pressure.

B1.29.7 SPECTROSCOPY AT HIGH PRESSURES

Almost every modern spectroscopic approach can be used to study matter at high pressures. Early experiments include NMR [67], ESR [68]; vibrational infrared [33] and Raman [69]; electronic absorption, reflection and emission [23, 24 and 25, 70]; x-ray absorption [71] and scattering [72], Mössbauer [73] and gc-ms analysis of products recovered from high-pressure photochemical reactions [74]. The literature contains too many studies to do justice to these fields by describing particular examples in detail, and only some general rules, appropriate to many situations, are given.

The frequencies of vibrational modes usually increase with increasing pressure because the corresponding potential wells become narrower and the force constants increase. In wavenumber terms, these increases range up to the order of 10 cm⁻¹ GPa⁻¹. A notable exception is the stretching mode of an O-H...O hydrogen bond. Other instances of modes whose frequencies decrease with increasing pressure suggest molecular or lattice instabilities that lead to phase transitions at higher pressures. Usually, the transition occurs before the frequency of the mode reaches zero.

-8-

Most electronic valence transitions shift to longer wavelengths at higher pressures: that is, the gap between the highest occupied orbital and lowest unoccupied orbital tends to decrease upon compression. The rates of shift usually are larger (1) for pure materials than for solutes in a solvent and (2) for stronger (more allowed) transitions. However, these correlations are not quantitative, and many transitions shift in the opposite direction. The largest shifts are of the magnitude 0.1 eV GPa⁻¹. Many d-d bands of transition element compounds vary linearly with the fifth power of the metal-ligand distance.

New methods appear regularly. The principal challenges to the ingenuity of the spectroscopist are availability of appropriate radiation sources, absorption or distortion of the radiation by the windows and other components of the high-pressure cells, and small samples. Lasers and synchrotron radiation sources are especially valuable, and use of beryllium gaskets for diamond-anvil cells will open new applications. Impulse-stimulated Brillouin [75], coherent anti-Stokes Raman [76, 77], picosecond kinetics of shocked materials [78], visible circular and x-ray magnetic circular dichroism [79, 80] and x-ray emission [72] are but a few recent spectroscopic developments in static and dynamic high-pressure research.

An especially interesting recent example is Benedetti *et al*'s use of circular dichroism (CD) spectroscopy to detect a pressure-induced change of the configuration at the metal centre of the octahedral chiral Δ- and Λ-tris

[cyclo O,O' 1(R),2(R)-dimethylethylene dithosphato] chromium(III) [79]. The pressure medium was Nujol[®]. To measure the CD spectrum, they had to overcome the birefringence of the strained diamond windows of the high-pressure cell. They did this by recording and averaging spectra of the sample—and of a blank cell filled with Nujol[®]—for each of four 90° rotations of the cell around the axis normal to the windows. The measurements for the blank showed that the baseline obtained by this averaging procedure was close to ideal, although a small further correction was required.

REFERENCES

- [1] Yoo C S and Nicol M F 1986 Chemical reactions and new phases of solid C₂N₂ at high pressure *J. Phys. Chem.* **90** 6726
 - [2] Yoo C S and Nicol M F 1986 Kinetics of the polymerization of solid C₂N₂ near 10 GPa *J. Phys. Chem.* **90** 6732
 - [3] Katz A I, Schiferl D and Mills R L 1981 New phases and chemical reactions in solid carbon monoxide under pressure *J. Phys. Chem.* **88** 3176
 - [4] Johnson Q and Mitchell A C 1972 First x-ray diffraction evidence for a phase transition during shock-wave compression *Phys. Rev. Lett.* **29** 1369
 - [5] Riter J R Jr 1973 Shock-induced graphite to wurtzite phase transformation in boron nitride and implications for stacking graphitic boron nitride *J. Chem. Phys.* **59** 1538
 - [6] Bundy F P 1980 The P, T phase and reaction diagram for elemental carbon, 1979 *J. Geophys. Res.* **85** 6930
 - [7] Shimizu K, Eremets M I, Suhara K and Amaya K 1998 Oxygen under high pressure—temperature dependence of electrical resistivity *Koatsuryoku no Kagaku to Gijutsu (Proc. Int. Conf. AIRAPT-16 and HPCJ-38 on High Pressure Science and Technology, 1997)* vol 7, p 1040
-
- 9-
- [8] Reichlin R, Schiferl D, Martin S, Vanderborgh C and Mills R L 1985 Optical studies of nitrogen to 130 GPa *Phys. Rev. Lett.* **55** 1464
 - [9] Goettel K A, Eggert J H and Silvera I F 1989 Optical evidence for the metallization of xenon at 132(5) GPa *Phys. Rev. Lett.* **62** 665
 - [10] Reichlin R, Ross M, Martin S and Goettel K A 1986 Metallization of cesium iodide *Phys. Rev. Lett.* **56** 2858
 - [11] Weir S T, Mitchell A C and Nellis W J 1996 Metallization of fluid molecular hydrogen at 140 GPa (1.4 Mbar) *Phys. Rev. Lett.* **76** 1860
 - [12] Hemley R J and Mao H K 1990 Critical behavior in the hydrogen insulator–metal transition *Science* **249** 391
 - [13] Lorenzana H E, Silvera I F and Goettel K A 1990 Order parameter and a critical point on the megabar-pressure hydrogen-A phase line *Phys. Rev. Lett.* **65** 1901
 - [14] Takemura K, Minomura S, Shimomura O, Fujii Y and Axe J D 1982 Structural aspects of solid iodine associated with metallization and molecular dissociation under high pressure *Phys. Rev. B* **26** 998
 - [15] Buontempo U, Degiorgi E and Postorino P 1998 Towards the metallization transition in liquid I₂: a spectroscopic study *Nuovo. Cimento. D* **20** 573
 - [16] Ross M 1968 Shock compression of argon and xenon. IV. Conversion of xenon to a metal-like state *Phys. Rev.* **171** 777

- [17] Goettel K A, Eggert J H, Silvera I F and Moss W C 1989 Optical evidence for the metallization of xenon at 132(5) GPa *Phys. Rev. Lett.* **62** 665
- [18] Reichlin R, Brister K E, McMahan A K, Ross M, Martin S, Vohra Y K and Ruoff A L 1989 Evidence for the insulator–metal transition in xenon from optical, x-ray, and band-structure studies to 170 GPa *Phys. Rev. Lett.* **26** 669
- [19] Nellis W J, Holmes N C, Mitchell A C and Van Thiel M 1984 Phase transition in fluid nitrogen at high densities and temperatures *Phys. Rev. Lett.* **53** 1661
- [20] Hamilton D C, Mitchell A C and Nellis W J 1986 Electrical conductivity measurements in shock compressed liquid nitrogen *Shock Waves in Condensed Matter (Proc. 4th Am. Phys. Soc. Top. Conf.)* p 473
- [21] Eremets M I, Shimizu K, Kobayashi T and Amaya K 1998 Metallic CsI at pressures of up to 220 gigapascals *Science* **281** 1333
- [22] Shimizu K, Suhara K, Ikumo M, Eremets M I and Amaya K 1998 Superconductivity in oxygen *Nature* **393** 767
- [23] Drickamer H G and Frank C W 1973 *Electronic Transitions and the High-Pressure Chemistry and Physics of Solids* (London: Chapman and Hall)
- [24] Drickamer H G 1986 Pressure tuning spectroscopy *Acc. Chem. Res.* **19** 329
- [25] Drickamer H G 1990 Forty years of pressure tuning spectroscopy *Ann. Rev. Mater. Sci.* **20** 1
- [26] Bridgman P W 1914 The coagulation of albumin by pressure *Proc. Am. Acad. Arts Sci.* **49** 627
- [27] Heydemann P L M 1997 The Bi I–II transition pressure measured with a dead-weight piston gauge *J. Appl. Phys.* **38** 2640
Getting I 1998 New determination of the bismuth I–II equilibrium pressure—a proposed modification to the practical pressure scale *Metrologica* **35** 119
- [28] Jayaraman A 1983 Diamond anvil cell and high-pressure physical investigations *Rev. Mod. Phys.* **55** 65

- [29] Jayaraman A 1984 The diamond-anvil high-pressure cell *Sci. Am.* **250** 54
- [30] Jayaraman A 1986 Ultrahigh pressures *Rev. Sci. Instrum.* **57** 1013
- [31] Bridgman P W 1941 Explorations towards the limit of utilizable pressures *J. Appl. Phys.* **12** 461
- [32] Jamieson J C, Lawson A W and Nachtreib N D 1959 New device for obtaining X-ray diffraction patterns from substances exposed to high pressures *Rev. Sci. Instrum.* **30** 1016
- [33] Weir C E, Lippencott E R, Van Valkenberg A and Bunting E N 1959 Infrared studies in the 1–15 micron region to 30,000 atmospheres *J. Res. Natl Bur. Stds A* **63** 55
- [34] Da Silva L B *et al* 1997 Absolute equation of state measurements on shocked liquid deuterium up to 200 GPa (2 Mbar) *Phys. Rev. Lett.* **78** 483
- [35] Trunin R F 1998 *Shock Compression of Condensed Matter* (Cambridge: Cambridge University Press)
- [36] Mao H K, Bell P M, Shaner J W and Steinberg D J 1978 Specific volume measurements of Cu, Mo, Pd, and Ag, and calibration of the ruby R_1 fluorescence pressure gauge from 0.06 to 1 Mbar *J. Appl. Phys.* **49** 3276
- [37] Zhao Y, Barvosa-Carter W, Theiss S D, Mitha S, Aziz M J and Schiferl D 1998 Pressure measurement at high temperature using ten Sm:YAG fluorescence peaks *J. Appl. Phys.* **84** 4049
- [38] Schmidt S C, Schiferl D, Zinn A S, Ragan D D and Moore D S 1991 Calibration of the nitrogen vibron pressure scale for use at high temperatures and pressures *J. Appl. Phys.* **69** 2793

- [39] Schiferl D, Nicol M, Zaug J M, Sharma S K, Cooney T F, Wang S-Y, Anthony T R and Fleischer J F 1997 The diamond $^{13}\text{C}/^{12}\text{C}$ isotope Raman pressure sensor system for high-temperature/pressure diamond-anvil cells with aqueous and other chemically reactive samples *J. Appl. Phys.* **82** 3256
- [40] Bundy F P, Hall H T, Strong H M and Wentorf R H 1955 Man-made diamonds *Nature* **176** 51
- [41] DeCarli P S and Jamieson J C 1961 Formation of diamond by explosive shock *Science* **133** 1821
- [42] DuPont 1965 Synthetic diamonds *UK Patent* 1115648
- [43] Wentorf R H 1957 Cubic form of boron nitride *J. Chem. Phys.* **26** 956
- [44] Hazen R M 1993 *The New Alchemists. Breaking Through The Barriers of High Pressure* (New York: Times Books)
- [45] Yagi T and Utsumi W 1993 Direct conversion of graphite into hexagonal diamond under high pressure *New Funct. Mater.* C 99
- [46] Wigner E and Huntington H B 1965 *J. Chem. Phys.* **3** 764
- [47] Narayana C, Luo H, Orloff J and Ruoff A L 1998 Solid hydrogen at 342 GPa: no evidence for an alkali metal *Nature* **393** 46
- [48] Hensel F and Franck E U 1966 Electric conductivity and density of supercritical, gaseous mercury at high pressures *Ber. Bunsenges. Phys. Chem.* **70** 1154
- [49] Hensel F and Franck E U 1968 Metal–nonmetal transition in dense mercury vapor *Rev. Mod. Phys.* **40** 697
- [50] Renkert H, Hensel F and Franck E U 1969 Metal–nonmetal transition in dense cesium vapor *Phys. Lett. A* **30** 494
-

-11-

- [51] Renkert H, Hensel F and Franck E U 1971 Electrical conductivity of liquid and gaseous cesium to 2000 deg. and 1000 bars *Ber. Bunsenges. Phys. Chem.* **75** 507
- [52] Olijnyk H and Jephcoat A P 1998 Vibrational studies on CO_2 up to 40 GPa by Raman spectroscopy at room temperature *Phys. Rev. B* **57** 879
- [53] Yoo C S, Cynn H and Nicol M, Carbon combustion at high pressures and temperatures: evidence for CO^+CO_3^- , an ionic dimer of carbon dioxide, in preparation
- [54] Agnew S F, Swanson B I, Jones L H, Mills R L and Schiferl D 1983 Chemistry of nitrogen oxide (N_2O_4) at high pressure: observation of a reversible transformation between molecular and ionic crystalline forms *J. Phys. Chem.* **87** 5065
- [55] Mitchell A C and Nellis W J 1982 Equation of state and electrical conductivity of water and ammonia shocked to the 100 GPa (1 Mbar) pressure range *J. Chem. Phys.* **76** 6273
- [56] See, Ree F and Vanthiel M 1986 Modeling explosive behavior *Energy Technol. Rev.* UCRL-52000-86-3, 41
- [57] Holzapfel W B and Franck E U 1966 Conductance and ion dissociation of water up to 1000° and 100 kilobars *Ber. Bunsenges. Phys. Chem.* **70** 1105
- [58] Nellis W J, Hamilton D C, Holmes N C, Radosky H B, Ree F H, Mitchell A C and Nicol M 1988 The nature of the interior of Uranus based on studies of planetary ices at high dynamic pressure *Science* **240** 779
- [59] Ancilotto F, Chiarotti G L, Scandolo S and Tosatti E 1997 Dissociation of methane into hydrocarbons at extreme (planetary) pressure and temperature *Science* **275** 1288
- [60] Ross M 1981 The ice layer in Uranus and Neptune. Diamonds in the sky? *Nature* **292** 435

- [61] Hubbard W B 1984 *Planetary Interiors* (New York: Van Nostrand-Reinhold)
- [62] Grzegory I, Jun J, Bockowski M, Krukowski S, Wroblewski M, Lucznik B and Porowski S 1995 III–V nitrides—thermodynamics and crystal growth at high N₂ pressure *J. Phys. Chem. Solids* **56** 639
- [63] Wallace C H, Rao L, Kim S-H, Heath J R, Nicol M and Kaner R B 1998 Solid-state metathesis reactions under pressure: a rapid route to crystalline gallium nitride *Appl. Phys. Lett.* **72** 596
- [64] Yoo C S, Akella J and Nicol M 1996 Chemistry at high pressures and temperatures; *in-situ* synthesis and characterization of β -Si₃N₄ by DAC x-ray/laser-heating studies *Advanced Materials '96* ed M Akaishi *et al* (Tsukuba: National Institute for Research in Inorganic Materials) p 175
- [65] Yoo C S, Akella J, Nicol M and Cynn H 1997 Direct elementary synthesis of hexagonal and cubic boron nitrides at high pressures and temperatures *Phys. Rev. B* **56** 140
- [66] Wallace C H 1998 The rapid solid-state synthesis of group III and transition metal nitrides at ambient and high pressures *PhD Dissertation* University of California, Los Angeles
- [67] Doverspike M A, Liu S B, Ennis P, Johnson T, Conradi M S, Luszczynski K and Norberg R E 1986 NMR in high-pressure phases of solid ammonia and ammonia-d₃ *Phys. Rev. B* **33** 14
- [68] Johansen C R, Nelson H M and Gardner J H 1968 Method for measuring magnetization to high pressures *J. Appl. Phys.* **39** 2152
-

-12-

- [69] Asell J F and Nicol M 1968 Raman spectrum of α -quartz at high pressures *J. Chem. Phys.* **49** 5395
- [70] Sonnenschein R, Syassen K and Otto A 1981 Effect of pressure on the first singlet exciton in crystalline anthracene *J. Chem. Phys.* **74** 4315
- [71] Ingalls R, Crozier E D, Whitmore J E, Seary A J and Tranquada J M 1980 Extended x-ray absorption fine structure of sodium bromide and germanium at high pressure *J. Appl. Phys.* **51** 3158
- [72] Kao C-C, Rueff J P, Strushkin V V, Shu J, Hemley R and Mao H-K 1999 Private communication
- [73] Ingalls R, Drickamer H G and De Pasquali G 1967 Isomer shift of iron-57 in transition metals under pressure *Phys. Rev.* **155** 165
- [74] Yin G Z and Nicol M 1985 Photochemistry of naphthalene in alcohol or alkane solutions at high pressures *J. Phys. Chem.* **89** 1171
- [75] Brown J M, Slutsky L J, Nelson K A and Cheng L T 1988 Velocity of sound and equations of state for methanol and ethanol in a diamond-anvil cell *Science* **241** 65
- [76] Schmidt S C, Moore D S, Schiferl D, Chatelet M, Turner T P, Shaner J W, Shampine D L and Holt W T 1986 Coherent and spontaneous Raman spectroscopy in shocked and unshocked liquids *NATO ASI Series C* **184** 425
- [77] Hare D E, Franken J and Dlott D D 1995 A new method for studying picosecond dynamics of shocked solids: application to crystalline energetic materials *Chem. Phys. Lett.* **244** 224
- [78] Chronister E L and Crowell R A 1991 Time-resolved coherent Raman spectroscopy of low-temperature molecular solids in a high-pressure diamond anvil cell *Chem. Phys. Lett.* **182** 27
- [79] Benedetti M, Biscarini P and Brillante A, The effect of pressure on circular dichroism spectra of chiral transition metal complexes *Physica B* **265** 1
- [80] Baudalet F, Odin S, Giorgetti C, Dartyge E, Itie J P, Polian A, Pizzini S, Fontaine A and Kappler J P 1997 PtFe₃ Invar studied by high pressure magnetic circular dichroism *J. Physique IV* **C7** 441

FURTHER READING

This brief essay could neither deal with all fields of high-pressure science nor any one field in depth. The literature of articles, reviews, books, and conference proceedings for this field is extensive and only a few are cited in the references. Here, we suggest a few books and conference proceedings for the reader interested in exploring the field further. This somewhat arbitrary selection is the author's; it emphasizes recent books and a few classic works.

Overviews and experimental methods

Asay J R and Shahinpour M (eds) 1993 *High-Pressure Shock Compression of Solids* (New York: Springer)

Bridgman P W 1958 *The Physics of High Pressure* (London: G Bell and Sons)

-13-

Brooks H, Birch F, Holton G and Paul W (eds) 1964 *Collected Experimental Papers of PW Bridgman* vol I–VII (Cambridge: Harvard University Press)

Chéret R 1992 *Detonation of Condensed Explosives* (New York: Springer)

Drickamer H G and Frank C W 1973 *Electronic Transitions and the High-Pressure Chemistry and Physics of Solids* (London: Chapman and Hall)

Eremets M I 1996 *High Pressure Experimental Methods* (New York: Oxford University Press)

Graham R A 1993 *Solids under High-Pressure Shock Compression* (New York: Springer)

Horie Y and Sawaoka A B 1993 *Shock Compression Chemistry of Materials* (Tokyo: KTK Scientific)

Sawaoka A B (ed) 1993 *Shock Waves in Materials Science* (New York: Springer)

Trunin R F 1998 *Shock Compression of Condensed Materials* (Cambridge: Cambridge University Press)

Phase diagrams

Liu L-G and Bassett W A 1986 *Elements, Oxides, Silicates* (New York: Oxford University Press)

Young D A 1991 *Phase Diagrams of the Elements* (Los Angeles: University of California Press)

Conference proceedings

Hocheimer H D and Eters R D 1991 *Frontiers of High-Pressure Research* (New York: Plenum)

Polian A, Loubeyre P and Boccara N (eds) 1989 *Simple Molecular Systems at Very High Density* (New York: Plenum)

Winter R and Jonas J (eds) 1993 *High Pressure Chemistry, Biochemistry and Materials Science* (Dordrecht: Kluwer)

Proceedings of the annual conference of the European High Pressure Research Group, the most recent of which is:

Isaacs N S (ed) 1998 *High Pressure Food Science, Bioscience and Chemistry (Spec. Publ. vol 222)* (Cambridge, UK: Royal Society of Chemistry)

Proceedings of the biannual conference of AIRAPT, the International Association for the Advancement of High Pressure Research and Technology, the most recent of which is:

Nakahara M (ed) 1998 *Koatsuryoku no Kagaku to Gijutsu (Proc. Int. Conf. AIRAPT-16 and HPCJ-38 on High Pressure Science and Technology, 1997)* vol 7 (Kyoto: Japan Society of High Pressure Science and Technology)

Proceedings of the biannual Conference of the American Physical Society Topical Group on Shock Compression Science, the most recent of which is:

Schmidt S C, Dandekar D P and Forbes J W (eds) 1998 *Shock Compression of Condensed Matter, 1997 (AIP Conf. Proc. vol 429)* (College

-14-

Park, MD: American Institute of Physics)

Proceedings of the US–Japan Seminars on High Pressure-Temperature Research, the most recent (fifth) of which is:

Manghnani M H and Yagi T (eds) 1998 *Properties of Earth and Planetary Materials* (Washington, DC: American Geophysical Union)

B 2.1 Ultrafast spectroscopy

Warren F Beck

B2.1.1 INTRODUCTION

The development of the millisecond and microsecond flash photolysis experiments by George Porter and co-workers [1, 2] in the 1950s marks the true birth of time-resolved spectroscopy. Porter's work, which provided for the first time a way to capture the absorption spectrum of a short-lived kinetic intermediate in a photochemical reaction, helped to start a new era in physical chemistry, one that was focused on the mechanism and dynamics of chemical reactions. Owing to the subsequent development of mode-locked laser sources, beginning with the picosecond ruby and neodymium–glass lasers in the 1960s, the sub-picosecond passively mode-locked dye laser in the late 1970s and, most recently, the femtosecond self-mode-locked Ti–sapphire laser in the early part of this decade, the time resolution for spectroscopic measurements has advanced three orders of magnitude, from the 10 ps to the 10 fs regime [3]. It is now possible to conduct a wide variety of spectroscopies with ultrashort laser pulses of photons selectable over the entire spectral range from the x-ray region [4, 5] to the terahertz or far-infrared (IR) region [6, 7, 8, 9 and 10]. A variety of robust methods have been developed to probe the time evolution of populations and coherences. The shortest time scale that is now routinely accessible is comparable to or shorter than the period of molecular vibrations, the fundamental time scale of chemistry.

This chapter focuses on the primary experimental methods of ultrafast spectroscopy, as discussed in terms of studies on intramolecular dynamics in the condensed phase or in proteins. Ultrafast spectroscopy generally denotes spectroscopy that exploits the time resolution obtainable with mode-locked laser sources. The ultrafast regime encompasses electronic and vibrational energy transfer, charge transfer and structural dynamics involving isomerization and the breaking of bonds. In many cases, these processes can be optically triggered so that the time course can be studied with a delayed probing or gating pulse. The initial and delayed pulses are derived from a pulse train emitted by a single mode-locked light source; the ultrashort timing of the experiment is usually derived from the distance of flight of the optical pulses using a technique that is reminiscent of interferometry. After a discussion of the current ideas in producing and characterizing tunable ultrashort optical pulses for spectroscopy, the chapter discusses methods for time-resolved fluorescence spectroscopy, pump–probe methods for time-resolved absorption spectroscopy and multipulse photon-echo techniques for the measurement of coherence. The chapter closes with a brief discussion of the use of phase-controlled, multiple-pulse sequences for advanced, highly selective spectroscopies and for the control of chemical dynamics.

B2.1.2 FEMTOSECOND LIGHT SOURCES

The development of ultrafast spectroscopy has paralleled progress in the technical aspects of pulse formation [11]. Because mode-locked laser sources are tunable only with difficulty, until recently the most heavily studied physical and chemical systems were those that had strong electronic absorption spectra in the neighbourhood of conveniently produced wavelengths.

As one important example, the introduction of the prism-controlled, colliding-pulse, mode-locked (CPM) dye laser [12, 13] led almost immediately to developments in measurement technique with pulses of less than 100

fs duration; Shank and co-workers used an amplified CPM laser [14] in their work with 6 fs pulses in 1987 [15]. Until recently, the pulses used in those experiments were the shortest optical pulses characterized. The transition-state spectroscopy of Zewail and Bernstein [16, 17, 18, 19, 20, 21 and 22] exploited an amplified CPM laser after frequency doubling and/or continuum generation. The chemical systems that were most easily studied, however, were those that could be stimulated either by the 620 nm output of the CPM directly or after frequency doubling to 310 nm. In addition, the CPM laser and its contemporary, more tunable alternative, the pulse-compressed, synchronously pumped dye laser [11], were tools that could be effectively used only by researchers with extensive backgrounds in lasers and optics.

These limitations have recently been eliminated using *solid-state* sources of femtosecond pulses. Most of the femtosecond dye laser technology that was in wide use in the late 1980s [11] has been rendered obsolete by three technical developments: the self-mode-locked Ti-sapphire oscillator [23, 24, 25, 26 and 27], the chirped-pulse, solid-state amplifier (CPA) [28, 29, 30 and 31], and the non-collinearly pumped optical parametric amplifier (OPA) [32, 33 and 34]. Moreover, although a number of investigators still construct home-built systems with narrowly chosen capabilities, it is now possible to obtain versatile, nearly state-of-the-art apparatus of the type described below from commercial sources. Just as home-built NMR spectrometers capable of multidimensional or solid-state spectroscopies were still being home built in the late 1970s and now are almost exclusively based on commercially prepared apparatus, it is reasonable to expect that ultrafast spectroscopy in the next decade will be conducted almost exclusively with apparatus from commercial sources based around entirely solid-state systems.

Figure B2.1.1 depicts an instrument that takes advantage of many of the most recent technical developments. The best strategy for generating wavelength-tunable ultrashort laser pulses for time-resolved spectroscopy involves use of an OPA as the only wavelength-tunable element. This approach is organized around the principle that extremely stable, fixed wavelength, high-energy pulse trains can now be generated using an amplified Ti-sapphire-based system. The chief advantage of instruments like the one shown in figure B2.1.1 is that experimental demands for specific operating wavelengths are met by adjustment of the *last* device in the pulse-forming chain, the OPA. One can expect such a design to be considerably more robust and user-friendly than systems based on wavelength tunable oscillators, which demand manipulation of *every* device in the instrument in response to tuning to a new wavelength.

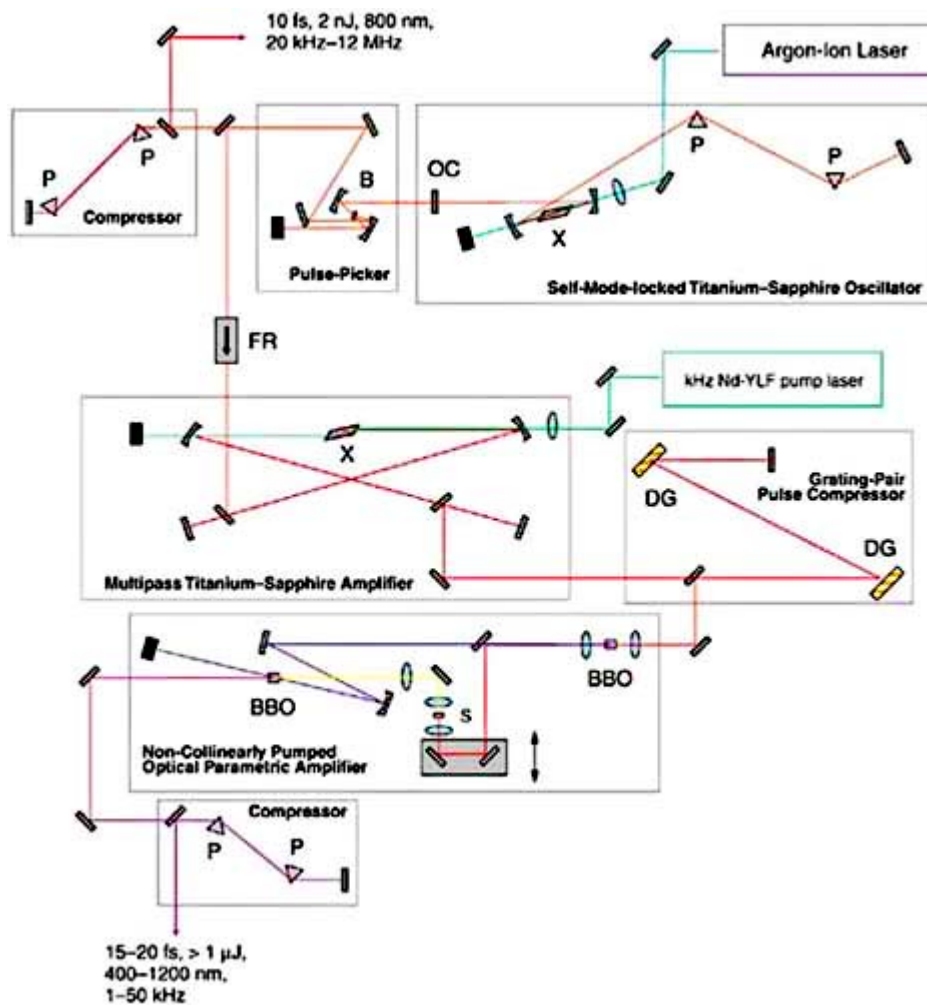


Figure B2.1.1 Femtosecond light source based on an amplified titanium–sapphire laser and an optical parametric amplifier. Symbols used: P, Brewster dispersing prism; X, titanium–sapphire crystal; OC, output coupler; B, acousto-optic pulse selector (Bragg cell); FR, Faraday rotator and polarizer assembly; DG, diffraction grating; BBO, β -barium borate nonlinear crystal.

B2.1.2.1 OSCILLATORS

The most commonly used femtosecond oscillator at this point is the self-mode-locked Ti–sapphire laser [23, 24, 25, 26 and 27], shown in figure B2.1.1 which can *routinely* produce pulses of light with durations adjustable over the 10–150 fs range. The wavelength of the Ti–sapphire oscillator can be tuned over the 700–1100 nm range using an intracavity slit or birefringent filter, providing pulse durations that are essentially limited by the bandwidth of the filtering element. (It should be emphasized, however, that tuning an oscillator of this type is not as routinely done as is tuning an OPA.) The pulse energy that can be directly obtained from the oscillator is typically limited to the 2 nJ/pulse regime, but the oscillator emits pulses at a high repetition rate, typically 75–100 MHz, depending on the cavity dimensions.

The Ti–sapphire oscillator is extremely useful as a stand-alone source of femtosecond pulses in the near-IR region of the spectrum. Some ultrafast experiments, especially of the pump–probe variety (see below), can be conducted with pulses obtained directly from the oscillator or after pulse selection at a lower repetition rate. Far-IR (terahertz) radiation is usually generated using a semiconductor (usually GaAs) substrate and focused Ti–sapphire oscillator pulses [7]. If somewhat higher-energy pulses are required for an experiment, the Ti–sapphire oscillator can be *cavity dumped* by an intracavity acousto-optical device known as a *Bragg cell*,

producing perhaps 50 nJ pulses but at a reduced repetition rate, typically between 200 kHz and 1 MHz [35]. The energy available from a cavity-dumped Ti-sapphire laser is intense enough to generate a continuum-like source in a single-mode optical fibre. Wiersma and co-workers [36] generated 5 fs pulses by compressing these continuum pulses with a sequence of gratings and prisms. Comparable oscillators have been described that exploit other types of solid-state gain media. Although Ti-sapphire crystals are widely used because the absorption spectrum overlaps favourably with the output spectra of argon-ion and frequency-doubled Nd:YVO₄ continuous-wave lasers, Cr-LISAF may be favoured in the future as a gain medium for femtosecond oscillators because it can be pumped by continuous-wave GaAs diode lasers [37, 38].

B2.1.2.2 AMPLIFICATION

Although many useful femtosecond spectroscopic experiments on condensed-phase targets can be easily performed with low-energy pulses, in the 100 pJ to 1 nJ regime, higher-energy pulses are required if wavelength tunability is desired. A femtosecond continuum [39] can be generated in water or sapphire if the pulse energy is higher than 200 nJ; OPA sources require even higher energies, in excess of 1 μJ/pulse. Amplification of Ti-sapphire oscillators is at this point routinely performed, with excellent commercial systems readily available of the regenerative amplifier type [28, 30, 31, 40, 41], and there are simple multipass amplifier designs [42, 43] that are easily constructed in the laboratory.

Pulses are selected for amplification from the oscillator's 75–100 MHz pulse train at a much lower repetition rate, ranging in published designs from 10 Hz to 250 kHz, either by a Pockels cell or a Bragg cell (as shown in [figure B2.1.1](#)). The selected pulse train is amplified in Ti-sapphire gain media using a method known as *chirped-pulse amplification* [28, 29, 30 and 31]. In this scheme, oscillator pulses are stretched temporally well into the picosecond regime prior to amplification so that the damage threshold for the gain crystal is not exceeded [44]. If the amplifier is designed to operate in the >10kHz regime, like the one depicted in [figure B2.1.1](#) a stretcher may not be required. A grating-pair pulse compressor [45] is used to compress the pulse back nearly to its original duration after it emerges from the amplifier. Regenerative amplifiers capable of producing 75–150 fs pulses are the most common systems in use [30, 40], but recently a multipass ring amplifier has been described that produces 20 fs pulses [43]. The multipass amplifier depicted in [figure B2.1.1](#) is a non-ring design that permits a more facile input and extraction of the amplified pulse.

The most common commercially prepared amplifier systems are pumped by frequency-doubled Nd:YAG or Nd:YLF lasers at a 1–5 kHz repetition rate; a continuously pumped amplifier that operates typically in the 250 kHz regime has been described and implemented commercially [40]. The average power of all of the commonly used types of Ti-sapphire amplifier systems approaches 1 W, so the energy per pulse required for an experiment effectively determines the repetition rate.

B2.1.2.3 OPTICAL PARAMETRIC AMPLIFICATION

Perhaps the ultimate femtosecond light source, the OPA exploits a nonlinear parametric process to amplify a portion of

a femtosecond continuum [32, 33 and 34, 46, 47 and 48]. In most designs, a portion of the output of a regenerative or multipass Ti-sapphire amplifier is frequency doubled in a nonlinear crystal to prepare an intense *pump* pulse. Less than 1 μJ/pulse of the amplifier's output is reserved to seed a single-filament continuum [47] in a thin sapphire crystal. A second nonlinear crystal is used as the gain medium for the parametric process, which splits an input pump photon at frequency ω_3 into two output photons, a signal photon at frequency ω_1 and an idler photon at frequency ω_2 , with energy conserved ($\omega_3 = \omega_1 + \omega_2$). The parametric process is greatly enhanced by the presence of *seed* light at either ω_1 or ω_2 , which is supplied by the continuum [46]. The apparatus can be adjusted to select a certain range of frequencies from the continuum

for amplification in the nonlinear crystal, allowing the production of wavelength-tunable output pulses derived either from the signal or idler with adjustable pulse durations. At this point, β -barium borate (BBO) is the material of choice for the nonlinear crystal.

The OPA should not be confused with an optical parametric oscillator (OPO), a resonant-cavity parametric device that is synchronously pumped by a femtosecond, mode-locked oscillator. 14 fs pulses, tunable over much of the visible regime, have been obtained by Hache and co-workers [49, 50] with a BBO OPO pumped by a self-mode-locked Ti-sapphire oscillator.

Shortly after the development of high-energy/pulse Ti-sapphire regenerative amplifier systems, a number of investigators reported progress in using OPAs in producing tunable sources of very short pulses. Wilson and co-workers [46] showed early on that an *experimentally* useful source for femtosecond spectroscopy with <50 fs pulses was obtained through the use of continuum seeding of a type I nonlinear OPA crystal, which was pumped by the *fundamental* output of an amplified Ti-sapphire laser. The main problem with the early systems was inherent to the physics of *collinearly pumped* parametric amplification: the signal, idler and pump frequencies have different group velocities (see below) in the nonlinear crystal, which limits the amount and frequency bandwidth of the parametric gain. In other language, the phase-matching condition for the collinearly pumped OPA works only over a small bandwidth, which tends to limit the pulse duration to fairly long (100 fs) pulses, and tuning of the OPA to different signal wavelengths requires reoptimization of the crystal's orientation. The design advanced by Wilson's group takes advantage of the smaller mismatch in group velocities in the near-IR part of the spectrum; other designs employing pumping with the second harmonic of the Ti-sapphire laser provide direct access to visible signal pulses but with significantly longer durations (150 fs) [48].

Very recently, Hache and co-workers [49] found that a *non-collinear* pumping of an OPO crystal produces a phase-matching condition that is *independent of signal wavelength* over a very broad bandwidth. This discovery makes it possible to obtain very high parametric gain in an OPA with a single pass through the crystal and adjustable signal bandwidths. The result is a source of light with wide tunability and adjustable pulse durations, the ultimate femtosecond light source. The design for the OPA depicted in [figure B2.1.1](#) is an adaptation of that recently described by Riedle and co-workers [32]. 400 nm light obtained by frequency doubling the output of a regenerative Ti-sapphire amplifier is overlapped at an angle of 3.7° with the femtosecond continuum light. This angle produces in BBO the wavelength-independent phase-matching condition noted by Hache and co-workers. Riedle and co-workers demonstrated that signal pulses of 15–20 fs duration could be obtained from this system, with tunability over most of the 400–800 nm visible range. Using a similar approach, but with some changes in the details of producing the continuum seed light that were intended to produce as broad a signal bandwidth as possible, De Silvestri and co-workers [34] subsequently showed that signal pulses as short as 7.2 fs in the visible regime could be produced. In contemporaneous work, Kobayashi and co-workers [33] obtained sub-10 fs pulses from the visible signal *and* near-IR idler from a comparable apparatus. These latter two results have nearly matched the legendary performance of the pulse-compressed CPM dye laser of Shank and co-workers [15].

A surprising aspect of the recent work on non-collinearly pumped OPA systems is that it appears that fairly long pump pulses, in the 150 fs regime, are to be preferred over shorter pulses if sub-10 fs signal pulses are desired. Mature commercial regenerative amplifier designs operating over the 1–250 kHz repetition-rate range are already capable of producing these pulses. Thus, the experimental motivation for developing amplifier systems capable of producing very short pulses *directly* has vanished for *spectroscopic* applications; however, there are a number of important applications for short, high-energy pulses, such as driving the formation of femtosecond pulses of x-rays [5]. High-energy physics applications for short, amplified pulses have been reviewed recently by Mourou and co-workers [51].

B2.1.2.4 PULSE COMPRESSION

Owing to its wide spectral bandwidth, a short optical pulse is distorted temporally by passing through the beamsplitters, lenses, filters, etc, that are required for an ultrafast spectroscopic experiment. The distortion arises from the dispersion of the speed of light in a medium as a function of wavelength, or group-velocity dispersion (GVD). The sweep of frequencies observed at a given spatial position is known as chirp (in analogy to the sound of a frequency-swept audio pulse) or group-delay dispersion (GDD). This phenomenon is usually described in terms of a Taylor series expansion of the phase ϕ of the optical pulse around the centre frequency, ω_0 [11, 52]. The first derivative term represents the group delay, $t(\omega) = d\phi/d\omega$. The quadratic term, corresponding to the *linear* sweep in the group delay with respect to wavelength, can be corrected by an optical delay line with *negative* GDD (alternatively speaking, *anomalous dispersion*) [53], constructed from a pair of diffraction gratings [45], a pair of dispersing prisms with Brewster angled surfaces [54] or, less often, by a Gires–Tournois interferometer (GTI) [55, 56]. The amount of negative GDD is determined by the distance of separation between the two prisms or gratings or between the two reflective films in the interferometer.

In most femtosecond experiments, a double-passed pair of prisms is inserted into the beam prior to its reaching the measurement apparatus. In [figure B2.1.1](#) a pair of prisms is depicted after the OPA and after the pulse-picker (for use in oscillator-only experiments). This practice allows one to *precompensate* for the GDD imparted to the beam by the optics in the measurement apparatus so that the pulses are as short as possible when they arrive at the sample's position. The prisms are typically manipulated by translating one or both of the prisms normal to the base so as to increase or decrease the amount of prism glass traversed; this permits a small amount of positive GDD to be added or subtracted. The pulse duration at the sample's position is usually determined with an intensity autocorrelation measurement, performed by replacing the sample with a nonlinear crystal, such as potassium dihydrogen phosphate (KDP) or BBO. An even more significant problem associated with GDD is that the formation of a short pulse in an oscillator requires many round-trips through the gain medium, so it is routine to place a pair of prisms or a GTI in the oscillator's cavity to repair the damage suffered by the pulse on each trip. The design of Murnane and co-workers [25] depicted for the oscillator in [figure B2.1.1](#) exemplifies this practice, which was first established in the final design for the CPM dye laser by Valdmanis and Fork [13]. In the future, oscillators may not require prism pairs; it is now possible to fabricate *chirped mirrors* that employ multiple layers of dielectric coatings to provide for a precise compensation of quadratic and cubic phase distortions that arise from the gain medium of a femtosecond oscillator [26]. In fact, a commercial design employing a chirped-mirror, Ti-sapphire oscillator and a solid-state continuous-wave Nd-YVO₄ pump laser in a single, compact, sealed (no user controls!) enclosure has just become available. Such a source would be expected to be unusually robust.

The shortest optical pulses actually used so far (1998) in ultrafast spectroscopic experiments were obtained by Shank and co-workers from an amplified CPM laser [57]. In these extraordinary experiments, a sequence of a pair of prisms

and a pair of gratings was employed. The reason for this additional complexity was that the cubic-term phase distortion imparted by the prism pair is of opposite sign of that imparted by the grating pair. As a result, it was possible to null simultaneously the quadratic and cubic dispersion terms at the position of the sample [15]. Most optical materials exhibit *normal* dispersion; the index of refraction increases monotonically with respect to frequency. Some materials, however, exhibit an anomalous dispersion regime in the near-IR part of the spectrum. Hochstrasser and co-workers exploited the finding that the mid-IR substrates CaF₂ and BaF₂, used, for instance, for beamsplitters and windows, imparts GDD to a transmitted beam that is of the opposite sign of that imparted by Ge, which was used as a long-pass filter [58]. Fused silica optical fibres exhibit anomalous dispersion at wavelengths above 1.3 μm [59, 60]. The implication of this phenomenon is that positively chirped IR pulses can be compressed by being propagated in a fibre. Further, short IR pulses can be propagated in a fibre for arbitrarily long distances without suffering a change in pulse duration [61, 62].

B2.1.3 FEMTOSECOND TIME-RESOLVED SPECTROSCOPY

In lieu of electronic timing mechanisms, a form of interferometry is usually employed in ultrafast spectroscopy to obtain time resolution on the picosecond or shorter time scales. The simplest ultrafast measurement apparatus is that shown in figure B2.1.2 an *autocorrelator*. This instrument is used in various forms to measure the duration of an ultrashort laser pulse [63]. A pulse of light is first split into equal portions by a beamsplitter; the two parts are then recombined after they travel in the two arms of an interferometer (here, derived from the Michelson design) so that they are focused onto a thin (100 μm or less in thickness) nonlinear crystal. When the two pulses overlap *temporally and spatially*, the second harmonic is emitted along the direction that conserves momentum, between the direction of the two emerging fundamental pulses [64]. By moving one of the retroreflectors in the interferometer along the beam path towards or away from the beamsplitter, one of the pulses is made to scan *temporally* through the other pulse in the nonlinear crystal. Thus, the ultrafast timing of the experiment is provided by a simple measurement of displacement and knowledge of the speed of light. Reproducible displacements of the order of 1 μm , corresponding to a time-of-flight displacement of 6.67 fs in the interferometer, are routinely performed with computer-controlled linear actuators.

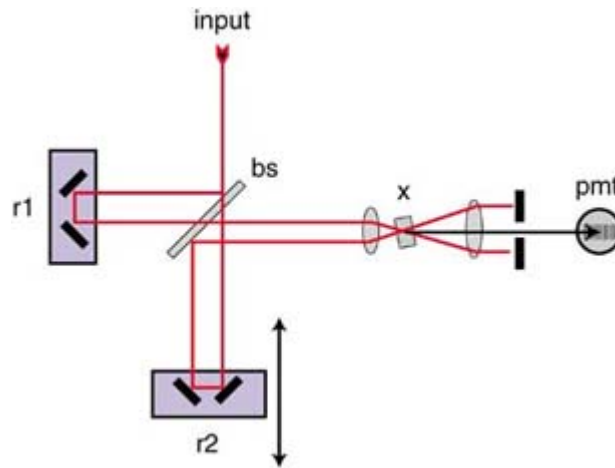


Figure B2.1.2 Modified Michelson interferometer for non-collinear intensity autocorrelation. Symbols used: r1, r2, retroreflecting mirror pair mounted on a translation stage; bs, beamsplitter; x, nonlinear crystal; pmt, photomultiplier tube.

The intensity of this *upconverted* light detected by a photodetector is described by the equation $I(\tau) = \int_{-\infty}^{+\infty} I(t - \tau)I(t) dt$, where $I(t)$ is the intensity of the pulse as a function of time t at a given point in the nonlinear crystal, and τ is the shift in time of the variably delayed or *gating* pulse, as controlled by one of the retroreflectors in the interferometer. The photodetector integrates the upconverted intensity for a given delay τ ; an example of the signal obtained by scanning τ is shown in figure B2.1.3(a) with the input pulse train provided by a self-mode-locked Ti-sapphire laser. The symmetrical shape of the autocorrelation trace arises from the convolution of the input $I(t)$ pulse shapes. The true shape of $I(t)$ and the dependence of the phase on wavelength can be obtained in a slightly more elaborate experiment called frequency-resolved optical gating (FROG) [65, 66, 67 and 68], a form of which can be performed by recording the autocorrelation traces as a function of wavelength by dispersing the upconverted light in a spectrometer placed before the photodetector [69].

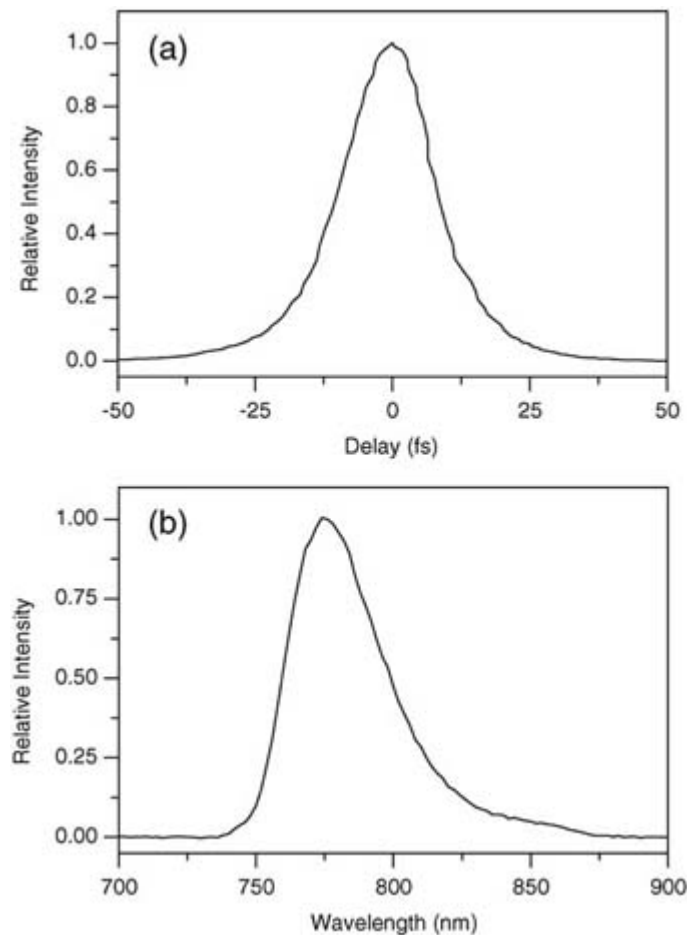


Figure B2.1.3 Output of a self-mode-locked titanium–sapphire oscillator: (a) non-collinear intensity autocorrelation signal, obtained with a 100 μm β -barium borate nonlinear crystal; (b) intensity spectrum.

The spectrum of the femtosecond pulse provides some information on whether the input pulse is chirped, however, causing the temporal width of $I(t)$ to be broader than expected from the Heisenberg indeterminacy relationship.

-9-

The full width at half maximum of the autocorrelation signal, 21 fs, corresponds to a pulse width of 13.5 fs if a sech^2 shape for the $I(t)$ function is assumed. The corresponding output spectrum shown in [figure B2.1.3\(b\)](#) exhibits a width at half maximum of approximately 700 cm^{-1} . The time–bandwidth product $\Delta \tau \Delta \nu$ is close to 0.3. This result implies that the pulse was compressed nearly to the Heisenberg indeterminacy (or Fourier transform) limit [53] by the double-passed prism pair placed in the beam path prior to the autocorrelator.

The intensity autocorrelation measurement is comparable to all of the *spectroscopic* experiments discussed in the sections that follow because it exploits the use of a variably delayed, *gating* pulse in the measurement. In the autocorrelation experiment, the gating pulse is just a replica of the time-fixed pulse. In the spectroscopic experiments, the gating pulse is used to interrogate the populations and coherences established by the time-fixed pulse.

B2.1.3.1 FLUORESCENCE UPCONVERSION SPECTROSCOPY

Time-resolved fluorescence is perhaps the most direct experiment in the ultrafast spectroscopist’s palette. Because only one laser pulse interacts with the sample, the method is essentially free of the problems with field-matter time orderings that arise in all of the subsequently discussed multipulse methods. The signal

detected is usually directly proportional to the population in the resonantly prepared excited state alone. In systems that exhibit photochemistry, for instance, the time evolution of the fluorescence provides a direct view of the decay of the photochemically active state.

Fluorescence spectroscopy can be performed in the ultrafast regime with a nonlinear crystal and a short gating pulse in an experiment known as *fluorescence upconversion* [70, 71]. The interferometer shown in [figure B2.1.2](#) is modified so that an incident pulse is split into two pulses: a weaker pulse that is used to excite a sample and a stronger, gating pulse that is overlapped spatially in a nonlinear crystal with the fluorescence that is collected from the sample. Sum-frequency generation in the nonlinear crystal produces output photons with frequency ω_3 from the input gate photons of frequency ω_1 and fluorescence photons of frequency ω_2 ; since $\omega_3 = \omega_1 + \omega_2$, the output photons are *upconverted*, or transferred to a higher frequency by the gate pulse. The intensity of the beam of output photons is proportional to the product of the intensity of the gating and fluorescence photons; the gate photons slice out only those fluorescence photons that are temporally overlapped with the short gating pulse in the nonlinear crystal. This permits the time course of the fluorescence intensity at a particular frequency ω_2 to be mapped out by scanning the time delay for the gate pulse.

The frequency ω_2 that is detected in the fluorescence can be largely selected by adjusting the phase-matching condition for the nonlinear crystal. A double monochromator (or, alternatively, a prism and a single-stage monochromator) is used to discriminate between the upconverted fluorescence and the background interference from the second harmonic of the gating light. Even though the nonlinear crystal is angle-tuned to optimize the intensity of the upconverted fluorescence at the wavelength chosen by the monochromator, the strong gate pulse generates a significant second-harmonic signal that can often be as strong or stronger than the gated fluorescence. If the gate pulse is short, with a concomitantly large spectral bandwidth, the second-harmonic background may significantly overlap the fluorescence spectrum of the sample under study. In many respects this problem is comparable to that encountered in discriminating Rayleigh scattering from Raman scattering in the low Raman frequency regime. In Raman spectroscopy, however, the vibrational line shapes are usually much narrower than any fluorescence background; in fluorescence upconversion spectroscopy, the second-harmonic background from the gate pulse is often as broad as the fluorescence signal, making things comparably more difficult [71].

In [figure B2.1.4](#) a design for a fluorescence upconversion spectrometer is depicted that is based on one discussed by Jimenez and Fleming [71]. In the most versatile instrument, an OPA would be used to generate the wavelength-tunable excitation pulse, while a portion of the direct output of the amplified Ti-sapphire laser that pumps the OPA would be reserved for use as a strong gate pulse. This practice has several advantages: the gating and excitation pulses are implicitly time-synchronized and the excitation pulse can be well removed in wavelength from the region of the fluorescence photon, in order to minimize the second-harmonic gate background mentioned above. The experimental set-up depicted in [figure B2.1.4](#) employs two off-axis parabolic reflectors to collect and collimate the fluorescence emitted by the sample and then to focus it onto the nonlinear crystal. Other designs employ a single elliptical reflector to perform both tasks. The design shown here may permit low-temperature fluorescence studies to be executed, however, since the crystal and sample can be well separated on the optical table. Jimenez and Fleming [71] note that very few femtosecond fluorescence upconversion experiments have so far been attempted at low temperature.

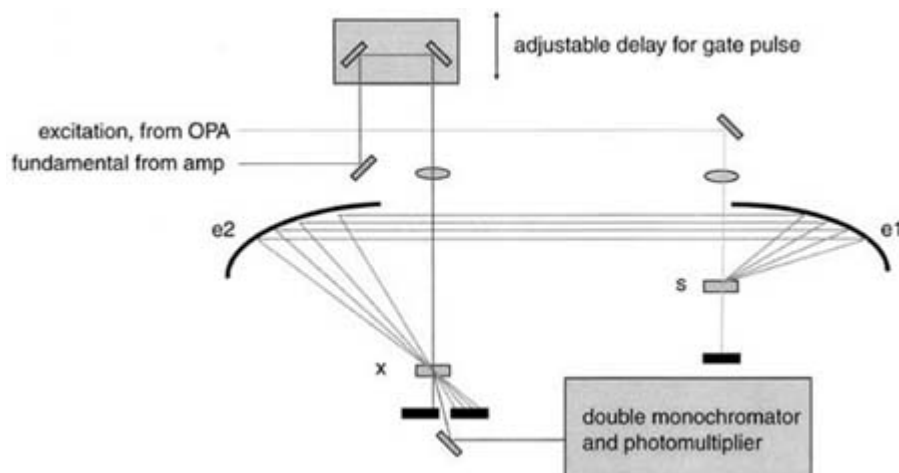


Figure B2.1.4 Fluorescence upconversion spectrometer based on the use of off-axis elliptical reflectors for the collection and focusing of fluorescence. Symbols used: e1, e2, off-axis elliptical reflectors; s, sample; x, nonlinear crystal. (After Jimenez and Fleming [71].)

Over the last few years, the time resolution attainable in fluorescence upconversion has reached the sub-100 fs regime. The main problems associated with pushing the time resolution down further are sensitivity and GDD. The problem of collecting enough fluorescence photons and imaging them onto the nonlinear crystal, where detection occurs, is at odds with obtaining pulse-width-limited time resolution. Large elliptical or parabolic reflectors enhance the number of photons collected, but the time-of-flight dispersion increases with the surface area used on the reflector, so in many implementations the reflectors are masked. Lenses were used for collection and focusing in early experiments, as in the design discussed by Barbara and co-workers [70], but the time resolution obtainable is typically limited to the 250 fs regime owing to the GDD suffered by the fluorescence photons in being transmitted through the material used to make the lens. In principle, spherical confocal reflective optics might be used without time-of-flight dispersion. Even so, the nonlinear crystal itself imparts GDD and distorts the time response.

The instrument response function (IRF) for the fluorescence upconversion experiment, then, cannot be shorter than the intensity *cross-correlation* function, which can be obtained using an instrument like that shown in figure B2.1.4

-11-

in which the excitation and gate pulses used in the upconversion experiment are injected separately into the two arms of the interferometer. In the simplest case, the actual time response of the fluorescence resembles the *integral* of the IRF, $S(\tau) = \int_{-\infty}^{+\infty} I(t - \tau)F(t) dt$. The fluorescence signal $F(\tau)$ at the detection frequency ω_2 is, in the absence of other dynamics and rapid relaxation, a step function; the excitation pulse transfers a fraction of the ground-state population to the resonant excited state, which then emits fluorescence. As indicated in the above equation, the upconversion signal $S(\tau)$ responds as the gate pulse $I(t)$ is integrated by the step-response of $F(\tau)$. Of course, if ground-state recovery or energy-transfer processes deplete the resonant excited-state population, then $F(\tau)$ decays accordingly. The upconversion signal $S(\tau)$ then exhibits a shape that is effectively a *convolution* of the IRF and the excited-state population response $F(\tau)$. Since many experiments are intended to *determine* $F(\tau)$ via measurement of $S(\tau)$, one is often faced with *deconvolution* of the IRF from the measured signal in the course of data analysis. This problem has been extensively discussed in the literature [72]; the best solution, is, if possible, to make the IRF much shorter than the dynamics exhibited by $F(\tau)$ so that the measured signal is not significantly distorted.

An important extension to the simplest upconversion experiment at a single detection frequency ω_2 is the practice of measuring *time-resolved fluorescence spectra*, that is, the shape of the fluorescence spectrum

emitted by the sample at a given gate delay τ , $S(\omega_2, \tau)$. In most reported work, the $S(\omega_2, \tau)$ spectrum is built up by obtaining a family of single-wavelength transients, scanned at a given ω_2 as a function of τ . If the instrument provides for computer control of the angle tuning of the nonlinear crystal and the detection monochromator, it would be more efficient, with respect to experimental time, to directly scan ω_2 at a given time delay τ .

Perhaps the best example of a situation requiring knowledge of the time evolution of the fluorescence spectrum is that associated with *dynamic solvation*, the time-dependent reorganization of solvent dipoles in response to a light-induced change in the dipole moment of a dissolved chromophore [73, 74, 75, 76, 77 and 78]. As an example, figure B2.1.5 shows two upconversion traces obtained with the dye phenoxazone dissolved in methanol at room temperature. When the observation wavelength is tuned to the blue edge of the emission spectrum, at 570 nm, the fluorescence is observed to decay with a time constant of several picoseconds. If the observation wavelength is tuned to the red edge of the absorption spectrum at 650 nm, the fluorescence transient exhibits a rise with a similar time constant. These two upconversion transients evidence a blue-to-red shift of the time-resolved fluorescence spectrum owing to dynamic solvation on the picosecond time scale.

-12-

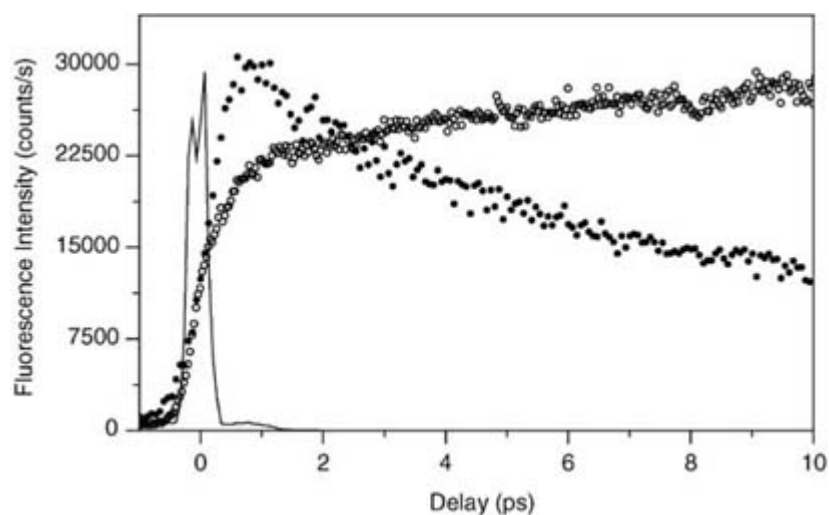


Figure B2.1.5 Fluorescence upconversion traces obtained at two observation wavelengths (full circles, 570 nm; open circles, 650 nm) at room temperature with an oxazine dye, phenoxazone, in methanol solvent. Figure courtesy of Professor S Rosenthal (Vanderbilt University).

The upconversion method can also be applied to the measurement of *anisotropy* [79], which returns information on the time dependence of the orientation of the excited-state transition-dipole moment with respect to time following excitation with a linearly polarized pulse of light. Anisotropy information can be used to study rotational diffusion [80, 81] in liquids, energy transfer between chromophores in proteins [79, 82] and excited-state isomerization [83, 84], to name just three common applications. The method takes advantage of photoselection of those molecules whose transition-dipole moments are aligned with the excitation pulse's plane of polarization to prepare an essentially polarized excited-state orientational distribution initially. The fluorescence that is emitted initially by this distribution is highly polarized; as time progresses, any mechanism that causes rotation of the excited-state transition-dipole moment, such as rotational diffusion, energy transfer and isomerization, will cause depolarization of the fluorescence.

The anisotropy function $r(t) = (I_{\parallel}(t) - I_{\perp}(t))/(I_{\parallel}(t) + 2I_{\perp}(t))$ is determined by two polarized fluorescence transients $I_{\parallel}(t)$ and $I_{\perp}(t)$ observed parallel and perpendicular, respectively, to the plane of polarization of the excitation pulse. In the upconversion experiment, the two measurements are most conveniently made by rotating the plane of polarization of the excitation pulse with respect to the fixed orientation of the input plane

of the nonlinear crystal. Because the photoselected angular distribution for a set of isolated transition dipoles exhibits a $\cos(\theta)$ dependence with respect to the angle θ between the plane of polarization of the excitation pulse and the observation plane, the anisotropy $r(t)$ decays from an initial value of 0.4 as the depolarization proceeds. The time course of the anisotropy is essentially a measure of the time-correlation function $\hat{\mu}_2(t)$ that describes the memory of the initial, photoselected dipole orientation $\hat{\mu}_1(0)$ as correlated with the probed dipole direction $\hat{\mu}_2(t)$ as a function of time [79, 84].

In certain situations involving coherently interacting pairs of transition dipoles, the initial fluorescence anisotropy value is expected to be larger than 0.4. As indicated by the theory described by Wynne and Hochstrasser [85, 86] and by Knox and Gülen [87, 88], the initial anisotropy expected for a pair of coupled dipoles oriented 90° apart, as an example,

-13-

is 0.7, and the decay of the anisotropy from 0.7 to 0.4 is a measure of the time scale for the decay of the electronic coherence between the two states. This theory has been applied to the interpretation of the decay of anisotropy in the analogous anisotropy measurement obtained using pump-probe (see below) stimulated-emission measurements in magnesium tetraphenylporphyrin [89] and in exciton-coupled chromophore pairs in cyanobacterial light-harvesting proteins [90, 91].

B2.1.3.2 PUMP-PROBE SPECTROSCOPY

An interferometric method was first used by Porter and Topp [1, 92] to perform a *time-resolved absorption* experiment with a Q -switched ruby laser in the 1960s. The nonlinear crystal in the autocorrelation apparatus shown in figure B2.1.2 is replaced by an absorbing sample, and then the transmission of the variably delayed pulse of light is measured as a function of the delay τ . This approach is known today as a *pump-probe* experiment; the first pulse to arrive at the sample transfers (*pumps*) molecules to an excited energy level and the delayed pulse *probes* the population (and, possibly, the coherence) so prepared as a function of time.

The pump-probe concept can be extended, of course, to other methods for detection. Zewail and co-workers [16, 18, 19 and 20, 93] have used the probe pulse to drive population from a reactive state to a state that emits fluorescence [94, 95, 96, 97 and 98] or photodissociates, the latter situation allowing the use of mass spectrometry as a sensitive and selective detection method [99, 100].

Pump-probe absorption experiments on the femtosecond time scale generally fall into two effective types, depending on the duration and spectral width of the pump pulse. If the pump spectrum is significantly narrower in width than the electronic absorption line shape, *transient hole-burning spectroscopy* [101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112 and 113] can be performed. The second type of experiment, *dynamic absorption spectroscopy* [57, 114, 115, 116, 117, 118, 119, 120, 121 and 122], can be performed if the pump and probe pulses are short compared to the period of the vibrational modes that are coupled to the electronic transition.

Figure B2.1.6 depicts a standard type of apparatus used for the hole-burning type of time-resolved absorption experiment [112, 113, 123]. A pulse train from an amplified laser is split into two portions. The minor portion is used directly as a source of pump photons; the major portion is used to generate a broad-band probe source derived from a femtosecond continuum. In this application, the continuum is typically generated by focusing a $> 1 \mu\text{J}$ pulse of light into flowing water or ethylene glycol in a cuvette [39]; a continuum with particularly good optical properties can be generated in a thin sapphire crystal [47]. After the pump and probe pulses are overlapped in the sample, the transmitted probe light is dispersed in a monochromator and then detected either by a photodiode or by a multichannel detector, such as a charge-coupled device (CCD). The most common detection scheme involves using a mechanical chopper to modulate the intensity of the pump beam; the pump-induced changes in the transmission of the probe beam are then detected by using a lock-in amplifier.

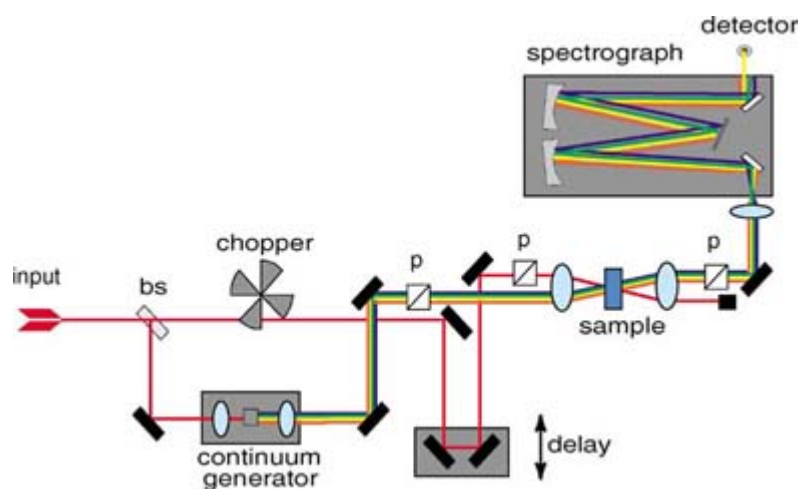


Figure B2.1.6 Femtosecond spectrometer for transient hole-burning spectroscopy with a continuum probe. Symbols used: bs, 10% reflecting beamsplitter; p, polarizer. The continuum generator consists of a focusing lens, a cell containing flowing water or ethylene glycol or, alternatively, a sapphire crystal and a recollimating lens.

As an example, a series of transient hole-burning spectra obtained with a chirp-compensated continuum probe with a light-harvesting protein is shown in [figure B2.1.7](#) [112]. As the probe delay increases, the initially narrow transmission hole increases in width owing to vibrational redistribution and shifts about 500 cm^{-1} to the red owing to dynamic solvation. Analysis of this series of spectra was made in terms of overlapping spectral contributions from ground-state depletion, stimulated emission, and excited-state absorption. The time resolution of the experiment is limited by the width of the pump–probe cross-correlation function, which can be conveniently determined in this case using either FROG or a wavelength-resolved optical Kerr measurement [124].

The main cost of this enhanced time resolution compared to fluorescence upconversion, however, is the aforementioned problem of time ordering of the photons that arrive from the pump and probe pulses. When the probe pulse either precedes or trails the arrival of the pump pulse by a time interval that is significantly longer than the pulse duration, the action of the probe and pump pulses on the populations resident in the various resonant states is unambiguous. When the pump and probe pulses temporally overlap in the sample, however, all possible time orderings of field–molecule interactions contribute to the response and complicate the interpretation. Double-sided Feynman diagrams, which provide a pictorial view of the density matrix’s time evolution under the action of the laser pulses, can be used to determine the various contributions to the sample response [125].

The part of the response arising from a coherent interaction between the temporally overlapped probe and pump pulses is the so-called *coherence spike* [126, 127], which makes its appearance in the zero-delay region in [figure B2.1.7](#) essentially confined to the spectral region coinciding with the pump-pulse spectrum. Accordingly, the overall pump–probe signal’s temporal shape when viewed at a single-probe wavelength is not just the integral of the convolution of the pump and probe pulse temporal shapes. The intensity of the coherence spike is strongly dependent on the duration of the laser pulses employed in the experiment and the time scale of dephasing [127, 128].

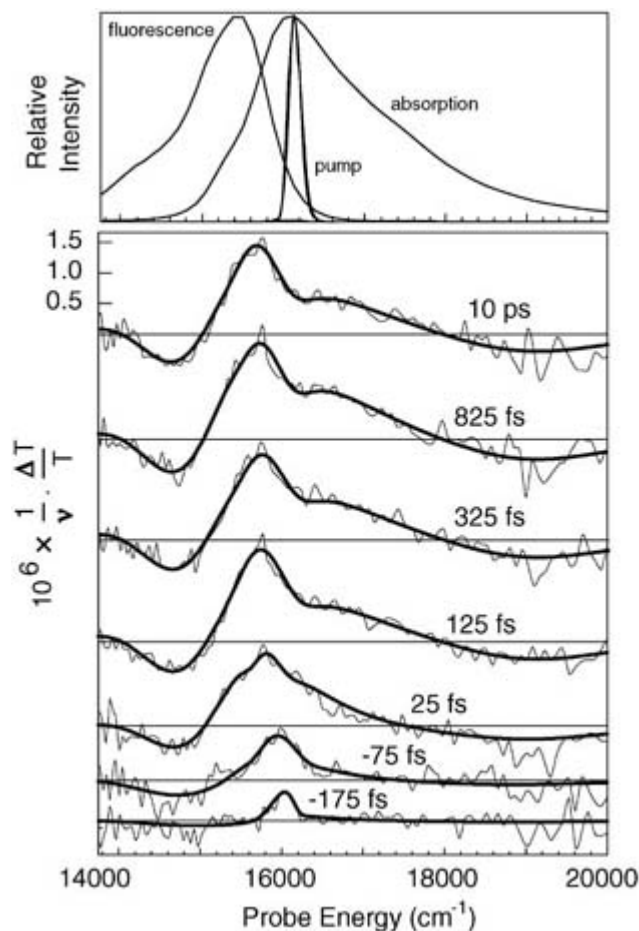


Figure B2.1.7 Transient hole-burned spectra obtained at room temperature with a tetrapyrrole-containing light-harvesting protein subunit, the α subunit of *C*-phycoerythrin. Top: fluorescence and absorption spectra of the sample superimposed with the spectrum of the 80 fs pump pulses used in the experiment, which were obtained from an amplified CPM dye laser operating at 620 nm. Bottom: absorption-difference spectra obtained at a series of probe time delays.

The first *dynamic absorption* studies that afforded a view of a molecular response stimulated by *impulsive* excitation with femtosecond laser pulses were performed by Shank and co-workers in 1988 with 6 fs pulses from a fibre-grating pulse-compressed CPM laser [57]. This experiment might be called a degenerate pump–probe experiment since the 6 fs pulses were used both for pumping and probing; the broad spectral bandwidth had been used previously as a short, chirp-free, continuum-like probe in time-resolved hole-burning spectroscopy [101]. Under the conditions of impulsive excitation, where the pulse duration is short relative to the period of the coupled vibrations, vibrational wavepackets [129, 130] are created on structurally displaced excited-state potential energy surfaces. The wavepackets move back and forth under the forces of the excited-state potential and cause modulation of the stimulated-emission contribution to the pump–probe signal. A second pump–field–molecular interaction causes a stimulated transition of some of the moving excited-state wavepacket back down to the ground state at a position that is displaced from the equilibrium

molecular geometry, causing a ground-state wavepacket motion; that corresponds to the signal detected in conventional resonance Raman spectroscopy [130, 131]. Accordingly, the modulation of the ground-state depletion signal owing to the wavepacket motion on the ground-state surface is termed resonant-impulsive stimulated-Raman scattering (RISRS) [132]. Champion and co-workers [133, 134 and 135] have made extensive use of RISRS in their studies of heam motions in response to ligand photodissociation. Nelson and co-workers [132] have used RISRS to look at collective motions in molecular crystals.

[Figure B2.1.8](#) shows dynamic absorption results obtained with an IR dye in solution. The experiment was conducted with 13 fs pulses from a pulse-picked Ti–sapphire laser and a rapid-scanning pump–probe interferometer [136]. The single-wavelength transient was obtained by dispersing the transmitted probe light in a monochromator and monitoring at a single narrow range of wavelengths. The transient exhibits a modulation signal arising from excited-state vibrational wavepacket motions that is sustained for at least a picosecond. Modulations of this type can be frequency analysed using either Fourier transformation or a linear-prediction, singular-value decomposition (LPSVD) method, as was done in [figure B2.1.8\(b\)](#). The LPSVD method [137] fits the modulation pattern to a series of damped cosinusoids. The representation shown in [figure B2.1.8\(b\)](#) uses the frequencies and damping (dephasing) times to construct a spectral representation that resembles a conventional Raman spectrum. The modulation spectrum shown in [figure B2.1.8\(b\)](#) evidences contributions from several vibrational modes over the 100–1200 cm^{-1} range. The intensity of modulation for a given frequency is observed to depend strongly on the pulse duration; as the pulse duration is made shorter, the frequency window that provides impulsive excitation broadens. This windowing has been described by Lotshaw and McMorro in their discussion of non-resonant optical Kerr effect studies, where impulsive excitation of intramolecular and intermolecular vibrational modes in neat liquids can be studied through the use of orthogonally polarized pump and probe beams and optically heterodyned detection methods [138, 139, 140, 141 and 142].

Excited-state vibrational coherence of the type observed in [figure B2.1.8](#) is potentially a very important tool for the elucidation of excited-state reaction dynamics. The most well known example of this type of work is that of Mathies, Shank and co-workers on the dynamics of rhodopsin [118, 143]. Elsaesser and co-workers [144] used a two-colour dynamic absorption technique to study ultrafast intramolecular proton transfer in a benzotriazole dye in solution. Wynne and Hochstrasser [145] observed vibrational coherence associated with charge transfer in a contact ion-pair in solution. Diffey *et al* [122] used a comparison of excited-state and RISRS wavepacket modulation patterns to study ultrafast charge transfer in a bacteriochlorophyll dimer system isolated from a purple-bacterial light-harvesting chromophore. Vos and co-workers [146, 147 and 148] have observed excited-state vibrational coherence in purple-bacterial reaction centres; Stanley and Boxer observed analogous signals using fluorescence upconversion [149].

So far we have exclusively discussed time-resolved absorption spectroscopy with *visible* femtosecond pulses. It has become recently feasible to perform time-resolved spectroscopy with femtosecond IR pulses. Hochstrasser and co-workers [58, 150, 151, 152, 153, 154, 155, 156 and 157] have worked out methods to employ IR pulses to monitor chemical reactions following electronic excitation by visible pump pulses; these methods were applied in work on the light-initiated charge-transfer reactions that occur in the photosynthetic reaction centre [156, 157] and on the excited-state isomerization of the retinal pigment in bacteriorhodopsin [155]. Walker and co-workers [158] have recently used femtosecond IR spectroscopy to study vibrational dynamics associated with intramolecular charge transfer; these studies are complementary to those performed by Barbara and co-workers [159, 160], in which ground-state RISRS wavepackets were monitored using a dynamic-absorption technique with visible pulses.

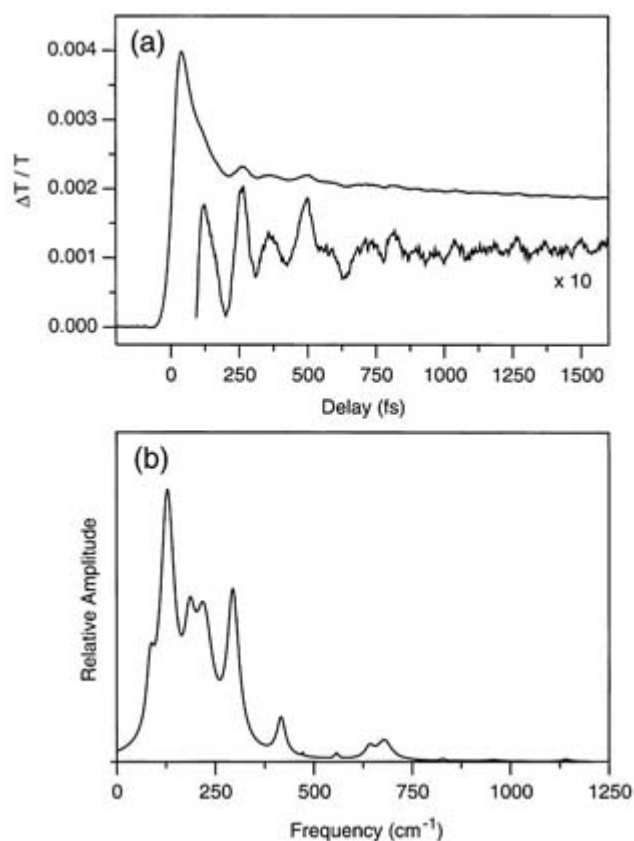


Figure B2.1.8 Dynamic absorption trace obtained with the dye IR144 in methanol, showing oscillations arising from coherent wavepacket motion: (a) transient observed at 775 nm; (b) frequency analysis of the oscillations obtained using a linear prediction, singular-value-decomposition method.

In some extremely innovative recent experiments, Hochstrasser and co-workers [58] have described IR transient hole-burning experiments focused on characterizing inhomogeneous broadening in the amide I transition in several small polypeptides. 180 fs IR pulses centred at 1650 cm^{-1} were generated from a BBO OPA source through the use of difference-frequency mixing of the signal and idler pulses in a AgGaS_2 crystal. Narrower, 1 ps duration IR pulses were produced from the spectrally broad femtosecond IR pulses using a mechanically scannable Fabry–Perot etalon. The results were presented using a novel two-dimensional representation as shown in [figure B2.1.9](#) in which the time-resolved hole-burned spectra were plotted against the centre frequency of the pump spectrum that burned the hole. The results provide information on the extent of delocalization of the amide vibrational wavefunction along the peptide backbone.

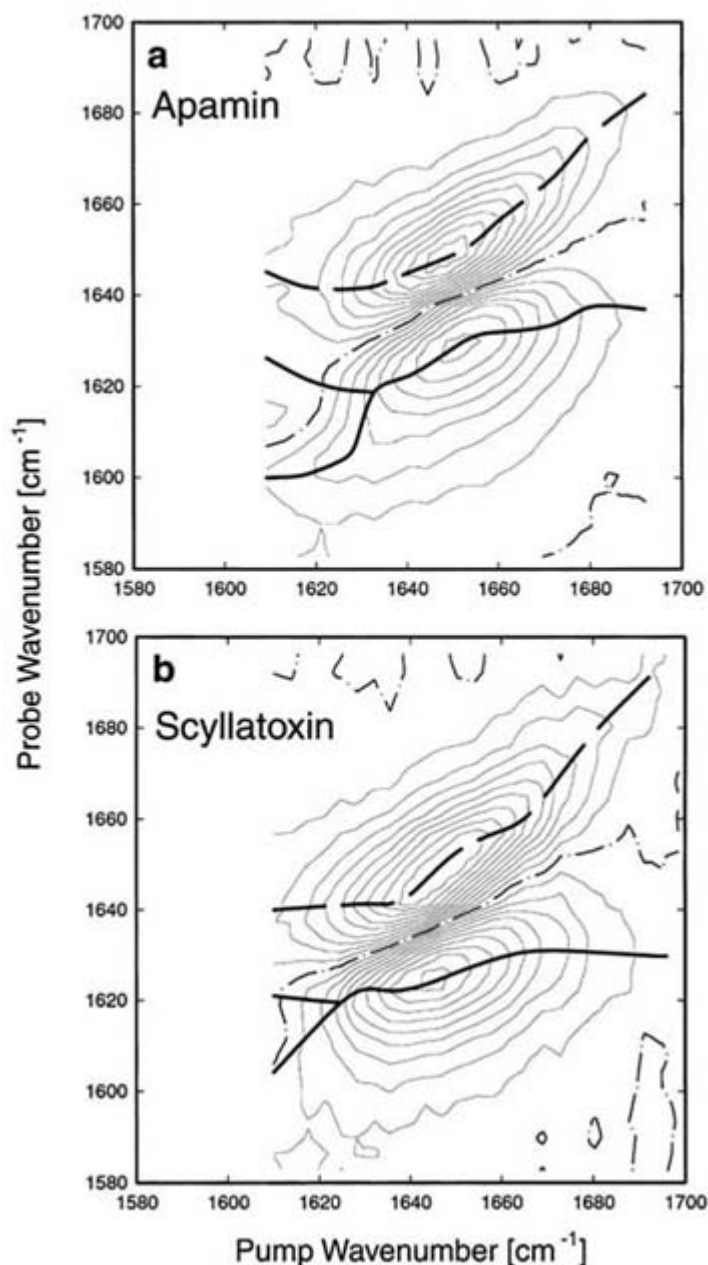


Figure B2.1.9 Two-dimensional time-resolved IR holeburning spectra obtained with two small polypeptides, apamin and scyllatoxin, by Hochstrasser and co-workers [58]. Figure courtesy Professor R M Hochstrasser (University of Pennsylvania).

B2.1.3.3 PHOTON-ECHO AND TRANSIENT-GRATING SPECTROSCOPY

The methods discussed so far, fluorescence upconversion, the various pump–probe spectroscopies, and the polarized variations for the measurement of anisotropy, are essentially conventional spectroscopies adapted to the femtosecond regime. At the simplest level of interpretation, the information content of these conventional time-resolved methods pertains to *populations* in resonantly prepared or probed states. As applied to chemical kinetics, for most *slow* reactions (on the ten picosecond and longer time scales), populations adequately specify the position of the reaction coordinate; intermediates and products show up as time-delayed spectral entities, and assignment of the transient spectra to chemical structures follows, in most cases, the same principles used in spectroscopic experiments performed with continuous wave or nanosecond pulsed lasers.

The multiple-pulse methods discussed in this section, in contrast, can be used to obtain information on the time evolution of electronic or vibrational coherence, the correlation of *phase* between two states. In *fast*, sub-picosecond chemical reactions and in energy transfer, as examples, knowledge of the time evolution of coherence *and* population is essential if a correct physical model is to be established. For example, in the purple-bacterial photosynthetic reaction centre, the 3 ps time scale for the charge transfer from the bacteriochlorophyll dimer that serves as the primary electron donor to the pheophytin that serves as the electron acceptor is comparable to that of vibrational dephasing and just longer than the fast part of electronic dephasing [146, 149, 156, 161, 162, 163, 164, 165, 166 and 167]. It is certainly inadequate to describe this situation just in terms of populations and the states of the *individual* macrocycles that are involved in the charge-transfer reaction. A similar problem applies to the photophysics of retinal in the visual and proton-pumping proteins rhodopsin and bacteriorhodopsin. The light-initiated isomerization of retinal in these proteins occurs on the sub-picosecond time scale; it involves a time evolution from an excited electronic state to a isomerized ground state on a time scale that is shorter than that involved in vibrational dephasing [114, 117, 118, 120, 143].

Photon-echo and transient-grating spectroscopy exploit, in general, time-ordered interactions between three or more optical pulses [125, 168]. The principles involved are analogous to those that are well established for pulsed experiments in nuclear and electron magnetic resonance [169]. Although the methods discussed below were applied first for the study of electronic coherence, as picosecond and femtosecond IR sources were developed a set of corresponding methods were applied to the *direct* study of vibrational coherence. Simple multiple-pulse sequences, with control over relative delays between pulses, can be formed using interferometers with three (or more) arms. Importantly, the pulses can be aimed spatially so that the phase-matched (momentum-conserving) outgoing directions for the signal (echo or diffracted) pulses are spatially resolved from the transmitted excitation pulses, affording zero-background detection [170].

Transient-grating spectroscopy is performed using two pump pulses that are arranged to arrive at the sample at the same time but aimed to overlap at an angle [171]. A *transmission diffraction grating* is formed owing to the spatial interference between the two pump pulses. The grating consists of alternating regions of excited-state molecules, formed in the bright regions where the two incoming beams interfere constructively, and of ground-state molecules, left undisturbed in the dark regions where the two beams interfere destructively. A third, probing laser pulse is diffracted by the population grating into a direction that is resolved from the pump directions, allowing the intensity of the population grating to be measured as a function of time. Any physical process that causes the spatial pattern of ground- and excited-state molecules to fade away or become less distinct will be detected in terms of a decrease in the diffracted beam's intensity. Thus, in addition to being sensitive to population decay, the transient-grating experiment

-20-

can be used to detect spatial motion of molecules owing to transport or diffusion. For example, Miller and co-workers [172, 173 and 174] have employed transient-grating spectroscopy to study protein motions in hem proteins. Further, since the diffraction efficiency is sensitive to the orientation of the transition-dipole moments of the excited-state molecules, the transient-grating method can be used to characterize rotational diffusion. Fayer and co-workers have also exploited polarization gratings, formed by making the two coincident incoming beams be orthogonally polarized [171].

The simplest echo experiment, the two-pulse photon echo, reports information on the decay of coherence in terms of echo intensity. As in the corresponding spin-echo experiment in magnetic resonance, the first incident laser pulse rotates the Bloch vector [168, 175] from pure population in the ground state to an orientation corresponding to a coherent superposition state; the density matrix now exhibits off-diagonal elements as well as diagonal elements. During the waiting period t between the first and second pulses, the off-diagonal (coherence) elements in the density matrix decay according to the dephasing time, T_2 , while the on-diagonal (population) elements decay according to the lifetime of the resonantly prepared state, T_1 . The

second laser pulse rotates the Bloch vectors again so that the phase of rotation of the Bloch vectors is inverted, which leads to refocusing. The vectors are maximally refocused at an interval t after the second pulse, so that the *spontaneous* emission arising from the ensemble of molecules is radiated with spatial coherence, forming an *echo* pulse. Thus, the intensity of the echo as a function of the waiting time t can be described by an exponential decay with time constant T_2 [176]. If the two input pulses are directed along the directions \vec{k}_1 and \vec{k}_2 , echo signal beams are emitted along the $2\vec{k}_1 - \vec{k}_2$ and $2\vec{k}_2 - \vec{k}_1$ directions. In practice, the input and echo beams are recollimated by a single lens after they emerge from the sample; iris apertures are used to spatially isolate the direction of either of the echo beams so that the emerging input beams are blocked.

In liquids, the Bloch equations (single-dephasing time scale) picture [175] used above to describe the formation of echo signals is apparently inadequate. It is now known that electronic dephasing occurs over a distribution of time scales, so a single time constant T_2 is insufficient to describe all of the line-broadening dynamics [177]. The two-pulse echo method described above only is sensitive to the fastest of processes; in organic molecules in solution, the two-pulse echoes typically decay on the 20 fs or shorter time scale [176]. This is the time scale usually assigned to homogeneous line broadening. The slower electronic dephasing processes that contribute to inhomogeneous line broadening, involving solvent-induced fluctuations or radiationless decay between uncorrelated states, extend over the 10 fs to 100 ps (or longer) time scales in liquids and proteins [77, 78, 177].

A three-pulse or stimulated photon-echo experiment can be employed to characterize dephasing on a much longer time scale than is accessible to the two-pulse photon echo experiment. Figure B2.1.10 shows a three-pulse interferometer and beam-input geometry employed by Fleming and co-workers [170, 178, 179, 180 and 181]. The modified forward-box beam-input geometry allows three-pulse echoes (or grating signals) to be detected in the phase-matched $\vec{k}_1 - \vec{k}_2 + \vec{k}_3$ and $-\vec{k}_1 + \vec{k}_2 + \vec{k}_3$ directions. As in the Hahn stimulated spin-echo sequence [169], for a given waiting time t between the first two pulses, a plot of the echo intensity as a function of the time period T between the second and third pulses returns just the lifetime T_1 of the resonantly prepared state. At a given delay T , a plot of the intensity as a function t returns a decay related to the dephasing time T_2 , but there is a subtle change in the shape of the intensity envelope that is discernible as T is varied [170, 179, 182, 183, 184 and 185]. Figure B2.1.11 shows the results obtained from a three-pulse photon-echo experiment on a small protein subunit that binds an extended tetrapyrrole chromophore [186]. At early delays T , the shape is asymmetrical, with the maximum intensity shifted away from $t = 0$. As T is increased, so that the ensemble evolves for longer time periods prior to rephasing by the third pulse, the envelope becomes more symmetrical, and the maximum shifts back to near $t = 0$. The asymmetrical shape observed at early

delays T reports the presence of an echo, but the symmetrical signal observed at longer delays T arises from a free-induction decay only.

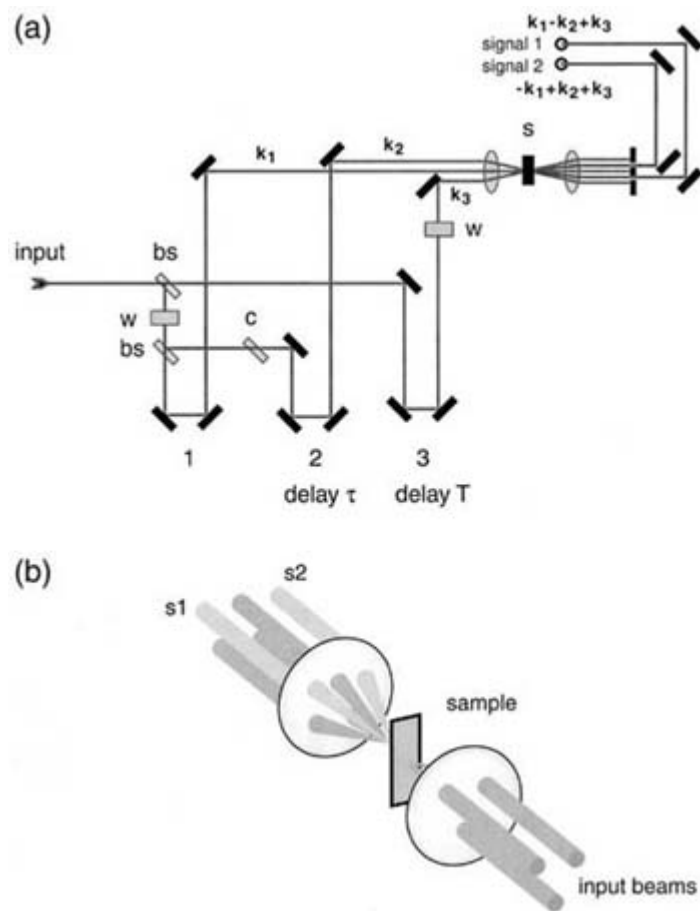


Figure B2.1.10 Stimulated photon-echo peak-shift (3PEPS) signals. Top: pulse sequence and interpulse delays t and T . Bottom: echo signals scanned as a function of delay t at three different population periods T , obtained with samples of a tetrapyrrole-containing light-harvesting protein subunit, the α subunit of *C*-phycoyanin.

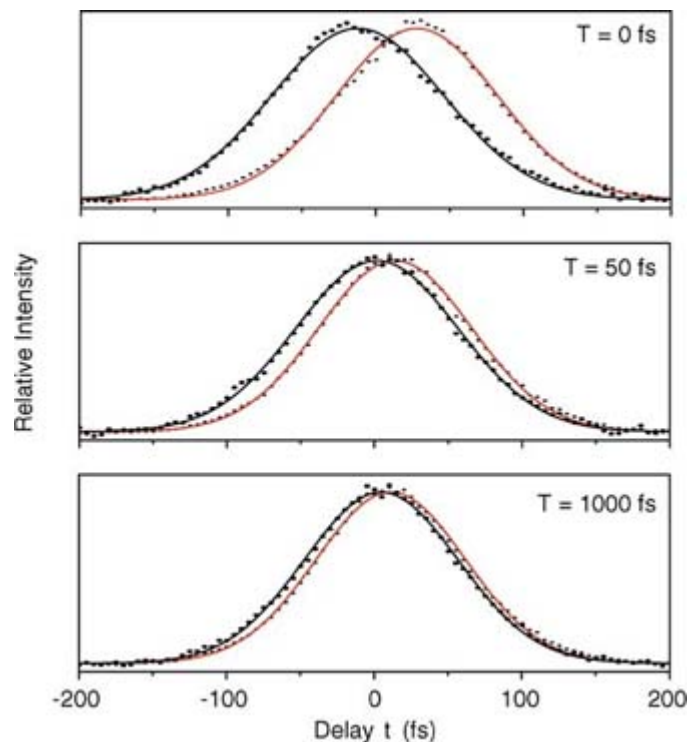


Figure B2.1.11 3PEPS profile obtained at room temperature with samples of a tetrapyrrole-containing light-harvesting protein subunit, the α subunit of *C*-phycoyanin, as used in the previous figure.

This experiment is now known as a three-pulse stimulated photon-echo peak-shift (3PEPS) experiment. Weiner and Ippen [182] were the first to describe the echo-envelope-shifting phenomenon and its relationship to inhomogeneous line broadening; application of the 3PEPS method to problems of dynamic solvation has been popularized especially by the groups of Fleming [177, 178 and 179, 184, 185] and of Wiersma, who has advanced gated versions of the experiment that actually time-resolves the echo in order to obtain additional information [187, 188 and 189]. The 3PEPS method returns, in general, superior information on solvation dynamics as compared to that returned by dynamic Stokes shift measurements by fluorescence upconversion or transient hole-burning spectroscopy because no line shape assumptions have to be made; in fact, a full analysis of the time-correlation and line-broadening functions obtained from the 3PEPS experiment can be used to obtain all pertinent spectroscopic observables, including the absorption and fluorescence line shapes. Owing to the mapping of coherence into population by the second pulse in the sequence, dephasing can be studied using the 3PEPS method over an enormous time scale, generally as long as the lifetime of the resonant state [177]. Figure B2.1.12 shows the entire 3PEPS profile obtained from a series of experiments on the sample used for figure B2.1.11 conducted with a series of T delays. Several different time scales that contribute to electronic dephasing are notable, corresponding to a very fast decay on the <20 fs time scale, a roughly exponential decay on the 100 fs time scale, and a slower decay to a long-lived offset over the 200 fs to 1 ps time scale [186]. The magnitude of the peak shift for each component is proportional to the strength of coupling of solvent fluctuations on a given time scale to the electronic dipole of the resonant electronic state that is used as a probe [177].

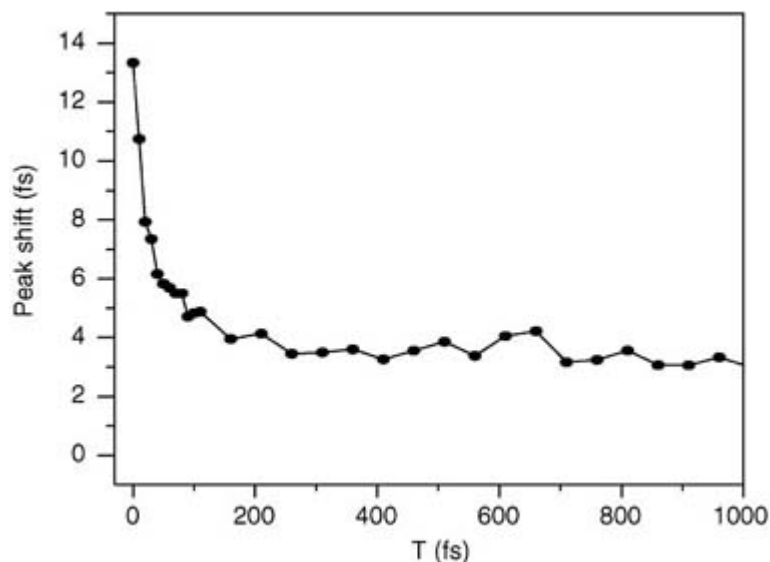


Figure B2.1.12. 3PEPS profile obtained at room temperature with samples of a tetrapyrrole-containing light-harvesting protein subunit, the a subunit of C-phycoerythrin, as used in the previous figure.

As implied above, comparable two- and three-pulse echo experiments can be conducted with femtosecond IR pulses in order to study *vibrational* dephasing. Notable work in this area has been conducted by Fayer and co-workers with picosecond IR pulses obtained from a free-electron laser at Stanford University [190, 191, 192, 193, 194 and 195]. Vibrational states can also be prepared with visible femtosecond pulses through the use of stimulated Raman *coherences*. In this family of methods, two laser pulses with frequencies ω_1 and ω_2 act simultaneously to transfer population to a vibrational level of frequency $\omega_1 - \omega_2$. The chief advantage of using pairs of visible pulses is that a wider frequency range becomes accessible owing to the availability of very short visible pulses; experiments conducted with IR pulses are limited by time-bandwidth considerations to lower frequencies. At present, however, the stimulated Raman methods have only been used *non-resonantly*, so very intense femtosecond laser pulses are required, typically in the $\mu\text{J}/\text{pulse}$ regime. In contrast, photon-echo experiments conducted with resonant electronic states are generally conducted with pulse energies in the low nJ regime.

One important stimulated Raman method, known as the Raman echo experiment, is analogous to the three-pulse or stimulated photon-echo discussed above. Two pairs of visible pulses prepare and map vibrational coherences into population, respectively, and after a waiting period a fifth pulse is used to stimulate rephasing and echo formation. Berg and co-workers have used the Raman echo to probe the vibrational dephasing of rotating methyl groups in different solvent environments [196, 197]. Tokmakoff, Fleming, and their co-workers [198, 199, 200 and 201] have exploited the intrinsic two-dimensionality of the Raman echo experiment to explore anharmonic coupling between intermolecular modes in liquid CS_2 ; each axis of the experiment returns information that is analogous to that available from conventional stimulated coherent Raman spectroscopy, but in the two-dimensional representation cross-peaks appear that directly report coupling between two vibrational modes.

B2.1.3.4 COHERENT CONTROL AND FUTURE ULTRAFAST SPECTROSCOPIES

A number of investigators are now developing pulse-shaping and modulation techniques that are useful with ultrashort laser pulses. These methods will permit preparation of precisely timed *and phased* multipulse sequences of arbitrary complexity for use in nonlinear spectroscopy. In addition, rather than just exploiting pulse sequences to project coherences into echo intensities and time shifts for spectroscopic purposes, as in the methods discussed above, several investigators are devising pulse sequences to focus wavefunctions onto

potential surfaces in a non-statistical manner. This concept, known generally as *coherent control* [202, 203 and 204], refers to attempts to control chemical reactions with specially constructed sequences of ultrashort laser pulses of known phase evolution and duration.

Scherer *et al* [205, 206] showed how to prepare, using interferometric methods, pairs of laser pulses with known relative phasing. These pulses were employed in experiments on vapour phase I₂, in which wavepacket motion was detected in terms of fluorescence emission. A more general approach, which can be used in principle to generate pulse sequences of any type, is to transform a single input pulse into a *shaped* output profile, with the intensity and phase of the output under control throughout. The idea being exploited by a number of investigators, notably Warren and Nelson, is to use a programmable dispersive delay line constructed from a pair of diffraction gratings spaced by an *active* device that is used either to absorb or phase shift selectively the frequency-dispersed wavefront. The approach favoured by Warren and co-workers exploits a Bragg cell driven by a radio-frequency signal obtained from a frequency synthesizer and a computer-controlled arbitrary waveform generator [207]. Nelson and co-workers use a computer-controlled liquid-crystal pixel array as a mask [208]. In the future, it is likely that one or both of these approaches will allow execution of currently impossible nonlinear spectroscopies with highly selective information content. One can take inspiration from the complex pulse sequences used in modern multiple-dimension NMR spectroscopy to suppress unwanted interfering resonances and to enhance selectively the resonances from targeted nuclei.

A simple example of what is possible now, even with two-pulse sequences, is the work by Shank and co-workers on focused RISRS wavepackets. Bardeen *et al* [209] used pump pulses prepared on purpose with a linear, negative chirp to enhance the magnitude of wavepackets driven to the ground state by impulsive stimulated Raman scattering. In the simplest application, this kind of approach might be used to make it easier to detect weakly displaced normal modes in dynamic absorption spectroscopy. In a mode-selective chemistry example, focusing of wavepackets [204, 210] might be used to prepare a certain vibrational superposition state, which could be subsequently excited by another laser pulse to produce an enhanced product yield [211]. The goal of this kind of work is to drive chemical reactions in directions that are normally not possible along normal kinetically and energetically controlled routes [204].

REFERENCES

- [1] Porter G 1992 Chemistry in microtime *The Chemical Bond: Structure and Dynamics* ed A Zewail (Boston: Academic) pp 113–48
- [2] Porter G 1995 Flash photolysis into the femtosecond—a race against time *Femtosecond Chemistry* ed J Manz and L Wöste (New York: VCH) pp 3–13
- [3] Shank C V 1986 Investigation of ultrafast phenomena in the femtosecond domain *Science* **233** 1276–80

- [4] Raksi F, Wilson K R, Jiang Z, Iklef A, Côté C Y and Kieffer J-C 1996 Ultrafast x-ray absorption probing of a chemical reaction *J. Chem. Phys.* **104** 6066–99
- [5] Schoenlein R W, Leeman W P, Chin A H, Volfbein P, Glover T E, Balling P, Zolotorev M, Kim K-J, Chattopadhyay S and Shank C V 1996 Femtosecond x-ray pulses at 0.4 Å generated by 90° Thomson scattering: a tool for probing the structural dynamics of materials *Science* **274** 236–8

- [6] Fattinger C and Grischkowsky D 1989 Terahertz beams *Appl. Phys. Lett.* **54** 490–2
- [7] Katzenellenbogen N and Grischkowsky D 1991 Efficient generation of 380 fs pulses of THz radiation by ultrafast laser pulse excitation of a biased metal–semiconductor interface *Appl. Phys. Lett.* **58** 222–4
- [8] Pedersen J E and Keiding S R 1992 THz time-domain spectroscopy of non-polar liquids *IEEE J. Quantum. Electron.* **28** 2518–22
- [9] Harde H, Katzenellenbogen N and Grischkowsky D 1995 Line-shape transition of collision broadened lines *Phys. Rev. Lett.* **74** 1307–10
- [10] Flanders B N, Cheville R A, Grischkowsky D and Scherer N F 1996 Pulsed terahertz transmission spectroscopy of liquid CHCl_3 , CCl_4 , and their mixtures *J. Phys. Chem.* **100** 11 824–35
- [11] Shank C V 1988 Generation of ultrashort optical pulses *Ultrashort Laser Pulses and Applications* ed W Kaiser (Berlin: Springer) pp 5–34
- [12] Valdmanis J A, Fork R L and Gordon J P 1985 Generation of optical pulses as short as 27 femtoseconds directly from a laser balancing self phase modulation, group velocity dispersion, saturable absorption and saturable gain *Opt. Lett.* **10** 131–3
- [13] Valdmanis J A and Fork R L 1986 Design considerations for a femtosecond pulse laser balancing self phase modulation, group velocity dispersion, saturable absorption, and saturable gain *IEEE J. Quantum. Electron.* **22** 112–18
- [14] Knox W H, Downer M C, Fork R L and Shank C V 1984 Amplified femtosecond optical pulses and continuum generation at 5 kHz repetition rate *Opt. Lett.* **9** 552–4
- [15] Fork R L, Brito Cruz C H, Becker P C and Shank C V 1987 Compression of optical pulses to six femtoseconds by using cubic phase compensation *Opt. Lett.* **12** 483–5
- [16] Zewail A H 1988 Laser femtochemistry *Nature* **328** 760–1
- [17] Khundkar L R and Zewail A H 1990 Ultrafast molecular reaction dynamics in real-time: progress over a decade *Annu. Rev. Phys. Chem.* **41** 15–60
- [18] Zewail A H 1990 The birth of molecules *Sci. Am.* **263** 76–82
- [19] Gruebele M and Zewail A H 1990 Ultrafast reaction dynamics *Phys. Today* **43** 24–33
- [20] Zewail A H 1991 Femtosecond transition-state dynamics *Faraday Discuss. Chem. Soc.* **91** 207–37
- [21] Polanyi J C and Zewail A C 1995 Direct observation of the transition state *Acc. Chem. Res.* **28** 119–32

- [22] Zewail A H 1995 Femtochemistry: concepts and applications *Femtosecond Chemistry* ed J Manz and L Wöste (New York: VCH) pp 15–128
- [23] Spence D E, Kean P N and Sibbett W 1991 60 fs pulse generation from a self-mode-locked Ti:sapphire laser *Opt. Lett.* **16** 42–4
- [24] Huang C-P, Asaki M T, Backus S, Murnane M M and Kapteyn H C 1992 17 fs pulses from a self-mode-locked Ti:sapphire laser *Opt. Lett.* **17** 1289–91

- [25] Asaki M T, Huang C-P, Garvey D, Zhou J, Kapteyn H C and Murnane M M 1993 Generation of 11 fs pulses from a self-mode-locked Ti:sapphire laser *Opt. Lett.* **977** 977–9
- [26] Stingl A, Lenzner M, Spielmann C, Krausz F and Szepes R 1996 Sub-10 fs mirror-dispersion-controlled Ti:sapphire laser *Opt. Lett.* **20** 602–4
- [27] Jung I D, Kärtner F X, Matuschek N, Sutter D H, Morier-Genoud F, Zhang G, Keller U, Scheuer V, Tilsch M and Tschudi T 1997 Self-starting 6.5 fs pulses from a Ti:sapphire laser *Opt. Lett.* **22** 1009–11
- [28] Bado P, Bourvier M and Coe J S 1987 Nd:YLF mode-locked oscillator and regenerative amplifier *Opt. Lett.* **12** 319–21
- [29] Pessot M, Squier J, Mourou G and Harter D 1989 Chirped-pulse amplification of 100 fsec pulses *Opt. Lett.* **14** 797–9
- [30] Vaillancourt G, Norris T B, Coe J S, Bado P and Mourou G 1990 Operation of a 1 kHz pulse-pumped Ti:sapphire regenerative amplifier *Opt. Lett.* **15** 317–19
- [31] Rudd J V, Korn G, Kane S, Squier J, Mourou G and Bado P 1993 Chirped-pulse amplification of 55 fs pulses at a 1 kHz repetition rate in a TiAl_2O_3 regenerative amplifier *Opt. Lett.* **18** 2044–6
- [32] Wilhelm T, Piel J and Riedle E 1997 Sub-20 fs pulses tunable across the visible from a blue-pumped single-pass noncollinear parametric converter *Opt. Lett.* **22** 1494–6
- [33] Shirakawa A, Sakane I and Kobayasi T 1998 Pulse-front-matched optical parametric amplification for sub-10 fs pulse generation tunable in the visible and near infrared *Opt. Lett.* **23** 1292–4
- [34] Cerullo G, Nisoli M, Stagira S and De Silvestri S 1998 Sub-8 fs pulses from an ultrabroadband optical parametric amplifier in the visible *Opt. Lett.* **23** 1283–5
- [35] Pshenichnikov M S, de Boeij W P and Wiersma D A 1994 Generation of 13 fs, 5 MW pulses from a cavity-dumped Ti:sapphire laser *Opt. Lett.* **19** 572–4
- [36] Baltuska A, Wei Z, Pshenichnikov M S and Wiersma D A 1997 Optical pulse compression to 5 fs at a 1 MHz repetition rate *Opt. Lett.* **22** 102–4
- [37] Evans J M, Spence D E, Sibbett W, Chai B H T and Miller A 1992 50 fs pulse generation from a self-mode-locked Cr:LiSrAlF_6 laser *Opt. Lett.* **17** 1447–9

- [38] Valentine G J, Hopkins J-M, Loza-Alvarez P, Kennedy G T, Sibbett W, Burns D and Valster A 1997 Ultralow-pump-threshold, femtosecond $\text{Cr}^{3+}:\text{LiSrAlF}_6$ laser pumped by a single narrow-stripe AlGaInP laser diode *Opt. Lett.* **22** 1639–41
- [39] Fork R L, Shank C V, Hirlimann C, Yen R and Tomlinson W J 1983 Femtosecond white-light continuum pulses *Opt. Lett.* **8** 1–3
- [40] Norris T B 1992 Femtosecond pulse amplification at 250 kHz with a Ti:sapphire regenerative amplifier and application to continuum generation *Opt. Lett.* **17** 1009–11
- [41] Joo T, Jia Y and Fleming G R 1995 Ti:sapphire regenerative amplifier for ultrashort high-power multikilohertz pulses without an external stretcher *Opt. Lett.* **20** 389–91
- [42] Le Blanc C, Grillon G, Chambaret J P, Migus A and Antonetti A 1993 Compact and efficient multipass Ti:sapphire system for femtosecond chirped-pulse amplification at the terawatt level *Opt. Lett.* **18** 140–

- [43] Backus S, Peatross J, Huang C P, Murnane M M and Kapteyn H C 1995 Ti:sapphire amplifier producing millijoule-level, 21 fs pulses at 1 kHz *Opt. Lett.* **20** 2000–2
- [44] Kane S, Squier J, Rudd J V and Mourou G 1994 Hybrid grating-prism stretcher-compressor system with cubic phase and wavelength tunability and decreased alignment sensitivity *Opt. Lett.* **19** 1876–8
- [45] Treacy E B 1969 Optical pulse compression with diffraction gratings *IEEE J. Quantum. Electron.* **5** 454–8
- [46] Yakovlev V V, Kohler B and Wilson K R 1994 Broadly tunable 30 fs pulses produced by optical parametric generation *Opt. Lett.* **19** 2000–2
- [47] Reed M K, Steiner-Shepard M K and Negus D K 1994 Widely tunable femtosecond optical parametric amplifier at 250 kHz with a Ti:sapphire regenerative amplifier *Opt. Lett.* **19** 1855–7
- [48] Greenfield S R and Wasielewski M R 1995 Near-transform-limited visible and near-IR femtosecond pulses from optical parametric amplification using Type II β -barium borate *Opt. Lett.* **20** 1394–6
- [49] Gale G M, Cavallari M, Driscoll T J and Hache F 1995 Sub-20 fs tunable pulses in the visible from an 82 MHz optical parametric oscillator *Opt. Lett.* **20** 1562–4
- [50] Hache F, Zéboulon A, Gallot G and Gale G M 1995 Cascaded second-order effects in the femtosecond regime in β -barium borate: self-compression in a visible femtosecond optical parametric oscillator *Opt. Lett.* **20** 1556–8
- [51] Umstadter D P, Barty C, Perry M and Mourou G A 1998 Tabletop, ultrahigh intensity lasers: dawn of nonlinear relativistic optics *Opt. Photon. News* **9** 41
- [52] Johnson A M and Shank C V 1989 Pulse compression in single-mode fibres—picoseconds to femtoseconds *The Supercontinuum Laser Source* ed R R Alfano (New York: Springer) pp 399–449
- [53] Ippen E 1997 Characterizing optical components for ultrafast laser applications *Optics 1997/98 Catalog* (Irvine, CA: Newport Corp.) pp 8-2–8-3
- [54] Fork R L, Martinez O E and Gordon J P 1984 Negative dispersion using pairs of prisms *Opt. Lett.* **9** 150–2

- [55] Gires F and Tournois P 1964 Interféromètre utilisable pour la compression d'impulsions lumineuses modulées en fréquence *Compte Rendue Acad. Sci. Paris* **258** 6112–15
- [56] Kuhl J and Heppner J 1986 Compression of femtosecond optical pulses with dielectric multilayer interferometers *IEEE J. Quantum. Electron.* **22** 182–5
- [57] Fragnito H L, Bigot J-Y, Becker P C and Shank C V 1989 Evolution of the vibronic absorption spectrum in a molecule following impulsive excitation with a 6 fs optical pulse *Chem. Phys. Lett.* **160** 101–4
- [58] Hamm P, Lim M and Hochstrasser R M 1998 Structure of the amide I band of peptides measured by femtosecond nonlinear-infrared spectroscopy *J. Phys. Chem. B* **102** 6123–38
- [59] Mollenauer L F, Stolen R H and Gordon J P 1980 Experimental observation of picosecond pulse narrowing and solitons in optical fibres *Phys. Rev. Lett.* **45** 1095–7
- [60] Agrawal G P 1989 Ultrashort pulse propagation in nonlinear dispersive fibers *The Supercontinuum*

Laser Source ed R R Alfano (New York: Springer) pp 91–116

- [61] Hasegawa A 1983 Amplification and reshaping of optical solitons in glass fibre–IV *Appl. Phys. Lett.* **8** 650–2
- [62] Mollenauer L F, Gordon J P and Islam M N 1986 Soliton propagation in long fibers with periodically compensated loss *IEEE J. Quantum. Electron.* **22** 157–73
- [63] Fleming G R 1986 *Chemical Applications of Ultrafast Spectroscopy* (New York: Oxford University Press)
- [64] Maznev A A, Crimmins T F and Nelson K A 1998 How to make femtosecond pulses overlap *Opt. Lett.* **23** 1378–80
- [65] Trebino R and Kane D J 1993 Using phase retrieval to measure the intensity and phase of ultrafast pulses: frequency-resolved optical gating *J. Opt. Soc. Am. A* **10** 1101–11
- [66] Kane D J and Trebino R 1993 Characterization of arbitrary femtosecond pulses using frequency-resolved optical gating *IEEE J. Quantum. Electron.* **29** 571–9
- [67] Kane D J and Trebino R 1993 Single-shot measurement of the intensity and phase of an arbitrary ultrashort pulse by using frequency-resolved optical gating *Opt. Lett.* **18** 823–5
- [68] Kane D J, Taylor A J, Trebino R and DeLong K W 1994 Single-shot measurement of the intensity and phase of a femtosecond UV laser pulse with frequency-resolved optical gating *Opt. Lett.* **19** 1061–3
- [69] DeLong K W, Trebino R, Hunter J and White W E 1994 Frequency-resolved optical gating with the use of second-harmonic generation *J. Opt. Soc. Am. B* **11** 2206–15
- [70] Walker G C, Jarzeba W, Kang T J, Johnson A E and Barbara P F 1990 Ultraviolet femtosecond fluorescence spectroscopy: techniques and applications *J. Opt. Soc. Am. B* **7** 1521–7
- [71] Jimenez R and Fleming G R 1996 Ultrafast spectroscopy of photosynthetic systems *Biophysical Techniques in Photosynthesis* ed J Amesz and A J Hoff (Dordrecht: Kluwer) pp 63–73
- [72] Cross A J and Fleming G R 1984 Analysis of time-resolved fluorescence anisotropy decays *Biophys. J.* **46** 45–56

- [73] Simon J D 1988 Time-resolved studies of solvation in polar media *Acc. Chem. Res.* **21** 128–34
- [74] Jarzeba W, Walker G C, Johnson A E and Barbara P F 1991 Nonexponential solvation dynamics of simple liquids and mixtures *Chem. Phys.* **152** 57–68
- [75] Maroncelli M 1993 The dynamics of solvation in polar solvents *J. Mol. Liq.* **57** 1–37
- [76] Jimenez R, Fleming G R, Kumar P V and Maroncelli M 1994 Femtosecond solvation dynamics of water *Nature* **369** 471–3
- [77] Stratt R M and Cho M 1994 The short-time dynamics of solvation *J. Chem. Phys.* **100** 6700–8
- [78] Stratt R M and Maroncelli M 1996 Nonreactive dynamics in solution: the emerging molecular view of solvation dynamics and vibrational relaxation *J. Phys. Chem.* **100** 12 981–96
- [79] van Amerongen H and Struve W S 1995 Polarized optical spectroscopy of chromoproteins *Methods Enzymol.* **246** 259–83

- [80] Tao T 1969 Time-dependent fluorescence depolarization and Brownian rotational diffusion coefficients of macromolecules *Biopolymers* **8** 609–32
- [81] Cross A J, Waldeck D H and Fleming G R 1983 Time resolved polarization spectroscopy: level kinetics and rotational diffusion *J. Chem. Phys.* **78** 6455–67
- [82] Matro A and Cina J A 1995 Theoretical study of time-resolved fluorescence anisotropy from coupled chromophore pairs *J. Phys. Chem.* **99** 2568–82
- [83] Abrash S, Repinec S and Hochstrasser R M 1990 The viscosity dependence and reaction coordinate for isomerization of *cis*-stilbene *J. Chem. Phys.* **93** 1041–53
- [84] Hochstrasser R M *et al* 1991 Anisotropy studies of ultrafast dipole reorientations *Proc. Indian Acad. Sci. (Chem. Sci.)* **103** 351–62
- [85] Wynne K and Hochstrasser R M 1993 Coherence effects in the anisotropy of optical experiments *Chem. Phys.* **171** 179–88
- [86] Wynne K and Hochstrasser R M 1995 Anisotropy as an ultrafast probe of electronic coherence in degenerate systems exhibiting Raman scattering, fluorescence, transient absorption and chemical reactions *J. Raman Spectrosc.* **26** 561–9
- [87] Rahman T S, Knox R S and Kenkre V M 1979 Theory of depolarization of fluorescence in molecular pairs *Chem. Phys.* **44** 197–211
- [88] Knox R S and Gülen D 1993 Theory of polarized fluorescence from molecular pairs *Photochem. Photobiol.* **57** 40–3
- [89] Galli C, Wynne K, LeCours S, Therien M J and Hochstrasser R M 1993 Direct measurement of electronic dephasing using anisotropy *Chem. Phys. Lett.* **206** 493–9
-

- [90] Edington M D, Riter R E and Beck W F 1995 Evidence for coherent energy transfer in allophycocyanin trimers *J. Phys. Chem.* **99** 15 699–704
- [91] Riter R E, Edington M D and Beck W F 1997 Isolated-chromophore and exciton-state photophysics in C-phycocyanin trimers *J. Phys. Chem. B* **101** 2366–71
- [92] Porter G and Topp M R 1968 Nanosecond flash photolysis and the absorption spectra of excited singlet states *Nature* **220** 1228-9
- [93] Zewail A and Bernstein R 1992 Real-time laser femtochemistry: viewing the transition from reagents to products *The Chemical Bond: Structure and Dynamics* ed A Zewail (San Diego, CA: Academic) pp 223–79
- [94] Scherer N F, Khundkar L R, Bernstein R B and Zewail A H 1987 Real-time picosecond clocking of the collision complex in a bimolecular reaction: the birth of OH from H + CO₂ *J. Chem. Phys.* **87** 1451–3
- [95] Dantus M, Rosker M J and Zewail A H 1987 Real-time femtosecond probing of 'transition states' in chemical reactions *J. Chem. Phys.* **87** 2395–7
- [96] Dantus M, Rosker M J and Zewail A H 1987 Femtosecond real-time probing of reactions. II. The dissociation reaction of ICN *J. Chem. Phys.* **89** 6128–40
- [97] Rosker M J, Dantus M and Zewail A H 1988 Femtosecond real-time probing of reactions. I. The

technique *J. Chem. Phys.* **89** 6113–27

- [98] Rosker M J, Dantus M and Zewail A H 1988 Femtosecond clocking of the chemical bond *Science* **241** 1200–2
- [99] Dantus M, Janssen M H M and Zewail A H 1991 Femtosecond probing of molecular dynamics by mass-spectrometry in a molecular beam *Chem. Phys. Lett.* **181** 281–7
- [100] Pedersen S, Herek J L and Zewail A H 1994 The validity of the 'Diradical' hypothesis: direct femtosecond studies of the transition-state structures *Science* **266** 1359–64
- [101] Brito Cruz C H, Fork R L, Knox W H and Shank C V 1986 Spectral hole burning in large molecules probed with 10 fs optical pulses *Chem. Phys. Lett.* **132** 341–5
- [102] Loring R F, Yan Y J and Mukamel S 1987 Time-resolved fluorescence and hole-burning line shapes of solvated molecules: longitudinal dielectric relaxation and vibrational dynamics *J. Chem. Phys.* **87** 5840–57
- [103] Vogel W, Welsch D-G and Wilhelmi B 1988 Time-resolved spectral hole burning *Chem. Phys. Lett.* **153** 376–8
- [104] Brito Cruz C H, Gordon J P, Becker P C, Fork R L and Shank C V 1988 Dynamics of spectral hole burning *IEEE J. Quantum. Electron.* **24** 261–6
- [105] Kinoshita S 1989 Theory of transient hole-burning spectrum for molecules in solution *J. Chem. Phys.* **91** 5175–84
- [106] Kang T J, Yu J and Berg M 1990 Rapid solvation of a nonpolar solute measured by ultrafast transient hole burning *Chem. Phys. Lett.* **174** 476–80

-31-

- [107] Kang T J, Yu J and Berg M 1991 Limitations on measuring solvent motion with ultrafast transient hole burning *J. Chem. Phys.* **94** 2413–24
- [108] Yu J and Berg M 1992 Solvent-electronic state interactions measured from the glassy to the liquid state. I. Ultrafast transient and permanent hole burning in glycerol *J. Chem. Phys.* **96** 8741–9
- [109] Murakami H, Kinoshita S, Hirata Y, Okada T and Mataga N 1992 Transient hole-burning and time-resolved fluorescence spectra of dye molecules in solution: evidence for ground-state relaxation and hole-filling effect *J. Chem. Phys.* **97** 7881–8
- [110] Ma J, Bout D V and Berg M 1995 Solvation dynamics studied by ultrafast transient hole burning *J. Mol. Liq.* **65/66** 301–4
- [111] Kovalenko S A, Ernsting N P and Ruthmann J 1996 Femtosecond hole-burning spectroscopy of the dye DCM in solution: the transition from the locally excited to a charge-transfer state *Chem. Phys. Lett.* **258** 445–54
- [112] Riter R E, Edington M D and Beck W F 1996 Protein-matrix solvation dynamics in a subunit of C-phycocyanin *J. Phys. Chem.* **100** 14 198–205
- [113] Edington M D, Riter R E and Beck W F 1997 Femtosecond transient hole-burning detection of interexciton-state radiationless decay in allophycocyanin trimers *J. Phys. Chem. B* **101** 4473–7
- [114] Dexheimer S L, Wang Q, Peteanu L A, Pollard W T, Mathies R A and Shank C V 1992 Femtosecond impulsive excitation of nonstationary vibrational states in bacteriorhodopsin *Chem. Phys. Lett.* **188** 61–6

- [115] Pollard W T, Dexheimer S L, Wang Q, Peteanu L A, Shank C V and Mathies R A 1992 Theory of dynamic absorption spectroscopy of nonstationary states. 4. Application to 12 fs resonant Raman spectroscopy of bacteriorhodopsin *J. Phys. Chem.* **96** 6147–58
- [116] Bardeen C J and Shank C V 1993 Femtosecond electronic dephasing in large molecules in solution using mode suppression *Chem. Phys. Lett.* **203** 535–9
- [117] Schoenlein R W, Peteanu L A, Wang Q, Mathies R A and Shank C V 1993 Femtosecond dynamics of cis-trans isomerization in a visual pigment analog: isorhodopsin *J. Phys. Chem.* **97** 12 087–92
- [118] Peteanu L A, Schoenlein R W, Wang Q, Mathies R A and Shank C V 1993 The first step in vision occurs in femtoseconds: complete blue and red spectral studies *Proc. Natl Acad. Sci. USA* **90** 11 762–6
- [119] Bardeen C J and Shank C V 1994 Ultrafast dynamics of the solvent-solute interaction measured by femtosecond four-wave mixing: LD690 in n-alcohols *Chem. Phys. Lett.* **226** 310–16
- [120] Wang Q, Schoenlein R W, Peteanu L A, Mathies R A and Shank C V 1994 Vibrationally coherent photochemistry in the femtosecond primary event of vision *Science* **266** 422–4
- [121] Wang Q, Kochendoerfer G G, Schoenlein R W, Verdegem P J E, Lugtenburg J, Mathies R A and Shank C V 1996 Femtosecond spectroscopy of a 13-demethylrhodopsin visual pigment analogue: the role of nonbonded interactions in the isomerization process *J. Phys. Chem.* **100** 17 388–94
- [122] Diffey W M, Homoelle B J, Edington M D and Beck W F 1998 Excited-state vibrational coherence and anisotropy decay in the bacteriochlorophyll a dimer protein B820 *J. Phys. Chem. B* **102** 2776–86

-32-

- [123] Edington M D, Riter R E and Beck W F 1996 Interexciton-state relaxation and exciton localization in allophycocyanin trimers *J. Phys. Chem.* **100** 14 206–17
- [124] Yamaguchi S and Hamaguchi H 1995 Convenient method of measuring the chirp structure of femtosecond white-light continuum pulses *Appl. Spectrosc.* **49** 1513–15
- [125] Mukamel S 1995 *Principles of Nonlinear Optical Spectroscopy* (New York: Oxford University Press)
- [126] Balk M W and Fleming G R 1985 Dependence of the coherence spike on the material dephasing time in pump-probe experiments *J. Chem. Phys.* **83** 4300–7
- [127] Cong P, Deuhl H P and Simon J D 1993 Using optical coherence to measure the ultrafast electronic dephasing of large molecules in room-temperature liquids *Chem. Phys. Lett.* **211** 367–73
- [128] Cong P, Simon J D and Yan Y 1995 Probing the molecular dynamics of liquids and solutions *Ultrafast Processes in Chemistry and Photobiology* ed M A El-Sayed, I Tanaka and Y Molin (Oxford: Blackwell) pp 53–82
- [129] Lee S-Y and Heller E J 1979 Time-dependent theory of Raman scattering *J. Chem. Phys.* **71** 4777–88
- [130] Heller E J, Sundberg R L and Tannor D 1982 Simple aspects of Raman scattering *J. Phys. Chem.* **86** 1822–33
- [131] Myers A B and Mathies R A 1987 Resonance Raman intensities: a probe of excited-state structure and dynamics *Biological Applications of Raman Spectroscopy* vol 2, ed T G Spiro (New York: Wiley-Interscience) pp 1–58
- [132] Yan Y-X, Cheng L-T and Nelson K A 1988 Impulsive stimulated light scattering *Advances in Non-*

linear Spectroscopy ed R J H Clark and R E Hester (Chichester: Wiley) pp 299–355

- [133] Zhu L, Li P, Huang M, Sage J T and Champion P M 1994 Real time observation of low frequency heme protein vibrations using femtosecond coherence spectroscopy *Phys. Rev. Lett.* **72** 301–4
- [134] Zhu L, Sage J T and Champion P M 1994 Observation of coherent reaction dynamics in heme proteins *Science* **266** 629–32
- [135] Zhu L, Wang W, Sage J T and Champion P M 1995 Femtosecond time-resolved vibrational spectroscopy of heme proteins *J. Raman Spectrosc.* **26** 527–34
- [136] Diffey W M and Beck W F 1997 Rapid-scanning interferometer for ultrafast pump–probe spectroscopy with phase-sensitive detection *Rev. Sci. Instrum.* 3296–300
- [137] Johnson A E and Myers A B A 1996 A comparison of time- and frequency-domain resonance Raman spectroscopy in triiodide *J. Chem. Phys.* **104** 2497–507
- [138] McMorrow D and Lotshaw W T 1990 The frequency response of condensed-phase media to femtosecond optical pulses: spectral-filter effects *Chem. Phys. Lett.* **174** 85–94
- [139] McMorrow D and Lotshaw W T 1991 Intermolecular dynamics in acetonitrile probed with femtosecond Fourier transform Raman spectroscopy *J. Phys. Chem.* **95** 10 395–406
- [140] McMorrow D and Lotshaw W T 1991 Dephasing and relaxation in coherently excited ensembles of intermolecular oscillators *Chem. Phys. Lett.* **178** 69–74

- [141] McMorrow D and Lotshaw W T 1993 Evidence for low-frequency ($\approx 15 \text{ cm}^{-1}$) collective modes in benzene and pyridine liquids *Chem. Phys. Lett.* **201** 369–76
- [142] Lotshaw W T, McMorrow D, Thantu N, Melinger J S and Kitchenbaum R 1995 Intermolecular vibrational coherence in molecular liquids *J. Raman Spectrosc.* **26** 571–83
- [143] Schoenlein R W, Peteanu L A, Mathies R A and Shank C V 1991 The first step in vision: femtosecond isomerization of rhodopsin *Science* **254** 412–15
- [144] Chudoba C, Riedle E, Pfeiffer M and Elsaesser T 1996 Vibrational coherence in ultrafast excited-state proton transfer *Chem. Phys. Lett.* **263** 622–8
- [145] Wynne K, Galli C and Hochstrasser R M 1994 Ultrafast charge transfer in an electron donor-acceptor complex *J. Chem. Phys.* **100** 4796–810
- [146] Vos M H, Rappaport F, Lambry J-C, Breton J and Martin J-L 1993 Visualization of the coherent nuclear motion in a membrane protein by femtosecond spectroscopy *Nature* **363** 320–5
- [147] Vos M H, Jones M R, Hunter C N, Breton J and Martin J-L 1994 Coherent nuclear dynamics at room temperature in bacterial reaction centers *Proc. Natl Acad. Sci. USA* **91** 12 701–5
- [148] Vos M H, Jones M R, Breton J, Lambry J-C and Martin J-L 1996 Vibrational dephasing of long- and short-lived primary donor states in mutant reaction centers of *Rhodobacter sphaeroides* *Biochemistry* **35** 2687–92
- [149] Stanley R J and Boxer S G 1995 Oscillations in the spontaneous fluorescence from photosynthetic reaction centers *J. Phys. Chem.* **99** 859–63
- [150] Walker G C, Maiti S, Cowen B R, Moser C C, Dutton P L and Hochstrasser R M 1994 Time resolution of electronic transitions of photosynthetic reaction centers in the infrared *J. Phys. Chem.*

- [151] Owrutsky J C, Raftery D and Hochstrasser R M 1994 Vibrational relaxation dynamics in solution *Annu. Rev. Phys. Chem.* **45** 519–55
- [152] Lian T, Locke B, Kholodenko Y and Hochstrasser R M 1994 Energy flow from solute to solvent probed by femtosecond IR spectroscopy: malachite green and heme protein solutions *J. Phys. Chem.* **98** 11 648–56
- [153] Owrutsky J C, Li M, Locke B and Hochstrasser R M 1995 Vibrational relaxation of the CO stretch vibration in hemoglobin-CO, myoglobin-CO, and protoheme-CO *J. Phys. Chem.* **99** 4842–6
- [154] Lian T, Kholodenko Y and Hochstrasser R M 1995 Infrared probe of the solvent response to ultrafast solvation processes *J. Phys. Chem.* **99** 2546–51
- [155] Diller R, Maiti S, Walker G C, Cowen B R, Pippenger R, Bogomolni R A and Hochstrasser R M 1995 Femtosecond time-resolved infrared laser study of the J–K transition of bacteriorhodopsin *Chem. Phys. Lett.* **241** 109–15
- [156] Haran G, Wynne K, Moser C C, Dutton P L and Hochstrasser R M 1996 Level mixing and energy redistribution in bacterial photosynthetic reaction centers *J. Phys. Chem.* **100** 5562–9

- [157] Wynne K, Haran G, Reid G D, Moser C C, Dutton P L and Hochstrasser R M 1996 Femtosecond infrared spectroscopy of low-lying excited states in reaction centers of *Rhodobacter sphaeroides* *J. Phys. Chem.* **100** 5140–8
- [158] Wang C, Akhremitchev B and Walker G C 1997 Femtosecond infrared and visible spectroscopy of photoinduced intermolecular electron transfer dynamics and solvent–solute reaction geometries: Coumarin 337 in dimethylaniline *J. Phys. Chem. A* **101** 2735–8
- [159] Walker G C, Barbara P F, Doorn S K, Dong Y and Hupp J T 1991 Ultrafast measurements on direct photoinduced electron transfer in a mixed-valence complex *J. Phys. Chem.* **95** 5712–15
- [160] Tominaga K, Kliner D A V, Johnson A E, Levinger N E and Barbara P F 1993 Femtosecond experiments and absolute rate calculations on intervalence electron transfer of mixed-valence compounds *J. Chem. Phys.* **98** 1228–43
- [161] Boxer S G, Goldstein R A, Lockhart D J, Middendorf T R and Takiff L 1989 Excited states, electron-transfer reactions, and intermediates in bacterial photosynthetic reaction centers *J. Phys. Chem.* **93** 8280–94
- [162] Kirmaier C and Holten D 1988 Subpicosecond spectroscopy of charge separation in *Rhodobacter capsulatus* reaction centers *Isr. J. Chem.* **28** 79–85
- [163] Jean J M, Chan C-K and Fleming G R 1988 Electronic energy transfer in photosynthetic bacterial reaction centers *Isr. J. Chem.* **28** 169–75
- [164] Breton J, Martin J-L, Fleming G R and Lambry J-C 1988 Low-temperature femtosecond spectroscopy of the initial step of electron transfer in reaction centers from photosynthetic purple bacteria *Biochemistry* **27** 8276
- [165] Holzappel W, Finkle U, Kaiser W, Oesterhelt D, Scheer H, Stolz H U and Zinth W 1989 Observation of a bacteriochlorophyll anion radical during the primary charge separation in a reaction center *Chem. Phys. Lett.* **160** 1–7
- [166] Holzappel W, Finkle U, Kaiser W, Oesterhelt D, Scheer H, Stolz H U and Zinth W 1990 Initial

electron-transfer in the reaction center from *Rhodobacter sphaeroides* *Proc. Natl Acad. Sci. USA* **87** 5168–72

- [167] Stanley R J, King B and Boxer S G 1996 Excited state energy transfer pathways in photosynthetic reaction centers. 1. Structural symmetry effects *J. Phys. Chem.* **100** 12 052–9
- [168] Levenson M D and Kano S S 1988 *Introduction to Nonlinear Laser Spectroscopy* (San Diego, CA: Academic)
- [169] Slichter C P 1980 *Principles of Magnetic Resonance* (Berlin: Springer)
- [170] Joo T and Albrecht A C 1993 Electronic dephasing studies of molecules in solution at room temperature by femtosecond degenerate four wave mixing *Chem. Phys.* **176** 233–47
- [171] Fourkas J T and Fayer M D 1992 The transient grating: a holographic window to dynamic processes *Acc. Chem. Res.* **25** 227–33
- [172] Genberg L, Richard L, McLendon G and Miller R J D 1991 Direct observation of global protein motion in hemoglobin and myoglobin on picosecond time scales *Science* **251** 1051–6

-35-

- [173] Miller R J D 1994 Energetics and dynamics of deterministic protein motion *Acc. Chem. Res.* **27** 145–50
- [174] Deak J, Richard L, Pereira M, Chui H-L and Miller R J D 1994 Picosecond phase grating spectroscopy: applications to bioenergetics and protein dynamics *Meth. Enzymol.* **232** 322–60
- [175] Allen L and Eberly J H 1975 *Optical Resonance and Two-Level Atoms* (New York: Wiley)
- [176] Becker P C, Fragnito H L, Bigot J-Y, Brito Cruz C H, Fork R L and Shank C V 1989 Femtosecond photon echoes from molecules in solution *Phys. Rev. Lett.* **63** 505–7
- [177] Fleming G R and Cho M 1996 Chromophore-solvent dynamics *Annu. Rev. Phys. Chem.* **47** 109–34
- [178] Joo T, Jia Y, Yu J-Y, Jonas D M and Fleming G R 1996 Dynamics in isolated bacterial light-harvesting antenna (LH2) of *Rhodobacter sphaeroides* at room temperature *J. Phys. Chem.* **100** 2399–409
- [179] Joo T, Jia Y, Yu J-Y, Lang M J and Fleming G R 1996 Third-order nonlinear time domain probes of solvation dynamics *J. Chem. Phys.* **104** 6089–108
- [180] Passino S A, Nagasawa Y, Joo T and Fleming G R 1997 Three-pulse echo peak shift studies of polar solvation dynamics *J. Phys. Chem. A* **101** 725–31
- [181] Nagasawa Y, Passino S A, Joo T and Fleming G R 1997 Temperature dependence of optical dephasing in an organic polymer glass *J. Chem. Phys.* **106** 4840–52
- [182] Weiner A M and Ippen E P 1985 Femtosecond excited state relaxation of dye molecules in solution *Chem. Phys. Lett.* **114** 456–60
- [183] Bigot J-Y, Portella M T, Schoenlein R W, Bardeen C J, Migus A and Shank C V 1991 Non-Markovian dephasing of molecules in solution measured with three-pulse femtosecond photon echoes *Phys. Rev. Lett.* **66** 1138–41
- [184] Cho M, Yu J-Y, Joo T, Nagasawa Y, Passino S A and Fleming G R 1996 The integrated photon echo and solvation dynamics *J. Phys. Chem.* **100** 11 944–53

- [185] Passino S A, Nagasawa Y, Joo T and Fleming G R 1996 Photon echo measurements in liquids using pulses longer than the electronic dephasing time *Ultrafast Phenomena X* ed P Barbara, W Knox, W Zinth and J Fujimoto (Berlin: Springer) pp 199–200
- [186] Homoelle B J, Edington M D, Diffey W M and Beck W F 1998 Stimulated photon-echo and transient-grating studies of protein-matrix solvation dynamics and interexciton-state radiationless decay in a phycocyanin and allophycocyanin *J. Phys. Chem. B* **102** 3044–52
- [187] de Boeij W P, Pshenichnikov M S and Wiersma D A 1995 Phase-locked heterodyne-detected stimulated photon echo. A unique tool to study solute-solvent interactions *Chem. Phys. Lett.* **238** 1
- [188] Pshenichnikov M S, Duppen K and Wiersma D A 1995 Time-resolved femtosecond photon echo probes bimodal solvent dynamics *Phys. Rev. Lett.* **74** 674–7
- [189] de Boeij W P, Pshenichnikov M S and Wiersma D A 1996 Mode suppression in the non-Markovian limit by time-gated stimulated photon echo *J. Chem. Phys.* **105** 2953–60
-

-36-

- [190] Tokmakoff A, Zimdars D, Urdahl R S, Francis R S, Kwok A S and Fayer M D 1995 Infrared vibrational photon echo experiments in liquids and glasses *J. Phys. Chem.* **99** 13 310–20
- [191] Tokmakoff A and Fayer M D 1995 Homogeneous vibrational dynamics and inhomogeneous broadening in glass-forming liquids: infrared photon echo experiments from room temperature to 10 K *J. Chem. Phys.* **103** 2810–26
- [192] Tokmakoff A and Fayer M D 1995 Infrared photon echo experiments: exploring vibrational dynamics in liquids and glasses *Acc. Chem. Res.* **28** 439–45
- [193] Tokmakoff A and Fayer M D 1995 Infrared photon echo experiments: exploring vibrational dynamics in liquids and glasses *Acc. Chem. Res.* **28** 437–45
- [194] Rella C W, Rector K D, Kwok A, Hill J R, Schwettman H A, Dlott D D and Fayer M D 1996 Vibrational echo studies of myoglobin–CO *J. Phys. Chem.* **100** 15 620–29
- [195] Rector K D, Rella C W, Hill J R, Kwok A S, Sligar S G, Chien E Y T, Dlott D D and Fayer M D 1997 Mutant and wild-type myoglobin–CO protein dynamics: vibrational echo experiments *J. Phys. Chem. B* **101** 1468–75
- [196] Vanden Bout D and Berg M 1995 Ultrafast Raman echo experiments in liquids *J. Raman Spectrosc.* **26** 503–11
- [197] Berg M and Vanden Bout D A 1997 Ultrafast Raman echo measurements of vibrational dephasing and the nature of solvent–solute interactions *Acc. Chem. Res.* **30** 65–71
- [198] Tokmakoff A, Lang M J, Larsen D S and Fleming G R 1997 Intrinsic optical heterodyne detection of a two-dimensional fifth order Raman response *Chem. Phys. Lett.* **272** 48–54
- [199] Tokmakoff A and Fleming G R 1997 Two-dimensional Raman spectroscopy of the intermolecular modes of liquid CS₂ *J. Chem. Phys.* **106** 2569–82
- [200] Tokmakoff A, Lang M J, Larsen D S, Fleming G R, Chernyak V and Mukamel S 1997 Two-dimensional Raman spectroscopy of vibrational interactions in liquids *Phys. Rev. Lett.* **79** 2702–5
- [201] Tokmakoff A, Lang M J, Jordanides X J and Fleming G R 1998 The intermolecular interaction mechanisms in liquid CS₂ at 295 and 165 K probed with two-dimensional Raman spectroscopy *Chem. Phys.* **233** 231–42

- [202] Warren W S 1993 Coherent control: the dream is alive *Science* **259** 1581
- [203] Nelson K A 1994 Coherent control: optics, molecules, and materials *Ultrafast Phenomena IX* ed P F Barbara, W H Knox, G A Mourou and A H Zewail (Berlin: Springer) pp 47
- [204] Kohler B, Krause J L, Raksi F, Wilson K R, Yakovlev V V, Whitnell R M and Yan Y 1995 Controlling the future of matter *Acc. Chem. Res.* **28** 133–40
- [205] Scherer N F, Carlson R J, Matro A, Du M, Ruggiero A J, Romero-Rochin V, Cina J A, Fleming G R and Rice S A 1991 Fluorescence-detected wave packet interferometry: time resolved molecular spectroscopy with sequences of femtosecond phase-locked pulses *J. Chem. Phys.* **95** 1487–511
- [206] Scherer N F, Matro A, Ziegler L D, Du M, Carlson R J, Cina J A and Fleming G R 1992 Fluorescence-detected wave packet interferometry. II. Role of rotations and determination of the susceptibility *J. Chem. Phys.* **96** 4180
-

-37-

- [207] Dugan M A, Tull J X and Warren W S 1997 High-resolution acousto-optic shaping of unamplified and amplified femtosecond laser pulses *J. Opt. Soc. Am. B* **14** 2348–58
- [208] Wefers M M and Nelson K A 1995 Analysis of programmable ultrashort waveform generation using liquid-crystal spatial light modulators *J. Opt. Soc. Am. B* **12** 1343–62
- [209] Bardeen C J, Wang Q and Shank C V 1998 Femtosecond chirped pulse excitation of vibrational wave packets in bacteriorhodopsin *J. Phys. Chem. A* **102** 2759–66
- [210] Krause J L, Whitnell R M, Wilson K R, Yan Y and Mukamel S 1993 Optical control of molecular dynamics: molecular cannons, reflectrons, and wave-packet focusers *J. Chem. Phys.* **99** 6562–78
- [211] Bardeen C J, Che J, Wilson K R, Yakovlev V V, Cong P, Kohler B, Krause J L and Messina M 1997 Quantum control of NaI photodissociation reaction product states by ultrafast tailored light pulses *J. Phys. Chem. A* **101** 3815–22
-

FURTHER READING

Manz J and Wöste L (eds) 1995 *Femtosecond Chemistry* (New York, VCH)

Theory and experimental techniques for study of chemical reaction dynamics with ultrafast spectroscopic methods.

El-Sayed M A, Tanaka I and Molin Y (eds) 1995 *Ultrafast Processes in Chemistry and Photobiology* (Oxford: Blackwell)

Applications of ultrafast spectroscopy to chemical dynamics, especially in the condensed phase and in proteins.

Mukamel S 1995 *Principles of Nonlinear Optical Spectroscopy* (New York: Oxford University Press)

A comprehensive theoretical treatment of nonlinear spectroscopy, with an emphasis on theory applicable to ultrafast nonlinear spectroscopy.

Fleming G R 1986 *Chemical Applications of Ultrafast Spectroscopy* (New York: Oxford University Press)

Fundamentals of technique and theory for ultrafast experiments in chemistry, written before the titanium–sapphire revolution but still indispensable.

Kaiser W (ed) 1988 *Ultrashort Laser Pulses and Applications* (Berlin: Springer)

Applications of ultrafast laser techniques for studies in solids, optoelectronics, condensed phase, and in biological systems.

Rullière C (ed) 1998 *Femtosecond Laser Pulses* (Berlin: Springer)

A current description of femtosecond laser technology, with a discussion of ultrafast spectroscopic applications.

-1-

B 2.2 Electron, ion and atom scattering

M R Flannery

B2.2.1 INTRODUCTION

This chapter deals with quantal and semiclassical theory of heavy-particle and electron–atom collisions. Basic and useful formulae for cross sections, rates and associated quantities are presented. A consistent description of the mathematics and vocabulary of scattering is provided. Topics covered include collisions, rate coefficients, quantal transition rates and cross sections, Born cross sections, quantal potential scattering, collisions between identical particles, quantal inelastic heavy-particle collisions, electron–atom inelastic collisions, semiclassical inelastic scattering and long-range interactions.

B2.2.2 COLLISIONS

B2.2.2.1 DIFFERENTIAL AND INTEGRAL CROSS SECTIONS

A uniform monoenergetic beam of *test* or *projectile* particles A with number density N_A and velocity v_A is incident on a single *field* or *target* particle B of velocity v_B . The direction of the relative velocity $\varpi = v_A - v_B$ is along the Z-axis of a Cartesian XYZ frame of reference. The incident current (or intensity) is then $j_i = N_A v_A$, which is the number of test particles crossing unit area normal to the beam in unit time. The differential cross section for scattering of the test particles into unit solid angle $d\Omega = d(\cos \psi) d\phi$ about the direction $\hat{v}'(\psi, \phi)$ of the final relative motion is

$$\frac{d\sigma(v; \psi, \phi)}{d\Omega} = \frac{\text{Number of test particles scattered by one field particle into unit solid angle per unit time}}{\text{Current } j_i \text{ of incident beam}}$$

The number of particles scattered per unit time by the field particle and detected per unit time is then

$$\frac{dN_d}{dt} = j_i \frac{d\sigma}{d\Omega} d\Omega = N_A v \frac{d\sigma}{d\Omega} d\Omega = N_A v \frac{d\sigma}{d\Omega} \frac{dA}{r^2}$$

where the detector, located along the scattered direction $\hat{v}(\psi, \phi)$, subtends an angle $d\Omega = dA/r^2$ at the scattering centre and projects an area $dA = r^2 d(\cos \theta) d\phi$ normal to the scattered beam. Thus $[d\sigma/d\Omega] d\Omega$ is the cross-sectional area of the beam that is intercepted by one target particle and scattered into the solid angle dA/r^2 of a cone with axis along $\hat{v}(\psi, \phi)$ and vertical angle $d\psi$. In classical terms (figure B2.2.1), the number of particles detected per second about direction (ψ, ϕ) is the number $N_A v(bdb d\phi)$ of incident particles crossing the initial areal element $bdb d\phi$ per second.

-2-

Hence

$$\frac{d\sigma}{d\Omega} = \frac{b db}{d(\cos \psi)}.$$

For an incident current flowing between two cylinders of radii b and $b + db$, then $j_i 2\pi [d\sigma/d\Omega] d(\cos \psi)$ is the number of particles scattered per second between the two cones of semivertical angles $\psi, \psi + d\psi$ (figure B2.2.1).

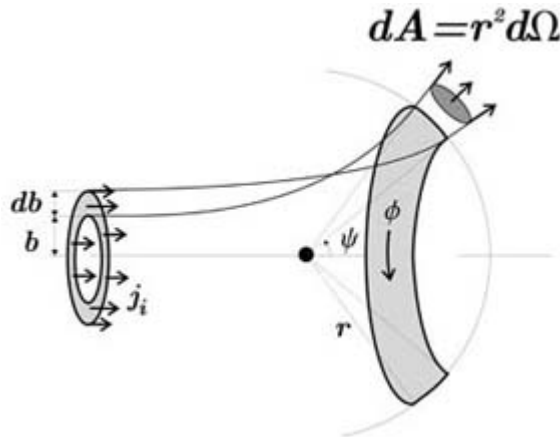


Figure B2.2.1. Scattering of a beam with current $j_i = N_A v$ particles per unit area incident between two cylinders of radii b and $b + db$ by one particle at rest in the laboratory.

The integral cross section for scattering over all directions is

$$\sigma(v) = \int_{-1}^{+1} d(\cos \psi) \int_0^{2\pi} \left[\frac{d\sigma}{d\Omega}(v; \psi, \phi) \right] d\phi.$$

The integral cross section is therefore the effective area presented by each field particle B for scattering of the test particles A into all directions. The *probability* that the test particles are scattered into a given direction $\hat{v}(\psi, \phi)$ is the ratio

$$\mathcal{P}(v; \psi, \phi) = \frac{d\sigma(v; \psi, \phi)}{d\Omega} / \sigma(v)$$

of the differential-to-integral cross sections.

B2.2.2.2 COLLISION RATES, COLLISION FREQUENCY AND PATH LENGTH

An electron or atomic beam of (projectile or test) particles A with density N_A of particles per cm^3 travels with speed v and energy E through an infinitesimal thickness dx of (target or field) gas particles B at rest with density N_B particles per cm^3 . The particles are scattered out of the beam by A–B collisions with integral cross section $\sigma(E)$ at a rate ($\text{cm}^{-3} \text{s}^{-1}$) given by the total number of collisions between A and B particles

$$\begin{aligned} \frac{dN_A(E)}{dt} &= -[N_A(E)v\sigma(E)]N_B \\ &= -k(E)N_A(E)N_B \\ &= -v_B(E)N_A(E) \end{aligned}$$

in unit time and unit volume. The *microscopic rate coefficient* ($\text{cm}^3 \text{s}^{-1}$) for the scattering of one test particle by one field particle is $k(E) = v\sigma(E)$. The frequency (s^{-1}) of collision between one test particle and N_B field particles (cm^{-3}) is $v_B = k(E)N_B$. Since $v = dx/dt$, the variation with x of intensity j_i of the attenuated beam is governed by

$$\frac{dj_i}{dx} = -[N_B\sigma(E)]j_i(x).$$

For constant density N_B and speed v , the solution is

$$\begin{aligned} j_i(E, x) &= j_i(E, 0) \exp[-N_B\sigma(E)x] \\ &= j_i(E, 0) \exp(-x/\lambda) \end{aligned}$$

where $\lambda \equiv 1/N_B\sigma(E) = v/v_B$ is the path length between collisions. Since $j_i = N_A v$, the density $N_A(E, x)$ obeys a similar equation. These equations describe the attenuation of a particle beam A travelling through a target gas B. For target gas particles with a distribution $f_B(v_B) dv_B$ in velocities v_A , the microscopic rate then becomes

$$k(E) = \int |v_A - v_B| \sigma(|v_A - v_B|) f_B(v_B) dv_B$$

where $E = \frac{1}{2}M_A v_A^2$ is the kinetic energy of the projectile beam. For an isothermal beam with an energy distribution $f_A(E) dE$ at temperature T , the macroscopic rate coefficient ($\text{cm}^3 \text{s}^{-1}$) or thermal rate constant is

$$k(T) = \int_0^\infty k(E) f_A(E) dE.$$

B2.2.2.3 ENERGY AND ANGULAR MOMENTUM: CENTRE OF MASS AND RELATIVE VELOCITY

The velocity of the centre of mass (CM) of the projectile and target particles of respective masses M_A and M_B is

$$\mathbf{V} = (M_A \mathbf{v}_A + M_B \mathbf{v}_B) / (M_A + M_B).$$

The relative velocity is

$$\mathbf{v} = \mathbf{v}_A - \mathbf{v}_B.$$

The velocities of A and B in terms of \mathbf{V} and \mathbf{v} are

$$\begin{aligned} \mathbf{v}_A &= \mathbf{V} + \frac{M_B}{M_A + M_B} \mathbf{v} \\ \mathbf{v}_B &= \mathbf{V} - \frac{M_A}{M_A + M_B} \mathbf{v}. \end{aligned}$$

The total kinetic energy then decomposes into the sum

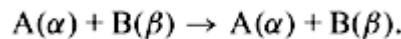
$$\begin{aligned} E &= \frac{1}{2} M_A v_A^2 + \frac{1}{2} M_B v_B^2 = \frac{1}{2} M V^2 + \frac{1}{2} M_{AB} v^2 \\ &= E_{CM}(A + B) + E_{rel}(AB) \end{aligned}$$

of the energy $E_{CM} = \frac{1}{2} M V^2$ of the CM with mass $M = (M_A + M_B)$, and the energy $E_{rel} = \frac{1}{2} M_{AB} v^2$ of relative motion, where the reduced mass M_{AB} is $M_A M_B / (M_A + M_B)$. Let \mathbf{R} be the position of the CM relative to a fixed origin O and \mathbf{r} be the inter-particle separation. The total angular momentum about O similarly decomposes into the sum

$$\begin{aligned} \mathbf{L} &= \mathbf{R} \times M \mathbf{V} + \mathbf{r} \times M_{AB} \mathbf{v} \\ &= \mathbf{L}_{CM}(A + B) + \mathbf{L}_{rel}(AB) \end{aligned}$$

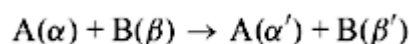
of angular momenta of the CM and of relative motion. For any collision in the absence of any external field, the energy E_{CM} and angular momentum \mathbf{L}_{CM} of the CM are always conserved for all types of collision. The two species, A and B, may be electrons, ions, atoms or molecules, with or without any internal structure and may therefore possess internal energy and angular momentum which must be taken into account. For structured particles E_{rel} and \mathbf{L}_{rel} can change in a collision.

B2.2.2.4 ELASTIC SCATTERING



Elastic scattering involves no permanent changes in the internal structures (states α and β) of A and B. Both the energy E_{rel} and angular momentum $\mathbf{L}_{rel}(AB)$ of relative motion are therefore all conserved.

B2.2.2.5 INELASTIC SCATTERING



Inelastic scattering produces a permanent change in the internal energy and angular momentum state of one or both structured collision partners A and B, which retain their original identity after the collision. For inelastic $i \equiv (\alpha, \beta) \rightarrow f \equiv (\alpha', \beta')$ collisional transitions, the energy $E_{i,f} = \frac{1}{2} M_{AB} v_{i,f}^2$ of relative motion, before (i) and after (f) the collision satisfies the energy conservation condition,

$$E_i + \epsilon_\alpha(\mathbf{A}) + \epsilon_\beta(\mathbf{B}) = E_f + \epsilon_{\alpha'}(\mathbf{A}) + \epsilon_{\beta'}(\mathbf{B})$$

where $\epsilon_{A,B}$ are the internal energies of A and B. The maximum amount of kinetic energy that can be transferred to internal energy is limited to the initial kinetic energy of relative motion, $E_{\text{rel}}(\mathbf{AB}) = \frac{1}{2} M_{AB} v_i^2$.

Excitation implies $\epsilon_i \equiv \epsilon_\alpha(\mathbf{A}) + \epsilon_\beta(\mathbf{B}) < \epsilon_{\alpha'}(\mathbf{A}) + \epsilon_{\beta'}(\mathbf{B}) \equiv \epsilon_f$ de-excitation (or superelastic) implies $\epsilon_f < \epsilon_i$ and energy resonance or excitation transfer implies $\epsilon_i = \epsilon_f$. Changes in angular momentum are limited by the conservation requirement that

$$L_{\text{rel}}(i) + L_\alpha(\mathbf{A}) + L_\beta(\mathbf{B}) = L_{\text{rel}}(f) + L_{\alpha'}(\mathbf{A}) + L_{\beta'}(\mathbf{B})$$

where $L_{\alpha,\beta}$ denotes the internal angular momentum of each isolated species. Collisions, in which only angular momentum is transferred without any energy change, are called *quasi-elastic* collisions.

B2.2.2.6 REACTIVE SCATTERING



Reactive scattering or a chemical reaction is characterized by a rearrangement of the component particles within the collision system, thereby resulting in a change of the physical and chemical identity of the original collision reactants A + B into different collision products C + D. Total mass is conserved. The reaction is exothermic when $E_{\text{rel}}(\mathbf{CD}) > E_{\text{rel}}(\mathbf{AB})$ and is endothermic when $E_{\text{rel}}(\mathbf{CD}) < E_{\text{rel}}(\mathbf{AB})$. A threshold energy is required for the endothermic reaction.

B2.2.2.7 CENTRE-OF-MASS TO LABORATORY CROSS SECTION CONVERSION

Theorists calculate cross sections in the CM frame while experimentalists usually measure cross sections in the laboratory frame of reference. The laboratory (Lab) system is the coordinate frame in which the target particle B is at rest before the collision i.e. $\mathbf{v}_B = 0$. The centre of mass (CM) system (or barycentric system) is the coordinate frame in which the CM is at rest, i.e. $\mathbf{v} = 0$. Since each scattering of projectile A into (ψ, ϕ) is accompanied by a recoil of target B into $(\pi - \psi, \phi + \pi)$ in the CM frame, the cross sections for scattering of A and B are related by

$$\left\{ \frac{d\sigma(\psi, \phi)}{d\Omega} \right\}_{\text{CM}} \equiv \left\{ \frac{d\sigma_A(\psi, \phi)}{d\Omega} \right\}_{\text{CM}} = \left\{ \frac{d\sigma_B(\pi - \psi, \phi + \pi)}{d\Omega} \right\}_{\text{CM}}$$

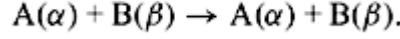
$$\left\{ \frac{d\sigma_B(\psi, \phi)}{d\Omega} \right\}_{\text{CM}} = \left\{ \frac{d\sigma(\pi - \psi, \phi - \pi)}{d\Omega} \right\}_{\text{CM}}$$

In the *Lab* frame, the projectile is scattered by θ_A and the target, originally at rest, recoils through angle θ_B . The number of particles scattered into each solid angle in each frame remains the same, the relative speed v is now v_A and $j_i = N_A v$ in each frame. Hence

$$\left\{ \frac{d\sigma_A(\theta_A, \phi)}{d\Omega_A} \right\}_{\text{Lab}} = \left\{ \frac{d\Omega}{d\Omega_A} \right\} \left[\frac{d\sigma(\psi, \phi)}{d\Omega} \right]_{\text{CM}}$$

$$\left\{ \frac{d\sigma_B(\theta_B, \phi)}{d\Omega_B} \right\}_{\text{Lab}} = \left\{ \frac{d\Omega}{d\Omega_B} \right\} \left[\frac{d\sigma(\psi, \phi)}{d\Omega} \right]_{\text{CM}}.$$

(A) TWO-BODY ELASTIC SCATTERING



The scattering and recoil angles θ_A and θ_B in the Lab frame are related to the CM scattering angle ψ by

$$\tan \theta_A = \frac{\sin \psi}{1 + \gamma \cos \psi} \quad \gamma = M_A/M_B$$

$$\theta_B = \frac{1}{2}(\pi - \psi) \quad 0 \leq \theta_B \leq \frac{1}{2}\pi.$$

The elastic cross sections for scattering and recoil in the Lab-frame are related to the cross section in the CM-frame by

$$\left\{ \frac{d\sigma_A(\theta_A, \phi)}{d\Omega_A} \right\}_{\text{Lab}} = \frac{(1 + \gamma^2 + 2\gamma \cos \psi)^{3/2}}{|1 + \gamma \cos \psi|} \left[\frac{d\sigma(\psi, \phi)}{d\Omega} \right]_{\text{CM}}$$

$$\left\{ \frac{d\sigma_B(\theta_B, \phi)}{d\Omega_B} \right\}_{\text{Lab}} = |4 \sin \frac{1}{2}\psi| \left[\frac{d\sigma(\psi, \phi)}{d\Omega} \right]_{\text{CM}}.$$

(B) TWO-BODY INELASTIC OR REACTIVE SCATTERING PROCESS $A + B \rightarrow C + D$

The energies E_i and E_f of relative motion of A and B and of C and D, respectively satisfy $E_f/E_i = 1 - \epsilon_{fi}/E_i$, where $\epsilon_{fi} = \epsilon_f - \epsilon_i$ is the increase in internal energy. The scattering and recoil angles are

$$\tan \theta_C = \frac{\sin \psi}{(\gamma_C + \cos \psi)} \quad \gamma_C = \left[\frac{M_A M_C}{M_B M_D} \right]^{1/2} \left(\frac{E_i}{E_f} \right)^{1/2}$$

$$\tan \theta_D = \frac{\sin \psi}{(|\gamma_D| - \cos \psi)} \quad \gamma_D = - \left[\frac{M_A M_D}{M_B M_C} \right]^{1/2} \left(\frac{E_i}{E_f} \right)^{1/2}.$$

-7-

The Lab and CM cross sections are then related by

$$\left\{ \frac{d\sigma_j(\theta_j, \phi)}{d\Omega_j} \right\}_{\text{Lab}} = \frac{[1 + 2\gamma_j \cos \psi + \gamma_j^2]^{3/2}}{|1 + \gamma_j \cos \psi|} \left[\frac{d\sigma(\psi, \phi)}{d\Omega} \right]_{\text{CM}}$$

where j denotes C or D. The scattering of a beam from a stationary target is governed by these equations. A crossed beam experiment in which two beams intersect at an angle is not in the Lab-frame. In this case the measured quantities can be similarly transformed [1] to CM for comparison with theoretical calculations.

B2.2.3 MACROSCOPIC RATE COEFFICIENTS

B2.2.3.1 SCATTERING RATE

$$\frac{dN_A}{dt} = -kN_A(t)N_B(t) = -\nu_B N_A(t).$$

A distribution $f_A(\mathbf{v}_A)$ of $N_A(t)$ test particles (cm^{-3}) of species A in a beam collisionally interacts with a distribution $f_B(\mathbf{v}_B)$ of $N_B(t)$ field particles of species B. Collisions with B will scatter A out of the beam at the loss rate ($\text{cm}^{-3} \text{s}^{-1}$)

$$k (\text{cm}^3 \text{s}^{-1}) = \int f_A(\mathbf{v}_A) d\mathbf{v}_A \int [v\sigma(v)] f_B(\mathbf{v}_B) d\mathbf{v}_B$$

The macroscopic rate coefficient $k (\text{cm}^3 \text{s}^{-1})$ for elastic collisions between the ensembles A and B is

$$k (\text{cm}^3 \text{s}^{-1}) = \tilde{\nu}_{AB} \int_0^\infty \sigma(\tilde{\epsilon}_{\text{rel}}) \tilde{\epsilon}_{\text{rel}} \exp(-\tilde{\epsilon}_{\text{rel}}) d\tilde{\epsilon}_{\text{rel}}$$

in terms of the integral cross section $\sigma(v)$ for A–B elastic scattering at relative speed $v = |\mathbf{v}_A - \mathbf{v}_B|$. The microscopic rate coefficient is $v\sigma(v)$. The frequency $\nu_B (\text{s}^{-1})$ of collision between one test particle A with N_B field particles is kN_B .

The rate coefficient for elastic scattering between two species with non-isothermal Maxwellian distributions is then

$$\tilde{\nu}_{AB} = \left[\frac{8k_B}{\pi} \left(\frac{T_A}{M_A} + \frac{T_B}{M_B} \right) \right]^{1/2}$$

-8-

where

$$\tilde{\epsilon}_{\text{rel}} = \frac{1}{2} \frac{M_A M_B v^2}{k_B (M_A T_B + M_B T_A)}.$$

and

$$k(T) = \langle \nu_{AB} \rangle \int_0^\infty \sigma(\epsilon) \epsilon \exp(-\epsilon) d\epsilon \quad (\text{cm}^3 \text{s}^{-1})$$

For isothermal distributions $T_A = T_B = T$, the rate is

$$k(T_c) = \langle \nu_c \rangle \int_0^\infty \sigma(\epsilon_c) \epsilon_c \exp(-\epsilon_c) d\epsilon_c \quad (\text{cm}^3 \text{s}^{-1})$$

where $\epsilon = \frac{1}{2}M_{AB}v^2/k_B T$ and $\langle v_{AB} \rangle = (8k_B T/\pi M_{AB})^{1/2}$. The rate of collisions of electrons A at temperature T_e with a gas of heavy-particles B at temperature T_B is

$$\frac{d}{dt}[N_A \langle \mathcal{E}_{AB} \rangle] = -k_E N_A(t) N_B(t) = -v_{EB} N_A(t)$$

where $\epsilon_e = \frac{1}{2}m_e v^2/k_B T_e$ and $\langle v_e \rangle = (8k_B T_e/\pi m_e)^{1/2}$.

B2.2.3.2 ENERGY TRANSFER RATE

Each of the species A transfers energy ϵ_{AB} to each species B. The amount of energy transferred per unit volume in unit time from ensemble A to ensemble B is

$$k_E = \int f_A(\mathbf{v}_A) d\mathbf{v}_A \int f_B(\mathbf{v}_B) d\mathbf{v}_B \int \mathcal{E}_{AB}(\mathbf{v}_A, \mathbf{v}_B; \psi, \phi) v \left(\frac{d\sigma}{d\Omega} \right) d\Omega.$$

where the macroscopic rate coefficient k_E (energy $\text{cm}^3 \text{s}^{-1}$) for the averaged energy loss $\langle \epsilon_{AB} \rangle$ is

The amount of energy lost in unit time, the energy-loss frequency, is $v_{EB} = k_E N_B(t)$. The energy-loss rate coefficient for two-temperature Maxwellian distributions is

$$k_E(T_A, T_B) = \frac{2M_A M_B}{(M_A + M_B)^2} k_B (T_A - T_B) \tilde{v}_{AB} \int_0^\infty \sigma_D(\tilde{\epsilon}_{\text{rel}}) (\tilde{\epsilon}_{\text{rel}})^2 \exp(-\tilde{\epsilon}_{\text{rel}}) d\tilde{\epsilon}_{\text{rel}}$$

-9-

where $\sigma_D(\tilde{\epsilon}_{\text{rel}})$ is the momentum transfer cross section at reduced energy $\tilde{\epsilon}_{\text{rel}}$. For isothermal distributions, $T_A = T_B$ and the energy rate coefficient k_E of course then vanishes.

B2.2.3.3 TRANSPORT CROSS SECTIONS AND COLLISION INTEGRALS

Transport cross sections are defined for integer $n = 1, 2, 3, \dots$, as

$$\sigma^{(n)}(E) = 2\pi \left[1 - \frac{1 + (-1)^n}{2(n+1)} \right]^{-1} \int_{-1}^{+1} [1 - \cos^n \theta] \frac{d\sigma}{d\Omega}(\cos \theta).$$

The diffusion and viscosity cross sections are given by the transport cross sections $\sigma^{(1)}$ and $\frac{2}{3}\sigma^{(2)}$, respectively.

Collision integrals are defined for integer $s = 0, 1, 2, \dots$, as

$$\begin{aligned} \Omega^{(n,s)}(T) &= [(s+1)!(k_B T)^{s+2}]^{-1} \int_0^\infty \sigma^{(n)}(E) E^{s+1} \exp(-E/k_B T) dE \\ &= [(s+1)!]^{-1} \int_0^\infty \sigma^{(n)}(\epsilon) \epsilon^{s+1} \exp(-\epsilon) d\epsilon \end{aligned}$$

where $\epsilon = \frac{1}{2}M_{AB}v^2/k_B T$. The external factors are chosen so that these expressions for $\sigma^{(n)}$ and $\Omega^{(n,s)}$ reduce to πd^2 for classical rigid spheres of diameter d . The rate coefficient k ($\text{cm}^3 \text{s}^{-1}$) for scattering can then be expressed, in terms of the collision integral, as equal to $\bar{v}_{AB}\Omega^{(0,0)}$. The amount of energy lost per cm^3 per second by collision can be expressed in terms of $\Omega^{(1,1)}$. Tables of transport cross sections and collision integrals for $(n, 6, 4)$ ion–neutral interactions are available [2, 3].

(C) CHAPMAN–ENSKOG MOBILITY FORMULA

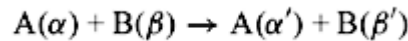
When ions move under equilibrium conditions in a gas and an external electric field, the energy gained from the electric field \mathbf{E} between collisions is lost to the gas upon collision so that the ions move with a constant drift speed $\mathbf{v}_d = K \mathbf{E}$. The mobility K of ions of charge e in a gas of density N is given in terms of the collision integral by the Chapman–Enskog formula [2]

$$K = \frac{3e}{16N} \left(\frac{\pi}{2Mk_B T} \right)^{1/2} [\Omega^{(1,1)}(T)]^{-1}.$$

B2.2.4 QUANTAL TRANSITION RATES AND CROSS SECTIONS

B2.2.4.1 MICROSCOPIC RATE OF TRANSITIONS

In the general elastic/inelastic collision process



the external scattering or deflection of a beam of projectile particles A (electrons, ions, atoms) by target particles B (atoms, molecules) is accompanied by transitions (electronic, vibrational, rotational) within the internal structure of either or both collision partners. For a beam with incident momentum $\mathbf{p}_i = \hbar \mathbf{k}_i$ in the range $(\mathbf{p}_i, \mathbf{p}_i + d\mathbf{p}_i)$ or directed energy $E_i \equiv (E_i, \hat{\mathbf{p}}_i)$ in the range $(E_i, E_i + dE_i)$, the translational states representing the A–B relative or external motion undergo free–free transitions $(E_i, E_i + dE_i) \rightarrow (E_f, E_f + dE_f)$ within the translational continuum, while the structured particles undergo bound–bound (excitation, de-excitation, excitation transfer) or bound–free (ionization, dissociation) transitions $i \equiv (\alpha, \beta) \rightarrow f \equiv (\alpha', \beta')$ in their internal electronic, vibrational or rotational structure. The transition frequency (s^{-1}) for this collision is

$$\frac{dW_{if}}{d\hat{\mathbf{p}}_f} (\text{s}^{-1}) = \frac{2\pi}{\hbar} \frac{1}{g_i} \sum_{i,f} |V_{fi}|^2 \rho_f(E_f)$$

which is an average over the g_i initial degenerate internal states i and a sum over all g_f final degenerate internal states f of the isolated systems A and B. It is therefore the probability per unit time for scattering from a specified E_i —(external) continuum state into unit solid angle $d\hat{\mathbf{p}}_f$ accompanied by a transition from any one of the g_i initial states (α, β) to all final internal states (α', β') of degeneracy g_f and to all final translational states $\rho_f(E_f) dE_f$ of relative motion consistent with energy conservation. The double summation $\sum_{i,f}$ is over the g_i initial and g_f final internal states of A and B with total energy ϵ_i and ϵ_f respectively.

Check: The dimension of $[|V_{ij}|^2 \rho]$ is E , $[\hbar] = Et$ so that $dW_{ij}/d\hat{\mathbf{p}}_f$ indeed has the correct dimension of t^{-1} .

(A) INTERACTION MATRIX ELEMENT

The matrix element

$$V_{fi} = \langle N_f \Phi_f | V(\mathbf{r}_A, \mathbf{r}_B, \mathbf{R}) | N_i \Psi_i^+ \rangle_{\mathbf{r}, \mathbf{R}} = V_{if}^*$$

-11-

is an integration over the internal coordinates $\mathbf{r} \equiv \mathbf{r}_A, \mathbf{r}_B$ of the electrons of A and B and over the channel vector \mathbf{R} for A–B relative motion. The matrix element of the mutual electrostatic interaction $V(\mathbf{r}_A, \mathbf{r}_B, \mathbf{R})$ couples the eigenfunction $N_i \Psi_i^+(\mathbf{R}, \mathbf{r}_A, \mathbf{r}_B)$ of $[\hat{H}_{\text{rel}} + \hat{H}_{\text{int}} + V]$ of $[\hat{\mathbf{H}}_{\text{rel}} + \hat{\mathbf{H}}_{\text{int}} + V]$ for the complete collision system for all \mathbf{R} to the final $R \rightarrow \infty$ asymptotic state $N_f \Phi_f(\mathbf{R}, \mathbf{r}_A, \mathbf{r}_B)$, which is an eigenfunction only of the unperturbed Hamiltonian $[\hat{\mathbf{H}}_{\text{rel}} + \hat{\mathbf{H}}_{\text{int}}]$. The wavefunction

$$\begin{aligned} \Psi_i^+(\mathbf{R}, \mathbf{r}_A, \mathbf{r}_B) &= \sum_j \Phi_j^+(\mathbf{R}) \psi_\alpha(\mathbf{r}_A) \phi_\beta(\mathbf{r}_B) \\ &\equiv \sum_j \Phi_j^+(\mathbf{R}) \psi_j^{\text{int}}(\mathbf{r}) \end{aligned}$$

for the full collision system with Hamiltonian $\hat{\mathbf{H}}_{\text{rel}} + \hat{\mathbf{H}}_{\text{int}} + V$ tends at asymptotic R to

$$\Psi_i^+ \sim \sum_j \left[e^{i\mathbf{k}_j \cdot \mathbf{R}} \delta_{ij} + f_{ij}(\theta, \phi) \frac{e^{i\mathbf{k}_j \cdot \mathbf{R}}}{R} \right] \psi_j^{\text{int}}(\mathbf{r})$$

which represents an incoming plane wave of unit amplitude in the incident elastic channel i and an outgoing spherical waves of amplitude f_{ij} in all channels j , including i . The Kronecker symbol means $\delta_{ij} = 1, i = j$ and $\delta_{ij} = 0, i \neq j$. The final state at infinite separation R is

$$\Phi_f(\mathbf{R}, \mathbf{r}_A, \mathbf{r}_B) = e^{i\mathbf{k}_f \cdot \mathbf{R}} \psi_{\alpha'}(\mathbf{r}_A) \phi_{\beta'}(\mathbf{r}_B) \equiv e^{i\mathbf{k}_f \cdot \mathbf{R}} \psi_f^{\text{int}}(\mathbf{r})$$

which is an eigenfunction only of $\hat{\mathbf{H}}_{\text{rel}} + \hat{\mathbf{H}}_{\text{int}}$. The plane wave of unit amplitude describes the external relative motion with Hamiltonian $\hat{\mathbf{H}}_{\text{rel}}$ and $\phi_{\alpha\beta}(\mathbf{r}_A) \psi_{\beta\alpha}(\mathbf{r}_B)$ describes the internal, isolated, normalized atomic eigenstates of A and B with internal Hamiltonian $\hat{\mathbf{H}}_{\text{int}}$. The factors $N_{i,f}$ provide the possibility of having translational (scattering) states with arbitrary amplitudes which are not necessarily unity.

(B) TRANSITION OPERATOR

The interaction matrix element can also be written as

$$V_{fi} = \langle N_f \Phi_f | \hat{T} | N_i \Phi_i \rangle$$

where the transition operator, \hat{T} , is defined by $\hat{T}\Phi = V\Psi$. The transition operator \hat{T} therefore couples states which are eigenfunctions of the same unperturbed Hamiltonian $\hat{\mathbf{H}}_{\text{rel}} + \hat{\mathbf{H}}_{\text{int}}$, in contrast to V which couples

states Ψ_i^+ and Φ_f belonging to different Hamiltonians.

-12-

B2.2.4.2 DETAILED BALANCE BETWEEN RATES

The frequency (number per second) of $i \rightarrow f$ transitions from all g_i degenerate initial internal states and from the $\rho_i d\mathbf{E}_i$ initial external translational states is equal to the reverse frequency from the g_f degenerate final internal states and the $\rho_f d\mathbf{E}_f$ final external translational states. The *detailed balance relation* between the forward and reverse frequencies is therefore

$$[g_i \rho_i d\mathbf{E}_i d\hat{\mathbf{p}}_i] \left(\frac{dW_{if}}{d\hat{\mathbf{p}}_f} \right) d\hat{\mathbf{p}}_f = [g_f \rho_f d\mathbf{E}_f d\hat{\mathbf{p}}_f] \left(\frac{dW_{fi}}{d\hat{\mathbf{p}}_i} \right) d\hat{\mathbf{p}}_i$$

since $V_{if} = V_{fi}^*$. From energy conservation $\epsilon_i + E_i = \epsilon_f + E_f$ then $dE_i = dE_f$. The differential frequencies

$$\frac{dR_{if}}{d\hat{\mathbf{p}}_f} \equiv g_i \rho_i \frac{dW_{if}}{d\hat{\mathbf{p}}_f} = g_f \rho_f \frac{dW_{fi}}{d\hat{\mathbf{p}}_i} \equiv \frac{dR_{fi}}{d\hat{\mathbf{p}}_i}$$

for the forward and reverse transitions, $i \rightleftharpoons f$, are therefore equal.

B2.2.4.3 ENERGY DENSITY OF CONTINUUM STATES

The continuum wavefunctions $\phi_p(\mathbf{R})$ for the states of the $A-B$ relative motion satisfy the orthonormality condition

$$\int \rho(\mathbf{E}) d\mathbf{E} \int \phi_p(\mathbf{R}) \phi_{p'}^*(\mathbf{R}) d\mathbf{R} = 1.$$

The number of translational states per unit volume $d\mathbf{R}$ with directed energies $\mathbf{E} \equiv (E, \hat{\mathbf{p}})$ in the range $[\mathbf{E}, \mathbf{E} + d\mathbf{E}]$ is $\rho(\mathbf{E}) d\mathbf{E}$. This orthonormality condition for continuum states is analogous to the condition $\sum_j |\langle \phi_j | \phi_i \rangle|^2 = 1$ for bound states. For plane waves, $\phi_p(\mathbf{R}) = N \exp(i\mathbf{p} \cdot \mathbf{R}/\hbar)$, then

$$\begin{aligned} \langle \phi_{p'} | \phi_p \rangle &= |N|^2 (2\pi\hbar)^3 \delta(\mathbf{p} - \mathbf{p}') = |N|^2 (2\pi)^3 \delta(\mathbf{k} - \mathbf{k}') \\ &= |N|^2 \frac{(2\pi\hbar)^3}{mp} \delta(E - E') = \frac{1}{\rho(\mathbf{E})} \delta(E - E'). \end{aligned}$$

Note, irrespective of the method chosen to normalize the wavefunctions, that

$$|N|^2 \rho(\mathbf{E}) = \frac{mp}{(2\pi\hbar)^3}$$

always. The amplitude $|N|$ does, however, depend on the choice of normalization.

-13-

(1) For momentum normalized states, $\langle \phi_{\mathbf{p}} | \phi_{\mathbf{p}'} \rangle = \delta(\mathbf{p} - \mathbf{p}')$, $|N| = (2\pi\hbar)^{-3/2}$ and the density of states $\rho(\mathbf{E}) = mp$.

(2) For wavevector normalized states, $\langle \phi_{\mathbf{p}} | \phi_{\mathbf{p}'} \rangle = \delta(\mathbf{k} - \mathbf{k}')$, $|N| = (2\pi)^{-3/2}$ and the density of states $\rho(\mathbf{E}) = (mp/\hbar^3)$.

(3) For energy-normalized states, $\langle \phi_{\mathbf{p}} | \phi_{\mathbf{p}'} \rangle = \delta(E - E')$, $\rho(\mathbf{E}) = 1$ and $|N| = (mp/\hbar^3)^{1/2}$.

(4) For waves with unit amplitude, $|N| = 1$ and $\rho(\mathbf{E}) = (mp/\hbar^3)$.

Note that $\langle \phi_{\mathbf{p}} | \phi_{\mathbf{p}'} \rangle \rho(\mathbf{E}) d\mathbf{E}$ is dimensionless for all cases and yields unity for a single particle when integrated over all \mathbf{E} . The number of states in the phase-space element $d\mathbf{E} d\mathbf{R}$ is

$$dn = |N|^2 \rho(\mathbf{E}) d\mathbf{E} d\mathbf{R} = d\mathbf{p} d\mathbf{R} / (2\pi\hbar)^3$$

i.e. each translational state occupies a cell of phase volume $(2\pi\hbar)^3$. The density of states in the interval $[E, E + dE]$ is $\rho(E) = 4\pi\rho(\mathbf{E})$. The number of translational states per unit volume with energy in the scalar range $[E, E + dE]$ is

$$|N|^2 \rho(\mathbf{E}) d\mathbf{E} = \frac{2}{\sqrt{\pi}} \frac{(2\pi m)^{3/2}}{h^3} E^{1/2} dE.$$

Check. The number of free particles with all momenta \mathbf{p} in equilibrium with a gas bath of volume v at temperature T is the translational partition function Z_t . Since the fraction of particles with energy E is $\exp(-E/k_B T)/Z_t$, the *Maxwell distribution*

$$\begin{aligned} f_M(E) dE &= \frac{|N|^2 v \rho(\mathbf{E}) d\mathbf{E}}{Z_t} \exp(-E/k_B T) \\ &= \frac{2}{\sqrt{\pi}} (E/k_B T)^{1/2} \exp(-E/k_B T) d(E/k_B T) \end{aligned}$$

is then recovered.

CURRENT

Current is the number of particles crossing unit area in unit time. The current in a beam with directed energy E within the range $(E, E + dE)$ is

$$j dE = v |N|^2 \rho(\mathbf{E}) d\mathbf{E} = (p^2/h^3) dE.$$

-14-

The current per unit dE is the *current density* $j = (p^2/h^3)$. The *quantal expression for current*

$$\mathbf{J} \equiv \frac{\hbar}{2mi} [\phi_{\mathbf{p}}^* \nabla \phi_{\mathbf{p}} - \phi_{\mathbf{p}} \nabla \phi_{\mathbf{p}}^*]$$

when applied to the plane wave $\phi_{\mathbf{p}} = N \exp(i\mathbf{p} \cdot \mathbf{R}/\hbar)$, gives $\mathbf{j} = |N|^2 \mathbf{v}$. The current in a $(E, E + dE)$ -beam of

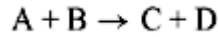
plane waves is then $j[\rho(\mathbf{E})d\mathbf{E}]$ so that the current density is $j(\mathbf{E}) = J\rho(\mathbf{E}) = |N|^2 v\rho(\mathbf{E})$, as before.

B2.2.4.4 INELASTIC CROSS SECTIONS

The differential cross section $d\sigma_{if}/d\hat{\mathbf{p}}_f$ for $i \rightarrow f$ transitions from any one of the g_i initial states is defined as $[dR_{if}/d\hat{\mathbf{p}}_f]/g_i j_i$, the transition frequency per unit incident current. Since current is the number of particles crossing unit area in unit time, the cross section is therefore the effective area presented by the target towards $i \rightarrow f$ internal transitions in the internal structures of the collision partners which are scattered into unit solid angle $d\hat{\mathbf{p}}_f$ about direction $\hat{\mathbf{p}}_f$ in the CM-frame.

(A) BASIC EXPRESSION FOR CROSS SECTION

The differential cross section for



collisions is therefore defined as

$$\frac{d\sigma_{if}}{d\hat{\mathbf{p}}_f} = \frac{1}{g_i j_i} \frac{dR_{if}}{d\hat{\mathbf{p}}_f} = \frac{2\pi}{\hbar} \left(\frac{\rho_i \rho_f}{j_i} \right) \frac{1}{g_i} \sum_{i,f} |\langle N_f \Phi_f | V(\mathbf{r}_A, \mathbf{r}_B, \mathbf{R}) | N_i \Psi_i^+ \rangle|^2$$

which is an average over the g_i initial internal degenerate states and a sum over the g_f final degenerate states. Since $j_i = |N_i|^2 v_i \rho_i = p_i^2/h^3$, an alternative form [4] for the cross section is

$$\frac{d\sigma_{if}}{d\hat{\mathbf{p}}_f} = \frac{2\pi}{\hbar v_i} \left(\frac{\rho_f}{g_i} \right) \sum_{i,f} |\langle N_f \Phi_f | V(\mathbf{r}_A, \mathbf{r}_B, \mathbf{R}) | \Psi_i^+ \rangle|^2.$$

B2.2.4.5 DETAILED BALANCE BETWEEN CROSS SECTIONS

When cast in terms of cross sections, the detailed balance relation in [section B2.2.4.2](#) is

$$g_i j_i(E_i) \frac{d\sigma_{if}}{d\hat{\mathbf{p}}_f} = g_f j_f(E_f) \frac{d\sigma_{fi}}{d\hat{\mathbf{p}}_i}.$$

-15-

The basic relationship satisfied by the differential cross sections for the forward and reverse $i \rightleftharpoons f$ transitions is

$$g_i p_i^2 \frac{d\sigma_{if}(E_i)}{d\hat{\mathbf{p}}_f} = g_f p_f^2 \frac{d\sigma_{fi}(E_f)}{d\hat{\mathbf{p}}_i}.$$

(A) COLLISION STRENGTHS

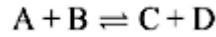
Collision strengths Ω_{if} exploit this detailed balance relation by being defined as

$$\Omega_{if} = g_i p_i^2 \sigma_{if}(E_i) = g_f p_f^2 \sigma_{fi}(E_f) = \Omega_{fi}.$$

They are therefore symmetrical in i and f .

(B) REACTIVE PROCESSES

For any reactive process



the detailed balance relations involving differential/integral cross sections are

$$g_A g_B p_{AB}^2 \left[\frac{d\sigma_{if}(E_{AB})}{d\hat{p}_{CD}} \right] = g_C g_D p_{CD}^2 \left[\frac{d\sigma_{fi}(E_{CD})}{d\hat{p}_{AB}} \right]$$

$$g_A g_B p_{AB}^2 \sigma_{if}(E_{AB}) = g_C g_D p_{CD}^2 \sigma_{fi}(E_{CD})$$

where $p_{JK}^2 = 2M_{JK}E_{JK}$, in terms of the reduced mass M_{JK} and relative energy E_{JK} of species J and K.

B2.2.4.6 EXAMPLES OF DETAILED BALANCE

(A) EXCITATION-DE-EXCITATION

$$e^-(E_i) + A_i \rightleftharpoons e^-(E_f) + A_f$$

$$\sigma_{if}(E_i) = \left(\frac{g_f}{g_i} \right) \left(\frac{E_f}{E_i} \right) \sigma_{fi}(E_f).$$

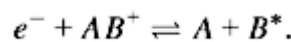
-16-

With energy conservation, $E_i = E_f + (\epsilon_f - \epsilon_i) \equiv E_f + \epsilon_{fi}$ the cross section for superelastic collisions ($E_f > E_i$) can be obtained from σ_{if} at energy E_i via the relation

$$\sigma_{fi}(E_i - \epsilon_{fi}) = \left(1 - \frac{\epsilon_{fi}}{E_i} \right)^{-1} \left(\frac{g_i}{g_f} \right) \sigma_{if}(E_i).$$

(B) DISSOCIATIVE RECOMBINATION/ASSOCIATIVE IONIZATION

Dissociative recombination and associative ionization are represented by the forward and backward directions of



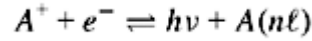
The respective cross sections σ_{DR} and σ_{AI} are related by

$$\sigma_{DR}(E_e) = \left(\frac{g_A g_B}{g_e g_{AB^+}} \right) \left(\frac{k_{AB}^2}{k_e^2} \right) \sigma_{AI}(E_{AB})$$

where the statistical weight of each species j involved is denoted by g_j .

(C) RADIATIVE RECOMBINATION/PHOTOIONIZATION

Similarly, the cross sections for radiative recombination (RR) and for photoionization (PI), the forward and reverse directions of



are related by

$$\sigma_{\text{RR}}(E_c) = \left(\frac{g_A g_\nu}{g_e g_{A^+}} \right) \left(\frac{p_\nu^2}{p_e^2} \right) \sigma_{\text{PI}}(h\nu).$$

The photon statistical weight is $g_\nu = 2$, corresponding to the two directions of polarization of the photon. The photon energy E is related to its momentum p_ν and wavenumber k_ν and to the ionization energy $I_{n\ell}$ of the atom $A(n\ell)$ by

$$E = h\nu = p_\nu c = \hbar k_\nu c = I_{n\ell} + E_c$$

-17-

where c is the speed of light. This ratio is

$$\frac{p_\nu^2}{p_e^2} = \frac{(h\nu)^2}{(2E_c m_e c^2)} = \frac{\alpha^2 (h\nu)^2}{2E_c \epsilon_0}.$$

B2.2.4.7 FOUR USEFUL EXPRESSIONS FOR THE CROSS SECTION

The final expressions to be used for the calculation of cross sections depend on the particular choice of normalization of the continuum wavefunction for relative motion. Since it is often a vexing problem and is a continued source of confusion and error in the literature, these final expressions are worked out below. The external relative-motion part of the system wavefunction $N\Psi^+ \equiv \Psi_p(\mathbf{R}, \mathbf{r}_A, \mathbf{r}_B)$ is $N\phi_p(\mathbf{R})$. Since $|N|^2 \rho = mp/\hbar^3$, the density $\rho(\mathbf{E})$ of continuum states therefore depends on the choice of normalization factor N adopted for the continuum wave. For future reference, the amplitude N and the energy densities $\rho(\mathbf{E})$ associated with four common methods adopted for normalization of continuum waves are summarized in table B2.2.1. Also included is the amplitude N_ℓ of the corresponding radial partial wave

$$R_{\ell\ell}(r) \sim \frac{N_\ell}{r} \sin \left(kr - \frac{1}{2} \ell\pi + \eta_\ell \right)$$

of section B2.2.6.1. The external multiplicative factors $\gamma_{if} = (2\pi/\hbar)(\rho_i \rho_f / j_i)$ in the basic formula in section B2.2.4.4 for the cross section are also summarized in table B2.2.1 for the various normalization schemes. The reduced masses before and after the collision are $m_i = M_A M_B / (M_A + M_B)$ and $m_f = M_C M_D / (M_C + M_D)$, respectively.

(A) ENERGY-NORMALIZED INITIAL AND FINAL STATES

The wavefunctions

$$\chi_p = \rho^{1/2} N \Psi = (m p / 2\pi \hbar^3)^{1/2} \Psi_p(\mathbf{R}, \mathbf{r}_A, \mathbf{r}_B)$$

are energy-normalized according to

$$\langle \chi_p | \chi_{p'} \rangle = \delta(\mathbf{E} - \mathbf{E}').$$

The basic formula in [section B2.2.4.4](#) with $j_i = p_i^2 / (2\pi \hbar)^3$ yields

$$\frac{d\sigma_{if}}{d\hat{p}_f} = \frac{\pi}{k_i^2} \frac{1}{g_i} |T_{if}|^2 = \left(\frac{\hbar^2}{8\pi m_i E_i} \right) \frac{1}{g_i} |T_{if}|^2.$$

-18-

The transition probability is

$$P_{if} = |T_{if}|^2 = \sum_{i,f} \int |2\pi \langle \tilde{\chi}_f | V | \chi_i^+ \rangle|^2 d\hat{k}_i$$

the magnitude squared of the element T_{if} of the transition matrix T between χ_i^+ and $\tilde{\chi}_f$, the two energy-normalized eigenfunctions of $\hat{H}_{\text{rel}} + \hat{H}_{\text{int}} + V$ and $\hat{H}_{\text{rel}} + \hat{H}_{\text{int}}$, respectively. The detailed balance relation in this case is simply

$$|T_{if}|^2 = |T_{fi}|^2$$

thereby verifying that $|T_{if}|^2$ is indeed the $i \rightarrow f$ transition probability for transitions between all g_i initial and g_f final states. This type of normalization is convenient for *rearrangement collisions* such as *dissociative*, *radiative* and *dielectronic recombination*.

(B) UNIT AMPLITUDE INITIAL AND FINAL STATES

Here the initial and final wavefunctions with unit amplitude are Ψ_i^+ and Φ_f . They are each normalized according to

$$\langle \Psi_{p'} | \Psi_p \rangle = (2\pi)^3 \delta(\mathbf{p} - \mathbf{p}').$$

The basic expression in [section B2.2.4.4](#) with $j_i = |N_i|^2 v_i \rho_i$ and $|N_f|^2 \rho_f = m_f p_f / \hbar^3$ reduces to

$$\frac{d\sigma_{if}}{d\hat{p}_f} = \frac{v_f}{v_i} |f_{if}(\theta, \varphi)|^2$$

where the scattering amplitude is

$$f_{if} = -\frac{1}{4\pi}(2m_f/\hbar^2)\langle\Phi_f|V|\Psi_i^+\rangle$$

which couples scattering states Ψ_i^+ and Φ_f of unit amplitude. This expression is also applicable for rearrangement collisions $A + B \rightarrow C + D$ by including the reduced mass $m_f = M_C M_D / (M_C + M_D)$ of the reacted species after the collision. The integral cross section consistent with the above scattering amplitude is,

$$\sigma_{if}(E) = \frac{v_f}{v_i} \int_0^\pi d(\cos\theta) \int_0^{2\pi} |f_{if}(\theta, \varphi)|^2 d\varphi$$

-19-

at relative energy $E = k_i^2 \hbar^2 / 2M_{AB}$. The scattering amplitude consistent with the common use of

$$\sigma_{if}(E) = \frac{k_f}{k_i} \int_0^\pi d(\cos\theta) \int_0^{2\pi} |\tilde{f}_{if}(\theta, \varphi)|^2 d\varphi$$

for rearrangement collisions is

$$\tilde{f}_{if} = -\frac{1}{4\pi}(2\sqrt{m_i m_f}/\hbar^2)\langle\Phi_f|V|\Psi_i^+\rangle.$$

Both conventions are identical only for direct collisions $A(\alpha) + B(\beta) \rightarrow A(\alpha') + B(\beta')$. This normalization is customary [5] for *elastic* and *inelastic* scattering processes.

For symmetrical potentials $V(r)$ scattering is confined to a plane and f_{ij} depends only on scattering angle $\theta = \hat{\mathbf{k}}_i \cdot \hat{\mathbf{k}}_f$.

(C) MOMENTUM-NORMALIZED INITIAL AND FINAL STATES

Here the initial and final wavefunctions $\xi_{\mathbf{p}} = (2\pi\hbar)^{-3/2}\Psi_i^+$ and $\xi_{\mathbf{p}'} = (2\pi\hbar)^{-3/2}\Phi_f$ are normalized according to

$$\langle\xi_{\mathbf{p}'}|\xi_{\mathbf{p}}\rangle = \delta(\mathbf{p} - \mathbf{p}').$$

The cross section B2.2.4.4 is then

$$\frac{d\sigma_{if}}{d\hat{\mathbf{p}}_f} = \frac{v_f}{v_i} |f_{if}(\theta, \varphi)|^2$$

where the scattering amplitude [6, 7 and 8] is now

$$f_{if} = -(2\pi^2\hbar^3)(2m_f/\hbar^2)|\langle\xi_f|V|\xi_i^+\rangle|.$$

(D) ENERGY-NORMALIZED FINAL AND UNIT AMPLITUDE INITIAL STATES

Here the basic formula B2.2.4.4 yields

$$\frac{d\sigma_{if}}{d\hat{p}_f} = \frac{2\pi}{\hbar v_i} |\langle \tilde{\chi}_f | V | \Phi_i^+ \rangle|^2$$

-20-

which couples the initial scattering state Ψ_i^+ of unit amplitude with the energy-normalized final state $\tilde{\chi}_f = \rho_f^{1/2} N_f \Phi_f$. This normalization is customary for photoionization problems.

B2.2.5 BORN CROSS SECTIONS

Here an (undistorted) plane wave of unit amplitude is adopted for the channel wavefunction $\Psi_i^+ = \sum_j \Phi_j(\mathbf{R}) \psi_j^{\text{int}}(\mathbf{r})$ for the complete system. The differential cross section for elastic ($i = f$) or inelastic scattering ($i \neq f$) into $\hat{\mathbf{k}}_f(\theta, \phi)$ is then

$$\frac{d\sigma_{if}}{d\Omega} = \frac{v_f}{v_i} |f_{if}(\theta)|^2.$$

The Born scattering amplitude for A–B collisions is

$$f_{if}^{(B)}(\mathbf{K}) = -\frac{1}{4\pi} \frac{2M_{AB}}{\hbar^2} \int V_{fi}(\mathbf{R}) \exp(i\mathbf{K} \cdot \mathbf{R}) d\mathbf{R}$$

which is the Fourier transform of the interaction potential

$$V_{fi}(\mathbf{R}) = \langle \psi_f^{\text{int}}(\mathbf{r}) | V(\mathbf{r}, \mathbf{R}) | \psi_i^{\text{int}}(\mathbf{r}) \rangle$$

which couples the initial and final isolated states $\psi_j^{\text{int}}(\mathbf{r}) = \phi_j(\mathbf{r}_A) \psi_j(\mathbf{r}_B)$ of the atoms. The diagonal potential $V_{ii}(\mathbf{R})$ is the static interaction for elastic scattering. The Born scattering amplitude is a pure function only of the collisional momentum change

$$\mathbf{q} = \hbar \mathbf{K} = M_{AB}(\mathbf{v}_i - \mathbf{v}_f) = \hbar(\mathbf{k}_i - \mathbf{k}_f)$$

where \mathbf{v} is the A–B relative velocity. Since $K^2 = k_i^2 + k_f^2 - 2k_i k_f \cos\theta$, the *Born integral cross section* is

$$\begin{aligned} \sigma_{if}^B(k_i) &= \frac{2\pi}{M_{AB}^2 v_i^2} \int_{q_-}^{q_+} |f_{if}^{(B)}(q)|^2 q dq \\ &= \frac{2\pi}{(k_i a_0)^2} \int_{K_- a_0}^{K_+ a_0} |f_{if}^{(B)}(K)|^2 (K a_0) d(K a_0) \end{aligned}$$

where $q_{\pm} = \hbar K_{\pm} = \hbar|k_i \pm k_f|$ are the maximum and minimum momentum changes consistent with energy conservation. For symmetric interactions $V_{fi}(R)$, then

$$f_{if}^B(K) = -\frac{2M_{AB}}{\hbar^2} \int V_{fi}(R) \frac{\sin KR}{KR} R^2 dR.$$

B2.2.5.1 FERMİ GOLDEN RULES

Rule A. The transition rate (probability per unit time) for a transition from state Φ_i of a quantum system to a number $\rho(E) dE$ of continuum states Φ_E by an *external* perturbation V is

$$w_{if} = \frac{2\pi}{\hbar} |\langle \Phi_E | V | \Phi_i \rangle|^2 \rho_f(E) \equiv \frac{2\pi}{\hbar} |V_{i\epsilon}|^2 \rho_f(E)$$

to first order in V . Since $|V_{i\epsilon}|^2 \rho_f$ has the dimension of energy, w_{if} has the dimension t^{-1} .

Rule B. When the direct coupling $V_{i\epsilon}$ from only the initial state to the continuum vanishes, but the coupling $V_{n\epsilon} \neq 0$ for $n \neq i$, the transition can then occur via the intermediate states n at the rate

$$w_{if} = \frac{2\pi}{\hbar} \sum_n \left| \frac{V_{in} V_{n\epsilon}}{E - E_n} \right|^2 \rho_f(E).$$

These rules, A and B (which are not exact) are useful for both scattering and radiative processes and are often referenced as Fermi's Rules 2 and 1, respectively.

SCATTERING EXAMPLE

The cross section for inelastic scattering of beam of particles by potential $V(\mathbf{r}, \mathbf{R})$ is

$$\frac{d\sigma_{if}}{d\hat{\mathbf{p}}_f} = \frac{w_{if}}{J_i}.$$

A plane-wave monoenergetic beam, $\Phi_{i,f} = N_i, f \exp(i\mathbf{p}_{i,f} \cdot \mathbf{R}/\hbar) \phi_{i,f}(\mathbf{r})$ has current $j_i = |N_i|^2 v_i$ and density determined from $|N_f|^2 \rho_f(E) = M_{AB} p_f / (2\pi\hbar)^3$. Hence

$$\frac{d\sigma}{d\hat{\mathbf{p}}_f} = \frac{v_f}{v_i} \left| \frac{1}{4\pi} \frac{2M_{AB}}{\hbar^2} \int V_{fi}(\mathbf{r}) e^{i(\mathbf{p}_i - \mathbf{p}_f) \cdot \mathbf{r}/\hbar} d\mathbf{r} \right|^2.$$

Since this agrees with the first Born differential cross section for (in)elastic scattering, Fermi's Rule 2 is therefore valid to first order in the interaction V .

B2.2.5.2 ION (ELECTRON)–ATOM COLLISIONS

The electrostatic interaction between a structureless projectile ion P of charge $Z_P e$ and an atom A with nuclear charge $Z_A e$ is

$$V(\mathbf{r}, \mathbf{R}) = \frac{Z_A Z_P e^2}{R} - \sum_{j=1}^{N_A} \frac{Z_P e^2}{|\mathbf{R} - \mathbf{r}_j|}.$$

With the use of Bethe's integral

$$\int \frac{e^{i\mathbf{K} \cdot \mathbf{R}}}{|\mathbf{R} - \mathbf{r}_j|} d\mathbf{R} = \frac{4\pi}{K^2} e^{i\mathbf{K} \cdot \mathbf{r}_j}$$

the Born scattering amplitude (see B2.2.5) reduces to

$$|f_{if}^{(B)}(q)|^2 = \frac{4M_{PA}^2 Z_P^2 e^4}{q^4} |Z_A \delta_{if} - F_{if}^A(q)|^2$$

which is a function only of momentum transfer $q = \hbar K$. The dimensionless inelastic *form factor* for $i \rightarrow f$ inelastic transitions between states $\phi_{i,f}$ of atom A with Z_A electrons is defined as

$$F_{fi}^A(q) = \langle \phi_f(\mathbf{r}) | \sum_{j=1}^{Z_A} e^{i\mathbf{q} \cdot \mathbf{r}_j / \hbar} | \phi_i(\mathbf{r}) \rangle$$

where the integration is over all electron positions denoted collectively by $\mathbf{r} \equiv \mathbf{r}_j$. The integrated cross section is

$$\begin{aligned} \sigma_{if}(v_{PA}) &= \frac{8\pi Z_P^2 e^4}{v_{PA}^2} \int_{q_-}^{q_+} |Z_A \delta_{if} - F_{if}^A(q)|^2 \frac{dq}{q^3} \\ &= \frac{8\pi Z_P^2 a_0^2}{(v_{PA}/v_0)^2} \int_{K_- a_0}^{K_+ a_0} |Z_A \delta_{if} - F_{if}^A(K)|^2 \frac{d(Ka_0)}{(Ka_0)^3} \end{aligned}$$

-23-

where $v_0 = e^2/\hbar$ is the atomic unit (au) of velocity. The dimensionless momentum change $q/m_e v_0$ is Ka_0 . In the heavy-particle or high-energy limit, $q_+ \rightarrow \infty$ and

$$q_- \approx \frac{|\Delta E_{fi}|}{v_{PA}} \left[1 + \frac{\Delta E_{fi}}{2M_{PA} v_{PA}^2} \right]$$

where $\Delta E_{fi} = E_f - E_i$ is the energy lost by the projectile. Since

$$f_{if}^B(q) = f_C^{Z_P Z_A}(q) \delta_{if} + f_C^{Z_P}(q) F_{if}(q)$$

can be expressed in terms of the individual two-body amplitudes $f_{\mathbf{c}}^{z_1 z_2}$ for Coulomb elastic scattering between particles of charges z_1 and z_2 , the Born cross section for inelastic collisions can be written [9, 11, 28] in the useful form

$$\sigma_{if}^B(v_i) = \frac{2\pi}{M_{\text{eP}}^2 v_i^2} \int_{q_-}^{q_+} P_{fi}(q) \left(\frac{d\sigma}{d\Omega} \right)_{\text{el}} q \, dq$$

where $P_{fi}(q) = |F_{fi}^A(q)|^2$ is the transition probability for which the impulsive transfer of momentum q to atom A and where

$$\left(\frac{d\sigma}{d\Omega} \right)_{\text{el}} = \frac{4M_{\text{eP}}^2 Z_p^2 e^4}{q^4}$$

is the differential cross section for elastic (Coulomb) scattering with momentum q transferred from the projectile of charge $Z_p e$ to one electron of atom A.

B2.2.5.3 ATOM-ATOM COLLISIONS

The Born integral cross section for specific $(\alpha\beta) \rightarrow (\alpha'\beta')$ transitions in the collision



in terms of the atomic form factors is

$$\sigma_{\alpha\alpha'}^{\beta\beta'}(v_i) = \frac{8\pi a_0^2}{(v_i/v_0)^2} \int_{K-a_0}^{K+a_0} |Z_A \delta_{\alpha\alpha'} - F_{\alpha\alpha'}^A(K)|^2 |Z_B \delta_{\beta\beta'} - F_{\beta\beta'}^B(K)|^2 \frac{d(Ka_0)}{(Ka_0)^3}.$$

B2.2.5.4 QUANTAL AND CLASSICAL IMPULSE CROSS SECTIONS

In the impulse approximation [6, 9], the integral cross section for $(\alpha, \beta) \rightarrow (\alpha', \beta')$ transitions in A is

$$\sigma_{if}(v_i) = \frac{2\pi}{M_{\text{eB}}^2 v_i^2} \int_{q_-}^{q_+} |f_{\text{eB}}^{\beta\beta'}(q)|^2 |F_{\alpha\alpha'}^A(q)|^2 q \, dq$$

where $f_{\text{eB}}^{\beta\beta'}$ is the scattering amplitude for elastic $\beta = \beta'$ or inelastic $\beta \neq \beta'$ collisions between projectile B and an orbital electron of A. For structureless ions B, the Coulomb $f_{\text{eB}}^{\beta\beta'}(q)$ for elastic electron-ion collisions reproduces the Born approximation for B-A collisions. When Born amplitudes $f_{\text{eB}}^{\beta\beta'}(q)$ are used for fast atom B-e collisions, then the Born approximation for atom-atom collisions is also recovered for general scattering amplitudes $f_{\text{eB}}^{\beta\beta'}$. For slow atoms B, f_{eB} is dominated by s-wave elastic scattering so that $f_{\text{eB}} = -a$ and $\sigma_{\text{eB}} = 4\pi a^2$ where a is the scattering length. Then

$$\sigma_{if}(v_i) = \frac{2\pi a^2}{(v_i/v_0)^2} \int_{K-a_0}^{K+a_0} |F_{if}^A(K)|^2 (K a_0) d(K a_0)$$

which is a good approximation for collisional transitions $n'l \rightarrow n'l'$ in Rydberg atoms A. The full quantum impulse cross section [6, 9] for general $f_{\mathbf{eB}}^{\beta\beta'}$ has recently been presented in a valuable new form [28] which is the appropriate representation for direct classical correspondence. The classical impulse cross section was then defined [28] to yield the first general expression for the classical impulse cross section for $n\ell - n'\ell'$ and $n\ell - \ell'$ electronic transitions. The cross section satisfies the optical theorem and detailed balance. Direct connection with the classical binary encounter approximation (BEA) was established and the derived $n\ell - n'$ and $n\ell - \ell$ cross sections reproduce the standard BEA cross sections.

B2.2.5.5 ATOMIC FORM FACTOR AND GENERALIZED OSCILLATOR STRENGTH

In terms of the form factor $F_{if}(K)$, the generalized oscillator strength is defined as

$$f_{if}(K) = \left(\frac{2m_e E_{fi}}{q^2} \right) |F_{if}(q)|^2 = \frac{2E_{fi}^{\text{a.u.}}}{(K a_0)^2} |F_{if}(K)|^2$$

which tends to the dipole oscillator strength in the $K \rightarrow 0$ limit.

-25-

(A) SUM RULES

$$\begin{aligned} \sum_f f_{if}(K) &= \sum_f f_{if}(K) + \int_0^\infty \frac{df_{iE}}{dE} dE = N \\ \sum_f |F_{if}(K)|^2 &= \sum_f F_{if}(K) + \int_0^\infty \frac{dF_{iE}}{dE} dE \\ &= N + \sum_{j < k}^N |\langle \Psi_i | \exp(i\vec{K} \cdot (\vec{r}_j - \vec{r}_k)) | \Psi_i \rangle|^2 \end{aligned}$$

where N is the number of electrons. The summation \sum_f extends over all discrete and continuum states.

(B) ENERGY-CHANGE MOMENTS

The energy-change moments are defined as

$$\begin{aligned} S(\alpha, K) &= \sum_{f \neq i} (2\Delta E_{fi}^{\text{a.u.}})^\alpha f_{if}(K) \\ &= \sum_{f \neq i} (2\Delta E_{fi}^{\text{a.u.}})^{\alpha+1} |F_{if}(K)|^2 (K a_0)^{-2}. \end{aligned}$$

The exact energy-change moments for H(1s) are

$$\begin{aligned}
S(-1, K) &= \{1 - [1 + \frac{1}{4}(Ka_0)^2]^{-4}\}(Ka_0)^{-2} \\
S(0, K) &= 1 \\
S(1, K) &= (Ka_0)^2 + \frac{4}{3} \\
S(2, K) &= (Ka_0)^4 + 4(Ka_0)^2 + \frac{16}{3}.
\end{aligned}$$

B2.2.5.6 FORM FACTORS FOR ATOMIC HYDROGEN

The probability of a transition $i \rightarrow f$ resulting from any external perturbation which impulsively transfers momentum \mathbf{q} to the internal momenta of the electrons of the target system is

$$P_{if}(\mathbf{q}) = |F_{fi}(\mathbf{q})|^2.$$

-26-

The impulse can be due to sudden collision with particles or to exposure to electromagnetic radiation. The physical significance of the form factor is that P_{if} is the impulsive transition probability for any atom. For $nl \rightarrow n'l'$ transitions in atomic hydrogen,

$$P_{nl,n'l'}(\mathbf{q}) = \sum_{m,m'} |\langle \Psi_{nlm}(\mathbf{r}) | e^{i\mathbf{q}\cdot\mathbf{r}/\hbar} | \Psi_{n'l'm'}(\mathbf{r}) \rangle|^2,$$

with $\Psi(\mathbf{r}) = R_{nl}(r)Y_{lm}(\hat{\mathbf{r}})$, can be decomposed as

$$P_{nl,n'l'}(\mathbf{q}) = (2l+1)(2l'+1) \sum_{L=|l-l'|}^{l+l'} (2L+1) \begin{pmatrix} L & l & l' \\ 0 & 0 & 0 \end{pmatrix}^2 [f_{nl,n'l'}^{(L)}(q)]^2$$

where (\dots) is the Wigner's $3j$ -symbol and $f_{nl,n'l'}^{(L)}(q)$ is the radial integral

$$f_{nl,n'l'}^{(L)}(q) = \int_0^\infty R_{nl}(r)R_{n'l'}(r)j_L(qr)r^2 dr$$

where j_L is the modified Bessel function. For $nlm \rightarrow n'l'm'$ subshell transitions, the amplitude decomposes as

$$F_{nlm,n'l'm'}(\mathbf{q}) = 4\pi \sum_{L=|l-l'|}^{l+l'} i^L w_{lm'l'm'}^{(L)} f_{nl,n'l'}^{(L)}(q) Y_{L,M}(\hat{\mathbf{q}})$$

where $M = m - m'$ and where the coefficients

$$w_{lm'l'm'}^{(L)} = \left[\frac{(2l+1)(2l'+1)(2L+1)}{4\pi} \right]^{\frac{1}{2}} \begin{Bmatrix} L & l & l' \\ M & -m & m' \end{Bmatrix} \begin{Bmatrix} L & l & l' \\ 0 & 0 & 0 \end{Bmatrix}.$$

Exact algebraic expressions for the probability

$$P_{n,n'}(q) = \sum_{l'l'} \sum_{m,m'} |\langle n'l'm' | e^{iq \cdot r/\hbar} | nlm \rangle|^2$$

of $n \rightarrow n'$ transitions in atomic hydrogen, have been recently derived [11] as analytical functions of n and n' .

B2.2.5.7 ROTATIONAL EXCITATION

For ion-point dipole D interactions, only $\Delta J = \pm 1$ transitions are allowed. For ion-point quadrupole Q interactions only $\Delta J = 0, \pm 2$ transitions are allowed. The Born differential cross sections for $j \rightarrow J$ transitions are

-27-

$$\begin{aligned} \frac{d\sigma^{(d)}}{d\hat{k}_f}(J \rightarrow J+1) &= \frac{4}{3} \frac{k_f}{k_i} \left(\frac{J+1}{2J+1} \right) \frac{D^2}{K^2} \\ \frac{d\sigma^{(q)}}{d\hat{k}_f}(J \rightarrow J) &= \frac{4}{45} \frac{J(J+1)}{(2J-1)(2J+3)} Q^2 \\ \frac{d\sigma^{(q)}}{d\hat{k}_f}(J \rightarrow J+2) &= \frac{2}{15} \frac{k_f}{k_i} \frac{(J+1)(J+2)}{(2J+1)(2J+3)} Q^2 \end{aligned}$$

which are all spherical symmetrical. The sum

$$\sum \frac{d\sigma^{(q)}}{d\hat{k}_f}(J \rightarrow J, J \pm 2) = \frac{4}{45} Q^2$$

is independent of the initial value of J . The integral cross sections

$$\begin{aligned} \sigma^{(d)}(J \rightarrow J+1) &= \frac{8\pi}{3k_i^2} \left(\frac{J+1}{2J+1} \right) \ln \left(\frac{k_i+k_f}{k_i-k_f} \right) D^2 \\ \sigma^{(q)}(J \rightarrow J, J+2) &= \frac{8\pi}{15} \frac{k_f}{k_i} \frac{(J+1)(J+2)}{(2J+1)(2J+3)} Q^2 \end{aligned}$$

all satisfy the detailed balance relation

$$k_i^2(2J_i+1)\sigma(J_i \rightarrow J_f) = k_f^2(2J_f+1)\sigma(J_f \rightarrow J_i).$$

The summed diffusion cross sections are

$$\begin{aligned} \sigma^{(d)} &= \int \left[\sum_{J \pm 1} \frac{d\sigma}{d\hat{k}_f}(J \rightarrow J') \right] (1 - \cos \theta) d\hat{k}_f \\ &= \left(\frac{8\pi}{3k_i^2} \right) D^2 \\ \sigma^{(q)} &= (16\pi/45) Q^2. \end{aligned}$$

B2.2.5.8 LIST OF BORN CROSS SECTIONS FOR MODEL POTENTIALS

$$k^2 = (2M_{AB}/\hbar^2)E \quad K = 2k \sin \frac{1}{2}\theta$$

$$U = (2M_{AB}/\hbar^2)V \quad U/k^2 = V/E.$$

-28-

For a symmetric potential, the scattering amplitude is

$$f_B(K) = - \int U(R) \frac{\sin KR}{KR} R^2 dR.$$

The Born integral cross section is

$$\sigma_{if}^B(E) = \frac{2\pi}{k^2} \int_{K_-}^{K_+} |f_{if}^{(B)}(K)|^2 K dK$$

which is independent of the sign of the potential V .

(A) EXPONENTIAL

$$V(R) = V_0 \exp(-\alpha R)$$

$$f_B(K) = - \frac{2\alpha U_0}{(\alpha^2 + K^2)^2}$$

$$\sigma_B(E) = \frac{16}{3} \pi U_0^2 \left[\frac{3\alpha^4 + 12\alpha^2 k^2 + 16k^4}{\alpha^4(\alpha^2 + 4k^2)^3} \right] \xrightarrow{E \rightarrow \infty} \frac{4}{3} \pi \left(\frac{V_0}{E} \right) \left(\frac{U_0}{\alpha^4} \right).$$

(B) GAUSSIAN

$$V(R) = V_0 \exp(-\alpha^2 R^2)$$

$$f_B(K) = - \left(\frac{\pi^{1/2} U_0}{4\alpha^2} \right) \exp(-K^2/4\alpha^2)$$

$$\sigma_B(E) = \left(\frac{\pi^2 U_0}{8\alpha^4} \right) \left(\frac{V_0}{E} \right) [1 - \exp(-2k^2/\alpha^2)].$$

(C) SPHERICAL WELL/BARRIER

$$V(R) = V_0 \text{ for } R < a, \quad V(R) = 0 \text{ for } R > a, \quad U_0 = (2M_{AB}/\hbar^2)V_0$$

$$f_B(K) = - \frac{U_0}{K^3} [\sin Ka - Ka \cos Ka],$$

$$\sigma_B(E) = \frac{\pi}{2} \frac{V_0}{E} (U_0 a^4) [1 - (ka)^{-2} + (ka)^{-3} \sin 2ka - (ka)^{-4} \sin^2 2ka].$$

At low energies, $f_B \rightarrow (2M_{AB}/\hbar^2)V_0 a^3/3$ and the scattering is isotropic. At high energies, $\sigma_B(E) E^{-1}$.

-29-

(D) SCREENED COULOMB INTERACTION

$$V(R) = V_0 \exp(-\alpha R)/R.$$

$$f_B(K) = -\frac{U_0}{\alpha^2 + K^2}$$
$$\sigma_B(E) = \frac{4\pi U_0^2}{\alpha^2(\alpha^2 + 4k^2)}$$

where $U_0 = 2Z/a_0$. At low energies, $f_B = -U_0/\alpha^2$ is isotropic. At high energies, $\sigma_B \rightarrow \pi (V_0/E) (U_0\alpha^2)$.

(E) ELECTRON-ATOM MODEL STATIC INTERACTION

$$V(R) = -N(e^2/a_0)[Z + a_0/R] \exp(-2ZR/a_0).$$

$$f_B(\theta) = \frac{2N}{a_0} \left[\frac{2\alpha^2 + K^2}{(\alpha^2 + K^2)^2} \right] \quad \alpha = 2Z/a_0$$
$$\sigma_B(E) = \frac{\pi a_0^2 N^2 [12Z^4 + 18Z^2 k^2 a_0^2 + 7k^4 a_0^2]}{3Z^2 (Z^2 + k^2 a_0^2)^3}.$$

For atomic H (1s), $N = 1$ and $Z = 1$. For He ($1s^2$), the approximate parameters are $N = 2$ and $Z = 27/16$.

(F) POLARIZATION POTENTIAL

$$V(R) = V_0/(R^2 + R_0^2)^2$$

$$f_B(K) = -\frac{1}{4}\pi \left(\frac{U_0}{R_0} \right) \exp(-K R_0)$$
$$\sigma_B(E) = \left(\frac{\pi^3 U_0}{32 R_0^4} \right) \left(\frac{V_0}{E} \right) [1 - (1 + 4k R_0) \exp(-4k R_0)].$$

B2.2.6 QUANTAL POTENTIAL SCATTERING

The *Schrödinger* equation

$$\left(-\frac{\hbar^2}{2M_{AB}} \nabla_r^2 + V(\mathbf{r}) \right) \Psi_k^+(\mathbf{r}) = E \Psi_k^+(\mathbf{r})$$

-30-

solved subject to the asymptotic condition

$$\Psi_k^+(\mathbf{r}) \sim \exp(i\mathbf{k} \cdot \mathbf{r}) + \frac{1}{r} f(\theta, \phi) \exp(ikr)$$

for outgoing spherical waves is equivalent to the solution of the Lippman–Schwinger integral equation

$$\Psi_{\mathbf{k}}^+(\mathbf{r}) = \Phi_{\mathbf{k}}^+(\mathbf{r}) + \int G(\mathbf{r}, \mathbf{r}') U(\mathbf{r}') \Psi_{\mathbf{k}}^+(\mathbf{r}') d\mathbf{r}'$$

where the outgoing Green's function for a free particle is

$$G(\mathbf{r}, \mathbf{r}') = -\frac{1}{4\pi} \frac{\exp\{ik|\mathbf{r} - \mathbf{r}'|\}}{|\mathbf{r} - \mathbf{r}'|}.$$

Solution of the scattering amplitude may then be determined from the asymptotic form of $\Psi_{\mathbf{k}}^+(\mathbf{r})$ directly or from the integral representation

$$f(\theta, \phi) = -\frac{1}{4\pi} (2M_{AB}/\hbar^2) (\exp(i\mathbf{k}_f \cdot \mathbf{r}) |V(\mathbf{r})| \Psi_i^+).$$

The differential cross section for elastic scattering is

$$\frac{d\sigma}{d\Omega} = |f(\theta, \phi)|^2.$$

B2.2.6.1 PARTIAL WAVE EXPANSION

A plane wave of unit amplitude can be decomposed according to

$$\Phi_{\mathbf{k}}(\mathbf{r}) = \exp(i\mathbf{k} \cdot \mathbf{r}) = 4\pi \sum_{\ell, m} i^\ell j_\ell(kr) Y_{\ell m}^*(\hat{\mathbf{k}}) Y_{\ell m}(\hat{\mathbf{r}})$$

where j_ℓ is the spherical Bessel function which varies asymptotically as

$$j_\ell(kr) \sim \frac{1}{kr} \sin\left(kr - \frac{1}{2}\ell\pi\right).$$

-31-

The addition theorem for spherical harmonics is

$$4\pi \sum_m Y_{\ell m}^*(\hat{\mathbf{k}}) Y_{\ell m}(\hat{\mathbf{r}}) = (2\ell + 1) P_\ell(\hat{\mathbf{k}} \cdot \hat{\mathbf{r}}).$$

Another useful identity is,

$$\frac{\sin Kr}{Kr} = 4\pi \sum_{\ell, m} j_\ell(k_i r) j_\ell(k_f r) Y_{\ell m}(\hat{\mathbf{k}}_i) Y_{\ell m}^*(\hat{\mathbf{k}}_f)$$

where $K^2 = k_i^2 + k_f^2 - 2k_i k_f \cos \theta$. The system wavefunction $\Psi_{\mathbf{k}}^+(\mathbf{r}) \sim N \Phi_{\mathbf{k}}$ with amplitude N is expanded according to

$$\begin{aligned}\Psi_{\mathbf{k}}^+(\mathbf{r}) &= \sum_{\ell,m} i^\ell e^{i\eta_\ell} R_{\ell\ell}(r) Y_{\ell m}^*(\hat{\mathbf{k}}) Y_{\ell m}(\hat{\mathbf{r}}) \\ &= \frac{4\pi N}{kr} \sum_{\ell,m} i^\ell F_{\ell\ell}(r) Y_{\ell m}^*(\hat{\mathbf{k}}) Y_{\ell m}(\hat{\mathbf{r}}).\end{aligned}$$

The radial wave F_ℓ is the solution of the radial Schrödinger equation

$$\frac{d^2 F_\ell}{dr^2} + \left[k^2 - \left\{ U(r) + \frac{\ell(\ell+1)}{r^2} \right\} \right] F_\ell(r) = 0.$$

The reduced potential and energy are $U(r) = (2M_{AB}/\hbar^2)V(r)$ and $K^2 = (2M_{AB}/\hbar^2)E$, respectively. They both have dimensions of $[a_0^{-2}]$. Also $(ka_0)^2 = (2E/\varepsilon_0)(M_{AB}/m_e)$. Each ℓ -partial wave is separately scattered since the angular momentum of relative motion is conserved for central forces. The radial waves $R_{\ell\ell}(r)$ and $F_{\ell\ell}(r)$ vary asymptotically as

$$\begin{aligned}R_{\ell\ell}(r) &\sim \frac{N_\ell}{r} \sin(kr - \frac{1}{2}\ell\pi + \eta_\ell) \\ F_{\ell\ell}(r) &\sim e^{i\eta_\ell} \sin(kr - \frac{1}{2}\ell\pi + \eta_\ell).\end{aligned}$$

The amplitude of the partial radial wave $R_{\ell\ell}(r)$ is $N_\ell = 4\pi N/k$. In table B2.2.1 are displayed the amplitudes N and N_ℓ appropriate to various choices for normalization of the continuum wavefunctions $\Psi_{\mathbf{k}}(r)$.

Table B2.2.1 Continuum wavefunction normalization, density of states and cross section factors.

Type	$\langle \Phi_{\mathbf{k}'} \Phi_{\mathbf{k}} \rangle$	N	N_ℓ	$\rho(E)$	γ_{if}
Unit amplitude	$\left\{ \begin{array}{l} (2\pi)^3 \delta(\mathbf{k} - \mathbf{k}') \\ (2\pi\hbar)^3 \delta(\mathbf{p} - \mathbf{p}') \end{array} \right\}$	1	$\frac{4\pi}{k}$	mp/\hbar^3	$\frac{v_f}{v_i} \left(\frac{1}{4\pi} \right)^2 \left(\frac{2m_f}{\hbar^2} \right)^2$
Wavenumber	$\delta(\mathbf{k} - \mathbf{k}')$	$(2\pi)^{-3/2}$	$\left(\frac{2}{\pi} \right)^{1/2} \frac{1}{k}$	mp/\hbar^3	$\frac{v_f}{v_i} (2\pi^2)^2 \left(\frac{2m_f}{\hbar^2} \right)^2$
Momentum	$\delta(\mathbf{p} - \mathbf{p}')$	$(2\pi\hbar)^{-3/2}$	$\left(\frac{2}{\pi\hbar} \right)^{1/2} \frac{1}{k}$	mp	$\frac{v_f}{v_i} (2\pi^2\hbar^3)^2 \left(\frac{2m_f}{\hbar^2} \right)^2$
Directed energy	$\delta(E - E')$	$(mp/\hbar^3)^{1/2}$	$\left(\frac{2m}{\hbar^2} \frac{1}{\pi k} \right)^{1/2}$	1	$\frac{(2\pi)^4}{k_i^2}$

B2.2.6.2 SCATTERING AMPLITUDES

For symmetric interactions $V = V(r)$, the wavefunctions $\Psi_i^+ = \sum_j \Phi_j(\mathbf{R}) \psi_j^{\text{int}}(\mathbf{r})$ and $\exp(i\mathbf{k}_f \cdot \mathbf{r})$ are

decomposed into partial waves. From their asymptotic forms, the following partial wave expansions for the scattering amplitude

$$f(\theta) = \frac{1}{2ik} \sum_{\ell=0}^{\infty} (2\ell+1) [\exp(2i\eta_{\ell}) - 1] P_{\ell}(\cos \theta)$$

$$f(\theta) = \frac{1}{2ik} \sum_{\ell=0}^{\infty} (2\ell+1) [S_{\ell}(k) - 1] P_{\ell}(\cos \theta)$$

$$f(\theta) = \frac{1}{2ik} \sum_{\ell=0}^{\infty} (2\ell+1) T_{\ell}(k) P_{\ell}(\cos \theta)$$

can be deduced. The *scattering*, *transition* and *reactance* matrix elements are defined, in terms of the *phase shift* η_{ℓ} suffered by each partial wave, as

$$S_{\ell}(k) = \exp(2i\eta_{\ell})$$

$$T_{\ell}(k) = 2i \sin \eta_{\ell} \exp(i\eta_{\ell})$$

$$K_{\ell}(k) = \tan \eta_{\ell}.$$

The asymptotic ($kr \rightarrow \infty$) form of F_{ℓ} may then be written in terms of the following linear combinations:

-33-

$$R_{\ell}(r) \sim \frac{N_{\ell}}{r} \sin(kr - \frac{1}{2}\ell\pi + \eta_{\ell})$$

$$F_{\ell}(r) \sim e^{i\eta_{\ell}} \sin(kr - \frac{1}{2}\ell\pi + \eta_{\ell}).$$

$$F_{\ell}(kr) \sim \sin(kr - \ell\pi/2) + \left(\frac{T_{\ell}}{2i}\right) e^{i(kr - \ell\pi/2)}$$

$$= -\frac{1}{2i} [e^{-i(kr - \ell\pi/2)} - S_{\ell} e^{i(kr - \ell\pi/2)}]$$

$$= e^{i\eta_{\ell}} \cos \eta_{\ell} [\sin(kr - \ell\pi/2) + K_{\ell} \cos(kr - \ell\pi/2)]$$

expressed as a combinations of standing waves (trigonometric functions), of incoming (-) and outgoing (+) spherical waves (exponential functions) and of a standing wave and an outgoing spherical wave. The physical significance of the admixture coefficients S_{ℓ} , T_{ℓ} and K_{ℓ} is then transparent. The elements are connected by

$$S_{\ell} = 1 + T_{\ell} = (1 + iK_{\ell}) / (1 - iK_{\ell})$$

K_{ℓ} is real while both S_{ℓ} and T_{ℓ} are complex. In term of the full solutions F_{ℓ} of the radial *Schrödinger* equation, the *T*-matrix element for elastic scattering is

$$T_{\ell} = -\frac{2i}{k} \int_0^{\infty} F_{\ell}^{(0)}(r) U(r) F_{\ell}(r) dr$$

where $F_{\ell}^{(0)} = (kr) j_{\ell}(kr)$ is the radial component of the final plane wave. The Born approximation to T_{ℓ} is

obtained upon the substitution $F_\ell^{(0)} = F_\ell$.

B2.2.6.3 INTEGRAL CROSS SECTIONS

$$\begin{aligned}\sigma(E) &= \frac{4\pi}{k^2} \sum_{\ell=0}^{\infty} (2\ell + 1) \sin^2 \eta_\ell \\ &= \frac{\pi}{k^2} \sum_{\ell=0}^{\infty} (2\ell + 1) |T_\ell|^2 \\ &= \frac{2\pi}{k} \sum_{\ell=0}^{\infty} (2\ell + 1) [1 - \text{Re}S_\ell].\end{aligned}$$

-34-

The semiclassical version is obtained by the substitution $mvb = (\ell + \frac{1}{2})\hbar$ so that $K^2 b^2 = (\ell + \frac{1}{2})^2$ in terms of the impact parameter b . Regarding ℓ as a continuous variable,

$$\sigma(E) = \frac{\pi}{k^2} \int_0^\infty (2\ell + 1) |T_\ell|^2 d\ell = 2\pi \int_0^\infty |T(b)|^2 b db.$$

The transition matrix $|T(b)|^2$ is therefore the probability of scattering particles with impact parameter b .

B2.2.6.4 DIFFERENTIAL CROSS SECTIONS

The differential cross section for elastic scattering is

$$\frac{d\sigma}{d\Omega} = |f(\theta)|^2 = A(\theta)^2 + B(\theta)^2$$

where the real and imaginary parts of $f(\theta)$ are, respectively,

$$\begin{aligned}A(\theta) &= \frac{1}{2k} \sum_{\ell=0}^{\infty} (2\ell + 1) \sin 2\eta_\ell P_\ell(\cos \theta) \\ B(\theta) &= \frac{1}{2k} \sum_{\ell=0}^{\infty} (2\ell + 1) [1 - \cos 2\eta_\ell] P_\ell(\cos \theta).\end{aligned}$$

Their individual contributions to the integral cross sections are

$$\begin{aligned}\int A(\theta)^2 d\Omega &= \frac{4\pi}{k^2} \sum_{\ell=0}^{\infty} (2\ell + 1) \sin^2 \eta_\ell \cos^2 \eta_\ell \\ \int B(\theta)^2 d\Omega &= \frac{4\pi}{k^2} \sum_{\ell=0}^{\infty} (2\ell + 1) \sin^4 \eta_\ell.\end{aligned}$$

(A) EXPANSION IN LEGENDRE POLYNOMIALS

When expanded as a series of Legendre polynomials $P_L(\cos \theta)$, the differential cross section has the following form

$$\frac{d\sigma(E, \theta)}{d\Omega} = \frac{1}{k^2} \sum_{L=0}^{\infty} a_L(E) P_L(\cos \theta)$$

-35-

where the coefficients

$$a_L = \sum_{\ell=0}^{\infty} \sum_{\ell'=|\ell-L|}^{\ell+L} (2\ell+1)(2\ell'+1)(\ell\ell'00 | \ell\ell'L0)^2 \sin \eta_{\ell} \sin \eta_{\ell'} \cos(\eta_{\ell} - \eta_{\ell'})$$

are determined by the phase shifts η_{ℓ} and the Clebsch–Gordon coefficients $(\ell\ell'mm' | \ell\ell'LM)$.

(B) EXAMPLE: THREE-TERM EXPANSION IN COS θ

The differential cross section can be expanded as

$$\frac{d\sigma(E, \theta)}{d\Omega} = \frac{1}{k^2} [(a_0 - \frac{1}{2}a_2) + a_1 \cos \theta + \frac{3}{2}a_2 \cos^2 \theta].$$

The coefficients are

$$\begin{aligned} a_0 &= \sum_{\ell=0}^{\infty} (2\ell+1) \sin^2 \eta_{\ell} \\ a_1 &= 6 \sum_{\ell=0}^{\infty} (\ell+1) \sin \eta_{\ell} \sin \eta_{\ell+1} \cos(\eta_{\ell+1} - \eta_{\ell}) \\ a_2 &= 5 \sum_{\ell=0}^{\infty} [b_{\ell} \sin^2 \eta_{\ell} + c_{\ell} \sin \eta_{\ell} \sin \eta_{\ell+2} \cos(\eta_{\ell+2} - \eta_{\ell})] \end{aligned}$$

where

$$\begin{aligned} b_{\ell} &= \frac{\ell(\ell+1)(2\ell+1)}{(2\ell+1)(2\ell+3)} \\ c_{\ell} &= \frac{3(\ell+1)(\ell+2)}{2\ell+3}. \end{aligned}$$

(C) EXAMPLE: S- AND P-WAVE CONTRIBUTIONS

The combined S-, P-wave ($\ell = 0, 1$) contributions to the differential and integral cross sections are

$$\frac{d\sigma}{d\Omega} = \frac{1}{k^2} [\sin^2 \eta_0 + [6 \sin \eta_0 \sin \eta_1 \cos(\eta_1 - \eta_0)] \cos \theta + 9 \sin^2 \eta_1 \cos^2 \theta]$$

$$\sigma(E) = \frac{4\pi}{k^2} [\sin^2 \eta_0 + 3 \sin^2 \eta_1].$$

-36-

For pure S-wave scattering, the differential cross section (DCS) is isotropic. For pure P-wave scattering, the DCS is symmetric about $\theta = \pi/2$, where it vanishes; the DCS rises to equal maxima at $\theta = 0, \pi$. For combined S- and P-wave scattering, the DCS is asymmetric with forward-backward asymmetry.

B2.2.6.5 OPTICAL THEOREM

The optical theorem relates the integral cross section to the imaginary part of the forward scattering amplitude by

$$\sigma(E) = (4\pi/k) \text{Im} f(0).$$

This relation is a direct consequence of the conservation of flux. The target casts a shadow in the forward direction where the intensity of the incident beam becomes reduced by just that amount which appears in the scattered wave. This decrease in intensity or shadow results from interference between the incident wave and the scattered wave in the forward direction. Figure B2.2.2 for the density $|\Psi_k^*(\mathbf{r})|$ of section B2.2.6 illustrates how this interference tends to illuminate the shadow region at the right-hand side of the target. Flux conservation also implies that the phase shifts η_ℓ are always real. Thus

$$|S_\ell|^2 = 1 \quad |T_\ell|^2 = \text{Im} T_\ell.$$

B2.2.6.6 LEVINSON'S THEOREM

For a local potential $V(r)$ which supports η_ℓ bound states of angular momentum ℓ and energy $E_n < 0$, the phase shift $\lim_{k \rightarrow 0} \eta_\ell(k)$ tends in the limit of zero collision energy to $\eta_\ell \pi$. When the well becomes deep enough so as to introduce an additional bound level $E_{n+1} = 0$ at zero energy, then $\lim_{k \rightarrow 0} \eta_0(k) = (n_0 + \frac{1}{2})\pi$.

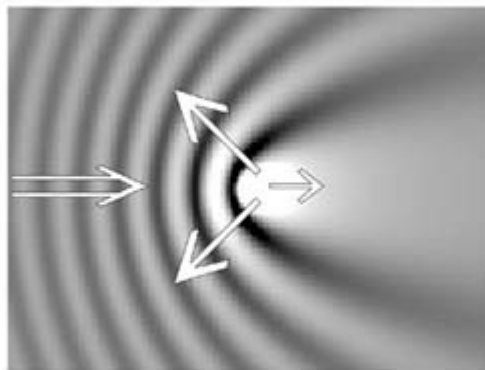


Figure B2.2.2. Scattering of an incident plane wave.

B2.2.6.7 PARTIAL WAVE EXPANSION FOR TRANSPORT CROSS SECTIONS

The transport cross sections

$$\sigma^{(n)}(E) = 2\pi \left[1 - \frac{1 + (-1)^n}{2(n+1)} \right]^{-1} \int_{-1}^{+1} [1 - \cos^n \theta] \frac{d\sigma}{d\Omega}(\cos \theta)$$

for $n = 1-4$ have the following phase shift expansions

$$\begin{aligned} \sigma^{(1)}(E) &= \frac{4\pi}{k^2} \sum_{\ell=0}^{\infty} (\ell+1) \sin^2(\eta_{\ell} - \eta_{\ell+1}) \\ \sigma^{(2)}(E) &= \frac{4\pi}{k^2} \left(\frac{3}{2}\right) \sum_{\ell=0}^{\infty} \frac{(\ell+1)(\ell+2)}{(2\ell+3)} \sin^2(\eta_{\ell} - \eta_{\ell+2}) \\ \sigma^{(3)}(E) &= \frac{4\pi}{k^2} \sum_{\ell=0}^{\infty} \frac{(\ell+1)}{(2\ell+5)} \left[\frac{(\ell+2)(\ell+3)}{(2\ell+3)} \sin^2(\eta_{\ell} - \eta_{\ell+3}) + \frac{3(\ell^2+2\ell-1)}{(2\ell-1)} \sin^2(\eta_{\ell} - \eta_{\ell+1}) \right] \\ \sigma^{(4)}(E) &= \frac{4\pi}{k^2} \left(\frac{5}{4}\right) \sum_{\ell=0}^{\infty} \frac{(\ell+1)(\ell+2)}{(2\ell+3)(2\ell+7)} \left[\frac{(\ell+3)(\ell+4)}{(2\ell+5)} \sin^2(\eta_{\ell} - \eta_{\ell+4}) + \frac{2(2\ell^2+6\ell-3)}{(2\ell-1)} \sin^2(\eta_{\ell} - \eta_{\ell-2}) \right]. \end{aligned}$$

The momentum-transfer or diffusion cross section is $\sigma^{(1)}$ and the viscosity cross section is $\frac{2}{3}\sigma^{(2)}$.

B2.2.6.8 BORN PHASE SHIFTS

For a symmetric interaction, the Born amplitude is

$$f_B(K) = - \int U(R) \frac{\sin KR}{KR} R^2 dR$$

where $U(r) = (2M_{AB}/\hbar^2)V(R)$. Comparison with the partial wave expansion for $f_B(K)$ and

$$\frac{\sin KR}{KR} = \sum_{\ell=0}^{\infty} (2\ell+1) [j_{\ell}(kR)]^2 P_{\ell}(\cos \theta)$$

provides the Born phase shift

$$\tan \eta_{\ell}^B(k) = -k \int_0^{\infty} U(R) [j_{\ell}(kR)]^2 R^2 dR.$$

(A) EXAMPLES OF THE BORN S-WAVE PHASE SHIFT

$$\tan \eta_0^B(k) = -\frac{1}{k} \int_0^\infty U(R) \sin^2(kR) dR.$$

For the potential $U = U_0 \frac{e^{-\alpha R}}{R}$

$$\tan \eta_0^B = -\frac{U_0}{4k} \ln[1 + 4k^2/\alpha^2].$$

For the potential $U = \frac{U_0}{(R^2 + R_0^2)^2}$

$$\tan \eta_0^B = -\frac{\pi U_0}{4k R_0^3} [1 - (1 + 2kR_0) e^{-2kR_0}].$$

(B) BORN PHASE SHIFTS (LARGE ℓ)

For $\ell \gg ka$,

$$\tan \eta_\ell^B = -\frac{k^{2\ell+1}}{[(2\ell+1)!!]^2} \int_0^\infty U(R) R^{2\ell+2} dR$$

valid only for finite-range interactions $U(R > a) = 0$. If $U = -U_0$, $R \leq a$ and $U = 0$, $R > a$, then

$$\tan \eta_\ell^B (\ell \gg ka) = U_0 a^2 \frac{(ka)^{2\ell+1}}{[(2\ell+1)!!]^2 (2\ell+3)}.$$

The ratio $\eta_{\ell+1}/\eta_\ell \sim (ka/2\ell)^2$.

B2.2.6.9 COULOMB SCATTERING

For elastic scattering by the interaction $V(r) = Z_A Z_B e^2/r$, the Coulomb wave can be decomposed as

$$\Psi_k^{(C)}(\mathbf{r}) = \frac{4\pi}{kr} \sum_{\ell, m} i^\ell e^{i\eta_\ell} F_{\ell\ell}(r) Y_{\ell m}^*(\hat{\mathbf{k}}) Y_{\ell m}(\hat{\mathbf{r}})$$

where the radial wave varies asymptotically as

$$F_\ell \sim \sin(kR - \frac{1}{2}\ell\pi + \eta_\ell^{(C)} - \beta \ln 2kR)$$

where the parameter β is $Z_A Z_B e^2/\hbar v$. The Coulomb phase shift is

$$\eta_\ell^{(C)} = \arg \Gamma(\ell + 1 + i\beta) = \text{Im} \ln \Gamma(\ell + 1 + i\beta)$$

to give the Coulomb S -matrix element

$$S_\ell^{(C)} = \exp[2i\eta_\ell^{(C)}] = \frac{\Gamma(\ell + 1 + i\beta)}{\Gamma(\ell + 1 - i\beta)}.$$

The Coulomb scattering amplitude is

$$f_C(\theta) = -\frac{\beta \exp[2i\eta_\ell^{(C)} - i\alpha \ln(\sin^2 \frac{1}{2}\theta)]}{2k \sin^2 \frac{1}{2}\theta}.$$

The Coulomb differential cross section $|f_C|^2$ is

$$\left(\frac{d\sigma}{d\Omega}\right)_{\text{Coul}} = \frac{Z_A^2 Z_B^2 e^4}{16E^2} \text{cosec}^4 \frac{1}{2}\theta.$$

This is the Rutherford scattering cross section. It is interesting to note that Born and classical theory also reproduce this cross section. Moreover,

$$\left(\frac{d\sigma}{d\Omega}\right)_{\text{Coul}}(q) = \frac{4M_{AB}^2 Z_A^2 Z_B^2}{q^4} = 4a_0^2 (M_{AB}/m_e)^2 \left[\frac{Z_A^2 Z_B^2}{(K a_0)^4} \right]$$

is a function only of the momentum transferred $q = \hbar K = 2\hbar k \sin \frac{1}{2}\theta$ in the collision. Note that $q^2 = 8M_{AB}E \sin^2 \frac{1}{2}\theta$.

B2.2.7 COLLISIONS BETWEEN IDENTICAL PARTICLES

The identical colliding particles, each with spin s , are in a resolved state with total spin S_t in the range $(0 \rightarrow 2s)$. The spatial wavefunction with respect to particle interchange satisfies $\Psi(\mathbf{R}) = (-1)^{S_t} \Psi(-\mathbf{R})$. Wavefunctions for identical particles with even or odd total spin S_t are therefore symmetric (S) or antisymmetric (A) with respect to particle

-40-

interchange. The appropriate combinations are $\Psi_{S,A}(\mathbf{R}) = \Psi(\mathbf{R}) \pm \Psi(-\mathbf{R})$, where the positive sign (symmetric wavefunction S) and the negative sign (antisymmetric wavefunction A) are associated with even and odd values of the total spin S_t , respectively. The scattering wavefunction for a pair of identical particles in spatially symmetric (+) or antisymmetric (-) states behaves asymptotically as

$$\Psi_{S,A}(\mathbf{R}) \rightarrow [\exp(i\mathbf{k} \cdot \mathbf{R}) \pm \exp(-i\mathbf{k} \cdot \mathbf{R})] + [f(\theta, \phi) \pm f(\pi - \theta, \phi + \pi)] \frac{\exp(i\mathbf{k} \cdot \mathbf{R})}{R}.$$

The differential cross section for scattering of *both* the projectile and target particles into direction θ is

$$\left(\frac{d\sigma}{d\Omega}\right)_{S,A} = |f(\theta, \phi) \pm f(\pi - \theta, \phi + \pi)|^2$$

in the CM-frame where scattering of the projectile into polar direction $(\pi - \theta, \phi + \pi)$ is accompanied by scattering of the identical target particle into direction (θ, ϕ) . This is related to the probability that both identical particles are scattered into θ . In the classical limit, where the particles are distinguishable, the classical cross section is

$$\left(\frac{d\sigma}{d\Omega}\right)_C = |f(\theta, \phi)|^2 + |f(\pi - \theta, \phi + \pi)|^2$$

the sum of the cross sections for observation of the projectile and target particles in the direction (θ, ϕ) . Since $P_\ell[\cos(\pi - \theta)] = (-1)^\ell P_\ell(\cos \theta)$, the differential cross section for ϕ -independent amplitudes f is then

$$\left(\frac{d\sigma}{d\Omega}\right)_{S,A} = \frac{1}{4k^2} \left| \sum_{\ell=0}^{\infty} \omega_\ell (2\ell + 1) [\exp 2i\eta_\ell - 1] P_\ell(\cos \theta) \right|^2.$$

For scattering in the symmetric (S) channel where S_t is even, $\omega_\ell = 2$ for ℓ even and $\omega_\ell = 0$ for ℓ odd. For scattering in the antisymmetric channel where S_t is odd, $\omega_\ell = 0$ for ℓ even and $\omega_\ell = 2$ for ℓ odd. The integral cross section is

$$\sigma_{S,A}(E) = \frac{8\pi}{k^2} \sum_{\ell=0}^{\infty} \omega_\ell (2\ell + 1) \sin^2 \eta_\ell.$$

Let g_A and g_S be the fractions of states with odd and even total spins $S_t = 0, 1, 2, \dots, 2s$. When the $2s + 1$ spin-states S_t are unresolved, the appropriate combination of symmetric and antisymmetric cross sections is the weighted mean

-41-

$$\begin{aligned} \frac{d\sigma}{d\Omega} &= g_S \left(\frac{d\sigma}{d\Omega}\right)_S + g_A \left(\frac{d\sigma}{d\Omega}\right)_A \\ \sigma(E) &= g_S \sigma_S(E) + g_A \sigma_A(E). \end{aligned}$$

B2.2.7.1 FERMION AND BOSON SCATTERING

(A) FERMIONS

For fermions with half-integral spin s , the statistical weights are $g_S = s/(2s + 1)$ and $g_A = (s + 1)/(2s + 1)$. The differential cross section for fermion-fermion scattering is then

$$\frac{d\sigma_F}{d\Omega} = |f(\theta)|^2 + |f(\pi - \theta)|^2 - \left(\frac{2}{2s + 1}\right) \text{Re}[f(\theta)f^*(\pi - \theta)].$$

The integral cross section fermion-fermion collisions is

$$\sigma_F = \frac{1}{2}[\sigma_S + \sigma_A] - \frac{1}{2}[\sigma_S - \sigma_A]/(2s + 1)$$

which reduces, for fermions with spin- $\frac{1}{2}$, to

$$\sigma_F(E) = \frac{2\pi}{k^2} \left[\sum_{\ell=\text{even}}^{\infty} (2\ell + 1) \sin^2 \eta_\ell + 3 \sum_{\ell=\text{odd}}^{\infty} (2\ell + 1) \sin^2 \eta_\ell \right].$$

(B) BOSONS

The statistical weights for bosons with integral spin s , are $g_S = (s + 1)/(2s + 1)$ and $g_A = s/(2s + 1)$. The differential cross section for boson–boson scattering is

$$\frac{d\sigma_B}{d\Omega} = |f(\theta)|^2 + |f(\pi - \theta)|^2 + \left(\frac{2}{2s + 1} \right) \text{Re}[f(\theta)f^*(\pi - \theta)].$$

The integral cross section boson–boson collisions is

$$\sigma_B = \frac{1}{2}[\sigma_S + \sigma_A] + \frac{1}{2}[\sigma_S - \sigma_A]/(2s + 1)$$

-42-

which reduces, for bosons with zero spin, to

$$\sigma_B(E) = \frac{8\pi}{k^2} \left[\sum_{\ell=\text{even}}^{\infty} (2\ell + 1) \sin^2 \eta_\ell \right].$$

Symmetry oscillations therefore appear in the differential cross sections for fermion–fermion and boson–boson scattering. They originate from the interference between unscattered incident particles in the forward ($\theta = 0$) direction and backward scattered particles ($\theta = \pi$, $\ell = 0$). A general differential cross section for scattering of spin- s particles is

$$\frac{d\sigma}{d\Omega} = |f(\theta)|^2 + |f(\pi - \theta)|^2 + \frac{(-1)^{2s}}{2s + 1} 2 \text{Re}[f(\theta)f^*(\pi - \theta)].$$

B2.2.7.2 COULOMB SCATTERING OF TWO IDENTICAL PARTICLES

(A) TWO SPIN-ZERO BOSONS

Two spin-zero bosons (e.g. ${}^4\text{He}$ – ${}^4\text{He}$)

$$\frac{d\sigma}{d\Omega} = \frac{\beta^2}{4k^2} [\text{cosec}^4 \frac{1}{2}\theta + \sec^4 \frac{1}{2}\theta + 2\text{cosec}^2 \frac{1}{2}\theta \sec^2 \frac{1}{2}\theta \cos \gamma].$$

(B) TWO SPIN- $\frac{1}{2}$ FERMIONS

Two spin- $\frac{1}{2}$ fermions (e.g. H^+-H^+ , $e^\pm-e^\pm$)

$$\frac{d\sigma}{d\Omega} = \frac{\beta^2}{4k^2} [\operatorname{cosec}^4 \frac{1}{2}\theta + \sec^4 \frac{1}{2}\theta - \operatorname{cosec}^2 \frac{1}{2}\theta \sec^2 \frac{1}{2}\theta \cos \gamma].$$

(C) TWO SPIN-1 BOSONS

Two spin-1 bosons (e.g. deuteron–deuteron)

$$\frac{d\sigma}{d\Omega} = \frac{\beta^2}{4k^2} [\operatorname{cosec}^4 \frac{1}{2}\theta + \sec^4 \frac{1}{2}\theta + \frac{2}{3} \operatorname{cosec}^2 \frac{1}{2}\theta \sec^2 \frac{1}{2}\theta \cos \gamma].$$

(a)–(c) are the Mott formulae, where $\beta = (Ze)^2/\hbar v$ and $\gamma = 2\beta \ln(\tan \frac{1}{2}\theta)$.

B2.2.7.3 SCATTERING OF IDENTICAL ATOMS

Two ground-state hydrogen atoms, for example, interact via the $X^1\Sigma_g^+$ and $b^3\Sigma_u^+$ electronic states of H_2 . The nuclei are interchanged by rotating the atom pair by π , then by reflecting the electrons first through the midpoint of R and then through a plane perpendicular to the original axis of rotation. The mid-point reflection changes the sign only of the ungerade state wavefunction and both Σ^+ states are symmetric with respect to the plane reflection.

The cross section for scattering by the gerade potential is then the combination

$$\left(\frac{d\sigma}{d\Omega}\right)_g = \frac{1}{4} \left(\frac{d\sigma}{d\Omega}\right)_S + \frac{3}{4} \left(\frac{d\sigma}{d\Omega}\right)_A$$

of S and A cross sections which involve the phase shifts η_l^S calculated under the singlet interaction. For scattering by the ungerade triplet interaction

$$\left(\frac{d\sigma}{d\Omega}\right)_u = \frac{1}{4} \left(\frac{d\sigma}{d\Omega}\right)_A + \frac{3}{4} \left(\frac{d\sigma}{d\Omega}\right)_S$$

where the S and A cross sections involve the phase shifts η_l^T calculated under the triplet interaction. Since the electrons have statistical weights $\frac{1}{4}$ and $\frac{3}{4}$ for the Σ_g^+ and Σ_u^+ states, the differential cross section for $H(1s) - H(1s)$ scattering by both potentials is

$$\frac{d\sigma}{d\Omega} = \frac{1}{4} \left(\frac{d\sigma}{d\Omega}\right)_g + \frac{3}{4} \left(\frac{d\sigma}{d\Omega}\right)_u.$$

These combinations also hold for the integral cross sections.

(C) SCATTERING OF INCIDENT BEAM ALONE

Since the current of incident particles $j_i = 2v$, the cross sections presented by the target (i.e. the number of incident particles removed from the beam in unit time per unit incident current) are 1/2 of all those above. For example,

$$\left(\frac{d\sigma}{d\Omega}\right)_{S,A}^I = \frac{1}{2}|f(\theta) \pm f(\pi - \theta)|^2 = \frac{1}{2}\left(\frac{d\sigma}{d\Omega}\right)_{S,A}$$

and

$$\sigma_{S,A}^I(E) = \frac{1}{2}\sigma_{S,A}(E).$$

B2.2.8 QUANTAL INELASTIC HEAVY-PARTICLE COLLISIONS

The wavefunction for the complete A–B collision system satisfies the Schrödinger equation

$$\begin{aligned} \mathcal{H}(\mathbf{r}, \mathbf{R})\Psi(\mathbf{r}, \mathbf{R}) &= \left[\hat{H}_{\text{int}}(\mathbf{r}) - \frac{\hbar^2}{2M_{AB}}\nabla_{\mathbf{R}}^2 + V(\mathbf{r}, \mathbf{R}) \right] \Psi(\mathbf{r}, \mathbf{R}) \\ &= E\Psi(\mathbf{r}, \mathbf{R}) \end{aligned}$$

where the internal Hamiltonian is the sum $\hat{H}_{\text{int}}(\mathbf{r}) = H_A(\mathbf{r}_A) + H_B(\mathbf{r}_B)$ of individual Hamiltonians $H_{A,B}$ for each isolated atomic or molecular species. The total energy (internal plus relative)

$$E = \frac{\hbar^2 k_i^2}{2M_{AB}} + \epsilon_i = \frac{\hbar^2 k_f^2}{2M_{AB}} + \epsilon_f$$

remains constant for all channels f throughout the collision. The combined internal energy ϵ_i of A and B at infinite separation R is $\epsilon_i(A) + \epsilon_i(B)$ which are the eigenvalues of the internal Hamiltonian \hat{H}_{int} corresponding to the combined eigenstates $\Phi_A(\mathbf{r}_A)\Phi_B(\mathbf{r}_B)$. There are two limiting formulations (*diabatic* and *adiabatic*) for describing the relative motion. These depend on whether the mutual electrostatic interaction $V(\mathbf{r}, \mathbf{R})$ between A and B at nuclear separation R , or the variation in the kinetic energy of relative motion, is considered to be a perturbation to the system, i.e. on whether the incident speed v_i is fast or slow in comparison with the internal motions, e.g. with the electronic speed of the electrons bound to A and B.

B2.2.8.1 ADIABATIC FORMULATION (KINETIC COUPLING SCHEME)

When relaxation of the internal motion during the collision is fast compared with the slow collision speed v_i , or when the relaxation time is short compared with the collision time, the kinetic energy operator $(2M_{AB}/\hbar^2)\nabla_{\mathbf{R}}^2$ is then considered as a small perturbation to the quasi-molecular A–B system at fixed \mathbf{R} . The system wavefunction $\Psi(\mathbf{r}, \mathbf{R}) = \sum_n F_n(\mathbf{R})\Phi_n(\mathbf{r}, \mathbf{R})$ can therefore be expanded in terms of the known ‘adiabatic’ molecular wavefunctions $\Phi_n(\mathbf{r}, \mathbf{R})$ for the quasi-molecule AB at fixed nuclear separation \mathbf{R} . This set of orthonormal eigenfunctions satisfies

$$[\hat{H}_{\text{int}}(\mathbf{r}) + V(\mathbf{r}, \mathbf{R})]\Phi_n(\mathbf{r}, \mathbf{R}) = E_n(\mathbf{R})\Phi_n(\mathbf{r}, \mathbf{R}).$$

As $R \rightarrow \infty$, both $\Phi_n(\mathbf{r}, \mathbf{R})$ and the eigenenergies $E_n(\mathbf{R})$ tend, in the limit of infinite nuclear separation \mathbf{R} , to the (diabatic) eigenfunctions $\Phi_n(\mathbf{R}_A, \mathbf{R}_B) = \psi_i(\mathbf{R}_A)\phi_j(\mathbf{R}_B)$, of \hat{H}_{int} with eigenenergies ϵ_n , respectively. The substitution $\Psi(\mathbf{r}, \mathbf{R}) = \sum_n F_n(\mathbf{R})\Phi_n(\mathbf{r}, \mathbf{R})$ into the Schrödinger equation results in the following set

$$[\nabla_{\mathbf{R}}^2 + \kappa_n^2(\mathbf{R})]F_n(\mathbf{R}) = \sum_j [X_{nj} \cdot \nabla_{\mathbf{R}} + T_{nj}(\mathbf{r})]F_j(\mathbf{R})$$

-45-

of coupled equations for the relative motion functions F_n . The local momentum K_n is determined from $\kappa_n^2 = 2M_{AB}[E - E_n(\mathbf{R})]/\hbar^2$ and the coupling matrix elements are

$$X_{nj}(\mathbf{R}) = -2\langle \Phi_n(\mathbf{r}, \mathbf{R}) | \nabla_{\mathbf{R}} | \Phi_j(\mathbf{r}, \mathbf{R}) \rangle_{\mathbf{r}}$$

and

$$T_{nj}(\mathbf{R}) = -\langle \Phi_n(\mathbf{r}, \mathbf{R}) | \nabla_{\mathbf{R}}^2 | \Phi_j(\mathbf{r}, \mathbf{R}) \rangle_{\mathbf{r}}.$$

Solution of this set for $F_n(\mathbf{R})$ represents the *adiabatic close-coupling method*. The adiabatic states are normally determined (via standard computational techniques of quantum chemistry) relative to a set of axes (X', Y', Z') with the Z'-axis directed along the nuclear separation \mathbf{R} . On transforming to this set which rotates during the collision, then $\psi(\mathbf{r}', \mathbf{R}')$, for the diatomic A–B case, satisfies

$$\left[\hat{H}_0(\mathbf{r}') + V(\mathbf{r}', \mathbf{R}') - \frac{\hbar^2}{2M_{AB}R'^2} \hat{K} \right] \Psi(\mathbf{r}', \mathbf{R}') = E\Psi(\mathbf{r}', \mathbf{R}')$$

where the perturbation operator to the molecular wavefunctions in the rotating frame is

$$\hat{K} = \frac{\partial}{\partial R'} \left(R'^2 \frac{\partial}{\partial R'} \right) - (\hat{L}_{X'} - \hat{J}_{X'})^2 - (\hat{L}_{Y'} - \hat{J}_{Y'})^2$$

in terms of the operators \hat{L} and \hat{J} for the total and internal angular momentum \mathbf{L} and \mathbf{j} respectively of the collision system. Note $L_{Z'} = J_{Z'}$, for diatoms. An advantage of using this rotating system in the adiabatic treatment is that radial perturbations, which cause vibrational $v \rightarrow v'$ and electronic $nl \rightarrow n'l$ transitions, originate from the first term (radial) of \hat{K} while angular perturbations (torques) which causes rotational $j \rightarrow J'$ and electronic $nl \rightarrow n'l$ transitions originate from the angular momentum operator products $[\hat{L}_{X'}\hat{J}_{X'} + \hat{L}_{Y'}\hat{J}_{Y'}]$. The use of a rotating frame causes some complication, however, to the direct use of the asymptotic boundary condition for $\Psi(\mathbf{r}', \mathbf{R}')$.

B2.2.8.2 DIABATIC FORMULATION (POTENTIAL COUPLING SCHEME)

When relaxation of the internal motion is slow compared with the fast relative speed v_j , then Ψ is expanded in terms of the known unperturbed (diabatic) orthonormal eigenstates $\Phi_j(\mathbf{r}_A, \mathbf{r}_B) = \psi_i(\mathbf{r}_A)\phi_k(\mathbf{r}_B)$ of \hat{H}_{int} according to

$$\Psi(\mathbf{r}, \mathbf{R}) = \sum_j F_j(\mathbf{R}) \Phi_j(\mathbf{r}).$$

-46-

Substituting into the Schrödinger equation, multiplying by $\Phi_n^*(\mathbf{r})$ and integrating over \mathbf{r} , shows that the unknown functions $F_n(\mathbf{R})$ for the relative motion in channel n satisfy the infinite set of coupled equations

$$[\nabla_{\mathbf{R}}^2 + \mathcal{K}_n^2(\mathbf{R})]F_n(\mathbf{R}) = \sum_{j \neq n} U_{nj}(\mathbf{R})F_j(\mathbf{R}).$$

The reduced potential matrix elements which couple the internal states n and j are

$$U_{nj}(\mathbf{R}) = \frac{2M_{AB}}{\hbar^2} V_{nj}(\mathbf{R}) = U_{jn}^*(\mathbf{R})$$

where the electrostatic interaction averaged over states n and j is

$$V_{nj}(\mathbf{R}) = \int \Phi_n^*(\mathbf{r}) V(\mathbf{r}, \mathbf{R}) \Phi_j(\mathbf{r}) d\mathbf{r}.$$

The local wavenumber K_n of relative motion under the static interaction V_{nn} is given by

$$\mathcal{K}_n^2(\mathbf{R}) = k_n^2 - U_{nn}(\mathbf{R}).$$

The diagonal elements U_{nn} are the *distortion* matrix elements which distort the relative motion from plane waves in elastic scattering, while the off-diagonal matrix elements, U_{ij} and U_{ji} , U_{jf} which couple states i and f either directly or via intermediate channels j cause inelastic scattering and polarization contributions to elastic scattering. In contrast to the *adiabatic formulation*, radial and angular transitions originate in the *diabatic formulation* from the radial and angular components to the potential coupling elements $V_{nj}(\mathbf{R})$. The set of coupled are solved subject to the usual asymptotic ($R \rightarrow \infty$) requirement that

$$F_j(\mathbf{R}) \sim \exp(ik_i Z) \delta_{ij} + f_{ij} \exp(ik_j R)/R$$

for the elastic $i = j$ and inelastic $i \neq j$ scattered waves. In terms of the amplitude f_{ij} for scattering into direction (θ, ϕ) , the differential and integral cross sections for $i \rightarrow j$ transitions are

$$\frac{d\sigma_{ij}}{d\Omega} = \frac{v_j}{v_i} |f_{ij}(\theta, \phi)|^2$$

and

$$\sigma_{ij} = \frac{v_j}{v_i} \int_0^\pi d(\cos \theta) \int_0^{2\pi} |f_{ij}(\theta, \phi)|^2 d\phi.$$

As well as obtaining the scattering amplitude from the above asymptotic boundary conditions, f_{if} can also be obtained from the *integral representation* for the scattering amplitude is

$$f_{if}(\theta) = \langle \Phi_f(\mathbf{r}) \exp(i\mathbf{k}_f \cdot \mathbf{R}) | V(\mathbf{r}, \mathbf{R}) | \Psi(\mathbf{r}, \mathbf{R}) \rangle_{\mathbf{r}, \mathbf{R}}$$

B2.2.8.3 INELASTIC SCATTERING BY A CENTRAL FIELD

When the atom–atom or atom–molecule interaction is spherically symmetric in the channel vector \mathbf{R} , i.e. $V(\mathbf{r}, \mathbf{R}) = V(r, R)$, then the orbital l and rotational j angular momenta are each conserved throughout the collision so that an ℓ -partial wave decomposition of the translational wavefunctions for each value of j is possible. The translational wave is decomposed according to

$$F_j(\mathbf{R}) = \frac{4\pi N}{k_i R} \sum_{\ell, m} i^\ell F_{j\ell}(R) Y_{\ell m}^*(\hat{\mathbf{k}}) Y_{\ell m}(\hat{\mathbf{R}})$$

and inserted into the *adiabatic* set of coupled equations (of [section B2.2.8.2](#)). The radial wavefunction $F_{j\ell}$ is then the solution of

$$\frac{d^2 F_{j\ell}}{dR^2} + \left[k_i^2 - \left\{ U_{ii}(R) + \frac{\ell(\ell+1)}{R^2} \right\} \right] F_{j\ell}(R) = \sum_{j \neq i} U_{ij}(R) F_{j\ell}(R)$$

which is the direct generalization of the quantal radial equation for potential scattering to directly include other channels $j \neq i$. The coupled equations are now solved subject to the requirements that

$$F_{i\ell}(k_i R) \sim \sin(k_i R - \ell\pi/2) + \left\{ \frac{T_{ii}^\ell}{2i} \right\} e^{i(k_i R - \ell\pi/2)}$$

for the elastic scattered wave and

$$F_{j\ell}(k_j R) \sim \left(\frac{k_i}{k_j} \right)^{\frac{1}{2}} \left\{ \frac{T_{ij}^\ell}{2i} \right\} e^{i(k_j R - \ell\pi/2)}$$

for the inelastic wave. The transition-matrix elements for elastic and inelastic scattering are

$$T_{ij}^\ell = -\frac{2i}{(k_i k_j)^{\frac{1}{2}}} \int_0^\infty F_{j\ell}^{(0)}(r) U_{ji}(r) F_{i\ell}(r) dr$$

where $F_{j\ell}^{(0)} = (k_f r) j_\ell(k_f r)$ and $F_{i\ell}(r)$ are the solutions of these coupled radial equations. The differential cross section for inelastic scattering is

$$\frac{d\sigma_{ij}}{d\Omega} = (1/4k_i^2) \left| \sum_{\ell=0}^{\infty} (2\ell + 1) T_{ij}^{\ell} P_{\ell}(\cos \theta) \right|^2.$$

The integral inelastic cross section is

$$\sigma_{ij}(E) = \frac{\pi}{k_i^2} \sum_{\ell=0}^{\infty} (2\ell + 1) |T_{ij}^{\ell}|^2.$$

The transition matrix $\mathbf{T}^{\ell} = \{T_{ij}^{\ell}\}$ is symmetrical, $T_{ij}^{\ell} = T_{ji}^{\ell}$, and the cross sections satisfy detailed balance. Each transition matrix element $|T_{ij}^{\ell}|^2$ is the probability of an $i \rightarrow j$ transition in the target for each value ℓ of the (orbital) angular momentum of relative motion.

B2.2.8.4 TWO-STATE TREATMENT

Here all couplings are ignored except the direct couplings between the initial and final states as in a two-level atom. The coupled equations to be solved are

$$\begin{aligned} [\nabla^2 + k_i^2 - U_{ii}(\mathbf{R})]\psi_i(\mathbf{R}) &= U_{if}(\mathbf{R})\psi_f(\mathbf{R}) \\ [\nabla^2 + k_f^2 - U_{ff}(\mathbf{R})]\psi_f(\mathbf{R}) &= U_{fi}(\mathbf{R})\psi_i(\mathbf{R}). \end{aligned}$$

(A) DISTORTED-WAVE APPROXIMATION

Here all matrix elements in the two-level equations (section B2.2.8.4) are included, except the back coupling $V_{if}\Psi_f$ term which provides the influence of the inelastic channel on the elastic channel and is required to conserve probability. Distortion of the elastic and outgoing inelastic waves by the averaged (static) interactions V_{ii} and V_{ff} respectively is therefore included. The two-state equations can then be decoupled and effectively reduced to one-channel problems. An analogous static-exchange distortion approximation, where exchange between the incident and one of the target particles also follows from the two-level treatment.

(B) BORN APPROXIMATION

Here the distortion (diagonal) and back coupling matrix elements in the two-level equations (section B2.2.8.4) are ignored so that $\psi_i(\mathbf{R}) = \exp(i\mathbf{k}_i \cdot \mathbf{R})$ remains an undistorted plane wave. The asymptotic solution for ψ_f when compared with the asymptotic boundary condition then provides the Born elastic ($i = f$) or inelastic scattering amplitudes

$$f_{ij}^B(\theta, \phi) = -\frac{1}{4\pi} \frac{2M_{AB}}{\hbar^2} \int V_{fi}(\mathbf{R}) e^{i\mathbf{k} \cdot \mathbf{R}} d\mathbf{R}.$$

The momentum change resulting from the collision is $\mathbf{Q} = \hbar\mathbf{K}$ where $\mathbf{k} = \mathbf{k}_i - \mathbf{k}_f$. The Born amplitude also follows by inserting $\psi(\mathbf{r}, \mathbf{R}) = \Phi_1(\mathbf{r}) \exp i(\mathbf{k}_1 \cdot \mathbf{R})$ in the integral representation. Comparison with potential scattering shows that the elastic scattering of structured particles occurs in the Born approximation via the averaged electrostatic interaction $V_{ii}(\mathbf{R})$.

For electron-ion or ion-ion collisions, the plane waves $\exp(i\mathbf{k}_i \cdot \mathbf{R})$ are simply replaced by Coulomb waves to

provide the Coulomb–Born approximation.

B2.2.8.5 EXACT RESONANCE

The two-state equations of section B2.2.8.4 cannot, in general, be solved analytically except for the specific case of *exact resonance* when $k_i = k_f = k$ and $U_{ii} = U_{ff} = U$, $U_{if} = U_{fi}$. Then the equations can be decoupled by introducing the linear combinations $\psi^\pm(\mathbf{R}) = \frac{1}{\sqrt{2}}[\psi_i(\mathbf{R}) \pm \psi_f(\mathbf{R})]$, so the two-state set can be converted to two one-channel decoupled equations

$$[\nabla^2 + k^2 - (U \pm U_{if})]\psi^\pm(\mathbf{R}) = 0.$$

The problem has therefore been reduced to potential scattering by the interactions $U_\pm = (U \pm U_{if})$ associated with elastic scattering amplitudes f^\pm . Hence the elastic ($i = f$) and ‘inelastic’ ($i \neq f$) amplitudes are

$$f_{ii} = (f^+ + f^-)/2 \quad f_{if} = (f^+ - f^-)/2.$$

In terms of the phase shifts η_l^\pm associated with potential scattering by U_\pm , the amplitudes for elastic and inelastic scattering are then

$$f_{in}(\theta) = \frac{1}{2ik} \sum_{l=0}^{\infty} (2l+1) [(e^{2i\eta_l^+} + e^{2i\eta_l^-})/2 - 1] P_l(\cos \theta)$$

and

$$f_{if}(\theta) = \frac{1}{2ik} \sum_{l=0}^{\infty} (2l+1) [(e^{2i\eta_l^+} - e^{2i\eta_l^-})/2 - 1] P_l(\cos \theta).$$

The corresponding differential cross sections $|f_{if}|^2$ will therefore exhibit interference oscillations. The integral cross sections are

$$\sigma_{ii} = \frac{4\pi}{k^2} \sum_{l=0}^{\infty} (2l+1) \left[\left(\frac{1}{2} \sin^2 \eta_l^+ + \sin^2 \eta_l^- \right) / - \frac{1}{4} \sin^2(\eta_l^+ - \eta_l^-) / 4 \right]$$

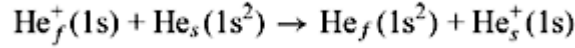
and

$$\sigma_{if} = \frac{\pi}{k^2} \sum_{l=0}^{\infty} (2l+1) \sin^2(\eta_l^+ - \eta_l^-)$$

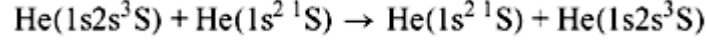
respectively.

(A) EXAMPLES: ATOMIC COLLISIONS WITH IDENTICAL NUCLEI

Important cases of *exact resonance* are the *symmetrical resonance charge transfer* collision



which converts a fast ion beam f to a fast neutral beam and the excitation transfer collision



which transfers the internal excitation in the projectile beam fully to the target atom. The electronic molecular wavefunctions divide into even (gerade) or odd (ungerade) classes upon reflection about the mid-point of the internuclear line ($\mathbf{R} \rightarrow -\mathbf{R}$). In the separated atom limit, $\psi_{g,u} \sim \phi(r_A) \pm \phi(r_B)$. The potentials U_{\pm} in the former case are the gerade and ungerade interactions $V_{g,u}$. The phase shifts for elastic scattering by the resulting gerade (g) and ungerade (u) molecular potentials of A_2^+ are, respectively, η_{ℓ}^g and η_{ℓ}^u . The charge transfer (X) and transport cross sections are then

$$\begin{aligned}\sigma_X(E) &= \frac{\pi}{k^2} \sum_{\ell=0}^{\infty} (2\ell + 1) \sin^2(\eta_{\ell}^g - \eta_{\ell}^u) \\ \sigma_{u,g}^{(1)}(E) &= \frac{4\pi}{k^2} \sum_{\ell=0}^{\infty} (\ell + 1) \sin^2(\beta_{\ell} - \beta_{\ell+1}) \\ \sigma_{u,g}^{(2)}(E) &= \frac{4\pi}{k^2} \left(\frac{3}{2}\right) \sum_{\ell=0}^{\infty} \frac{(\ell + 1)(\ell + 2)}{(2\ell + 3)} \sin^2(\beta_{\ell} - \beta_{\ell+2}).\end{aligned}$$

-51-

For ungerade potentials, $\beta_{\ell} = \eta_{\ell}^g$ for ℓ even and η_{ℓ}^u for ℓ odd. For gerade potentials, $\beta_{\ell} = \eta_{\ell}^u$ for ℓ even and η_{ℓ}^g for ℓ odd. The diffusion cross section $\sigma_{u,g}^{(1)}$ contains (g/u) interference. The viscosity cross section $\sigma_{u,g}^{(2)}$ does not. For charge transfer between the heavier rare gas ions Rg^+ with their parent atoms Rg , the degenerate states at large internuclear separations are not s states but p states. The states are then $\Sigma_{g,u}$ which arise from the p state with $m = 0$ and $\Pi_{g,u}$ which arises from $m = \pm 1$ with space quantization along the molecular axis. Since there is no coupling between molecular states of different electronic angular momentum, the scattering by the $^2\Sigma_{g,u}$ pair and the $^2\Pi_{g,u}$ pair of Ne_2^+ potentials (for example) is independent. The cross section is therefore the combination

$$\sigma_{el,X}(E) = \frac{1}{3}\sigma_{\Sigma}(E) + \frac{2}{3}\sigma_{\Pi}(E)$$

of cross sections σ_{Σ} and σ_{Π} for the individual contributions arising from the isolated $^2\Sigma_{g,u}$ and $^2\Pi_{g,u}$ states to elastic el or charge-transfer X scattering. See [12, 13] for further details on excitation-transfer and charge-transfer collisions.

(B) SINGLET-TRIPLET SPIN-FLIP CROSS SECTION

This cross section is

$$\sigma_{\text{ST}}(E) = \frac{\pi}{k^2} \sum_{\ell=0}^{\infty} (2\ell + 1) \sin^2(\ell^s - \ell^t)$$

where $\eta_{\ell}^{s,t}$ are the phase shifts for individual potential scattering by the singlet and triplet potentials, respectively.

B2.2.8.6 PARTIAL WAVE ANALYSIS

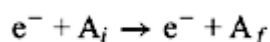
In order to reduce the three-dimensional *diabatic* or *adiabatic* set of coupled equations for atom–atom and atom–molecule scattering to a corresponding working set of coupled radial equations, analogous to those in [section B2.2.8.3](#), the orbital angular momentum l of relative motion must be distinguished from the combined internal angular momentum j associated with the internal (rotational and electronic) degrees of freedom of the partners A and B at rest at infinite separation R . Both the orbital angular momentum l of relative motion and the internal angular momentum j of the atomic electrons or of molecular rotation are in general coupled. The total angular momentum $J = l + j$ and its component J_z along some fixed direction (of incidence) are each conserved. Angular momentum may therefore be exchanged between the internal (rotational) and translational (orbital) degrees of freedom via the couplings $V_{\text{nm}}(\mathbf{R})$ or $\hat{\mathbf{k}}$. Partial wave analysis is an exercise in angular momentum coupling and is well-established (e.g. [14]) for both the diabatic and adiabatic treatments of heavy-particle collisions.

-52-

B2.2.9 ELECTRON–ATOM INELASTIC COLLISIONS

B2.2.9.1 CLOSE-COUPPLING EQUATIONS FOR ELECTRON–ATOM (ION) COLLISIONS

A partial wave decomposition provides the full close-coupling quantal method for treating A–B collisions, electron–atom, electron–ion or atom–molecule collisions. The method [15] is summarized here for the inelastic processes



at collision speeds less or comparable with those target electrons actively involved in the transition. It is based upon an expansion of the total wavefunction Ψ for the $(e^- - A)$ - multi-electron system in terms of a sum of products of the known atomic target state wavefunctions $|\Phi_i\rangle$ and the unknown functions $F_i(r)$ for the relative motion. Here

$$\Psi^\Gamma(\mathbf{r}_1, \dots, \mathbf{r}_N; \mathbf{r}) = \mathcal{A} \sum_i \Phi_i^\Gamma(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N; \hat{\mathbf{r}}) \frac{1}{r} F_i^\Gamma(r)$$

involves a sum over all discrete and an integral over the continuum states of the target. The operator \mathcal{A} antisymmetrizes the summation with respect to exchange of all pairs of electrons in accordance with the Pauli exclusion principle. The angular and spin momenta (denoted collectively by $\hat{\mathbf{r}}$) of the projectile electron have been coupled with the orbital and spin angular momenta of the target states $|\Phi_i\rangle$ to produce the ‘channel functions’ $\Phi_i^{LS\pi}(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_i; \hat{\mathbf{r}})$ which are eigenstates of the total orbital L , total spin S angular momentum, their Z -components M_L, M_S and parity π . The set $\Gamma \equiv LSM_L M_S \pi$ of quantum numbers are therefore conserved throughout the collision. By substituting the expansion for Ψ^Γ into the Schrödinger equation.

$$H_{N+1}\Psi^\Gamma = \left[\sum_{i=1}^{N+1} \left(-\frac{1}{2}\nabla_i^2 - \frac{Z}{r_i} \right) + \sum_{i>j=1}^{N+1} \frac{1}{r_{ij}} \right] \Psi^\Gamma$$

expressed in *atomic units*, the radial functions for the motion of the scattered electron satisfy the infinite set of coupled *integro-differential* equations

$$\left[\frac{d^2}{dr^2} + k_i^2 - \frac{\ell_i(\ell_i + 1)}{r^2} + \frac{2(Z - N)}{r} \right] F_i^\Gamma(r) = 2 \sum_j [V_{ij}^\Gamma(r) + W_{ij}^\Gamma(r)] F_j^\Gamma(r).$$

The direct potential couplings are represented by

$$V_{ij}^\Gamma(r) = Z_P \left[\frac{Z_i}{r} \delta_{ij} + \sum_{k=1}^N \left\langle \Phi_i^\Gamma \left| \frac{1}{|\mathbf{r}_k - \mathbf{r}|} \right| \Phi_j^\Gamma \right\rangle \right].$$

-53-

The non-local exchange couplings are represented by

$$W_{ij} F_j^\Gamma(r) = \sum_{k=1}^N \left\langle \Phi_i^\Gamma \left| \frac{1}{|\mathbf{r}_k - \mathbf{r}|} \right| (\mathcal{A} - 1) \Phi_j^\Gamma F_j^\Gamma \right\rangle.$$

The direct potential gives rise to the long-range polarization attraction which is very important for low-energy scattering. The exchange potentials are short range and are extremely complicated. Additional non-local potentials that arise from various correlations (which cannot be included directly but which can be constructed from pseudostates) can also be added to the right-hand side of the equations.

Numerical solution of this set of *close-coupled equations* is feasible only for a limited number of close target states. For each N , several sets of independent solutions F_{ij} of the resulting close-coupled equations are determined subject to $F_{ij} = 0$ at $r = 0$ and to the reactance \mathbf{K} -matrix asymptotic boundary conditions,

$$F_{ij}^\Gamma \sim \sin \theta_i \delta_{ij} + K_{ij}^\Gamma \cos \theta_i$$

for n open channels characterized by $k_i^2 = 2(E - E_i) > 0$. The argument is

$$\theta_i = k_i r - \frac{1}{2} \ell_i \pi + \frac{Z - N}{k_i} \ln(2k_i r) + \sigma_i$$

where ℓ_i is the orbital angular momentum of the scattered electron and where $\sigma_i = \arg \Gamma[\ell_i + 1 - i(Z - N)k_i]$ is the Coulomb phase. For closed channels, $k_i^2 < 0$ $F_{ij} \sim C_{ij} \exp(-|k_i|r)$ as $r \rightarrow \infty$. The scattering amplitude can then be expressed in terms of the elements T_{ij} of the $(n \times n)$ \mathbf{T} -matrix which is related to the \mathbf{K} and \mathbf{S} matrices by,

$$\mathbf{T}^\Gamma = \frac{2i\mathbf{k}^\Gamma}{\mathbf{I} - i\mathbf{K}^\Gamma} = \mathbf{S}^\Gamma - \mathbf{I}.$$

The integral cross section for the transition $i \equiv \alpha_a L_i S_i \rightarrow f \equiv \alpha_f L_f S_f$ in the target atom, where α denotes the additional quantum numbers required to completely specify the state, is then

$$\sigma_{if}(k_i^2) = \frac{\pi}{k_i^2} \sum_{L, S, \pi, \ell_i, \ell_f} \frac{(2L+1)(2S+1)}{2(2L_i+1)(2S_i+1)} |T_{ij}^\Gamma|^2.$$

According to detailed balance, the collision strength

$$\Omega_{if} = k_i^2 (2L_i + 1)(2S_i + 1) \sigma_{if}(k_i^2)$$

-54-

is therefore dimensionless and is symmetric with respect to $i \rightarrow f$ interchange. Further extensions, simplifications and calculational schemes of the basic close-coupling and related methods are found in [15, 16 and 17].

With modern high-speed computers, it is feasible to solve the coupled set of radial equations only for a restricted basis set of unperturbed states $\Phi_n(\mathbf{r})$ regarded as being closely and strongly coupled. For electron-atom (molecule) collisions at low energies E , the full quantal close-coupling method is extremely successful in predicting the cross sections and shapes and widths of resonances which appear at energies E just below the various thresholds for excitation of the various excited levels. As E increases past the threshold for ionization, it becomes less successful, and is plagued by problems with convergence both in the number of the basis states and in the number of partial waves used in the expansion for ψ^Γ . Other methods for intermediate and high energies are therefore preferable. For heavy-particle collisions and for electron collisions at high energies, semiclassical versions (in section B2.2.10) of the close-coupling equations can be derived.

B2.2.9.2 CLOSE COUPLING WITH PSEUDOSTATES AND CORRELATION

(A) PSEUDOSTATES

A partial acknowledgment of the influence of higher discrete and continuum states, not included within the wavefunction expansion, is to add, to the truncated set of basis states, functions of the form $\Psi_p(\mathbf{r})\Phi_p(\mathbf{r})$ where Φ_p is not an eigenfunction of the internal Hamiltonian \hat{H}_{int} but is chosen so as to represent some appropriate average of bound and continuum states. These pseudostates can provide full polarization distortion to the target by incident electrons and allows flux to be transferred from the the open channels included in the truncated set.

(B) CORRELATION

When the initial and final internal states of the system are not well-separated in energy from other states then the closed-coupling calculation converges very slowly. An effective strategy is to add a series of correlation terms involving powers of the distance r_{ij} between internal particles of projectile and target to the truncated close-coupling expansion which already includes the important states.

B2.2.9.3 THE R-MATRIX METHOD

This method, introduced originally in an analysis of nuclear resonance reactions, has been extensively developed [15, 16 and 17] over the past 20 years as a powerful *ab initio* calculational tool. It partitions configuration space into two regions by a sphere of radius $r = a$, where r is the scattered electron coordinate. In the internal region $r > a$, the electron-atom complex behaves almost as a bound state so that a configuration

interaction expansion of the total wavefunction ψ , as in atomic structure calculations, is appropriate. In the external region the scattered electron moves in the long-range multipole potential contained in the direct electrostatic interaction, and can be accurately represented by a perturbation approach. See [15, 16 and 17] for further details, for other modern quantal approximations and for various computational methods useful for electron–atom collisions over a wide energy range.

-55-

B2.2.9.4 ELECTRON–MOLECULE COLLISIONS

The close-coupling equations are also applicable to electron–molecule collision but severe computational difficulties arise due to the large number of rotational and vibrational channels that must be retained in the expansion for the system wavefunction. In the fixed nuclei approximation, the Born–Oppenheimer separation of electronic and nuclear motion permits electronic motion and scattering amplitudes $f_{nn'}(\mathbf{R})$ to be determined at fixed internuclear separations R . Then in the adiabatic nuclear approximation the scattering amplitude for $i \equiv n, v, J \rightarrow n', v', J' \equiv f$ transitions is

$$f_{if}(\mathbf{0}) = \langle \chi_{n'v'}(\mathbf{R}) Y_{J'M'}(\hat{\mathbf{R}}) | f_{nn'}(\mathbf{r}) | \chi_{nv}(\mathbf{R}) Y_{JM}(\hat{\mathbf{R}}) \rangle$$

and cross sections can be obtained. See [15] for further details.

B2.2.10 SEMICLASSICAL INELASTIC SCATTERING

The term *semiclassical* is used in scattering theory to denote many different situations.

- (a) The use of some time-dependent classical path $\mathbf{R}(t)$ within a time-dependent quantal treatment of the response of the internal degrees of freedom of A and B to the time-varying field $V(\mathbf{R}(t))$ created by the approach of A towards B along the classical trajectory $\mathbf{R}(t)$. This procedure generalizes classical theory for potential scattering to structured collision partners and inelastic transitions.
- (b) The use of the three-dimensional *eikonal-phase* $S(\mathbf{R})$, which is the solution of the *Hamilton–Jacobi equation*, for the channel wavefunction $\psi^+(\mathbf{R})$, within the full quantal expression for the cross section.
- (c) The use of JWKB approximate solutions of the radial *Schrödinger equation* for the radial wavefunction $R_{\epsilon, \ell}$ for A–B relative motion within the full quantum treatment of the A–B collision.

B2.2.10.1 CLASSICAL PATH THEORY

The basic assumption here is the existence over the inelastic scattering region of a common classical trajectory $\mathbf{R}(t)$ for the relative motion under an appropriately averaged *central* potential $\bar{v}[R(t)]$. The interaction $V[\mathbf{r}, \mathbf{R}(t)]$ between A and B may then be considered as time-dependent. The system wavefunction therefore satisfies

$$i\hbar \frac{\partial \Psi(\mathbf{r}, t)}{\partial t} = [\hat{H}_{\text{int}}(\mathbf{r}) + V(\mathbf{r}, \mathbf{R}(t))] \Psi(\mathbf{r}, t)$$

-56-

and can be expanded in terms of the eigenfunctions Φ_n of \hat{H}_{int} as

$$\Psi(\mathbf{r}, t) = \sum_n A_n(t) \Phi_n(\mathbf{r}) \exp(-iE_n t).$$

The transition amplitudes A_n then satisfy the set

$$i\hbar \frac{\partial A_n(b, t)}{\partial t} = \sum_n A_j(b, t) V_{nj}(\mathbf{R}(t)) \exp(i\omega_{nj} t)$$

of first-order equations coupled by the matrix elements $V_{nj}(\mathbf{R})$ between states n and j with energy separation $\hbar\omega_{nj} = E_n - E_j$. Once the classical trajectory $\mathbf{R} \equiv (R(t), \theta(t), \phi = \text{constant})$ is determined from the classical equations

$$\begin{aligned} \frac{dR}{dt} &= \pm v[1 - b^2/R^2 - \bar{V}(R)/E]^{1/2} \\ \frac{db}{dt} &= \frac{vb}{R^2} \end{aligned}$$

of motion for impact parameter b and kinetic energy $E = \frac{1}{2}M_{\text{AB}}v^2$, the coupled equations are solved subject to the requirement $A_n(b, t \rightarrow -\infty) = \delta_{ni}$. Since the probability for an $i \rightarrow f$ transition is $P_{if} = |A_f(b, t \rightarrow \infty)|^2$, the differential cross section for inelastic scattering is

$$\frac{d\sigma_{if}}{d\Omega} = \sum_n P_{if}(b_n, \phi) \left\{ \frac{d\sigma_{\text{el}}}{d\Omega} \right\}$$

where $d\sigma_{\text{el}}/d\Omega$ is the differential cross section $|bdb/d(\cos\theta)|$ for elastic scattering by $\bar{V}(R)$ and where the summation is over all trajectories b_n which pass through (θ, ϕ) . The integral cross section is

$$\sigma_{if} = 2\pi \int_0^\infty |A_f(b, \infty)|^2 b db.$$

IMPACT PARAMETER METHOD

This normally refers to the use of the straight-line trajectory $R(t) = (b^2 + v^2 t^2)^{1/2}$, $\theta(t) = \arctan(b/vt)$ within the classical path treatment. See Bates [18, 19] for examples and further discussion.

B2.2.10.2 LANDAU-ZENER CROSS SECTION

The Landau-Zener transition probability is derived from an approximation to the full two-state impact-parameter treatment of the collision. The single passage probability for a transition between the diabatic surfaces $H_{11}(\mathbf{R})$ and $H_{22}(\mathbf{R})$ which cross at R_X is the Landau-Zener transition probability

$$P_{12}(R_X; b) = 1 - \exp(-2\pi |H_{12}(R_X)|^2 / \hbar v_X |H'_{11} - H'_{22}|)$$

where H_{12} is the interaction coupling states 1 and 2. The diabatic curves are assumed to have linear shapes in the vicinity of the crossing at R_X , i.e. ($H'_{11} - H'_{22} = \Delta F$) and H_{12} is assumed constant. The adiabatic surfaces

$$W^\pm = \frac{1}{2}(H_{11} + H_{22}) \pm \frac{1}{2}[(H_{11} - H_{22})^2 + 4H_{12}]$$

do not cross (avoided crossing). They are separated at R_X by $w^+ - w^- = 2H_{12}(R_X)$. The probability for remaining on the adiabatic surface is $p_{12}(R_X)$. The probability for remaining on the diabatic surface or for pseudocrossing between the adiabatic curves is $1 - P_{12}(R_X)$. The overall transition probability for both the incoming and outgoing legs of the trajectory $\mathbf{R}(t)$ is then

$$\mathcal{P}_{12} = 2P_{12}(1 - P_{12}).$$

The Landau–Zener cross section is

$$\sigma_{12} = 4\pi \int_0^{R_X} P_{12}(1 - P_{12})b \, db$$

where the variation of P_{12} on impact parameter b arises from the speed $v_X \approx v_0(1 - b^2/R_X^2)^{1/2}$ at the crossing point R_X . For rectilinear trajectories $\mathbf{R} = \mathbf{b} + v_0 \mathbf{t}$,

$$\sigma_{12}(v_0) = 4\pi R_X^2 [E_3(\alpha) - E_3(2\alpha)]$$

where $E_n(\alpha) = \int_1^\infty y^{-n} \exp(-\alpha y) dy$ is the exponential integral with argument $\alpha = 2\pi |H_{12}|^2 / \hbar V_0 \Delta f$. See Nikitin [20, 21] for more elaborate models which include interference effects arising from the phases or eikonals associated with the incoming and outgoing legs of the trajectory.

B2.2.10.3 EIKONAL THEORIES

Here the relative motion wavefunction $F_n(\mathbf{R})$ is decomposed as [22]

$$F_n(\mathbf{R}) = A_n(\mathbf{R}) \exp iS_n(\mathbf{R}) \exp(-\chi_n(\mathbf{R})).$$

-58-

The classical action, or solution of the Hamilton–Jacobi equation $\nabla S_n(\mathbf{R}) = \kappa_n(\mathbf{R})$, for relative motion under the channel interaction $V_{nn}(\mathbf{R})$, is

$$S_n(\mathbf{R}) = \mathbf{k}_n \cdot \mathbf{R} + \int_{R_0}^R (\kappa_n - \mathbf{k}_n) \cdot d\mathbf{R}_n$$

where R_0 is the initial point on the associated trajectory $\mathbf{R}_n(t)$ where $\kappa_n = \mathbf{k}_n$. The current \mathbf{j}_n in channel n , assumed elastic, satisfies the conservation condition $\nabla \cdot \mathbf{j}_n = 0$, so that χ_n is the solution of

$$\nabla_{\mathbf{R}}^2 S_n - 2(\nabla_{\mathbf{R}} S_n) \cdot (\nabla_{\mathbf{R}} \chi_n) = 0.$$

Flux in channel n is therefore lost only via transition to another state f with a probability controlled solely by A_f . When many wavelengths of relative motion can be accommodated within the range of V_{nn} , as at the higher energies favoured by the diabatic scheme, the fast \mathbf{R} -variation of F_n is mainly controlled by S_n , and the original diabatic set of coupled equations then reduce to the simpler set

$$i\hbar \frac{\partial A_f(t)}{\partial t} = \sum_{n \neq f} A_n(t) V_{fn}(\mathbf{R}_f(t)) \exp i(S_n - S_f) \exp -(\chi_n - \chi_f)$$

of first-order coupled equations. When a common trajectory $\mathbf{R}_n(t) = \mathbf{R}(t)$ under some averaged interaction $\bar{V}(\mathbf{R})$ can be assumed for all channels n then

$$S_n(\mathbf{R}) - S_f(\mathbf{R}) = \omega_{fn} t + \hbar^{-1} \int_{t_0}^t [V_{ff}(\mathbf{R}(t)) - V_{nn}(\mathbf{R}(t))] dt$$

and the classical path equations are recovered [22].

(A) AVERAGED POTENTIAL

The orbit common to all channels is found by choosing the potential governing the relative motion as the average [22]

$$\mathcal{V}(\mathbf{R}) = \langle \Psi(\mathbf{r}, \mathbf{R}) | \hat{H}_{\text{int}}(\mathbf{r}) + V(\mathbf{r}, \mathbf{R}) | \Psi(\mathbf{r}, \mathbf{R}) \rangle_{\mathbf{r}}.$$

Hamilton's equations of motion for this interaction

$$\mathcal{V}(\mathbf{R}) = \sum_n \left[|A_n|^2 + \sum_f A_f^* A_n V_{fn} \exp i(S_n - S_f) \right]$$

-59-

are therefore coupled to the set of first-order equations for the transition amplitudes $A_f(\mathbf{R})$. An essential feature is that total energy is always conserved, being continually redistributed between the relative motion and internal degrees of freedom, as motion along the trajectory proceeds. In terms of the solutions $A_f(b_j, t)$ and the differential cross section $d\sigma_{\text{el}}^{(j)}$ for elastic scattering of particles with impact parameter $b_j(\theta)$ through θ by $\bar{V}(\mathbf{R})$, the semiclassical scattering amplitude is

$$f_{if}^j(\theta) = A_f[b_j(\theta), \infty] \exp iS'_f(b_j(\theta), \infty) \{d\sigma_{\text{el}}^{(j)}/d\Omega\}^{1/2}.$$

The accumulated classical action for orbit $b_j(\theta)$ is

$$S'_f(b_j, t) = \int_{\mathbf{R}_0}^{\mathbf{R}(t)} [\kappa_f(\mathbf{R}) - k_n(\mathbf{R})] \cdot d\mathbf{R}.$$

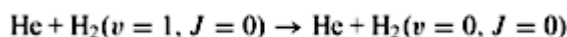
When the same scattering angle originates from more than one impact parameter b_j , then interference effects originate from the different actions associated with the different orbits $b_j(\theta)$. The contributions arising from N -

orbits which are well-separated combine according to

$$f_{if}(\theta) = -i \sum_{j=1}^N \alpha_j \beta_j f_{if}^j(\theta).$$

The coefficients $\alpha_j = \exp\pm\pi/4$ depends on whether the scattered particle emerges on the same side (+) of the axis as it entered, as in a collision overall repulsive, or on the opposite side (-), as in an overall attractive collision. The coefficients β_j is $\exp\pm\pi/4$ according to whether the sign of $db/d\theta$ is (+) or (-). The differential cross section will therefore exhibit characteristic oscillations, directly attributable to interference between the action phases $S_r^j(b_j)$ associated with each contributing classical path $b_j(\theta)$. The analysis can be extended, as in the uniform Airy function approximation to cover orbits which are not widely separated, as for the case of rainbow scattering or of caustics, in general, where the density of paths become infinite. This theory provides the basis of the multistate orbital treatment [22] which is successful for rotational and vibrational excitation in atom-molecule and ion-molecule collisions at higher energies $E_{AB} \geq 11$ eV. Other semiclassical treatments based on the JWKB approximation to the corresponding set of coupled equations for the radial wavefunction for relative motion can be found in [23, 24 and 25].

In figure B2.2.3 cross sections for the quenching process



for collision energies E ranging from the ultracold to 1 keV are displayed. The full quantal results [26] are shown together with those calculated [27] from the semiclassical multistate orbital method [22]. It is seen that results from both methods complement and connect with each other very well, in that the quantal treatment is computationally feasible up to $E \sim 1$ eV while semiclassical procedures are feasible at the higher collision energies.

-60-

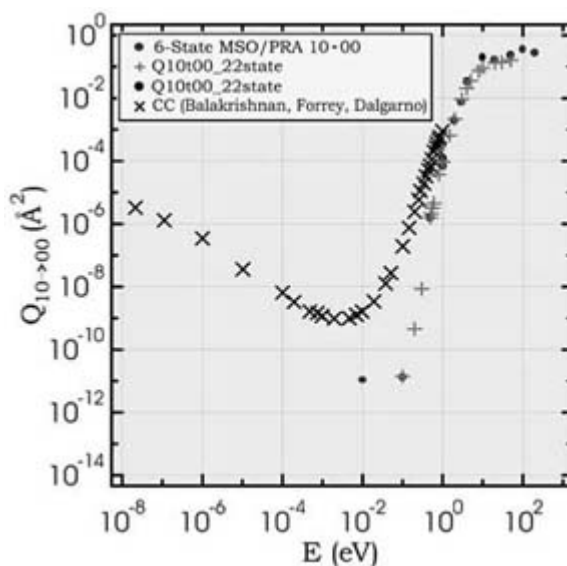


Figure B2.2.3. Vibrational relaxation cross sections (quantal and semiclassical) as a function of collision energy E .

(B) MULTICHANNEL EIKONAL METHOD

For electronic transitions in electron–atom and heavy-particle collisions at high impact energies, the major contribution to inelastic cross sections arises from scattering in the forward direction. The trajectories implicit in the action phases and set of coupled equations can be taken as rectilinear. The integral representation

$$f_{if}(\theta) = \langle \phi_f(\mathbf{r}) \exp(i\mathbf{k}_f \cdot \mathbf{R}) | V(\mathbf{r}, \mathbf{R}) | \Psi(\mathbf{r}, \mathbf{R}) \rangle_{\mathbf{r}, \mathbf{R}}$$

for the scattering amplitude, where

$$\Psi(\mathbf{r}, \mathbf{R}) = \sum_n A_n(\mathbf{R}) \phi_n(\mathbf{r}) \exp iS_n(\mathbf{R})$$

then provides the basis of the multichannel eikonal treatment [28] valuable, in particular, for heavy-particle collisions and for electron (ion)–excited atom collisions where, due to the large effect of atomic polarization (charge-induced dipole), the collision is dominated by scattering in the forward direction.

B2.2.11 LONG-RANGE INTERACTIONS

B2.2.11.1 POLARIZATION, ELECTROSTATIC AND DISPERSION INTERACTIONS

The long-range interaction $V(\mathbf{R})$ between two atomic/molecular species can be decomposed into

-61-

$$V_{\text{polarization}}(\mathbf{R}) + V_{\text{electrostatic}}(\mathbf{R}) + V_{\text{dispersion}}(\mathbf{R}).$$

The polarization interaction arises from the interaction between the ion of charge Ze and the multipole moments it induces in the atom or molecule AB. The dominant polarization interaction is the ion-induced dipole interaction

$$V_{\text{pol}}(Ze; \text{ind}D) = -\frac{\alpha_d Ze^2}{2R^4} [1 + (\alpha'_d/\alpha_d) P_2(\hat{\mathbf{s}} \cdot \hat{\mathbf{R}})]$$

where the averaged dipole polarizability is $\alpha_d = (\alpha_{\parallel} + 2\alpha_{\perp})/3$ and α_{\parallel} and α_{\perp} are the polarizabilities of AB in the directions parallel and perpendicular to the molecular axis $\hat{\mathbf{s}}$ of AB. The anisotropic polarizability is $\alpha'_d = 2(\alpha_{\parallel} - \alpha_{\perp})/3$. The next polarization interaction is the charge-induced quadrupole interaction, averaged over all molecular orientations

$$V_{\text{pol}}(Ze; \text{ind}Q) = -\bar{\alpha}_q Ze^2 / 2R^6$$

where $\bar{\alpha}_q$ is the averaged quadrupole polarizability. Additional polarization terms arise from permanent multipole moments of one partner and the dipole (or multipole) it induces in the other, averaged over all directions. The leading term is

$$V_{\text{pol}}(D; \text{ind}D) = -\frac{1}{R^6} (D_n^2 \bar{\alpha}_{di} + D_i^2 \bar{\alpha}_{dn})$$

where the subscripts i and n label the permanent dipole moments D and the dipole-polarizabilities $\alpha_{,i}$ of the

ion and neutral, respectively. The variation R_{-6} is similar to that for the charge-induced quadrupole interaction.

The electrostatic interaction results from the interaction of the ion with the permanent multipole moments of the neutral. For cylindrically symmetric neutrals or linear molecules, the ion–neutral multipole interaction is

$$V_{el}(Ze; \mathbf{D}, \mathbf{Q}) = -\frac{(Ze)D_n}{R^2} P_1(\hat{\mathbf{s}} \cdot \hat{\mathbf{R}}) + \frac{(Ze)Q_n}{R^3} P_2(\hat{\mathbf{s}} \cdot \hat{\mathbf{R}})$$

where $D_n = \int \mathbf{r} \rho(\mathbf{r}) d\mathbf{r}$ and $Q_n = \int (3z^2 - r^2) \rho(\mathbf{r}) d\mathbf{r}$ are the permanent dipole and quadrupole moments of the neutral. The ion dipole–neutral dipole interaction is

$$V_{el}(\mathbf{D}_i; \mathbf{D}_n) = -\frac{D_i D_n}{R^3} [2 \cos \theta_i \cos \theta_n - \sin \theta_i \sin \theta_n \cos(\phi_i - \phi_n)]$$

where θ_i and θ_n are the angles made by the ionic and molecular dipoles \mathbf{D}_i and \mathbf{D}_n and the line \mathbf{R} of centres and ϕ_i and ϕ_n are the azimuthal angles of rotation about the line of centres. The dipole–molecular quadrupole interaction is

$$V_{el}(\mathbf{D}_i, \mathbf{Q}_n) = \frac{3D_i Q_n}{2R^4} [(3 \cos^2 \theta_n - 1) \cos \theta_i - 2 \sin \theta_i \sin \theta_n \cos \theta_n \cos(\phi_i - \phi_n)].$$

-62-

The dispersion interaction arises between the fluctuating multipoles and the moments they induce and can occur even between spherically symmetric ions and neutrals. Thus,

$$V_{\text{dispersion}} \sim -\frac{C_6}{R^6} - \frac{C_8}{R^8} - \frac{C_{10}}{R^{10}} \dots$$

represents the interaction of the fluctuating dipole interacting with the induced dipole C_6 term and quadrupole C_8 term, respectively. The leading R^{-6} term represents the van der Waal's attraction.

REFERENCES

- [1] Catchen G L, Husain J and Zare R N 1978 *J. Chem. Phys.* **69** 1737
- [2] Mason E A and McDaniel E W 1988 *Transport Properties of Ions in Gases* (New York: Wiley)
- [3] Viehland L A, Mason E A, Morrison W F and Flannery M R 1975 *At. Data Nucl. Data Tables* **16** 495
- [4] Rodberg L S and Thaler R M 1967 *Introduction to the Quantum Theory of Scattering* (New York: Academic) p 226
- [5] Mott N F and Massey H S W 1965 *The Theory of Atomic Collisions* 3rd edn (Oxford: Clarendon)
- [6] Goldberger M L and Watson K M 1964 *Collision Theory* (New York: Wiley)
- [7] Newton R G (ed) 1966 *Scattering Theory of Waves and Particles* (New York: McGraw-Hill)
- [8] Joachain C J 1975 *Quantum Collision Theory* (Amsterdam: North-Holland) p 383
- [9] Flannery M R 1980 *Phys. Rev. A* **22** 2408
- [10] Flannery M R and Vrinceanu D 2000 *Phys. Rev. Lett.* **85**
- [11] Vrinceanu D and Flannery M R 1999 *Phys. Rev. A* **6** 1053
- [12] Bransden B H and McDowell M R C 1992 *Charge Exchange and the Theory of Ion–Atom Collisions* (Oxford: Clarendon)
- [13] Bransden B H 1983 *Atomic Collision Theory* 2nd edn (Menlo Park, CA: Benjamin-Cummings)
- [14] Child M S 1996 *Molecular Collision Theory* (New York: Dover)

- [15] Burke P G 1996 *Atomic, Molecular and Optical Physics Handbook* ed G W Drake (New York: American Institute of Physics Press) ch 45
- [16] Bartschat K (ed) 1996 *Computational Atomic Physics* (New York: Springer)
- [17] McCarthy I E and Weigold E 1995 *Electron-Atom Collisions* (Cambridge: Cambridge University Press)
- [18] Bates D R 1961 *Quantum Theory I. Elements* ed D R Bates (New York: Academic) ch 8
- [19] Bates D R (ed) 1962 *Atomic and Molecular Processes* (New York: Academic) ch 14
- [20] Nikitin E E 1996 *Atomic, Molecular and Optical Physics Handbook* ed G W Drake (American Institute of Physics Press) ch 47
- [21] Nikitin E E and Umanskiĭ S Ya 1984 *Theory of Slow Atomic Collisions* (Berlin: Springer)
- [22] McCann K J and Flannery M R 1978 *J. Chem. Phys.* **12** 5275
McCann K J and Flannery M R 1975 *J. Chem. Phys.* **63** 4695
- [23] Bates D R and Crothers D S F 1970 *Proc. R. Soc. A* **315** 465
- [24] Crothers D S F 1971 *Adv. Phys.* **20** 405
- [25] Child M S 1991 *Semiclassical Mechanics with Molecular Applications* (Oxford: Clarendon)
-

-63-

- [26] Balakrishnan N, Forrey R C and Dalgarno A 1998 *Phys. Rev. Lett.* **80** 3224
- [27] Flannery M R and McCann K J, in preparation (unpublished)
- [28] Flannery M R and McCann K J 1974 *Phys. Rev.* **9** 1947
-

FURTHER READING

- Drake G W (ed) 1996 *Atomic, Molecular and Optical Physics Handbook* (American Institute of Physics Press)
- McDaniel E W and Mansky E J 1994 Guide to bibliographies, books, reviews and compendia of data on atomic collisions *Advances in Atomic and Molecular Physics* vol 33, ed B Bederson and H Walther p 389
- Bates D R and Estermann I (eds) 1965 *Advances in Atomic and Molecular Physics* vol 1
- Bates D R and Estermann I (eds) 1973 *Advances in Atomic and Molecular Physics* vol 1
- Bates D R and Bederson B (eds) 1974 *Advances in Atomic and Molecular Physics* vol 10
- Bates D R and Bederson B (eds) 1988 *Advances in Atomic and Molecular Physics* vol 25
- Bates D R and Bederson B (eds) 1989 *Advances in Atomic Molecular and Optical Physics* vol 26
- Bates D R and Bederson B (eds) 1993 *Advances in Atomic Molecular and Optical Physics* vol 31
- Bederson B and Dalgarno A (eds) 1994 *Advances in Atomic Molecular and Optical Physics* vol 32
- Bederson B and Walther H (eds) 1994 *Advances in Atomic Molecular and Optical Physics* vol 33
- Bederson B and Walther H (eds) 1998 *Advances in Atomic Molecular and Optical Physics* vol 38
- Scoles G (ed) 1988 *Atomic and Molecular Beam Methods* (New York: Oxford University Press)
- Baer M (ed) 1985 *Theory of Chemical Reaction Dynamics* (Boca Raton, FL: CRC Press) vols 1-4
- Bernstein R B (ed) 1979 *Atom-Molecule Collision Theory: A Guide for the Experimentalist* (New York: Plenum)
- Miller W H (ed) 1976 *Dynamics of Molecular Collisions, Parts A and B* (New York: Plenum)
- McDaniel E W and McDowell M R C (eds) 1969 *Case Studies in Atomic Collision Physics* (Amsterdam: North-Holland) vol 1

McDaniel E W and McDowell M R C (eds) 1972 *Case Studies in Atomic Collision Physics* (Amsterdam: North-Holland) vol 2

McDaniel E W, Čermák V, Dalgarno A, Ferguson E E and Friedman L (eds) 1970 *Ion-Molecule Reactions* (New York: Wiley)

Bates D R (ed) 1962 *Atomic and Molecular Processes* (New York: Academic)

Bates D R (ed) 1961 *Quantum Theory I. Elements* (New York: Academic)

-64-

Massey H S W, Burhop E H S and Gilbody H B (eds) 1969–74 *Electronic and Ionic Impact Phenomena* (Oxford: Clarendon) vols 1–5

McDaniel E W 1989 *Atomic Collisions: Electron and Photon Projectiles* (New York: Wiley)

McDaniel E W, Mitchell J B A and Rudd M E 1993 *Atomic Collisions: Heavy Particle Projectiles* (New York: Wiley)

-1-

B2.3 Reactive scattering

Paul J Dagdigan

B2.3.1 INTRODUCTION

Reactive scattering is one of a number of gas-phase phenomena included in the field of molecular collision dynamics, which is the study of the molecular mechanism of elementary physical and chemical rate processes. Other such dynamical processes include photodissociation, vibrational and rotational energy transfer, electronic quenching, unimolecular decay, reactions within weakly bound complexes and gas–surface interactions. The object of studying the dynamics of these processes is to gain an understanding of the behaviour of a system at the molecular level. We would like to unravel the forces exerted on the nuclei, as described by the potential energy surface (PES) of interaction, during the collisional encounter. We also wish to learn whether the system has jumped to another PES through an electronically non-adiabatic transition. In this section, techniques appropriate to the study of the dynamics of chemical reactions are emphasized. However, these techniques are generally applicable to the study of a variety of gas-phase collisional processes.

The implementation of molecular beam techniques and introduction of laser-based detection methods has allowed chemical reaction dynamics to be elucidated in far greater detail than is possible from inferences based on the temperature dependence of reaction rate constants. In an ideal crossed-beam reactive scattering experiment, which is illustrated schematically in [figure B2.3.1](#) two collimated beams of the reagents, of well defined velocities, are crossed in a collision centre, and the flux of reaction products in specified internal vibration–rotation quantum states scattered into particular solid angles is determined. In this way, the differential cross section for scattering of the reaction product in a given internal quantum state into a given solid angle element can be measured. There have been relatively few studies of reaction dynamics which have closely approximated this ideal, nor is the ideal experiment usually required in order to infer what one would like to understand about the dynamics of a particular reaction. While this section describes experimental techniques as applied to the study of chemically reactive collisions, these methods have also been applied to

the study of a wide range of collisional phenomena, including non-reactive energy transfer collisions, photodissociation, and gas–surface scattering.

Two radically different approaches have been taken for the study of reactive scattering. In the first, which approximates the ideal reactive scattering experiment, collimated beams are crossed in a high-vacuum chamber, and the products are detected with a rotatable detector. In most scattering experiments, a mass spectrometer, with electron bombardment ionization and mass-resolved detection of the ions transmitted through a radio-frequency (RF) electric quadrupole [1], is employed as the detector, and the products are identified from the mass-to-charge ratio of the detected ion, either the parent or a fragment ion. This detection method is ‘universal’ in that every atom or molecule can, in principle, be detected mass spectrometrically. An excellent description of such a molecular beam apparatus is given by Lee *et al* [2].

-2-

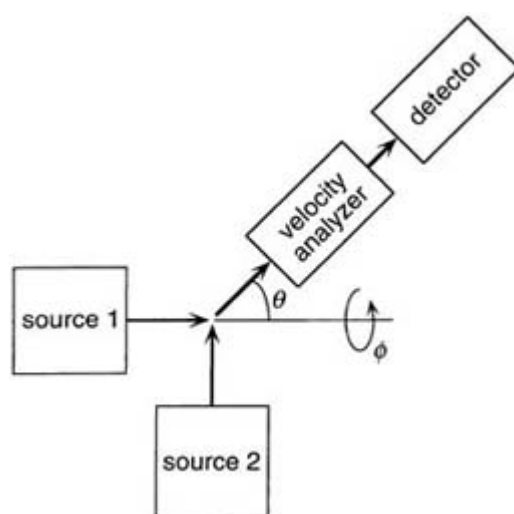


Figure B2.3.1. Schematic diagram of an idealized molecular beam scattering experiment.

In order to determine the partitioning of the energy available to the products into internal (vibrational and rotational) excitation and relative translational recoil energy of the products, the velocity of the detected products is determined, usually by a time-of-flight method [3]. In this way, the translational energy of the products can be determined. The mass spectrometer is essentially insensitive to the degree of internal excitation of the product, however, and the internal excitation of the products can be only determined indirectly, through energy conservation with the knowledge of the total energy available to the products (reaction exoergicity + translational and internal energy of the reagents).

The second approach to the study of reactive scattering involves the use of some spectroscopic method for the detection of the products in specified internal quantum states. Molecular spectroscopy is well suited to the determination of the relative populations in individual states since the quantum numbers of the upper and lower states of a molecular line in an assigned transition are known. Moreover, the intensities may be directly related to concentrations of specific internal states. The original implementation of this approach for the study of reactive scattering involved observation of spontaneous infrared emission from the radiative decay of vibrationally excited products [4, 5]. This approach is still being employed, however now usually with detection of the emission with Fourier transform [6], rather than grating-tuned spectrometers. In some cases, emission from electronically excited products can be observed for highly exothermic reactions.

For many reaction products and for the detection of molecules in their ground vibrational level, some laser-based spectroscopic method must be employed, rather than observation of spontaneous emission. The simplest spectroscopic method for determining concentrations of specified product internal states would involve the

application of the Beer–Lambert law on resolved molecular lines in direct absorption. However, the optical density of the product will be very small and limited by the requirement that the nascent reaction products do not undergo any secondary, relaxing collisions before being detected. In the gas phase, the collision frequency can be conveniently reduced by changing the density. The average time between collisions is increased by reducing the total pressure, and hence the concentration of the products. Very recently [7], an ultrasensitive absorption method has been developed and applied for the detection of reaction products.

-3-

A more sensitive method of detecting absorption, through observation of a so-called ‘action’ or ‘excitation’ spectrum, has been mainly employed for the detection of the reaction products. In most such experiments, a wavelength-tunable laser, usually a dye laser, is scanned over an electronic band system of the reaction product in question, and a signal indicative of molecular absorption is recorded. The relative intensities of spectral lines or bands are then converted into relative populations of the reaction product in specified internal quantum states. In this way the disposal of the available reaction energy into the internal degrees of freedom of this product is directly determined. If the accompanying product is an atom or a molecule with little internal excitation, the relative translational energy of the products can be obtained from energy conservation and knowledge of the total available energy. In this second approach, the angular distribution of the products is usually not determined. However, recent experiments employing Doppler resolution of isolated spectral lines have allowed determination of a low-resolution angular distribution of the product.

Most often, fluorescence excitation to an excited electronic state with the fundamental or frequency-doubled output of a wavelength-tunable dye laser has been employed for laser-based detection of the reaction products. In this method, the total, spectrally unresolved, photon emission from the detection zone is monitored as the laser wavelength is scanned over a molecular transition. Such an excitation spectrum provides the same information as would be available from an absorption spectrum, but with much higher detection sensitivity. This increased sensitivity arises from two factors. Fluorescence detection is a ‘zero-background’ technique and is limited only by background due to scattered light and quantum counting statistics, while absorption requires the measurement of ratios of signals. In addition, the sensitivity of fluorescence detection is greatly enhanced by the high spectral intensity of laser radiation, as opposed to incoherent radiation from lamps. Product internal state distributions have been determined with laser fluorescence detection in both beam- and bulb-type experiments.

The first half of this section discusses the use of the crossed beams method for the study of reactive scattering, while the second half describes the application of laser-based spectroscopic methods, including laser-induced fluorescence and several other laser-based optical detection techniques. Further discussion of both non-optical and optical methods for the study of chemical reaction dynamics can be found in articles by Lee [8] and Dagdigian [9].

B2.3.2 CROSSED-BEAMS METHOD

B2.3.2.1 THE BASIC SCATTERING EXPERIMENT AND SIGNAL INTENSITY

An ideal scattering experiment requires that the velocity spread of the reagent beams be narrow, so that the relative translational energy of the reagents is well defined. Effusive beams have a very broad velocity distribution, and their use in a scattering experiment usually required the insertion of a slotted-disk velocity selector [10] to reduce the velocity spread to a reasonable width. The flux in an effusive beam is not large, and the insertion of a velocity selector reduces the reagent beam flux significantly. The introduction of supersonic beam sources, with a dramatic narrowing of the velocity spread [11], radically increased the flux from beam sources and made the modern era of crossed-beam reactions possible. The term ‘supersonic’ refers to the fact that the molecular velocities in such a source are greater than the local speed of sound. In such a source and in

contrast to an effusive source, the backing pressure is high enough that the mean free path within the source is much smaller than the orifice diameter so that the gas behaves as a hydrodynamic fluid. Under ideal conditions the enthalpy is converted to net motion during the expansion of the gas into vacuum, and the local temperature becomes very low, leading to a very small spread of velocities about the mean.

-4-

In addition to obviating the need of velocity selection, the increased backing pressure over that attainable with an effusive source leads to significantly higher downstream beam densities.

A molecular beam scattering experiment usually involves the detection of low signal levels. Thus, one of the most important considerations is whether a sufficient flux of product molecules can be generated to allow a precise measurement of the angular and velocity distributions. The rate of formation of product molecules, dN/dt , can be expressed as

$$dN/dt = n_1 n_2 v_{rel} \sigma V_{coll} \quad (B2.3.1)$$

where the number densities of the reagent beams in the collision volume are given by n_1 and n_2 ; and v_{rel} , σ , V_{coll} are the relative velocity between the reagents, the integral reaction cross section, and the scattering volume, respectively. (Equation B2.3.1 is just a re-expression of the law of mass action since the product $v_{rel} \sigma$ is the microcanonical rate constant.) In an experiment with one supersonic beam and a velocity-selected effusive beam, typical values for the reagent beam densities are 10^{12} and 10^{10} molecules cm^{-3} , respectively, in a collision volume of 10^{-2} cm^3 . The relative velocity will be approximately 10^5 cm s^{-1} , and reaction cross section 10^{-15} cm^2 . This leads to an estimate of 10^{10} molecules s^{-1} for the total rate of production formation dN/dt .

The product molecules will scatter into a range of laboratory angles, depending upon the exoergicity of the reaction, the reaction dynamics, and kinematics, which is a function of the masses of the reagents and products. If we assume that the scattering is confined to 1 sr of solid angle (out of the total 4π), then the detector will receive $\sim 3 \times 10^6$ molecules s^{-1} if it subtends 1° in both directions (i.e. an angular acceptance of $1/3000$ sr). If, instead, the molecules were isotropically scattered, the detector would see a considerably smaller flux of $\sim 2 \times 10^5$ molecules s^{-1} . Of course, we desire not only the angular distribution of the products, but also the velocity distribution at each scattering angle, for a fuller understanding of the reaction dynamics.

If the molecules could be detected with 100% efficiency, the fluxes quoted above would lead to impressive detected signal levels. The first generation of reactive scattering experiments concentrated on reactions of alkali atoms, since surface ionization on a hot-wire detector is extremely efficient. Such detectors have been superseded by the 'universal' mass spectrometer detector. For electron-bombardment ionization, the rate of formation of the molecular ions can be written as

$$d[M^+]/dt = I_e \sigma [M] \quad (B2.3.2)$$

where I_e is the intensity of the electron beam (typically 10 mA cm^{-2} , or 6×10^{16} electrons $\text{cm}^2 \text{s}^{-1}$) and σ is the ionization cross section (typically 10^{-16} cm^2 for electrons of 150 eV energy). This leads to an estimated ionization rate of $k_i = I_e \sigma = 6$ s^{-1} . The molecules are, of course, not stationary in the ionization region but are travelling with a typical velocity of $\sim 5 \times 10^4$ cm s^{-1} . With an ionization region of length ~ 1 cm, the probability of ionization is thus estimated to be $\sim 10^{-4}$, which is a low detection efficiency. For example, the above quoted product flux of 3×10^6 molecules s^{-1} leads to only 360 detection ions s^{-1} . This ion count rate, or rates as low as even 1 ion s^{-1} , can be measured with good statistics in minutes if the background signal is not much larger than this. Hence, such beam-scattering experiments have been successful mainly through

careful reduction of the background ion count rate. This discussion of expected

-5-

product signal levels follows a discussion of intensities in crossed-beam reactive scattering experiments by Lee [8].

The background ion signal arises from two sources of molecules, namely the inherent background of molecules in a vacuum chamber and molecules effusing from the collision chamber into the detector chamber while the beams are on. The former arises from outgassing from the materials employed in the construction of the apparatus and the limitations in the pumps. The latter requires the careful design of differential pumping of the sources and the detector [2, 12].

B2.3.2.2 LABORATORY TO CENTRE-OF-MASS TRANSFORMATION

In a crossed-beam experiment the angular and velocity distributions are measured in the laboratory coordinate system, while scattering events are most conveniently described in a reference frame moving with the velocity of the centre-of-mass of the system. It is thus necessary to transform the measured velocity flux contour maps into the center-of-mass coordinate (CM) system [13]. Figure B2.3.2 illustrates the reagent and product velocities in the laboratory and CM coordinate systems. The CM coordinate system is travelling at the velocity c of the centre of mass

$$c = [m_1 v_1 + m_2 v_2] / (m_1 + m_2) \quad (\text{B2.3.3})$$

where the velocities of the reagents are v_1 and v_2 , with masses m_1 and m_2 , respectively. Thus, the velocities in the two coordinate systems are related by

$$v_i = c + u_i \quad \text{for } i = 1, 2. \quad (\text{B2.3.4})$$

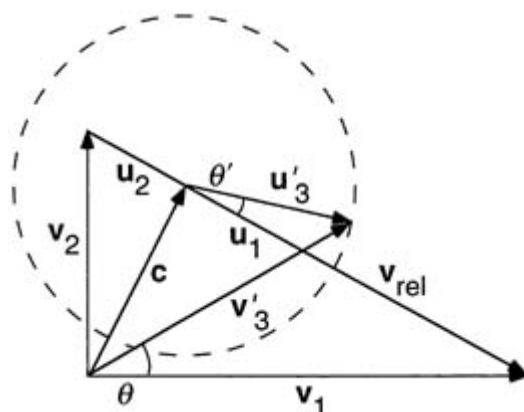


Figure B2.3.2. Velocity vector diagram for a crossed-beam experiment, with a beam intersection angle of 90° . The laboratory velocities of the two reagent beams are v_1 and v_2 , while the corresponding velocities in the centre-of-mass coordinate system are u_1 and u_2 , respectively. The laboratory and CM velocities for one of the products (assumed here to be in the plane of the reagent velocities) are denoted v'_3 and u'_3 , respectively. The dashed circle denotes the possible laboratory velocities v'_3 for the full range of CM scattering angles θ' .

-6-

The CM velocities are given by

$$\mathbf{u}_1 = m_2 \mathbf{v}_{\text{rel}} / (m_1 + m_2) \quad (\text{B2.3.5a})$$

$$\mathbf{u}_2 = -m_1 \mathbf{v}_{\text{rel}} / (m_1 + m_2) \quad (\text{B2.3.5b})$$

where $v_{\text{rel}} = v_1 - v_2$ is the relative velocity.

The relative translational energy of the reagents is given by

$$E_{\text{trans}} = \frac{1}{2} \mu v_{\text{rel}}^2. \quad (\text{B2.3.6})$$

The energy available to the product equals

$$E'_{\text{tot}} = E_{\text{trans}} + E_{\text{int}} + \Delta E \quad (\text{B2.3.7})$$

where E_{int} is the internal excitation energy of the reagents and ΔE is the reaction exoergicity. The energy E'_{tot} can be partitioned between translational and internal excitation of the products. The CM speed of one of the products can be expressed as

$$u'_3 = \frac{m_4}{m_3 + m_4} \left(\frac{2}{\mu'} (E'_{\text{tot}} - E'_{\text{int}}) \right)^{1/2}. \quad (\text{B2.3.8})$$

where m_3 is the mass of this product, m_4 is the mass of the other product, $\mu' = m_3 m_4 / (m_3 + m_4)$ is the reduced mass of the products, and E'_{int} is the internal excitation energy of the products.

It can be seen from [figure B2.3.2](#) that scattering angles (relative to the direction of one of the reagent beams) are different in the laboratory and CM coordinate systems. If the detected product is very heavy compared with its partner, or if its translational energy is very small, then its speed will be small compared with the speed of the centre of mass of the system. In this case, the product is scattered into a small range of scattering angles about c , and determination of the CM angular distribution will be difficult. Moreover, the scattered intensity at one laboratory scattering angle can come from two CM scattering angles, as can be seen in [figure B2.3.2](#). From the intensity estimates presented in [section B2.3.2.1](#), this concentration of the scattered product into a small laboratory angular range will facilitate detection of the product molecules. By contrast, if the product CM speed is large, then the product can be scattered into all laboratory angles.

In addition to transforming the velocities and scattering angles between the laboratory and CM frames, we must also consider the transformation of the cross sections

$$\left(\frac{d^2 \sigma}{dv' d\Omega} \right)_{\text{lab}} = \left(\frac{d^2 \sigma}{du' d\omega} \right)_{\text{CM}} \frac{du' d\omega}{dv' d\Omega} \quad (\text{B2.3.9})$$

where the last term on the right-hand side is the Jacobian of the transformation. Because the internal energies of the products are quantized, the velocity of a product scattered into a specific direction can have only discrete values. Equation (B2.3.9) is written with the velocities as continuous variables since the experimental resolution in reactive scattering experiments is not sufficient to resolve the discrete product velocities, due to the spread in the reagent beam velocities and angles. A detailed derivation of the Jacobian has been presented [13], and we obtain

$$\frac{du' d\omega}{dv' d\Omega} = \frac{v'^2}{u^2} \quad (\text{B2.3.10})$$

Equation (B2.3.10) shows that the scattered intensity observed in the laboratory is distorted from that in the CM coordinate system. Those products which have a larger laboratory velocity or a smaller CM velocity will be observed in the laboratory with a greater intensity.

The detection technique can also have an effect upon the angle- and velocity-dependent intensities. Cross sections refer to fluxes of molecules into a given range of velocities and angles. The commonly employed technique of mass spectrometric detection provides a measure of the density in the ionization region. Since density and flux are related by the velocity, we must include a factor of $1/v'$ in making the transformation indicated in equation (B2.3.10) from the CM cross sections to the measured laboratory intensities.

If the reagent velocity and angular spreads are sufficiently small, one can infer the CM angle-velocity distributions, i.e. the CM differential cross section on the right-hand side of equation (B2.3.10), directly from the measured laboratory intensities, by simply transforming the velocities to the CM frame and removing the transformation of the Jacobian, with inclusion of the velocity-dependent detection efficiency. On the other hand, as the scattering experiments are often carried out with limited resolution, it is usually necessary to deconvolute the results over the experimental spread [14]. More commonly, a forward convolution technique is employed, in which the CM angle-velocity distribution is adjusted until the laboratory distribution calculated by transformation of the coordinate system and convoluted over the experimental spreads agrees with the measured laboratory distribution.

B2.3.2.3 BEAM SOURCES

Many reactive scattering experiments involve the reaction of an atomic species, such as hydrogen, oxygen, a halogen, or a metal atom, with a stable molecular reagent. A variety of techniques have been employed for the generation of the reagent atomic beam. Beams of halogen atoms have been prepared by thermal dissociation. At room temperature, these elements exist as diatomic molecules, while the equilibrium is shifted toward the monatomic species at sufficiently high temperatures. A detailed description of such a source for the production of Cl, Br, and I beams is given by Valentini *et al* [15]. The atomic beam is prepared by heating the halogen molecule, diluted in a rare gas, to 2000 °C in a graphite tube. At this temperature, dissociation to atoms is essentially complete. In order to reduce the spread in velocities, the gas mixture is expanded supersonically into vacuum. Problems with materials corrosion have, until recently, limited the intensities of atomic fluorine beams. Use of a nickel tube limits the temperature to 700 °C, for

-8-

which dissociation yields of <15% are obtained. Recently a F atom source employing a tube made of single-crystal MgF₂ has been constructed, and dissociation fractions of ~80% have been achieved at tube temperatures near 1000 °C [16, 17].

Thermal dissociation is not suitable for the generation of beams of oxygen atoms, and RF [18] and microwave [19] discharges have been employed in this case. The first excited electronic state, O(¹D), has a different spin multiplicity than the ground O(³P) state and is electronically metastable. The collision dynamics of this very reactive state have also been studied in crossed-beam reactions with a RF discharge source which has been optimized for production of O(¹D) [20].

Beams of metal atoms have been prepared by many researchers through thermal vaporization from a heated crucible. An example of such a source, employed for the generation of beams of alkaline earth atoms, is described by Irvin and Dagdigian [21]. By striking an electrical discharge within this source, beams

containing electronically excited metastable atoms could be prepared. For Ca, the conversion efficiency to the $3s3p\ ^3P$ and $3s4d\ ^1D$ metastable excited states was of the order of 80%. Laser ablation from a solid has been widely used to generate atomic atoms of refractory elements [22]. A detailed description of such a source for the production of a beam of atomic carbon has been given [23]. This source has been employed in crossed-beam studies of reactions of carbon atoms.

Laser photolysis of a precursor may also be used to generate a reagent. In a crossed-beam study of the $D + H_2$ reaction [24], a hyperthermal beam of deuterium atoms (0.5 to 1 eV translational energy) was prepared by 248 nm photolysis of DI. This preparation method has been widely used for the preparation of molecular free radicals, both in beams and in experiments in a cell, with laser detection of the products. Laser photolysis as a method to prepare reagents in experiments in which the products are optically detected is further discussed below.

In most reactive scattering experiments, the reagent beam sources, which are housed in differentially pumped enclosures, are fixed and cross at a 90° intersection angle, while the detector is rotated about the scattering centre. The stable molecular co-reagent is usually produced in an effusive or supersonic source of the pure reagent. Care must be taken to ensure that no clusters are formed in the beam source, for example by heating the source or by limiting the total pressure behind the source orifice.

B2.3.2.4 AN EXAMPLE REACTION: $F + H_2 \rightarrow HF + H$

This reaction has been intensively studied because of its accessibility to both experimental and theoretical treatments. This reaction is also important because it is the pumping mechanism for the hydrogen fluoride infrared chemical laser. We present some data from the extensive study in 1985 by Lee and co-workers (Neumark *et al* [25, 26]) and recent, higher resolution experiments by Faubel *et al* [27, 28]. Figure B2.3.3 presents a schematic diagram of the apparatus in which Lee and co-workers carried out their experiments [29]. An effusive beam of fluorine atoms was prepared by thermal dissociation of F_2 in a nickel tube at $650\ ^\circ C$. The velocity spread was reduced to 11% by passage through a slotted-disk chopper. The fluorine atom beam was crossed with a supersonic molecular hydrogen beam, and the incident relative translational energy was varied by changing the temperature of the molecular beam source. The products were detected in a triply differentially pumped mass spectrometer employing electron-impact ionization. Cryogenically cooled surfaces also provided additional pumping to reduce the background signal, primarily from diffuse scattering from surfaces. The laboratory velocity distributions of the product at various laboratory scattering angles were measured by a time-of-flight method with a mechanical chopper. Crossed-beam scattering of a normal

hydrogen beam, a *para*-hydrogen beam (all molecules in the $j = 0$ rotational level) [25], and beams of D_2 and HD [26] was studied.

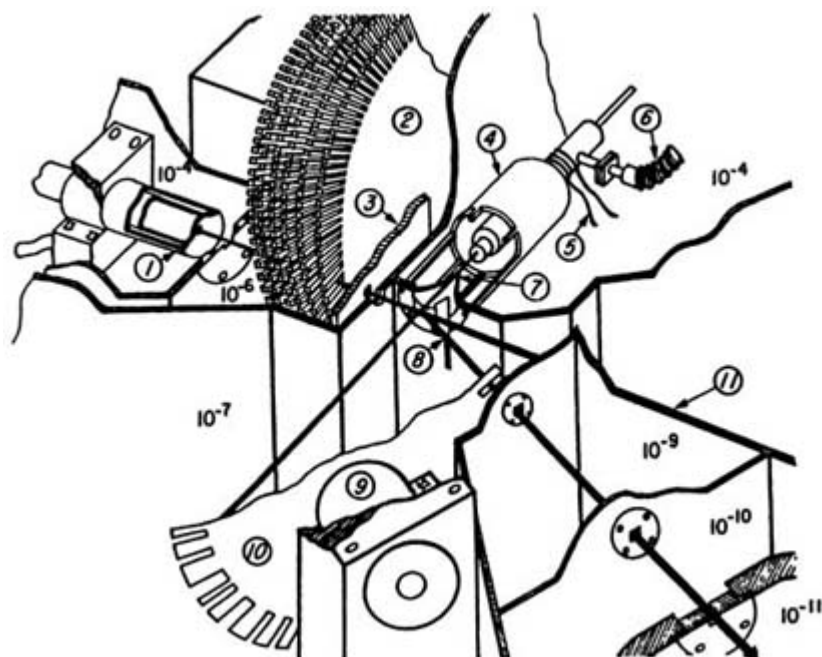


Figure B2.3.3. Crossed-molecular beam apparatus employed for the study of the $F + D_2 \rightarrow DF + D$ reaction. Indicated in the figure are: (1) the effusive F atom source; (2) slotted-disk velocity selector; (3) liquid-nitrogen-cooled trap; (4) D_2 beam source; (7) skimmer; (8) chopper; (9) cross-correlation chopper for product velocity analysis; and (11) rotatable, ultrahigh-vacuum, triply differentially pumped, mass spectrometer detector chamber. Reprinted with permission from Lee [29]. Copyright 1987 American Association for the Advancement of Science.

We present results on the $F + D_2$ reaction. Study of the reaction of the D_2 isotopic reagent was easier because the background in the mass spectrometer was smaller at mass 21 (DF) than for mass 20 (HF). Moreover, the masses of the DF and D are less dissimilar, so that the DF product is less kinematically constrained in the range of accessible laboratory scattering angles. [Figure B2.3.4](#) presents the laboratory angular distribution and a velocity vector diagram for the reaction, showing the accessible angular ranges for the product vibrational levels. From this plot, it already appears that the bulk of the DF products is made in the $v = 3$ and 4 vibrational levels that are scattered backward in the CM frame with respect to the incident F atom beam.

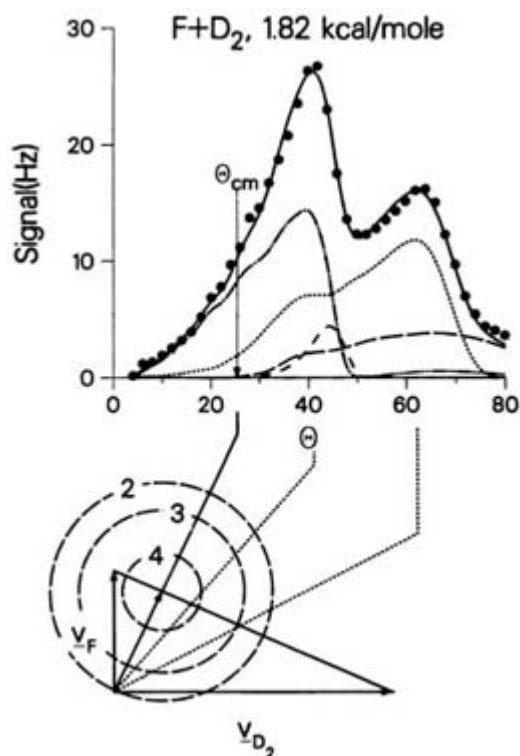


Figure B2.3.4. Laboratory angular distribution of DF products from the $F + D_2$ reaction at an incident relative translational energy of $1.82 \text{ kcal mol}^{-1}$ [26]. The full curve shows the fit with the derived CM angle-velocity contour. The angular distributions for the $v = 1, 2, 3,$ and 4 vibrational levels are indicated by (— —), (— —), (.....), and (— · —), respectively. (By permission from AIP.)

The above conjectures need to be verified by measurement of the doubly differential cross sections in angle and velocity. Typical time-of-flight distributions at several laboratory scattering angles from the more recent, higher resolution experiments of Faubel *et al* [27, 28] are presented in [figure B2.3.5](#). We see that products formed in the different vibrational levels appear at distinct time intervals, corresponding to different laboratory velocities. From data such as those presented in [figure B2.3.5](#) and after transformation from the laboratory to the CM frame, the CM velocity flux contour map is obtained. [Figure B2.3.6](#) displays such a contour plot derived by Lee and co-workers (Neumark *et al* [26]) for the $F + D_2$ reaction at one collision energy. It can be seen that all DF vibrational levels are predominantly scattered into the backward hemisphere. The CM angular spread is seen to be larger for the highest energetically accessible level, $v = 4$.

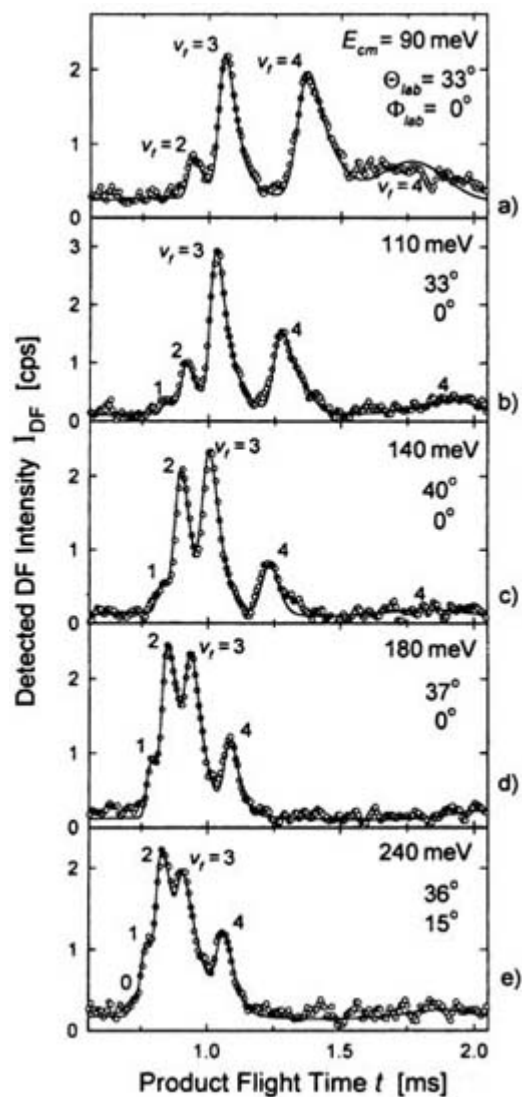


Figure B2.3.5. Typical time-of-flight spectra of DF products from the $F + D_2$ reaction [28]. The collision energies and in-plane (Θ_{lab}) and out-of-plane (Φ_{lab}) laboratory scattered angles are given in each panel. The DF product vibrational quantum number ν_f associated with each peak is indicated. Reprinted with permission from Faubel *et al* [28]. Copyright 1997 American Chemical Society.

-12-



Figure B2.3.6. CM angle–velocity contour plot for the $F + D_2$ reaction at an incident relative translational energy of 1.82 kcal mol [26]. Contours are given at equally spaced intensity intervals. This CM differential cross section was used to generate the calculated laboratory angular distributions given in figure B2.3.4. (By permission from AIP.)

Keil and co-workers (Dharmasena *et al* [16]) have combined the crossed-beam technique with a state-selective detection technique to measure the angular distribution of HF products, in specific vibration–rotation states, from the $F + H_2$ reaction. Individual states are detected by vibrational excitation with an infrared laser and detection of the deposited energy with a bolometer [30].

B2.3.2.5 PROBLEMS WITH PRODUCT IDENTIFICATION

It is well known that the electron-impact ionization mass spectrum contains both the parent and fragment ions. The observed fragmentation pattern can be useful in identifying the parent molecule. This ion fragmentation also occurs with mass spectrometric detection of reaction products and can cause problems with identification of the products. This problem can be exacerbated in the mass spectrometric detection of reaction products because these internally excited molecules can have very different fragmentation patterns than thermal molecules. The parent molecules associated with the various fragment ions can usually be sorted out by comparison of the angular distributions of the detected ions [8].

Many of the problems associated with electron impact ionization, such as the formation of fragment ions, can be alleviated by the use of photoionization detection. When the photon wavelength is tuned below the dissociative ionization threshold, it is possible to ionize the molecule ‘softly’, with the formation of parent ions only. This advantage for photoionization arises because the cross sections generally rise rapidly from the energetic threshold, namely the ionization potential. Recently, a molecular beam scattering apparatus using photoionization mass spectrometric detection of the products has been constructed [12]. This apparatus, shown schematically in figure B2.3.7 takes advantage of the intense vacuum ultraviolet (VUV) radiation available from a third-generation synchrotron radiation source at a national facility, the advanced light source, at the Lawrence Berkeley National Laboratory. In most scattering experiments, the detector is rotated about the crossing point of the fixed reagent beams. In this newly constructed apparatus, the detector includes an electron storage ring and cannot be rotated; in this case, it is the

sources that are rotated about the scattering centre. This apparatus makes liberal use of turbomolecular pumps, which can be positioned in any orientation and are convenient for vacuum pumps on rotating assemblies.

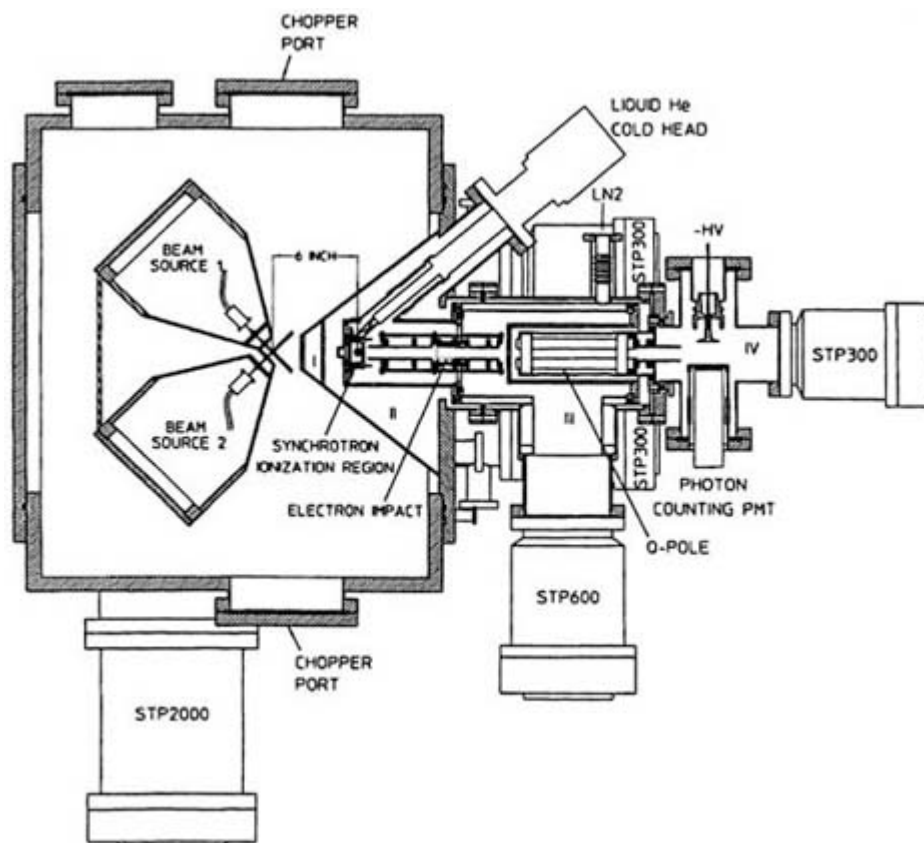


Figure B2.3.7. Schematic apparatus of crossed molecular beam apparatus with synchrotron photoionization mass spectrometric detection of the products [12]. To vary the scattering angle, the beam source assembly is rotated in the plane of the detector. (By permission from AIP.)

B2.3.3 OPTICAL DETECTION OF THE REACTION PRODUCTS

Optical methods, in both bulb and beam experiments, have been employed to determine the relative populations of individual internal quantum states of products of chemical reactions. Most commonly, such methods employ a transition to an excited electronic, rather than vibrational, level of the molecule. Molecular electronic transitions occur in the visible and ultraviolet, and detection of emission in these spectral regions can be accomplished much more sensitively than in the infrared, where vibrational transitions occur. In addition to their use in the study of collisional reaction dynamics, laser spectroscopic methods have been widely applied for the measurement of temperature and species concentrations in many different kinds of reaction media, including combustion media [31] and atmospheric chemistry [32].

B2.3.3.1 LASER-INDUCED FLUORESCENCE DETECTION

The most widely employed optical method for the study of chemical reaction dynamics has been laser-induced fluorescence. This detection scheme is schematically illustrated in the left-hand side of figure B2.3.8. A tunable laser is scanned through an electronic band system of the molecule, while the fluorescence emission is detected. This maps out an ‘action spectrum’ that can be used to determine the relative concentrations of the various vibration–rotation levels of the molecule.

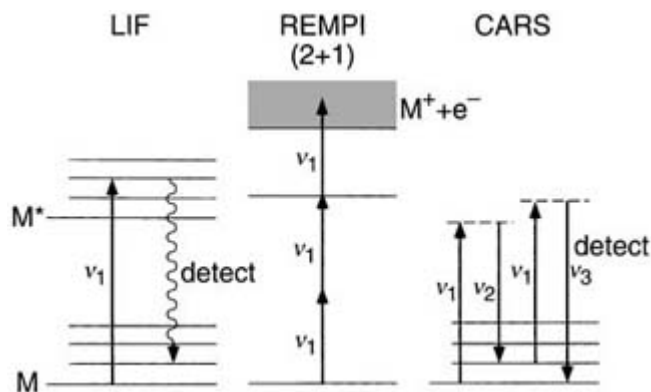


Figure B2.3.8. Energy-level schemes describing various optical methods for state-selectively detecting chemical reaction products: left-hand side, laser-induced fluorescence (LIF); centre, resonance-enhanced multiphoton ionization (REMPI); and right-hand side, coherent anti-Stokes Raman spectroscopy (CARS). The ionization continuum is denoted by a shaded area. The dashed lines indicate virtual electronic states. Straight arrows indicate coherent radiation, while a wavy arrow denotes spontaneous emission.

There are several requirements for this to be a suitable detection method for a given molecule. Obviously, the molecule must have a transition to a bound, excited electronic state whose wavelength can be reached with tunable laser radiation, and the band system must have been previously spectroscopically assigned. If the molecules are formed with considerable vibrational excitation, the available spectroscopic data may not extend up to these vibrational levels. Transitions in the visible can be accessed directly by the output of a tunable dye laser, while transitions in the ultraviolet can be reached by frequency-doubled radiation. The excited state must also have a reasonably short radiative lifetime (say $<10^{-5}$ s) with a near 100% fluorescence quantum yield (preferably independent of internal state). Finally, assignments of the individual rotational lines and the vibrational bands within the electronic transition must be available. These restrictions place considerable limits on the molecules which can be detected in this way—mainly diatomics and some triatomics. For incisive interpretation of the experimental observations, it is precisely those reactions involving small molecules whose collision dynamics can be treated theoretically with modern quantum mechanical methods.

Figure B2.3.9 presents a schematic diagram of a typical laser fluorescence experiment. In this apparatus, one of the reagents is prepared by photolysis of a suitable precursor [33] using radiation from an excimer laser (usually 248 nm from a KrF laser or 193 nm from a ArF laser). The tunable laser employed for fluorescence excitation counter-propagates along the beam of the excimer laser. Fluorescence of the product molecules is collected with a telescope and is imaged onto a photomultiplier. Because of their greater coverage of wavelengths, pulsed, rather than continuous (cw), lasers are almost universally employed. Thus, the photomultiplier output signal will typically appear at the 10–50

Hz repetition rate of the lasers and is usually sampled with a gated integrator, whose output is recorded with a laboratory computer. Analogue, rather than digital, electronics is usually employed because of pile-up of the detected photon counts in an experiment with reasonable product intensities.

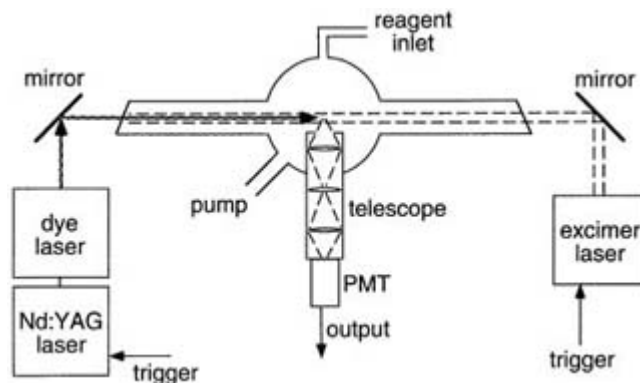


Figure B2.3.9. Schematic diagram of an apparatus for laser fluorescence detection of reaction products. The dye laser is synchronized to fire a short delay after the excimer laser pulse, which is used to generate one of the reagents photolytically.

The principal source of background in laser fluorescence detection is laser light scattered diffusely from optical elements such as windows, and from surfaces such as the walls of the apparatus. The windows for entry and exit of the laser beam, which are significant sources of scattered light, are usually mounted on long (0.4–0.8 m) sidearms. Baffles are installed in the sidearms to prevent light from scattering off the inside of the sidearms. Further reduction of scattered light can be achieved through the use of imaging optics (as illustrated in figure B2.3.9 to relay fluorescence from the excitation zone to the photomultiplier detector, so that emission from only a well defined volume is detected. If the fluorescence is mainly at wavelengths greatly different from the excitation wavelength, as would be the case, for example, for a molecule with significantly different equilibrium internuclear separations in the ground and excited electronic state, and hence a very non-diagonal Franck–Cordon array, then spectral filtering can provide further reduction in the background signal.

B2.3.3.2 DETERMINATION OF PRODUCT INTERNAL STATE DISTRIBUTIONS

Considerable spectroscopic data are required for the determination of the relative populations in the various internal quantum levels of the product from the relative intensities of various lines, or bands, in a spectrum. As discussed above, the spectrum must be assigned, i.e. the quantum numbers of the upper and lower levels of the spectral lines must be available. In addition to the line positions, intensity information is also required.

To compare the relative populations of vibrational levels, the intensities of vibrational transitions out of these levels are compared. Figure B2.3.10 displays typical potential energy curves of the ground and an excited electronic state of a diatomic molecule. The intensity of a (v', v'') vibrational transition can be written as

$$I(v', v'') = C N_{v''} p_{v', v''} \quad (\text{B2.3.11})$$

where $N_{v''}$ is the desired density of product molecules in the vibrational level v'' , $p_{v', v''}$ is the vibrational band strength [34, 35], and C is a proportionality constant. If the electronic transition moment [36] is constant as a function of the internuclear separation, then $p_{v', v''}$ is proportional to the Franck–Cordon factor $q_{v', v''}$, i.e. the square of the overlap integral of the upper and lower vibrational wavefunctions (see figure B2.3.10). The band strengths are usually determined experimentally from measurement of the decay lifetime of the excited vibrational level and the branching of emission into the various ground state vibrational levels v'' , as illustrated for the A–X transition of the hydroxyl radical [37]. Alternatively, if the potential energy curves of the two electronic states can be calculated from spectroscopic data, e.g. by the RKR method [38], then Franck–Cordon factors can be computed [35] and used to estimate the relative band strengths.

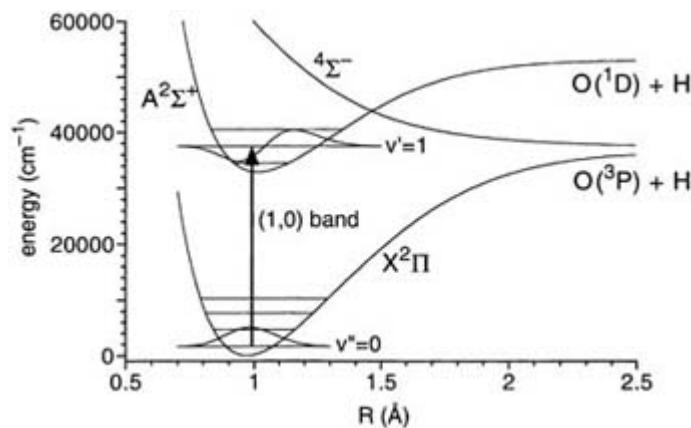


Figure B2.3.10. Potential energy curves [42] of the ground $X^2\Pi$ and excited $A^2\Sigma^+$ electronic states of the hydroxyl radical. Several vibrational levels are explicitly drawn in each electronic state. One vibrational transition is explicitly indicated, and the upper and lower vibrational wavefunctions are plotted. The upper and lower state vibrational quantum numbers are denoted v' and v'' , respectively. Also shown is one of the three repulsive potential energy curves which correlate with the ground $O(^3P) + H$ dissociation asymptote. These cause predissociation of the higher rotational and vibrational levels of the $A^2\Sigma^+$ state.

The $H + NO_2 \rightarrow OH + NO$ reaction provides an excellent example of the use of laser fluorescence detection for the elucidation of the dynamics of a chemical reaction. This reaction is a prototype example of a radical-radical reaction in that the reagents and products are all open-shell free radical species. Both the hydroxyl and nitric oxide products can be conveniently detected by electronic excitation in the UV at wavelengths near 226 and 308 nm, respectively. Atlases of rotational line positions for the lowest electronic band systems of these molecules ($A^2\Sigma^+ - X^2\Pi$ for both) are available [39, 40], and accurate band strengths for transition between various vibrational levels in the ground and excited electronic states have been reported [37, 41]. Because it is crossed by repulsive electronic states correlating with the ground state atoms $O(^3P) + H$ (see figure B2.3.10), the $OH(A^2\Sigma^+)$ state has low fluorescence quantum yields for rotational levels $N' > 25, 17,$ and 4 for vibrational levels $v' = 0, 1,$ and 2 , respectively [42]. This causes some problems for the detection of higher vibrational levels of the OH product since the intensities of the vibrational bands are strongest for the so-called diagonal bands, i.e. those for which $\Delta v = v' - v'' = 0$ [37]. Because of the excited-state predissociation, the higher levels must be detected through the weaker off-diagonal bands ($\Delta v < 0$).

The dynamics of the $H + NO_2$ reaction have been studied by several different techniques including laser fluorescence excitation of the products, infrared chemiluminescence, crossed-beam scattering, and electron paramagnetic resonance. We highlight the laser fluorescence studies by Sauder and Dagdigian [43] and by Irvine *et al* [44]. In the former experiment, the NO products, in vibrational levels $v \leq 2$, were detected at the intersection of an atomic hydrogen beam, generated by a microwave discharge source, with a pulsed beam of NO_2 diluted in Ar. Figure B2.3.11 illustrates an excitation spectrum for the detection of NO products in the ground ($v = 0$) vibrational level. The structure in the spectrum results from the energy differences between the rotational/fine-structure levels in the upper and lower electronic states and the degree of internal excitation of the products. Figure B2.3.12 illustrates the rotational transitions allowed in a $2\Sigma^+ - 2\Pi$ electronic transition. With such a diagram it is possible to determine the rotational/fine-structure level being detected through the excitation of a specific rotational line in the spectrum.

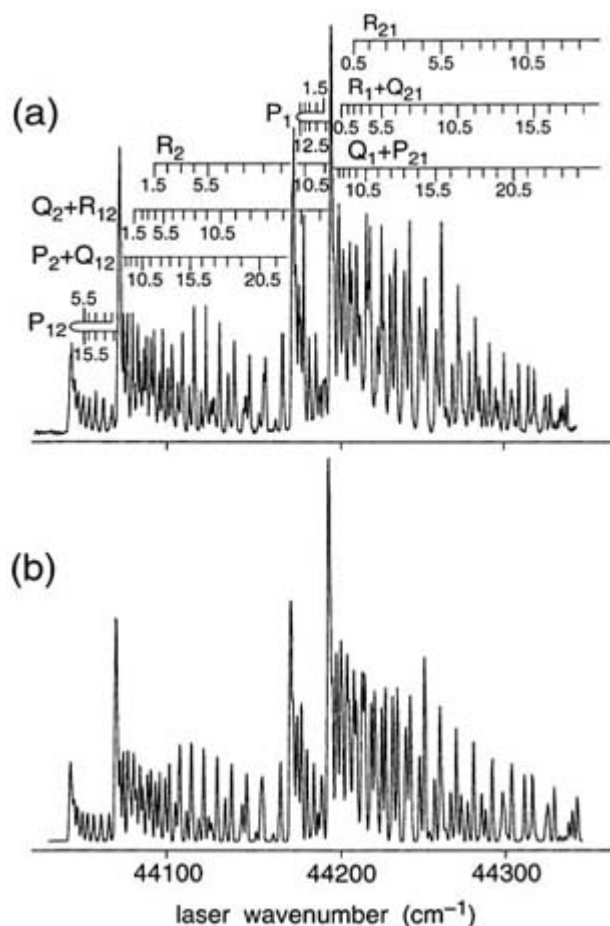


Figure B2.3.11. (a) Experimental laser fluorescence excitation spectrum of the $A\ ^2\Sigma^+ - X\ ^2\Pi\ (0,0)$ band for the NO product from the $H + NO_2$ reaction [43]. Individual lines in the various rotational branches are denoted by the total angular momentum J of the lower state. (b) Simulated spectrum with the NO rotational state populations adjusted to reproduce the spectrum in (a). (By permission from AIP.)

Both OH and NO are open-shell free radicals, with doublet electron spin multiplicity. Consequently, the coupling of the angular momentum of the unpaired electron with the angular momentum N of nuclear rotation leads to a more complicated rotational energy level pattern than for a closed-shell molecule ($^1\Sigma^+$ electronic state) [45]. For the upper, $^2\Sigma^+$ electronic state, the electron spin $S = \frac{1}{2}$ can couple with the rotational angular momentum to yield two fine-structure levels, with total angular momenta $J = N + \frac{1}{2}$ and $N - \frac{1}{2}$. These are conventionally [34] denoted F_1 and F_2 , respectively, in order of increasing energy for a given value of J . The rotational energy is given approximately by $BN(N + 1)$, where B is the rotational constant, and the splitting of the fine-structure levels is usually much smaller and grows with increasing N . This pattern of rotational/fine-structure levels is illustrated in the upper portion of figure B2.3.12.

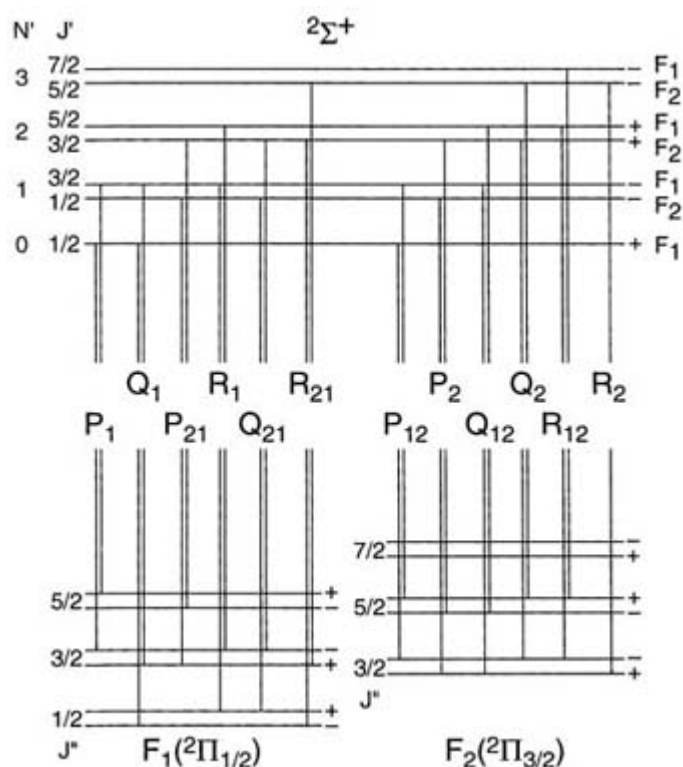


Figure B2.3.12. Rotational transitions between a specific pair of vibrational levels in a ${}^2\Sigma^+ - {}^2\Pi$ electronic transition. The total angular momentum J and the parity of the lowest rotational levels in each state are given. Hund's case (a) coupling is assumed for the ${}^2\Pi$ state. Conventional spectroscopic designations [34] are given for the allowed rotational transitions.

For the ${}^2\Pi$ state, the projection $\Lambda = 1$ of the electron orbital angular momentum along the internuclear axis can couple with the projection $\Sigma = \pm \frac{1}{2}$ to yield two spin-orbit levels, ${}^2\Pi_{\Omega}$, with $\Omega = \frac{1}{2}$ and $\frac{3}{2}$. The NO($X^2\Pi$) state follows so-called Hund's case (a) coupling [34], for which the spin-orbit splitting is much larger than the rotational energy. In this case, the rotational energy within each spin-orbit level is given approximately by $B[J(J+1) - \Omega^2]$, with half-integral $J \geq \Omega$. It should be noted that a Π state is orbitally degenerate, i.e. it has two components of the same energy. In Cartesian notation, these are often indicated as Π_x and Π_y . As a result, the rotational/fine-structure levels appear as nearly

degenerate pairs, with opposite parity, or symmetry with respect to reflection of all the coordinates through the space-fixed origin. These pairs of levels are called Λ -doublets.

For high rotational levels, or for a molecule like OH, for which the spin-orbit splitting is small, even for low J , the pattern of rotational/fine-structure levels approaches the Hund's case (b) limit. In this situation, it is not meaningful to speak of the projection quantum number Ω . Rather, we first consider the rotational angular momentum N exclusive of the electron spin. This is then coupled with the spin to yield levels with total angular momentum $J = N + \frac{1}{2}$ and $N - \frac{1}{2}$. As before, there are two nearly degenerate pairs of levels associated with each value of J .

The rotational/fine-structure levels of the lower, ${}^2\Pi$ electronic state in [figure B2.3.12](#) are drawn for a molecule near the case (a) limit since NO falls into this coupling scheme. Also indicated in the figure are the electric-dipole allowed rotational lines, indicated with conventional spectroscopic notation [34]. In the

spectrum displayed in [figure B2.3.11](#) the individual rotational lines appear to pile up into so-called ‘heads’ [34] at four distinct wavenumbers. The splitting between the two pairs of heads can be roughly identified with the NO(X ²Π) spin-orbit splitting.

In a conventional spectroscopic experiment, the intensity of a rotational transition within a given vibrational band can be written as

$$I(J', J'') = C' [N_{J''} (2J'' + 1)^{-1}] S_{J', J''} \quad (\text{B2.3.12})$$

where $N_{J''}$ is the density of the rotational/fine-structure level, J'' is its total angular momentum, $S_{J', J''}$ is the rotational line strength factor [34, 45], and C' is a proportionality constant. The relative intensities of the rotational lines can be used with equation (B2.3.12) to derive the rotational/fine-structure state distribution associated with a given vibrational level. Zare [45] presents a detailed discussion of the calculation of rotational line strength factors for diatomic electronic transitions.

Strictly speaking, equation (B2.3.12) does not apply to a measurement of the concentration through laser-induced fluorescence detection, as would be observed in the apparatus schematically illustrated in [figure B2.3.9](#). The rotational line strength factors $S_{J', J''}$ apply to the situation of isotropic irradiation and detection, which is clearly not the case for irradiation with a unidirectional polarized laser and detection of fluorescence emitted into a specific solid angle. Greene and Zare [46] have considered in detail the correct relationship between the molecular density and the intensity for arbitrary fluorescence excitation and detection geometries. In practice, because of the large angular momentum J often found for reaction products, the factors $S_{J', J''}$ follow fairly closely the J -dependence of the correct line strength factors, as long as lines in the same rotational branch are employed.

An additional inadequacy of equation (B2.3.12) is the assumption of an isotropic M_J distribution of product molecules in the detected rotational level. The product could be aligned because of the dynamics of the reaction. An extreme case is that of alignment imposed by kinematics for the mass combination $H + HL \rightarrow HH + L$, where H and L represent heavy and light atoms, respectively. From angular momentum conservation, we have $\mathbf{J}_{\text{tot}} = \mathbf{L}_i + \mathbf{J}_i = \mathbf{L}_f + \mathbf{J}_f$ where \mathbf{J}_{tot} is the total angular momentum of the system. Here, the vectors \mathbf{L} and \mathbf{J} are the orbital and rotational angular momenta, respectively, and the subscripts i and f denote the reagents and products, respectively. Because of the small moment of inertia of the HL reagent, \mathbf{J}_i will be small. Moreover, \mathbf{L}_f will also be small because of the small reduced mass of the $HH-L$ combination of the products. Thus, we have $\mathbf{L}_i = \mathbf{J}_f$. This then implies that the product rotational angular

-20-

momentum \mathbf{J}_f must be strongly polarized, with an anisotropic M_J distribution, since \mathbf{L}_i is perpendicular to the initial relative velocity.

The anisotropy of the product rotational state distribution, or the polarization of the rotational angular momentum, is most conveniently parametrized through multipole moments of the M_J distribution [45]. Odd multipoles, such as the dipole, describe the orientation of the angular momentum \mathbf{J} , i.e. which way the tips of the \mathbf{J} vectors preferentially point. Even multipoles, such as the quadrupole, describe the alignment of \mathbf{J} , i.e. the spatial distribution of the \mathbf{J} vectors, regarded as a collection of double-headed arrows. Orr-Ewing and Zare [47] have discussed in detail the measurement of orientation and alignment in products of chemical reactions and what can be learned about the reaction dynamics from these measurements.

At low laser powers, the fluorescence signal is linearly proportional to the power. However, the power available from most tunable laser systems is sufficient to cause partial saturation of the transition, with the result that the fluorescence intensity is no longer linearly proportional to the probe laser power. While more

elaborate treatments have been given [48, 49], saturation can be simply described by a rate-equation model of radiative transitions with the help of the 2-level diagram in figure B2.3.13. If the laser pulse can be approximated as a rectangular pulse of length T , then the fraction f of molecules originally in the lower state which are excited is

$$f = [W_{12}/(W_{12} + W_{21} + A_{21})][1 - \exp(-\{W_{12} + W_{21} + A_{21}\}T)]. \quad (\text{B2.3.13})$$

The fluorescence signal is linearly proportional to the fraction f of molecules excited. The absorption rate W_{12} and the stimulated emission rate W_{21} are proportional to the laser power. In the limit of low laser power, f is proportional to the laser power, while this is no longer true at high powers ($W_{12}, W_{21} \gg A_{21}$). Care must thus be taken in a laser fluorescence experiment to be sure that one is operating in the linear regime, or that proper account of saturation effects is taken, since transitions with different strengths reach saturation at different laser powers.

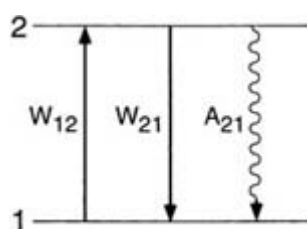


Figure B2.3.13. Model 2-level system describing molecular optical excitation, with first-order excitation rate constant W_{12} proportional to the laser power, and spontaneous (first-order rate constant A_{21}) and stimulated (first-order rate constant W_{21} proportional to the laser power) emission pathways.

Following the procedures outlined above, internal state distributions for the products of the $\text{H} + \text{NO}_2$ reaction have been determined [43, 44, 50]. Comparison of the intensities of various bands of the $\text{NO } A^2\Sigma^+ - X^2\Pi$ electronic transitions, through equation (B2.3.11), allows determination of the ratio of the populations of the vibrational levels of the NO product. From spectra such as that in figure B2.3.11 the rotational/fine-structure state distribution of the NO product in a particular vibrational level can be deduced. Figure B2.3.14 presents the vibration-rotation state distribution

derived for the NO product [43]. The vibrational state populations monotonically decrease with increasing v , up to $v = 3$, the highest detected level [50].

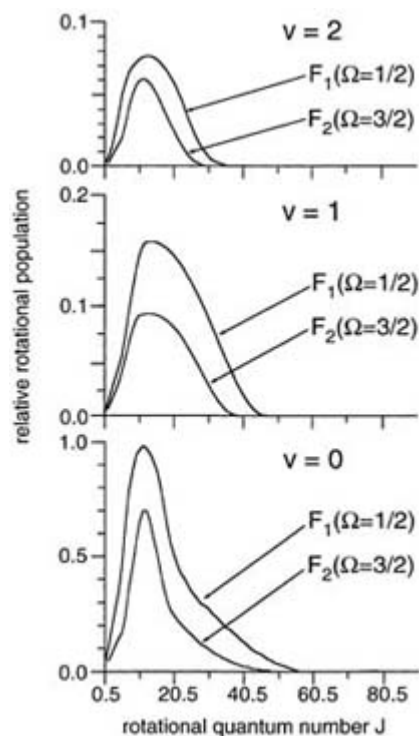


Figure B2.3.14. Experimentally derived vibration–rotation populations for the NO product from the $\text{H} + \text{NO}_2$ reaction [43]. The fine-structure labels F_1 and F_2 refer to the two ways that the projections Σ and Λ of the electron spin and orbital angular momenta along the internuclear axis of this open-shell can be coupled ($\Omega = \Lambda + \Sigma$). (By permission from AIP.)

In the work of Irvine *et al* [44], the OH product was detected, as illustrated by the fluorescence excitation spectrum in [figure B2.3.15](#). Since the rotational constant of OH is much larger than that of NO, the spectrum is much less congested. Since $\text{OH}(X^2\Pi)$ follows Hund’s case (b) coupling, the spin–orbit splitting is not directly reflected in any separations between rotational lines. The distribution in the product OH rotational/fine-structure levels was determined by the same methods as employed for the analysis of the NO spectrum. The degree of product OH vibrational excitation was found to be significantly greater than for the NO product. The $\text{H} + \text{NO}_2$ reaction proceeds on the ground state HONO PES, which has a fairly deep well corresponding to the stable nitrous acid molecule. Because of this well, the collision complex has a transient existence. However, its lifetime is not sufficiently long that the available energy is randomized through all the degrees of freedom. The ‘new’ OH bond is found to have more vibrational energy than the ‘old’ NO bond.

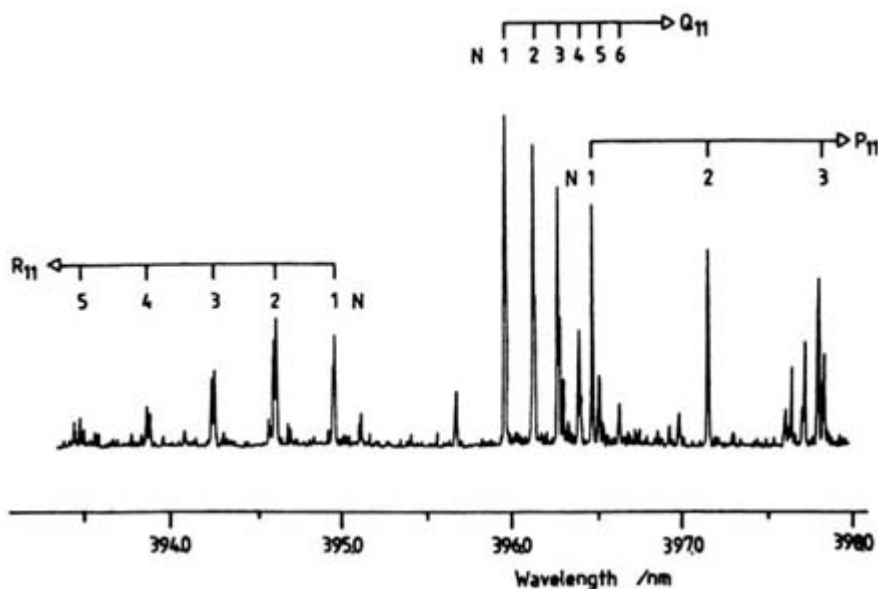


Figure B2.3.15. Laser fluorescence excitation spectrum of the $A\ ^2\Sigma^+-X\ ^2\Pi(1,3)$ band for the OH product, in the $v = 3$ vibrational level, from the $H + NO_2$ reaction [44]. (By permission from AIP.)

B2.3.3.3 PREPARATION OF REAGENTS

In most reactive scattering experiments in which the products are detected optically, a thermodynamically labile species is allowed to react with a stable molecular reagent. In many experiments, this involves allowing a beam of the unstable species, prepared in a separately pumped vacuum chamber, to impinge upon the scattering partner in a so-called beam-gas scattering arrangement. The beam is usually prepared by one of the methods described in [section B2.3.2.3](#), e.g. a high-temperature source of a beam of metal atoms or a microwave discharge source for a beam of hydrogen atoms. In some cases, two beams are crossed, and the products are detected in the collision zone [43, 51, 52]. In a few cases, the product of a reaction of two labile reagents have been studied, e.g. the OD product from the $O(^3P) + ND_2$ reaction [53]. In this study, the oxygen atoms were prepared in a microwave discharge source while the ND_2 reagent was prepared by laser photolysis of ND_3 .

Many optical studies have employed a quasi-static cell, through which the photolytic precursor of one of the reagents and the stable molecular reagent are slowly flowed. The reaction is then initiated by laser photolysis of the precursor, and the products are detected a short time after the photolysis event. To avoid collisional relaxation of the internal degrees of freedom of the product, the products must be detected in a shorter time when compared to the time between gas-kinetic collisions, that depends inversely upon the total pressure in the cell. In some cases, for example in case of the stable NO product from the $H + NO_2$ reaction discussed in [section B2.3.3.2](#), the products are not removed by collisions with the walls and may have long residence times in the apparatus. Study of such reactions are better carried out with pulsed introduction of the reagents into the cell or under crossed-beam conditions.

B2.3.3.4 EXTRACTION OF ANGULAR INFORMATION (CORRELATIONS)

With spectroscopic detection of the products, the angular distribution of the products is usually not measured. In principle, spectroscopic detection of the products can be incorporated into a crossed-beam scattering experiment of the type described in [section B2.3.2](#). There have been relatively few examples of such studies because of the great demands on detection sensitivity. The recent work of Keil and co-workers (Dharmasena *et al* [16]) on the $F + H_2$ reaction, mentioned in [section B2.3.3](#), is an excellent example of the implementation

of state-selective optical detection in the measurement of the angular distribution of a reaction product.

The use of photolytically generated reagents in a cell, combined with sub-Doppler detection, has allowed the extraction of information on the angular distribution and also the alignment of the products in experiments carried out in a cell [54]. The theoretical treatment of Shafer *et al* [55] shows how, in principle, the reaction product angular distribution can be extracted from measurement of its laboratory velocity distribution when one of the reagents is prepared by photolysis. It is well known that the angular distribution of photolytically formed fragments can be expressed as [45]

$$P(\theta_{\text{phot}}) = [1 + \beta P_2(\cos \theta_{\text{phot}})]/(4\pi) \quad (\text{B2.3.14})$$

where θ_{phot} is the angle between the \mathbf{E} vector of the photolysis laser and the fragment recoil direction, $P_2(x)$ is the second-order Legendre polynomial, and β is the recoil anisotropy parameter. The parameter β can vary from -1 (perpendicular-type transition) to 2 (parallel-type transition), where the type of transition refers to the direction of fragment recoil relative to the electronic transition moment of the dissociation transition [45].

In the bimolecular collision of the photolytically generated reagent, assumed to have a mass m_1 and laboratory speed v_1 , the centre-of-mass speed will be

$$c = m_1 v_1 / (m_1 + m_2) \quad (\text{B2.3.15})$$

if the velocity of the co-reagent (mass m_2) can be neglected. In this case, the relative velocity vector v_{rel} and c are parallel to the laboratory velocity v_1 of the photolytically generated reagent. The CM speed u'_3 of the detected product can be computed with equation (B2.3.8). Figure B2.3.16 illustrates how the laboratory velocity v'_3 of this product is related to the CM scattering angle, θ' . Shafer *et al* [55] show that the laboratory velocity distribution $f(v'_3)$ of the product is related to the CM differential cross section by

$$f(v'_3) = (2 v'_3 c u'_3)^{-1} \left(\frac{d\sigma}{d\omega} \right)_{\text{CM}} [1 + \beta P_2(\cos \alpha) P_2(\cos \theta'_{\text{phot}})] \quad \text{for } |c - u'_3| < v'_3 < (c + u'_3) \quad (\text{B2.3.16})$$

$$f(v'_3) = 0 \quad \text{for } v'_3 < |c - u'_3| \text{ or } v'_3 > (c + u'_3)$$

where

$$\cos \alpha = [v_3'^2 + c^2 - u_3'^2] / (2 v_3' c) \quad (\text{B2.3.17})$$

and θ'_{phot} is the angle between v'_3 and the \mathbf{E} vector of the photolysis laser.

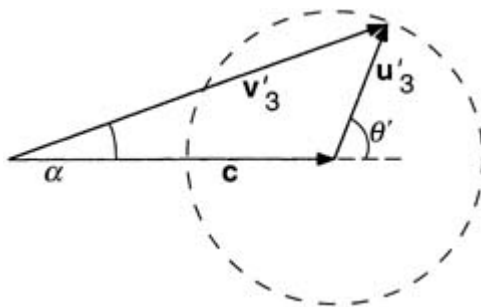


Figure B2.3.16. Velocity diagram for the reaction of a photolytically generated reagent with an assumed stationary co-reagent. In this case, the relative velocity v_{rel} of the reagents is parallel to the velocity c of the centre of mass.

There are several practical limitations to the use of [Equation \(B2.3.16\)](#) for the determination of CM angular distributions. The optimum kinematics for the use of this equation is the case where the speed c of the centre of mass is approximately equal to the product CM speed u'_3 . In the limiting case where the latter is small, the product laboratory distribution is dominated by the angular distribution of the velocity c of the centre of mass and nothing can be learned about the product CM angular distribution. In the opposite limiting case where u'_3 is much larger than c , the angular distribution of v'_3 is limited by the angular distribution of the photolytically-prepared reagent. [Equation \(B2.3.16\)](#) assumes that the velocity of the co-reagent can be neglected. This applies to the situation where the reagents are pre-cooled in a supersonic beam expansion [56]. The effects of thermal averaging have been considered to describe photo-initiated reactors in room-temperature cells [54]. Information on the anisotropy of the rotational angular momentum can also be determined through the study of photo-initiated reactions by variation of the direction of the E vector of the probe laser [54, 57].

B2.3.3.5 OTHER SPECTROSCOPIC TECHNIQUES

In addition to laser fluorescence excitation, several other laser spectroscopic methods have been found to be useful for the state-selective and sensitive detection of products of reactive collisions: resonance-enhanced multiphoton ionization [58], coherent anti-Stokes Raman scattering [59], bolometric detection with laser excitation [30], and direct infrared absorption [7]. Several additional laser techniques have been developed for use in spectroscopic studies or for diagnostics in reacting systems. Of these, four-wave mixing [60] is applicable to studies of reaction dynamics although it does have a somewhat lower sensitivity than the techniques mentioned above.

The most widely used of these techniques is resonance-enhanced multiphoton ionization (REMPI) [58]. A schematic energy-level diagram of the most commonly employed variant (2 + 1) of this detection scheme is illustrated in the

centre of [figure B2.3.8](#). The molecules are irradiated with the focused output of a tunable laser. As the wavelength of the laser is tuned through a 2-photon transition in the molecule, there will be some electronic excitation. If the photon energy is sufficiently large that the ionization continuum can be reached by absorption of an additional photon by the molecule, then molecular ions can be efficiently produced. While non-resonant laser ionization is possible, the efficiency of ionization will be strongly enhanced at a 2-photon resonance in the molecule. This particular resonant ionization scheme is called 2 + 1 REMPI. Other ionization schemes are possible, with different numbers of photons required for electronic excitation and ionization of the molecule.

A REMPI spectrum is usually recorded by monitoring the molecular ion signal as the laser wavelength is scanned, although it is also possible to record the spectrum by monitoring the photoelectron signal. In most applications of REMPI, the laser-produced ions are mass analysed in a time-of-flight mass spectrometer (TOFMS) [61] in order to detect the desired molecular ion in the presence of background ions, for example from non-resonant ionization of other species in the reaction chamber. This mass discrimination is particularly important in probing chemical reaction products since the product ion signal could be obscured by a small degree of non-resonant ionization of reagent molecules present at much higher concentrations than the product.

This technique can be used both to permit the spectroscopic detection of molecules, such as H_2 and HCl , whose first electronic transition lies in the vacuum ultraviolet spectral region, for which laser excitation is possible but inconvenient [62], or molecules such as CH_3 that do not fluoresce. With 2-photon excitation, the required wavelengths are in the ultraviolet, conveniently generated by frequency-doubled dye lasers, rather than 1-photon excitation in the vacuum ultraviolet. [Figure B2.3.17](#) displays 2 + 1 REMPI spectra of the HCl and DCl products, both in their $\nu = 0$ vibrational levels, from the $\text{Cl} + (\text{CH}_3)_3\text{CD}$ reaction [63]. For some electronic states of HCl/DCl , both parent and fragment ions are produced, and the spectrum in [figure B2.3.17](#) for the DCl product was recorded by monitoring mass 2 (D^+) ions. In this case, both isotopomers (D^{35}Cl and D^{37}Cl) are detected.

In the ideal case for REMPI, the efficiency of ion production is proportional to the line strength factors for 2-photon excitation [64], since the ionization step can be taken to have a wavelength- and state-independent efficiency. In actual practice, fragment ions can be produced upon absorption of a fourth photon, or the ionization efficiency can be reduced through predissociation of the electronically excited state. It is advisable to employ experimentally measured ionization efficiency line strength factors to calibrate the detection sensitivity. With sufficient knowledge of the excited molecular electronic states, it is possible to understand the state dependence of these intensity factors [65].

Product angular and velocity distributions can be measured with REMPI detection, similar to Doppler probing in a laser-induced fluorescence experiment discussed in [section B2.3.3.5](#). With appropriate time- and space-resolved ion detection, it is possible, in principle, to determine the three-dimensional velocity distribution of a product (see [equation \(B2.3.16\)](#)). The time-of-arrival of a particular mass in the TOFMS will be broadened by the velocity of the neutral molecule being detected. In some modes of operation of a TOFMS, e.g. space-focusing conditions [61], the shift of the arrival time from the centre of a mass peak is proportional to the projection of the molecular velocity along the TOFMS axis. In addition, Doppler tuning of the probe laser allows one component of the velocity perpendicular to the TOFMS axis to be determined. A more general approach for the two-dimensional velocity distribution in the plane perpendicular to the TOFMS direction involves the use of imaging detectors [66].

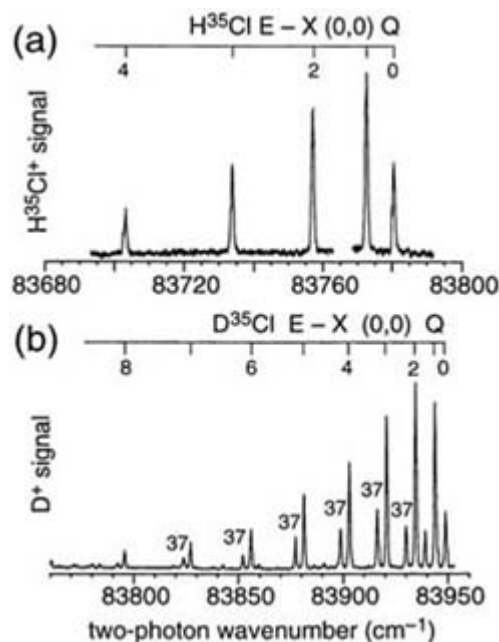
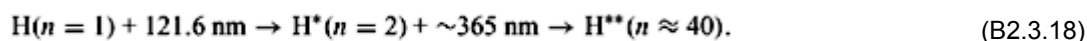


Figure B2.3.17. REMPI spectra of the HCl and DCl products from the reaction of Cl atoms with (CH₃)₃CD [63]. The mass 36 and 2 ion signals are plotted as a function of the 2-photon wavenumber. Assignments of the *Q*-branch lines ($\Delta J = 0$) of the E $1\Sigma^+ - X 1\Sigma^+ (0,0)$ bands of H³⁵Cl and D³⁵Cl are given. In (b), both the D³⁵Cl and D³⁷Cl isotopomers are observed since D⁺ ions are monitored.

Welge and co-workers (Schnieder *et al* [67]) have developed a resonant ionization technique for hydrogen atoms which allows the determination of the velocity to $\sim 0.3\%$ by a time-of-flight method. The hydrogen atoms are sequentially irradiated in the detection zone with a 121.6 nm laser, which is resonant with the $n = 2 \leftarrow n = 1$ transition, and ~ 365 nm light to produce high- n Rydberg atoms:



The Rydberg atoms are allowed to drift through a ~ 1 m flight path, after which they are field ionized with a strong electric field, and the resulting ions collected with a particle detector. The key to the high-velocity resolution of this technique is to ionize the atoms far from the laser interaction region. By contrast, when the ions are produced in this region, space-charge effects can lead to a significant velocity spread. This detection technique has been applied to the study of the H + D₂ reaction through measurement of the velocity distribution of the D atom products [68]. The laboratory velocities of D atoms formed in coincidence with specific HD vibration-rotation states were resolved, and angularly resolved differential cross sections for the formation of HD products in specific states were determined. This H atom detection technique has also been extensively employed for the study of the dynamics of the photodissociation of hydride molecules [69].

Coherent anti-Stokes Raman spectroscopy (CARS) [59] has also found utility in the determination of the internal state distributions of products of chemical reactions. This is one of several coherent Raman spectroscopies based on the

existence of vibrational resonances in four-wave mixing. As illustrated in the schematic energy level diagram in the right-hand side of figure B2.3.8 electric fields at three frequencies are mixed to produce a fourth field. In most CARS experiments, ν_1 is held fixed, usually at 532 nm, the second harmonic of a Nd:YAG laser output, while ν_2 is scanned. The intensity of the output field at ν_3 is enhanced whenever the difference $\nu_1 - \nu_2$ equals the energy difference between two molecular levels connected by a Raman transition. Unlike the

normal, spontaneous Raman process, this mixing is coherent, and the output light is a coherent beam propagating in a particular direction. The high intensity and directionality provides a great increase in detection sensitivity. In reactive scattering experiments, the CARS technique has been employed for the determination of the vibration–rotation state distributions of H₂ and HD products in reactions yielding hydrogen molecular products, e.g. the H + D₂ [70] and H + HX (X = halogen) [71] reactions.

Recently, the state-selective detection of reaction products through infrared absorption on vibrational transitions has been achieved and applied to the study of HF products from the F + H₂ reaction by Nesbitt and co-workers (Chapman *et al* [7]). The relatively low sensitivity for direct absorption has been circumvented by the use of a multi-pass absorption arrangement with a narrow-band tunable infrared laser and dual beam differential detection of the incident and transmission beams on matched detectors. A particular advantage of probing the products through absorption is that the absolute concentration of the product molecules in a given vibration–rotation state can be determined.

B2.3.4 CONCLUSION

The molecular beam and laser techniques described in this section, especially in combination with theoretical treatments using accurate PESs and a quantum mechanical description of the collisional event, have revealed considerable detail about the dynamics of chemical reactions. Several aspects of reactive scattering are currently drawing special attention. The measurement of vector correlations, for example as described in [section B2.3.3.5](#), continue to be of particular interest, especially the interplay between the product angular distribution and rotational polarization.

In most theoretical treatments of the collision dynamics, the reaction is assumed to proceed on a single PES. However, reactions involving open-shell reagents or products will involve several PESs. For example, in the F + H₂ reaction, discussed in [section B2.3.2.4](#), three PESs emanate from the separated reagents, of which only one leads to the H + HF products. The F atom ground ²P_{3/2} spin–orbit state is connected to the products by this reactive PES, while the excited ²P_{1/2} state is not. Nevertheless, the ²P_{1/2} state does react with F₂ to a small extent because of non-adiabatic transitions between the PESs. There is considerable current interest in elucidating the role of such non-adiabatic transitions in collision dynamics.

The reaction of an atom with a diatomic molecule is the prototype of a chemical reaction. As the dynamics of a number of atom–diatom reactions are being understood in detail, attention is now being turned to the study of the dynamics of reactions involving larger molecules. The reaction of Cl atoms with small aliphatic hydrocarbons is an example of the type of polyatomic reactions which are now being studied [56, 63, 72, 73]. The idea of controlling the outcome of a chemical reaction by exciting a particular bond in a reagent has long held considerable appeal. Such bond-selected chemistry has been achieved with simple triatomic reagents such as partially deuterated water, HOD, by preparation of the reagent in a suitable vibrational level [74]. Current interest is focused on the extension to larger reagents. This is more difficult than in triatomics because of intramolecular redistribution of the initial excitation [75], which becomes more rapid in larger molecules.

REFERENCES

- [1] Dawson P H 1976 *Quadrupole Mass Spectrometry and Its Applications* (Amsterdam: Elsevier)
- [2] Lee Y T, McDonald J D, LeBreton P R and Herschbach D R 1969 Molecular beam reactive scattering apparatus with electron bombardment detector *Rev. Sci. Instrum* **40** 1402–8
- [3] Auerbach D J 1988 Velocity measurements by time-of-flight methods *Atomic and Molecular Beam Methods* vol 1, ed G Scoles *et al* (New York: Oxford University Press) pp 362–79

- [4] Carrington T and Polanyi J C 1972 Chemiluminescent reactions *Chemical Kinetics, Int. Rev. Sci. Physical Chemistry* series 1, vol 9, ed J C Polanyi (London: Butterworths) pp 135–71
- [5] Leone S R 1983 Infrared fluorescence: a versatile probe of state-selected chemical dynamics *Acc. Chem. Res.* **16** 88–95
- [6] Sloan J J 1992 Fourier-transform methods: infrared *Atomic and Molecular Beam Methods* vol 2, ed G Scoles, D Lainé and U Valbusa (New York: Oxford University Press) pp 309–23
- [7] Chapman W B, Blackman B W, Nizkorodov S and Nesbitt D J 1998 Quantum-state resolved reactive scattering of $F + H_2$ in supersonic jets: Nascent $HF(v, J)$ rovibrational distributions via IR laser direct absorption methods *J. Chem. Phys.* **109** 9306–17
- [8] Lee Y T 1988 Reactive scattering I: nonoptical methods *Atomic and Molecular Beam Methods* vol 1, ed G Scoles *et al* (New York: Oxford University Press) pp 553–68
- [9] Dagdigian P J 1988 Reactive scattering II: optical methods *Atomic and Molecular Beam Methods* vol 1, ed G Scoles *et al* (New York: Oxford University Press) pp 596–629
- [10] van den Meijdenberg C J N 1988 Velocity selection by mechanical methods *Atomic and Molecular Beam Methods* vol 1, ed G Scoles *et al* (New York: Oxford University Press) pp 345–61
- [11] Miller D R 1988 Free jet sources *Atomic and Molecular Beam Methods* vol 1, ed G Scoles *et al* (New York: Oxford University Press) pp 14–53
- [12] Yang X, Lin J, Lee Y T, Blank D A, Suits A G and Wodtke A M 1997 Universal crossed molecular beams apparatus with synchrotron photoionization mass spectrometric product detection *Rev. Sci. Instrum.* **68** 3317–26
- [13] Catchen G L, Husain J and Zare R N 1978 Scattering kinematics: transformation of differential cross sections between two moving frames *J. Chem. Phys.* **69** 1737–41
- [14] Siska P E 1973 Iterative unfolding of intensity data, with application to molecular beam scattering *J. Chem. Phys.* **59** 6052–60
- [15] Valentini J J, Coggiola M J and Lee Y T 1977 Supersonic atomic and molecular halogen nozzle beam source *Rev. Sci. Instrum.* **48** 58–63
- [16] Dharmasena G, Copeland K, Young J H, Lasell R A, Phillips T R, Parker G A and Keil M 1997 Angular dependence for v', j' -resolved states in $F + H_2 \rightarrow HF(v', j') + H$ reactive scattering using a new atomic beam source *J. Phys. Chem. A* **101** 6429–40
- [17] Faubel M, Martinez-Haya B, Rusin L Y, Tappe U and Toennies J P 1996 An intense fluorine atom beam source *J. Phys. D: Appl. Phys.* **29** 1885–93
- [18] Sibener S J, Buss R J, Ng C Y and Lee Y T 1980 Development of a supersonic $O(^3P_J)$, $O(^1D_2)$ atomic oxygen nozzle beam source *Rev. Sci. Instrum.* **51** 167–82
- [19] Gorry P A and Grice R 1979 Microwave discharge source for the production of supersonic atom and free radical beams *J. Phys. E: Sci.* **12** 857–60

-29-

- [20] Casavecchia P, Balucani N and Volpi G G 1993 Reaction dynamics of $O(^3P)$, $O(^1D)$ and $OH(X^2II)$ with simple molecules *Research in Chemical Kinetics* vol 1, ed R G Compton and G Hancock (Amsterdam: Elsevier) pp 1–63
- [21] Irvin J A and Dagdigian P J 1980 Chemiluminescence from the $Ca(4s3d^1D) + O_2$ reaction: absolute cross sections, photon yield, and CaO dissociation energy *J. Chem. Phys.* **73** 176–82
- [22] Dietz T G, Duncan M A, Powers D E and Smalley R E 1981 Laser production of supersonic metal cluster beams *J. Chem. Phys.* **74** 6511–12
- [23] Kaiser R I and Suits A G 1995 A high-intensity, pulsed supersonic carbon source with $C(^3P_J)$ kinetic energies of 0.08–0.7 eV for crossed beam experiments *Rev. Sci. Instrum.* **66** 5405–11
- [24] Continetti R E, Balko B A and Lee Y T 1990 Crossed molecular beams study of the reaction $D + H_2 \rightarrow DH + H$ at collision energies of 0.53 and 1.01 eV *J. Chem. Phys.* **93** 5719–40
- [25] Neumark D M, Wodtke A M, Robinson G N, Hayden C C and Lee Y T 1985 Molecular beam studies of the $H + H_2$ reaction *J. Chem. Phys.* **82** 3045–66
- [26] Neumark D M, Wodtke A M, Robinson G N, Hayden C C, Shobatake K, Sparks R K, Schafer T P and Lee Y T 1985 Molecular beam studies of the $F + D_2$ and $F + HD$ reactions *J. Chem. Phys.* **82** 3067–77
- [27] Faubel M, Rusin L, Schlemmer S, Sondermann F, Tappe U and Toennies J P 1994 A high resolution

crossed molecular beam investigation of the absolute cross sections and product rotational states for the reaction $F + D_2(v_f = 0, j_f = 0, 1) \rightarrow DF(v_f, j_f) + D$ *J. Chem. Phys.* **101** 2106–25

- [28] Faubel M, Martinez-Haya R, Rusin L Y, Tappe U and Toennies J P 1997 Experimental absolute cross sections for the reaction $F + D_2$ at collision energies 90–240 meV *J. Phys. Chem. A* **101** 6415–28
- [29] Lee Y T 1987 Molecular beam studies of elementary chemical processes *Science* **236** 793–8
- [30] Miller R E 1995 Near-infrared laser optothermal techniques *Laser Techniques in Chemistry* vol 23, ed A B Myers and T R Rizzo (New York: Wiley) pp 43–69
- [31] Kohse-Höinghaus K 1994 Laser techniques for the quantitative detection of reactive intermediates in combustion systems *Proc. Energy Combust. Sci.* **20** 203–79
- [32] Crosley D R 1995 The measurement of OH and HO₂ in the atmosphere *J. Am. Sci.* **52** 3299–314
- [33] Okabe H 1978 *Photochemistry of Small Molecules* (New York: Wiley)
- [34] Herzberg G 1950 *Molecular Spectra and Molecular Structure I. Spectra of Diatomic Molecules* 2nd edn (Princeton: Van Nostrand)
- [35] Zare R N 1964 Calculation of intensity distribution in the vibrational structure of electronic transitions: the B ³Π_{0+u}–X ¹Σ_{0+g} resonance series of molecular iodine *J. Chem. Phys.* **40** 1934–44
- [36] Whiting E E, Schadee A, Tatum J B, Hougen J T and Nicholls R W 1980 Recommended conventions for defining transition moments and intensity factors in diatomic molecular spectra *J. Molec. Spectrosc.* **80** 249–56
- [37] Chidsey I L and Crosley D R 1980 Calculated rotational transition probabilities for the A–X system of OH *J. Quant. Spectrosc. Radiat. Transfer* **23** 187–99
- [38] Tellinghuisen J A 1974 A fast quadrature method for computing diatomic RKR potential energy curves *Comput. Phys. Commun.* **6** 221–8
- [39] Dieke G H and Crosswhite H M 1963 The ultraviolet bands of OH: fundamental data *J. Quant. Spectrosc. Radiat. Transfer* **2** 97–199
- [40] Engleman R Jr, Rouse P E, Peek H M and Biamonte V D 1970 Beta and gamma band systems of nitric oxide *Los Alamos Scientific Laboratory Report* no LA-4364

-30-

- [41] Piper L G and Cowles L M 1986 Einstein coefficients and transition moment variation for the NO(A ²Σ⁺–X ²Π) transition *J. Chem. Phys.* **85** 2419–22
- [42] Yarkony D R 1992 A theoretical treatment of the predissociation of the individual rovibronic levels of OH/OD (A ²Σ⁺) *J. Chem. Phys.* **97** 1838–49
- [43] Sauder D G and Dagdigian P J 1990 Determination of the internal state distribution of NO produced from the H + NO₂ reaction *J. Chem. Phys.* **92** 2389–96
- [44] Irvine A M L, Smith I W M, Tuckett R P and Yang X-F 1990 A laser-induced fluorescence determination of the complete internal state distribution of OH produced in the reaction: $H + NO_2 \rightarrow OH + NO$ *J. Chem. Phys.* **93** 3177–86
- [45] Zare R N 1988 *Angular Momentum* (New York: Wiley)
- [46] Greene C H and Zare R N 1983 Determination of product population and alignment using laser-induced fluorescence *J. Chem. Phys.* **78** 6741–53
- [47] Orr-Ewing A J and Zare R N 1995 Orientation and alignment of the products of bimolecular reactions *The Chemical Dynamics and Kinetics of Small Radicals* vol 2, ed K Liu and A Wagner (Singapore: World Scientific) pp 936–1063
- [48] Altkorn R and Zare R N 1984 Effects of saturation on laser-induced fluorescence measurements of population and polarization *Ann. Rev. Phys. Chem.* **35** 265–89
- [49] Hefter U and Bergmann K 1988 Spectroscopic detection methods *Atomic and Molecular Beam Methods* vol 1, ed G Scoles *et al* (New York: Oxford University Press) pp 193–253
- [50] Irvine A M L, Smith I W M and Tuckett R P 1990 A laser-induced fluorescence determination of the internal state distribution of NO produced in the reaction $H + NO_2 \rightarrow OH + NO$ *J. Chem. Phys.* **93** 3187–95
- [51] Liu K, Macdonald R G and Wagner A F 1990 Crossed-beam investigations of state-resolved collision dynamics of simple radicals *Int. Rev. Phys. Chem.* **9** 187–225

- [52] Scott D C, Winterbottom F, Scholfield M R, Goyal S and Reisler H 1994 Kinetic energy effects on product state distributions in the $C(^3P) + N_2O (X^1 \Sigma^+)$ reaction: energy partitioning between the $NO(X^2 \Pi)$ and $CN(X^2 \Sigma^+)$ products *Chem. Phys. Lett.* **222** 471–80
- [53] Patel-Misra D, Sauder D G and Dagdigian P J 1991 Internal state distribution of OD produced from the $O(^3P) + ND_2$ reaction *J. Chem. Phys.* **95** 955–62
- [54] Alexander A J, Brouard M, Kalogerakis K S and Simons J P 1998 Chemistry with a sense of direction—the stereodynamics of bimolecular reactions *Chem. Soc. Rev.* **27** 405–15
- [55] Shafer N E, Orr-Ewing A J, Simpson W R, Xu H and Zare R N 1993 State-to-state differential cross sections from photoinitiated bulk reactions *Chem. Phys. Lett.* **212** 155–162
- [56] Simpson W R, Orr-Ewing A J and Zare R N 1993 State-to-state differential cross sections for the reaction $Cl(^2P_{3/2}) + CH_4(v_3 = 1, J = 1) \rightarrow HCl(v' = 1, J') + CH_3$ *Chem. Phys. Lett.* **212** 163–71
- [57] Rakitzis T P, Kandel S A and Zare R N 1997 Determination of differential-cross-section moments from polarization-dependent product velocity distributions of photoinitiated bimolecular reactions *J. Chem. Phys.* **107** 9382–91
- [58] Ashfold M N R and Howe J D 1994 Multiphoton spectroscopy of molecular species 1994 *Ann. Rev. Phys. Chem.* **45** 57–82
- [59] Valentini J J 1985 Coherent anti-Stokes spectroscopy *Spectrometric Techniques* vol 4, ed G A Vanasse (New York: Academic) pp 1–62
- [60] Vaccaro P H 1995 Resonant four-wave mixing spectroscopy: a new probe for vibrationally-excited species *Molecular Dynamics and Spectroscopy by Stimulated Emission Pumping (Advances in Chemistry Series)* vol 7, ed H-L Dai and R W Field (Singapore: World Scientific) p 1

- [61] Wiley W C and McLaren I H 1955 Time-of-flight mass spectrometer with improved resolution *Rev. Sci. Instrum* **26** 1150–7
- [62] Hepburn J W 1995 Generation of coherent vacuum ultraviolet radiation: applications to high-resolution photoionization and photoelectron spectroscopy *Laser Techniques in Chemistry* vol 23, ed A B Myers and T R Rizzo (New York: Wiley) pp 149–83
- [63] Varley D F and Dagdigian P J 1996 Product state resolved study of the $Cl + (CH_3)_3 CD$ reaction: comparison of the dynamics of abstraction of primary vs tertiary hydrogens *J. Phys. Chem.* **100** 4365–74
- [64] Kummel A C, Sitz G O and Zare R N 1986 Determination of population and alignment of the ground state using two-photon nonresonant excitation *J. Chem. Phys.* **85** 6874–97
- [65] Dagdigian P J, Varley D F, Liyanage R, Gordon R J and Field R W 1996 Detection of DCl by multiphoton ionization and determination of DCl and HCl internal state distributions *J. Chem. Phys.* **106** 10 251–62
- [66] Heck A J R and Chandler D W 1995 Imaging techniques for the study of chemical reaction dynamics *Ann. Rev. Phys. Chem.* **46** 335–72
- [67] Schnieder L, Maier W, Welge K H, Ashfold M N R and Western C M 1990 Photodissociation dynamics of H_2S at 121.6 nm and a determination of the potential energy function of $SH(A^2 \sigma^+)$ *J. Chem. Phys.* **92** 7027–37
- [68] Schnieder L, Seekamp-Rahn K, Wiede E and Welge K H 1997 Experimental determination of quantum state resolved differential cross sections for the hydrogen exchange reaction $H + D_2 \rightarrow HD + D$ *J. Chem. Phys.* **107** 6175–95
- [69] Ashfold M N R, Mordaunt D H and Wilson S H S 1996 Photodissociation dynamics of hydride molecules: H atom photofragment translational spectroscopy *Adv. Photochem.* **21** 217–95
- [70] Gerrity D P and Valentini J J 1984 Experimental study of the dynamics of the $H + D_2 \rightarrow HD + D$ reaction at collision energies of 0.55 and 1.30 eV *J. Chem. Phys.* **81** 1298–313
- [71] Aker P M, Germann G J and Valentini J J 1989 State-to-state dynamics of $H + HX$ collisions. I. The $H + HX \rightarrow H_2 + X$ ($X = Cl, Br, I$) abstraction reactions at 1.6 eV collision energy *J. Chem. Phys.* **90** 4795–808
- [72] Hemmi N and Suits A G 1998 The dynamics of hydrogen abstraction reactions: crossed-beam reaction $Cl + n-C_5H_{12} \rightarrow C_5H_{11} + HCl$ *J. Chem. Phys.* **109** 5338–43
- [73] Kandel S A and Zare R N 1998 Reaction dynamics of atomic chlorine with methane: importance of methane bending and torsional excitation in controlling reactivity *J. Chem. Phys.* **109** 9719–27

[74] Crim F F 1996 Bond-selected chemistry: vibrational state control of photodissociation and bimolecular reaction *J. Phys. Chem.* **100** 12 725–34

[75] Nesbitt D J and Field R W 1998 Vibrational energy flow in highly excited molecules: role of intramolecular vibrational redistribution *J. Chem. Phys.* **100** 12 735–56

FURTHER READING

Scoles G, Bassi D, Buck U and Lainé D (eds) 1988 *Atomic and Molecular Beam Methods* vol 1 (New York: Oxford University Press)

This book presents an extensive and detailed description of basic techniques for the generation and detection of atomic and molecular beams, as well as beam techniques for the study of molecular scattering processes.

-32-

Zare R N 1988 *Angular Momentum* (New York: Wiley)

This book presents a detailed exposition of angular momentum theory in quantum mechanics, with numerous applications and problems in chemical physics. Of particular relevance to the present section is an elegant and clear discussion of molecular wavefunctions and the determination of populations and moments of the rotational state distributions from polarized laser fluorescence excitation experiments.

Herzberg G 1989 *Molecular Spectra and Molecular Structure. I. Spectra of Diatomic Molecules* reprint (Malabar, FL: Krieger)

This book, originally published in 1950, is the first of a classic three-volume set on molecular spectroscopy. A rather complete discussion of diatomic electronic spectroscopy is presented. Volumes II (1945) and III (1967) discuss infrared and Raman spectroscopy and polyatomic electronic spectroscopy, respectively.

-1-

B2.4 NMR methods for studying exchanging systems

Alex D Bain

B2.4.1 INTRODUCTION

No molecule is completely rigid and fixed. Molecules vibrate, parts of a molecule may rotate internally, weak bonds break and re-form. Nuclear magnetic resonance spectroscopy (NMR) is particularly well suited to observe an important class of these motions and rearrangements. An example is the restricted rotation about bonds, which can cause dramatic effects in the NMR spectrum (figure B2.4.1).

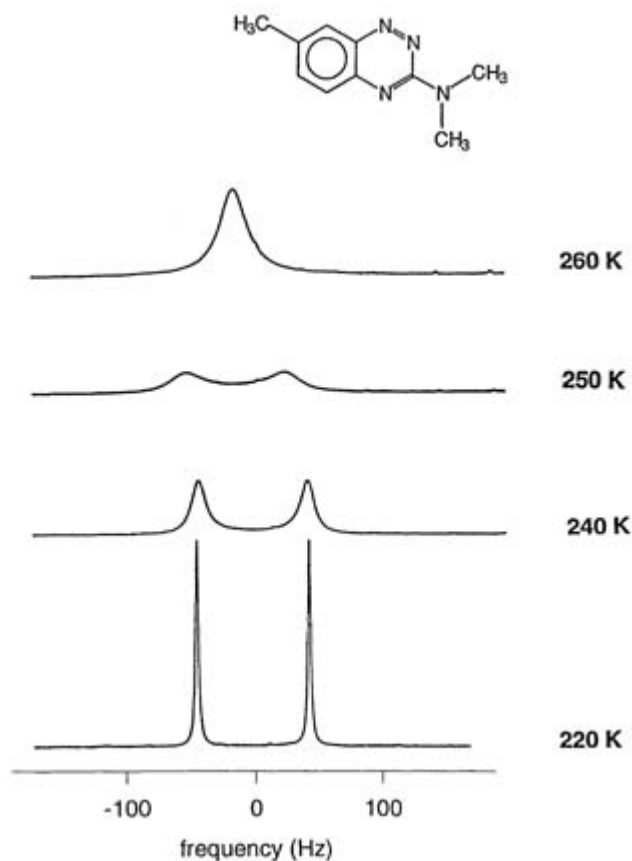


Figure B2.4.1. Proton NMR spectra of the *N,N'*-dimethyl groups in 3-dimethylamino-7-methyl-1,2,4-benzotriazine, as a function of temperature. Because of partial double-bond character, there is restricted rotation about the bond between the dimethylamino group and the ring. As the temperature is raised, the rate of rotation around the bond increases and the NMR signals of the two methyl groups broaden and coalesce.

-2-

These exchanges often occur while the system is in macroscopic equilibrium—the sample itself remains the same and the dynamics may be invisible to other techniques. It is merely the environment of a given nucleus that changes. Since NMR follows an individual nucleus, it can easily follow these dynamic processes. This is just one of several reasons that the study of chemical exchange by NMR is important.

First is the observation of the phenomenon itself—the exchange of axial and equatorial ligands in trigonal bipyramidal species, the scrambling of carbonyl ligands in metal complexes and dynamic behaviour of rings was mainly revealed and studied in detail by NMR methods. Not only does NMR give a detailed picture of the mechanism of the exchange, but it also provides excellent ways of measuring the reaction rate.

Secondly, NMR is a good example of spectroscopy in general. The spectroscopic transition probability can be shown to have a simple physical interpretation in NMR: the total magnetization is divided amongst individual observable transitions and the intensity of a transition is related to its share of the total. This can be further generalized to exchanging systems, in which the transition probability now becomes a complex number. The exchange lineshapes can be decomposed into a sum of transitions, whose phase, intensity, position and linewidths are governed by the real and imaginary parts of the transition probability.

Finally, exchange is a kinetic process and governed by absolute rate theory. Therefore, study of the rate as a function of temperature can provide thermodynamic data on the transition state, according to equation (B2.4.1)). This equation, in which k is Boltzmann's constant and h is Planck's constant, relates the observed rate to the Gibbs free energy of activation, ΔG^\ddagger .

$$\text{Rate} = \frac{kT}{h} e^{-\Delta G^\ddagger/RT} = \frac{kT}{h} e^{-\Delta H^\ddagger/RT} e^{\Delta S^\ddagger/R}. \quad (\text{B2.4.1})$$

In order to separate the enthalpy and the entropy of activation, the rate is measured as a function of temperature. These data should give a straight line on an Eyring plot of $\log(\text{rate}/T)$ against $(1/T)$ (figure B2.4.2). The slope of the line gives ΔH^\ddagger , and the intercept at $1/T = 0$ is related to ΔS^\ddagger . A unimolecular reaction, such as many cases of exchange, might be expected to have a very small entropy change on going to the transition state. However, several systems have shown significant entropy contributions—entropy can make up more than 10% of the barrier. It is therefore important to measure the rates over as wide a range of temperatures as possible to obtain reliable thermodynamic data on the transition state.

There are several ways of measuring exchange rates with NMR, each with its own optimum range. Figure B2.4.2 illustrates this. A combination of spin–spin relaxation time (T_2) measurements at high temperature, bandshape analysis in the intermediate regime and selective inversions at low temperature, gives rates over a range of about five orders of magnitude. This provides a very well defined Eyring plot. The first fully analysed example of chemical exchange in NMR was the proton spectrum of dimethylamides [1], observed at about the same time (and in the same laboratory) as the phenomenon of scalar coupling between nuclei. Figure B2.4.1 illustrates this type of behaviour. If there is no rotation about the bond joining the N, N'-dimethyl group to the ring, the proton NMR signals of the two methyl groups will have different chemical shifts. If the rotation were very fast, then the two methyl environments would be exchanged very quickly and only a single, average, methyl peak would appear in the proton NMR spectrum. Between these two extremes, spectra like those in figure B2.4.1 are observed. At low temperature, when the rate is slow, two

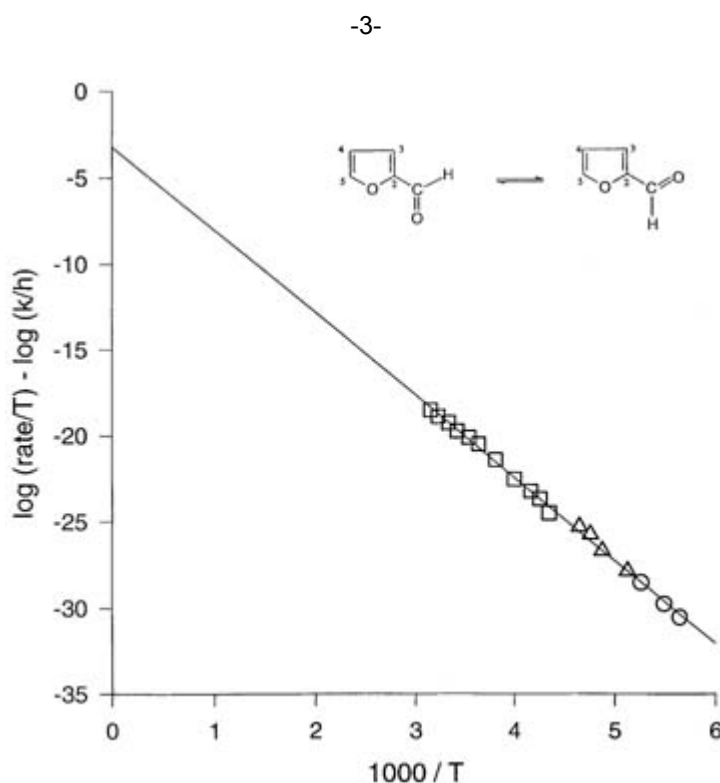


Figure B2.4.2. Eyring plot of $\log(\text{rate}/T)$ versus $(1/T)$, where T is absolute temperature, for the *cis–trans* isomerism of the aldehyde group in furfural. Rates were obtained from three different experiments: T_2 measurements (squares), bandshapes (triangles) and selective inversions (circles). The line is a linear regression to the data. The slope of the line is $\Delta H^\ddagger/R$, and the intercept at $1/T = 0$ is $\Delta S^\ddagger/R$, where R is the gas constant. ΔH^\ddagger and ΔS^\ddagger are the enthalpy and entropy of activation, according to equation (B2.4.1)

sharp lines are seen. As the temperature is increased, the exchange becomes faster, the lines broaden, coalesce into a single line and finally become a single sharp line at the average chemical shift. This is perhaps the most familiar manifestation of chemical exchange in NMR.

The different types of chemical exchange in NMR are classified according to the rate relative to some NMR timescale. The example in [figure B2.4.1](#) is called intermediate exchange, in which the exchange rate is comparable to the chemical shift differences and coupling constants. Intermediate exchange gives an array of unusual and characteristic lineshapes in the spectrum, which can be explained quite neatly in terms of a generalized transition probability. Fast exchange is the regime well after coalescence, when only a single line is observed. There is still observable broadening due to exchange, and rates are measured from the linewidth, or equivalently, the spin–spin relaxation time, T_2 . In slow exchange, no dramatic line broadening is observed, but the exchange rate is comparable to the reciprocal of the spin–lattice relaxation time, T_1 . In this regime, modifications of the inversion-recovery experiment, or techniques related to the nuclear Overhauser effect (NOE), are used to measure rates.

The timescale is just one sub-classification of chemical exchange. It can be further divided into coupled *versus* uncoupled systems, mutual or non-mutual exchange, inter- or intra-molecular processes and solids *versus* liquids. However, all of these can be treated in a consistent and clear fashion.

-4-

The NMR experimental methods for studying chemical exchange are all fairly routine experiments, used in many other NMR contexts. To interpret these results, a numerical model of the exchange, as a function of rate, is fitted to the experimental data. It is therefore necessary to look at the theory behind the effects of chemical exchange. Much of the theory is developed for intermediate exchange, and this is the most complex case. However, with this theory, all of the rest of chemical exchange can be understood.

B2.4.2 INTERMEDIATE EXCHANGE

B2.4.2.1 INTRODUCTION

[Figure B2.4.1](#) shows the lineshape for intermediate chemical exchange between two equally populated sites without scalar coupling. For more complicated spin systems, the lineshapes are more complicated as well, since a spin may retain its coupling information even though its chemical shift changes in the exchange.

[Figure B2.4.3](#) shows an example of this in the aldehyde proton spectrum of ^{15}N -labelled formamide. Some lines in the spectrum remain sharp, while others broaden and coalesce. There is no fundamental difference between the lineshapes in [figures B2.4.1](#) and [figures B2.4.3](#)—only a difference in the size of the matrices involved. First, the uncoupled case will be discussed, then the extension to coupled spin systems.

-5-

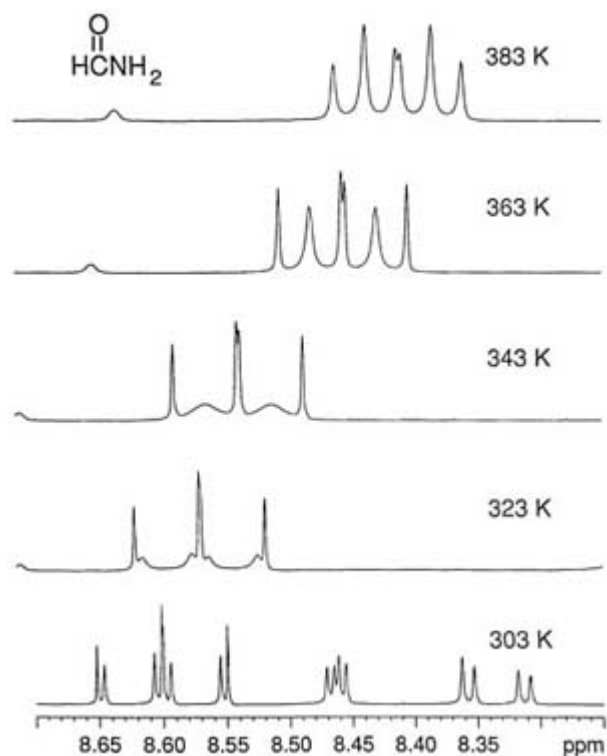
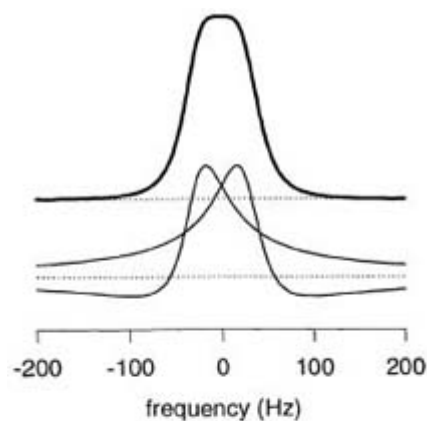
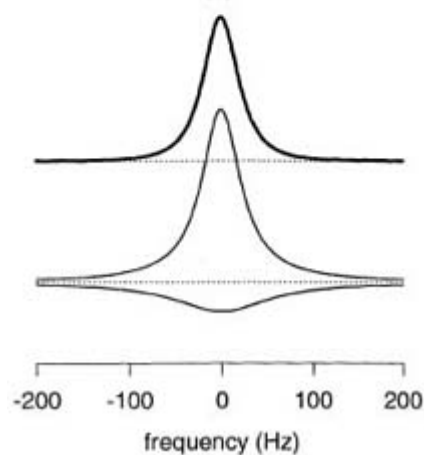


Figure B2.4.3. Proton NMR spectrum of the aldehyde proton in ^{15}N -labelled formamide. This proton has couplings of 1.76 Hz and 13.55 Hz to the two amino protons, and a coupling of 15.0 Hz to the ^{15}N nucleus. The outer lines in the spectrum remain sharp, since they represent the sum of the couplings, which is unaffected by the exchange. The inner lines of the multiplet broaden and coalesce, as in [figure B2.4.1](#). The other peaks in the 303 K spectrum are due to the NH_2 protons, whose chemical shifts are even more temperature dependent than that of the aldehyde proton.

The original analysis of the spectra in [figure B2.4.1](#) was done by the groups of Gutowsky [1] and McConnell [2], both of whom treated the spectrum as a whole in the frequency domain. Reeves showed [3] somewhat later that the two-site exchange lineshape ([figure B2.4.4\(a\)](#) and [figure B2.4.4\(b\)](#)) can be deconstructed into two transitions. More recently, it was demonstrated [4] that these transitions are defined by a generalized transition probability. This transition probability (now a complex number) which is just the product of how much coherence a transition receives at the start of an experiment, times how much the transition contributes to the total signal. This leads naturally into a discussion of the xy magnetizations in the time domain. As with most NMR, the choice of the time domain or the frequency domain depends on the problem.



(a)



(b)

Figure B2.4.4. The two-site equally populated exchange lineshape (figure B2.4.1) decomposed into two individual transitions. The bottom spectrum (a) is the situation before coalescence: two symmetrically out-of-phase lines. In slow exchange, these become the signals of the two sites. The top spectrum (b) is after coalescence: the lineshape is made up of two central lines, one positive and one negative. In fast exchange, the negative line broadens and loses intensity, to leave a single positive line at the average chemical shift.

B2.4.2.2 THE BLOCH EQUATIONS APPROACH

The Bloch equations for the motion of the x and y magnetizations (usually called the u - and v -mode signals), in the presence of a weak radiofrequency (RF) field, B_1 , are given in equation (B2.4.2)).

$$\frac{du}{dt} + \frac{u}{T_2} - (\omega_0 - \omega)v = 0$$

$$\frac{dv}{dt} + \frac{v}{T_2} + (\omega_0 - \omega)u = \gamma B_1 M_z.$$
(B2.4.2)

In this equation, ω is the frequency of the RF irradiation, ω_0 is the Larmor frequency of the spin, T_2 is the spin–spin relaxation time and M_z is the z magnetization of the spin system. The notation can be simplified somewhat by defining a complex magnetization, M , as in equation (B2.4.3).

$$M = u + iv. \quad (\text{B2.4.3})$$

With this definition, the Bloch equations can be written as in equation (B2.4.4).

$$\frac{dM}{dt} + i(\omega_0 - \omega)M + \frac{1}{T_2}M = i\gamma B_1 M_z. \quad (\text{B2.4.4})$$

In chemical exchange, the two exchanging sites, A and B, will have different Larmor frequencies, ω_A and ω_B . Assuming equal populations in the two sites, and the rate of exchange to be k , the two coupled Bloch equations for the two sites are given by equation (B2.4.5).

$$\begin{aligned} \frac{dM_A}{dt} + i(\omega_A - \omega)M_A + \frac{1}{T_2}M_A - kM_B + kM_A &= i\gamma B_1 M_{zA} \\ \frac{dM_B}{dt} + i(\omega_B - \omega)M_B + \frac{1}{T_2}M_B - kM_A + kM_B &= i\gamma B_1 M_{zB}. \end{aligned} \quad (\text{B2.4.5})$$

The observable NMR signal is the imaginary part of the sum of the two steady-state magnetizations, M_A and M_B . The steady state implies that the time derivatives are zero and a little further calculation (and neglect of T_2 terms) gives the NMR spectrum of an exchanging system as equation (B2.4.6).

$$v = \frac{1}{2}\gamma B_1 M_z \frac{k(\omega_A - \omega_B)^2}{(\omega_A - \omega)^2(\omega_B - \omega)^2 + 4k^2(\omega - (\omega_A + \omega_B)/2)^2}. \quad (\text{B2.4.6})$$

B2.4.2.3 MATRIX FORMULATION OF CHEMICAL EXCHANGE

Equation (B2.4.5) can be re-written in a matrix form [5] as equation (B2.4.7).

$$\frac{d}{dt} \begin{pmatrix} M_A \\ M_B \end{pmatrix} + i\mathbf{L} \begin{pmatrix} M_A \\ M_B \end{pmatrix} + \mathbf{R} \begin{pmatrix} M_A \\ M_B \end{pmatrix} + \mathbf{K} \begin{pmatrix} M_A \\ M_B \end{pmatrix} = i\gamma B_1 \begin{pmatrix} M_{zA} \\ M_{zB} \end{pmatrix}. \quad (\text{B2.4.7})$$

In this equation, the matrices \mathbf{L} , \mathbf{R} and \mathbf{K} are given by equation (B2.4.8), equation (B2.4.9) and equation (B2.4.10).

$$\mathbf{L} = \begin{pmatrix} \omega_A - \omega & 0 \\ 0 & \omega_B - \omega \end{pmatrix} \quad (\text{B2.4.8})$$

$$\mathbf{R} = \begin{pmatrix} \frac{1}{T_2} & 0 \\ 0 & \frac{1}{T_2} \end{pmatrix} \quad (\text{B2.4.9})$$

$$\mathbf{K} = \begin{pmatrix} k & -k \\ -k & k \end{pmatrix}. \quad (\text{B2.4.10})$$

The steady-state solution without saturation to this equation is obtained by setting the time derivatives to zero and taking the terms linear in B_1 , as in equation (B2.4.11).

$$\begin{pmatrix} M_A \\ M_B \end{pmatrix} = (i\mathbf{L} + \mathbf{R} + \mathbf{K})^{-1} \begin{pmatrix} M_{zA} \\ M_{zB} \end{pmatrix}. \quad (\text{B2.4.11})$$

Recall that \mathbf{L} contains the frequency ω (equation (B2.4.8)). To trace out a spectrum, equation (B2.4.11) is solved for each frequency. In order to obtain the observed signal v , the sum of the two individual magnetizations can be written as the dot product of two vectors, equation (B2.4.12).

$$v = (1 \ 1) \begin{pmatrix} M_A \\ M_B \end{pmatrix}. \quad (\text{B2.4.12})$$

This apparently artificial way of re-writing the Bloch equations is important, since this form applies to all exchanging systems—coupled or uncoupled—in the frequency domain. The description starts with the equilibrium z magnetizations. These are affected by all the NMR interactions: chemical shifts, relaxation and exchange. Finally, the observed signal is detected. This is the standard preparation–evolution–detection paradigm used in multi-dimensional NMR. There may be algebraic and numerical complications in setting up and solving the equations for different systems, but the form remains the same for all frequency-domain calculations.

B2.4.2.4 CHEMICAL EXCHANGE IN THE TIME DOMAIN

If the magnetizations, M_A and M_B , are created (by a pulse) at time zero, and then the B_1 magnetic field is turned off, equation (B2.4.7) can be simplified to equation (B2.4.13). Note that ω in the matrix \mathbf{L} (equation (B2.4.8)) is also zero.

-9-

$$\frac{d}{dt} \begin{pmatrix} M_A \\ M_B \end{pmatrix} = -(i\mathbf{L} + \mathbf{R} + \mathbf{K}) \begin{pmatrix} M_A \\ M_B \end{pmatrix}. \quad (\text{B2.4.13})$$

Equation (B2.4.13) is a pair of first-order differential equations, so its formal solution is given by equation (B2.4.14), in which $\exp()$ means the exponential of a matrix.

$$\begin{pmatrix} M_A(t) \\ M_B(t) \end{pmatrix} = \exp(-[i\mathbf{L} + \mathbf{R} + \mathbf{K}]t) \begin{pmatrix} M_A(0) \\ M_B(0) \end{pmatrix}. \quad (\text{B2.4.14})$$

This is the description of NMR chemical exchange in the time domain. Note that this equation and equation (B2.4.11) are Fourier transforms of each other. The time-domain and frequency-domain pictures are always related in this way.

In practice, the matrix $(i\mathbf{L} + \mathbf{R} + \mathbf{K})$ is diagonalized first, with a matrix of eigenvectors, \mathbf{U} , as in equation (B2.4.15), to give a diagonal matrix, $\mathbf{\Lambda}$, with the eigenvalues, λ^i , of \mathbf{L} down the diagonal.

$$\mathbf{\Lambda} = \mathbf{U}^{-1} (i\mathbf{L} + \mathbf{R} + \mathbf{K}) \mathbf{U}. \quad (\text{B2.4.15})$$

equation (B2.4.14) becomes equation (B2.4.16).

$$\begin{pmatrix} M_A(t) \\ M_B(t) \end{pmatrix} = \mathbf{U} \exp(-\mathbf{\Lambda}t) \mathbf{U}^{-1} \begin{pmatrix} M_A(0) \\ M_B(0) \end{pmatrix}. \quad (\text{B2.4.16})$$

The exponential of a diagonal matrix is again a diagonal matrix with exponentials of the diagonal elements, equation (B2.4.17)).

$$\begin{pmatrix} M_A(t) \\ M_B(t) \end{pmatrix} = \mathbf{U} \begin{pmatrix} e^{-\lambda_1 t} & 0 \\ 0 & e^{-\lambda_2 t} \end{pmatrix} \mathbf{U}^{-1} \begin{pmatrix} M_A(0) \\ M_B(0) \end{pmatrix}. \quad (\text{B2.4.17})$$

As was mentioned above, the observed signal is the imaginary part of the sum of M_A and M_B , so equation (B2.4.17)) predicts that the observed signal will be the sum of two exponentials, evolving at the complex frequencies λ_1 and λ_2 . This is the free induction decay (FID). In the limit of no exchange, the two frequencies are simply $i\omega_A$ and $i\omega_B$, as expected. When k is non-zero, the situation is more complex.

Without relaxation and exchange, \mathbf{L} is a Hermitian matrix with real eigenvalues and eigenvectors. However, when the exchange contributes significantly, the Hermitian character is lost and the eigenvalues and eigenvectors have both real and imaginary parts. The eigenvalues are given by the roots of the characteristic equation, (B2.4.18), in which δ is $(\omega_A - \omega_B)/2$.

-10-

$$\begin{vmatrix} i\delta + \frac{1}{T_2} + k - \lambda & -k \\ -k & -i\delta + \frac{1}{T_2} + k - \lambda \end{vmatrix} = 0. \quad (\text{B2.4.18})$$

The eigenvalues of equation (B2.4.16)) are given in equation (B2.4.19)).

$$\lambda = \left(\frac{1}{T_2} + k \right) \pm \sqrt{k^2 - \delta^2}. \quad (\text{B2.4.19})$$

These eigenvalues are the (complex) frequencies of the lines in the spectrum: the imaginary part gives the oscillation frequency and the real part gives the rate of decay. If $k < \delta$ (slow exchange) then there are two different imaginary frequencies, which become $\pm\delta$ in the limit of small k . Figure B2.4.4 a) shows this decomposition. In fast exchange, when k exceeds the shift difference, δ , the quantity in the square root in equation (B2.4.19) becomes positive, so the roots are pure real. This means that the spectrum is still two lines, but they are both at the average chemical shift (offset of zero) and have different widths (figure B2.4.4(b)).

It is convenient, for simple systems, to have explicit expressions for equation (B2.4.17). Since the original matrix is non-Hermitian, the matrix formed by the eigenvectors will not be unitary, and will have four independent complex elements. Let them be a , b , c and d , so that \mathbf{U} is given by equation (B2.4.20).

$$\mathbf{U} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}. \quad (\text{B2.4.20})$$

Regardless of whether \mathbf{U} is unitary, its inverse is given by equation (B2.4.21), where Δ is the determinant of the matrix.

$$\mathbf{U}^{-1} = \frac{1}{\Delta} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}. \quad (\text{B2.4.21})$$

Equation (B2.4.16) then says that the signal is given by equation (B2.4.22), regardless of slow or fast exchange.

$$\text{Signal} = \frac{(a+c)(d-b)}{\Delta} e^{\lambda_1 t} + \frac{(b+d)(-c+a)}{\Delta} e^{\lambda_2 t}. \quad (\text{B2.4.22})$$

For slow exchange, a convenient matrix of eigenvectors is given by equation (B2.4.23).

$$\begin{pmatrix} k & i(\sqrt{\delta^2 - k^2} + \delta) \\ -i(\sqrt{\delta^2 - k^2} + \delta) & k \end{pmatrix}. \quad (\text{B2.4.23})$$

-11-

After coalescence, a possible set of eigenvectors is given in equation (B2.4.24). If these are substituted into (B2.4.22), the results are pure real, reflecting the fact that $k^2 - \delta^2$ is now positive.

$$\begin{pmatrix} \sqrt{k^2 - \delta^2} - i\delta & -\sqrt{k^2 - \delta^2} - i\delta \\ k & k \end{pmatrix}. \quad (\text{B2.4.24})$$

Because of the role of the eigenvectors in [equation \(B2.4.16\)](#), the factor (amplitude) multiplying the complex exponential is itself complex. The magnitude of the complex amplitude gives the intensity of the line and its phase gives the phase of the line (the mixture of absorption and dispersion). In slow exchange, the two lines have the same real part, but the imaginary parts have opposite signs, so the phase distortion is opposite, as in [\(figure B2.4.4\(a\)\)](#). The sum of these distorted lineshapes gives the familiar coalescence spectrum. In fast exchange, the two lines are both in phase, but one line is negative [\(figure B2.4.4\(b\)\)](#). This negative line is very broad, and decreases in absolute intensity as the rate increases, leaving only the single, positive, in-phase line for fast exchange.

B2.4.2.5 CHEMICAL EXCHANGE IN COUPLED SPIN SYSTEMS

The development given in the previous section is simply a special case of the general density matrix treatment of chemical exchange. In an uncoupled system, the whole of the coherence from one site is transferred to the other, since the signal is directly associated with a given nucleus. This simplifies the calculation. In a coupled spin system, particularly a strongly coupled system, this is no longer true. The relation between the lines in the spectrum and individual nuclei can be much more complicated. Furthermore, the amount of ‘mixing’ of nuclei in a given spectrum depends on the chemical shifts and couplings. Therefore, when a nucleus exchanges in a coupled system, coherence that was associated with a single line in one site may be distributed amongst several lines in the other site. In dealing with chemical exchange in coupled systems, it is necessary to keep track of the details of each of the lines in the spectrum, but the fundamental approach is the same.

There is an important special case, called mutual exchange. In all exchange phenomena, a specific nucleus experiences a different magnetic environment when it moves from one site to the other. However, in many cases, the new arrangement is simply a permutation of the old, as in the case of formamide in [\(figure B2.4.3\)](#). The two amide protons have switched places, but the chemical shifts and couplings are the same. All that has changed is the nuclei associated with each of them. This can be treated in the same way as all other exchanges, as two different sites. However, the permutation symmetry of the problem means that this is equivalent to copies of a single, mutual, exchange. The matrices are then reduced in size. It is possible to have quite complex permutations, so the analysis must be done carefully and systematically. It is therefore

important to identify mutual exchange and treat it appropriately.

Binsch [6] provided the standard way of calculating these lineshapes in the frequency domain, and implemented it in the program DNMR3 [7]. Formally, it is the same as the matrix description given in [section \(B2.4.2.3\)](#). The calculation of the matrices **L**, **R** and **K** is more complex for a coupled spin system, but that should not interfere with the understanding of how the method works. This work will be discussed later, but first the time-domain approach will be developed.

The basic equation [8] is the equation of motion for the density matrix, ρ , given in [equation \(B2.4.25\)](#), in which H is the Hamiltonian.

-12-

$$i \frac{\hbar}{2\pi} \frac{\partial}{\partial t} \rho = [H, \rho]. \quad (\text{B2.4.25})$$

It is more convenient to re-express this equation in Liouville space [8, 9 and 10], in which the density matrix becomes a vector, and the commutator with the Hamiltonian becomes the Liouville superoperator. In this formulation, the lines in the spectrum are some of the elements of the density ‘matrix’ vector, and what happens to them is described by the superoperator matrix. [equation \(B2.4.25\)](#) becomes [\(B2.4.26\)](#).

$$i \frac{\hbar}{2\pi} \frac{\partial}{\partial t} \rho = \mathbf{L} \rho. \quad (\text{B2.4.26})$$

This Liouville-space equation of motion is exactly the time-domain Bloch equations approach used in [equation \(B2.4.13\)](#). The magnetizations are arrayed in a vector, and anything that happens to them is represented by a matrix. In frequency units ($\hbar/2\pi = 1$), the formal solution to [equation \(B2.4.26\)](#) is given by [equation \(B2.4.27\)](#) (compare [equation \(B2.4.14\)](#)).

$$\rho(t) = \exp(-i\mathbf{L}t) \rho(0). \quad (\text{B2.4.27})$$

For a coupled spin system, the matrix of the Liouvillian must be calculated in the basis set for the spin system. Usually this is a simple product basis, often called product operators, since the vectors in Liouville space are spin operators. The matrix elements can be calculated in various ways. The Liouvillian is the commutator with the Hamiltonian, so matrix elements can be calculated from the commutation rules of spin operators. Alternatively, the angular momentum properties of Liouville space can be used. In either case, the chemical shift terms are easily calculated, but the coupling terms (since they are products of operators) are more complex. In [section B2.4.2.7](#), the Liouville matrix for the single-quantum transitions for an AB spin system is presented.

Relaxation or chemical exchange can be easily added in Liouville space, by including a Redfield matrix, **R**, for relaxation, or a kinetic matrix, **K**, to describe exchange. The equation of motion for a general spin system becomes [equation \(B2.4.28\)](#).

$$\rho(t) = \exp(-i\mathbf{L} - \mathbf{R} - \mathbf{K})t \rho(0). \quad (\text{B2.4.28})$$

In NMR, the magnetization in the xy plane is detected, so it is the expectation value of the I_x operator that is measured. This is just the unweighted sum of all the I_{x_i} operators for the individual spins i . It may be a function of several time variables (multi-dimensional experiments), including the time during the acquisition,

but it is always given by equation (B2.4.29).

$$\langle I_x(t) \rangle = \text{trace}(I_x \rho(t)). \quad (\text{B2.4.29})$$

In Liouville space, both the density matrix and the I_x operator are vectors. The dot product of these Liouville space

-13-

vectors is the trace of their product as operators. Therefore, the NMR signal, S , as a function of a single time variable, t , is given by equation (B2.4.30), in which the parentheses denote a Liouville space scalar product (compare equation (B2.4.12)).

$$S(t) = (I_x | \rho(t)). \quad (\text{B2.4.30})$$

The experiment starts at equilibrium. In the high-temperature approximation, the equilibrium density operator is proportional to the sum of the I_z operators, which will be called F_z . If there are multiple exchanging sites with unequal populations, p_i , the sum is a weighted one, as in equation (B2.4.31).

$$F_x = \sum_{i=1}^n p_i I_{xi}. \quad (\text{B2.4.31})$$

A simple, non-selective pulse starts the experiment. This rotates the equilibrium z magnetization onto the x axis. Note that neither the equilibrium state nor the effect of the pulse depend on the dynamics or the details of the spin Hamiltonian (chemical shifts and coupling constants). The equilibrium density matrix is proportional to F_z . After the pulse the density matrix is therefore given by F_x and it will evolve as in equation (B2.4.27). If (B2.4.28) is substituted into (B2.4.30), the NMR signal as a function of time t , is given by (B2.4.32). In this equation there is a distinction between the sum of the operators weighted by the equilibrium populations, F_x , from the unweighted sum, I_x . The detector sees each spin (but not each coherence!) equally well.

$$S(t) = (I_x | \exp([-i\mathbf{L} - \mathbf{R} - \mathbf{K}]t) F_x). \quad (\text{B2.4.32})$$

As with the uncoupled case, one solution involves diagonalizing the Liouville matrix, $i\mathbf{L} + \mathbf{R} + \mathbf{K}$. If \mathbf{U} is the matrix with the eigenvectors as columns, and Λ is the diagonal matrix with the eigenvalues down the diagonal, then (B2.4.32) can be written as (B2.4.33). This is similar to other eigenvalue problems in quantum mechanics, such as the transformation to normal co-ordinates in vibrational spectroscopy.

$$S(t) = (I_x | \mathbf{U} \exp(-i\Lambda t) \mathbf{U}^{-1} | F_x). \quad (\text{B2.4.33})$$

Note that the Liouville matrix, $i\mathbf{L} + \mathbf{R} + \mathbf{K}$ may not be Hermitian, but it can still be diagonalized. Its eigenvalues and eigenvectors are not necessarily real, however, and the inverse of \mathbf{U} may not be its complex-conjugate transpose. If complex numbers are allowed in it, equation (B2.4.33) is a general result. Since Λ is a diagonal matrix it can be expanded in terms of the individual eigenvalues, λ_j . The inverse matrix \mathbf{U}^{-1} can be applied ('backwards') to I_x , and we obtain equation (B2.4.34).

$$S(t) = \sum_j (\mathbf{U}^{-1} I_x)_j^* (\mathbf{U} F_x)_j e^{i\lambda_j t}. \quad (\text{B2.4.34})$$

In this equation, the index j runs over all the transitions and the exponents have both real and imaginary parts, which

-14-

give the linewidth and position of the lines. The terms before the exponential are also complex, giving the intensity and phase.

For a general system, these sets of equations are huge, as written. For n spins $-1/2$, the density matrix has 2^{2n} elements, so the Liouville matrix has 2^{4n} elements. However, the density matrix elements can be sorted according to coherence level—the number of quanta associated with the transition. In this case, the matrices block, and the largest single block is the one corresponding to the single-quantum transitions. Its size is the binomial coefficient $(2n)!/(n+1)!(n-1)!$. This can be further divided into blocks based on the factoring from spectral analysis. A transition can change the z quantum number of the spin wavefunction by only ± 1 . For three spins $-1/2$, the eight wave functions are divided as follows: one with z quantum number $+3/2$, three with $+1/2$, three with $-1/2$ and one with $-3/2$. Therefore, the 15 possible transitions are divided into two groups of three ($+3/2 \rightarrow +1/2$, and $-1/2 \rightarrow -3/2$), and a group of nine ($+1/2 \rightarrow -1/2$). Further reductions can be achieved using weak coupling approximations and magnetic equivalence. In practice, system of five or six spins can be treated with modern computers.

B2.4.2.6 GENERALIZED TRANSITION PROBABILITIES

The quantities $(\mathbf{U}^{-1}I_x)_j$ and $(\mathbf{U}F_x)_j$ in B2.4.34 are projections of the eigenvector j along I_x . From the above equations, this can be interpreted as follows. The term $(\mathbf{U}F_x)_j$ is the amount that the transition j received from the total x magnetization created from the equilibrium state and $(\mathbf{U}^{-1}I_x)_j$ is how much that transition contributes to the observed signal. These two terms may not be equal, as will be seen in exchanging systems. An informal way of thinking about these terms is to consider the transition moment, $\langle \phi_f | I_x | \phi_i \rangle$. If the ket–bra operator $|\phi_f\rangle \langle \phi_i|$ represents the transition, then the transition moment is the projection of the transition operator along the I_x operator.

In the usual preparation–evolution–detection paradigm, neither the preparation nor the detection depend on the details of the Hamiltonian, except in special cases. Starting from equilibrium, a hard pulse gives a density matrix that is just proportional to F_z . The detector picks up only the unweighted sum of the spin operators, I_x . It is only during an evolution (perhaps between sampling points in an FID) that these totals need be divided amongst the various lines in the spectrum. Therefore, one of the factors in the transition probability represents the conversion from preparation to evolution; the other factor represents the conversion back from evolution to detection.

Equation (B2.4.33) and equation (B2.4.34) are the basic equations for a time-domain description. For instance, they say that any time-domain NMR signal is the sum of decaying oscillations. This is obvious from the fact that it is described by a first-order differential equation, but (B2.4.34) gives a way of calculating the values of these exponentials for any system, static or dynamic. The distinctions amongst different types of spectrum lie in the eigenvalues and eigenvectors of the Liouville matrix $i\mathbf{L}+\mathbf{R}+\mathbf{K}$. equation (B2.4.34) describes static spectra, spin relaxation and spectra showing the effects of chemical exchange or T_2 relaxation, in a single, unified picture.

B2.4.2.7 EXAMPLE OF THE AB SPIN SYSTEM

For example, the observed transitions of an AB spin system have a Liouville matrix given in equation (B2.4.35). The coupling constant is J , and it is assumed that $\omega_B = -\omega_A = -\delta/2$, so that δ is the frequency difference between the two sites. The angle, θ , is defined for the AB system by the equation $\tan(\theta)=J/2\delta$. The Liouville space basis used here is the superspin equivalent of the four product operators $(I_x^A, I_x^A I_z^B, I_x^B, I_x^B I_z^A)$,

and a set of rules for calculating these elements is given elsewhere [12].

-15-

$$\begin{pmatrix} i\omega_A & iJ/2 & 0 & -iJ/2 \\ iJ/2 & i\omega_A & -iJ/2 & 0 \\ 0 & -iJ/2 & i\omega_B & iJ/2 \\ -iJ/2 & 0 & iJ/2 & i\omega_B \end{pmatrix}. \quad (\text{B2.4.35})$$

The four eigenvalues, which give the positions of the lines, are $\pm J/2 \pm ((J/2)^2 + (\delta/2)^2)^{1/2}$, as expected for an AB system. The matrix of eigenvectors as columns is given in equation (B2.4.36), in which $c = \cos(\theta)$, $s = \sin(\theta)$ and δ is defined above.

$$\text{Eigenvectors} = \begin{pmatrix} c & s & c & s \\ c & s & -c & -s \\ -s & c & s & -c \\ -s & c & -s & c \end{pmatrix}. \quad (\text{B2.4.36})$$

In the basis used in (B2.4.35), the total x magnetization is proportional to the vector (1, 0, 1, 0). Taking the dot product with the eigenvectors shows that the outer lines receive $(\cos\theta - \sin\theta)$ from the total, whereas the inner lines receive $(\cos\theta + \sin\theta)$. The squares of these terms give the familiar AB system intensities: $(1 - \sin 2\theta)$ and $(1 + \sin 2\theta)$.

Once the AB spin system is defined, the effects of chemical exchange can be calculated. This can be either non-mutual or mutual exchange. In the case of non-mutual exchange, there are two blocks, one for each site, and the exchange connecting them, as in equation (B2.4.37). For a simple product basis, the exchange always has this form: the off-diagonal blocks are themselves diagonal and the sum of the exchange contributions in any column must be zero, to preserve the number of spins. In this equation, zeros have been replaced by dots to emphasize the form of the matrix.

$$\begin{pmatrix} i\omega_A - k' & iJ/2 & \cdot & -iJ/2 & k & \cdot & \cdot & \cdot \\ iJ/2 & i\omega_A - k' & -iJ/2 & \cdot & \cdot & k & \cdot & \cdot \\ \cdot & -iJ/2 & i\omega_B - k' & iJ/2 & \cdot & \cdot & k & \cdot \\ -iJ/2 & \cdot & iJ/2 & i\omega_B - k' & \cdot & \cdot & \cdot & k \\ k' & \cdot & \cdot & \cdot & i\omega'_A - k & iJ'/2 & \cdot & -iJ'/2 \\ \cdot & k' & \cdot & \cdot & iJ'/2 & i\omega'_A - k & -iJ'/2 & \cdot \\ \cdot & \cdot & k' & \cdot & \cdot & -iJ'/2 & i\omega'_B - k & iJ'/2 \\ \cdot & \cdot & \cdot & k' & -iJ'/2 & \cdot & iJ/2 & i\omega'_B - k \end{pmatrix}. \quad (\text{B2.4.37})$$

In this equation, the primes on the imaginary parts indicate that the Larmor frequencies and coupling constants will be different. Also, if the equilibrium constant for the exchange is not 1, then the forward and reverse rates will not be equal. Note that the 1,2 block, in the top right, represents the rate from site 2 into site 1.

B2.4.2.8 MUTUAL EXCHANGE IN THE AB SYSTEM

In the case of mutual AB exchange this matrix can be simplified. The equilibrium constant must be 1, so $k = k'$. Also, ω^A is equal to ω_B and vice versa, and the coupling constant is the same. For instance, if \mathbf{L} is the Liouville matrix for one site, then the Liouville matrix for the other site is $\mathbf{P}^{-1}\mathbf{L}\mathbf{P}$, where \mathbf{P} is the matrix describing the permutation.

The exchange matrix, \mathbf{K} , is just the rate, k , times the unit matrix. In block form, the full matrix for two sites is given in the eigenvalue equation, (B2.4.38).

$$\begin{pmatrix} i\mathbf{L} - \mathbf{K} & \mathbf{K} \\ \mathbf{K} & i\mathbf{P}^{-1}\mathbf{L}\mathbf{P} - \mathbf{K} \end{pmatrix} \begin{pmatrix} a \\ \mathbf{P}^{-1}a \end{pmatrix} = \lambda \begin{pmatrix} a \\ \mathbf{P}^{-1}a \end{pmatrix}. \quad (\text{B2.4.38})$$

This equation is equivalent to the pair of equations in (B2.4.39).

$$\begin{aligned} i\mathbf{L}a - \mathbf{K}(1 - \mathbf{P}^{-1})a &= \lambda a \\ i\mathbf{P}^{-1}\mathbf{L}a - \mathbf{K}(\mathbf{P}^{-1} - 1)a &= \lambda \mathbf{P}^{-1}a. \end{aligned} \quad (\text{B2.4.39})$$

Since \mathbf{K} is a multiple of the unit matrix and the permutation is its own inverse, the two equations (B2.4.39) are the same. The Liouvillian for a single site is set up and the exchange is described by $\mathbf{K}(1 - \mathbf{P}^{-1})$.

Application of this approach to [equation \(B2.4.37\)](#) gives equation (B2.4.40). If $\omega_B = -\omega_A = -\delta/2$, the symmetry of the matrix and one additional transformation means that it can be broken into two 2×2 complex matrices, which can be diagonalized analytically. The resulting lineshapes match the published solutions [13].

$$\begin{pmatrix} i\omega_A - k & iJ/2 & k & -iJ/2 \\ iJ/2 & i\omega_A - k & -iJ/2 & k \\ k & -iJ/2 & i\omega_B - k & iJ/2 \\ -iJ/2 & k & iJ/2 & i\omega_B - k \end{pmatrix}. \quad (\text{B2.4.40})$$

B2.4.2.9 INTERMOLECULAR EXCHANGE

The phenomenon of intermolecular exchange is very common. The loss of couplings to hydroxyl protons in all but the very purest ethanol samples was observed at a very early stage. Proton transfer reactions are still probably the most carefully studied [14] class of intermolecular exchange.

In classical kinetics, intermolecular exchange processes are quite different from the unimolecular, first-order kinetics associated with intramolecular exchange. However, the NMR of chemical exchange can still be treated as pseudo-first-order kinetics, and all the previous results apply. One way of rationalizing this is as follows. NMR follows a particular nucleus, but typically only 1 in 10^5 nuclei is ‘visible’ (due to the small Boltzmann population difference). When a visible nucleus exchanges with another nucleus on another molecule, the probability is that the other nucleus is invisible. The exchange partners vastly overwhelm the visible nuclei.

However, all the nuclei have spin. An example of this occurs in the intermolecular exchange of an AB spin system, as in equation (B2.4.41).



In this case, a spin A that was coupled to the α orientation of the B spin may end up, after the exchange, coupled to either α or β . Because of the Boltzmann distribution, the amounts of α and β orientation are each

half of the sample. The first exchange is degenerate, but the second is a change of the B spin. This can be treated as exchange with a site in which the shifts are the same, but the coupling constant is of the opposite sign. If these spin parameters are used in [equation \(B2.4.37\)](#), then the lineshapes in [figure B2.4.5](#) are obtained.

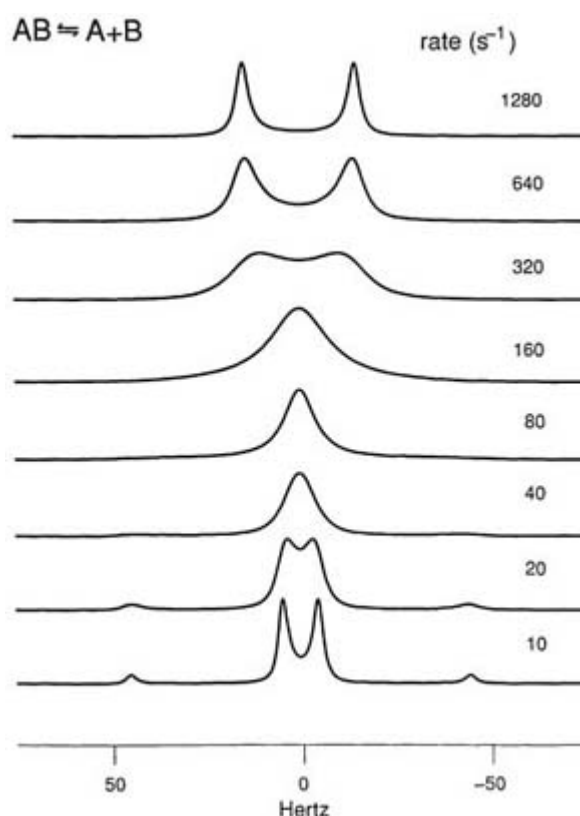


Figure B2.4.5. Simulated lineshapes for an intermolecular exchange reaction in which the bond joining two strongly coupled nuclei breaks and re-forms at a series of rates, given beside the lineshape. In slow exchange, the typical spectrum of an AB spin system is shown. In the limit of fast exchange, the spectrum consists of two lines at the two chemical shifts and all the coupling has disappeared.

B2.4.2.10 CALCULATION OF THE SPECTRUM

Once the basic work has been done, the observed spectrum can be calculated in several different ways. If the problem is solved in the time domain, then the solution provides a list of transitions. Each transition is defined by four quantities: the integrated intensity, the frequency at which it appears, the linewidth (or decay rate in the time domain) and the phase. From this list of parameters, either a spectrum or a time-domain FID can be calculated easily. The spectrum has the advantage that it can be directly compared to the experimental result. An FID can be subjected to some sort of apodization before Fourier transformation to the spectrum; this allows additional line broadening to be added to the spectrum independent of the simulation.

The Bloch equation approach ([equation \(B2.4.6\)](#)) calculates the spectrum directly, as the portion of the spectrum that is linear in a B_1 observing field. Binsch generalized this for a fully coupled system, using an exact density-matrix approach in Liouville space. His expression for the spectrum is given by [equation \(B2.4.42\)](#). Note that this is formally the Fourier transform of [equation \(B2.4.32\)](#), so the time domain and frequency domain are connected as usual.

$$S(\omega) = \text{Re}[F_-(i\mathbf{L} - \mathbf{R} - \mathbf{K})^{-1}M_0]. \quad (\text{B2.4.42})$$

In practice, the spectrum is usually calculated by diagonalizing the matrix first [15], as was done in the time domain. This means that the large matrix does not have to be inverted for each point in the spectrum. However, for very large matrices, it may be numerically more efficient not to diagonalize, but rather to invert at each data point. For a six-spin system, the full matrix is 792×792 , and each additional spin multiplies each dimension by roughly a factor of 4. Since the time for a diagonalization scales roughly as the cube of the dimension of the matrix, larger spin systems become impractical. However, modern sparse-matrix methods for matrix inversion do not suffer from the same dramatic scaling, and so will become more efficient for larger spin systems.

The method for studying intermediate exchange in NMR is to obtain an excellent equilibrium spectrum of the system as a function of temperature. Then the theoretical apparatus developed above can be used to simulate and to fit the experimental data, in order to obtain the rate data.

B2.4.3 FAST EXCHANGE

B2.4.3.1 INTRODUCTION

In the limit of fast exchange, the lineshape of chemical exchange quickly becomes a single Lorentzian line. In [figure B2.4.4\(b\)](#) the negative line broadens directly as the rate, and loses absolute intensity as well. This combination of the increasing width and decreasing integral means that the negative line quickly becomes irrelevant to the experimental lineshape. This becomes a pure Lorentzian, whose width is proportional to $(\Delta\omega)^2/k$, where $\Delta\omega$ is the difference in Larmor frequency of the two sites and k is the exchange rate. Measuring the rate is then equivalent to measuring the spin–spin relaxation time, T_2 . The problem is that unless there is an estimate of $\Delta\omega$ (from a spectrum that is ‘frozen out’) an absolute value of the rate cannot be measured. However, T_2 measurements themselves have an associated timescale. If that can match the exchange rate, then an absolute rate can be measured.

B2.4.3.2 T_2 MEASUREMENTS

In principle, T_2 can be measured directly from the linewidth of the spectrum. However, since experimental linewidths are also governed by inhomogeneous broadening (magnetic field inhomogeneities etc), careful T_2 determinations require methods that cancel out inhomogeneous effects [16]. Three techniques are commonly available: the Carr–Purcell–Meiboom–Gill (CPMG) spin echo experiment, the $T_{1\rho}$ experiment and the offset-saturation method [17]. All three can be implemented easily on modern spectrometers.

The Hahn spin echo, with a single refocusing pulse, eliminates effects due to magnetic field inhomogeneities, so a study of the echo intensity as a function of the echo time should yield T_2 . However, if there is a field gradient present, diffusion in the sample can also attenuate the echo. Even with modern well shimmed magnets, the gradients are still large enough to affect a T_2 measurement in water. In the CPMG experiment, a series of closely spaced refocusing pulses suppresses diffusion effects. In this case, the echo intensity is measured as a function of echo number, and reliable values of T_2 can be obtained.

The parameter $T_{1\rho}$ is the longitudinal relaxation time of a magnetization which is spin locked along a radiofrequency magnetic field. The magnetization is flipped onto the x axis (for instance) by an RF pulse along y . The RF phase is changed by 90° , and the magnetization is then spin locked by the RF field, now along x . When the RF is shut off, the remaining xy magnetization can be detected directly. An analysis of the signal as a function of the spin-locking time yields T_2 . The offset-saturation experiment [18] consists of

irradiating the spin system with a known RF field at some offset from resonance until a steady state is achieved. The z magnetization is then measured by a non-selective observe pulse. A plot (figure B2.4.6) of the partially saturated z magnetization against offset from resonance will show a dip at resonance. The width of this dip is given by equation (B2.4.43).

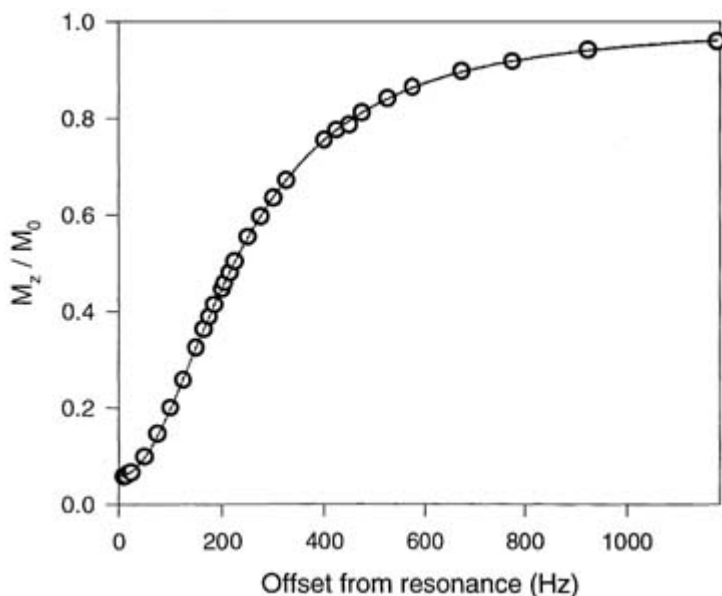


Figure B2.4.6. Results of an offset-saturation experiment for measuring the spin–spin relaxation time, T_2 . In this experiment, the signal is irradiated at some offset from resonance until a steady state is achieved. The partially saturated z magnetization is then measured with a $\pi/2$ pulse. This figure shows a plot of the z magnetization as a function of the offset of the saturating field from resonance. Circles represent measured data; the line is a non-linear least-squares fit. The signal is normal when the saturation is far away, and dips to a minimum on resonance. The width of this dip gives T_2 , independent of magnetic field inhomogeneity.

$$\text{Dip width} = (\gamma B_2) \sqrt{T_1/T_2}. \quad (\text{B2.4.43})$$

-20-

If the strength of the saturating RF, ΔB_2 , and the spin–lattice relaxation time, T_1 , are known, then T_2 can be measured, again free of magnetic field inhomogeneities.

These T_2 measurements can allow reaction rates of 10^5 s^{-1} or more to be measured. In combination with slow- and intermediate-exchange methods, this means that rates can be measured over a range of more than five orders of magnitude. This means that excellent thermodynamic parameters can be obtained. [Figure B2.4.2](#) shows some results on furfural, a system which has unequally populated sites. For this case, the range of rates over which lineshape methods are useful is quite small. In the case of two-site, unequally populated exchange, the minor peak broadens faster than the major peak (the relative rate of broadening is the ratio of the major population to the minor). The minor peak disappears into the baseline quickly, but T_2 measurements can still provide good data.

These experiments yield T_2 which, in the case of fast exchange, gives the ratio $(\Delta\omega)^2/k$. However, since the experiments themselves have an implicit timescale, absolute rates can be obtained in favourable circumstances. For the CPMG experiment, the timescale is the repetition time of the refocusing pulse; for the $T_1\rho$ experiment, it is the rate of precession around the effective RF field. If this timescale is fast with respect to the exchange rate, then the experiment effectively measures T_2 in the absence of exchange. If the timescale is slow, the apparent T_2 contains the effects of exchange. Therefore, the apparent T_2 shows a dispersion as the

timescale of the measurement method is changed [19]. Practical spectrometer considerations of RF heating and duty cycle usually limit these timescales to tens of kilohertz. However, if the conditions are appropriate, this dispersion curve yields an absolute rate.

B2.4.4 SLOW EXCHANGE

B2.4.4.1 INTRODUCTION

The term ‘slow’ in this case means that the exchange rate is much smaller than the frequency differences in the spectrum, so the lines in the spectrum are not significantly broadened. However, the exchange rate is still comparable with the spin–lattice relaxation times in the system. Exchange, which has many mathematical similarities to dipolar relaxation, can be observed in a NOESY-type experiment (sometimes called EXSY). The rates are measured from a series of EXSY spectra, or by performing modified spin–lattice relaxation experiments, such as those pioneered by Hoffman and Forsen [20].

In the absence of exchange (and ignoring dipolar relaxation), each z magnetization will relax back to equilibrium at a rate governed by its own T_1 , as in (B2.4.44).

$$\frac{d}{dt}[M(t) - M(\infty)] = -\frac{1}{T_1}[M(t) - M(\infty)]. \quad (\text{B2.4.44})$$

If there are two sites, A and B, then an analogous equation can be written, as in (B2.4.45).

$$\frac{d}{dt} \begin{bmatrix} M_A(t) - M_A(\infty) \\ M_B(t) - M_B(\infty) \end{bmatrix} = \begin{pmatrix} \frac{1}{T_1^A} & 0 \\ 0 & \frac{1}{T_1^B} \end{pmatrix} \begin{bmatrix} M_A(t) - M_A(\infty) \\ M_B(t) - M_B(\infty) \end{bmatrix}. \quad (\text{B2.4.45})$$

-21-

If the two sites exchange with rate k during the relaxation, then a spin can relax either through normal spin–lattice relaxation processes, or by exchanging with the other site. [equation \(B2.4.45\)](#) becomes (B2.4.46).

$$\frac{d}{dt} \begin{bmatrix} M_A(t) - M_A(\infty) \\ M_B(t) - M_B(\infty) \end{bmatrix} = \begin{pmatrix} -\frac{1}{T_1^A} - k & k \\ k & -\frac{1}{T_1^B} - k \end{pmatrix} \begin{bmatrix} M_A(t) - M_A(\infty) \\ M_B(t) - M_B(\infty) \end{bmatrix}. \quad (\text{B2.4.46})$$

This equation is very similar to (B2.4.13). The basic situation is just as in intermediate exchange, except that it describes z magnetizations rather than xy . The frequencies are zero, and the matrix now has pure real eigenvalues, but the approach is the same. The time domain is a natural one for slow exchange, since a relaxation experiment follows the z magnetizations as a function of time. As before, the time dependence is obtained by diagonalizing the relaxation/exchange matrix and calculating the magnetizations for each time at which they are sampled. In this case, the solution is given by [equation \(B2.4.47\)](#), the same as [equation \(B2.4.14\)](#), except there are no imaginary terms.

$$\begin{bmatrix} M_A(t) - M_A(\infty) \\ M_B(t) - M_B(\infty) \end{bmatrix} = \exp(-[R + K]t) \begin{pmatrix} M_A(0) - M_A(\infty) \\ M_B(0) - M_B(\infty) \end{pmatrix}. \quad (\text{B2.4.47})$$

B2.4.4.2 TWO-DIMENSIONAL METHODS

There are two main applications of slow chemical exchange: one is to determine the qualitative mechanism, and the other is to measure the rates of the processes as accurately as possible. For the first case, in which we have a spectrum in slow exchange, we need to establish the mechanism: which site is exchanging with which. For this purpose, the homonuclear two-dimensional experiment EXSY (the same pulse sequence as NOESY, but involving exchange) is by far the best technique to use. Exchange between sites leads to a pair of symmetrical cross-peaks joining the diagonal peaks of the same site, so the mechanism is very obvious.

The EXSY pulse sequence starts with two $\pi/2$ pulses separated by the incrementable delay, t_1 . This modulates the z magnetizations, so that the relaxation that occurs during the mixing time which follows, t_m , is frequency labelled. Finally, the z magnetizations are sampled with a third $\pi/2$ pulse. Magnetization from a different site that enters via exchange will have a different frequency label. A two-dimensional Fourier transform then produces the spectrum. The initial rate of increase of the cross-peak gives the rate of exchange. A series of EXSY experiments as a function of mixing times will define the mechanism and give an estimate of the rates.

However, care must be taken in choosing the mixing time if there are multiple exchange processes. If the mixing time is too long, there is a substantial probability that a spin may have exchanged twice in that time, leading to spurious cross-peaks. Orrell and his group [21] have solved this problem by treating the effect of exchange on the z magnetizations correctly, and have written a program which simulates the two-dimensional spectrum as a function of the mixing time.

Since exchange and coupled relaxation have the same mathematical form, both may contribute to a NOESY/EXSY spectrum, as in [figure B2.4.7](#). For small molecules, since the NOE is positive and exchange creates saturation transfer (like a negative NOE), the NOESY and EXSY cross peaks have opposite signs. For macromolecules in the spin-diffusion limit, the peaks have the same sign, but exchange cross-peaks can usually be distinguished by their much stronger temperature dependence.

-22-

Figure B2.4.7. Contour plot of a phase-sensitive NOESY/EXSY spectrum of the derivative of TEMPO. The peaks are positive, with the exception of the circled peaks, which are negative. The spectrum shows the exchange of the four methyl groups in this molecule. A combination of a ring-flip and inversion at nitrogen means that the axial methyl on one side of the ring exchanges with the equatorial methyl on the other side. The positive cross-peaks show the exchange of the two sets of methyls. However, there are also NOE cross-peaks between methyl groups on the same side of the ring. Since this is a small molecule, the NOE peaks are of opposite sign to the exchange peaks. There are two NOE cross-peaks for each methyl, since the exchange process is relatively fast, and distributes the NOE between the two exchange partners.

B2.4.4.3 ONE-DIMENSIONAL SELECTIVE-INVERSION METHODS

For careful rate measurements, once the mechanism is established, it is our opinion that one-dimensional methods are superior to quantitative 2D ones. Apart from the fact that 1D spectra can be integrated more easily, there is also more control over the experiment. Modern spectrometers can create almost any type of selective excitation, so that there is control of the conditions at the start of the relaxation. For two sites, a non-selective inversion that inverts both sites equally will mask most of the exchange effects and the relaxation will be dominated by T_1 . However, if one site is inverted selectively, then that site can regain equilibrium by either T_1 processes or by exchanging with the other site that was left at equilibrium [20]. The inverted signal will relax at roughly the sum of the exchange and spin-lattice relaxation rate, while the signal that was unperturbed at the start of the experiment shows a characteristic transient, as in [figure B2.4.8](#). These one-dimensional selective-inversion experiments have been widely used in systems without scalar coupling, such as methyl groups or ^{13}C spectra.

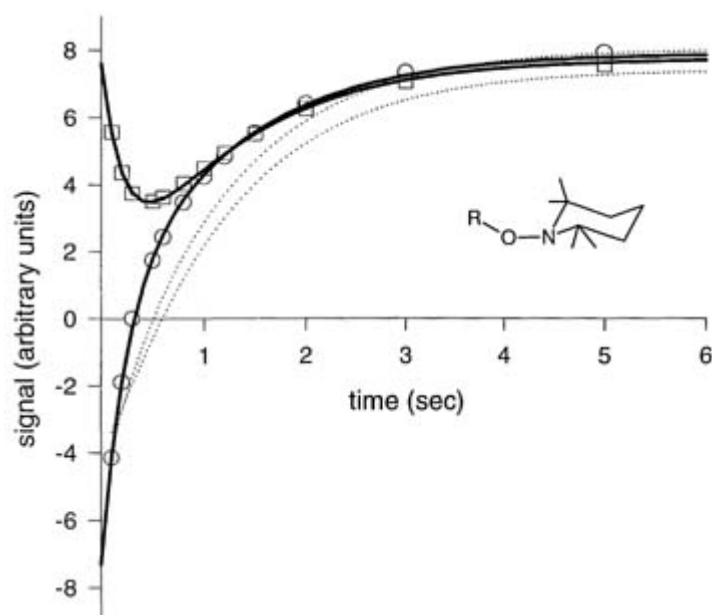


Figure B2.4.8. Relaxation of two of the exchanging methyl groups in the TEMPO derivative in [figure B2.4.7](#). The dotted lines show the relaxation of the two methyl signals after a non-selective inversion pulse (a typical T_1 experiment). The heavy solid line shows the recovery after the selective inversion of one of the methyl signals. The inverted signal (circles) recovers more quickly, under the combined influence of relaxation and exchange with the non-inverted peak. The signal that was not inverted (squares) shows a characteristic transient. The lines represent a non-linear least-squares fit to the data.

For multiple sites, a wide range of initial conditions is available. For instance, in a three-site exchange amongst A, B and C, the signal due to A can be inverted selectively. This will provide rate information on the A–B exchange and the A–C exchange, but relatively little about B–C. This is one example of using the initial conditions to suppress or enhance the observation of particular processes. The definition of how selective a selective inversion may be gives an added degree of control over the experiment.

The selective-inversion experiment gives excellent rate data. The description of the time evolution of the z magnetizations in [equation \(B2.4.47\)](#) is exact. There are no assumptions about short mixing times or initial rates. Standard non-linear least-squares methods allow fitting these curves to the measured data and deriving values for the rates involved. Believing in the error estimates of these multi-parameter fitting procedures can be dangerous, however. A more reliable error estimate can be obtained by ‘profiling’. In this procedure, a global fit to all parameters is done first. Then the rate (or any other parameter of interest) is fixed at a different value, and the data are re-fitted using all the other parameters. As the rate is moved from the optimum value, the fit will become worse, as measured by the sum of the squares of the deviations between real data and the model. When the ‘badness of fit’ exceeds a critical level of the F statistic, the value of the rate is at the end of a confidence interval. This confidence interval can be several times larger than the one calculated from the usual standard deviation in a nonlinear least-squares fit. Even with this error estimate, it is possible to measure rates with errors of less than 10%. [Figure B2.4.8](#) shows the quality of result that is possible.

In a selective-inversion experiment, it is the relaxation of the z magnetizations that is being studied. For a system without scalar coupling, this is straightforward: a simple pulse will convert the z magnetizations directly into observable signals. For a coupled spin system, this relation between the z magnetizations and the observable transitions is much more complex [22].

In a coupled spin system, the number of observed lines in a spectrum does not match the number of independent z magnetizations and, furthermore, the spectra depend on the flip angle of the pulse used to observe them. Because of the complicated spectroscopy of homonuclear coupled spins, it is only recently that selective inversions in simple coupled spin systems [23] have been studied. This means that slow chemical exchange can be studied using proton spectra without the requirement of single characteristic peaks, such as methyl groups.

The z magnetizations of the spin system are key to the problem, since the exchange is measured in competition with their relaxation processes. However, for a coupled spin system, the lines in the spectrum do not directly reflect the z magnetizations. Even for two weakly coupled spins, there are four lines in the spectrum (two doublets), but there are only three independent z magnetizations. There are indeed four energy levels, but the sum of their populations must be constant. There are three independent quantities: the total magnetization of A , the total X magnetization, and a shared $I_z J_z$ magnetization. For coupled systems, especially those with strong coupling, the relation of the z magnetizations to the observed spectrum can be quite complex.

There are several complications. One is the flip angle dependence of the spectra [22, 24]. For a non-equilibrium state of a coupled spin system, the observed intensities of the lines depend on the flip angle used to observe them. In particular, spectra are only 'true' reflections of the z magnetizations in the limit of small flip angles. A further complication arises because the z magnetizations are part of a larger manifold of coherences that also includes the zero-quantum transitions. Both the z magnetizations and the zero-quantum transitions have coherence level zero and they cannot be separated by pulses or phase cycling [11]. This is the problem with the zero-quantum coherences in NOESY, for instance. For instance, for three spins there are eight z magnetizations, one of which is fixed as the total number of spins. However, there are 20 coherence level zero density matrix elements, leaving six pairs of zero quantum transitions. The 15 observable xy magnetizations for a three-spin system cannot correspond directly to the eight z magnetizations. Provided that these complications are recognized, they can be treated easily with standard spin-dynamics techniques.

The xy magnetizations can also be complicated. For n weakly coupled spins, there can be $n^* 2^{n-1}$ lines in the spectrum and a strongly coupled spin system can have up to $(2n)!/((n-1)!(n+1)!)$ transitions. Because of small couplings, and because some lines are weak combination lines, it is rare to be able to observe all possible lines. It is important to maintain the distinction between mathematical and practical relationships for the density matrix elements.

These complications require some careful analysis of the spin systems, but fundamentally the coupled spin systems are treated in the same way as uncoupled ones. Measuring the z magnetizations from the spectra is more complicated, but the analysis of how they relax is essentially the same.

B2.4.5 EXCHANGE IN SOLIDS

Exchange in the solid state follows the same basic principles as in liquids. The classic Cope re-arrangement of bullvalene occurs in both the liquid and solid state [25], and the lineshapes in the spectra are similar. However, because of chemical shielding anisotropy (CSA) and quadrupolar and dipolar effects, the Larmor

frequency of a given spin depends on the orientation of the individual molecule. In a liquid, where there is isotropic tumbling, these effects average out and exchange is only evident if there is a change in isotropic chemical shift. In a solid, almost any type of molecular motion can cause lineshape and other effects. Furthermore, many NMR spectra of solids are run under the conditions of magic-angle spinning. This introduces a further timescale into the spectroscopy: the spinning rate, which can now go up to 25 kHz or more. The basic principles are the same, but the systems studied and the observed phenomena can be quite different.

Intermediate exchange is the regime in which lineshape changes are the most obvious manifestation. In ^{13}C magic-angle spinning (MAS) spectra, there can be all the liquid-like coalescence phenomena [26]. These can be analysed just as before. Another type occurs when a molecule re-orientates in the crystal lattice. Since the magnetic environment of the nucleus is anisotropic, the Larmor frequency of the spin changes. One of the most familiar examples is the effect of dynamics on powder patterns in deuterium spectra [27]. In a typical carbon–deuterium bond, the quadrupole coupling of the deuterium nucleus is about 160 kHz. This defines a timescale that is very useful for polymers and biological membranes, and deuterium spectra are widely used. Quite detailed information is available: lineshapes are significantly different for twofold jumps, threefold jumps or continuous diffusion. For instance, in a *tert*-butyl group, overall rotation can be distinguished from rotations of individual methyl groups.

Magic-angle spinning of solid samples provides an experimental parameter not available in liquids. A full analysis of the combined effects of dynamics and MAS for all relative timescales is a very complex problem, but for slow exchange there are techniques that are intuitive. The first is EXSY, which works well in MAS spectra, although the interpretation is confused by the phenomenon of spin diffusion. Cross-peaks can be due to the exchange of spin polarization (rather than the nuclei themselves) via the dipolar interaction. One-dimensional methods that use MAS are also available. The TOSS pulse sequence will eliminate spinning sidebands, provided there is no internal dynamics in the sample. If there is exchange on the timescale of a rotor period, the careful cancellation will no longer work, and sidebands will reappear [28]. Another use of spinning is the ODESSA [29] pulse sequence, which selectively inverts some of the sidebands, which then relax back due to chemical exchange.

B2.4.6 CONCLUSIONS

In order to study chemical exchange in NMR, it is necessary to have a scale against which to measure it. In fast and intermediate exchange, the timescale is the difference in Larmor frequency between the two sites. As the exchange rate approaches this timescale, the lines in the one-dimensional spectrum broaden, coalesce and sharpen into single lines. For intermediate exchange, the lineshape provides the best information. In fast exchange, the rate is starting to dominate the timescale, but useful information can still be extracted from T_2 measurements. In slow exchange, the spin–lattice relaxation provides a timescale. In this regime, modifications of methods for measuring T_1 and the NOE

are used. In MAS spectra of solids, the spinning rate provides another timescale. When the timescales match, dramatic effects can often be observed.

Once the exchange has been established, it is necessary to measure its rate. This is usually done by simulating the NMR experiment with a mathematical model and adjusting the rate in the model until it matches experiment. This means simulating the lineshape in intermediate exchange, or simulating the coupled relaxation of the z magnetizations in slow exchange. In both these cases, the model is similar. There is a matrix which describes each site in the exchange. These matrices form the diagonal of a larger block-diagonal matrix. The blocks are then connected by the exchange process, to form one large matrix. The exchange is

described by the eigenvalues and eigenvectors of the single large matrix.

The exact form of the matrices depend on the situation. In slow exchange, the matrices are real and they model the multi-exponential relaxation of the z magnetizations. In intermediate exchange, the matrix has both real and imaginary parts, as do the eigenvalues and eigenvectors. The model in this case produces a series of transitions, whose intensity, phase, position and width are given by a complex-valued transition probability. The intermediate-exchange spectrum is just the sum of these transitions.

The NMR methods for studying chemical exchange are fundamentally no different from standard NMR methods. Chemical exchange effects appear in the spectrum and in measurements of the relaxation times, so careful measurement of these will provide good exchange data. Perhaps this is the single conclusion: apart from some algebraic and numerical details, chemical exchange is identical to 'normal' NMR.

REFERENCES

- [1] Gutowsky H S and Holm C H 1956 Rate processes and nuclear magnetic resonance spectra. II. Hindered internal rotation of amides *J. Chem. Phys.* **25** 1228–34
- [2] McConnell H M 1958 Reaction rates by nuclear magnetic resonance *J. Chem. Phys.* **28** 430–1
- [3] Reeves L W and Shaw K N 1970 Nuclear magnetic resonance studies of multi-site chemical exchange. I. Matrix formulation of the Bloch equations *Can. J. Chem.* **48** 3641–53
- [4] Bain A D and Duns G J 1996 A unified approach to dynamic NMR based on a physical interpretation of the transition probability *Can. J. Chem.* **74** 819–24
- [5] Sack R A 1958 A contribution to the theory of the exchange narrowing of spectral lines *Mol. Phys.* **1** 163–7
- [6] Binsch G 1969 A unified theory of exchange effects on nuclear magnetic resonance lineshapes *J. Am. Chem. Soc.* **91** 1304–9
- [7] Kleier D A and Binsch G 1970 General theory of exchange-broadened NMR line shapes. II. Exploitation of invariance properties *J. Magn. Reson.* **3** 146–60
- [8] Ernst R R, Bodenhausen G and Wokaun A 1987 *Principles of Nuclear Magnetic Resonance in One and Two Dimensions* (Oxford: Clarendon)

- [9] Fano U 1964 Liouville representation of quantum mechanics with application to relaxation processes *Lectures on the Many Body Problem* vol 2, ed E R Caianiello (New York: Academic) pp 217–39
- [10] Bain A D 1988 The superspin formalism for pulse NMR *Prog. Nucl. Magn. Reson. Spectrosc.* **20** 295–315
- [11] Bain A D 1984 Coherence levels and coherence pathways in NMR. A simple way to design phase cycling procedures *J. Magn. Reson.* **56** 418–27
- [12] Banwell C N and Primas H 1963 On the analysis of high-resolution nuclear magnetic resonance spectra. I. Methods of calculating NMR spectra *Mol. Phys.* **6** 225–56
- [13] Alexander S 1962 Exchange of interacting nuclear spin in nuclear magnetic resonance. I. Intramolecular exchange *J. Chem. Phys.* **37** 967–74
- [14] Limbach H H 1991 Dynamic NMR spectroscopy in the presence of kinetic hydrogen/deuterium isotope effects *NMR Basic Principles and Progress* vol 23, ed P Diehl, E Fluck, H Günther, R Kosfeld and J Seelig (Berlin:

Springer) p 63–164

- [15] Gordon R G and McGinnis R P 1968 Lineshapes in molecular spectra *J. Chem. Phys.* **49** 2455–6
- [16] Freeman R and Hill H D W 1975 Determination of spin–spin relaxation time in high-resolution NMR *Dynamic Nuclear Magnetic Resonance Spectroscopy* ed L M Jackman and F A Cotton (New York: Academic) p 131–62
- [17] Bain A D, Duns G J, Ternieden S, Ma J and Werstik N H 1994 The barrier to internal rotation and chemical exchange in N-acetylpyrrole. A study based on novel NMR methods and molecular modelling *J. Phys. Chem.* **98** 7458–63
- [18] Bain A D and Duns G J 1994 Simultaneous determination of spin–lattice (T1) and spin–spin (T2) relaxation times in NMR; a robust and facile method for measuring T2. Optimization and data analysis of the offset-saturation experiment *J. Magn. Reson. A* **109** 56–64
- [19] Deverell C, Morgan R E and Strange J H 1970 Studies of chemical exchange by nuclear magnetic relaxation in the rotating frame *Mol. Phys.* **18** 553–9
- [20] Hoffman R A and Forsen S 1966 Transient and steady-state Overhauser experiments in the investigation of relaxation processes. Analogies between chemical exchange and relaxation *J. Chem. Phys.* **45** 2049–60
- [21] Abel E W, Coston T P J, Orrell K G, Sik V and Stephenson D 1986 Two-dimensional NMR exchange spectroscopy. Quantitative treatment of multisite exchanging systems *J. Magn. Reson.* **70** 34–53
- [22] Schäublin S, Höhener A and Ernst R R 1974 Fourier spectroscopy of non-equilibrium states. Application to CIDNP, Overhauser experiments and relaxation time measurements *J. Magn. Reson.* **13** 196–216
- [23] McClung R E D and Aarts G H M 1995 Multisite magnetization transfer in strongly coupled spin systems *J. Magn. Reson. A* **115** 145–54
- [24] Bain A D and Martin J S 1978 FT NMR of non-equilibrium states of complex spin systems: I. A Liouville space description *J. Magn. Reson.* **29** 125–35
- [25] Olivier L, Poupko R, Zimmermann H and Luz Z 1996 Bond shift tautomerism of bibullvalenyl in solution and in the solid-state—a C-13 NMR-study *J. Phys. Chem.* **100** 17 995–18 003
- [26] Lyster J R, Yannoni C S and Fyfe C A 1982 Chemical applications of variable-temperature CPMAS NMR spectroscopy in solids *Accounts Chem. Res.* **15** 208–16

-28-

- [27] Vold R R and Vold R L 1991 Deuterium relaxation in molecular solids *Adv. Magn. Opt. Reson.* **16** 85–171
- [28] Yang Y, Schuster M, Blümich B and Spiess H W 1987 Dynamic magic-angle spinning NMR spectroscopy: exchange-induced sidebands *Chem. Phys. Lett.* **139** 239–43
- [29] Gerardy-Montouillout V, Malveau C, Tekely P, Olender Z and Luz Z 1996 Odessa, a new 1D NMR exchange experiment for chemically equivalent nuclei in rotating solids *J. Magn. Reson. A* **123** 7–15

FURTHER READING

Sandstrom J 1982 *Dynamic NMR Spectroscopy* (London: Academic)

An excellent introductory text.

Jackman L M and Cotton F A 1975 *Dynamic Nuclear Magnetic Resonance Spectroscopy* (New York: Academic)

A collection of reviews, covering essentially all of dynamic NMR up to about 1974.

Johnson C S 1965 Chemical rate processes and magnetic resonance *Adv. Magn. Reson.* **1** 33–102

One of the first complete reviews of the subject, and still very useful.

Orrell K G, Sik V and Stephenson D 1990 Quantitative investigations of molecular stereodynamics by 1D and 2D NMR methods *Prog. Nucl. Magn. Reson. Spectrosc.* **22** 141–208

Perrin C L and Dwyer T 1990 Application of two-dimensional NMR to kinetics of chemical exchange *Chem. Rev.* **90** 935–67

These two reviews (Orrell *et al* and Perrin and Dwyer) cover the modern pulse and two-dimensional NMR techniques for studying exchange.

-1-

B2.5 Gas-phase kinetics studies

David Luckhaus and Martin Quack

B2.5.1 INTRODUCTION

The key to experimental gas-phase kinetics arises from the measurement of time, concentration, and temperature. Chemical kinetics is closely linked to time-dependent observation of concentration or amount of substance. Temperature is the most important single statistical parameter influencing the rates of chemical reactions (see [chapter A3.4](#) for definitions and fundamentals).

The rich history of experimental chemical kinetics can be broadly classified according to various conceptual phases. The starting point of quantitative chemical kinetics was the formulation, in 1850 by Wilhelmy [1], of the time dependence of concentrations by a differential equation corresponding to a pseudo-first-order rate law for the hydrolysis ('inversion') of cane sugar. The observation of the concentration of cane sugar was carried out spectroscopically in the early experiments by following the time-dependent rotation of the plane of polarized light by the reaction mixture after mixing the reactant and the catalyst (the acid). During the following half-century, until about 1900, the nature of such phenomenological rate laws was clarified, as was the role of the rate constant and its temperature dependence. This epoch is characterized by the concepts introduced by van't Hoff 2 and Arrhenius 3. It became clear that a distinction must be made between *phenomenological rate laws* for reactions following from a compound mechanism, such as the inversion of cane sugar, and rate laws and rate constants for *elementary reactions*, which can be combined into a compound mechanism. These new concepts characterize the second phase of chemical kinetics studies. For about half a century from 1900 to 1950, experimental investigations concentrated on elementary reactions and the mechanisms in which they are combined. The fathers of gas-phase kinetics, such as Bodenstein, Lindemann, and Hinshelwood, may be named as the representatives of this period.

During the course of these studies the necessity arose to study ever-faster reactions in order to ascertain their elementary nature. It became clear that the mixing of reactants was a major limitation in the study of fast elementary reactions. Fast mixing had reached its high point with the development of the accelerated and stopped-flow techniques [4, 5], reaching effective time resolutions in the millisecond range. Faster reactions were then frequently called 'immeasurably fast reactions' [1].

The new concept overcoming this limitation in the third phase of experimental kinetics started around 1950,

and consisted of initiating a chemical reaction by a very fast physical perturbation and measuring the subsequent relaxation kinetics without a mixing step in the experiment. Various schools developed techniques along these lines, such as Norrish and Porter with flash photolysis [7, 8], and Eigen's school with *T*-jump and other relaxation techniques [6]. Weller's kinetics of fluorescence change can be classified as an indirect technique along these lines [9, 10 and 11], and the Davidson and Jost schools developed shock-wave methods [12]. While the ideas for such techniques can be traced to earlier theoretical papers by Nernst and Einstein, the actual experimental developments started around 1950, initiating what has been called 'a race against time' [7, 8]. In particular, in relation to the laser-flash photolysis and pump-probe techniques developed after 1960, the essence of this race is to generate ever shorter laser pulses for 'pumping' a sample and well-controlled pulse delays for probing the sample's time evolution, where the lengths of the pulses define

-2-

the limits of the time resolution of the techniques. Nanosecond resolution was available by about 1966, picosecond resolution around 1970, and by about 1985 the domain around 10 fs had been reached. Since that time progress by these techniques towards shorter times has slowed down somewhat, a typical value being about 5 fs today (see [chapter B2.1](#)). One might mention, however, a paper on the '0 fs' pulse [13] (dated 1 April 1990).

In parallel with this race against time, new experimental concepts were introduced, which escape the race by switching to a conceptually new approach. Among these are the molecular-beam-scattering techniques developed by Datz, Taylor, Martin, Herschbach, Lee, and others [14, 15], measuring reaction cross sections without time resolution instead of time evolution and reaction rates. NMR line-shape methods proposed by Gutowsky and Holm [16, 17] are further examples for alternative techniques, where time-dependent rates are calculated from time independent information. Neither of these techniques, however, is very well suited to study very fast intramolecular primary processes, which mark the starting point of all chemical reactions and may be considered to characterize the present fourth phase of experimental chemical kinetics. The new concept making these processes accessible to experimental investigation starts from high-resolution molecular spectra in the frequency domain (stationary or at least without short-time resolution) to derive ultimately the full molecular quantum-chemical kinetics in the time ranges from nanoseconds to attoseconds [18]. This experimental approach was, in fact, historically the first one to provide non-trivial three-dimensional molecular quantum wave packet dynamics [18, 19, and 20], some time even before the first one-dimensional molecular quantum wave packet kinetics became available by short-pulse techniques (see [7, 21, 22 and 23] and references cited therein). The scope of the present chapter is to cover the most important experimental techniques in current use, including some of the well-established but also some of the most recent ones, together with current developments and to illustrate them with a few typical examples of results. The presentation by necessity is exemplary and not exhaustive.

On a modest level of detail, kinetic studies aim at determining overall phenomenological rate laws. These may serve to discriminate between different mechanistic models. However, to *it prove* a compound reaction mechanism, it is necessary to determine the rate constant of each elementary step individually. Many kinetic experiments are devoted to the investigations of the temperature dependence of reaction rates. In addition to the obvious practical aspects, the temperature dependence of rate constants is also of great theoretical importance. Many statistical theories of chemical reactions are based on thermal equilibrium assumptions. Non-equilibrium effects are not only important for theories going beyond the classical transition-state picture. Eventually they might even be exploited to control chemical reactions [24]. This has led to the increased importance of energy or even quantum-state-resolved kinetic studies, which can be directly compared with detailed quantum-mechanical models of chemical reaction dynamics [25, 26].

Many experimental methods may be distinguished by whether and how they achieve time resolution—directly or indirectly. Indirect methods avoid the requirement for fast detection methods, either by determining relative rates from product yields or by transforming from the time axis to another coordinate, for example the distance or flow rate in flow tubes. Direct methods include (laser-) flash photolysis [27], pulse radiolysis [28]

(see also [chapter A3.5](#) on ion reactions), and the important relaxation techniques, which study the relaxation of a reacting system back into equilibrium after a perturbation [29]. Here one distinguishes two types of perturbation methods: (i) small perturbations from equilibrium leading to relaxation kinetics in the narrower sense with generalized first-order kinetics [6] and (ii) large perturbations from equilibrium leading to generally nonlinear rate laws. Typical examples for large perturbations are temperature jumps achieved through laser heating or in shock-tube experiments [30].

The time resolution of these methods is determined by the time it takes to initiate the reaction, for example the mixing time in flow tubes or the laser pulse width in flash photolysis, and by the time resolution of the detection. Relatively

-3-

slow reactions can be monitored by taking samples and quickly quenching the reaction by cooling to low temperature or by dilution. The samples can then be analysed using conventional analytical techniques [31]. For time-dependent monitoring, any physical–chemical property changing during the course of a reaction can be used. Perhaps most important are spectroscopic detection techniques, particularly IR (infrared) and UV–VIS (ultraviolet–visible) absorption and fluorescence techniques. Directly recording spectroscopic signals as a function of time can achieve a time resolution of about 0.1 ns in favourable cases. Even higher time resolution of a few femtoseconds can be realised in pump–probe, laser flash-photolysis experiments [22, 32] (see also [chapter B2.1](#) on ultrafast spectroscopy).

A completely different approach, in particular for fast unimolecular processes, extracts state-resolved kinetic information from molecular spectra without using any form of time-dependent observation. This includes conventional line-shape methods, as well as the quantum-dynamical analysis of rovibrational overtone spectra [18, 33, 34 and 35].

At this point, we only mention the very important molecular-beam techniques [15, 27, 36, 37 and 38] that allow the study of isolated molecules, largely without thermal congestion. They are ideally suited to the investigation of unimolecular processes, in particular dissociation reactions and energy redistribution processes (see [chapter A3.13](#) on energy redistribution in reacting systems). The determination of state-resolved cross sections for bimolecular reactions in crossed molecular beams has paved the way for mechanistic investigations of elementary processes in the greatest possible detail (see [chapter B2.3](#) *Reactive scattering*).

B2.5.2 FLOW TUBES

[Figure B2.5.1](#) schematically illustrates a typical flow-tube set-up. In gas-phase studies, it serves mainly two purposes. On the one hand it allows highly reactive shortlived reactant species, such as radicals or atoms, to be prepared at well-defined concentrations in an inert buffer gas. On the other hand, the flow replaces the time dependence, t , of a reaction by the dependence on the distance x from the point where the reactants are mixed by the simple transformation with the flow velocity v_f :

$$t - t_0 = \frac{x - x_0}{v_f}. \quad (\text{B2.5.1})$$

Instead of shifting the detector position, as indicated in [figure B2.5.1](#) one often varies the location of the reactant mixing region using moveable injectors. This allows complex, possibly slow, but powerful, analytical techniques to be used for monitoring gas-phase reactions. In combination with mass-spectrometric detection,

both reactants and products can be monitored quantitatively [39, 40 and 41]. A further possibility consists of keeping the position $(x-x_0)$ in equation B2.5.1 constant and varying the flow velocity v_f thereby varying $(t-t_0)$. This technique is called ‘accelerated flow’.

-4-

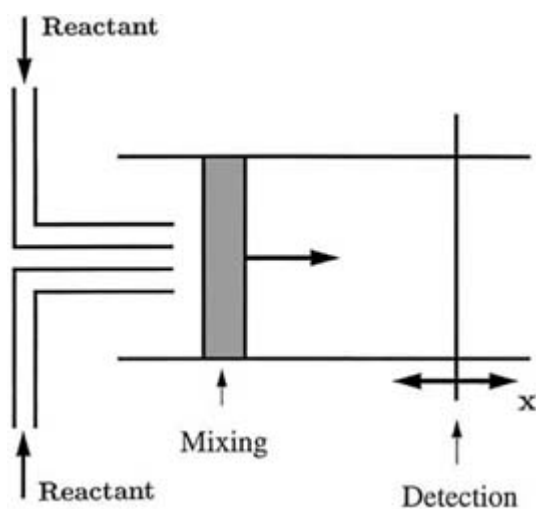


Figure B2.5.1. Schematic representation of a typical flow tube set-up with moveable detection. Adapted from [110].

The time-to-distance transformation requires fast mixing and a known flow profile, ideally a turbulent flow with a well-defined homogeneous composition perpendicular to the direction of flow (‘plug-flow’), as indicated by the shaded area in figure B2.5.1. More complicated profiles may require numerical transformations.

One of the major limiting factors for the time resolution of flow-tube experiments is the time required for mixing reactants and—to a lesser extent—the resolution of distance. With typical fast flow rates of more than 25 ms^{-1} [42, 43] the time resolution lies between milliseconds and microseconds.

Modern applications of the technique include kinetic studies of post-combustion processes and their complex reaction systems. The influence of traces of NO_x on the reaction kinetics of the H_2/O_2 and $\text{CO}/\text{H}_2\text{O}/\text{O}_2$ systems has recently been investigated in a high-pressure, turbulent-flow reactor at pressures up to 14 atm between 750 K and 1100 K [31, 44]. The reaction was monitored by taking samples at a fixed position and varying the location where the fuel ($\text{CO}/\text{H}_2\text{O}/\text{NO}_x$ or H_2/NO_x) is injected into a hot stream of O_2 . The samples were instantly quenched in the hot-water-cooled sampling probe and analysed with a variety of analytical techniques including Fourier-transform infrared spectroscopy. The results were interpreted in terms of a reaction mechanism including 52 elementary reactions, assuming instant mixing and homogeneous composition perpendicular to the flow direction.

NO generally catalyses ‘fuel consumption’ by transforming hydroperoxyl radicals into highly-reactive hydroxyl radicals:





-5-

As a stable radical, however, NO can also catalyze the recombination of radicals (X,Y) at higher concentrations, eventually inhibiting overall oxidation [45]:



The balance of these two effects was found to depend delicately on the stoichiometry, pressure, and temperature. The results were used to develop a more comprehensive CO/H₂O/O₂/NO_x reaction mechanism, incorporating the explicit fall-off behaviour of recombination reactions [46, 47].

B2.5.3 RELAXATION METHODS

Two types of relaxation techniques are distinguished, depending on whether the perturbation applied is small or large.

B2.5.3.1 RELAXATION AFTER A SMALL PERTURBATION FROM EQUILIBRIUM

Perturbation or relaxation techniques are applied to chemical reaction systems with a well-defined equilibrium. An ‘instantaneous’ change of one or several state functions causes the system to relax into its new equilibrium [29]. In gas-phase kinetics, the perturbations typically exploit the temperature (*T*-jump) and pressure (*P*-jump) dependence of chemical equilibria [6]. The relaxation kinetics are monitored by spectroscopic methods.

T-jump techniques can achieve fast heating of the reaction system by pulsed radiation, for example with a microwave source or an IR laser. In the latter case one often adds an efficient inert absorber, such as SF₆. The heating of the reaction system then results from fast collisional relaxation of the initially-excited absorber molecules [48, 49].

When the perturbation is small, the reaction system is always close to equilibrium. Therefore, the relaxation follows generalized first-order kinetics, even if bi- or trimolecular steps are involved (see [chapter A3.4](#)). Take, for example, the reversible bimolecular step



With equilibrium concentrations c^{eq} , the (small) deviation from equilibrium is given by

$$\Delta_c = c_A - c_A^{\text{eq}} = c_B - c_B^{\text{eq}} = c_C^{\text{eq}} - c_C = c_D^{\text{eq}} - c_D. \quad (\text{B2.5.91})$$

$$k_2 c_A^{\text{eq}} c_B^{\text{eq}} = k_{-2} c_C^{\text{eq}} c_D^{\text{eq}} \quad (\text{B2.5.10})$$

-6-

and neglecting terms quadratics in Δ_c leads to the approximate first-order rate law for this elementary step:

$$-\frac{dc_A}{dt} = -\frac{d\Delta_c}{dt} = k_{\text{eff}} \Delta_c \quad (\text{B2.5.11})$$

For this reaction alone, one would thus obtain a simple exponential relaxation with relaxation time

$$k_{\text{eff}} = \{k_2(c_A^{\text{eq}} + c_B^{\text{eq}}) + k_{-2}(c_C^{\text{eq}} + c_D^{\text{eq}})\}. \quad (\text{B2.5.12})$$

More generally, the relaxation follows generalized first-order kinetics with several relaxation times τ_i , as depicted schematically in figure B2.5.2 for the case of three well-separated time scales. The various relaxation times determine the turning points of the product concentration on a logarithmic time scale. These relaxation times are obtained from the eigenvalues of the appropriate rate coefficient matrix ([chapter A3.4](#)). The time resolution of T -jump relaxation techniques is often limited by the rate at which the system can be heated. With typical T -jumps of several Kelvin, the time resolution lies in the microsecond range.

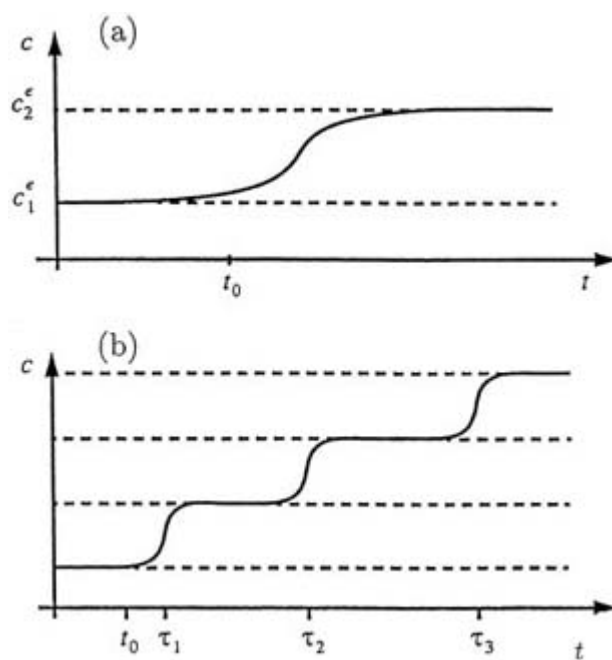


Figure B2.5.2. Schematic relaxation kinetics in a T -jump experiment. c measures the progress of the reaction, for example the concentration of a reaction product as a function of time t (abscissa with a logarithmic time scale). The reaction starts at t_0 . (a) Simple relaxation kinetics with a single relaxation time. (b) Complex reaction mechanism with several relaxation times τ_i . The different relaxation times τ_i are given by the turning points of c as a function of $\ln(t)$. Adapted from [110].

-7-

T -jump experiments are particularly well-suited to the study of the dissociation kinetics of weakly-bound

molecules or molecular complexes. Markwalder *et al* [49] used the laser-induced T -jump method to investigate the temperature and pressure dependence of NO_2 recombination kinetics:



With $\text{M} = \text{He}$, experiments were carried out between 255 K and 273 K with a few millibar NO_2 at total pressures between 300 mbar and 200 bar. Temperature jumps on the order of 1 K were effected by pulsed irradiation ($\ll 1 \mu\text{s}$) with a CO_2 laser at 9.2– 9.6 μm and with SiF_4 or perfluorocyclobutane as primary IR absorbers ($\ll 1$ mbar). Under these conditions, the dissociation of N_2O_4 occurs within the irradiated volume on a time scale of a few hundred microseconds. NO_2 and N_2O_4 were monitored simultaneously by recording the time-dependent UV absorption signal at 420 nm and 253 nm, respectively. The recombination rate constant k_{rec} can be obtained from the effective first-order relaxation time, τ_{R} . A derivation analogous to (equation (B2.5.9), equation (B2.5.10), equation (B2.5.11) and equation (B2.5.12)) yield

$$k_{\text{rec}} = \frac{\tau_{\text{R}}^{-1}}{K_{\text{c}} + 4[\text{NO}_2]} \quad (\text{B2.5.15})$$

where K_{c} is the equilibrium constant of equation (B2.5.14). k_{rec} , K_{c} , and $[\text{NO}_2]$ all refer to the final temperature. At 255 K, the authors obtained the typical fall-off curve depicted in figure B2.5.3. Even at 200 bar, the effective rate constant is still less than half the extrapolated high-pressure limit. The final results of the high- ($k_{\text{rec},\infty}$) and low-pressure ($k_{\text{rec},0}$) limiting rate constants (see chapter A3.4) were [49]

$$k_{\text{rec},\infty} = (2.2 \pm 0.2) \times 10^6 (T/\text{K})^{(2.3 \pm 0.2)} \text{ cm}^3 \text{ mol}^{-1} \text{ s}^{-1} \quad (\text{B2.5.16})$$

$$k_{\text{rec},0} = (7.5 \pm 0.8) \times 10^{35} (T/\text{K})^{(-9.0 \pm 0.9)} \times [\text{He}] \text{ cm}^6 \text{ mol}^{-2} \text{ s}^{-1} \quad (\text{B2.5.17})$$

where the temperature is given in Kelvin.

-8-

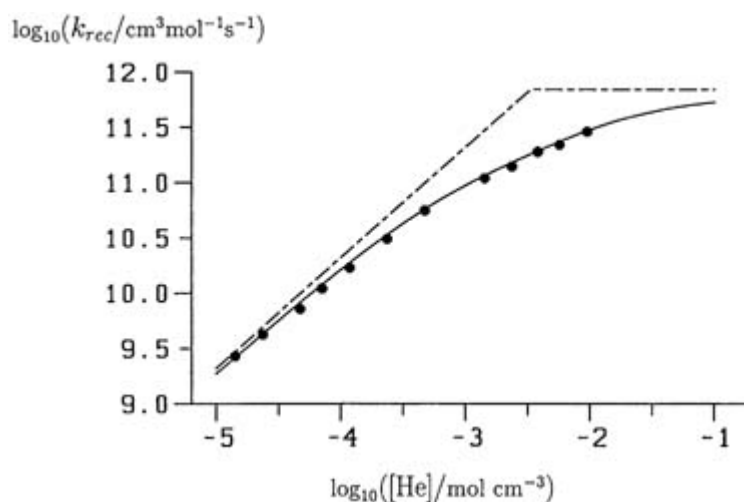


Figure B2.5.3. The fall-off curve of reaction (B2.5.14) with $\text{M} = \text{He}$ between 0.3 bar and 200 bar. The dashed lines represent the extrapolated low- and high-pressure limits. $k_{\text{rec},0} = (2.1 \pm 0.2) \times 10^{14} \times [\text{He}] \text{ cm}^6 \text{ mol}^{-2} \text{ s}^{-1}$ and $k_{\text{rec},\infty} = (7.0 \pm 0.7) \times 10^{11} \text{ cm}^3 \text{ mol}^{-1} \text{ s}^{-1}$ yield the best fit (full curve) to the experimental data (full circles). Adapted from [49].

B2.5.3.2 PERIODIC SMALL PERTURBATION FROM EQUILIBRIUM AND ULTRASOUND ABSORPTION

The previous subsection described single-experiment perturbations by T -jumps or P -jumps. By contrast, sound and ultrasound may be used to induce small periodic perturbations of an equilibrium system that are equivalent to periodic pressure and temperature changes. A temperature amplitude $\delta T \approx 0.002$ K and a pressure amplitude $\delta P \approx 30$ mbar are typical in experiments with high-frequency ultrasound. Figure B2.5.4 illustrates the situation for different rates of chemical relaxation with the angular frequency of the sound wave ω and the relaxation time τ_R :

$\omega\tau_R \ll 1$. The sample relaxes fast with the displacement from equilibrium synchronous with the sound wave.

$\omega\tau_R \approx 1$. Compared with the sound wave, the system relaxes slowly. It lags behind, the phase is shifted, and amplitudes are reduced by damping.

$\omega\tau_R \gg 1$. Very slow relaxation.

As an example for the mathematical treatment, we take the bimolecular reaction

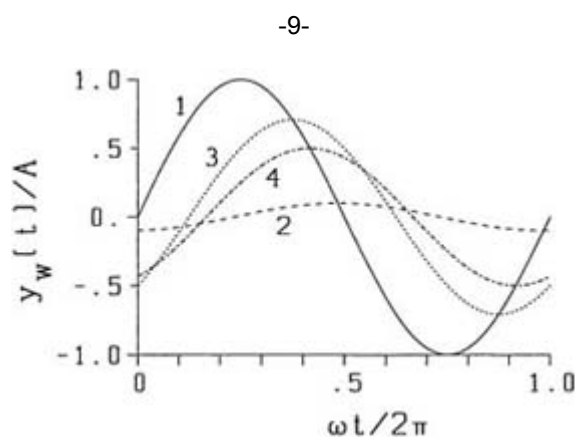


Figure B2.5.4. Periodic displacement from equilibrium through a sound wave. The full curve represents the temporal behaviour of pressure, temperature, and concentrations in the case of a very fast relaxation. The other lines illustrate various situations, with $\omega\tau_R$ according to table B2.5.1. ω is the angular frequency of the sound wave and τ_R is the chemical relaxation time. Adapted from [110].

Table B2.5.1 Form of the ‘chemical wave’ $y_\omega(t)$ (equation B2.5.24) for the various cases depicted in figure B2.5.4.

Case		Amplitude	Phase shift	$y_\omega(t)$
1	$\omega\tau_R \ll 1$	$\approx a$	≈ 0	$a \sin(\omega t)$
2	$\omega\tau_R = 10 \gg 1$		$\approx \pi/2$	

		$\approx a / (\omega\tau_R)$		$-[a/(\omega\tau_R)] \cos(\omega t)$
3	$\omega\tau_R = 1$	$a/\sqrt{2}$	$\pi/4$	$(a/\sqrt{2}) \sin(\omega t - \pi/4)$
4	$\omega\tau_R = \sqrt{3}$	$a/2$	$\pi/3$	$(a/2) \sin(\omega t - \pi/3)$

The turnover variable

$$x = c_A(t=0) - c_A = c_B(t=0) - c_B \quad (\text{B2.5.19})$$

-10-

obeys a first-order rate law near the equilibrium value, x_{eq} , which, in turn, depends on the temperature change ΔT induced by the sound wave:

$$\frac{dx}{dt} = k_{\text{eff}}(x_{\text{eq}}(T + \Delta T) - x). \quad (\text{B2.5.20})$$

For small ΔT , the temperature dependence of the effective first-order rate constant, k_{eff} , can be neglected. With $y = x_{\text{eq}}(T) - x$ and with the shift of the equilibrium, $\Delta x_{\text{eq}} = x_{\text{eq}}(T) - x_{\text{eq}}(T + \Delta T)$, one obtains

$$-\frac{d\Delta y}{dt} = (y - \Delta x_{\text{eq}})\{k_b + k_a(c_A^{\text{eq}} + c_B^{\text{eq}})\} = \frac{y - \Delta x_{\text{eq}}}{\tau_R}. \quad (\text{B2.5.21})$$

As long as ΔT , Δx_{e} , and Δx remain small, they will be proportional to the sinusoidal pressure wave. In particular

$$\Delta x_{\text{eq}} = a \sin(\omega t). \quad (\text{B2.5.22})$$

This leads to

$$y(t) + \tau_R \frac{dy(t)}{dt} = a \sin(\omega t) \quad (\text{B2.5.23})$$

with the general solution

$$y(t) = \left[y(0) + \frac{a\omega\tau_R}{1 + \omega^2\tau_R^2} \right] \exp\left(\frac{-t}{\tau_R}\right) + \frac{a}{1 + \omega^2\tau_R^2} \sin(\omega t) - \frac{a\omega\tau_R}{1 + \omega^2\tau_R^2} \cos(\omega t). \quad (\text{B2.5.24})$$

For sufficiently long times (index w), the exponential can be neglected, leaving an oscillation of the turnover variable phase shifted with respect to the sound wave and with its amplitude reduced by the finite relaxation time τ_R :

$$y_w(t) = \frac{a}{\sqrt{1 + \omega^2 \tau_R^2}} \sin(\omega t - \arctan(\omega \tau_R)). \quad (\text{B2.5.25})$$

The easily accessible frequency range of sound and ultrasound waves confines the range of applicability of this technique to relaxation times, τ_R , between 10^{-4} s and 10^{-9} s. The derivation given here is, of course, independent of the underlying chemical process, as long as it is characterized by a single relaxation time. In general a complex relaxation spectrum is possible, so that the method reaches its limit for complex reactions as the interpretation of the results may become ambiguous. Extensive descriptions of relaxation experiments with small perturbations—both single and periodic—can be found in [29, 50]. Here one can also find numerous relaxation-time expressions for various equilibrium systems (uni-, bi-, trimolecular and reverse).

-11-

B2.5.3.3 RELAXATION AFTER LARGE PERTURBATION: SHOCK-WAVE EXPERIMENTS

A general limitation of the relaxation techniques with small perturbations from equilibrium discussed in the previous section arises from the restriction to systems starting at or near equilibrium under the conditions used. This limitation is overcome by techniques with large perturbations. The most important representative of this class of relaxation techniques in gas-phase kinetics is the shock-tube method, which achieves T -jumps of some 1000 K (accompanied by corresponding P -jumps) [30, 51, 52 and 53]. Shock tubes are particularly useful for measuring the temperature dependence of reaction rates up to high temperatures. Figure B2.5.5 shows a schematic representation of the experimental set-up. The shock tube consists of a high- and a low-pressure part (R), separated by a diaphragm (d). The latter is filled with the reaction mixture and is operated either as a static cell or as a low-pressure flow tube. The high-pressure part is filled with a light, inert gas, such as H_2 or He, whose pressure is increased until the diaphragm breaks. This creates a shock wave travelling through the reaction mixture with supersonic speed. Behind the shock front the temperature can jump by more than 1000 K within $1\ \mu\text{s}$ or less. After passing the detection zone, where the relaxation is followed spectroscopically, the wave front is reflected back at the end of the tube. The resulting change of the temperature as a function of time is depicted in figure B2.5.6. The temperature T_2 after the wave front is determined indirectly from the speed, u , at which the wave front travels through the tube. For a highly-dilute reaction mixture in a monoatomic gas with atomic mass M one obtains

$$T_2 = T_1 \left(\frac{Mu^2}{5RT_1} - \frac{1}{3} \right) \left(\frac{Mu^2 + 1}{5RT_1} \right) \left(\frac{16Mu^2}{5RT_1} \right)^{-1} \quad (\text{B2.5.26})$$

where T_1 is the temperature before the wave front has passed.

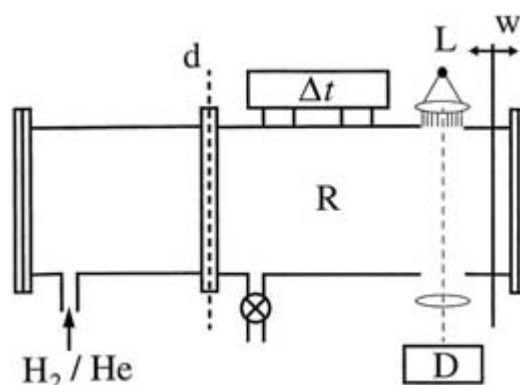


Figure B2.5.5. Schematic representation of a shock-tube apparatus. The diaphragm d separates the high-

pressure part from the low-pressure reaction chamber R. The speed of the shock wave is determined by the time Δt it takes to pass two observation points. The reaction itself is monitored spectroscopically using the light source L and a detector D close to the reflection wall W. Adapted from [110].

-12-

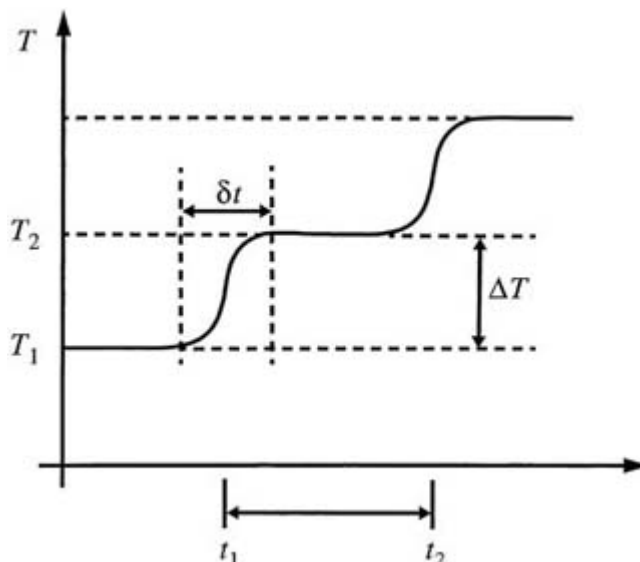


Figure B2.5.6. Temperature as a function of time in a shock-tube experiment. The first T -jump results from the incoming shock wave. The second is caused by the reflection of the shock wave at the wall of the tube. The rise time δt typically is less than $1 \mu\text{s}$, whereas the time delay between the incoming and reflected shock wave is on the order of several hundred microseconds. Adapted from [110].

A classic shock-tube study concerned the high-temperature recombination rate and equilibrium for methyl radical recombination [54, 55]. Methyl radicals were first produced in a fast decomposition of diazomethane at high temperatures ($T > 1000 \text{ K}$)



Subsequently, the recombination of methyl radicals was studied by the high-temperature UV absorption of the methyl radicals near 216 nm [54, 55].



-13-

Figure B2.5.7 shows the absorption traces of the methyl radical absorption as a function of time. At the time resolution considered, the appearance of CH_3 is practically instantaneous. Subsequently, CH_3 disappears by recombination (equation B2.5.28). At temperatures below 1500 K, the equilibrium concentration of CH_3 is negligible compared with C_2H_6 (left-hand trace): the recombination is complete. At temperatures above 1500 K (right-hand trace) the equilibrium concentration of CH_3 is appreciable, and thus the technique allows the determination of both the equilibrium constant and the recombination rate [54, 55]. This experiment resolved a famous controversy on the temperature dependence of the recombination rate of methyl radicals. While standard RRKM theories [57, 58] predicted an increase of the high-pressure recombination rate coefficient $k_{\text{rec}^\infty}(T)$ by a factor of 10–30 between 300 K and 1400 K, the statistical-adiabatic-channel model predicts a

slight decrease of $k_{\text{rec},\infty}(T)$ with increasing temperature [59, 60], in agreement with experiment [54, 55]. This temperature dependence of the high-pressure recombination rate coefficient for radical–radical association is now a generally accepted feature, frequently reconfirmed for other examples in this class of reaction. The secondary isotope effect for the recombination of CD_3 has also been studied for this reaction [54, 55].

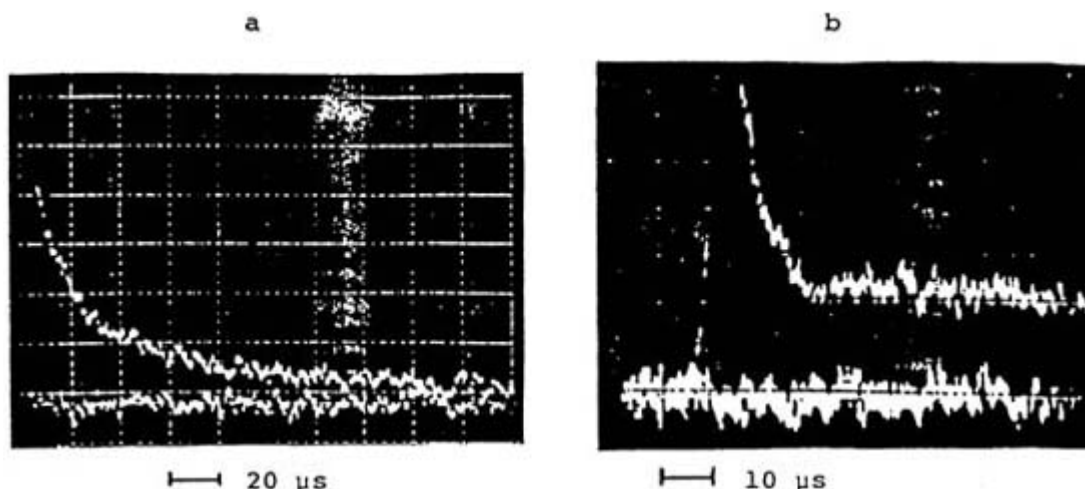


Figure B2.5.7. Oscilloscope trace of the UV absorption of methyl radical at 216 nm produced by decomposition of azomethane after a shock wave (after [54]): at (a) 1280 K and (b) 1575 K.

In a more recent example, a shock-tube experiment was used to study the thermal decomposition of methylamine between 1500 K and 2000 K [61, 62]:



$$k(T) = 8.17 \times 10^{16} \exp(-30710 \text{ K}/T) \text{ cm}^3 \text{ mol}^{-1} \text{ s}^{-1} (\pm 20\%). \quad (\text{B2.5.30})$$

The pyrolysis of CH_3NH_2 (<1 mbar) was performed at 1.3 atm in Ar, spectroscopically monitoring the concentration of NH_2 radicals behind the reflected shock wave as a function of time. The interesting aspect of this experiment was the combination of a shock-tube experiment with the particularly sensitive detection of the NH_2 radicals by frequency-modulated, laser-absorption spectroscopy [61]. Compared with ‘conventional’ narrow-bandwidth laser-absorption detection the signal-to-noise ratio could be increased by a factor of 20, with correspondingly more accurate values for the rate constant $k(T)$.

B2.5.4 FLASH PHOTOLYSIS WITH FLASH LAMPS AND LASERS

One of the most important techniques for the study of gas-phase reactions is flash photolysis [8, 63]. A reaction is initiated by absorption of an intense light pulse, originally generated from flash lamps (duration $\approx 1 \mu\text{s}$). Nowadays these have frequently been replaced by pulsed laser sources, with the shortest pulses of the order of a few femtoseconds [22, 64].

B2.5.4.1 FLASH PHOTOLYSIS WITH FLASH LAMPS

The absorption of a light pulse ‘instantaneously’ generates reactive species in high concentrations, either through the formation of excited species or through photodissociation of suitable precursors. The reaction can

then be followed spectroscopically by monitoring reactant and product concentrations. Among the classic studies using this technique, one may mention methyl radical spectroscopy used to study recombination at room temperature [56, 65, 66, 67, and 68].

A recent example of laser flash-lamp photolysis is given by Hippler *et al* [69], who investigated the temperature and pressure dependence of the thermal recombination rate constant k_{rec} for the reaction



The experiments were performed in a static reaction cell in a large excess of N_2 (2–200 bar). An UV laser pulse (193 nm, 20 ns) started the reaction by the photodissociation of N_2O to form O atoms in the presence of NO. The reaction was monitored via the NO_2 absorption at 405 nm using a Hg–Xe high-pressure arc lamp, together with direct time-dependent detection. With a 20–200-fold excess of NO, the formation of NO_2 followed a pseudo-first-order rate law:

$$[\text{NO}_2] = [\text{NO}_2]_{t=\infty} (1 - \exp\{-k_{\text{rec}}[\text{NO}]t\}). \quad (\text{B2.5.32})$$

Direct time-dependent detection is limited by the response time of detectors, which depends on the frequency range, and the electronics used for data acquisition. In the most favourable cases, modern detector/oscilloscope combinations achieve a time resolution of up to 100 ps, but 1 ns is more typical. Again, this reaction has been of fundamental theoretical interest for a long time [59, 60].

B2.5.4.2 LASER FLASH PHOTOLYSIS AND PUMP-PROBE TECHNIQUES

The so-called pump-probe technique uses a first photolysis pump pulse to generate reactive species and a second ‘probe’ pulse to detect reactant and product species. Figure B2.5.8 illustrates the experimental set-up. The time resolution is achieved through varying the delay between the pump pulse, which initiates the reaction, and the probe pulse, which monitors the reaction. Variable, short time delays in the picosecond range are conveniently realized through geometrical variation of the optical path length that the probe pulse travels. The probe pulse monitors reactant or product species either through direct absorption or by fluorescence excitation (laser-induced fluorescence, LIF), which generally is much more sensitive [70].

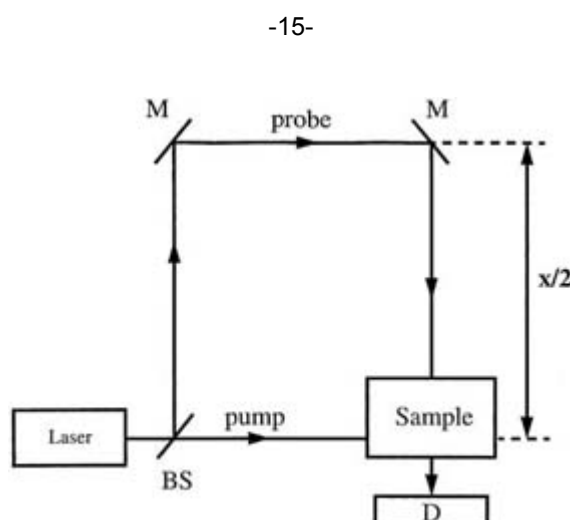


Figure B2.5.8. Schematic representation of laser-flash photolysis using the pump-probe technique. The beam splitter BS splits the pulse coming from the laser into a pump and a probe pulse. The pump pulse initiates a reaction in the sample, while the probe beam is diverted by several mirrors M through a variable delay line.

The detector D monitors the absorption of the probe beam as a function of the delay between the pulses given by $x/2c$, where c is the speed of light and x is the difference between the optical path travelled by the probe and by the pump pulse. Adapted from [110].

With the short pulses available from modern lasers, femtosecond time resolution has become possible [7, 71, 72 and 73]. Producing accurate time delays between pump and probe pulses on this time scale represents a major challenge for the experimentalist, since light only travels $0.3 \mu\text{m fs}^{-1}$. Table B2.5.2 summarizes typical laser pulses and characteristic times that are now available.

Table B2.5.2. Examples for pulsed lasers with different pulse durations and corresponding path lengths. For comparison the last column also gives the distance travelled by atoms with a velocity of 1000 ms^{-1} (in parentheses) [81].

Pulse duration	Laser	Availability	Optical path
100–200 ns	Atmospheric CO ₂ laser	Commercial	30–60 m (0.1–0.2 mm)
1–2 ns	Atmospheric CO ₂ laser, mode coupled with saturable absorber	Available	30–60 cm (1–2 μm)
100 fs–1 ps	Solid-state laser (e.g. Ti:sapphire), dye laser	Commercial	0.03–0.3 mm (100 pm–1 nm)
8 fs	Laser with subsequent pulse compression	World record [108]	2.4 μm (8.4 pm)
6.5 fs	Ti:sapphire, mode coupling, with saturable absorber (semiconductor)	World record [64]	2 μm (6.5 pm)

One of the early examples for kinetic studies on the femtosecond time scale is the photochemical predissociation of NaI [74]:

(B2.5.33)

The experiment is illustrated in figure B2.5.9. The initial pump pulse generates a localized wavepacket in the first excited S_1 state of NaI, which evolves with time. The potential well in the S_1 state is the result of an avoided crossing with the ground state. Every time the wavepacket passes this region, part of it crosses to the lower surface before the remainder is reflected at the outer wall of the S_1 potential. The crossing leads to ground-state dissociation products Na ($^2S_{1/2}$) + ($^2P_{3/2}$). The crossing is monitored with the time-delayed probe pulse, which excites ground-state Na ($^2P^0 \leftarrow ^2S_{1/2}$). A photomultiplier detects the fluorescence back to the ground state. Figure B2.5.10 shows the resulting LIF signal as a function of the probe delay. One clearly recognizes the signature of the oscillatory wavepacket motion with a relatively long oscillation period of 1 ps, resulting from the flat potential and the heavy masses.

Figure B2.5.9. Schematic representation of the potential curves for the photodissociation of NaI as a function of the interatomic distance R . L1 and L2 are the pump and probe laser pulses, respectively. The dissociation limits are (I) $\text{Na}(^2\text{S}_{1/2}) + \text{I}(^2\text{P}_{3/2})$, (II) $\text{Na}(^2\text{S}_{1/2}) + \text{I}(^2\text{P}_{1/2})$, (III) $\text{Na}^+(^1\text{S}_0) + \text{I}^-(^1\text{S}_0)$, and (IV) $\text{Na}(^2\text{P}) + \text{I}(^2\text{P}_{3/2})$.

E is the excitation energy relative to the lowest dissociation limit (I). Adapted from Rosker *et al* [74].

-17-

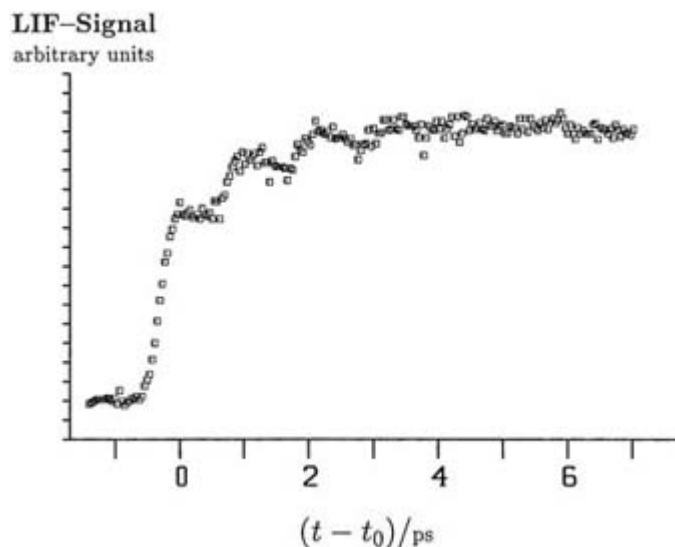


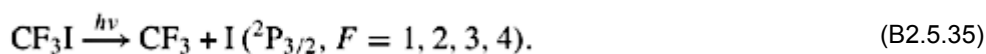
Figure B2.5.10. LIF signal of free Na atoms produced in the photodissociation of NaI. $t - t_0$ is the delay between the photolysis pulse (at t_0) and the probe pulse. Adapted from [111].

B2.5.4.3 THE PRINCIPLE OF CONTINUOUS DETECTION WITH UNCERTAINTY-LIMITED TIME AND FREQUENCY RESOLUTION

In this approach one uses narrow-band continuous wave (cw) lasers for continuous spectroscopic detection of reactant and product species with high time and frequency resolution. Figure B2.5.11 shows an experimental scheme using detection lasers with a 1 MHz bandwidth. Thus, one can measure the energy spectrum of reaction products with very high energy resolution. In practice, today one can achieve an uncertainty-limited resolution given by

$$\Delta\nu\Delta t \geq \frac{1}{4\pi}. \quad (\text{B2.5.34})$$

This technique with very high frequency resolution was used to study the population of different hyperfine structure levels of the iodine atom produced by the IR-laser-flash photolysis of organic iodides through multiphoton excitation:



-18-

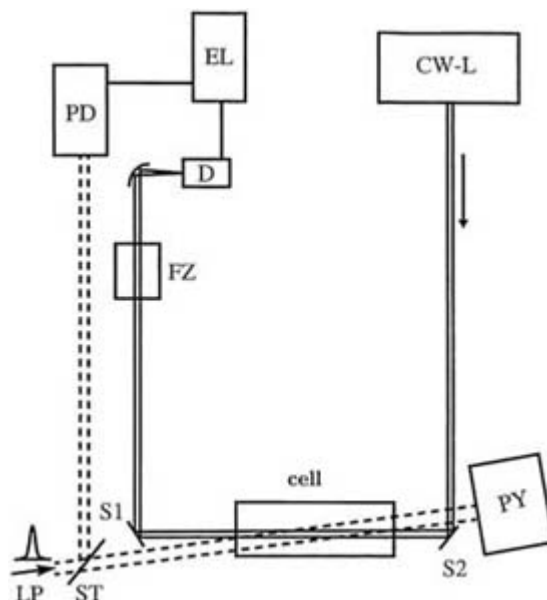


Figure B2.5.11. Schematic set-up of laser-flash photolysis for detecting reaction products with uncertainty-limited energy and time resolution. The excitation CO_2 laser pulse LP (broken line) enters the cell from the left, the tunable cw laser beam CW-L (full line) from the right. A filter cell FZ protects the detector D, which determines the time-dependent absorbance, from scattered CO_2 laser light. The pyroelectric detector PY measures the energy of the CO_2 laser pulse and the photon drag detector PD its temporal profile. A complete description can be found in [109].

Figure B2.5.12 shows the energy-level scheme of the fine structure and hyperfine structure levels of iodine. The corresponding absorption spectrum shows six sharp hyperfine structure transitions. The experimental resolution is sufficient to determine the Doppler line shape associated with the velocity distribution of the I atoms produced in the reaction. In this way, one can determine either the temperature in an oven—as shown in Figure B2.5.12—or the primary translational energy distribution of I atoms produced in photolysis, equation B2.5.35.

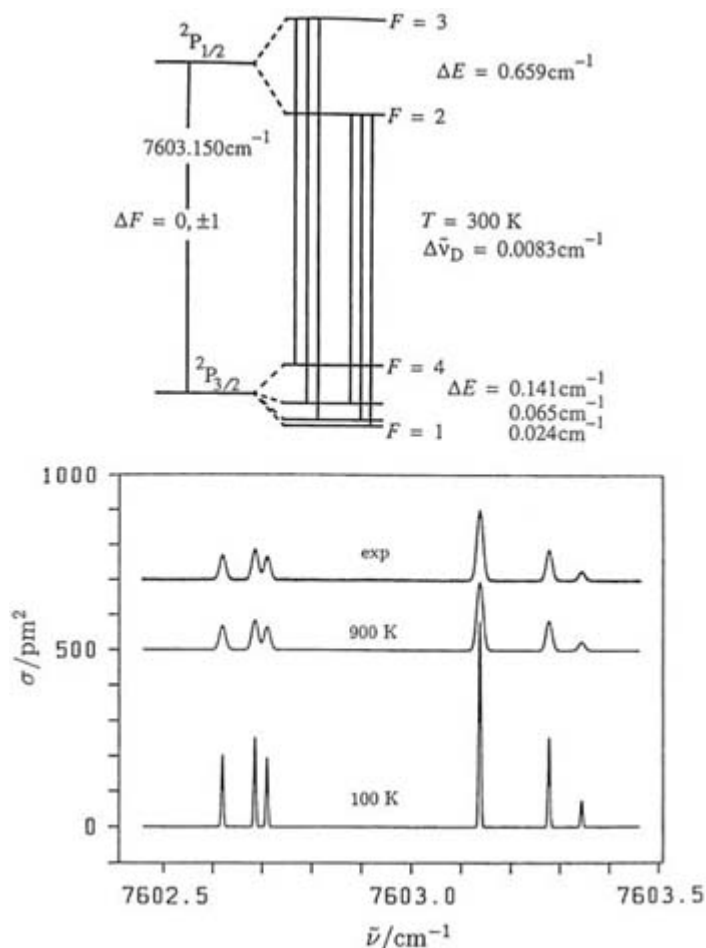


Figure B2.5.12. Hyperfine structure energy level scheme and spectrum for the $I(^2P_{3/2}) \leftrightarrow I(^2P_{1/2})$ fine structure transition [109].

B2.5.5 MULTIPHOTON EXCITATION

B2.5.5.1 MECHANISMS OF MULTIPHOTON EXCITATION

The common flash-lamp photolysis and often also laser-flash photolysis are based on photochemical processes that are initiated by the absorption of a photon, $h\nu$. The intensity of laser pulses can reach GW cm^{-2} or even TW cm^{-2} , where multiphoton processes become important. Figure B2.5.13 summarizes the different mechanisms of multiphoton excitation [75, 76, 112]. The direct multiphoton absorption of mechanism (i) requires an odd number of photons to reach an excited atomic or molecular level in the case of strict electric dipole and parity selection rules [117].

The Goeppert–Mayer two- (or multi-) photon absorption, mechanism (ii), may look similar, but it involves intermediate levels far from resonance with one-photon absorption. A third, quasi-resonant stepwise mechanism (iii), proceeds via single-photon excitation steps involving near-resonant intermediate levels. Finally, in mechanism (iv), there is the stepwise multiphoton absorption of incoherent radiation from thermal light sources or broad-band statistical multimode lasers. In principle, all of these processes and their combinations play a role in the multiphoton excitation of atoms and molecules, but one can broadly

distinguish two situations.

- (A) During the multiphoton excitation of molecular vibrations with IR lasers, many (typically 10–50) photons are absorbed in a quasi-resonant stepwise process until the absorbed energy is sufficient to initiate a unimolecular reaction, dissociation, or isomerization, usually in the electronic ground state. The record in the number of absorbed photons (about 500 photons of a CO₂ laser) was reached with the C₆₀ molecule [77]. This case proved an exception in that the primary reaction was ionization. The IR multiphoton excitation is the starting point for a new gas-phase photochemistry, IR laser chemistry, which encompasses numerous chemical processes.
- (B) The multiphoton excitation of electronic levels of atoms and molecules with visible or UV radiation generally leads to ionization. The mechanism is generally a combination of direct, Goeppert–Mayer, and quasi-resonant stepwise processes. Since ionization often requires only two or three photons, this type of multiphoton excitation is used for spectroscopic purposes in combination with mass-spectrometric detection of ions.

B2.5.5.2 IR MULTIPHOTON EXCITATION AND IR LASER CHEMISTRY

The most commonly used laser-light source in IR laser chemistry is the atmospheric CO₂ laser, with IR emission lines between 900 cm⁻¹ and 1100 cm⁻¹, in the fingerprint range of the IR spectrum, where characteristic molecular vibrations can be excited. With a photon energy of about 12 kJ mol⁻¹, on the order of 10–40 photons are needed to initiate a chemical reaction in the energy range of 100–500 kJ mol⁻¹. The laser pulses are 100 ns– μs long, but a series of 1–2 ns pulses can be generated by mode coupling. Typical intensities are 100 MW cm⁻². [Figure B2.5.14](#) schematically illustrates the photodissociation of CF₃I (equation [B2.5.35](#)) after multiphoton excitation via the CF stretching vibration at 1070 cm⁻¹. More than 17 photons are needed to break the C–I bond, a typical value in IR laser chemistry. Contributions from direct absorption (i) are insignificant, so that the process almost exclusively follows the quasi-resonant mechanism (iii), which can be treated by generalized first-order kinetics. As an example, [figure B2.5.15](#) illustrates the formation of I atoms (upper trace) during excitation with the pulse sequence of a mode-coupled CO₂ laser (lower trace). In addition to the intensity, *I*, the fluence, *F*, of radiation is a very important parameter in IR laser chemistry (and more generally in multiphoton excitation):

$$F(t) = \int_0^t I(t') dt'. \quad (\text{B2.5.36})$$

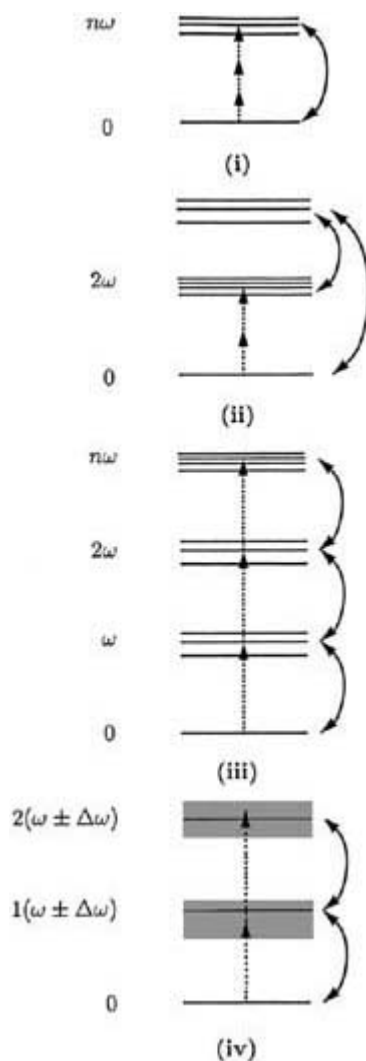


Figure B2.5.13. Schematic representation of the four different mechanisms of multiphoton excitation: (i) direct, (ii) Goeppert–Mayer (iii) quasi-resonant stepwise and (iv) incoherent stepwise. Full lines (right) represent the coupling path between the energy levels and broken arrows the photon energies with angular frequency ω ($\Delta\omega$ is the frequency width of the excitation light in the case of incoherent excitation), see also [112].

Consequently, the reaction yield F_p in [figure B2.5.15](#) is shown as a function of the fluence, F . At the end of a laser-pulse sequence with a typical fluence $F \simeq 3 \text{ J cm}^{-2}$, practically 100% of the CF_3I is photolysed. As described in [section B2.5.4.3](#), the product-level distribution of the iodine atoms formed in this type of reaction can be determined

spectroscopically. Table B2.5.3 shows the results of such an analysis of the population of hyperfine structure levels and of the translational energy distribution for the IR multiphoton dissociation of different organic iodides. The average product translational energy (in the centre-of-mass system) does not change much from the small CF_3I to the much larger $\text{C}_6\text{F}_5\text{I}$ molecule. Its relative share of the total energy, however, decreases: much more energy appears as the internal energy of the C_6F_5 fragment. This can be readily understood assuming a roughly statistical distribution over the large number of internal degrees of freedom. Such results are crucial for a more accurate dynamical understanding of the processes taking place during a chemical reaction.

Table B2.5.3. Product energy distribution for some IR laser chemical reactions. $\langle E_t \rangle$ is the average relative translational energy of fragments, $\langle E_{\text{int}} \rangle$ is the average vibrational and rotational energy of polyatomic fragments, and f_t is the fraction of the total product energy appearing as translational energy [109].

Reaction	$\langle E_t \rangle / (\text{kJ mol}^{-1})$	$\langle E_{\text{int}} \rangle / (\text{kJ mol}^{-1})$	f_t
$\text{CF}_3\text{I} \rightarrow \text{CF}_3 + \text{I}$	9.9	19.8	0.33
$\text{CF}_3\text{CHF}\text{I} \rightarrow \text{CF}_3\text{CHF} + \text{I}$	10.9	100.9	0.097
$\text{C}_6\text{F}_5\text{I} \rightarrow \text{C}_6\text{F}_5 + \text{I}$	13.5	233.0	0.055

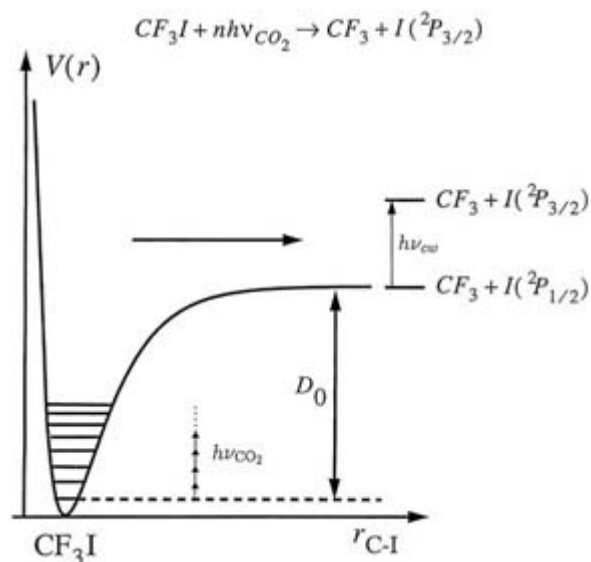


Figure B2.5.14. The IR laser chemistry of CF_3I excited up to the dissociation energy D_0 with about 17 quanta of a CO_2 laser, $h\nu_{\text{CO}_2}$. The dissociation is detected by uncertainty limited cw absorption ($h\nu_{\text{cw}}$), see figures B2.5.11 and B2.5.12. The energy levels of the C–I stretching vibration are not drawn to scale. In reality their separation is much smaller. Adapted from [109].

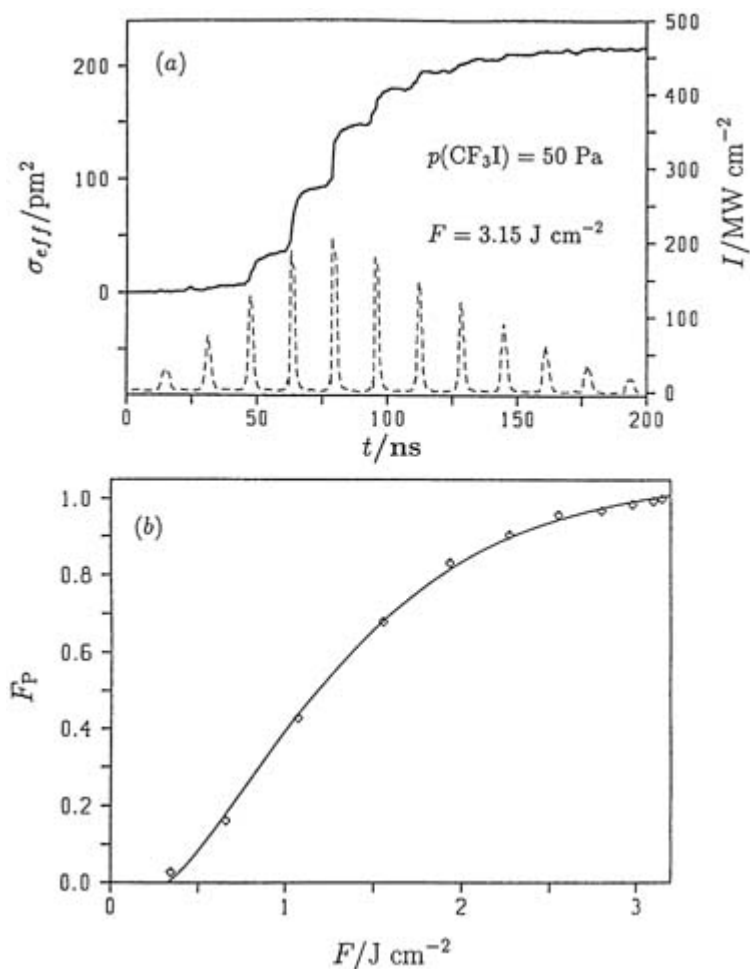


Figure B2.5.15. Iodine atom formation in the IR laser chemistry of CF_3I (excitation at 1074.65 cm^{-1} , probe on the $F = 4 \rightarrow F = 3$ hyperfine structure transition, see figure B2.5.12.) (a) The absorbance as a function of time (effective absorption cross section σ_{eff} full curve, left ordinate) shows clear steps at each maximum of the mode locked CO_2 laser pulse sequence (intensity, broken curve, right ordinate). (b) The fraction F_p of dissociating molecules as a function of fluence F .

In exceptional cases, the IR laser excitation can lead to ionization. An interesting example is the CO_2 -laser-induced ionization of C_{60} , where $n \geq 500$ photons are absorbed and vibrations are excited far beyond the ionization threshold of the molecule



The C_{60}^+ is excited further and decomposes stepwise into C_{58}^+ , C_{56}^+ , etc with the formation of C_2 units [77].

In contrast to the ionization of C_{60} after vibrational excitation, typical multiphoton ionization proceeds via the excitation of higher electronic levels. In principle, multiphoton ionization can either be used to generate ions and to study their reactions, or as a sensitive detection technique for atoms, molecules, and radicals in reaction kinetics. The second application is more common. In most cases of excitation with visible or UV laser radiation, a few photons are enough to reach or exceed the ionization limit. A particularly important technique is resonantly enhanced multiphoton ionization (REMPI), which exploits the resonance of monochromatic laser radiation with one or several intermediate levels (in one-photon or in multiphoton processes). The mechanisms are distinguished according to the number of photons leading to the resonant intermediate levels and to the final level, as illustrated in figure B2.5.16. Several lasers of different frequencies may be combined.

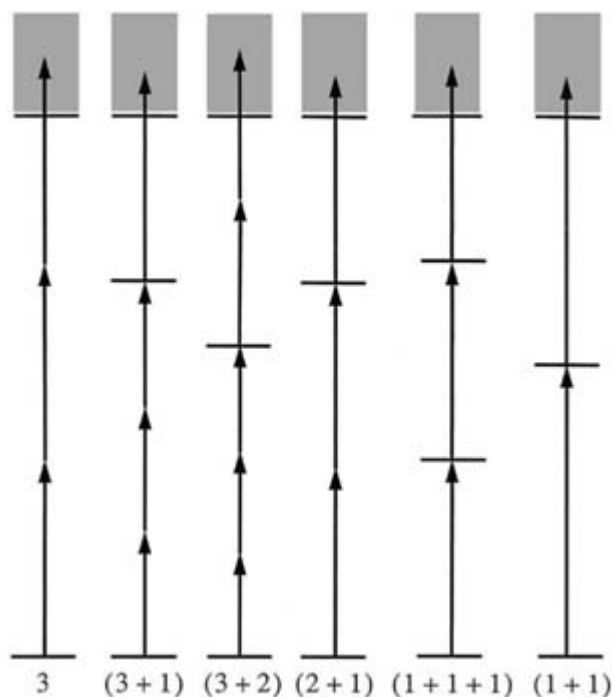


Figure B2.5.16. Different multiphoton ionization schemes. Each scheme is classified according to the number of photons that lead to resonant intermediate levels and to the ionization continuum (hatched area). Adapted from [110].

As an example, we mention the detection of iodine atoms in their $^2P_{3/2}$ ground state with a 3 + 2 multiphoton ionization process at a laser wavelength of 474.3 nm. Excited iodine atoms ($^2P_{1/2}$) can also be detected selectively as the resonance condition is reached at a different laser wavelength of 477.7 nm. As an example, [figure B2.5.17](#) shows REMPI iodine atom detection after IR laser photolysis of CF_3I . This ‘pump–probe’ experiment involves two, delayed, laser pulses, with a 200 ns IR photolysis pulse and a 10 ns probe pulse, which detects iodine atoms at different times during and after the photolysis pulse. This experiment illustrates a fundamental problem of product detection by multiphoton ionization: with its high intensity, the short-wavelength probe laser radiation alone can photolyse the

reactant CF_3I molecules. One cannot distinguish between iodine atoms produced by the photolysis pulse and those produced by the probe pulse. In the present example the problem is solved by the well-founded assumption that the photolysis of CF_3I by a visible probe pulse produces excited iodine atoms ($^2P_{1/2}$), whereas the IR photolysis pulse leads to ground-state iodine atoms ($^2P_{3/2}$). In general, however, significant perturbations of the reaction system are to be expected from the REMPI spectroscopic detection of products.

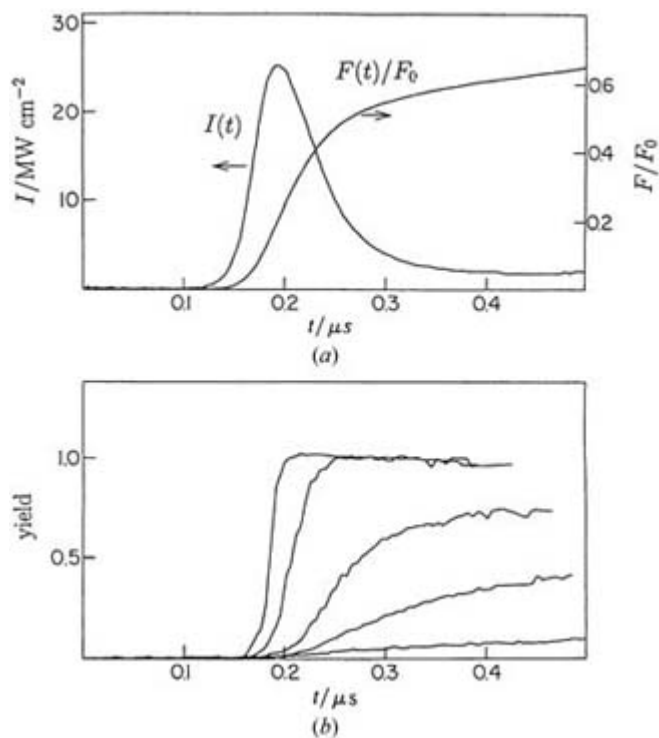


Figure B2.5.17. (a) Time-dependent intensity I and reduced fluence F/F_0 for a single-mode CO_2 laser pulse used in the IR laser photolysis of CF_3I . F_0 is the total fluence of the laser pulse. (b) VIS-REMPI iodine atom signals obtained with CO_2 laser pulses of different fluence (after [113]).

B2.5.5.4 LASER ISOTOPE SEPARATION AND MODE-SELECTIVE REACTIONS

Apart from the obvious property of defining pulses within short time intervals, the pulsed laser radiation used in reaction kinetics studies can have additional particular properties: (i) high intensity, (ii) high monochromaticity, and (iii) coherence. Depending on the type of laser, these properties may be more or less pronounced. For instance, the pulsed CO_2 lasers used in IR laser chemistry easily reach intensities between MW cm^{-2} and GW cm^{-2} . Special lasers used in nuclear fusion experiments may even reach $10^{21} \text{ W cm}^{-2}$ [78, 79]. Ideally the monochromaticity, $\Delta\nu$, is related to the pulse length, Δt , through

$$\Delta\nu\Delta t \simeq 1. \quad (\text{B2.5.39})$$

Although this limit is not always reached. The same is true for the coherence of the radiation. Each of these properties can be exploited for particular chemical applications. The monochromaticity can be used to initiate a chemical reaction of particular molecules in a mixture. The laser isotope separation of ^{12}C and ^{13}C in natural abundance exploits the isotope shift of molecular vibrational frequencies. At $10\text{--}50 \text{ cm}^{-1}$, the corresponding shift of IR absorption wavenumbers is large compared to the spectral width of the CO_2 laser pulse ($\leq 0.1 \text{ cm}^{-1}$), which makes the ^{13}C isotope separation relatively easy. Table B2.5.4 summarizes this and other similar applications [75, 80, 81]. The intermolecular selectivity of IR-multiphoton excitation can be greatly increased by two-frequency-two-step schemes such as in the new spectroscopic technique of IRLAPS (InfraRed Laser Assisted Photofragment Spectroscopy [115]).

Table B2.5.4 Laser isotope separation (see also [75]).

Isotope Source Comments

^2H	CHF_2Cl	High selectivity at room temperature
^{10}B	BCl_3	Early laser isotope separation after IR multiphoton excitation high selectivity at room temperature
^{13}C	CHF_2Cl	Two-step separation scheme ($220 \text{ mg } ^{13}\text{C h}^{-1}$)
$^{14}\text{N}, ^{15}\text{N}$	CH_3NO_2	Selectivity through two absorption bands
$^{16}\text{O}, ^{17}\text{O}$	OCS	IR–UV double resonance; also selective for S and C
$^{29}\text{Si}, ^{30}\text{Si}$	Si_2F_6	Reaction of both isotopes with high selectivity (high fluence)
^{34}S	SF_6	Early report of laser isotope separation
$^{35}\text{Cl}, ^{37}\text{Cl}$	CF_2Cl_2	Also selective with respect to C
Mo	MoF_6	Applied to several isotopes; low selectivity and yield
^{235}U	UF_6	Dissociation with two lasers at different wavelengths (two-colour dissociation)

[Figure B2.5.18](#) compares this *inter* molecular selectivity with *intra* molecular or mode selectivity. In an IR plus UV, two-photon process, it is possible to break either of the two bonds selectively in the same HOD molecule. Depending on whether the OH or the OD stretching vibration is excited, the products are either H + OD or HO + D [24]. In large molecules, *intramolecular* selectivity competes with fast *intramolecular* (i.e. unimolecular) vibrational energy redistribution (IVR) processes, which destroy the selectivity. In laser experiments with D-difluorobutane [82], it was estimated that, in spite of frequency selective excitation of the CHDF end group, no selective reaction would occur on time scales above 10^{-11} s, [figure B2.5.18](#). In contrast to IVR processes, which can be very fast, the *intermolecular* energy transfer processes, which may reduce intermolecular selectivity, are generally much slower, since they proceed via bimolecular energy exchange, which is limited by the collision frequency (see [chapter A3.13](#)).

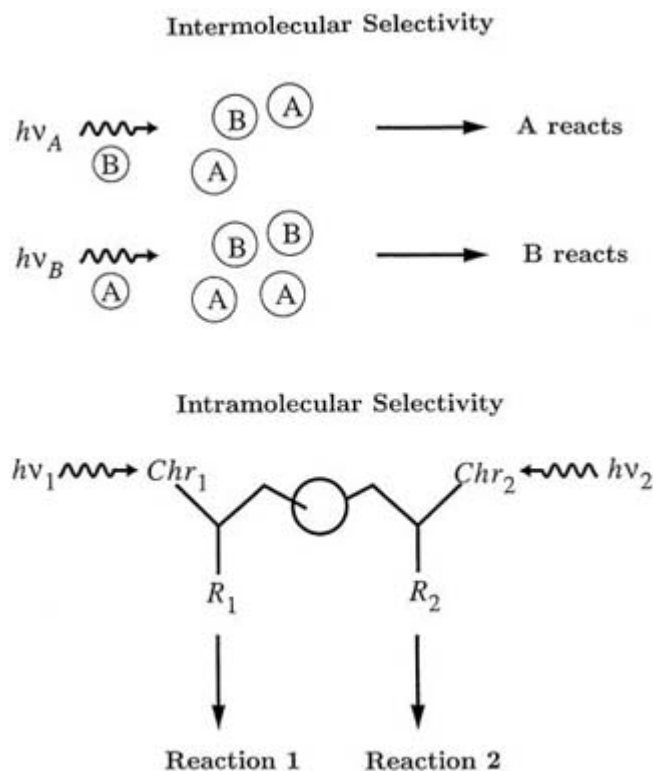


Figure B2.5.18. General scheme for *inter*- and *intramolecular* selectivity in laser chemistry. Intermolecular selectivity: a laser with frequency ν_A selectively excites molecules A, which subsequently react, in a mixture of A and B molecules. *Intramolecular* selectivity: a laser with frequency ν_1 (ν_2) selectively excites the chromophore Chr_1 (Chr_2) of a molecule which preferentially follows reaction 1 (2) at this position (after [75]).

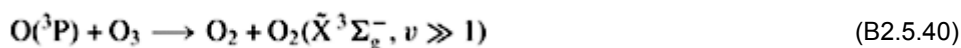
Strategies for achieving intra- and intermolecular selectivity are the subject of a very active field of current research with many open questions. Under the label ‘coherent control’ it includes approaches that exploit the coherence properties of laser radiation to control chemical reactions. Figure B2.5.18 summarizes the different schemes of intra- and intermolecular selectivity.

B2.5.6 CHEMICAL ACTIVATION

The formation of reactive species by photodissociation of a precursor through flash photolysis can be regarded as a special case of chemical activation. More generally, this technique exploits the enthalpy of a chemical reaction to generate species with a non-equilibrium energy distribution (relative to the ambient temperature). Using different reactions to produce the same reactive species allows one to study the energy dependence of the ensuing reaction kinetics (or collisional deactivation). Historically, the method has played a central role in the experimental study of collisional energy-transfer processes and non-equilibrium effects on chemical reaction rates [83, 84, 85, 86 and 87].

Although modern laser techniques can in principle achieve much narrower energy distributions, optical excitation is frequently not a viable method for the preparation of excited reactive species. Therefore chemical activation—often combined with (laser-) flash photolysis—still plays an important role in gas-phase kinetics, in particular of unstable species such as radicals [88]. Chemical activation also plays an important role in energy-transfer studies (see [chapter A3.13](#)).

A recent study of the vibrational-to-vibrational (V–V) energy transfer between highly-excited oxygen molecules and ozone combines laser-flash photolysis and chemical activation with detection by time-resolved LIF [89]. Partial laser-flash photolysis at 532 nm of pure ozone in the Chappuis band produces translationally-hot oxygen atoms $O(^3P)$. In the chemical-activation step they react with ozone to form an electronic ground-state $O_2(\tilde{X}^3\Sigma_g^-, v'')$ with up to $v'' = 27$ quanta of vibrational excitation in an excess of thermally populated ozone:



$$k = 3.9 \times 10^{-11} \text{ cm}^3 \text{ s}^{-1}. \quad (\text{B2.5.41})$$

The chemical-activation step is between one and two orders of magnitude faster than the subsequent collisional deactivation of vibrationally excited O_2 . Finally, the population of individual vibrational levels v'' of O_2 is probed through LIF in the Schumann–Runge band ($B^3\Sigma_u^- \leftarrow X^3\Sigma_g^-$): after exciting the oxygen molecules to the vibrational ground state of their first electronically excited state ($v' = 0$), the ensuing fluorescence back to the electronic ground state is detected by a photomultiplier tube and recorded as a function of time. The resulting collisional relaxation rate constants as a function of the vibrational excitation of O_2

$$\frac{d[O_2(v'')]}{dt} = k_{v''}(O_3)[O_2(v'')][O_3] \quad (\text{B2.5.42})$$

show a pronounced maximum near $v'' = 23$, as illustrated in [figure B2.5.19](#). At this value, the $v'' \rightarrow v'' - 1$ transition happens to be in almost perfect resonance with the symmetric stretch fundamental of O_3 . The resonance enhancement by one to two orders of magnitude is typical for collisional V–V energy transfer of highly-excited molecules.

-29-

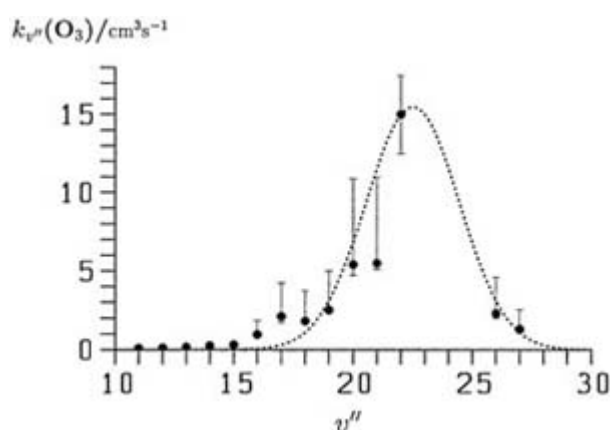


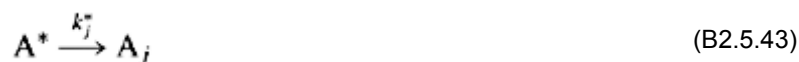
Figure B2.5.19. The collisional deactivation rate constant $k_{v''}(O_3)$ (equation B2.5.42) as a function of the vibrational level v'' . Adapted from [89]. Experimental data are represented by full circles with error bars. The broken curve is to serve as a guide to the eye.

B2.5.7 LINE-SHAPE METHODS

Energy (or frequency) spectra are fundamentally related to the underlying time-dependent processes through

the Fourier transformation. In practice, however, the relation between spectroscopically-observed line shapes and kinetic (reaction) processes is neither simple nor unambiguous [18]. There are many contributions to observed line shapes [33]. Apart from finite instrumental resolution, spectra may be inhomogeneously broadened through thermal congestion. A simple example is the Doppler broadening as a result of the Maxwell–Boltzmann velocity distribution leading to a Gaussian line shape.

Even if the homogeneous line shape can be extracted, many other processes can contribute. Every decay process contributes to the finite lifetime of an excited species, A^* , with an individual decay constant k_j^*



$$-\frac{d[A^*]}{dt} = \sum_j k_j^* [A^*] \quad (\text{B2.5.44})$$

$$k_{\text{eff}} = \sum_j k_j^*. \quad (\text{B2.5.45})$$

-30-

The exponential decay of the A^* population corresponds to a Lorentzian line shape for the absorption (or emission) cross section, σ , as a function of energy E . The lineshape is centred around its maximum at E_0 . The full-width at half-maximum (Γ) is proportional to k_{eff} :

$$\sigma(E) = \sigma(E_0) \frac{(\Gamma/2)^2}{(E - E_0)^2 + (\Gamma/2)^2} \quad (\text{B2.5.46})$$

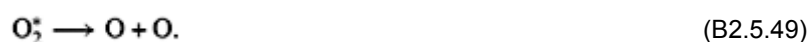
$$\Gamma = k_{\text{eff}} h / (2\pi). \quad (\text{B2.5.47})$$

Apart from the *natural lifetime* due to spontaneous emission, both uni- and bimolecular processes can contribute to the observed value of Γ . One important contribution k_{col} comes from *collisional broadening*, which can be distinguished by its pressure dependence (or dependence upon concentration $[M]$ of the collision partner):

$$k_{\text{col}} = \left(\frac{8k_B T}{\pi \mu} \right)^{1/2} \langle \sigma_{\text{col}} \rangle [M]. \quad (\text{B2.5.48})$$

Equation B2.5.48 introduces the effective average collision cross section $\langle \sigma_{\text{col}} \rangle$. Here, the lifetime broadening results from the (collisional) perturbation of A^* by collisions with M .

Lifetimes of 1 ps translate into linewidths of about 5 cm^{-1} . Thus, line-shape methods are ideally suited to measure very fast decay processes, in particular predissociation of excited species. An example is the predissociation of O_2 molecules excited above $50\,000 \text{ cm}^{-1}$, which gives rise to the broadening of the Schumann–Runge bands



This is the source of ozone, through the reaction $O_2 + O \xrightarrow{[M]} O_3$. One obtains a pronounced dependence of the

decay rate on the vibrational level of O_2^* and to a lesser extent on its rotational state [90, 91]. Typical decay rate constants for this reaction range from $1.5 \times 10^{11} \text{ s}^{-1}$ to $7.5 \times 10^{11} \text{ s}^{-1}$. Another important example is the predissociation of methyl radicals [54]



The results are summarized in [table B2.5.5](#). The rate constants k_j of individual decay channels may be obtained from the relative yields of all primary reaction products, which can be determined in stationary experiments.

-31-

Table B2.5.5. The photochemical decomposition of methyl radicals (UV excitation at 216 nm). $\bar{\Gamma}$ is the wavenumber linewidth of the methyl radical absorption and k is the effective first-order decay constant [54].

Decay process	$\bar{\Gamma}(\text{cm}^{-1})$	$\Delta\nu = hc\bar{\Gamma} (\text{s}^{-1})$	$k = 2\pi\Delta\nu (\text{s}^{-1})$	$\tau = 1/k(\text{fs})$
$\text{CH}_3^* \rightarrow \text{CH}_2 + \text{H}$	60	1.8×10^{12}	1.13×10^{13}	88
$\text{CD}_3^* \rightarrow \text{CD}_2 + \text{D}$	8	2.4×10^{11}	1.51×10^{12}	663

Similar considerations have been exploited for the systematic analysis of room-temperature and molecular-beam IR spectra in terms of intramolecular vibrational relaxation rates [33, 34, 92, 94] (see also [chapter A3.13](#)).

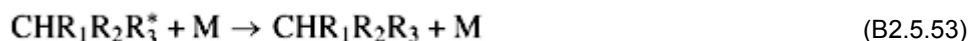
B2.5.8 INTRAMOLECULAR KINETICS FROM HIGH-RESOLUTION SPECTROSCOPY

Molecular spectroscopy offers a fundamental approach to intramolecular processes [18, 94]. The spectral analysis in terms of detailed quantum mechanical models in principle provides the complete information about the wave-packet dynamics on a level of detail not easily accessible by time-resolved techniques.

The approach is ideally suited to the study of IVR on fast timescales, which is the most important primary process in unimolecular reactions. The application of high-resolution rovibrational overtone spectroscopy to this problem has been extensively demonstrated. Effective Hamiltonian analyses alone are insufficient, as has been demonstrated by explicit quantum dynamical models based on *ab initio* theory [95]. The fast IVR characteristic of the CH chromophore in various molecular environments is probably the most comprehensively studied example of the kind [96] (see [chapter A3.13](#)). The importance of this question to chemical kinetics can perhaps best be illustrated with the following examples. The atom recombination reaction



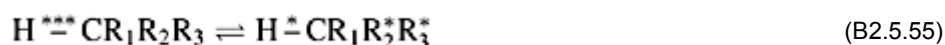
is well known to occur as a very slow trimolecular process. By contrast, the polyatomic recombination



happens quickly as a sequence of bimolecular recombination and collisional energy-transfer steps, with a relatively long-lived intermediate, $\text{CHR}_1\text{R}_2\text{R}_3^*$. The reason is the possibility of transferring energy intramolecularly from the

-32-

initially excited C–H bond to other parts of the polyatomic molecule, according to the scheme



This illustrates the steps of energy transfer from the initially highly-excited C–H bond to other parts of the molecule, subsequent concentration of energy in one part of the molecule (CR_3^{**}), and finally rupture of the corresponding bond. A typical example of this kind is the chemical activation reaction (abbreviated)



It is the first IVR step of B2.5.59 that is investigated by high-resolution spectroscopy. The analysis, outlined in some detail in [18], follows the scheme in figure B2.5.20. This kind of analysis has been applied to the evolution of entropy in the single, isolated molecule CHD_2F , as shown in figure B2.5.21. In this case, entropy is investigated as a relevant time-dependent observable of kinetics (see chapter A3.4). In the example, the question of time-reversal symmetry on the femtosecond timescale has been studied [18, 114, 116], but many other applications can be thought of.

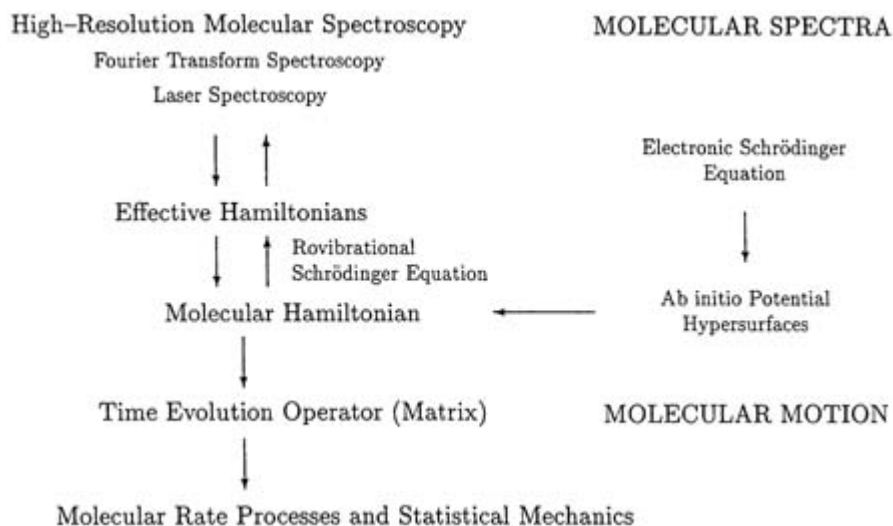


Figure B2.5.20. The combined experimental and theoretical approach ‘Molecular spectra and motion’ (after [18]).

-33-

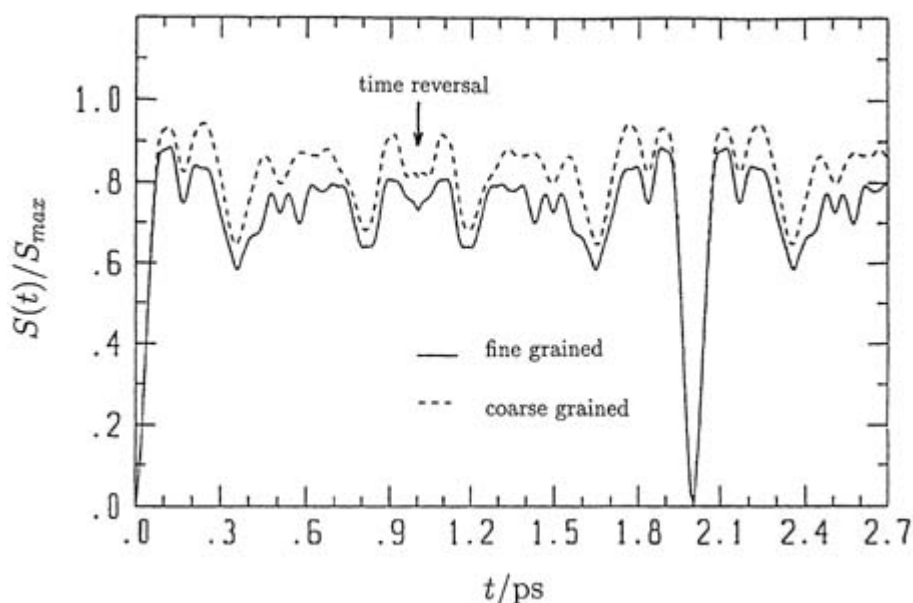


Figure B2.5.21. Time-dependent entropy $S(t)/S_{\max}$ of CHD_2F starting from a pure CH stretching excitation with six quanta ($\nu_s = 6$) at $t = 0$ fs. Time evolution with time reversal at $t = 1$ ps (after [114]).

This kind of ‘dynamical spectroscopic analysis’ is not restricted to fast primary IVR processes. It would apply just as well to the study of completely unimolecular reactions, viz isomerizations such as H-atom transfer reactions, for example $\text{CH}_2\text{O} \rightleftharpoons \text{HCHO}$ [97] $\text{HCN} \rightleftharpoons \text{HNC}$ [98], and references cited therein), and $\text{HCCH} \rightleftharpoons \text{H}_2\text{CC}$ [99] and references cited therein) (although the spectroscopic aspects have not been fully exploited in these cases), as well as the carefully studied $\text{NH}_2 \rightleftharpoons \text{NH}_3\text{O}$ [100, 101]. Recent studies on the tunnelling dynamics of hydrogen peroxide and aniline have actually carried through the method to a model for one of chemistry’s most fundamental processes: the stereomutation of chiral molecular structures [102, 103, 104, 105 and 106], see also [107]. Figure B2.5.22 illustrates the minimum energy path for the interconversion of the left- and right-handed forms of hydrogen peroxide, roughly corresponding to the torsion about the O–O bond. The quantum dynamics are governed by tunnelling through the low barrier in the *trans* configuration, even at

very high energies, a phenomenon readily understood in terms of an adiabatic picture of the stereomutation kinetics. The detailed model extracted from experimental spectra with the support of quantum-chemical calculations allows one to describe the observed mode specificity of the stereomutation in terms of the full six-dimensional quantum wavepacket dynamics. The time-dependent probability density in the reaction coordinate (figure B2.5.23) illustrates the acceleration of the stereomutation by IR excitation of the antisymmetric bend vibration (ν_6). An approximately Gaussian wavepacket initially localized on one side of the *trans* barrier (figure B2.5.22) moves periodically between the two potential wells. In the vibrational ground state ($\nu = 0$), this corresponds to the stereomutation of a chiral equilibrium structure through tunnelling to its enantiomer within 1.5 ps. Exciting the antisymmetric bend vibration ($\nu_6 = 1$) roughly halves the time required for stereomutation, an effect that could be

-34-

considered as catalysis by IR (vibrational) excitation. Figure B2.5.23 also illustrates the high degree of adiabaticity of this process: the initial form of the wavepacket in the reaction coordinate is found to be approximately conserved, even after about 10^2 tunnelling periods, although the spectroscopic result (as analysed by theory) is exact in full six-dimensional dynamics. While ν_6 can thus be considered to be a promoting mode for stereomutation, other vibrations of H_2O_2 (except torsion) have been shown to be inhibiting modes, slowing down the stereomutation process. Thus, one has cases of inhibition of a reaction by vibrational excitation. A certain degree of thermal averaging allows one to evaluate a relaxation time corresponding to rate constants more characteristic for ordinary racemization kinetics [103, 107] in contrast to the strictly periodic process shown in figure B2.5.23.

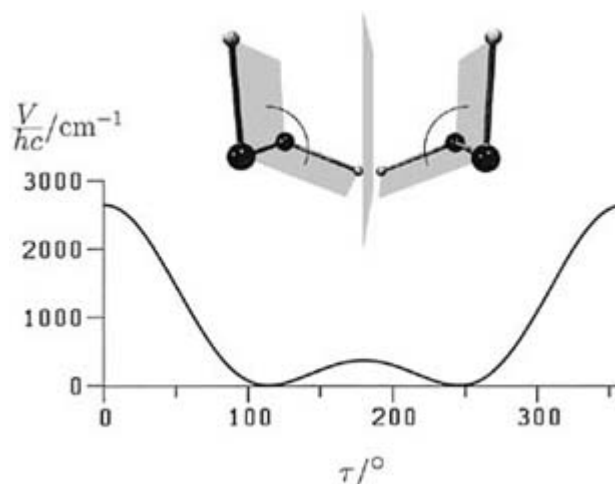


Figure B2.5.22. Potential V along the minimum energy path for the stereomutation of hydrogen peroxide. Adapted from [103].

Related results of promotion (catalysis) and inhibition of stereomutation by vibrational excitation have also been obtained for the much larger molecule, aniline-NHD ($\text{C}_6\text{H}_5\text{NHD}$), which shows short-time chirality and stereomutation [104, 105]. This kind of study opens the way to a new look at kinetics, which shows ‘coherent’ and mode-selective dynamics, even in the absence of coherent external fields. The possibility of enforcing coherent dynamics by fields (‘coherent control’) is discussed in chapter A3.13.

-35-

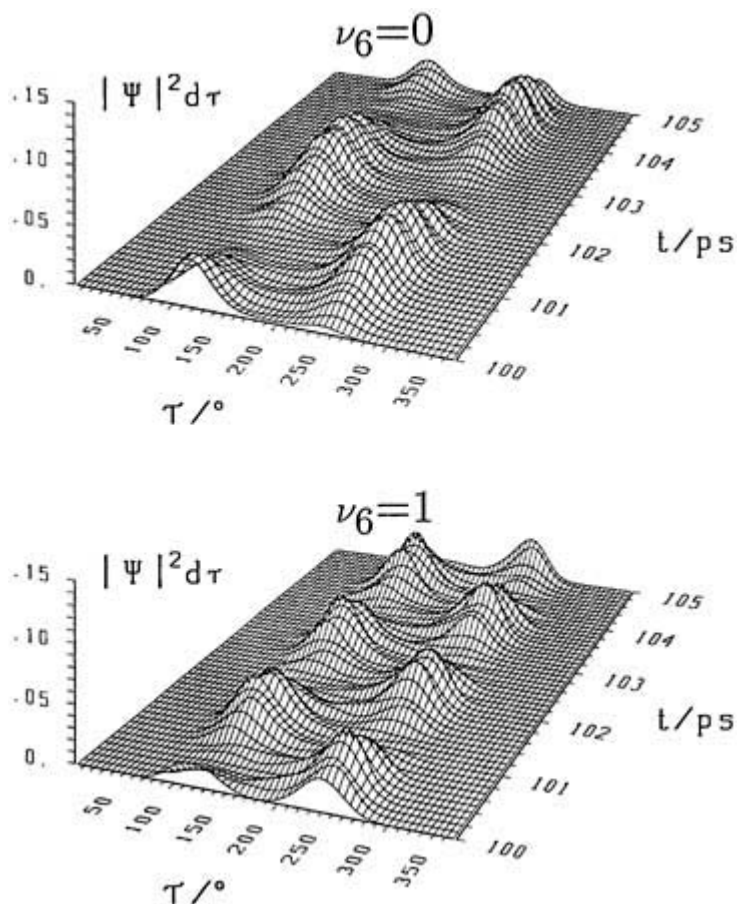


Figure B2.5.23. Mode-specific stereomutation tunnelling in hydrogen peroxide: time-dependent probability density $|\Psi|^2$ in the reaction coordinate τ (see figure B2.5.22). The probability density was integrated over the remaining coordinates (OH stretches, OH bends, OO stretch). The initial wavepacket at $t = 0$ was strictly localized on one side of the torsional barrier. $\nu_6 = 0$ refers to the vibrational ground state and $\nu_6 = 1$ to an initial state with one quantum of antisymmetric OOH bend excitation.

B2.5.9 SUMMARIZING OVERVIEW ON GAS-PHASE KINETICS STUDIES

Gas-phase kinetics studies are ideally concerned with the most fundamental events of chemical reactions related to ‘isolated, single molecules’ either as elementary unimolecular reactions, isolated bimolecular collisions, or trimolecular reactions. The experimental study of such fast elementary processes has progressed to a point where it is possible to ‘prove a reaction mechanism’ by identifying each elementary reaction contributing to the total reactive flux and by demonstrating that any conceivable additional contribution to the total reactive flux must be negligible. In fact gas-phase kinetics studies have even gone beyond this fundamental goal of reaction kinetics. By using the techniques of femtosecond spectroscopy and quantum-chemical kinetics from high-resolution spectroscopy it is possible to look into the very details of the primary processes that initiate chemical reactions. These fields are still in active development and most of the fruits from these fields still remain to be harvested.

- [1] Wilhelmy L 1850 über das Gesetz nach welchem die Einwirkung der Säuren auf Rohrzucker Stattfiudet *Ann. Physik* **81** 413–29
- [2] van't Hoff J H 1884 *Études de Dynamique Chimique* (Amsterdam: Müller)
- [3] Arrhenius S 1899 Zur Theorie der chemischen Reaktionsgeschwindigkeiten *Z. Physik. Chem.* **28** 318–35
- [4] Chance B 1949 The reaction of catalase and cyanide *J. Biol. Chem.* **179** 1299–341
- [5] Chance B 1951 Rapid and sensitive spectrophotometry. I. The accelerated and stopped-flow methods for the measurement of the reaction kinetics and spectra of unstable compounds in the visible region of the spectrum *Rev. Sci. Instrum* **22** 619–27
- [6] Eigen M 1996 *Die unmessbar schnellen Reaktionen (Ostwalds Klassiker der exakten Naturwissenschaften)* vol 281 (Thun und Frankfurt: Harri Deutsch)
- [7] Manz J and Woeste L (eds) 1995 *Femtosecond Chemistry Proc. Berlin Conf. Femtosecond Chemistry (Berlin, March 1993)* vol 1 (Weinheim: Verlag Chemie)
- [8] Porter G 1995 Flash photolysis into the femtosecond—a race against time *Femtosecond Chemistry Proc. Berlin Conf. Femtosecond Chemistry (Berlin, March 1993)* ed J Manz and L Woeste (Weinheim: Verlag Chemie) ch 1, pp 3–13
- [9] Weller A 1952 Quantitative Untersuchungen der Fluoreszenzumwandlung bei Naphtholen *Ber. Bunsenges. Phys. Chem.* **56** 662–8
- [10] Weller A 1957 Eine Verallgemeinerte Theorie diffusionsbestimmter Reaktionen und ihre Anwendung auf die Fluoreszenzlöschung *Z. Phys. Chem.* **13** 335–52
- [11] Weller A 1961 Fast reactions of excited molecules *Progress in Reaction Kinetics* (Oxford: Pergamon) pp 187–214
- [12] Jost W 1939 *Explosionen und Verbrennungsvorgänge in Gasen* (Berlin: Springer)
- [13] Knox W H, Knox R S, Hoose J F and Zare R N 1990 Observation of the 0 fs pulse *Opt. Photon. News* **1** 44–5
- [14] Herschbach D R 1987 Molecular dynamics of elementary chemical reactions *Angew. Chem.* **26** 1221–43
- [15] Lee Y T 1987 Molecular beam studies of elementary chemical processes *Angew. Chem.* **26** 939–51
- [16] Gutowsky H S and Holm C H 1956 Rate processes and nuclear magnetic resonance spectra. II. Hindered internal rotation of amides *J. Chem. Phys.* **25** 1228–34
- [17] Gutowsky H S and Holm C H 1975 Time-dependent magnetic perturbations *Dynamic Nuclear Magnetic Resonance Spectroscopy* ed L M Jackman and F A Cotton (New York: Academic) pp 1–21
- [18] Quack M 1995 Molecular femtosecond quantum dynamics between less than yoctoseconds and more than days: experiment and theory *Femtosecond Chemistry Proc. Berlin Conf. Femtosecond Chemistry (Berlin, March 1993)* ed J Manz and L Woeste (Weinheim: Verlag Chemie) ch 27, pp 781–818

- [19] Marquardt R, Quack M, Stohner J and Sutcliffe E 1986 Quantum-mechanical wavepacket dynamics of the CH group in the symmetric top X₃CH compounds using effective Hamiltonians from high-resolution spectroscopy *J. Chem. Soc. Faraday Trans. Series 2* **82** 1173–87
- [20] Marquardt R and Quack M 1991 The wavepacket motion and intramolecular vibrational redistribution

in CHX₃ molecules under infrared multiphoton excitation *J. Chem. Phys.* **95** 4854–67

- [21] Zewail A H 1995 Femto chemistry: concepts and applications *Femtosecond Chemistry Proc. Berlin Conf. Femtosecond Chemistry (Berlin, March 1993)* ed J Manz and L Woeste (Weinheim: Verlag Chemie) ch 2, pp 15–128
- [22] Zewail A H 2000 Femtochemistry: atomic-scale dynamics of the chemical bond using ultrafast lasers (Nobel lecture) *Angew. Chem.* **39** 2586–631
- [23] Gerber R B, McCoy A B and Garcia-Vela A 1995 Dynamics of photoinduced reactions in the van der Waals and in the hydrogen-bonded clusters *Femtosecond Chemistry Proc. Berlin Conf. Femtosecond Chemistry (Berlin, March 1993)* ed J Manz and L Woeste (Weinheim: Verlag Chemie) pp 499–531
- [24] Crim F F 1996 Bond-selected chemistry: vibrational state control of photodissociation and bimolecular reaction *J. Phys. Chem.* **100** 12 725
- [25] Fernández-Alonzo F, Bean B D, Ayers J D, Pomerantz A E, Zare R N, Bañares L and Aoiz F J 2000 Evidence for scattering resonances in the H + D₂ reaction *Angew. Chem. Int. Ed. (Eng.)* **39** 2748–52
- [26] Moore C B and Smith I W M 1996 State-resolved studies of reactions in the gas-phase *J. Phys. Chem.* **100** 12 848
- [27] Bernstein R B (ed) 1982 Chemical dynamics via molecular beam and laser techniques *The Hinshelwood Lectures (Oxford, 1980)* (Oxford: Oxford University Press)
- [28] Pagsberg P, Jodkowski J T, Ratajczak E and Sillesen A 1998 Experimental and theoretical studies of the reaction between CF₃ and NO₂ at 298 K *Chem. Phys. Lett.* **286** 138–44
- [29] Bernasconi C F (ed) 1976 *Relaxation Kinetics* (New York: Academic)
- [30] Sturtevant B, Shephard J E and Hornung H G (eds) 1996 *Proc. 20th Int. Symp. on Shock Waves* (Singapore: World Scientific)
- [31] Mueller M A, Yetter R A and Dryer F L 1999 Flow reactor studies and kinetic modelling of the H₂/O₂/NO_x reaction *Int. J. Chem. Kinetics* **31** 113–25
- [32] Fleming G R 1986 *Chemical Applications of Ultrafast Spectroscopy* (Oxford: Oxford University Press)
- [33] Quack M 1990 Spectra and dynamics of coupled vibrations in polyatomic molecules *Ann. Rev. Phys. Chem.* **41** 839–74
- [34] Lehmann K, Scoles G and Pate B H 1994 Intramolecular dynamics from Eigenstate-resolved infrared spectra *Ann. Rev. Phys. Chem.* **45** 241–74
- [35] Quack M 1995 Molecular infrared spectra and molecular motion *J. Mol. Struct.* **347** 245–66

- [36] Scoles G (ed) 1988 *Atomic and Molecular Beam Methods* (Oxford: Oxford University Press)
- [37] Faubel M and Toennies J P 1978 *Adv. Atom. Mol. Phys.* **13** 229
- [38] Levine R D and Bernstein R B (eds) 1989 *Molecular Reaction Dynamics and Chemical Reactivity* (Oxford: Oxford University Press)
- [39] Gehring M M, Hoyermann K, Schacke H and Wolfrum J 1973 *Proc. 14th Int. Symp. on Combustion* p

- [40] Bedjanian Y, Le Bras G and Poulet G 1999 Kinetic study of the reactions of Br₂ with OH and OD *Int. J. Chem. Kinetics* **31** 698–704
- [41] Lipson J B, Beiderhase T W, Molina L T and Molina M J 1999 Production of HCl in the OH + ClO reaction: laboratory measurements and statistical rate theory calculations *J. Phys. Chem. A* **103** 6540–51
- [42] Daugey N, Bergeat A, Schuck A, Caubet P and Dorthe G 1997 Vibrational distribution in CN($\chi^2\Sigma^+$) from the N + C₂ → CN + C reaction *Chem. Phys.* **222** 87–103
- [43] Bergeat A, Calvo T, Dorthe G and Loison J-C 1999 Fast-flow study of the CH + CH reaction products *J. Phys. Chem. A* **103** 6360–5
- [44] Mueller M A, Yetter R A and Dryer F L 1999 Flow reactor studies and kinetic modelling of the H₂/O₂/NO_x and CO/H₂O/O₂/NO_x reactions *Int. J. Chem. Kinetics* **31** 705–24
- [45] van den Bergh H and Troe J 1975 NO-catalyzed recombination of iodine atoms. Elementary steps of the complex mechanism *Chem. Phys. Lett.* **31** 351–4
- [46] Gilbert R G, Luther K and Troe J 1983 Theory of thermal unimolecular reactions in the fall-off range. II. Weak collision rate constants *Ber. Bunsenges. Phys. Chem.* **87** 169–77
- [47] Cobos C J, Hippler H and Troe J 1985 High-pressure falloff curves and specific rate constants for the reactions *J. Phys. Chem.* **89** 342–9
- [48] Quack M 1984 On the mechanism of reversible unimolecular reactions and the canonical ('high pressure') limit of the rate coefficient at low pressures *Ber. Bunsenges. Phys. Chem.* **88** 94–100
- [49] Markwalder B, Gozel P and van den Berg H 1992 Temperature-jump measurements on the kinetics of association and dissociation in weakly bound systems: N₂O₄ + M = NO₂ + NO₂ + M *J. Chem. Phys.* **97** 5472–9
- [50] Eigen M and de Maeyer L 1963 Relaxation methods *Technique of Organic Chemistry* vol 8, ed S L Friess, E S Lewis and A Weissberger (New York: Wiley) pp 895–1054
- [51] Troe J 1975 Shock wave studies of elementary chemical processes *Modern Developments in Shock Tube Research* ed G Kamimoto (Japan: Shock Tube Research Society) pp 29–54
- [52] Greene C H and Toennies J P 1964 *Chemical Reactions in Shock Waves* (London: Arnold)
- [53] Jaumotte A L (ed) 1971 *Chocs et Ondes de Choc* (Paris: Masson)
- [54] Glänzer K, Quack M and Troe J 1977 High temperature UV absorption and recombination of methyl radicals in shock waves *Proc. 16th Int. Symp. on Combustion* (Pittsburg, PA: The Combustion Institute) pp 949–60

- [55] Glänzer K, Quack M and Troe J 1976 A spectroscopic determination of the methyl radical recombination rate constant in shockwaves *Chem. Phys. Lett.* **39** 304–9
- [56] Herzberg G and Shoosmith 1956 Absorption spectrum of free CH₃ and CD₃ radicals *Can. J. Phys.* **34** 523–5

- [57] Burcat A, Skinner G B, Crossley R W and Scheller K 1973 High temperature decomposition of ethane *Int. J. Chem. Kinetics* **5** 345–52
- [58] Waage E V and Rabinovitch B S 1971 Some aspects of theory and experiment in the ethane–methyl radical system *Int. J. Chem. Kinetics* **3** 105–25
- [59] Quack M and Troe J 1974 Specific rate constants of unimolecular processes ii. Adiabatic channel model *Ber. Bunsenges. Phys. Chem.* **78** 240–52
- [60] Quack M and Troe J 1998 Statistical adiabatic channel models *Encyclopedia of Computational Chemistry* vol 4, ed P v R Schleyer *et al* (New York: Wiley) pp 2708–26
- [61] Votsmeier M, Song S, Davidson D F and Hanson R K 1999 Shock tube study of monomethylamine thermal decomposition and NH₂ high temperature absorption coefficients *Int. J. Chem. Kinetics* **31** 323–30
- [62] Votsmeier M, Song S, Davidson D F and Hanson R K 1999 Sensitive detection of NH₂ in shock tube experiments using frequency modulation spectroscopy *Int. J. Chem. Kinetics* **31** 445–53
- [63] Porter G 1950 The absorption spectroscopy of substances of short life *Discuss. Faraday Soc.* **9** 60–9
- [64] Jung D, Kärtner F X, Matuschek N, Suther D H, Morier-Genoud F, Zhang G, Keller U, Scheurer V, Tilsch M and Tschudi T 1997 Self-starting 6.5-fs pulses from a Ti:sapphire laser *Opt. Lett.* **22** 1009–11
- [65] Callear A B and Metcalfe M P 1976 *Chem. Phys.* **14** 275
- [66] van den Bergh H E, Callear A B and Norström R J 1969 An experimental determination of the oscillator strength of the 2160 Å band of the free methyl radical and a spectroscopic measurement of the combination rate *Chem. Phys. Lett.* **4** 101–2
- [67] Herzberg G 1961 The spectra and structures of free methyl and free methylene *Proc. R. Soc. A* **262** 291–317
- [68] Herzberg G 1971 *The Spectra and Structures of Simple Free Radicals. An Introduction to Molecular Spectroscopy* (Ithaca, NY: Cornell University Press)
- [69] Hippler H, Siefke M, Staerk H and Troe J 1999 New studies of the unimolecular reaction NO₂ O + NO. Part 1. High pressure range of the O + NO recombination between 200 and 400 K *Phys. Chem. Chem. Phys.* **1** 57–61
- [70] Sinha M P, Schulz A and Zare R N 1973 Internal state distribution of alkali dimers in supersonic nozzle beams *J. Chem. Phys.* **58** 549–56
- [71] Zewail A H 1993 Femtochemistry *J. Phys. Chem.* **97** 12 427–46
- [72] Zewail A H 1994 *Femtochemistry. Ultrafast Dynamics of the chemical Bond (World Scientific Series in 20th Century Chemistry, vol 3)* (Singapore: World Scientific)
- [73] Zewail A H 1995 Femtosecond dynamics of reactions: elementary processes of controlled solvation *Ber. Bunsenges. Phys. Chem.* **99** 474–7

- [74] Rosker M J, Rose T S and Zewail A 1988 Femtosecond real-time dynamics of photofragment-trapping resonances on dissociative potential-energy surfaces *Chem. Phys. Lett.* **146** 175–9

- [75] Lupo D W and Quack M 1987 IR-laser photochemistry *Chem. Rev.* **87** 181–216
- [76] Quack M 1982 Reaction dynamics and statistical mechanics of the preparation of highly excited states by intense infrared radiation *Adv. Chem. Phys.* **50** 395–473
- [77] Hippler M, Quack M, Schwarz R, Seyfang G, Matt S and Märk T 1997 Infrared multiphoton excitation, dissociation, and ionization of C₆₀ *Chem. Phys. Lett.* **278** 111–20
- [78] Ditmire T, Zweiback J, Yanovsky V P, Cowan T E, Hays G and Wharton K B 1999 Nuclear fusion from explosions of femtosecond laser-heated deuterium clusters *Nature* **389** 489–92
- [79] Pretzler G *et al* 1998 Neutron production by 200 mJ ultrashort laser pulses *Phys. Res. E* **58** 1165–8
- [80] Quack M 1989 Infrared laser chemistry and the dynamics of molecular multiphoton excitation *Infrared Phys.* **29** 441–66
- [81] Quack M 1995 IR laser chemistry *Infrared Phys. Technol.* **36** 365–80
- [82] Quack M and Thöne H J 1987 Absolute and relative rate coefficients in the IR-laser chemistry of bichromophoric fluorobutanes: tests for inter- and intra-molecular selectivity *Chem. Phys. Lett.* **135** 487–94
- [83] Rabinovitch B S and Flowers M C 1964 Chemical activation *Q. Rev. Chem. Soc.* **18** 122–67
- [84] Oref I and Rabinovitch B S 1979 Do highly excited polyatomic molecules behave ergodically? *Acc. Chem. Res.* **12** 166–75
- [85] Flowers M C and Rabinovitch B S 1985 Localization of excitation energy in chemically activated systems. 3-ethyl-2-methyl-2-pentyl radicals *J. Phys. Chem.* **89** 563–5
- [86] von E Doering W, Gilbert J C and Leermakes P A 1968 Symmetrical distribution of energy in initially unsymmetrically excited products. Reaction of dideuteriodiazomethane with allene, methylenecyclopropane, and vinylcyclopropane *Tetrahedron* **29** 6863–72
- [87] Setser D W 1972 *International Review of Science. Physical Chemistry* ed J C Polanyi (London: Butterworths)
- [88] Sang Kyu Kim, Ju Guo, Baskin J S and Zewail A H 1996 Femtosecond chemically activated reactions: concept of nonstatistical activation at high thermal energies *J. Phys. Chem.* **100** 9202–5
- [89] Mack J A, Mikulecky K and Wodtke A M 1997 Resonant vibration–vibration energy transfer between highly vibrationally excited O₂(X³Σ_g⁻, v = 15–20) and CO₂, N₂O, N₂, and O₃ *J. Chem. Phys.* **105** 4105–16
- [90] Ackermann M and Biauwe F 1970 Structure of the Schumann–Runge bands from the 0–0 to the 13–0 band *J. Mol. Spectrosc.* **35** 73–82
- [91] Cheung A S C, Yoshino K, Freeman D E, Friedman R S, Dalgarno A and Parkinson W H 1989 The Schumann–Runge absorption-bands of ¹⁶O¹⁸O in the wavelength region 175–205 nm and spectroscopic constants of isotopic oxygen molecules *J. Mol. Spectrosc.* **134** 362–89

- [92] von Puttkamer K, Dübal H-R and Quack M 1983 Time-dependent processes in polyatomic molecules during and after intense infrared irradiation *Faraday Discuss. Chem. Soc.* **75** 197–210

- [93] Quack M and Suhm M A 1991 Potential energy surfaces, quasiadiabatic channels, rovibrational spectra, and intramolecular dynamics of (HF)₂ and its isotopomers from quantum Monte Carlo calculations *J. Chem. Phys.* **95** 28–59
- [94] Quack M and Kutzelnigg W 1995 Molecular spectroscopy and molecular dynamics: theory and experiment *Ber. Bunsenges. Phys. Chem.* **99** 231–45
- [95] Beil A, Luckhaus D, Quack M and Stohner J 1997 Intramolecular vibrational redistribution and unimolecular reaction: concepts and new results on the femtosecond dynamics and statistics in CHFClBr *Ber. Bunsenges. Phys. Chem.* **101** 311–28
- [96] Quack M 1993 Molecular quantum dynamics from high resolution spectroscopy and laser chemistry *J. Mol. Struct.* **292** 171–96
- [97] Moore C B and Weisshaar J C 1983 Formaldehyde photochemistry *Ann. Rev. Phys. Chem.* **34** 525
- [98] Bowman J M and Gazdy B 1997 A new perspective on isomerization dynamics illustrated by HCN→HNC *J. Phys. Chem. A* **101** 6384–8
- [99] Kiefer J H, Mudipalli P S, Wagner A F and Harding L 1996 Importance of hindered rotations in the thermal dissociation of small unsaturated molecules: classical formulation and application to hcn and hcch *J. Chem. Phys.* **105** 1–22
- [100] Luckhaus D 1997 The rovibrational spectrum of hydroxylamine: a combined high resolution experimental and theoretical study *J. Chem. Phys.* **106** 8409–26
- [101] Luckhaus D 1997 The rovibrational dynamics of hydroxylamine *Ber. Bunsenges. Phys. Chem.* **101** 346–55
- [102] Kuhn B, Rizzo T R, Luckhaus D, Quack M and Suhm M A 1999 A new six-dimensional analytical potential up to chemically significant energies for the electronic ground state of hydrogen peroxide *J. Chem. Phys.* **111** 2565–87
- [103] Fehrensen B, Luckhaus D and Quack M 1999 Mode selective stereomutation tunnelling in hydrogen peroxide isotopomers *Chem. Phys. Lett.* **300** 312–20
Fehrensen B, Luckhaus D and Quack M 2001 to be published
- [104] Fehrensen B, Luckhaus D and Quack M 1999 Inversion tunneling in aniline from high resolution infrared spectroscopy and an adiabatic reaction path hamiltonian approach *Z. Phys. Chem.* **209** 1–19
- [105] Fehrensen B, Hippler M and Quack M 1998 Isotopomer selective overtone spectroscopy by ionization detected IR + UV double resonance jet-cooled aniline *Chem. Phys. Lett.* **298** 320–8
- [106] Luckhaus D 2000 6D vibrational quantum dynamics: generalized coordinate discrete variable representation and (a)diabatic contraction *J. Chem. Phys.* **113** 1329–47
- [107] Quack M 1989 Structure and dynamics of chiral molecules *Angew. Chem.* **28** 571–86
- [108] Shank C 1985 *Laser Focus* March
- [109] He Y, Pochert J, Quack M, Ranz R and Seyfang G 1995 Dynamics of unimolecular reactions induced by monochromatic infrared radiation: experiment and theory for C_nF_mXI→C_nF_mX + I probed with hyperfine-, Doppler- and uncertainty limited time resolution of iodine atom infrared absorption *J. Chem. Soc. Faraday Discuss.* **102** 275–300

- [110] Quack M and Jans-Bürli S 1986 *Molekulare Thermodynamik und Kinetik. Teil 1: Chemische Reaktionskinetik* (Zürich: Verlag der Fachvereine) (New English edition in preparation)
- [111] Zewail A H 1988 Laser femtochemistry *Science* **242** 1645–53
- [112] Quack M 1998 Multiphoton excitation *Encyclopedia of Computational Chemistry* vol 3, ed P v R Schleyer *et al* (New York: Wiley) pp 1775–91
- [113] Quack M, Sutcliffe E, Hackett P A and Rayner D M 1986 Molecular photofragmentation with many infrared photons. Absolute rate parameters from quantum dynamics, statistical mechanics, and direct measurement *Faraday Discuss. Chem. Soc.* **82** 229–40
- [114] Quack M and Stohner J 1993 Femtosecond quantum dynamics of functional groups under coherent infrared multiphoton excitation as derived from the analysis of high-resolution spectra *J. Phys. Chem.* **97** 12 574–90
- [115] Settle R D F and Rizzo T R 1992 *J. Chem. Phys.* **97** 2823
Boyarkine O V, Settle R D F and Rizzo T R 1995 Vibrational overtone spectra of jet-cooled CF₃H by infrared laser assisted photofragment spectroscopy *Ber. Bunsenges. Phys. Chem.* **99** 504–13
- [116] Quack M 1999 Intramolekulare Dynamik, Irreversibilität, Zeitumkehrsymmetrie und eine absolute Moleküluhr *Nova Acta Leopoldina NF* **81** 137–73
- [117] Douley E A, Marquardt R, Quack M, Stohner J, Thanopoulos I and Wallenborn E-U 2001 *Mol. Phys.* to be published
-

FURTHER READING

Bernasconi C F 1976 *Relaxation Kinetics* (New York: Academic)

Faraday Discussions of the Chemical Society **112** *Unimolecular Dynamics*

Fleming G R 1986 *Chemical Applications of Ultrafast Spectroscopy* (Oxford: Oxford University Press)

Johnston H S 1966 *Gas Phase Reaction Rate Theory* (Ronald)

Levine R D and Bernstein R B 1989 *Molecular Reaction Dynamics and Chemical Reactivity* (Oxford: Oxford University Press)

Lupo D W and Quack M 1987 IR-laser photochemistry *Chem. Rev.* **87** 181–216

Manz J and Woeste L (eds) 1995 *Femtosecond Chemistry Proc. Berlin Conf. Femtosecond Chemistry (Berlin, March 1993)* (Weinheim: Verlag Chemie)

Pilling M J and Smith I W M (eds) 1987 *Modern Gas Kinetics. Theory, Experiment and Application* (Oxford: Blackwell)

Quack M 1982 Reaction dynamics and statistical mechanics of the preparation of highly excited states by intense infrared radiation *Adv. Chem. Phys.* **50** 395–473

Quack M and Jans-Bürli S 1986 *Molekulare Thermodynamik und Kinetik. Teil 1. Chemische Reaktionskinetik* (Zürich: Verlag der Fachvereine)

Sandström J 1981 *Dynamic NMR Spectroscopy* (New York: Academic)

Steinfeld J I, Francisco S and Hase W L 1998 *Chemical Kinetics and Dynamics* 2nd edn (Englewood Cliffs, NJ: Prentice-Hall)

B3.1 Quantum structural methods for atoms and molecules

Jack Simons

B3.1.1 WHAT DOES QUANTUM CHEMISTRY TRY TO DO?

Electronic structure theory describes the motions of the electrons and produces energy surfaces and wavefunctions. The shapes and geometries of molecules, their electronic, vibrational and rotational energy levels, as well as the interactions of these states with electromagnetic fields lie within the realm of quantum structure theory.

B3.1.1.1 THE UNDERLYING THEORETICAL BASIS—THE BORN–OPPENHEIMER MODEL

In the Born–Oppenheimer [1] model, it is assumed that the electrons move so quickly that they can adjust their motions essentially instantaneously with respect to any movements of the heavier and slower atomic nuclei. In typical molecules, the valence electrons orbit about the nuclei about once every 10^{-15} s (the inner-shell electrons move even faster), while the bonds vibrate every 10^{-14} s, and the molecule rotates approximately every 10^{-12} s. So, for typical molecules, the fundamental assumption of the Born–Oppenheimer model is valid, but for loosely held (e.g. Rydberg) electrons and in cases where nuclear motion is strongly coupled to electronic motions (e.g. when Jahn–Teller effects are present) it is expected to break down.

This separation-of-time-scales assumption allows the electrons to be described by electronic wavefunctions that smoothly ‘ride’ the molecule’s atomic framework. These electronic functions are found by solving a Schrödinger equation whose Hamiltonian \hat{H}_e contains the kinetic energy T_e of the electrons, the Coulomb repulsions among all the molecule’s electrons V_{ee} , the Coulomb attractions V_{en} among the electrons and all of the molecule’s nuclei, treated with these nuclei held clamped, and the Coulomb repulsions V_{nn} among all of these nuclei, but it does not contain the kinetic energy T_N of all the nuclei. That is, this Hamiltonian keeps the nuclei held fixed in space. The electronic wavefunctions ψ_k and energies E_k that result

$$\hat{H}_e \psi_k = E_k \psi_k$$

thus depend on the locations $\{Q_i\}$ at which the nuclei are sitting. That is, the E_k and ψ_k are parametric functions of the coordinates of the nuclei, and, of course, the wavefunctions ψ_k depend on the coordinates of all of the electrons.

These electronic energies’ dependence on the positions of the atomic centres cause them to be referred to as electronic energy surfaces such as that depicted below in [figure B3.1.1](#) for a diatomic molecule. For nonlinear polyatomic molecules having N atoms, the energy surfaces depend on $3N - 6$ internal coordinates and thus can be very difficult to visualize. In [figure B3.1.2](#), a ‘slice’ through such a surface is shown as a function of two of the $3N - 6$ internal coordinates.

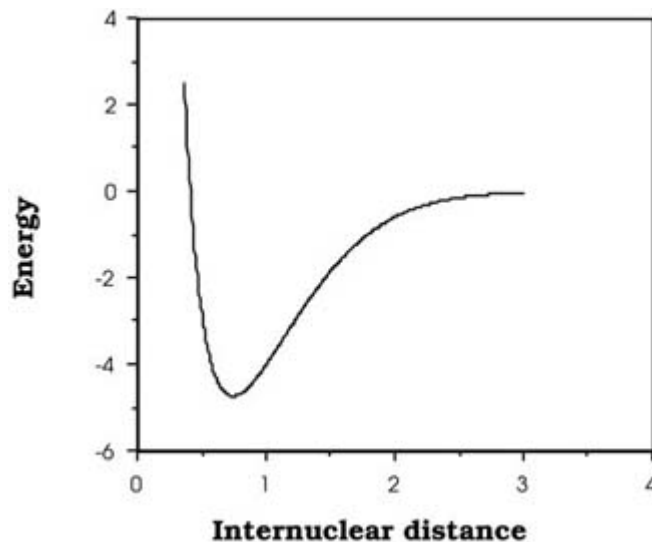


Figure B3.1.1. Energy as a function of internuclear distance for a typical bound diatomic molecule or ion.

The Born–Oppenheimer theory is soundly based in that it can be derived from a Schrödinger equation describing the kinetic energies of all electrons and of all N nuclei plus the Coulomb potential energies of interaction among all electrons and nuclei. By expanding the wavefunction Ψ that is an eigenfunction of this full Schrödinger equation in the complete set of functions $\{\psi_k\}$ and then neglecting all terms that involve derivatives of any ψ_k with respect to the nuclear positions $\{Q_i\}$, one can separate variables such that:

- (1) the electronic wavefunctions and energies obey

$$\hat{H}_e \psi_k = E_k \psi_k$$

- (2) the nuclear motion (i.e. vibration/rotation) wavefunctions obey

$$(\hat{T}_N + E_k) \chi_{k,L} = E_{k,L} \chi_{k,L}$$

where T_N is the kinetic energy operator for movement of all nuclei.

Each and every electronic energy state, labelled k , has a set, labelled L , of vibration/rotation energy levels $E_{k,L}$ and wavefunctions $\chi_{k,L}$.

B3.1.1.2 NON-BORN–OPPENHEIMER CORRECTIONS—RADIATIONLESS TRANSITIONS

Because the Born–Oppenheimer model is obtained from the full Schrödinger equation by making approximations, it is not exact. Thus, in certain circumstances it becomes necessary to correct the predictions of the Born–Oppenheimer theory (i.e. by including the effects of the neglected coupling terms using perturbation theory). For example, when developing a theoretical model to interpret the rate at which electrons are ejected from rotationally/vibrationally hot NH^- ions, we had to consider [3] coupling between:

- (1) $^2\Pi$ NH^- in its $\nu = 1$ vibrational level and in a high rotational level (e.g. $J > 30$) prepared by laser excitation of vibrationally ‘cold’ NH^- in $\nu = 0$ having high J (due to natural Boltzmann populations), see [figure B3.1.3](#) and
- (2) $^3\Sigma^-$ NH neutral plus an ejected electron in which the NH is in its $\nu = 0$ vibrational level (no higher level

is energetically accessible) and in various rotational levels (labelled N).

Because NH has an electron affinity of 0.4 eV, the total energies of the above two states can be equal only if the kinetic energy KE carried away by the ejected electron obeys

$$KE = E_{\text{vib/rot}}(\text{NH}^- (v = 1, J)) - E_{\text{vib/rot}}(\text{NH} (v = 0, N)) - 0.4 \text{ eV}.$$

In the absence of any coupling terms, no electron detachment would occur. It is only by the anion converting some of its vibration/rotation energy and angular momentum into electronic energy that the electron that occupies a bound N_{2p} orbital in NH^- can gain enough energy to be ejected.

My own research efforts [4] have, for many years, involved taking into account such non-Born–Oppenheimer couplings, especially in cases where vibration/rotation energy transferred to electronic motions causes electron detachment, as in the NH^- case detailed above. Professor Yngve Öhrn has been active [5] in attempting to avoid using the Born–Oppenheimer approximation and, instead, treating the dynamical motions of the nuclei and electrons simultaneously. Professor David Yarkony has contributed much [6] to the recent treatment of non-Born–Oppenheimer effects and to the inclusion of spin–orbit coupling in such studies.

B3.1.1.3 WHAT IS LEARNED FROM AN ELECTRONIC STRUCTURE CALCULATION?

The knowledge gained via structure theory is great. The electronic energies $E_k(Q)$ allow one to determine [7] the geometries and relative energies of various isomers that a molecule can assume by finding those geometries $\{Q_i\}$ at which the energy surface E_k has minima $\partial E_k / \partial Q_i = 0$, with all directions having positive curvature (this is monitored by considering the so-called Hessian matrix $H_{i,j} = \partial^2 E_k / \partial Q_i \partial Q_j$: if none of its eigenvalues are negative, all directions have positive curvature). Such geometries describe stable isomers, and the energy at each such isomer geometry gives the relative energy of that isomer. Professor Berny Schlegel [8] has been one of the leading figures in using gradient and Hessian information to locate stable structures and transition states. Professor Peter Pulay [9] has done as much as anyone to develop the theory that allows us to compute gradients and Hessians for most commonly used electronic structure methods.

There may be other geometries on the E_k energy surface at which all ‘slopes’ vanish $\partial E_k / \partial Q_i = 0$, but at which not all directions possess positive curvature. If the Hessian matrix has only one negative eigenvalue, there is only one direction leading downhill away from the point $\{Q_i\}$ of zero force; all the remaining directions lead uphill from this point. Such a geometry describes that of a *transition state*, and its energy plays a central role in determining the rates of reactions which pass through this transition state. The energy surface shown in [figure B3.1.2](#) displays such transition states, and it also shows a second-order saddle point (i.e. a point where the gradient vanishes and the Hessian has two directions of negative curvature).

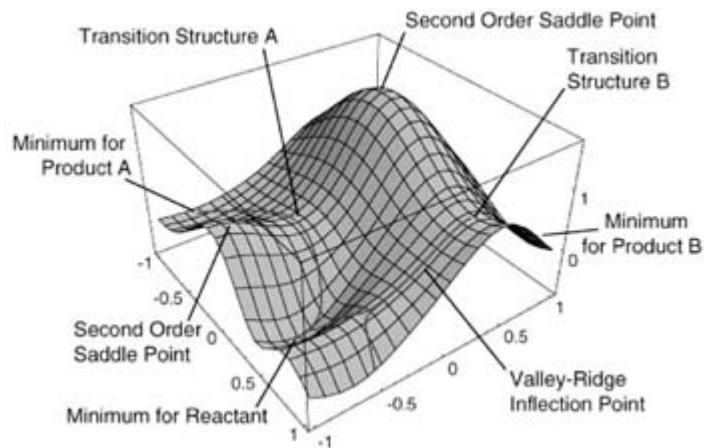


Figure B3.1.2. Two-dimensional slice through a $(3N - 6)$ -dimensional energy surface of a polyatomic molecule or ion. After [2].

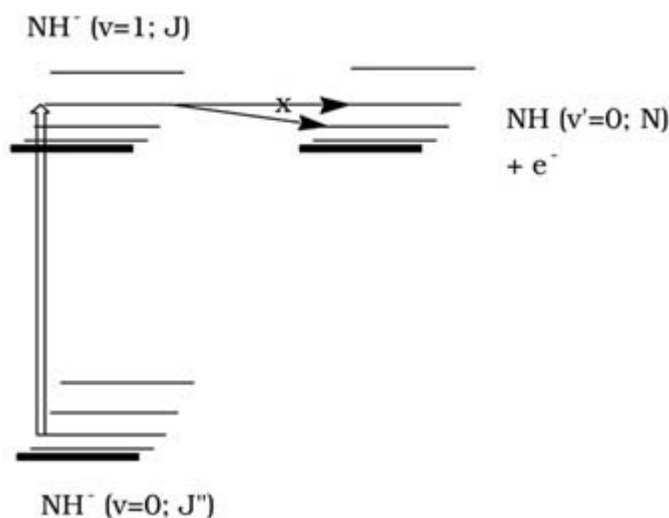


Figure B3.1.3. Energies of NH^- and of NH pertinent to the autodetachment of $v = 1, J$ levels of NH^- formed by laser excitation of $v = 0, J''$ NH^- .

At any geometry $\{Q_i\}$, the gradient vector having components $\partial E_k / \partial Q_i$ provides the forces ($F_i = -\partial E_k / \partial Q_i$) along each of the coordinates Q_i . These forces are used in molecular dynamics simulations which solve the Newton $\mathbf{F} = m\mathbf{a}$ equations and in molecular mechanics studies which are aimed at locating those geometries where the \mathbf{F} vector vanishes (i.e. the stable isomers and transition states discussed above).

Also produced in electronic structure simulations are the electronic wavefunctions $\{\psi_k\}$ and energies $\{E_k\}$ of each of the electronic states. The separation in energies can be used to make predictions on the spectroscopy of the system. The wavefunctions can be used to evaluate the properties of the system that depend on the spatial distribution of the electrons. For example, the z component of the dipole moment [10] of a molecule μ_z can be computed by integrating

the probability density for finding an electron at position \mathbf{r} multiplied by the z coordinate of the electron and the electron's charge e : $\mu_z = \int e \psi_k^* \psi_k z \, d\mathbf{r}$. The average kinetic energy of an electron can also be computed by carrying out such an average-value integral: $\int \psi_k^* (-\hbar^2 / 2m_e \nabla^2) \psi_k \, d\mathbf{r}$. The rules for computing the average value of any physical observable are developed and illustrated in popular undergraduate text books on

physical chemistry [11] and in graduate-level texts [12].

Not only can electronic wavefunctions tell us about the average values of all the physical properties for any particular state (i.e. ψ_k above), but they also allow us to tell us how a specific ‘perturbation’ (e.g. an electric field in the Stark effect, a magnetic field in the Zeeman effect and light’s electromagnetic fields in spectroscopy) can alter the specific state of interest. For example, the perturbation arising from the electric field of a photon interacting with the electrons in a molecule is given within the so-called electric dipole approximation [12] by:

$$\hat{H}_{\text{pert}} = \sum_j e^2 \mathbf{r}_j \cdot \mathbf{E}(t)$$

where \mathbf{E} is the electric field vector of the light, which depends on time t in an oscillatory manner, and \mathbf{r}_j gives the spatial coordinates of the j th electron. This perturbation, \hat{H}_{pert} can induce transitions to other states ψ_k , with probabilities that are proportional to the square of the integral:

$$\int \psi_k^* \hat{H}_{\text{pert}} \psi_k \, d\mathbf{r}.$$

So, if this integral were to vanish, transitions between ψ_k and ψ_k would not occur, and would be referred to as ‘forbidden’. Whether such integrals vanish or not often is determined by symmetry. For example, if ψ_k were of odd symmetry under a plane of symmetry σ_v of the molecule, while ψ_k were even under σ_v , then the integral would vanish unless one or more of the three Cartesian components of the dot product $\mathbf{r}_j \cdot \mathbf{E}$ were odd under σ_v . The general idea is that for the integral not to vanish, the direct product of the symmetries of ψ_k and of ψ_k must match the symmetry of at least one of the symmetry components present in \hat{H}_{pert} . Professor Poul Jørgensen [13] has been involved in developing such so-called response theories for perturbations that may be time dependent (e.g. as in the interaction of light’s electromagnetic radiation).

B3.1.1.4 SUMMARY

In summary, computational *ab initio* quantum chemistry attempts to solve the electronic Schrödinger equation for the $E_k(\mathbf{R})$ energy surfaces and wavefunctions $\psi_k(\mathbf{r};\mathbf{R})$ on a ‘grid’ of values for the ‘clamped’ nuclear positions. Because the Schrödinger equation produces wavefunctions, it has a great deal of predictive power. Wavefunctions contain all the information needed to compute dipole moments, polarizability, etc and transition properties such as the electric dipole transition strengths among states. They also permit the evaluation of system responses with respect to external perturbations such as geometrical distortions [9], which provides information on vibrational frequencies and reaction paths.

B3.1.2 WHY IS IT SO DIFFICULT TO CALCULATE ELECTRONIC ENERGIES AND WAVEFUNCTIONS WITH REASONABLE ACCURACY?

As a scientific tool, *ab initio* quantum chemistry is not yet as accurate as modern laser spectroscopic measurements, for example. Moreover, it is difficult to estimate the accuracies with which various methods predict bond energies and lengths, excitation energies and the like. In the opinion of the author, chemists who

rely on the results of quantum chemistry calculations must better understand what underlies the concepts and methods of this field. Only by so doing will they be able to judge for themselves the value of given quantum chemistry data to their own research. There exist a variety of sources of further information on the ‘jargon’, underlying theory, methodologies, and current strengths and weaknesses of *ab initio* quantum chemistry. In 1996, Head-Gordon [14] produced a nice overview entitled ‘Quantum chemistry and molecular processes’, Schaefer *et al* [15] offered a very good discussion in 1995; Simons [16] offered a somewhat earlier perspective in 1991. The present chapter includes many of the ideas contained in these and other earlier descriptions of this field’s impacts, but also attempts to extend the perspective to include more recent developments.

Returning now to the issue of the accuracy of various electronic structure predictions, it is natural to ask why it is so difficult to achieve reasonable accuracy (i.e. ca. 1 kcal mol⁻¹ in computed bond energies or activation energies) even with the most sophisticated and computer-resource-intensive quantum chemistry calculations. The reasons include the following.

- (A) *Many-body problems with R^{-1} potentials* are notoriously difficult. It is well known that the Coulomb potential falls off so slowly with distance that mathematical difficulties can arise. The $4\pi R^2$ dependence of the integration volume element, combined with the R^{-1} dependence of the potential, produce ill-defined interaction integrals unless attractive and repulsive interactions are properly combined. The classical or quantum treatment of ionic melts [17], many-body gravitational dynamics [18] and Madelung sums [19] for ionic crystals are all plagued by such difficulties.
- (B) *The electrons require quantal treatment and they are indistinguishable.* The electron’s small mass produces local de Broglie wavelengths that are long compared to atomic ‘sizes’, thus necessitating quantum treatment. Their indistinguishability requires that permutational symmetry be imposed on solutions of the Schrödinger equation.
- (C) *All mean-field models of electronic structure require large corrections.* Essentially all *ab initio* quantum chemistry approaches introduce a ‘mean field’ potential V_{mf} that embodies the average interactions among the N electrons. The difference between the mean-field potential and the true Coulombic potential is termed [20] the ‘*fluctuation potential*’. The solutions $\{\Psi_k, E_k\}$ to the true electronic Schrödinger equation are then approximated in terms of solutions $\{\Psi_k^0, E_k^0\}$ to the model Schrödinger equation in which V_{mf} is used. Improvements to the solutions of the model problem are made using perturbation theory or the variational method. Such approaches are expected to work when the difference between the starting model and the final goal is small in some sense.

The most elementary mean-field models of electronic structure introduce a potential that an electron at r_1 would experience if it were interacting with a *spatially averaged* electrostatic charge density arising from the $N - 1$ remaining electrons:

-7-

$$V_{\text{mf}}(\mathbf{r}_1) = \int \rho_{N-1}(\mathbf{r}') \frac{e^2}{|\mathbf{r}_1 - \mathbf{r}'|} d\mathbf{r}'.$$

Here $\rho_{N-1}(\mathbf{r}')$ represents the probability density for finding the $N - 1$ electrons at \mathbf{r}' , and $e^2 / |\mathbf{r}_1 - \mathbf{r}'|$ is the mutual Coulomb repulsion between electron density at r_1 and \mathbf{r}' .

The magnitude and ‘shape’ of such a mean-field potential is shown below [21] in figure B3.1.4 for the two 1s electrons of a beryllium atom. The Be nucleus is at the origin, and one electron is held fixed 0.13 Å from the nucleus, the maximum of the 1s orbital’s radial probability density. The Coulomb potential experienced by the second electron is then a function of the second electron’s position along the x -axis (connecting the Be nucleus and the first electron) and its distance perpendicular to the x -axis. For simplicity, this second electron

is arbitrarily constrained to lie on the x -axis. Along this direction, the Coulomb potential is singular, and hence the overall interactions are very large.

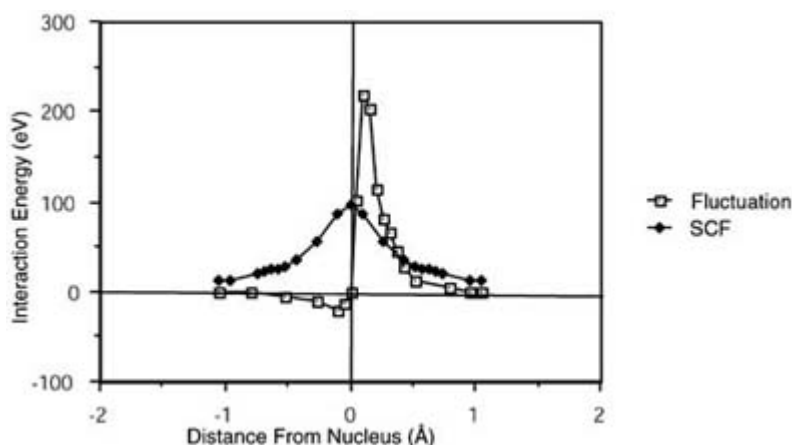


Figure B3.1.4. Fluctuation and mean-field SCF potentials for a 2s electron in Be.

On the ordinate, two quantities are plotted: (i) the mean-field potential between the second electron and the other 1s electron computed, via the self-consistent field (SCF) process (described later), as the interaction of the second electron with a spherical $|1s|^2$ charge density centred on the Be nucleus; and (ii) the fluctuation potential (F) of this average (mean-field) interaction.

As a function of the inter-electron distance, the fluctuation potential decays to zero more rapidly than does the mean-field potential. However, the magnitude of F is quite large and remains so over an appreciable range of inter-electron distances. The corrections to the mean-field picture are therefore quite large when measured in kcal mol^{-1} . For example, the differences (called pair correlation energies) ΔE between the true (state-of-the-art quantum chemical calculation as discussed later) energies of the interaction among the four electrons in the Be atom and the mean-field estimates of these interactions are given in [table B3.1.1](#) in electronvolts ($1 \text{ eV} = 23.06 \text{ kcal mol}^{-1}$).

Table 3.1.1 Pair correlation energies for the four electrons in Be.

Orbital pair	$1s_\alpha 1s_\beta$	$1s_\alpha 2s_\alpha$	$1s_\alpha 2s_\beta$	$1s_\beta 2s_\alpha$	$1s_\beta 2s_\beta$	$2s_\alpha 2s_\beta$
ΔE (eV)	1.126	0.022	0.058	0.058	0.022	1.234

Another example of the difficulty is offered in figure B3.1.5. Here we display on the ordinate, for helium's 1S ($1s^2$) state, the probability of finding an electron whose distance from the He nucleus is 0.13 \AA (the peak of the 1s orbital's density) and whose angular coordinate relative to that of the other electron is plotted on the abscissa. The He nucleus is at the origin and the second electron also has a radial coordinate of 0.13 \AA . As the relative angular coordinate varies away from 0° , the electrons move apart; near 0° , the electrons approach one another. Since both electrons have opposite spin in this state, their mutual Coulomb repulsion alone acts to keep them apart.

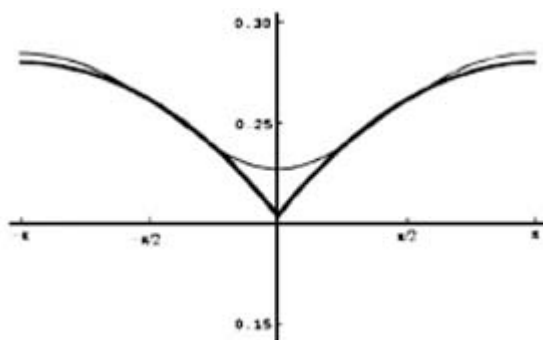


Figure B3.1.5. Probability (as a function of angle) for finding the second electron in He when both electrons are located at the maximum in the 1s orbital's probability density. The bottom line is that obtained using a Hylleraas-type function, and the other related to a highly-correlated multiconfigurational wavefunction. After [22].

What figure B3.1.5 shows is that, for a highly accurate wavefunction (one constructed using so-called Hylleraas functions [23] that depend explicitly on the coordinates of the two electrons as well as on their interparticle distance coordinate), one finds a 'cusp' in the probability density for finding one electron in the neighbourhood of another electron with the same spin. The probability plot for the Hylleraas function is the lower bold curve in figure B3.1.5. The line above the Hylleraas plot was extracted from a configuration interaction wavefunction for He obtained using a rather large atomic orbital (AO) basis set [22]. Even for such a sophisticated wavefunction (of the type used in many state-of-the-art *ab initio* calculations), the cusp in the relative probability distribution is, clearly, not well represented. Finally, the Hartree–Fock (HF) probability, which is not even displayed above, would, if plotted, be flat as a function of the angle shown above and thus clearly very much in error.

B3.1.2.1 SUMMARY

The above evidence shows why an *ab initio* solution of the Schrödinger equation is a very demanding task if high accuracy is desired. The HF potential takes care of 'most' of the interactions among the N electrons (which interact via long-range Coulomb forces and whose dynamics requires the application of quantum physics and permutational

-9-

symmetry). However, the residual fluctuation potential is large enough to cause significant corrections to the HF picture. The reality is that electrons in atoms and molecules undergo dynamical motions in which their Coulomb repulsions cause them to 'avoid' one another at every instant of time, not only in the average-repulsion manner that the mean-field models embody. The inclusion of instantaneous spatial correlations (usually called *dynamical correlations*) among electrons is necessary to achieve a more accurate description of the atomic and molecular electronic structure.

B3.1.3 WHAT ARE THE ESSENTIAL CONCEPTS OF *AB INITIO* QUANTUM CHEMISTRY?

The mean-field potential and the need to improve it to achieve reasonably accurate solutions to the true electronic Schrödinger equation introduce three constructs that characterize essentially all *ab initio* quantum chemical methods: *orbitals*, *configurations* and *electron correlation*.

B3.1.3.1 ORBITALS AND CONFIGURATIONS—WHAT ARE THEY (REALLY)?

(A) HOW THE MEAN-FIELD MODEL LEADS TO ORBITALS AND CONFIGURATIONS

The mean-field potentials that have proven most useful are all one-electron additive: $V_{\text{mf}}(\mathbf{r}) = \sum_j V_{\text{mf}}(\mathbf{r}_j)$. Since the electronic kinetic energy $\hat{T} = \sum_j \hat{T}_j$ operator is also one-electron additive, so is the mean-field

Hamiltonian $\hat{H}^0 = \hat{T} + \hat{V}_{\text{mf}}$. The additivity of \hat{H}^0 implies that the mean-field energies E_k^0 are additive and the wavefunctions $\{\Psi_k^0\}$ can be formed in terms of products of functions $\{\phi_k\}$ of the coordinates of the individual electrons.

Thus, it is the *ansatz* that V_{mf} is separable that leads to the concept of *orbitals*, which are the one-electron functions $\{\phi_j\}$ found by solving the one-electron Schrödinger equations: $(\hat{T}_1 + \hat{V}_{\text{mf}}(\mathbf{r}_1)) \phi_j(\mathbf{r}_1) = \epsilon_j \phi_j(\mathbf{r}_1)$; the eigenvalues $\{\epsilon_j\}$ are called *orbital energies*.

Given the complete set of solutions to this one-electron equation, a complete set of N -electron mean-field wavefunctions can be written. Each $\{\Psi_k^0\}$ is constructed by forming a product of N orbitals chosen from the set of $\{\phi_j\}$, allowing each orbital in the list to be a function of the coordinates of one of the N electrons (e.g. $\Psi_k^0 = |\phi_{k1}(\mathbf{r}_1)\phi_{k2}(\mathbf{r}_2)\phi_{k3}(\mathbf{r}_3) \dots \phi_{kN-1}(\mathbf{r}_{N-1})\phi_{kN}(\mathbf{r}_N)|$, as above). The corresponding mean-field energy is evaluated as the sum over those orbitals that appear in Ψ_k^0 : $E_k^0 = \sum_{j=1, N} \epsilon_{kj}$.

Because of the indistinguishability of the N electrons, the antisymmetric component of any such orbital product must be formed to obtain the proper mean-field wavefunction. To do so, one applies the so-called antisymmetrizer operator [24] $\hat{A} = \sum_P (-1)^P \hat{P}$, where the permutation operator \hat{P} runs over all $N!$ permutations of the N electrons. Application of \hat{A} to a product function does not alter the occupancy of the functions $\{\phi_{kj}\}$ in $\{\Psi_k^0\}$, it simply scrambles the order which the electrons occupy the $\{\phi_{kj}\}$ and it causes the resultant function (which is often denoted $|\phi_{k1}(\mathbf{r}_1)\phi_{k2}(\mathbf{r}_2)\phi_{k3}(\mathbf{r}_3) \dots \phi_{kN-1}(\mathbf{r}_{N-1})\phi_{kN}(\mathbf{r}_N)|$ and called a Slater determinant) to obey the Pauli exclusion principle.

Because the electrons also possess intrinsic spin, the one-electron functions $\{\phi_j\}$ used in this construction are taken to

-10-

be eigenfunctions of $(\hat{T}_1 + \hat{V}_{\text{mf}}(\mathbf{r}_1))$ multiplied by either an α or β spin function. This set of functions is called the set of mean-field *spin orbitals*.

By choosing to place N electrons into N specific spin orbitals, one specifies a *configuration*. By making other choices of which $N\phi_j$ to occupy, one describes other configurations. Just as the one-electron mean-field Schrödinger equation has a complete set of spin-orbital solutions $\{\phi_j$ and $\epsilon_j\}$, the N -electron mean-field Schrödinger equation has a complete set of antisymmetric N -electron Slater determinants. When these determinants are combined to generate functions that are eigenfunctions of the total S^2 and S_z and eigenfunctions of the molecule's point group symmetry (or 2 and z for atoms), one has what are called *configuration state functions* (CSFs) whose mean-field energies are also given by .

(B) THE SELF-CONSISTENT MEAN-FIELD (SCF) POTENTIAL

The one-electron additivity of the mean-field Hamiltonian \hat{H}^0 gives rise to the concept of spin orbitals for *any* additive $V_{\text{mf}}(\mathbf{r})$. In fact, there is no *single* mean-field potential; different scientists have put forth different suggestions for V_{mf} over the years. Each gives rise to spin orbitals and configurations that are specific to the particular V_{mf} . However, if the difference between any particular mean-field model and the full electronic

Hamiltonian is fully treated, corrections to all mean-field results should converge to the same set of exact states. In practice, one is never able to treat *all* corrections to any mean-field model. Thus, it is important to seek particular mean-field potentials for which the corrections are as small and straightforward to treat as possible.

In the most commonly employed mean-field models [25] of electronic structure theory, the configuration specified for study plays a central role in defining the mean-field potential. For example, the mean-field Coulomb potential felt by a $2p_x$ orbital's electron at a point \mathbf{r} in the $1s^2 2s^2 2p_x 2p_y$ configuration description of the carbon atom is:

The above mean-field potential is used to find the $2p_x$ orbital of the carbon atom, which is then used to define the mean-field potential experienced by, for example, an electron in the $2s$ orbital:

Notice that the orbitals occupied in the configuration under study appear in the mean-field potential. However, it is \hat{V}_{mf} that, through the one-electron Schrödinger equation, determines the orbitals. For these reasons, the solution of these

-11-

equations must be carried out in a so-called SCF manner. One begins with an approximate description of the orbitals in $\{\Psi_k^0\}$. These orbitals then define \hat{V}_{mf} and the equations $(\hat{T}_1 + \hat{V}_{\text{mf}}(\mathbf{r}_1))\phi_j(\mathbf{r}_1) = \epsilon_j \phi_j(\mathbf{r}_1)$ are solved for 'new' spin orbitals. These orbitals are then be used to define an improved \hat{V}_{mf} which gives another set of solutions to $(\hat{T}_1 + \hat{V}_{\text{mf}}(\mathbf{r}_1))\phi_j(\mathbf{r}_1) = \epsilon_j \phi_j(\mathbf{r}_1)$. This iterative process is continued until the orbitals used to define \hat{V}_{mf} are identical to those that result as solutions of $(\hat{T}_1 + \hat{V}_{\text{mf}}(\mathbf{r}_1))\phi_j(\mathbf{r}_1) = \epsilon_j \phi_j(\mathbf{r}_1)$. When this condition is reached, one has achieved 'self-consistency'.

B3.1.3.2 WHAT IS ELECTRON CORRELATION?

By expressing the mean-field interaction of an electron at \mathbf{r} with the $N - 1$ other electrons in terms of a probability density $\rho_{N-1}(\mathbf{r}')$ that is independent of the fact that another electron resides at \mathbf{r} , the mean-field models ignore spatial *correlations* among the electrons. In reality, as shown in [figure B3.1.5](#) the conditional probability density for finding one of $N - 1$ electrons at \mathbf{r}' , given that one electron is at \mathbf{r} depends on \mathbf{r} . The absence of a spatial correlation is a direct consequence of the spin-orbital *product nature* of the mean-field wavefunctions $\{\Psi_k^0\}$.

To improve upon the mean-field picture of electronic structure, one must move beyond the single-configuration approximation. It is essential to do so to achieve higher accuracy, but it is also important to do so to achieve a *conceptually* correct view of the chemical electronic structure. Although the picture of configurations in which N electrons occupy N spin orbitals may be familiar and useful for systematizing the electronic states of atoms and molecules, these constructs are approximations to the true states of the system. They were introduced when the mean-field approximation was made, and neither orbitals nor configurations can be claimed to describe the proper eigenstates $\{\Psi_k, E_k\}$. It is thus inconsistent to insist that the carbon atom be thought of as $1s^2 2s^2 2p^2$ while insisting on a description of this atom accurate to ± 1 kcal mol⁻¹.

B3.1.3.3 SUMMARY

The SCF mean-field potential takes care of 'most' of the interactions among the N electrons. However, for all

mean-field potentials proposed to date, the residual or fluctuation potential is large enough to require significant corrections to the mean-field picture. This, in turn, necessitates the use of more sophisticated and computationally taxing techniques (e.g., high-order perturbation theory or large variational expansion spaces) to reach the desired chemical accuracy.

For electronic structures of atoms and molecules, the SCF model requires quite substantial corrections to bring its predictions in line with experimental fact. Electrons in atoms and molecules undergo dynamical motions in which their Coulomb repulsions cause them to ‘avoid’ one another at every instant of time, not only in the average-repulsion manner of mean-field models. The inclusion of *dynamical correlations* among electrons is necessary to achieve a more accurate description of atomic and molecular electronic structure. No single spin-orbital product wavefunction is capable of treating electron correlation to *any* extent; its product nature renders it incapable of doing so.

-12-

B3.1.4 HOW TO INTRODUCE ELECTRON CORRELATION VIA CONFIGURATION MIXING

B3.1.4.1 THE MULTI-CONFIGURATION WAVEFUNCTION

In most of the commonly used *ab initio* quantum chemical methods [26], one forms a set of configurations by placing N electrons into spin orbitals in a manner that produces the spatial, spin and angular momentum symmetry of the electronic state of interest. The correct wavefunction Ψ is then written as a linear combination of the mean-field configuration functions $\{\Psi_k\}$: $\Psi = \sum_k C_k \Psi_k^0$. For example, to describe the

ground 1S state of the Be atom, the $1s^2 2s^2$ configuration is augmented by including other configurations such as $1s^2 3s^2$, $1s^2 2p^2$, $1s^2 3p^2$, $1s^2 2s 3s$, $3s^2 2s^2$, $2p^2 2s^2$, etc, all of which have overall 1S spin and angular momentum symmetry. The various methods of electronic structure theory differ primarily in how they determine the $\{C_k\}$ expansion coefficients and how they extract the energy E corresponding to this Ψ .

B3.1.4.2 THE PHYSICAL MEANING OF MIXING IN ‘EXCITED’ CONFIGURATIONS

When considering the ground 1S state of the Be atom, the following four antisymmetrized spin-orbital products are found to have the largest C_k amplitudes:

$$\Psi \cong C_1 |1s^2 2s^2| - C_2 [|1s^2 2p_x^2| + |1s^2 2p_y^2| + |1s^2 2p_z^2|].$$

The fact that the latter three terms possess the same amplitude C_2 is a result of the requirement that a state of 1S symmetry is desired. It can be shown [27] that this function is equivalent to

$$\begin{aligned} \Psi \cong \frac{1}{6} C_1 |1s\alpha 1s\beta| & \{ [(2s - a2p_x)\alpha(2s + a2p_x)\beta - (2s - a2p_x)\beta(2s + a2p_x)\alpha] \\ & + [(2s - a2p_y)\alpha(2s + a2p_y)\beta - (2s - a2p_y)\beta(2s + a2p_y)\alpha] \\ & + [(2s - a2p_z)\alpha(2s + a2p_z)\beta - (2s - a2p_z)\beta(2s + a2p_z)\alpha] \} \end{aligned}$$

where $a = \sqrt{3C_2/C_1}$.

Here two electrons occupy the $1s$ orbital (with opposite, α and β spins) while the other electron pair resides in $2s$ - $2p$ polarized orbitals in a manner that instantaneously correlates their motions. These *polarized orbital*

pairs ($2s \pm a2p_{x,y \text{ or } z}$) are formed by combining the 2s orbital with the $2p_{x,y \text{ or } z}$ orbital in a ratio determined by C_2/C_1 . This way of viewing an electron pair correlation forms the basis of the generalized valence bond (GVB) method that Professor Bill Goddard [28] pioneered.

This ratio C_2/C_1 can be shown to be proportional to the magnitude of the coupling $\langle 1s^2 2s^2 | \hat{H} | 1s^2 2p^2 \rangle$ between the two

-13-

configurations involved and inversely proportional to the energy difference ($\langle 1s^2 2s^2 | \hat{H} | 1s^2 2s^2 \rangle - \langle 1s^2 2p^2 | \hat{H} | 1s^2 2p^2 \rangle$) between these configurations. In general, configurations that have similar Hamiltonian expectation values and that are coupled strongly give rise to strongly mixed (i.e. with large $|C_2/C_1|$ ratios) polarized orbital pairs.

A set of polarized orbital pairs is described pictorially in figure B3.1.6. In each of the three equivalent terms in the above wavefunction, one of the valence electrons moves in a $2s+a2p$ orbital polarized in one direction while the other valence electron moves in the $2s-a2p$ orbital polarized in the opposite direction. For example, the first term $(2s-a2p_x)\alpha(2s+a2p_x)\beta - (2s-a2p_x)\beta(2s+a2p_x)\alpha$ describes one electron occupying a $2s-a2p_x$ polarized orbital while the other electron occupies the $2s+a2p_x$ orbital. The electrons thus reduce their Coulomb repulsion by occupying *different* regions of space; in the SCF picture $1s^2 2s^2$, both electrons reside in the same 2s region of space. In this particular example, the electrons undergo *angular correlation* to ‘avoid’ one another.

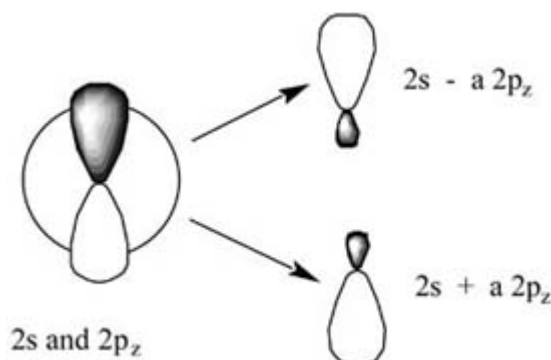


Figure B3.1.6. Polarized orbital pairs involving 2s and $2p_z$ orbitals.

Let us consider another example. In describing the π^2 electron pair of an olefin, it is important to mix in ‘doubly excited’ configurations of the form $(\pi^*)^2$. The physical importance of such configurations can again be made clear by using the identity

$$\begin{aligned}
 & C_1 | \dots \phi \alpha \phi \beta \dots | - C_2 | \dots \phi' \alpha \phi' \beta \dots | \\
 &= C_1 / 2 \{ | \dots (\phi - x \phi') \alpha (\phi + x \phi') \beta \dots | - | \dots (\phi - x \phi') \beta (\phi + x \phi') \alpha \dots | \}
 \end{aligned}$$

where $x = (C_2/C_1)^{1/2}$.

In this example, the two non-orthogonal ‘polarized orbital pairs’ involve mixing the π and π^* orbitals to produce two left–right polarized orbitals as depicted in figure B3.1.7. Here one says that the π^2 electron pair undergoes left–right correlation when the $(\pi^*)^2$ configuration is introduced.

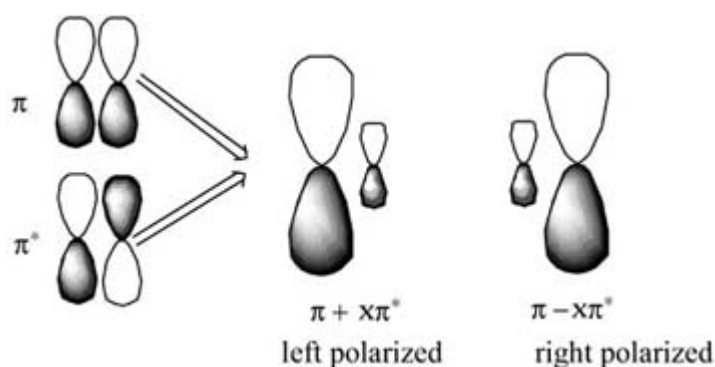


Figure B3.1.7. Left- and right-polarized orbital pairs involving π and π^* orbitals.

B3.1.4.3 ARE POLARIZED ORBITAL PAIRS HYBRID ORBITALS?

It should be stressed that these polarized orbital pairs are *not* the same as hybrid orbitals. The latter are used to describe directed bonding, but polarized orbital *pairs* are each a ‘mixture’ of two mean-field orbitals with amplitude $x = (C_2/C_1)^{1/2}$ and with a *single electron* in each, thereby allowing the electrons to be spatially correlated and to ‘avoid’ one another. In addition, polarized orbital pairs are *not generally orthogonal* to one another; hybrid orbital sets are.

B3.1.4.4 RELATIONSHIP TO THE GENERALIZED VALENCE BOND PICTURE

In these examples, the analysis allows one to *interpret* the combination of pairs of configurations that differ from one another by a ‘double excitation’ from one orbital (ϕ) to another (ϕ') as equivalent to a singlet coupling of two polarized orbitals ($\phi - a\phi'$) and ($\phi + a\phi'$). As mentioned earlier, this picture is closely related to the GVB model that Goddard [28] and Goddard and Harding [29] developed. In the simplest embodiment of the GVB model, each electron pair in the atom or molecule is correlated by mixing in a configuration in which that pair is ‘doubly excited’ to a correlating orbital. The direct product of all such pair correlations generates the simplest GVB-type wavefunction.

In most *ab initio* quantum chemical methods, the correlation calculation is actually carried out by forming a linear combination of the mean-field configuration state functions and determining the $\{C_k\}$ amplitudes by some procedure. The identities discussed in some detail above are then introduced merely to permit one to interpret the presence of configurations that are ‘doubly excited’ relative to the dominant mean-field configuration in terms of polarized orbital pairs.

B3.1.4.5 SUMMARY

The dynamical interactions among electrons give rise to instantaneous spatial correlations that must be handled to arrive at an accurate picture of the atomic and molecular structure. The single-configuration picture provided by the mean-field model is a useful starting point, but it is *incapable* of describing electron correlations. Therefore, improvements are needed. The use of doubly-excited configurations is a mechanism by which Ψ can place electron *pairs*, which in the mean-field picture occupy the same orbital, into different regions of space thereby lowering their mutual Coulombic repulsions. Such electron correlation effects are referred to as *dynamical electron correlation*; they are extremely important to include if one expects to achieve chemically meaningful accuracy.

B3.1.5 THE SINGLE-CONFIGURATION PICTURE AND THE HF APPROXIMATION

Given a set of N -electron space- and spin-symmetry-adapted configuration state functions $\{\Phi_j\}$ in terms of which Ψ is to be expanded as $\Psi = \sum_j C_j \Phi_j$, two primary questions arise: (1) how to determine the $\{C_j\}$ coefficients and the energy E and (2) how to find the ‘best’ spin orbitals $\{\phi_j\}$? Let us first consider the case where a single configuration is used so only the question of determining the spin orbitals exists.

B3.1.5.1 THE SINGLE-DETERMINANT WAVEFUNCTION

(A) THE CANONICAL SCF EQUATIONS

The simplest trial function employed in *ab initio* quantum chemistry is the single Slater determinant function in which N spin orbitals are occupied by N electrons:

$$\Psi = |\phi_1 \phi_2 \phi_3 \dots \phi_N|.$$

For such a function, variational optimization of the spin orbitals to make the expectation value $\langle \Psi | \hat{H} | \Psi \rangle$ stationary produces [30] the canonical HF equations

$$\hat{F} \phi_i = \varepsilon_i \phi_i$$

where the so-called Fock operator \hat{F} is given by

$$\hat{F} \phi_i = \hat{h} \phi_i + \sum_{j(\text{occupied})} [\hat{J}_j - \hat{K}_j] \phi_i.$$

The Coulomb (\hat{J}_j) and exchange (\hat{K}_j) operators are defined by the relations

$$\hat{J}_j \phi_i = \int \phi_j^*(r') \phi_j(r') / |r - r'| d\tau' \phi_i(r)$$

and

$$\hat{K}_j \phi_i = \int \phi_j^*(r') \phi_i(r') / |r - r'| d\tau' \phi_j(r)$$

the symbol \hat{h} denotes the sum of the electronic kinetic energy, and electron–nuclear Coulomb attraction operators. The $d\tau$ implies integration over the spin variables associated with the ϕ_i (and, for the exchange

operator, ϕ_i), as a result of which the exchange integral vanishes unless the spin function of ϕ_j is the same as that of ϕ_i ; the Coulomb integral is non-vanishing no matter what the spin functions of ϕ_j and ϕ_i .

(B) THE EQUATIONS HAVE ORBITAL SOLUTIONS FOR OCCUPIED AND UNOCCUPIED ORBITALS

The HF [31] equations $\hat{F}\phi_i = \varepsilon_i\phi_i$ possess solutions for the spin orbitals in Ψ (the *occupied* spin orbitals) as well as for orbitals not occupied in Ψ (the *virtual* spin orbitals) because the \hat{F} operator is Hermitian. Only the ϕ_i occupied in Ψ appear in the Coulomb and exchange potentials of the Fock operator.

(C) THE SPIN-IMPURITY PROBLEM

As formulated above, the HF equations yield orbitals that do not guarantee that Ψ has proper spin symmetry. To illustrate, consider an open-shell system such as the lithium atom. If $1s\alpha$, $1s\beta$, and $2s\alpha$ spin orbitals are chosen to appear in Ψ , the Fock operator will be

$$\hat{F} = \hat{h} + \hat{J}_{1s\alpha} + \hat{J}_{1s\beta} + \hat{J}_{2s\alpha} - [\hat{K}_{1s\alpha} + \hat{K}_{1s\beta} + \hat{K}_{2s\alpha}].$$

Acting on an α spin orbital $\phi_{k\alpha}$ with F and carrying out the spin integrations, one obtains

$$\hat{F}\phi_{k\alpha} = \hat{h}\phi_{k\alpha} + (2\hat{J}_{1s} + \hat{J}_{2s})\phi_{k\alpha} - (\hat{K}_{1s} + \hat{K}_{2s})\phi_{k\alpha}.$$

In contrast, when acting on a β spin orbital, one obtains

$$\hat{F}\phi_{k\beta} = \hat{h}\phi_{k\beta} + (2\hat{J}_{1s} + \hat{J}_{2s})\phi_{k\beta} - (\hat{K}_{1s})\phi_{k\beta}.$$

Spin orbitals of α and β type do *not* experience the same exchange potential in this model because Ψ contains two α spin orbitals and only one β spin orbital. A consequence is that the optimal $1s\alpha$ and $1s\beta$ spin orbitals, which are themselves solutions of $\hat{F}\phi_i = \varepsilon_i\phi_i$, do not have identical orbital energies (i.e. $\varepsilon_{1s\alpha} \neq \varepsilon_{1s\beta}$) and are not spatially identical. This resultant spin polarization of the orbitals gives rise to *spin impurities* in Ψ . The determinant $|1s\alpha 1s'\beta 2s\alpha|$ is not a

-17-

pure doublet spin eigenfunction, although it is an S_z eigenfunction with $M_s = 1/2$; it contains both $S = 1/2$ and $S = 3/2$ components. If the $1s\alpha$ and $1s'\beta$ spin orbitals were spatially identical, then $|1s\alpha 1s'\beta 2s\alpha|$ would be a pure spin eigenfunction with $S = 1/2$.

The above single-determinant wavefunction is referred to as being of the unrestricted Hartree–Fock (UHF) type because no restrictions are placed on the spatial nature of the orbitals in Ψ . In general, UHF wavefunctions are not of pure spin symmetry for any open-shell system or for closed-shell systems far from their equilibrium geometries (e.g. for H_2 or N_2 at long bond lengths) These are significant drawbacks of methods based on a UHF starting point. Such a UHF treatment forms the basis of the widely used and highly successful Gaussian 70 through Gaussian-9X series of electronic structure computer codes [32] which derive from Pople [32] and co-workers.

To overcome some of the problems inherent in the UHF method, it is possible to derive SCF equations based on minimizing the energy of a wavefunction formed by spin projecting a single Slater determinant starting

function (e.g. using $\{ |1s\alpha\ 2s\beta\rangle - |1s\beta\ 2s\alpha\rangle \} / 2^{1/2}$ for the singlet excited state of He rather than $|1s\alpha\ 2s\beta\rangle$). It is also possible for a trial wavefunction of the form $|1s\alpha\ 1s\beta\ 2s\alpha\rangle$ to constrain the $1s\alpha$ and $1s\beta$ orbitals to have exactly the same spatial form. In both cases, one then is able to carry out what are called restricted Hartree–Fock (RHF) calculations.

B3.1.5.2 THE LINEAR COMBINATIONS OF ATOMIC ORBITALS TO FORM MOLECULAR ORBITALS EXPANSION OF THE SPIN ORBITALS

The HF equations must be solved iteratively because the J_i and K_i operators in F depend on the orbitals ϕ_i for which solutions are sought. Typical iterative schemes begin with a ‘guess’ for those ϕ_i that appear in Ψ , which then allows \hat{F} to be formed. Solutions to $\hat{F}\phi_i = \varepsilon_i\phi_i$ are then found, and those ϕ_i which possess the space and spin symmetry of the occupied orbitals of Ψ and which have the proper energies and nodal character are used to generate a new \hat{F} operator (i.e. new \hat{J}_i and \hat{K}_i operators). This iterative HF SCF process is continued until the ϕ_i and ε_i do not vary significantly from one iteration to the next, at which time one says that the process has converged.

In practice, solution of $\hat{F}\phi_i = \varepsilon_i\phi_i$ as an integro-differential equation can be carried out only for atoms [34] and linear molecules [35] for which the angular parts of the ϕ_i can be exactly separated from the radial because of axial- or full-rotation group symmetry (e.g. $\phi_i = Y_{l,m}(\theta, \phi)R_{n,l}(r)$ for an atom and $\phi_i = \exp(im\phi)R_{n,l,m}(\rho, z)$ for a linear molecule).

In the procedures most commonly applied to nonlinear molecules, the ϕ_i are expanded in a *basis* χ_μ according to the linear combinations of AOs to form molecular orbitals (LCAO–MO) [36] procedure:

$$\phi_i = \sum_{\mu} C_{\mu,i} \chi_{\mu}$$

-18-

This reduces $\hat{F}\phi_i = \varepsilon_i\phi_i$ to a matrix eigenvalue-type equation:

$$\sum_{\nu} F_{\mu,\nu} C_{\nu,i} = \varepsilon_i \sum_{\nu} S_{\mu,\nu} C_{\nu,i}$$

where $S_{\mu,\nu} = \langle \chi_{\mu} | \chi_{\nu} \rangle$ is the overlap matrix among the AOs and

$$F_{\mu,\nu} = \langle \chi_{\mu} | \hat{h} | \chi_{\nu} \rangle + \sum_{\delta,\kappa} [\gamma_{\delta,\kappa} \langle \chi_{\mu} \chi_{\delta} | \hat{g} | \chi_{\nu} \chi_{\kappa} \rangle - \gamma_{\delta,\kappa}^{\text{ex}} \langle \chi_{\mu} \chi_{\delta} | \hat{g} | \chi_{\kappa} \chi_{\nu} \rangle]$$

is the matrix representation of the Fock operator in the AO basis. Here and elsewhere, the symbol \hat{g} is used to represent the electron–electron Coulomb potential $e^2/|\mathbf{r} - \mathbf{r}'|$.

The charge- and exchange-density matrix elements in the AO basis are:

$$\gamma_{\delta,\kappa} = \sum_{i(\text{occupied})} C_{\delta,i} C_{\kappa,i}$$

and

$$\gamma_{\delta,\kappa}^{\text{ex}} = \sum_{i(\text{occupied and same spin})} C_{\delta,i} C_{\kappa,i}$$

where the sum in $\gamma_{\delta,\kappa}^{\text{ex}}$ runs over those occupied spin orbitals whose m_s value is equal to that for which the Fock matrix is being formed (for a closed-shell species, $\gamma_{\delta,\kappa}^{\text{ex}} = 1/2\gamma_{\delta,\kappa}$).

It should be noted that by moving to a matrix problem, one does not remove the need for an iterative solution; the $F_{\mu,\nu}$ matrix elements depend on the $C_{\nu,i}$ LCAO–MO coefficients which are, in turn, solutions of the so-called Roothaan [30] matrix HF equations: $\sum_{\nu} F_{\mu,\nu} C_{\nu,i} = \epsilon_i \sum_{\nu} S_{\mu,\nu} C_{\nu,i}$. One should also note that, just as $\hat{F}\phi_i = \epsilon_i \phi_i$ possesses a complete set of eigenfunctions, the matrix $F_{\mu,\nu}$, whose dimension M is equal to the number of atomic basis orbitals, has M eigenvalues ϵ_j and M eigenvectors whose elements are the $C_{\nu,i}$. Thus, there are *occupied and virtual* MOs each of which is described in the LCAO–MO form with the $C_{\nu,i}$ coefficients obtained via solution of $\sum_{\nu} F_{\mu,\nu} C_{\nu,i} = \epsilon_i \sum_{\nu} S_{\mu,\nu} C_{\nu,i}$.

B3.1.5.3 AO BASIS SETS

(A) SLATER-TYPE ORBITALS AND GAUSSIAN-TYPE ORBITALS

The basis orbitals commonly used in the LCAO–MO process fall into two primary classes:

- (1) Slater-type orbitals (STOs) $\chi_{n,l,m}(r, \theta, \phi) = N_{n,l,m,\zeta} Y_{l,m}(\theta, \phi) r^{n-1} \exp(-\zeta r)$, are characterized by the quantum numbers n , l and m and the exponent (which characterizes the ‘size’) ζ . The symbol $N_{n,l,m,\zeta}$ denotes the normalization constant.
- (2) Cartesian Gaussian-type orbitals (GTOs) $\chi_{a,b,c}(r, \theta, \phi) = N'_{a,b,c,\alpha} x^a y^b z^c \exp(-\alpha r^2)$, are characterized by the quantum numbers a , b and c , which detail the angular shape and direction of the orbital, and the exponent α which governs the radial ‘size’.

For both types of orbitals, the coordinates r , θ and ϕ refer to the position of the electron relative to a set of axes attached to the centre on which the basis orbital is located. Although STOs have the proper ‘cusp’ behaviour near the nuclei, they are used primarily for atomic- and linear-molecule calculations because the multi-centre integrals which arise in polyatomic-molecule calculations cannot efficiently be performed when STOs are employed. In contrast, such integrals can routinely be done when GTOs are used. This fundamental advantage of GTOs has led to the dominance of these functions in molecular quantum chemistry.

To overcome the primary weakness of GTO functions (i.e. their radial derivatives vanish at the nucleus whereas the derivatives of STOs are non-zero), it is common to combine two, three, or more GTOs, with combination coefficients which are fixed and *not* treated as LCAO–MO parameters, into new functions called *contracted* GTOs or CGTOs. Typically, a series of tight, medium, and loose GTOs are multiplied by *contraction coefficients* and summed to produce a CGTO, which approximates the proper ‘cusp’ at the nuclear centre.

Although most calculations on molecules are now performed using Gaussian orbitals (STOs are still commonly employed in atomic calculations), it should be noted that other basis sets can be used as long as they span enough of the region of space (radial and angular) where significant electron density resides. In fact,

it is possible to use plane wave orbitals [37] of the form $\chi(r, \theta, \phi) = N \exp[i(k_x r \sin\theta \cos\phi + k_y r \sin\theta \sin\phi + k_z r \cos\theta)]$, where N is a normalization constant and k_x , k_y , and k_z are the quantum numbers detailing the momenta of the orbital along the x , y and z Cartesian directions. The advantage to using such ‘simple’ orbitals is that the integrals one must perform are much easier to handle with such functions; the disadvantage is that one must use many such functions to accurately describe sharply peaked charge distributions of, for example, inner-shell core orbitals.

(B) BASIS SET LIBRARIES

Much effort has been devoted to developing sets of STO or GTO basis orbitals for main-group elements and the lighter transition metals. This ongoing effort is aimed at providing standard basis set libraries which:

- (1) yield predictable chemical accuracy in the resultant energies;
- (2) are cost effective to use in practical calculations;

-20-

- (3) are relatively transferable so that a given atom’s basis is flexible enough to be used for that atom in various bonding environments.

The fundamental core and valence basis. In constructing an AO basis, one can choose from among several classes of functions. First, the size and nature of the primary core and valence basis must be specified. Within this category, the following choices are common.

- (1) A *minimal basis* in which the number of STO or CGTO orbitals is equal to the number of core and valence AOs in the atom.
- (2) A *double-zeta (DZ)* basis in which twice as many STOs or CGTOs are used as there are core and valence AOs. The use of more basis functions is motivated by a desire to provide additional variational flexibility so the LCAO–MO process can generate MOs of variable diffuseness as the local electronegativity of the atom varies.
- (3) A *triple-zeta (TZ)* basis in which three times as many STOs or CGTOs are used as the number of core and valence AOs (and, yes, there now are quadruple-zeta (QZ) and higher-zeta basis sets appearing in the literature).
- (4) Dunning and Dunning and Hay [38] developed CGTO bases which range from approximately DZ to substantially beyond QZ quality. These bases involve contractions of primitive uncontracted GTO bases which Huzinaga [39] had earlier optimized. These Dunning bases are commonly denoted as follows for first-row atoms: (10s,6p/5s,4p), which means that 10 s-type primitive GTOs have been contracted to produce five separate s-type CGTOs and that six primitive p-type GTOs were contracted into four separate p-type CGTOs in each of the x , y and z directions.
- (5) Even-tempered basis sets [40] consist of GTOs in which the orbital exponents α_k belonging to series of orbitals consist of geometrical progressions: $\alpha_k = a\beta^k$, where a and β characterize the particular set of GTOs.
- (6) STO-3G bases [41] were employed some years ago, but have recently become less popular. These bases are constructed by least-squares fitting GTOs to STOs which have been optimized for various electronic states of the atom. When three GTOs are employed to fit each STO, a STO-3G basis is formed.
- (7) 4-31G, 5-31G and 6-31G bases [42] employ a single CGTO of contraction length 4, 5, or 6 to describe the core orbital. The valence space is described at the DZ level with the first CGTO constructed from three primitive GTOs and the second CGTO built from a single primitive GTO.
- (8) More recently, the Dunning group has focused on developing basis sets that are optimal not for use in SCF-level calculations on atoms and molecules, but that have been optimized for use in correlated calculations. These so-called correlation-consistent bases [43] are now widely used because more and more *ab initio* calculations are being performed at a correlated level.
- (9) Atomic natural orbital (ANO) basis sets [44] are formed by contracting Gaussian functions so as to reproduce the natural orbitals obtained from correlated (usually using a configuration interaction with

single and double excitation (CISD) level wavefunction) calculations on atoms.

Optimization of the *orbital exponents* (ζ s or α s) and the GTO-to-CGTO *contraction coefficients* for the kind of bases described above have undergone explosive growth in recent years. As a result, it is not possible to provide a single or even a few literature references from which one can obtain the most up-to-date bases. However, the theory group at the Pacific Northwest National Laboratories (PNNL) offer a webpage [45] from which one can find (and even download in a form prepared for input to any of several commonly used electronic structure codes) a wide variety of Gaussian atomic basis sets.

Polarization functions. One usually enhances any core and valence functions with a set of so-called polarization

-21-

functions. They are functions of one higher angular momentum than appears in the atom's valence orbital space (e.g. d-functions for C, N and O and p-functions for H), and they have exponents (ζ or α) which cause their radial sizes to be similar to the sizes of the valence orbitals (i.e. the polarization p orbitals of the H atom are similar in size to the 1s orbital). Thus, they are *not* orbitals which describe the atom's valence orbital with one higher l value; such higher- l valence orbitals would be radially more diffuse.

The primary purpose of the polarization functions is to give additional angular flexibility to the LCAO–MO process in forming the valence MOs. This is illustrated below in figure B3.1.8 where polarization d_{π} orbitals are seen to contribute to formation of the bonding π orbital of a carbonyl group by allowing polarization of the carbon atom's p_{π} orbital toward the right and of the oxygen atom's p_{π} orbital toward the left.

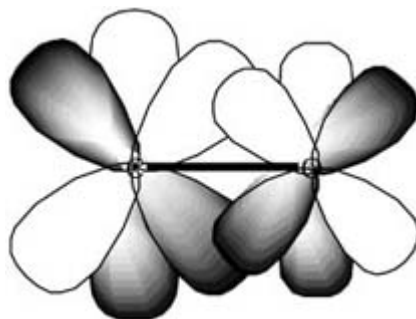


Figure B3.1.8. The role of d-polarization functions in the π bond between C and O.

The polarization functions are essential in strained ring compounds because they provide the angular flexibility needed to direct the electron density into the regions between the bonded atoms.

Functions with higher l values and with 'sizes' like those of lower- l valence orbitals are also used to introduce additional angular correlation by permitting angularly polarized orbital pairs to be formed. Optimal polarization functions for first- and second-row atoms have been tabulated and are included in the PNNL Gaussian orbital web site data base [45].

Diffuse functions. When dealing with anions or Rydberg states, one must further *augment* the basis set by adding so-called diffuse basis orbitals. The valence and polarization functions described above do not provide enough radial flexibility to adequately describe either of these cases. Once again, the PNNL web site data base [45] offers a good source for obtaining diffuse functions appropriate to a variety of atoms.

Once one has specified an AO basis for each atom in the molecule, the LCAO–MO procedure can be used to

determine the $C_{v,i}$ coefficients that describe the occupied and virtual orbitals. It is important to keep in mind that the basis orbitals are *not* themselves the SCF orbitals of the isolated atoms; even the proper AOs are combinations (with atomic values for the $C_{v,i}$ coefficients) of the basis functions. The LCAO–MO–SCF process itself determines the magnitudes and signs of the $C_{v,i}$; alternations in the signs of these coefficients allow radial nodes to form.

-22-

B3.1.5.4 THE PHYSICAL MEANING OF ORBITAL ENERGIES

The HF–SCF equations $\hat{F}\phi_i = \varepsilon_i\phi_i$ imply that ε_i can be written as

$$\begin{aligned}\varepsilon_i &= \langle \phi_i | \hat{F} | \phi_i \rangle = \langle \phi_i | \hat{h} | \phi_i \rangle + \sum_{j(\text{occupied})} \langle \phi_i | \hat{J}_j - \hat{K}_j | \phi_i \rangle \\ &= \langle \phi_i | \hat{h} | \phi_i \rangle + \sum_{j(\text{occupied})} [J_{i,j} - K_{i,j}].\end{aligned}$$

Thus ε_i is the average value of the kinetic energy plus the Coulombic attraction to the nuclei for an electron in ϕ_i plus the sum over all of the spin orbitals occupied in Ψ of the Coulomb minus exchange interactions. If ϕ_i is an occupied spin orbital, the term $[J_{i,i} - K_{i,i}]$ disappears and the latter sum represents the Coulomb minus exchange interaction of ϕ_i with all of the $N - 1$ *other* occupied spin orbitals. If ϕ_i is a virtual spin orbital, this cancellation does not occur, and one obtains the Coulomb minus exchange interaction of ϕ_i with all N of the occupied spin orbitals.

Hence the orbital energies of *occupied* orbitals pertain to interactions appropriate to a total of N electrons, while the orbital energies of *virtual* orbitals pertain to a system with $N + 1$ electrons. This usually makes SCF virtual orbitals not very good for use in subsequent correlation calculations or for use in interpreting electronic excitation processes. To correlate a pair of electrons that occupy a valence orbital requires double excitations into a virtual orbital of similar size; the SCF virtual orbitals are too diffuse. For this reason, significant effort has been devoted to developing methods that produce so-called ‘improved virtual orbitals’ (IVOs) [46] that are of more utility in performing correlated calculations.

(A) KOOPMANS’ THEOREM

Let us consider a model of the vertical (i.e. at fixed molecular geometry) detachment or attachment of an electron to an N -electron molecule.

- (1) In this model, *both* the parent molecule and the species generated by adding or removing an electron are treated at the single-determinant level.
- (2) The HF orbitals of the parent molecule are used to describe both species. It is said that such a model neglects ‘*orbital relaxation*’ (i.e. the reoptimization of the spin orbitals to allow them to become appropriate to the daughter species).

Within this model, the energy difference between the daughter and the parent can be written as follows (ϕ_k represents the particular spin orbital that is added or removed):

- (1) For electron detachment

$$E^{N-1} - E^N = -\varepsilon_k.$$

(2) For electron attachment

$$E^N - E^{N+1} = -\varepsilon_k.$$

-23-

So, within the limitations of the single-determinant, frozen-orbital model, the ionization potentials (IPs) and electron affinities (EAs) are given as the negative of the occupied and virtual spin-orbital energies, respectively. This statement is referred to as *Koopmans' theorem* [47]; it is used extensively in quantum chemical calculations as a means for estimating IPs and EAs and often yields results that are qualitatively correct (i.e., ± 0.5 eV).

(B) ORBITAL ENERGIES AND THE TOTAL ENERGY

The total SCF electronic energy can be written as

$$E = \sum_{i(\text{occupied})} \langle \phi_i | \hat{h} | \phi_i \rangle + \sum_{i>j(\text{occupied})} [J_{i,j} - K_{i,j}]$$

and the sum of the orbital energies of the occupied spin orbitals is given by

$$\sum_{i(\text{occupied})} \varepsilon_i = \sum_{i(\text{occupied})} \langle \phi_i | \hat{h} | \phi_i \rangle + \sum_{i,j(\text{occupied})} [J_{i,j} - K_{i,j}].$$

These two expressions differ in a very important way; the sum of occupied orbital energies double counts the Coulomb minus exchange interaction energies. Thus, within the HF approximation, the sum of the occupied orbital energies is *not* equal to the total energy.

B3.1.5.5 SOLVING THE Roothaan SCF Equations

Before moving on to discuss methods that go beyond the single-configuration mean-field model, it is important to examine some of the computational effort that goes into carrying out an SCF calculation.

Once atomic basis sets have been chosen for each atom, the *one- and two-electron integrals* appearing in $F_{\mu,\nu}$ must be evaluated. There are numerous, highly-efficient computer codes [48] which allow such integrals to be computed for s, p, d, f and even g, h and i basis functions. After executing one of these 'integral packages' for a basis with a total of P functions, one has available (usually on the computer's hard disk) of the order of $P^2/2$ one-electron ($\langle \chi_{\mu} | \hat{h} | \chi_{\nu} \rangle$ and $\langle \chi_{\mu} | \chi_{\nu} \rangle$) and $P^4/8$ two-electron ($\langle \chi_{\mu} \chi_{\delta} | \hat{g} | \chi_{\nu} \chi_{\kappa} \rangle$) integrals. When treating extremely large AO basis sets (e.g. 1000 or more basis functions), modern computer programs [49] calculate the requisite integrals, but never store them on the disk. Instead, their contributions to $F_{\mu,\nu}$ are accumulated 'on the fly' after which the integrals are discarded. Recently, much progress has been made towards achieving an evaluation of the non-vanishing (i.e. numerically significant) integrals [48] as well as solving the subsequent SCF equations in a manner whose effort *scales linearly* [50] with the number of basis functions for large P .

After the requisite integrals are available or are being computed on the fly, to begin the SCF process one must input into the computer routine which computes $F_{\mu,\nu}$ the *initial 'guesses'* for the $C_{\nu,i}$ values corresponding to the occupied

orbitals. These initial guesses are typically made as follows.

- (1) If one has available the $C_{v,i}$ values for the system from a calculation performed at a nearby geometry, one can use these $C_{v,i}$ values.
- (2) If one has $C_{v,i}$ values appropriate to fragments of the system (e.g. for C and O atoms if the CO molecule is under study or for CH₂ and O if H₂CO is being studied), one can use these.
- (3) If one has no other information available, one can carry out one iteration of the SCF process in which the two-electron contributions to $F_{\mu,\nu}$ are ignored (i.e. take $F_{\mu,\nu} = \langle \chi_\mu | h | \chi_\nu \rangle$) and use the resultant solutions to $\sum_\nu F_{\mu,\nu} C_{v,i} = \varepsilon_i \sum_\nu S_{\mu,\nu} C_{v,i}$ as initial guesses.

Once the initial guesses have been made for the $C_{v,i}$ of the occupied orbitals, the full $F_{\mu,\nu}$ matrix is formed and new ε_i and $C_{v,i}$ values are obtained by solving $\sum_\nu F_{\mu,\nu} C_{v,i} = \varepsilon_i \sum_\nu S_{\mu,\nu} C_{v,i}$. These new orbitals are then used to form a new $F_{\mu,\nu}$ matrix from which new ε_i and $C_{v,i}$ are obtained. This iterative process is carried on until the ε_i and $C_{v,i}$ do not vary (within specified tolerances) from iteration to iteration, at which time the SCF process has reached self-consistency.

B3.1.6 METHODS FOR TREATING ELECTRON CORRELATION

B3.1.6.1 AN OVERVIEW OF VARIOUS APPROACHES

There are numerous procedures currently in use for determining the ‘best’ wavefunction of the form

$$\Psi = \sum_I C_I \Phi_I$$

where Φ_I is a spin- and space-symmetry-adapted CSF consisting of determinants $|\phi_{I1}\phi_{I2}\phi_{I3}\dots\phi_{IN}|$ (see [14, 16, 26]). In all such wavefunctions there are two kinds of parameters that need to be determined—the C_I and the LCAO–MO coefficients describing the ϕ_{Ik} . The most commonly employed methods used to determine these parameters include the following.

(A) THE MULTICONFIGURATIONAL SELF-CONSISTENT FIELD METHOD

In this approach [51], the expectation value $\langle \Psi | \hat{H} | \Psi \rangle / \langle \Psi | \Psi \rangle$ is treated variationally and made stationary with respect to variations in the C_I and $C_{v,i}$ coefficients. The energy functional is a quadratic function of the C_I coefficients, and so one can express the stationary conditions for these variables in the secular form

$$\sum_J H_{I,J} C_J = E C_I.$$

However, E is a quartic function of the $C_{v,i}$ s because $H_{I,J}$ involves two-electron integrals $\langle \phi_i \phi_j | \hat{g} | \phi_k \phi_l \rangle$ that depend quartically on these coefficients.

It is well known that minimization of the function (E) of several nonlinear parameters (the $C_{v,i}$) is a difficult task that can suffer from poor convergence and may locate local rather than global minima. In a multiconfigurational self-consistent field (MCSCF) wavefunction containing many CSFs, the energy is only weakly dependent on the orbitals that appear in CSFs with small C_I values; in contrast, E is strongly dependent on those orbitals that appear in the CSFs with larger C_I values. One is therefore faced with minimizing a function of many variables that depends strongly on several of the variables and weakly on many others.

For these reasons, in the MCSCF method the number of CSFs is usually kept to a small to moderate number (e.g. a few to several thousand) chosen to describe *essential correlations* (i.e. configuration crossings, near degeneracies, proper dissociation, etc, all of which are often termed *non-dynamical correlations*) and important dynamical correlations (those electron-pair correlations of angular, radial, left–right, etc nature that are important when low-lying ‘virtual’ orbitals are present).

(B) THE CONFIGURATION INTERACTION METHOD

In this approach [52], the LCAO–MO coefficients are determined first via a single-configuration SCF calculation or an MCSCF calculation using a small number of CSFs. The C_I coefficients are subsequently determined by making the expectation value $\langle \Psi | \hat{H} | \Psi \rangle / \langle \Psi | \Psi \rangle$ stationary.

The CI wavefunction is most commonly constructed from CSFs of Φ_j that include:

- (1) all of the CSFs in the SCF or MCSCF wavefunction used to generate the molecular orbitals ϕ_i . These are referred to as the ‘reference’ CSFs;
- (2) CSFs generated by carrying out single-, double-, triple-, etc, level ‘excitations’ (i.e. orbital replacements) relative to reference CSFs. CI wavefunctions limited to include contributions through various levels of excitation are denoted S (singly), D (doubly), SD (singly and doubly), SDT (singly, doubly, and triply) excited.

The orbitals from which electrons are removed can be restricted to focus attention on the correlations among certain orbitals. For example, if the excitations from the core electrons are excluded, one computes the total energy that contains no core correlation energy. The number of CSFs included in the CI calculation can be far in excess of the number considered in typical MCSCF calculations. CI wavefunctions including 5000 to 50 000 CSFs are routine, and functions with one to several billion CSFs are within the realm of practicality [53].

The need for such large CSF expansions should not be surprising considering (i) that each electron pair requires *at least* two CSFs to form polarized orbital pairs, (ii) there are of the order of $N(N-1)/2 = X$ electron pairs for N electrons, hence (iii) the number of terms in the CI wavefunction scales as 2^X . For a molecule containing ten electrons, there could be $2^{45} = 3.5 \times 10^{13}$ terms in the CI expansion. This may be an overestimate of the number of CSFs needed, but it demonstrates how rapidly the number of CSFs can grow with the number of electrons.

The $H_{I,J}$ matrices are, in practice, evaluated in terms of one- and two-electron integrals over the MOs using the Slater–Condon rules [54] or their equivalent. Prior to forming the $H_{I,J}$ matrix elements, the one- and two-electron integrals,

which can be computed only for the atomic (e.g. STO or GTO) basis, must be transformed [55] to the MO

basis. This transformation step requires computer resources proportional to the fifth power of the number of basis functions, and thus is one of the more troublesome steps in most configuration interaction calculations.

For large CI calculations, the full $H_{I,J}$ matrix is *not* formed and stored in the computer's memory or on disk; rather, 'direct CI' methods [56] identify and compute non-zero $H_{I,J}$ and immediately add up contributions to the sum $\sum_j H_{I,J} C_j$. Iterative methods [57], in which approximate values for the C_j coefficients are refined through sequential application of $\sum_j H_{I,J}$ to the preceding estimate of the C_j vector, are employed to solve these large eigenvalue problems.

(C) THE MØLLER-PLESSET PERTURBATION METHOD

This method [58] uses the single-configuration SCF process to determine a set of orbitals $\{\phi_i\}$. Then, using an unperturbed Hamiltonian equal to the sum of the N electrons' Fock operators $\hat{H}^0 = \sum_{i=1,N} \hat{F}(i)$, perturbation theory is used to determine the C_1 amplitudes for the CSFs. The MPPT procedure [59] is a special case of many-body perturbation theory (MBPT) in which the UHF Fock operator is used to define \hat{H}^0 . The amplitude for the *reference* CSF is taken as unity and the other CSFs' amplitudes are determined by the Rayleigh–Schrödinger perturbation using $\hat{H} - \hat{H}^0$ as the perturbation.

In the MPPT/MBPT method, once the reference CSF is chosen and the SCF orbitals belonging to this CSF are determined, the wavefunction Ψ and energy E are determined in an order-by-order manner. The perturbation equations *determine* what CSFs to include and their particular order. This is one of the primary strengths of this technique; it does not require one to make further choices, in contrast to the MCSCF and CI treatments where one needs to choose which CSFs to include.

For example, the first-order wavefunction correction Ψ^1 is

$$\Psi^1 = - \sum_{i < j, m < n} [\langle i, j | \hat{g} | m, n \rangle - \langle i, j | \hat{g} | n, m \rangle] [\varepsilon_m - \varepsilon_i + \varepsilon_n - \varepsilon_j]^{-1} | \Phi_{i,j}^{m,n} \rangle$$

where the SCF orbital energies are denoted ε_k and $\Phi_{i,j}^{m,n}$ represents a CSF that is *doubly excited* (ϕ_i and ϕ_j are replaced by ϕ_m and ϕ_n) relative to Φ . Only doubly-excited CSFs contribute to the *first-order wavefunction*; the fact that the contributions from singly-excited configurations vanish in Ψ^1 is known as the *Brillouin theorem* [60].

The energy E is given through second order as

$$E = E_{\text{SCF}} - \sum_{i < j, m < n} | \langle i, j | \hat{g} | m, n \rangle - \langle i, j | \hat{g} | n, m \rangle |^2 / [\varepsilon_m - \varepsilon_i + \varepsilon_n - \varepsilon_j].$$

Both Ψ and E are expressed in terms of two-electron integrals $\langle i, j | \hat{g} | m, n \rangle$ coupling the virtual spin orbitals ϕ_m and ϕ_n to the spin orbitals from which the electrons were excited ϕ_i and ϕ_j as well as the orbital energy differences $[\varepsilon_m - \varepsilon_i + \varepsilon_n - \varepsilon_j]$ accompanying such excitations. Clearly, the major contributions to the correlation energy are made by double excitations into virtual orbitals $\phi_m \phi_n$ with large $\langle i, j | \hat{g} | m, n \rangle$ integrals and small orbital energy gaps $[\varepsilon_m - \varepsilon_i + \varepsilon_n - \varepsilon_j]$. In higher-order corrections, contributions from CSFs that are singly, triply, etc excited relative to Φ appear, and additional contributions from the doubly-excited CSFs also

enter.

(D) THE COUPLED-CLUSTER METHOD

In the coupled-cluster (CC) method [61], one expresses the wavefunction in a somewhat different manner:

$$\Psi = \exp(T)\Phi$$

where Φ is a single CSF (usually the UHF determinant) used in the SCF process to generate a set of spin orbitals. The operator \hat{T} is expressed in terms of operators that achieve spin-orbital excitations as follows:

$$T = \sum_{i,m} t_i^m \hat{m}^+ \hat{i} + \sum_{i,j}^{m,n} t_{i,j}^{m,n} \hat{m}^+ \hat{n}^+ \hat{j} \hat{i} + \dots$$

where the combination of operators $\hat{m}^+ \hat{i}$ denotes the *creation* of an electron in the virtual spin orbital ϕ_m and the *removal* of an electron from the occupied spin orbital ϕ_i to generate a single excitation. The operation $\hat{m}^+ \hat{n}^+ \hat{j} \hat{i}$ therefore represents a double excitation from $\phi_i \phi_j$ to $\phi_m \phi_n$.

The amplitudes t_i^m , $t_{i,j}^{m,n}$, etc, which play the role of the C_1 coefficients in CC theory, are determined through the set of equations generated by projecting the Schrödinger equation in the form

$$\exp(-T)\hat{H}\exp(T)\Phi = E\Phi$$

against CSFs which are single, double, etc, excitations relative to Φ :

$$\begin{aligned} \langle \Phi_i^m | \hat{H} + [\hat{H}, T] + \frac{1}{2}[[\hat{H}, T], T] + \frac{1}{6}[[[\hat{H}, T], T], T] + \frac{1}{24}[[[[\hat{H}, T], T], T], T] | \Phi \rangle &= 0 \\ \langle \Phi_{i,j}^{m,n} | \hat{H} + [\hat{H}, T] + \frac{1}{2}[[\hat{H}, T], T] + \frac{1}{6}[[[\hat{H}, T], T], T] + \frac{1}{24}[[[[\hat{H}, T], T], T], T] | \Phi \rangle &= 0 \\ \langle \Phi_{i,j,k}^{m,n,p} | \hat{H} + [\hat{H}, T] + \frac{1}{2}[[\hat{H}, T], T] + \frac{1}{6}[[[\hat{H}, T], T], T] + \frac{1}{24}[[[[\hat{H}, T], T], T], T] | \Phi \rangle &= 0 \end{aligned}$$

and so on for higher-order excited CSFs.

It can be shown [62] that the expansion of the exponential operators truncates exactly at the fourth power in T . As a result, the exact CC equations are *quartic equations* for the t_i^m , $t_{i,j}^{m,n}$, etc amplitudes. The matrix elements appearing in

the CC equations can be expressed in terms of one- and two-electron integrals over the spin orbitals including those occupied in Φ and the virtual orbitals not in Φ .

These quartic equations are solved in an iterative manner and, as such, are susceptible to convergence difficulties. In any such iterative process, it is important to start with an approximation reasonably close to the final result. In CC theory, this is often achieved by neglecting all of the terms that are nonlinear in the t amplitudes (because the t s are assumed to be less than unity in magnitude) and ignoring factors that couple different doubly-excited CSFs (i.e. the sum over i', j', m' and n'). This gives t amplitudes that are equal to the

amplitudes of the first-order MPPT/MBPT wavefunction:

$$t_{i,j}^{m,n} = -\langle i, j | \hat{g} | m, n \rangle' / [\varepsilon_m - \varepsilon_i + \varepsilon_n - \varepsilon_j].$$

As Bartlett [63] and Pople have both demonstrated [64], there is a close relationship between the MPPT/MBPT and CC methods when the CC equations are solved iteratively starting with such an MPPT/MBPT-like initial ‘guess’ for these double-excitation amplitudes.

(E) DENSITY FUNCTIONAL THEORIES

These approaches provide alternatives to the conventional tools of quantum chemistry. The CI, MCSCF, MPPT/MBPT, and CC methods move beyond the single-configuration picture by adding to the wavefunction more configurations whose amplitudes they each determine in their own way. This can lead to a very large number of CSFs in the correlated wavefunction and, as a result, a need for extraordinary computer resources.

The density functional approaches are different [65]. Here one solves a set of orbital-level equations

$$\left[-\hbar^2/2m_e \nabla^2 - \sum_A Z_A e^2 / |\mathbf{r} - \mathbf{R}_A| + \int \rho(\mathbf{r}') e^2 / |\mathbf{r} - \mathbf{r}'| d\mathbf{r}' + U(\mathbf{r}) \right] \phi_i = \varepsilon_i \phi_i$$

in which the orbitals $\{\phi_i\}$ ‘feel’ potentials due to the nuclear centres (having charges Z_A), Coulombic interaction with the *total* electron density $\rho(\mathbf{r}')$ and a so-called *exchange-correlation* potential denoted $U(\mathbf{r}')$. The particular electronic state for which the calculation is being performed is specified by forming a corresponding density $\rho(\mathbf{r}')$. Before going further in describing how density functional theory (DFT) calculations are carried out, let us examine the origins underlying this theory.

The so-called Hohenberg–Kohn [66] theorem states that the *ground-state* electron density $\rho(\mathbf{r})$ describing an N -electron system uniquely determines the potential $V(\mathbf{r})$ in the Hamiltonian

$$\hat{H} = \sum_j \left\{ -\hbar^2/2m_e \nabla_j^2 + V(r_j) + \frac{1}{2} \sum_{k \neq j} e^2 / r_{j,k} \right\}$$

and, because \hat{H} determines the ground-state energy and wavefunction of the system, the ground-state density $\rho(\mathbf{r})$

determines the ground-state properties of the system. The proof of this theorem proceeds as follows.

- (a) $\rho(\mathbf{r})$ determines N because $\int \rho(\mathbf{r}) d^3r = N$.
- (b) Assume that there are two distinct potentials (aside from an additive constant that simply shifts the zero of total energy) $V(\mathbf{r})$ and $V'(\mathbf{r})$ which, when used in \hat{H} and \hat{H}' , respectively, to solve for a ground state produce $E_0, \Psi(\mathbf{r})$ and $E'_0, \Psi'(\mathbf{r})$, $\Psi'(\mathbf{r})$ that have the same one-electron density: $\int |\Psi|^2, dr_2, dr_3 \dots dr_N = \rho(\mathbf{r}) = \int |\Psi'|^2, dr_2, dr_3 \dots dr_N$.
- (c) If we think of Ψ' as trial variational wavefunction for the Hamiltonian \hat{H} , we know that

$$\begin{aligned}
E_0 < \langle \Psi' | \hat{H} | \Psi' \rangle &= \langle \Psi' | \hat{H}' | \Psi' \rangle + \int \rho(\mathbf{r}) [V(\mathbf{r}) - V'(\mathbf{r})] d^3r \\
&= E'_0 + \int \rho(\mathbf{r}) [V(\mathbf{r}) - V'(\mathbf{r})] d^3r.
\end{aligned}$$

(d) Similarly, taking Ψ as a trial function for the H' Hamiltonian, one finds that

$$E'_0 < E_0 + \int \rho(\mathbf{r}) [V'(\mathbf{r}) - V(\mathbf{r})] d^3r.$$

(e) Adding the equations in (c) and (d) gives

$$E_0 + E'_0 < E_0 + E'_0.$$

A clear contradiction.

Hence, there cannot be two distinct potentials V and V' that give the same ground-state $\rho(\mathbf{r})$. So, the ground-state density $\rho(\mathbf{r})$ uniquely determines N and V , and thus \hat{H} , and therefore Ψ and E_0 . Furthermore, because Ψ determines all the properties of the ground state, then $\rho(\mathbf{r})$, in principle, determines all such properties. This means that even the kinetic energy and the electron–electron interaction energy of the ground state are determined by $\rho(\mathbf{r})$. It is easy to see that $\int \rho(\mathbf{r}) V(\mathbf{r}) d^3r = V[\rho]$ gives the average value of the electron–nuclear (plus any additional one-electron additive potential) interaction in terms of the ground-state density $\rho(\mathbf{r})$, but how are the kinetic energy $T[\rho]$ and the electron–electron interaction $V_{ee}[\rho]$ energy expressed in terms of ρ ?

The main difficulty with DFTs is that the Hohenberg–Kohn theorem shows that the *ground-state* values of T , V_{ee} , V , etc are all unique functionals of the *ground-state* ρ (i.e. that they can, in principle, be determined once ρ is given), but it does not tell us what these functional relations are.

To see how it might make sense that a property such as the kinetic energy, whose operator $(-\hbar^2/2m_e)\nabla^2$ involves derivatives, can be related to the electron density, consider a simple system of N non-interacting electrons moving in a three-dimensional cubic ‘box’ potential. The energy states of such electrons are known to be

-30-

$$E = (\hbar^2/8m_e L^2)(n_x^2 + n_y^2 + n_z^2)$$

where L is the length of the box along the three axes and n_x , n_y , and n_z are the quantum numbers describing the state. We can view $n_x^2 + n_y^2 + n_z^2 = R^2$ as defining the squared radius of a sphere in three dimensions, and we realize that the density of quantum states in this space is one state per unit volume in the n_x , n_y , and n_z space. Because n_x , n_y , and $n_z E = (\hbar^2/2m_e L^2)R^2$ is one-eighth the volume of the sphere of radius R :

$$\Phi(E) = \frac{1}{8}(4\pi/3)R^3 = (\pi/6)(8m_e L^2 E / \hbar^2)^{3/2}.$$

Since there is one state per unit of such volume, $\Phi(E)$ is also the number of states with energy less than or equal to E , and is called the *integrated density of states*. The number of states $g(E)dE$ with energy between E and $E + dE$, the *density of states*, is the derivative of Φ :

$$g(E) = d\Phi/dE = (\pi/4)(8m_e L^2 / \hbar^2)^{3/2} E^{1/2}.$$

If we calculate the total energy for N electrons, with the states having energies up to the so-called *Fermi energy* (E_F) (i.e. the energy of the highest occupied molecular orbital HOMO) doubly occupied, we obtain the ground-state energy:

$$E_0 = 2 \int_0^{E_F} g(E)E \, dE = (8\pi/5)(2m_e/h^2)^{3/2} L^3 E_F^{5/2}.$$

The total number of electrons N can be expressed as

$$N = 2 \int_0^{E_F} g(E) \, dE = (8\pi/3)(2m_e/h^2)^{3/2} L^3 E_F^{3/2}$$

which can be solved for E_F in terms of N to then express E_0 in terms of N instead of E_F :

$$E_0 = (3h^2/10m_e)(3/8\pi)^{2/3} L^3 (N/L^3)^{5/3}.$$

This gives the total energy, which is also the kinetic energy in this case because the potential energy is zero within the ‘box’, in terms of the electron density $\rho(x,y,z) = (N/L^3)$. It therefore may be plausible to express kinetic energies in terms of electron densities $\rho(\mathbf{r})$, but it is by no means clear how to do so for ‘real’ atoms and molecules with electron–nuclear and electron–electron interactions operative.

In one of the earliest DFT models, the *Thomas–Fermi* theory, the kinetic energy of an atom or a molecule is approximated using the above type of treatment on a ‘local’ level. That is, for each volume element in \mathbf{r} space, one

assumes the expression given above to be valid, and then one integrates over all \mathbf{r} to compute the total kinetic energy:

$$T_{\text{TF}}[\rho] = \int (3h^2/10m_e)(3/8\pi)^{2/3} [\rho(\mathbf{r})]^{5/3} \, d^3r = C_F \int [\rho(\mathbf{r})]^{5/3} \, d^3r$$

where the last equality simply defines the C_F constant (which is 2.8712 in atomic units). Ignoring the correlation and exchange contributions to the total energy, this T is combined with the electron–nuclear V and Coulombic electron–electron potential energies to give the Thomas–Fermi total energy:

$$E_{0,\text{TF}}[\rho] = C_F \int [\rho(\mathbf{r})]^{5/3} \, d^3r + \int V(\mathbf{r})\rho(\mathbf{r}) \, d^3r + e^2/2 \int \rho(\mathbf{r})\rho(\mathbf{r}')/|\mathbf{r} - \mathbf{r}'| \, d^3r \, d^3r'.$$

This expression is an example of how E_0 is given as a *local density functional approximation* (LDA). The term local means that the energy is given as a functional (i.e. a function of ρ) which depends only on $\rho(\mathbf{r})$ at the points in space, but not on $\rho(\mathbf{r})$ at more than one point in space.

Unfortunately, the Thomas–Fermi energy functional does not produce results that are of sufficiently high accuracy to be of great use in chemistry. What is missing in this theory are the exchange energy and the correlation energy; moreover, the kinetic energy is treated only in the approximate manner described.

In the book by Parr and Yang [67], it is shown how Dirac was able to address the exchange energy for the ‘uniform electron gas’ (N Coulomb *interacting* electrons moving in a uniform positive background charge whose magnitude balances the charge of the N electrons). If the exact expression for the exchange energy of the uniform electron gas is applied on a local level, one obtains the commonly used Dirac *local density approximation to the exchange energy*:

$$E_{\text{ex,Dirac}}[\rho] = -C_x \int [\rho(\mathbf{r})]^{4/3} d^3r$$

with $C_x = (3/4) (3/\pi)^{1/3} = 0.7386$ in atomic units. Adding this exchange energy to the Thomas–Fermi total energy $E_{0,\text{TF}}[\rho]$ gives the so-called Thomas–Fermi–Dirac (TFD) energy functional.

Because electron densities vary rather strongly spatially near the nuclei, corrections to the above approximations to $T[\rho]$ and $E_{\text{ex,Dirac}}$ are needed. One of the more commonly used so-called *gradient-corrected* approximations is that invented by Becke [68], and referred to as the Becke88 exchange functional:

$$E_{\text{ex}}(\text{Becke88}) = E_{\text{ex,Dirac}}[\rho] - \gamma \int x^2 \rho^{4/3} (1 + 6\gamma x \sinh^{-1}(x))^{-1} d\mathbf{r}$$

where $x = \rho^{-4/3} |\nabla\rho|$, and γ is a parameter chosen so that the above exchange energy can best reproduce the known exchange energies of specific electronic states of the inert gas atoms (Becke finds γ to equal 0.0042). A common

gradient correction to the earlier $T[\rho]$ is called the Weizsacker correction and is given by

$$\delta T_{\text{Weizsacker}} = (1/72)(\hbar/m_e) \int |\nabla\rho(\mathbf{r})|^2 / \rho(\mathbf{r}) d\mathbf{r}.$$

Although the above discussion suggests how one might compute the ground-state energy once the ground-state density $\rho(\mathbf{r})$ is given, one still needs to know how to obtain ρ . Kohn and Sham [69] (KS) introduced a set of so-called KS orbitals obeying the following equation:

$$\left\{ -(\hbar^2/2m_e)\nabla^2 + V(\mathbf{r}) + e^2/2 \int \rho(\mathbf{r}')/|\mathbf{r} - \mathbf{r}'| d\mathbf{r}' + U_{\text{xc}}(\mathbf{r}) \right\} \phi_j = \varepsilon_j \phi_j$$

where the so-called exchange-correlation potential $U_{\text{xc}}(\mathbf{r}) = \delta E_{\text{xc}}[\rho]/\delta\rho(\mathbf{r})$ could be obtained by functional differentiation if the exchange-correlation energy functional $E_{\text{xc}}[\rho]$ were known. KS also showed that the KS orbitals $\{\phi_j\}$ could be used to compute the density ρ by simply adding up the orbital densities multiplied by orbital occupancies n_j :

$$\rho(\mathbf{r}) = \sum_j n_j |\phi_j(\mathbf{r})|^2.$$

Here $n_j = 0, 1$ or 2 is the occupation number of the orbital ϕ_j in the state being studied. The kinetic energy

should be calculated as

$$T = \sum_j n_j \langle \phi_j(\mathbf{r}) | -(\hbar^2/2m_e)\nabla^2 | \phi_j(\mathbf{r}) \rangle.$$

The same investigations of the idealized ‘uniform electron gas’ that identified the Dirac exchange functional, found that the correlation energy (per electron) could also be written exactly as a *function* of the electron density ρ of the system, but only in two limiting cases—the high-density limit (large ρ) and the low-density limit. There still exists no exact expression for the correlation energy even for the uniform electron gas that is valid at arbitrary values of ρ . Therefore, much work has been devoted to creating efficient and accurate interpolation formulae connecting the low- and high-density uniform electron gas expressions (see appendix E in [67] for further details). One such expression is

$$E_C[\rho] = \int \rho(\mathbf{r}) \varepsilon_c(\rho) d\mathbf{r}$$

-33-

where

$$\varepsilon_c(\rho) = A/2 \{ \ln(x/X) + 2b/Q \tan^{-1}(Q/(2x+b)) - bx_0/X_0 [\ln((x-x_0)^2/X) + 2(b+2x_0)/Q \tan^{-1}(Q/(2x+b))] \}$$

is the correlation energy per electron. Here $x = r_s^{1/2}$, $X = x^2 + bx + c$, $X_0 = x_0^2 + bx_0 + c$ and $Q = (4c - b^2)^{1/2}$, $A = 0.0621814$, $x_0 = -0.409286$, $b = 13.0720$, and $c = 42.7198$. The parameter r_s is how the density ρ enters since $\frac{4}{3}\pi r_s^3$ is equal to $1/\rho$; that is, r_s is the radius of a sphere whose volume is the effective volume occupied by one electron. A reasonable approximation to the full $E_{xc}[\rho]$ would contain the Dirac (and perhaps gradient corrected) exchange functional plus the above $E_C[\rho]$, but there are many alternative approximations to the exchange-correlation energy functional [68]. Currently, many workers are doing their best to ‘cook up’ functionals for the correlation and exchange energies, but no one has yet invented functionals that are so reliable that most workers agree to use them.

To summarize, in implementing any DFT, one usually proceeds as follows.

- (1) An AO basis is chosen in terms of which the KS orbitals are to be expanded.
- (2) Some initial guess is made for the LCAO–KS expansion coefficients $C_{j,a} : \phi_j = \sum_a C_{j,a} \chi_a$.
- (3) The density is computed as $\rho(\mathbf{r}) = \sum_j n_j |\phi_j(\mathbf{r})|^2$. Often, $\rho(\mathbf{r})$ is expanded in an AO basis, which need not be the same as the basis used for the ϕ_j , and the expansion coefficients of ρ are computed in terms of those of the ϕ_j . It is also common to use an AO basis to expand $\rho^{1/3}(\mathbf{r})$ which, together with ρ , is needed to evaluate the exchange-correlation functional’s contribution to E_0 .
- (4) The current iteration’s density is used in the KS equations to determine the Hamiltonian $\{-1/2\nabla^2 + V(\mathbf{r}) + e^2/2 \int \rho(\mathbf{r}')/|\mathbf{r} - \mathbf{r}'| d\mathbf{r}' + U_{xc}(\mathbf{r})\}$ whose ‘new’ eigenfunctions $\{\phi_j\}$ and eigenvalues $\{\varepsilon_j\}$ are found by solving the KS equations.
- (5) These new ϕ_j are used to compute a new density, which, in turn, is used to solve a new set of KS equations. This process is continued until convergence is reached (i.e. until the ϕ_j used to determine the

current iteration's ρ are the same ϕ_j that arise as solutions on the next iteration).

(6) Once the converged $\rho(\mathbf{r})$ is determined, the energy can be computed using the earlier expression

$$E[\rho] = \sum_i n_i \langle \phi_j(\mathbf{r}) | -(\hbar^2/2m_e)\nabla^2 | \phi_j(\mathbf{r}) \rangle + \int V(\mathbf{r})\rho(\mathbf{r}) d\mathbf{r} + e^2/2 \int \rho(\mathbf{r})\rho(\mathbf{r}')/|\mathbf{r}-\mathbf{r}'| d\mathbf{r} d\mathbf{r}' + E_{xc}[\rho].$$

In closing this section, it should once again be emphasized that this area is currently undergoing explosive growth and much scrutiny [70]. As a result, it is nearly certain that many of the specific functionals discussed above will be replaced in the near future by improved and more rigorously justified versions. It is also likely that extensions of DFTs to excited states (many workers are actively pursuing this) will be placed on more solid ground and made applicable to molecular systems. Because the computational effort involved in these approaches scales much less strongly [71] with the basis set size than for conventional (MCSCF, CI, etc) methods, density functional methods offer great promise and are likely to contribute much to quantum chemistry in the next decade.

(F) EFFICIENT AND WIDELY DISTRIBUTED COMPUTER PROGRAMS EXIST FOR CARRYING OUT ELECTRONIC STRUCTURE CALCULATIONS

The development of electronic structure theory has been ongoing since the 1940s. At first, only a few scientists had access to computers, and they began to develop numerical methods for solving the requisite equations (e.g. the HF equations for orbitals and orbital energies, the configuration interaction equations for electronic state energies and wavefunctions). By the late 1960s, several research groups had developed reasonably efficient computer codes (written primarily in Fortran with selected subroutines that needed to be written especially efficiently in machine language), and the explosive expansion of this discipline was underway. By the 1980s and through the 1990s, these electronic structure programs began to be used by practicing 'bench chemists' both because they became easier to use and because their efficiency and the computers' speed grew to the point where modest to large molecules could be studied.

Web page links [72] to many of the more widely used programs offer convenient access. At present, more electronic structure calculations are performed by non-theorists than by practicing theoretical chemists, largely because of the proliferation of such programs. This does not mean that all that needs to be done in electronic structure theory is done. The rates at which improvements are being made in the numerical algorithms used to solve the problems as well as at which new models are being created remain as high as ever. For example, Professor Rich Friesner [73] has developed and Professor Emily Carter [74] has implemented, for correlated methods, a highly efficient way to replace the list of two-electron integrals $(\phi_i\phi_j|1/r_{12}|\phi_k\phi_l)$, which number N^4 , where N is the number of AO basis functions, by a much smaller list $(\phi_i\phi_j|g)$ from which the original integrals can be rewritten as

$$(\phi_i\phi_j|1/r_{12}|\phi_k\phi_l) = \sum_g (\phi_i(g)\phi_j(g)) \int d\mathbf{r} \phi_k(\mathbf{r})\phi_l(\mathbf{r})/|\mathbf{r}-\mathbf{g}|.$$

This tool, which they call *pseudospectral methods*, promises to reduce the CPU, memory and disk storage requirements for many electronic structure calculations, thus permitting their application to much larger molecular systems. In addition to ongoing developments in the underlying theory and computer

implementation, the range of phenomena and the kinds of physical properties that one needs electronic structure theory to address is growing rapidly. There is every reason to believe that this sub-discipline of theoretical chemistry is continuing to blossom.

B3.1.6.2 COMPUTATIONAL REQUIREMENTS, STRENGTHS AND WEAKNESSES OF VARIOUS METHODS

(A) COMPUTATIONAL STEPS

Essentially all of the techniques discussed above require the evaluation of one- and two-electron integrals over the N AO basis functions: $\langle \chi_a | \hat{f} | \chi_b \rangle$ and $\langle \chi_a \chi_b | \hat{g} | \chi_c \chi_d \rangle$. As mentioned earlier, there are of the order of $N^4/8$ such two-electron integrals that must be computed (and perhaps stored on disk); their computation and storage is a major consideration in performing conventional *ab initio* calculations. Much current research is being devoted to reducing the number of such integrals that must be evaluated using either the pseudo-spectral methods discussed earlier or methods that approximate

-35-

integrals between product distributions (one such distribution is $\chi_a \chi_c$ and another is $\chi_b \chi_d$ when the integral $\langle \chi_a \chi_b | \hat{g} | \chi_c \chi_d \rangle$ is treated) whenever the distributions involve orbitals on sites that are distant from one another.

Another step that is common to most, if not all, approaches that compute orbitals of one form or another is the solution of matrix eigenvalue problems of the form

$$\sum_{\nu} F_{\mu,\nu} C_{\nu,i} = \epsilon_i \sum_{\nu} S_{\mu,\nu} C_{\nu,i}.$$

The solution of any such eigenvalue problem requires a number of computer operations that scales as the dimension of the $F_{\mu,n}$ matrix to the third power. Since the indices on the $F_{\mu,n}$ matrix label AOs, this means that the task of finding all eigenvalues and eigenvectors scales as the cube of the number of AOs (N^3).

The DFT approaches involve basis expansions of orbitals $\phi_i = \sum_{\nu} C_{i,\nu} \chi_{\nu}$ and of the density ρ (or various fractional powers of ρ), which is a quadratic function of the orbitals ($\rho = \sum_i n_i |\phi_i|^2$). These steps require computational effort scaling only as N^2 , which is one of the most important advantages of these schemes. No cumbersome large CSF expansion and associated large secular eigenvalue problem arise, which is another advantage.

The more conventional quantum chemistry methods provide their working equations and energy expressions in terms of one- and two-electron integrals over the final MOs: $\langle \phi_i | \hat{f} | \phi_j \rangle$ and $\langle \phi_i \phi_j | \hat{g} | \phi_k \phi_l \rangle$. The MO-based integrals can only be evaluated by *transforming* the AO-based integrals [55]. Clearly, the N^5 scaling of the integral transformation process makes it an even more time-consuming step than the (N^4) atomic integral evaluation and a severe *bottleneck* to applying *ab initio* methods to larger systems. Much effort has been devoted to expressing the working equations of various correlated methods in a manner that does not involve the fully-transformed MO-based integrals.

Once the requisite one- and two-electron integrals are available in the MO basis, the multiconfigurational wavefunction and energy calculation can begin. Each of these methods has its own approach to describing the configurations $\{\Phi_i\}$ included in the calculation and how the $\{C_i\}$ amplitudes and the total energy E are to be

determined.

The *number of configurations* (N_C) varies greatly among the methods and is an important factor to keep in mind. Under certain circumstances (e.g. when studying reactions where an avoided crossing of two configurations produces an activation barrier), it may be *essential* to use more than one electronic configuration. Sometimes, one configuration (e.g. the SCF model) is adequate to capture the qualitative essence of the electronic structure. In all cases, many configurations will be needed if a highly accurate treatment of electron–electron correlations are desired.

The value of N_C determines how much computer time and memory is needed to solve the N_C -dimensional $\sum_J H_{I,J} C_J = E C_I$ secular problem in the CI and MCSCF methods. Solution of these matrix eigenvalue equations requires computer time that scales as N_C^2 (if few eigenvalues are computed) to N_C^3 (if most eigenvalues are obtained).

So-called *complete active space* (CAS) methods form *all* CSFs that

-36-

can be created by distributing N valence electrons among P valence orbitals. For example, the eight non-core electrons of H_2O might be distributed, in a manner that gives $M_S = 0$, among six valence orbitals (e.g. two lone-pair orbitals, two OH σ -bonding orbitals and two OH σ^* -antibonding orbitals). The number of configurations thereby created is 225. If the same eight electrons were distributed among ten valence orbitals 44 100 configurations result; for 20 and 30 valence orbitals, 23 474 025 and 751 034 025 configurations arise, respectively. Clearly, practical considerations dictate that CAS-based approaches be limited to situations in which a few electrons are to be correlated using a few valence orbitals.

(B) VARIATIONAL METHODS PROVIDE UPPER BOUNDS TO ENERGIES

Methods that are based on making the functional $\langle \Psi | \hat{H} | \Psi \rangle / \langle \Psi | \Psi \rangle$ stationary yield *upper bounds* to the lowest energy state having the symmetry of the CSFs in Ψ . The CI and MCSCF methods are of this type. They also provide approximate excited-state energies and wavefunctions in the form of other solutions of the secular equation [75] $\sum_J H_{I,J} C_J = E C_I$. Excited-state energies obtained in this manner obey the so-called *bracketing theorem*; that is, between any two approximate energies obtained in the variational calculation, there exists at least one true eigenvalue. These are strong attributes of the variational methods, as is the long and rich history of developments of analytical and computational tools for efficiently implementing such methods.

(C) VARIATIONAL METHODS ARE NOT SIZE-EXTENSIVE

Variational techniques suffer from a serious drawback, however: they are not necessarily *size extensive* [76]. The energy computed using these tools cannot be trusted to scale with the size of the system. For example, a calculation performed on two CH_3 species at large separation may not yield an energy equal to twice the energy obtained by performing the *same* kind of calculation on a single CH_3 species. Lack of size extensivity precludes these methods from use in extended systems (e.g. polymers and solids) where errors due to improper size scaling of the energy produce nonsensical results.

By carefully adjusting the variational wavefunction used, it is *possible* to circumvent size-extensivity problems for selected species. For example, the CI calculation on Be_2 using all $^1\Sigma_g$ CSFs formed by placing the four valence electrons into the $2\sigma_g$, $2\sigma_u$, $3\sigma_g$, $3\sigma_u$, $1\pi_u$, and $1\pi_g$ orbitals can yield an energy equal to twice that of the Be atom described by CSFs in which the two valence electrons of the Be atom are placed into the

2s and 2p orbitals in all ways consistent with a 1S symmetry. Such CAS-space MCSCF or CI calculations [77] are size extensive, but it is impractical to extend such an approach to larger systems.

(D) MOST PERTURBATION AND CC METHODS ARE SIZE-EXTENSIVE, BUT DO NOT PROVIDE UPPER BOUNDS AND THEY ASSUME THAT ONE CSF DOMINATES

In contrast to variational methods, perturbation theory and CC methods achieve their energies by projecting the Schrödinger equation against a reference function $\langle \Phi |$ to obtain [78] a *transition formula* $\langle \Phi | \hat{H} | \Psi \rangle$, rather than from an expectation value $\langle \Psi | \hat{H} | \Psi \rangle$. It can be shown that this difference allows non-variational techniques to yield size-extensive energies.

-37-

This can be seen by considering the second-order MPPT energy of two non-interacting Be atoms. The reference CSF is $\Phi = |1s_a^2 2s_a^2 1s_b^2 2s_b^2|$; as discussed earlier, only doubly-excited CSFs contribute to the correlation energy through second order. These ‘excitations’ can involve atom a, atom b, or both atoms. However, CSFs that involve excitations on both atoms (e.g. $|1s_a^2 2s_a 2p_a 1s_b^2 2s_b 2p_b|$) give rise to one- and two-electron integrals over orbitals on both atoms (e.g. $\langle 2s_a 2p_a | \hat{g} | 2s_b 2p_b \rangle$) that vanish if the atoms are far apart, so contributions due to such CSFs vanish. Hence, only CSFs that are excited on one or the other atom contribute to the energy. This, in turn, results in a second-order energy that is additive as required by any size-extensive method. In general, a method will be size extensive *if* its energy formula is additive *and* the equations that determine the C_J amplitudes are themselves separable. The MPPT/MBPT and CC methods possess these characteristics.

However, size-extensive methods have two serious weaknesses. Their energies do *not* provide upper bounds to the true energies of the system (because their energy functional is not of the expectation-value form for which the upper bound property has been proven). Moreover, they express the correct wavefunction in terms of corrections to a (presumed dominant) reference function which is usually taken to be a single CSF (although efforts have been made to extend the MPPT/MBPT and CC methods to allow for multiconfigurational reference functions, this is not yet standard practice). For situations in which two CSFs ‘cross’ along a reaction path, the single-dominant-CSF assumption breaks down, and these methods can have difficulty.

B3.1.7 THERE ARE METHODS THAT CALCULATE ENERGY DIFFERENCES RATHER THAN ENERGIES

In addition to the myriad of methods discussed above for treating the energies and wavefunctions as solutions to the electronic Schrödinger equation, there exists a family of tools that allow one to compute energy differences ‘directly’ rather than by first finding the energies of pairs of states and subsequently subtracting them. Various energy differences can be so computed: differences between two electronic states of the same molecule (i.e. electronic excitation energies ΔE), differences between energy states of a molecule and the cation or anion formed by removing or adding an electron (i.e. IPs and EAs).

Because of space limitations, we will not be able to elaborate much further on these methods. However, it is important to stress that:

- (1) these so-called *Greens function* or *propagator* methods [71] utilize essentially the same input

information (e.g. AO basis sets) and perform many of the same computational steps (e.g. evaluation of one- and two-electron integrals, formation of a set of mean-field MOs, transformation of integrals to the MO basis, etc) as do the other techniques discussed earlier;

- (2) these methods are now rather routinely used when ΔE , IP, or EA information is sought. In fact, the 1998 version of the Gaussian program includes an electron propagator option.

The basic ideas underlying most, if not all, of the energy-difference methods follow

-38-

- (1) One forms a *reference wavefunction* Ψ (this can be of the SCF, MPn, CC, etc variety); the energy differences are computed relative to the energy of this function.
- (2) One expresses the *final-state wavefunction* Ψ' (i.e. describing the excited, cation, or anion state) in terms of an operator Ω acting on the reference Ψ : $\Psi' = \Omega\Psi$. Clearly, the Ω operator must be one that removes or adds an electron when one is attempting to compute IPs or EAs, respectively.
- (3) One writes equations which Ψ and Ψ' are expected to obey. For example, in the early development of these methods [80], the Schrödinger equation itself was assumed to be obeyed, so $\hat{H}\Psi = E\Psi$ and $\hat{H}'\Psi' = E'\Psi'$ are the two equations (note that, in the IP and EA cases, the latter equation, and the associated Hamiltonian \hat{H}' , refer to one fewer and one more electrons than does the reference equation $\hat{H}\Psi = E\Psi$).
- (4) One combines $\Omega\Psi = \Psi'$ with the equations that Ψ and Ψ' obey to obtain an equation that Ω must obey. In the above example, one: (a) uses $\Omega\Psi = \Psi'$ in the Schrödinger equation for Ψ' , (b) allows Ω to act from the left on the Schrödinger equation for Ψ and (c) subtracts the resulting two equations to achieve $(\hat{H}'\hat{\Omega} - \hat{\Omega}\hat{H})\Psi = (E' - E)\Omega\Psi$ or, in commutator form $[\hat{H}', \hat{\Omega}]\Psi = \Delta E\hat{\Omega}\Psi$. By expressing the Hamiltonian in the second-quantization form, only one \hat{H} appears in this final so-called *equation of motion* (EOM) $[\hat{H}, \hat{\Omega}]\Psi = \Delta E\hat{\Omega}\Psi$ (i.e. in the second-quantized form, \hat{H}' and \hat{H} are one and the same).
- (5) One can, for example, express Ψ in terms of a superposition of configurations $\Psi = \sum_j C_j \Phi_j$ whose amplitudes C_j have been determined from an MCSCF, CI or MPn calculation and express Ω in terms of second-quantization operators $\{O_K\}$ that cause single-, double-, etc, level excitations (for the IP (EA) cases, Ω is given in terms of operators that remove (add), remove and singly excite (add and singly excite) electrons): $\Omega = \sum_K D_K \hat{O}_K$.
- (6) Substituting the expansions for Ψ and for $\hat{\Omega}$ into the EOM $[\hat{H}, \hat{\Omega}]\Psi = \Delta E\hat{\Omega}\Psi$, and then projecting the resulting equation on the left against a set of functions (e.g. $\{\hat{O}_{K'}|\Psi\rangle\}$ or $\{\hat{O}_{K'}|\Phi_0\rangle$, where Φ_0 is the dominant component of Ψ), gives a matrix eigenvalue–eigenvector equation:

$$\sum_K \langle \hat{O}_{K'}|\Psi|[\hat{H}, \hat{O}_K]\Psi\rangle D_K = \Delta E \sum_K \langle \hat{O}_{K'}|\Psi|\hat{O}_K\Psi\rangle \hat{D}_K$$

to be solved for the \hat{D}_K operator coefficients and the excitation energies ΔE . Such are the working equations of the EOM (or Greens function or propagator) methods.

In recent years, these methods have been greatly expanded and have reached a degree of reliability where they now offer some of the most accurate tools for studying excited and ionized states. In particular, the use of time-dependent variational principles have allowed the much more rigorous development of equations for energy differences and nonlinear response properties [81]. In addition, the extension of the EOM theory to include coupled-cluster reference functions [82] now allows one to compute excitation and ionization energies using some of the most accurate *ab initio* tools.

B3.1.8 SUMMARY OF AB INITIO METHODS

At this time, it may not be possible to say which method is preferred for applications where all are practical. Nor is it possible to assess, in a way that is applicable to most chemical species, the accuracies with which various methods

-39-

predict bond lengths and energies or other properties. However, there are reasons to recommend some methods over others in specific cases. For example, certain applications require a size-extensive energy (e.g. extended systems that consist of a large or macroscopic number of units or studies of weak intermolecular interactions), so MBPT/MPPT-, CC- or CAS-based MCSCF are preferred. Moreover, many chemical reactions and bond-breaking events require two or more ‘essential’ electronic configurations. For them, single-configuration-based methods such as conventional CC and MBTP/MPPT should be used only with caution; MCSCF or CI calculations are preferred. Very large molecules, in which thousands of AO basis functions are required, may be impossible to treat by methods whose effort scales as N^4 or higher; density functional methods would be the only choice then.

For all calculations, the choice of AO basis set must be made carefully, keeping in mind the N^4 scaling of the two-electron integral evaluation step and the N^5 scaling of the two-electron integral transformation step. Of course, basis functions that describe the essence of the states to be studied are essential (e.g. Rydberg or anion states require diffuse functions and strained rings require polarization functions).

As larger atomic basis sets are employed, the size of the CSF list used to treat a dynamic correlation increases rapidly. For example, many of the above methods use singly- and doubly-excited CSFs for this purpose. For large basis sets, the number of such CSFs (N_C) scales as the number of electrons squared n_e^2 times the number of basis functions squared N^2 . Since the effort needed to solve the CI secular problem varies as N_C^2 or N_C^3 (the latter being to find all eigenvalues and vectors), a dependence as strong as $n_e^6 N^6$ can result. To handle such large CSF spaces, all of the multiconfigurational techniques mentioned in this paper have been developed to the extent that calculations involving of the order of 100–5000 CSFs are routinely performed and calculations using even several billion CSFs are possible [53].

Some of the most significant advances that have been made recently in expanding the applicability of the *ab initio* methods to larger systems are based on recognizing that many of the two-electron integrals and one- and two-electron density matrix elements arising in the pertinent working equations vanish if expressed in terms of localized (atomic or molecular) orbitals. For example, in a polymer consisting of P monomer units (or a crystal composed of P unit cells), the integrals and density matrix elements indexed by monomer units far distant from one another are negligible. Thus, if a method whose effort scales as the k th power of the number of AOs (N) per monomer (or unit cell) is applied to a system having P units, the effort should *not* scale as $(PN)^k$ but, hopefully, as PN^k . Indeed, for the DFT ($k = 3$), SCF ($k = 4$) and MP2 ($k = 5$) methods, specialized techniques [50] have allowed for the implementation of codes scaling linearly (or nearly so for MP2) with the system ‘size’ P (i.e. the number of units).

Other methods, most of which can be viewed as derivatives of the techniques introduced above, have been and are still being developed; stimulated by the explosive growth in computer power and changes in computer architecture realized in recent years. All indications are that this growth pattern will continue; so *ab initio* quantum chemistry is likely to have an even larger impact on future chemistry research and education (through new insights and concepts). For many of the most commonly employed *ab initio* quantum chemistry tools, the computational efforts, as characterized by how they scale with the system size P (i.e. the number of units), with basis set size N and with the number of electronic configurations N_C , as well their variational nature and size extensivity are summarized in [table B3.1.2](#).

Table 3.1.2 Properties of commonly used methods.

Method	Variational/size extensive	Computational scaling
HF	Yes/Yes	N^4 integrals; N^3 eigenvalues; P^1
GVB	Yes/Yes	N^4 integrals N^4 (per electron pair) GVB equations
DFT	No/Yes	N^3 eigenvalues; N^2 integrals; P^1 N^3 orbital orthogonalization; P^1
MP2	No/Yes	N^5 ; P^2
CI	Yes/No	N^5 transformed integrals; N_C^2 to solve for <i>one</i> CI energy and eigenvector
CISD	Yes/No	N^5 transformed integrals; $n^2 N^4$ to solve for one CI energy and eigenvector
CAS-MCSCF	Yes/Yes	N^5 transformed integrals; N_C^2 to solve for CI energy; many iterations also needed
CCS	No/Yes	N^4
CCSD	No/Yes	N^6
CCSDT	No/Yes	N^8
CCSD(T)	No/Yes	N^7

Figure B3.1.9 [83] displays the errors (in picometres compared to experimental findings) in the equilibrium bond lengths for a series of 28 molecules obtained at the HF, MP2-4, CCSD, CCSD(T), and CISD levels of theory using three polarized correlation-consistent basis sets (valence DZ through to QZ).

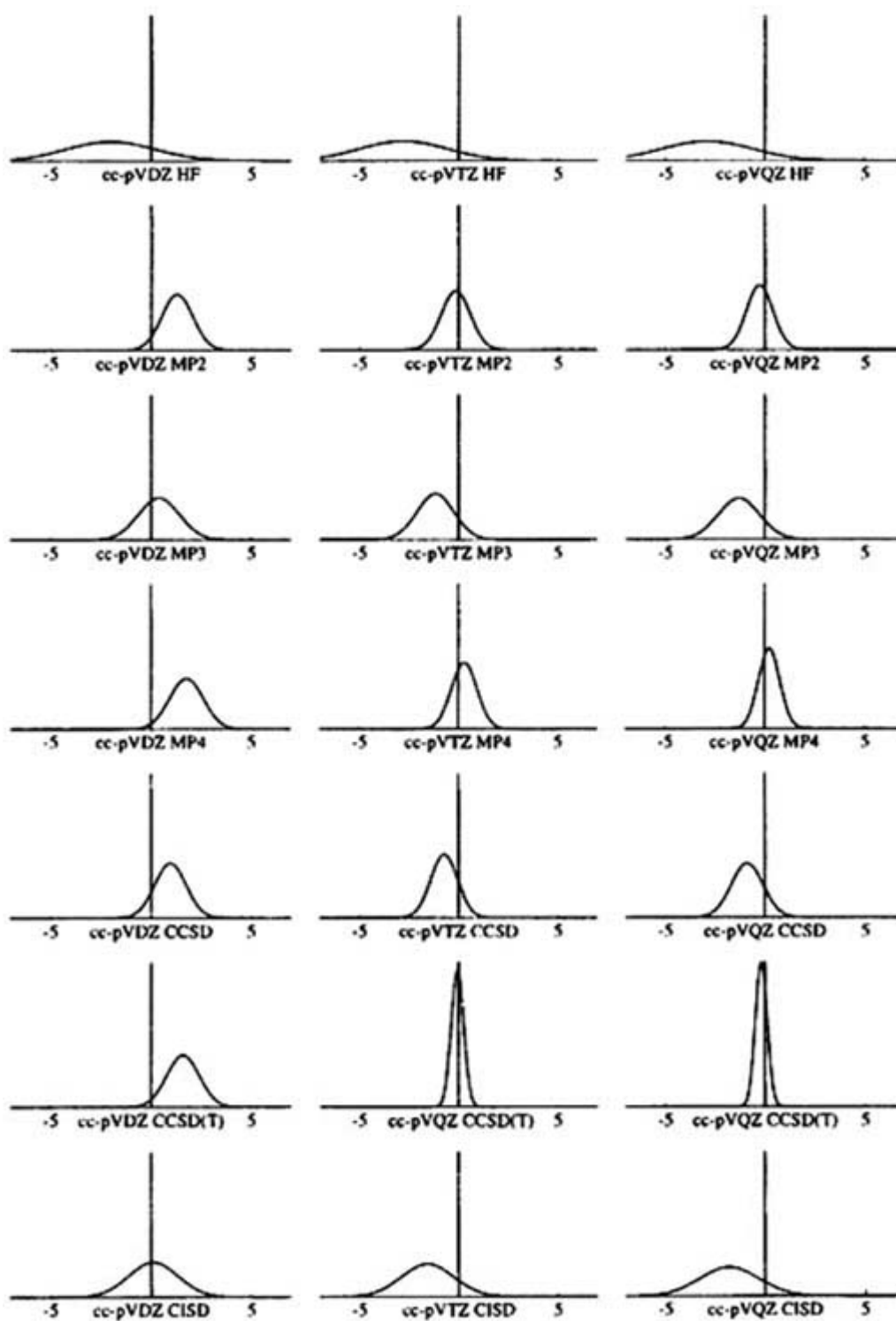


Figure B3.1.9. Distribution in errors (picometres) in calculated bond lengths for 28 test molecules.

Clearly, the HF method, independent of basis, systematically underestimates the bond lengths over a broad percentage range. The CISD method is neither systematic nor narrowly distributed in its errors, but the MP2 and MP4 (but not MP3) methods are reasonably accurate and have narrow error distributions if valence TZ or QZ bases are used. The CCSD(T), but not the CCSD, method can be quite reliable if valence TZ or QZ bases are used.

In closing this section and this chapter, I wish to remind the reader that my discussion has been limited to *ab initio* techniques; that is, to methods that begin with the electronic Schrödinger equation attempt to solve it without explicitly introducing any experimental data or any numerical results from another calculation. There exists a whole family of alternative approaches called *semi-empirical methods* [84] in which (a) overlaps between pairs of orbitals distant from one another are neglected, (b) many of the two-electron integrals appearing in *ab initio* methods are neglected (because they are ‘small’ in some sense) and (c) certain combinations of one- and two-electron integrals that can be (approximately) related to orbital energies of a constituent atom are not computed explicitly but are replaced by experimental data (or data from an *ab initio* calculation) on that atom. Interested readers in these approaches to electronic structure are referred to the articles given in [84].

REFERENCES

- [1] Born M and Oppenheimer J R 1927 *Ann. Phys., Lpz* **84** 457
It was then used within early quantum chemistry in the following references:
Kolos W and Wolniewicz L 1968 Improved theoretical ground-state energy of the hydrogen molecule *J. Chem. Phys.* **49** 404–10
Pack R T and Hirschfelder J O 1968 Separation of rotational coordinates from the *N*-electron diatomic Schrödinger equation *J. Chem. Phys.* **49** 4009
Pack R T and Hirschfelder J O 1970 Energy corrections to the Born–Oppenheimer approximation. The best adiabatic approximation *J. Chem. Phys.* **52** 521–34
- [2] Schlegel H B 1995 *Modern Electronic Structure Theory* ed D R Yarkony (Singapore: World Scientific) ch 8
- [3] Chalasinski G, Kendall R A, Taylor H and Simons J 1988 Propensity rules for vibration–rotation induced electron detachment of diatomic anions: application to $\text{NH}^- \rightarrow \text{NH} + \text{e}^-$ *J. Phys. Chem.* **92** 3086–91
- [4] Early treatments of molecules in which non-Born–Oppenheimer terms were included were made in:
Kolos W and Wolniewicz L 1963 Nonadiabatic theory for diatomic molecules and its application to the hydrogen molecule *Rev. Mod. Phys.* **35** 473–83
Kolos W and Wolniewicz L 1964 *J. Chem. Phys.* **41** 3663
Kolos W and Wolniewicz L 1964 *J. Chem. Phys.* **41** 3674
Kolos W and Wolniewicz L 1965 Potential energy curves for the $X^1\Sigma_g^+$, $b^3\Sigma_u^+$, and $C^1\Pi_u$ states of the hydrogen molecule *J. Chem. Phys.* **43** 2429–41
Some of the work my students and I have done in this area can be found in:
Simons J 1981 Propensity rules for vibration-induced electron detachment of anions *J. Am. Chem. Soc.* **103** 3971–6
Acharya P K, Kendall R A and Simons J 1984 Vibration-induced electron detachment in molecular anions *J. Am. Chem. Soc.* **106** 3402–7
Acharya P K, Kendall R A and Simons J 1985 Associative electron detachment: $\text{O}^- + \text{H} \rightarrow \text{OH} + \text{e}^-$ *J. Chem. Phys.* **83** 3888–93
O’Neal D and Simons J 1989 Vibration-induced electron detachment in acetaldehydeenolate anion *J. Phys. Chem.* **93** 58–61
Simons J 1989 Modified rotationally adiabatic model for rotational autoionization of dipole-bound molecular anion *J. Chem. Phys.* **91** 6858–68
- [5] Ohm Y 2000 The Quantum Theory Project at the University of Florida webpage <http://www.qtp.ufl.edu/~ohrn/>
-

- [6] Yarkony D 2000 webpage <http://jhuniverse.jhu.edu/~chem/yarkony.html>
For a good recent overview of some of his work, see:
Yarkony D R 1995 Electronic structure aspects of nonadiabatic processes in polyatomic systems *Modern Electronic Structure Theory* vol 2, ed D R Yarkony (Singapore: World Scientific) pp 642–721
- [7] Simons J 1983 *Energetic Principles of Chemical Reactions* (Boston, MA: Jones and Bartlett)
- [8] Schlegel B 2000 webpage <http://www.science.wayne.edu/~chem/schlegel.html>
For a good recent overview see:
Schlegel H B 1995 Geometry optimization on potential energy surfaces *Modern Electronic Structure Theory* vol 2, ed D R Yarkony (Singapore: World Scientific) pp 459–500

Strategies for 'walking' on potential energy surfaces are overviewed by Schlegel H B 1987 Optimization of equilibrium geometries and transition structures *Adv. Chem. Phys.* **67** 249–86
My own coworkers and I have also contributed to finding transition states, in particular. See, for example:
Simons J, Jørgensen P, Taylor H and Ozment J 1983 Walking on potential energy surfaces *J. Phys. Chem.* **87** 2745–53
Nichols J A, Taylor H, Schmidt P P and Simons J 1990 Walking on potential energy surfaces *J. Chem. Phys.* **92** 340–6
Simons J and Nichols J 1990 Strategies for walking on potential energy surfaces using local quadratic approximations *Int. J. Quant. Chem. S* **24** 263–76

- [9] Pulay P 1995 Analytical derivative techniques and the calculation of vibrational spectra *Modern Electronic Structure Theory* vol 2, ed D R Yarkony (Singapore: World Scientific) pp 1191–240
Much of the early work is described in:
Pulay P 1977 Direct use of the gradient for investigating molecular energy surfaces *Modern Theoretical Chemistry* vol 4, ed H F III Schaefer (New York: Plenum) pp 53–185
One of the earliest applications to molecular structure is given in:
Thomsen K and Swanstrøm P 1973 Calculation of molecular one-electron properties using coupled Hartree–Fock methods I. Computational scheme *Mol. Phys.* **26** 735–50
More recent contributions are summarized in:
Jørgensen P and Simons J (eds) 1986 *Geometrical Derivatives of Energy Surfaces and Molecular Properties* (Boston, MA: Reidel)
Pulay P 1987 Analytical derivative methods in quantum chemistry *Advances in Chemical Physics* vol LXIX, ed K P Lawley (New York: Wiley–Interscience) pp 241–86
Helgaker T and Jørgensen P 1988 Analytical calculation of geometrical derivatives in molecular electronic structure theory *Adv. Quantum Chem.* **19** 183–245
- [10] The dipole moment is discussed in:
Karplus M and Porter R N 1970 *Atoms and Molecules* (New York: Benjamin)
- [11] Atkins P W 1992 *The Elements of Physical Chemistry* (New York: Freeman)
Berry R S, Rice S A and Ross J 1980 *Physical Chemistry* (New York: Wiley)
and also webpage www.whfreeman.com/echem/index.html
- [12] Levine I N 1991 *Quantum Chemistry* 4th edn (Englewood Cliffs, NJ: Prentice-Hall)
McQuarrie D A 1983 *Quantum Chemistry* (Mill Valley, CA: University Science)
Simons J and Nichols J 1997 *Quantum Mechanics in Chemistry* (New York: Oxford University Press)
Atkins P W and Friedman R S 1997 *Molecular Quantum Mechanics* 3rd edn (Oxford: Oxford University Press)
Szabo A and Ostlund N S 1989 *Modern Quantum Chemistry* 1st edn (revised) (New York: McGraw-Hill)
and also webpage <http://www.emsl.pnl.gov:2080/docs/tms/quantummechanics/>
- [13] For a good recent overview, see:
Olsen J and Jørgensen P 1995 Time-dependent response theory with applications to self-consistent field and multiconfigurational self-consistent field wave functions *Modern Electronic Structure Theory* vol 2, ed D R Yarkony (Singapore: World Scientific) pp 857–990
An earlier introduction of molecular properties in terms of wavefunctions and energies and their responses to externally applied fields is given in:
Jørgensen P and Simons J 1981 *Second Quantization-Based Methods in Quantum Chemistry* (New York: Academic)
Jørgensen P and Simons J (eds) 1986 *Geometrical Derivatives of Energy Surfaces and Molecular Properties* (Boston, MA: Reidel)
Amos R D 1987 Molecular property derivatives *Advances in Chemical Physics* vol LXVII, ed K P Lawley, pp 99–153

- [14] Head-Gordon M 1996 Quantum chemistry and molecular processes *J. Phys. Chem.* **100** 13 213–25
- [15] Schaefer H F III, Thomas J R, Yamaguchi Y, DeLeeuw B J and Vacek G 1995 The chemical applicability of standard methods in *ab initio* molecular quantum mechanics *Modern Electronic Structure Theory* vol 2, ed D R Yarkony (Singapore: World Scientific) pp 3–54
Schaefer F 2000 webpage <http://zopyros.ccqc.uga.edu/>
- [16] Simons J 1991 An experimental chemist's guide to *ab initio* quantum chemistry *J. Phys. Chem.* **95** 1017–29
- [17] Levy H A and Danford M D 1964 *Molten Salt Chemistry* ed M Blander (New York: Interscience)
Bredig M A 1969 *Molten Salts* ed G M Mamantov (New York: Dekker)
- [18] See, for example:
Syngé J L and Griffith B A 1949 *Principles of Mechanics* (New York: McGraw-Hill)
- [19] Berry R S, Rice S A and Ross J 1980 *Physical Chemistry* (New York: Wiley) section 11.9
- [20] One of the earliest uses of this term was offered by:
Sinanoglu O 1961 Many-electron theory of atoms and molecules *Proc. US Natl Acad. Sci.* **47** 1217–26

- [21] Sinanoglu O 1962 Many-electron theory of atoms and molecules I. Shells, electron pairs vs many-electron correlations *J. Chem. Phys.* **36** 706–17
- [22] Helgaker T, Jørgensen P and Olsen J 1999 *Electronic Structure Theory* (New York: Wiley)
- [23] Hylleraas E A 1963 Reminiscences from early quantum mechanics of two-electron atoms *Rev. Mod. Phys.* **35** 421–31
- [24] Pilar F L 1968 *Elementary Quantum Chemistry* (New York: McGraw-Hill) section 11.5
- [25] Löwdin P O 1959 Correlation problem in many-electron quantum mechanics *Adv. Chem. Phys.* **2** 207–32
- [26] Excellent early overviews of many of these methods are included in:
 Schaefer H F III (ed) 1977 *Modern Theoretical Chemistry* vol 3 (New York: Plenum)
 Schaefer H F III (ed) 1977 *Modern Theoretical Chemistry* vol 4 (New York: Plenum)
 Lawley K P (ed) 1987 *Advances in Chemical Physics* vol LXVII (New York: Wiley–Interscience)
 Lawley K P (ed) 1987 *Advances in Chemical Physics* vol LXIX (New York: Wiley–Interscience)
 Yarkony D R (ed) 1995 *Modern Electronic Structure Theory* (Singapore: World Scientific)
- [27] Simons J 1983 *Energetic Principles of Chemical Reactions* (Boston, MA: Jones and Bartlett) p 129
- [28] Goddard W A III 2000 webpage <http://www.caltech.edu/~chemistry/Faculties/Goddard.html>
 When most quantum chemists were pursuing improvements in the molecular orbital method, he returned to the valence bond theory and developed the so-called GVB methods that allow electron correlation to be included within a valence bond framework
- [29] See, for example:
 Goddard W A III and Harding L B 1978 The description of chemical bonding from *ab initio* calculation *Ann. Rev. Phys. Chem.* **29** 363–96
- [30] The classic papers in which the SCF equations for closed- and open-shell systems are treated are:
 Roothaan C C J 1951 New developments in molecular orbital theory *Rev. Mod. Phys.* **23** 69–89
 Roothaan C C J 1960 Self-consistent field theory for open shells of electronic systems *Rev. Mod. Phys.* **32** 179–85
- [31] The original works by the Hartree father and son team appear in:
 Hartree D R 1928 The wave mechanics of an atom with a non-Coulomb central field. Part III. Term values and intensities in series in optical spectra *Proc. Camb. Phil. Soc.* **24** 426–37
 Hartree D R, Hartree W and Swirles B 1940 Self-consistent field including exchange and superposition of configurations with some results for oxygen *Phil. Trans. R. Soc. A* **238** 229–47
 The work of Fock is given in:
 Fock V 1930 *Z. Phys.* **61** 126
- [32] Frisch M J *et al* 1995 *Gaussian 94* Revision A.1 (Pittsburgh, PA: Gaussian)

- [33] Pople J 2000 webpage <http://www.chem.nwu.edu/brochure/pople.html>
 Pople made many developments leading to the suite of Gaussian computer codes that now constitute the most widely used electronic structure computer programs
- [34] Froese-Fischer C 1970 A multi-configuration Hartree–Fock program *Comput. Phys. Commun.* **1** 151–66
- [35] McCullough E A Jr 1975 The partial-wave self-consistent method for diatomic molecules computational formalism and results for small molecules *J. Chem. Phys.* **62** 3991–9
 Christiansen P A and McCullough E A Jr 1977 Numerical Hartree–Fock calculations for N₂, FH, and CO comparison with optimized LCAO results *J. Chem. Phys.* **67** 1877–82
- [36] Mulliken Prof. R S University of Chicago. He is the person who came up with the phrase ‘molecular orbital’ and the concepts behind it
- [37] Jordan Prof. K 2000 webpage <http://www.chem.pitt.edu/~jordan/index.html>
 Jordan compared the use of plane wave and conventional Gaussian basis orbitals within density functional calculations in:
 Nachtigall P, Jordan K D, Smith A and Jönsson H 1996 Investigation of the reliability of density functional methods reaction and activation energies for Si–Si bond cleavage and H₂ elimination from silanes *J. Chem. Phys.* **104** 148–58
- [38] Dunning T H Jr 1970 Gaussian basis functions for use in molecular calculations I. Contraction of (9s 5p) atomic basis sets for the first-row atoms *J. Chem. Phys.* **53** 2823–33
 Dunning T H Jr and Hay P J 1977 Gaussian basis sets for molecular calculations *Methods of Electronic Structure Theory* vol 3, ed H F III Schaefer (New York: Plenum) pp 1–27
- [39] Huzinaga S 1965 Gaussian-type functions for polyatomic system I. *J. Chem. Phys.* **42** 1293–1302

- [40] Schmidt M W and Ruedenberg K 1979 Effective convergence to complete orbital bases and to the atomic Hartree–Fock limit through systematic sequences of Gaussian primitives *J. Chem. Phys.* **71** 3951–62
- [41] Hehre W J, Stewart R F and Pople J A 1969 Self-consistent molecular-orbital method I. Use of Gaussian expansions of Slater-type atomic orbitals *J. Chem. Phys.* **51** 2657–64
- [42] Ditchfield R, Hehre W J and Pople J A 1971 Self-consistent molecular-orbital methods IX. An extended Gaussian-type basis for molecular-orbital studies of organic molecules *J. Chem. Phys.* **54** 724–8
Hehre W J, Ditchfield R and Pople J A 1972 Self-consistent molecular-orbital methods XII. Further extension of Gaussian-type basis sets for use in molecular orbital studies of organic molecules *J. Chem. Phys.* **56** 2257–61
Hariharan P C and Pople J A 1973 The influence of polarization functions on molecular orbital hydrogenation energies *Theoret. Chim. Acta.* **28** 213–22
Krishnan R, Binkley J S, Seeger R and Pople J A 1980 Self-consistent molecular orbital methods XX. A basis set for correlated wave functions *J. Chem. Phys.* **72** 650–4
- [43] Helgaker T and Taylor P R 1995 Gaussian basis sets and molecular integrals *Modern Electronic Structure Theory* vol 2, ed D R Yarkony (Singapore: World Scientific) section 5.4, pp 725–856
- [44] Helgaker T and Taylor P R 1995 Gaussian basis sets and molecular integrals *Modern Electronic Structure Theory* vol 2, ed D R Yarkony (Singapore: World Scientific) section 5.3, pp 725–856
- [45] Pacific Northwest National Laboratories 2000 webpage www.emsl.pnl.gov:2080/forms/basisform.html
- [46] Hunt W J and Goddard W A III 1969 Excited states of H₂O using improved virtual orbitals *Chem. Phys. Lett.* **3** 414–18
- [47] von Koopmans T 1934 Über die zuordnung von wellenfunktionen und eigenwerten zu den einzelnen elektronen eines atoms *Physica* **1** 104–13
- [48] Some of the integral packages and the techniques used to evaluate the integrals are described in:
Csizmadia I G, Harrison M C, Moscovitz J W and Sutcliffe B T 1966 Commentationes. Non-empirical LCAO–MO–SCF–CI calculations on organic molecules with Gaussian type functions. Part I. Introductory review and mathematical formalism *Theoret. Chim. Acta* **6** 191–216

Clementi E and Davis D R 1967 Electronic structure of large molecular systems *J. Comp. Phys.* **1** 223–44
Rothenberg S, Kollman P, Schwartz M E, Hays E F and Allen L C 1970 Mole. A system for quantum chemistry I. General description *Int. J. Quantum Chem. S* **3** 715–25
Hehre W J, Lathan W A, Ditchfield R, Newton M D and Pople J A 1971 *Program* No 236 (Bloomington, IN: Quantum Chemistry Program Exchange)
Dupuis M, Rys J and King H F 1976 Evaluation of molecular integrals over Gaussian basis functions *J. Chem. Phys.* **65** 111–16
McMurchie L E and Davidson E R 1978 One- and two-electron integrals over Cartesian Gaussian functions *J. Comp. Phys.* **26** 218–31
Gill P M W 1994 Molecular integrals over Gaussian basis functions *Adv. Quantum Chem.* **25** 141–205

- [49] This concept of ‘direct’ calculations in which integrals are not stored but used ‘on the fly’ is discussed in:
Almlöf J, Faegri K and Korsell K 1982 Principles for a direct SCF approach to LCAO–MO *ab initio* calculations *J. Comput. Chem.* **3** 385–99
A good recent overview of direct methods is given by:
Almlöf J 1995 Direct methods in electronic structure theory *Modern Electronic Structure Theory* vol 2, ed D R Yarkony (Singapore: World Scientific) pp 110–51
- [50] Strain M C, Scuseria G E and Frisch M J 1996 Linear scaling for the electronic quantum coulomb problem *Science* **271** 51–3

White C A, Johnson B G, Gill P M W and Head-Gordon M 1994 The fast multipole method *Chem. Phys. Lett.* **230** 8–16
White C A, Johnson B G, Gill P M W and Head-Gordon M 1996 Linear scaling density functional calculations via the continuous fast multipole method *Chem. Phys. Lett.* **253** 268–78
Saebo S and Pulay P 1993 Local treatment of electron correlation *Ann. Rev. Phys. Chem.* **44** 213–36

- [51] Werner H-J 1987 Matrix-formulated direct multiconfiguration self-consistent field and multiconfiguration reference configuration-interaction methods *Advances in Chemical Physics* vol LXIX, ed K P Lawley (New York: Wiley-Interscience) pp 1–62
Shepard R 1987 The multiconfiguration self-consistent field method *Advances in Chemical Physics* vol LXIX, ed K P Lawley (New York: Wiley-Interscience) pp 63–200
describe several of the advances that have been made in the MCSCF method, especially with respect to enhancing its rate and range of convergence
Wahl A C and Das G 1977 The multiconfiguration self-consistent field method *Modern Theoretical Chemistry* vol 3, ed H F III Schaefer (New York: Plenum) pp 51–78
covers the 'earlier' history on this topic
Bobrowicz F W and Goddard W A III 1977 The self-consistent field equations for generalized valence bond and open-shell Hartree-Fock wave functions *Modern Theoretical Chemistry* vol 3, ed H F III Schaefer (New York: Plenum) pp 97–127
provide, in *Modern Theoretical Chemistry* vol 3, an overview of the GVB approach, which can be viewed as a specific kind of MCSCF calculation
See also:
Dalgaard E and Jørgensen P 1978 Optimization of orbitals for multiconfigurational reference states *J. Chem. Phys.* **69** 3833–44
Jensen H J Aa, Jørgensen P and Ågren H 1987 Efficient optimization of large scale MCSCF wave functions with a restricted step algorithm *J. Chem. Phys.* **87** 451–66
Lengsfeld B H III and Liu B 1981 A second order MCSCF method for large CI expansions *J. Chem. Phys.* **75** 478–80
- [52] An early article on this method is:
Boys S F 1950 Electronic wave functions II. A calculation for the ground state of the beryllium atom *Proc. R. Soc. A* **201** 125–37
Shavitt I 1977 The method of configuration interaction *Modern Theoretical Chemistry* vol 3, ed H F III Schaefer (New York: Plenum) pp 189–275
Ross B O and Siegbahn P E M 1977 The direct configuration interaction method from molecular integrals *Modern Theoretical Chemistry* vol 3, ed H F III Schaefer (New York: Plenum) pp 277–318
give excellent overviews of the CI method
For a nice overview of recent work, see:
Schaefer H F III, Thomas J R, Yamaguchi Y, Deleuw B J and Vacek G 1995 The chemical applicability of standard methods in *ab initio* molecular quantum mechanics *Modern Electronic Structure Theory* vol 2, ed D R Yarkony (Singapore: World Scientific) pp 3–54
- [53] Olsen J, Roos B, Jørgensen P and Jensen H J Aa 1988 Determinant based configuration interaction algorithms for complete and restricted configuration interaction spaces *J. Chem. Phys.* **89** 2185–92
Olsen J, Jørgensen P and Simons J 1990 Passing the one-billion limit in full configuration-interaction (Fci) calculations *Chem. Phys. Lett.* **169** 463–72

- [54] The so-called Slater-Condon rules express the matrix elements of any one-electron (F) plus two-electron (G) additive operator between pairs of antisymmetrized spin-orbital products that have been arranged (by permuting spin-orbital ordering) to be in so-called maximal coincidence. Once in this order, the matrix elements between two such Slater determinants (labelled $|i\rangle$ and $|j\rangle$) are summarized as follows:
(i) if $|i\rangle$ and $|j\rangle$ are identical, then

$$\langle F + G \rangle = \sum_i \langle \phi_i | f | \phi_i \rangle + \sum_{i>j} [\langle \phi_i \phi_j | g | \phi_i \phi_j \rangle - \langle \phi_i \phi_j | g | \phi_j \phi_i \rangle]$$

where the sums over i and j run over all spin orbitals in $|\rangle$;

(ii) if $|\rangle$ and $|\prime\rangle$ differ by a single spin-orbital mismatch ($\phi_p \neq \phi'_p$),

$$\langle F + G \rangle = \langle \phi_p | f | \phi'_p \rangle + \sum_j [\langle \phi_p \phi_j | g | \phi'_p \phi_j \rangle - \langle \phi_p \phi_j | g | \phi_j \phi'_p \rangle]$$

where the sum over j runs over all spin orbitals in $|\rangle$ except ϕ_p ;

(iii) if $|\rangle$ and $|\prime\rangle$ differ by two spin orbitals ($\phi_p \neq \phi'_p$ and $\phi_q \neq \phi'_q$),

$$\langle F + G \rangle = \langle \phi_p | f | \phi'_p \rangle + \sum_j [\langle \phi_p \phi_j | g | \phi'_p \phi_j \rangle - \langle \phi_p \phi_j | g | \phi_j \phi'_p \rangle]$$

(note that the F contribution vanishes in this case);

(iv) if $|\rangle$ and $|\prime\rangle$ differ by three or more spin orbitals, then

$$\langle F + G \rangle = 0$$

- [55] Nesbet R K 1963 Computer programs for electronic wave-function calculations *Rev. Mod. Phys.* **35** 552–7
It would seem that the process of evaluating all N^4 of the $\langle \phi_i \phi_j | g | \phi_k \phi_l \rangle$, each of which requires N^4 additions and multiplications, would require computer time proportional to N^8 . However, it is possible to perform the full transformation of the two-electron integral list in a time that scales as N^5 by first performing a transformation of the $\langle \chi_a \chi_b | g | \chi_c \chi_d \rangle$ to an intermediate array $\langle \chi_a \chi_b | g | \chi_c \phi \rangle = \sum_d C_{d,i} \langle \chi_a \chi_b | g | \chi_c \chi_d \rangle$ which requires N^5 multiplications and additions. The list $\langle \chi_a \chi_b | g | \chi_c \phi \rangle$ is then transformed to a second-level transformed array $\langle \chi_a \chi_b | g | \phi_k \phi_l \rangle = \sum_c C_{c,k} \langle \chi_a \chi_b | g | \chi_c \phi \rangle$, which requires another N^5 operations. This sequential transformation is repeated four times until the final $\langle \phi_i \phi_j | g | \phi_k \phi_l \rangle$ array is in hand
- [56] For early perspectives, see, for example:
Nesbet, R K 1965 Algorithm for diagonalization of large matrices *J. Chem. Phys.* **43** 311–12
Davidson E R 1976 The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices *J. Comput. Phys.* **17** 87–94
Roos B O and Siegbahn P E M 1977 The direct configuration interaction method from molecular integrals *Modern Theoretical Chemistry* vol 3, ed H F III Schaefer (New York: Plenum) pp 277–318
Roos B 1972 A new method for large-scale CI calculations *Chem. Phys. Lett.* **15** 153–9
For a good review, see:
Saunders V R and Van Lenthe J H 1983 The direct CI method a detailed analysis *Mol. Phys.* **48** 923–54
- [57] Davidson E 2000 webpage <http://php.indiana.edu/~davidson/>
Professor Davidson has contributed as much as anyone both to the development of the fundamentals of electronic structure theory and its applications to many perplexing problems in molecular structure and spectroscopy
- [58] The essential features of the MPPT/MBPT approach are described in the following articles:
Pople J A, Krishnan R, Schlegel H B and Binkley J S 1978 Electron correlation theories and their application to the study of simple reaction potential surface *Int. J. Quantum Chem.* **14** 545–60
Bartlett R J and Silver D M 1975 Many-body perturbation theory applied to electron pair correlation energies I. Closed-shell first-row diatomic hydrides *J. Chem. Phys.* **62** 3258–68

Krishnan R and Pople J A 1978
Approximate fourth-order perturbation
theory of the electron correlation
energy *Int. J. Quantum Chem.* **14** 91–
100

- [59] Kelly H P 1963 Correlation effects in atoms *Phys. Rev* **131** 684–99
Møller C and Plesset M S 1934 Note on an approximation treatment for many
electron systems *Phys. Rev* **46** 618–22
- [60] Szabo A and Ostlund N S 1989 *Modern Quantum Chemistry* 1st edn (revised)
(New York: McGraw-Hill) p 128

- [61] The early work in chemistry on this method is described in:
Cizek J 1966 On the correlation problem in atomic and molecular systems. Calculation of wave function components in Ursell-type expansion using quantum-field theoretical methods *J. Chem. Phys.* **45** 4256–66
Paldus J, Cizek J and Shavitt I 1972 Correlation problems in atomic and molecular systems IV. Extended coupled-pair many-electron theory and its application to the BH_3 molecule *Phys. Rev A* **5** 50–67
Bartlett R J and Purvis G D 1978 Many-body perturbation theory coupled-pair many-electron theory and the importance of quadruple excitations for the correlation problem *Int. J. Quantum Chem.* **14** 561–81
Purvis G D III and Bartlett R J 1982 A full coupled-cluster singles and doubles model. The inclusion of disconnected triples *J. Chem. Phys.* **76** 1910
- [62] Jørgensen P and Simons J 1981 *Second Quantization Based Methods in Quantum Chemistry* (New York: Academic) ch 4
- [63] Bartlett R 2000 webpage <http://www.qtp.ufl.edu/~bartlett>
Professor Bartlett brought the CC method, developed earlier by others, into the mainstream of electronic structure theory. For a nice overview of his work on the CC method see:
Bartlett R J 1995 Coupled-cluster theory: an overview of recent developments *Modern Electronic Structure Theory* vol 2, ed D R Yarkony (Singapore: World Scientific) pp 1047–131
- [64] Bartlett R J and Purvis G D 1978 Many-body perturbation theory coupled-pair many-electron theory and the importance of quadruple excitations for the correlation problem *Int. J. Quantum Chem.* **14** 561–81
Pople J A, Krishnan R, Schlegel H B and Binkley J S 1978 Electron correlation theories and their application to the study of simple reaction potential surfaces *Int. J. Quantum Chem.* **14** 545–60
- [65] Parr B 2000 webpage <http://net.chem.unc.edu/faculty/rgrp/cfrgp01.html>
Professor Parr was among the first to push the density functional theory of Hohenberg and Kohn to bring it into the mainstream of electronic structure theory. For a good overview, see the book:
Parr R G and Yang W 1989 *Density Functional Theory of Atoms and Molecules* (New York: Oxford University Press)
- [66] Hohenberg P and Kohn W 1964 Inhomogeneous electron gas *Phys. Rev B* **136** 864–72
- [67] The Hohenberg–Kohn theorem and the basis of much of density functional theory are treated:
Parr R G and Yang W 1989 *Density-Functional Theory of Atoms and Molecules* (New York: Oxford University Press)
The original paper relating to this theory is [66]
- [68] Professor Axel Becke of Queens University, Belfast has been very actively involved in developing and improving exchange-correlation energy functionals. For a good recent overview, see:
Becke A D 1995 Exchange-correlation approximations in density-functional theory *Modern Electronic Structure Theory* vol 2, ed D R Yarkony (Singapore: World Scientific) pp 1022–46
Becke A D 1983 Numerical Hartree–Fock–Slater calculations on diatomic molecules *J. Chem. Phys.* **76** 6037–45
- [69] Kohn W and Sham L J 1965 Self-consistent equations including exchange and correlation effects *Phys. Rev A* **140** 1133–8
- [70] Many of the various density functional approaches that are under active development can be found in:
Jones R O 1987 Molecular calculations with the density functional formalism *Advances in Chemical Physics* vol LXVII, ed K P Lawley (New York: Wiley–Interscience) pp 413–37
Dunlap B I 1987 Symmetry and degeneracy in $X\alpha$ and density functional theory *Advances in Chemical Physics* vol LXIX, ed K P Lawley (New York: Wiley–Interscience) pp 287–318
Dahl J P and Avery J (eds) 1984 *Local Density Approximations in Quantum Chemistry Solid State Physics* (New York: Plenum)
Parr R G 1983 Density functional theory *Ann. Rev. Phys. Chem.* **34** 631–56
Salahub D R, Lampson S H and Messmer R P 1982 Is there correlation in $X\alpha$ analysis of Hartree–Fock and LCAO $X\alpha$ calculations for O_3 *Chem. Phys. Lett.* **85** 430–3
-

Ziegler T, Rauk A
and Baerends E J
1977 On the
calculation of
multiplet energies by
the Hartree–Fock–
Slater method *Theor.*
Chim. Acta **43** 261–
71
Becke A D 1983
Numerical Hartree–
Fock–Slater
calculations on
diatomic molecules
J. Chem. Phys. **76**
6037–45
Case D A 1982
Electronic structure
calculation using the
 X_α method *Ann.*
Rev. Phys. Chem. **33**
151–71
Labanowski J K and
Andzelm J W (eds)
1991 *Density*
Functional Methods
in Chemistry (New
York: Springer)
For a recent critical
evaluation of
situations where
current DFT
approaches
experience
difficulties, see:
Davidson E R 1998
How robust is
present-day DFT?
Int. J. Quantum
Chem. **69** 241–5

- [71] This is because no four-indexed two-electron integral like expressions enter into the integrals needed to compute the energy. All such integrals involve $\rho(\mathbf{r})$ or the product $\rho(\mathbf{r})\rho(\mathbf{r})$; because ρ is itself expanded in a basis (say of M functions), even the term $\rho(\mathbf{r})\rho(\mathbf{r})$ scales no worse than M^2 . The solution of the KS equations for the KS orbitals ϕ_i involves solving a matrix eigenvalue problem; this is expected to scale as M^3 . However, as discussed in [section \(B3.1.8\)](#), the scalings of the DFT, SCF, and MP2 methods have been reduced even further
- [72] Pacific Northwest National Laboratories is developing a suite of programs called NWChem
Pacific Northwest National Laboratories 2000 webpage <http://www.emsl.pnl.gov:2080/>
The MacroModel program of Professor C Still, Columbia University webpage
<http://www.cc.columbia.edu/~chempub/mmod/mmod.html>
The Gaussian suite of programs webpage <http://www.gaussian.com>
The GAMESS program webpage <http://www.msg.ameslab.gov/GAMESS/GAMESS.html>
The HyperChem programs of Hypercube, Inc webpage <http://www.hyper.com>
The CAChe software packages from Oxford Molecular webpage <http://www.oxmol.com/getinfo/eduf>
The MOPAC program of CambridgeSoft webpage <http://www.camsoft.com>
The Amber program of Professor Peter Kollman, University of California, San Francisco webpage
<http://www.amber.ucsf.edu/amber/amber.html>
The CHARMM program webpage charmm-bbs-request@emperor.harvard.edu
The programs of MSI, Inc webpage <http://www.msi.com/info/index.html>
The COLUMBUS program webpage shavitt@mps.ohio-state.edu
The CADPAC program of Dr Roger Amos webpage
<http://www.cray.com/PUBLIC/DAS/files/CHEMISTRY/CADPAC.txt>
The programs of Wavefunction, Inc webpage <http://wavefun.com/>

The ACES II program of Professor Rod Bartlett webpage <http://www.qtp.ufl.edu/Aces2/>
The MOLCAS program of Professor Bjorn Roos webpage teobor@garm.teokem.lu.se
A nice compendium of various softwares is given in the appendix of reviews in:
Lipkowitz K B and Boyd D B (eds) 1996 *Computational Chemistry* vol 7 (New York: VCH)

- [73] Friesner R 2000 webpage <http://www.columbia.edu/cu/chemistry/faculty/raf.html>
Professor Friesner built on earlier developments of:
Beebe N H F and Linderberg J 1977 Simplifications in the generation and transformation of two-electron integrals in molecular calculations *Int. J. Quantum Chem.* **12** 683–705
Feyereisen M, Fitzgerald G and Komornicki A 1993 Use of approximate integrals in *ab initio* theory an application in MP2 energy calculations *Chem. Phys. Lett.* **208** 359–63
to develop the pseudospectral methods that he and others now widely use. See:
Friesner R A 1987 Solution of the Hartree–Fock equations for polyatomic molecules by a pseudospectral method *J. Chem. Phys.* **86** 3522–31
- [74] Carter E 2000 webpage <http://www.chem.ucla.edu/dept/Faculty/carter.html>
For an overview of Professor Carter's group's work using pseudospectral methods, see:
Martinez T J and Carter E A 1995 Pseudospectral methods applied to the electron correlation problem *Modern Electronic Structure Theory* vol 2, ed D R Yarkony (Singapore: World Scientific) pp 1132–65
- [75] Hylleraas E A and Undheim B 1930 *Z. Phys.* **65** 759
MacDonald J K L 1933 Successive approximations by the Rayleigh–Ritz variation method *Phys. Rev* **43** 830–3
- [76] Pople J A 1973 Theoretical models for chemistry *Energy, Structure, and Reactivity* ed D W Smith and W B McRae (New York: Wiley) p 51–67

-50-

- [77] Roos B O, Taylor P R and Siegbahn P E M 1980 A complete active space SCF method (CASSCF) using a density matrix formulated super-CI approach *Chem. Phys.* **48** 157–73
Roos B O 1987 The complete active space self-consistent field method and its applications in electronic structure calculations *Adv. Chem. Phys.* **69** 399–445
- [78] Kelly H P 1963 Correlation effects in atoms *Phys. Rev* **131** 684–99
- [79] Good early overviews of the electron propagator (that is used to obtain IP and EA data) and of the polarization propagator are given in:
Jørgensen P and Simons J 1981 *Second Quantization Based Methods in Quantum Chemistry* (New York: Academic)
The very early efforts on these methods are introduced in:
Linderberg J and Öhrn Y 1973 *Propagator Methods in Quantum Chemistry* (New York: Academic)
More recent summaries include:
Cederbaum L S and Domcke W 1977 Theoretical aspects of ionization potentials and photoelectron spectroscopy a Green's function approach *Adv. Chem. Phys.* **36** 205–344
Oddershede J 1987 Propagator methods *Adv. Chem. Phys.* **69** 201–39
Ortiz J V 1997 The electron propagator picture of molecular electronic structure *Computational Chemistry: Reviews of Current Trends* vol 2, ed J Leszczynski (Singapore: World Scientific) pp 1–61
- [80] The introduction of EOMs for energy differences and for operators that connect two states appears first in the nuclear physics literature; see for example:
Rowe D J 1968 Equation-of-motion method and the extended shell model *Rev. Mod. Phys.* **40** 153–66
I applied these ideas to excitation energies in atoms and molecules in 1971; see equation (2.1)–(2.6) in:
Simons J 1971 Direct calculation of first- and second-order density matrices. The higher RPA method *J. Chem. Phys.* **55** 1218–30
In 1973, the EOM method was then extended to treat IP and EA cases:
Simons J 1973 Theory of electron affinities of small molecules *J. Chem. Phys.* **58** 4899–907
In a subsequent treatment from the time-dependent response point of view, connection with the Greens function methods was made:
Simons J 1972 Energy-shift theory of low-lying excited electronic states of molecules *J. Chem. Phys.* **57** 3787–92
A more recent overview of much of the EOM, Greens function, and propagator field is given in:
Oddershede J 1987 Propagator methods *Adv. Chem. Phys.* **69** 201–39
- [81] Olsen J and Jørgensen P 1995 Time-dependent response theory with applications to self-consistent field and multiconfigurational self-consistent field wave functions *Modern Electronic Structure Theory* vol 2, ed D R Yarkony (Singapore: World Scientific) pp 857–990
- [82] A good overview of the recent status is given in:
Bartlett R J 1995 Coupled-cluster theory: an overview of recent developments *Modern Electronic Structure Theory* vol 2, ed D R Yarkony (Singapore: World Scientific) pp 1047–131

- [83] Helgaker T, Gauss J, Jørgensen P and Olsen J 1997 The prediction of molecular equilibrium structures by the standard electronic wave functions *J. Chem. Phys.* **106** 6430–40
for a listing and for further details on this study
- [84] Two review papers that introduce and compare the myriad of semi-empirical methods:
Stewart J J P 1991 Semiempirical molecular orbital methods *Reviews in Computational Chemistry* vol 1, ed K B Lipkowitz and D B Boyd (New York: VCH) pp 45–81
Zerner M C 1991 Semiempirical molecular orbital methods *Reviews in Computational Chemistry* vol 2, ed K B Lipkowitz and D B Boyd (New York: VCH) 313–65
A very recent overview, including efforts to interface semi-empirical electronic structure with molecular mechanics treatments of some degrees of freedom is given by:
Thiel W 1996 Perspectives on semiempirical molecular orbital theory *New Methods in Computational Quantum Mechanics (Adv. Chem. Phys. XCIII)* ed I Prigogine I and S A Rice (New York: Wiley) pp 703–57
Earlier texts dealing with semi-empirical methods include:
Pople J A and Beveridge D L 1970 *Approximate Molecular Orbital Theory* (New York: McGraw-Hill)
Murrell J N, Kettle S F A and Tedder J M 1965 *Valence Theory* 2nd edn (London: Wiley)
-

-1-

B3.2 Quantum structural methods for the solid state and surfaces

Frank Starrost and Emily A Carter

B3.2.1 INTRODUCTION

We are entering an era when condensed matter chemistry and physics can be predicted from theory with increasing realism and accuracy. This is particularly important in cases where experiments lead to ambiguous conclusions, for regimes in which there still exists no experimental probe and for predictions of the properties of modern materials in order to select the most promising ones for synthesis and experimental testing. For example, continuing miniaturization in microelectronics heightens the importance of understanding of quantum effects, which computational materials theory is poised to provide, based to some degree on the methods presented here.

Our intention is to give a brief survey of advanced theoretical methods used to determine the electronic and geometric structure of solids and surfaces. The electronic structure encompasses the energies and wavefunctions (and other properties derived from them) of the electronic states in solids, while the geometric structure refers to the equilibrium atomic positions. Quantities that can be derived from the electronic structure calculations include the electronic (electron energies, charge densities), vibrational (phonon spectra), structural (lattice constants, equilibrium structures), mechanical (bulk moduli, elastic constants) and optical (absorption, transmission) properties of crystals. We will also report on techniques used to study solid surfaces, with particular examples drawn from chemisorption on transition metal surfaces.

In his chapter on the fundamentals of quantum mechanics of condensed phases ([A1.3](#)), James R Chelikowsky introduces the plane wave pseudopotential method. Here, we will complement his chapter by introducing in some detail tight-binding methods as the simplest pedagogical illustration of how one can construct crystal wavefunctions from atomic-like orbitals. These techniques are very fast but generally not very accurate. After reviewing some of the efforts made to improve upon the local density approximation (LDA, explained in [A1.3](#)), we will discuss general features of the technically more complex all-electron band structure methods, focusing on the highly accurate but not very fast linear augmented plane wave (LAPW) technique as an example. We will introduce the idea of orbital-free electronic structure methods based directly on density functional theory (DFT), the computational effort of which scales linearly with size, allowing very large

systems to be studied. The periodic Hartree–Fock (HF) method and the promising quantum Monte Carlo (QMC) techniques will be briefly sketched, representing many-particle approaches to the condensed phase electronic structure problem.

In the final section, we will survey the different theoretical approaches for the treatment of adsorbed molecules on surfaces, taking the chemisorption on transition metal surfaces, a particularly difficult to treat yet extremely relevant surface problem [1], as an example. While solid state approaches such as DFT are often used, hybrid methods are also advantageous. Of particular importance in this area is the idea of embedding, where a small cluster of surface atoms around the adsorbate is treated with more care than the surrounding region. The advantages and disadvantages of the approaches are discussed.

-2-

B3.2.2 TIGHT-BINDING METHODS

B3.2.2.1 TIGHT BINDING: FROM EMPIRICAL TO SELF-CONSISTENT

The wavefunction in a solid can be thought to originate from two different limiting cases. One extreme is the nearly free electron (NFE) approach. The idea here is that the valence electrons are hardly affected by the periodic potential of the atomic cores. Their wavefunctions can then be assumed to be easily described as linear combinations of the solutions for free electrons: the plane waves, $\exp(i\mathbf{k} \cdot \mathbf{r})$. The NFE approximation is particularly useful for so-called NFE metals, such as the alkali metals. At the other extreme, the solid can be viewed as constructed from individual atoms. The valence wavefunctions of the solid are then approximated as linear combinations of the wavefunctions of the valence electrons of the atoms (see also [section A1.3.5.6](#)). In this case, the electrons are considered to be ‘tightly bound’ to the atoms. This is a physically reasonable view of covalently bound solids and molecules, where localized chemical bonds are the norm (bulk silicon, organic or biomolecules etc). Methods which employ this view of the electrons in the solid are called tight-binding (TB) methods. The wavefunctions are generally expanded in atomic orbitals (in a linear combination of atomic orbitals (LCAO) formalism) or similarly localized functions.

An advantage of TB is that generally the number of basis functions linearly combined to give the wavefunctions is rather small. The solution of the Schrödinger equation in these bases is then fast because the matrices representing the operators are small. Also, the construction of the Hamiltonian matrix elements is fast, since generally a number of, sometimes drastic, approximations are made. At the same time, however, the small basis set generally limits the quality of the TB results, since the variational freedom for the solution of the Schrödinger equation is not as high as in other methods. The approximations of Hamiltonian matrix elements often further reduce the quality of the results.

Today, the term TB method is generally understood to refer to a technique using TB basis functions in which the Hamiltonian matrix elements are adjusted to reproduce results from experiments and/or from more sophisticated electronic structure methods [2]. Depending on the degree of dependence on external parameters, the methods are called empirical or semi-empirical TB. A number of approaches are used for the fitting of the TB parameters, generally a tough minimization task with many minima (using genetic algorithms has proved quite efficient [3, 4]). It has been noted that ‘great care is needed to test the resulting model for reasonable behavior outside the range of the fit’ [5, 6]. A disadvantage of the empirical methods is that it is difficult to distinguish to what extent the parametrization or the method itself is responsible for errors in the results.

Frequent approximations made in TB techniques in the name of achieving a fast method are the use of a minimal basis set, the lack of a self-consistent charge density, the fitting of matrix elements of the potential,

the assumption of an orthogonal overlap matrix, a cut-off radius used in the integration to determine matrix elements, and the neglect of matrix elements that require three-centre integrals and crystal-field terms. We will now provide more details on these approximations.

Generally, the following *ansatz* for the wavefunction is made:

$$\psi_i(\mathbf{r}) = \sum_{\alpha l} c_{\alpha l}^i \varphi_{\alpha l}(\mathbf{r}),$$

-3-

where $\varphi_{\alpha\lambda}(\mathbf{r}) = \langle \mathbf{r} | \varphi_{\alpha\lambda} \rangle$ represents an atomic orbital of symmetry α (such as s, p_x, p_y, p_z) at atom l .

This yields the generalized eigenvalue problem

$$\underline{H}\underline{c}^i = \epsilon_i \underline{S}\underline{c}^i, \quad (\text{B3.2.1})$$

with the elements of the Hamiltonian matrix $H_{\alpha\lambda\beta\mu} \equiv \langle \varphi_{\alpha\lambda} | \mathbf{H} | \varphi_{\beta\mu} \rangle$ and the overlap matrix $S_{\alpha\lambda\beta\mu} \equiv \langle \varphi_{\alpha\lambda} | \varphi_{\beta\mu} \rangle$. In the TB approximation, the basis functions are thought to be sufficiently localized such that contributions to the Hamiltonian matrix usually are accounted for only up to at most the third or fourth neighbour. Frequently a minimal basis set is used, i.e. a single orbital $\varphi_{\alpha\lambda}$ is used per atom and per orbital symmetry to expand the wavefunction.

In orthogonal TB methods, the overlap matrix is assumed to be diagonal, even though the basis functions of adjacent sites ordinarily are not orthogonal [6]. Harrison has shown that this approximation can be compensated for by adjustments to the Hamiltonian matrix elements (these adjustments are arrived at automatically in methods depending on fitting, for example, a DFT band structure) [7]. However, this approach reduces the transferability of the TB parameters to other structures [8]. Including the overlap matrix brings with it the additional cost of its calculation and solving the generalized eigenvalue problem, see equation (B3.2.1), rather than an ordinary eigenvalue problem.

One can construct an effective potential, written here in the DFT language (see, for example, [equation A1.3.38](#) of A1.3) as

$$v_{\text{eff}}(\mathbf{r}) = v_{\text{ext}}(\mathbf{r}) + v_{\text{H}}[\rho(\mathbf{r})] + v_{\text{xc}}[\rho(\mathbf{r})]. \quad (\text{B 3.2.2})$$

To rationalize the ‘two-centre approximation’, the effective potential is written as

$$v_{\text{eff}}(\mathbf{r}) = \sum_l v_{\text{eff},l}(|\mathbf{r} - \mathbf{R}_l|),$$

where $v_{\text{eff},l}$ is centred on the atom l and vanishes away from the atom, which need not involve any approximation.

In the calculation of the elements

$$H_{\alpha l \beta m} = \left\langle \varphi_{\alpha l} \left| T + \sum_n v_{\text{eff},n} \right| \varphi_{\beta m} \right\rangle$$

with $T = -\frac{1}{2}\nabla^2$ the kinetic energy operator, several types of potential matrix elements can be distinguished [6]:

- (1) Three-centre terms, i.e. $l \neq m \neq n$. These are frequently neglected, in what is called the two-centre approximation, based on the assumed strong localization of the orbitals $\varphi_{\alpha\lambda}(\mathbf{r})$.
-

-4-

- (2) Inter-atomic two-centre matrix elements $\langle \varphi_{\alpha\lambda} | v_{\text{eff},l} + v_{\text{eff},m} | \varphi_{\beta\mu} \rangle$. These matrix elements represent the hopping of electrons from one site to another. They can be described [7] as linear combinations of so-called Slater–Koster elements [9]. The coefficients depend only on the orientation of the atoms l and m in the crystal. For elementary metals described with s, p, and d basis functions there are ten independent Slater–Koster elements. In the traditional formulation, the orientation is neglected and the two-centre elements depend only on the distance between the atoms [6]. (In several models [6, 10], they have been made dependent on the environment of the atoms l and m .) These elements are generally fitted to reproduce DFT results such as the band structure or the values of DFT matrix elements in diatomics.
- (3) Intra-atomic matrix elements, or on-site terms, with $l = m$. Traditionally, the potential contributions from other atomic sites, $v_{\text{eff},n \neq \lambda = \mu}$, so-called crystal-field terms, are neglected [10]. In this case, then the only non-zero on-site terms have $\alpha = \beta$, since basis functions on the same site are orthogonal atomic orbitals. There are methods which include these crystal-field terms [11, 12]. Physically, these diagonal elements represent the energy required to place an electron in a specific orbital. In some implementations, they are set to the orbital energy values of the neutral free atom [13], guaranteeing the correct limit for isolated atoms. However, this approach ignores the potential contributions to the diagonal elements due to different environments in a molecule or crystal; these are taken into account in other variants of the method [6, 10, 11].

Most TB approaches are not charge self-consistent. This means that they do not ensure that the charge derived from the wavefunctions yields the effective potential v_{eff} assumed in their calculation. Some methods have been developed which yield charge densities consistent with the electronic potential [14, 15 and 16].

The localized nature of the atomic basis set makes it possible to implement a linear-scaling TB algorithm, i.e. a TB method that scales linearly with the number of electrons simulated [17]. (For more information on linear scaling methods, see [section B3.2.3.3](#).)

The accuracy of most TB schemes is rather low, although some implementations may reach the accuracy of more advanced self-consistent LCAO methods (for examples of the latter see [18, 19 and 20]). However, the advantages of TB are that it is fast, provides at least approximate electronic properties and can be used for quite large systems (e.g., thousands of atoms), unlike some of the more accurate condensed matter methods. TB results can also be used as input to determine other properties (e.g., photoemission spectra) for which high accuracy is not essential.

B3.2.2.2 APPLICATIONS OF TIGHT-BINDING METHODS

TB methods have been widely used to study properties of simple semiconductors such as Si [11] and GaN [16]. In the latter study, the effect of dislocations on the electronic structure of GaN was investigated with a view toward understanding how dislocations affect the material's optical properties. The large supercell of 224 atoms led to TB as the method of choice. This particular variant of TB fits TB matrix elements to DFT-LDA results and solves self-consistently for atomic charges. It has also been used to predict reaction energetics of organic molecules, the structure of large biomolecules and the surface geometry and band

structure of III–V semiconductors [15]. The TB method is expected to provide qualitatively reasonable results for systems where localized atomic charges make sense and hence is not expected to perform as well for metallic systems. Despite potential problems of TB for metals, the TB approach has also been used to study the phonon spectrum of the transition metal molybdenum [6], the elastic constants,

-5-

vacancies and surfaces of monatomic transition and noble metals and the Hall coefficient of complex perovskite crystals [10]. As an example of data available from a TB calculation, a TB variant of extended Hückel theory [21, 22] was used to describe the initial states in photoemission from GaN [23]. The parameters were fitted to the bulk band structure $E_n(\mathbf{k})$ (for a definition, see section A1.3.6). As displayed in figure (B3.2.1) good agreement is found for the occupied states (negative energies), while larger differences for the conduction bands (positive energies) reveal a typical problem of the TB methods: they are far less capable of describing the delocalized conduction band states (the same is true for delocalized valence states in a metal, as mentioned above). In figure (B3.2.2) we show a series of calculated photoemission spectra compared to experimental results [23]. The dispersion of the main peaks as a function of emission angle and photon energy agrees reasonably well in theory and experiment.

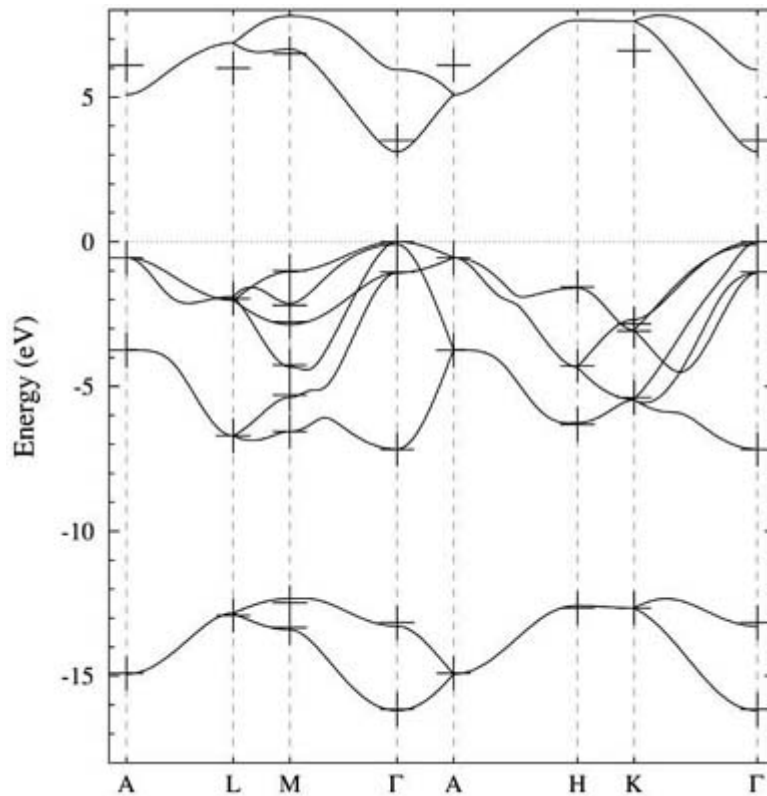


Figure B3.2.1. The band structure of hexagonal GaN, calculated using EHT-TB parameters determined by a genetic algorithm [23]. The target energies are indicated by crosses. The target band structure has been calculated with an *ab initio* pseudopotential method using a quasiparticle approach to include many-particle corrections [194].

-6-

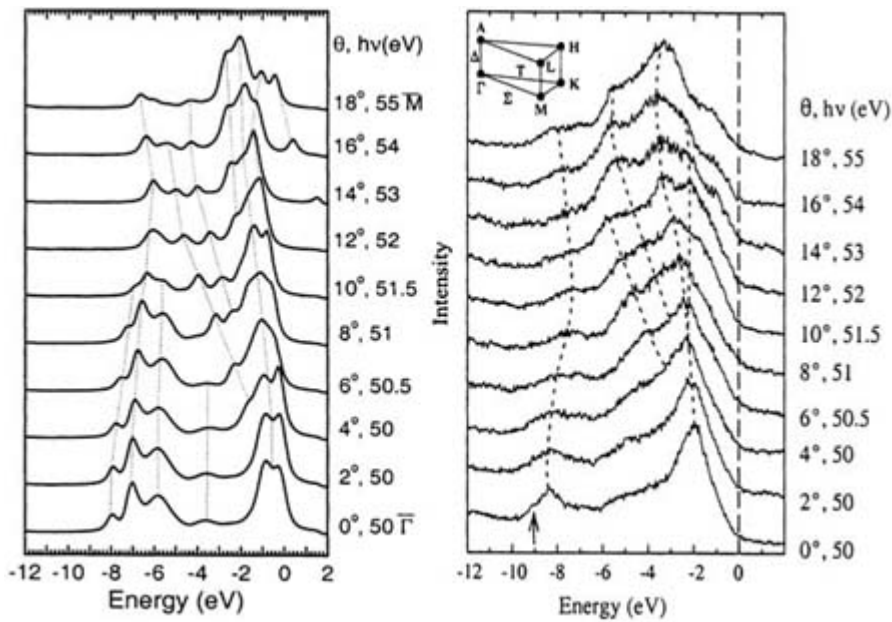


Figure B3.2.2. A series of photoemission spectra. The angles give the polar angle of electron emission at the stated photon energy scanning the surface Brillouin zone from $\bar{\Gamma}$ to \bar{M} . Left: A calculation using the tight-binding parametrization (given the band structure in figure (B3.2.1)) for the initial states [23]. Right: Experimental spectra by Dhesi *et al* [195]. The difference in binding energies is due to the experimental difficulty in determining the Fermi energy [23]. (Experimental figure by Professor K E Smith.)

B3.2.3 FIRST-PRINCIPLES ELECTRONIC STRUCTURE METHODS

In this section, we briefly review the basic elements of DFT and the LDA. We then focus on improvements suggested to remedy some of the shortcomings of the LDA (see section B3.2.3.1). A wide variety of techniques based on DFT have been developed to calculate the electron density. Many approaches do not calculate the density directly but rather solve for either a set of single-electron orbitals, or the Green's function, from which the density is derived.

In section B3.2.3.2, we introduce a number of techniques commonly referred to as *ab initio* all-electron electronic structure methods. *Ab initio* methods, in particular, aim at calculating the energies of electrons and their wavefunctions as accurately as possible, introducing as few adjustable parameters as possible. (Empirical or semi-empirical methods include the empirical pseudopotential approach (see section A1.3.5.5) and many TB techniques (see section B3.2.2).) Within the *ab initio* band structure approach, two communities exist that differ in their treatment of the singular nature of realistic, Coulomb-like crystal potentials. In the pseudopotential approach discussed by Chelikowsky in chapter A1.3, the Coulomb singularity ($-Z/r$) of the crystal potential is replaced by a smoother function, whereas in the so-called 'all-electron' approach, the Coulomb singularity is retained. The pseudopotential transformation limits the range of electron energies which can be accessed. However, since the pseudo-wavefunction is much smoother than the all-electron wavefunction (which has large oscillations near the nucleus), the pseudopotential allows the use of a plane

wave basis set, which is comparatively easy to handle. In principle, the all-electron methods have no limitation on the energy range of calculations. This is achieved by a sophisticated representation of the wavefunction.

The so-called orbital-free DFT technique, which aims to directly calculate the electron density for which the

total energy is minimal, is presented as an example of methods whose computational effort scales linearly with system size (see [section B3.2.3.3](#)). In [section B3.2.3.4](#), we discuss the periodic HF method, an alternative approach to DFT that offers a well defined starting point for many-particle corrections. Finally, the two most frequently used QMC techniques are described in [section B3.2.3.5](#).

B3.2.3.1 THE LOCAL DENSITY APPROXIMATION AND BEYOND

In DFT, the electronic density rather than the wavefunction is the basic variable. Hohenberg and Kohn showed [24] that all the observable ground-state properties of a system of interacting electrons moving in an external potential $v_{\text{ext}}(\mathbf{r})$ are uniquely dependent on the charge density $\rho(\mathbf{r})$ that minimizes the system's total energy. However, there is no known formula to calculate from the density the total energy of many electrons moving in a general potential. Hohenberg and Kohn proved that there exists a universal functional of the density, called $G[\rho]$, such that the expression

$$E[\rho] = \int v_{\text{ext}}(\mathbf{r})\rho(\mathbf{r}) d^3r + \frac{1}{2} \int \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d^3r d^3r' + G[\rho] \quad (\text{B 3.2.3})$$

has as its minimum value the correct ground-state energy associated with $v_{\text{ext}}(\mathbf{r})$. Here, the first term on the right-hand side represents the energy due to an external potential, including the electron–nuclear potential, while the second term is the classical Coulomb energy of the electronic system. The functional $G[\rho]$ is valid for any number of electrons and any external potential, but it is unknown and further steps are necessary to approximate it.

Kohn and Sham [25] decompose $G[\rho]$ into the kinetic energy of an analogous set of non-interacting electrons with the same density $\rho(\mathbf{r})$ as the interacting system,

$$T_s[\rho] = \sum_i \left\langle \psi_i \left| -\frac{1}{2} \nabla^2 \right| \psi_i \right\rangle$$

(where $\psi_i(\mathbf{r}) = \langle \mathbf{r} | \psi_i \rangle$ is the wavefunction of electron i), and the exchange and correlation energy of an interacting system with density $\rho(\mathbf{r})$, $E_{\text{xc}}[\rho]$. The functional $E_{\text{xc}}[\rho]$ is not known exactly. Physically, it represents all the energy corrections beyond the Hartree term to the independent-particle model, i.e. the non-classical many-body effects of exchange and correlation (xc) and the difference between the kinetic energy of the interacting electron system $T[\rho]$ and the analogous non-interacting system $T_s[\rho]$.

In the LDA, the exchange and correlation energy is approximated using the exchange and correlation energy of the homogeneous electron gas at the same density (see [section A1.3.3.3](#)). The crystal density is obtained by solving the single-particle Kohn–Sham equation

-8-

$$\left(-\frac{1}{2} \nabla^2 + v_{\text{eff}}(\mathbf{r}) \right) \psi_i(\mathbf{r}) = E_i \psi_i(\mathbf{r}), \quad (\text{B 3.2.4})$$

for a self-consistent potential v_{eff} i.e. a potential which is produced by the density ρ . In bulk crystal calculations, the index i runs over both the Bloch vector \mathbf{k} (see [section A1.3.4](#)) and the band index n (in a simple crystal, this band could be derived, for example, entirely from s states). The solutions to equation (B3.2.4) are often called Kohn–Sham orbitals. The crystal density is then

$$\rho(\mathbf{r}) = \sum_i \psi_i^*(\mathbf{r})\psi_i(\mathbf{r}).$$

The eigenenergy E_i can be defined as the derivative of the total energy of the many-electron system with respect to the occupation number of a specific orbital [26]. In HF theory (where equation (B3.2.4) applies and the v_{eff} contains a non-local exchange operator, see section A1.3.1.2 and [chapter B3.1](#)), Koopmans' theorem states that the single-particle eigenvalue is the negative of the ionization energy (neglecting the relaxation of the electronic system). In contrast, the identification of the highest occupied Kohn–Sham eigenvalue with the negative of the ionization energy is a controversial subject [27]. While there is no rigorous connection between eigenvalue differences and excitation energies in either HF or DF theory, comparisons of these values are common practice (see below for more appropriate methods). Relative differences among occupied single-particle energies often agree well with the experiment. Even though DFT only provides a solution for the ground state of the electronic system, the energy differences in the lower conduction bands, i.e. low-energy excited states, often are represented surprisingly well, too. However, in LDA calculations of semiconductors and insulators, almost always the size of the gap between the valence band maximum and the conduction band minimum is underestimated, since many-particle effects are incorrectly represented by the parametrized exchange–correlation energy (see, for example, [28]). One *ad hoc* remedy, which works well for many systems and which is employed in the examples presented here, is to use what is amusingly referred to as a *scissor operator*, i.e. a rigid shift, to correct the gap size [29, 30]. Typically the shift is determined by knowing, for example, the DFT error in predicting the measured optical band gap. The entire conduction band is shifted rigidly upward by the amount to match the experimental band gap.

More advanced techniques take into account quasiparticle corrections to the DFT-LDA eigenvalues. Quasiparticles are a way of conceptualizing the elementary excitations in electronic systems. They can be determined in band structure calculations that properly include the effects of exchange and correlation. In the LDA, these effects are modelled by the exchange–correlation potential $v_{\text{xc}}^{\text{LDA}}$. In order to more accurately account for the interaction between a particle and the rest of the system, the notion of a local potential has to be generalized and a non-local, complex and energy-dependent exchange–correlation potential has to be introduced, referred to as the self-energy operator $\Sigma(\mathbf{r}, \mathbf{r}'; E)$. The self-energy can be expanded in terms of the screened Coulomb potential W , where $W = \epsilon^{-1}v$ is the Coulomb interaction v screened by the inverse dielectric function ϵ^{-1} . In a lowest order expansion in W , the self-energy can be approximated as $\Sigma = GW$, giving the *GW* approximation [31]. Here G is the one-electron Green's function describing the propagation of an additional electron injected into a system of other electrons (it can also describe the extraction of an electron).

-9-

To be a bit more explicit (following [32, 33]), the quasiparticle energies and wavefunctions are given by

$$(T + v_{\text{ext}} + v_{\text{H}})\psi_{nk}(\mathbf{r}) + \int d\mathbf{r}' \Sigma(\mathbf{r}, \mathbf{r}'; E_{nk})\psi_{nk}(\mathbf{r}') = E_{nk}\psi_{nk}(\mathbf{r}),$$

where T is the kinetic energy operator, v_{ext} is the external potential due to the ions, and v_{H} is the Hartree Coulomb interaction. Since the self-energy operator in general is non-Hermitian, the quasiparticle energies E_{nk} are complex in general, and the imaginary part gives the lifetime of the quasiparticle. To first order in W , the self-energy is then given by

$$\Sigma(\mathbf{r}, \mathbf{r}'; E) = \frac{i}{2\pi} \int d\omega e^{-i\delta\omega} G(\mathbf{r}, \mathbf{r}'; E - \omega)W(\mathbf{r}, \mathbf{r}'; \omega)$$

where δ is a positive infinitesimal and ω corresponds to an excitation frequency. The inputs are the full interacting Green's function,

$$G(\mathbf{r}, \mathbf{r}'; E) = \sum_{nk} \frac{\psi_{nk}(\mathbf{r})\psi_{nk}^*(\mathbf{r}')}{E - E_{nk} - i\delta_{nk}},$$

where δ_{nk} is an infinitesimal and the dynamically screened Coulomb interaction,

$$W(\mathbf{r}, \mathbf{r}'; \omega) = \Omega^{-1} \int d\mathbf{r}'' \epsilon^{-1}(\mathbf{r}, \mathbf{r}''; \omega) v(\mathbf{r}'' - \mathbf{r}'),$$

where ϵ^{-1} is the inverse dielectric matrix, $v(\mathbf{r}) = 1/|\mathbf{r}|$ and Ω is the volume of the system. Usually the calculations start with the construction of the Green's function and the screened Coulomb potential from self-consistent LDA results. The self-energy Σ then has to be obtained together with G in a self-consistent procedure. However, due to the severe computational cost of this procedure, it is usually not carried out (see, for example, [34]). Instead, it is common practice to construct the self-energy operator non-self-consistently using the self-consistent LDA results to determine quasiparticle corrections to the LDA energies, resulting in the quasiparticle band structure. The GW approximation has been applied to a wide range of metals, semiconductors and insulators, where it has been found to lead to striking improvements in the agreement of optical excitation spectra with the experiment (see, for example [32, 35, 36 and 37]). Recent studies also found that the GW charge density is close to the experiment for diamond structure semiconductors [38], and lifetimes of low-energy electrons in metals have been calculated [39].

Another disadvantage of the LDA is that the Hartree Coulomb potential includes interactions of each electron with itself, and the spurious term is not cancelled exactly by the LDA self-exchange energy, in contrast to the HF method (see A1.3), where the self-interaction is cancelled exactly. Perdew and Zunger proposed methods to evaluate the self-interaction correction (SIC) for any energy density functional [40]. However, full SIC calculations for solids are extremely complicated (see, for example [41, 42 and 43]). As an alternative to the very expensive GW calculations, Pollmann *et al* have developed a pseudopotential built with self-interaction and relaxation corrections (SIRC) [44].

-10-

The pseudopotential is derived from an all-electron SIC-LDA atomic potential. The relaxation correction takes into account the relaxation of the electronic system upon the excitation of an electron [44]. The authors speculate that ‘...the ability of the SIRC potential to produce considerably better band structures than DFT-LDA may reflect an extra nonlocality in the SIRC pseudopotential, related to the nonlocality or orbital dependence in the SIC all-electron potential. In addition, it may mimic some of the energy and the non-local space dependence of the self-energy operator occurring in the GW approximation of the electronic many body problem’ [45].

The LDA also fails for strongly correlated electronic systems. Examples of such systems are the late 3d transition-metal mono-oxides MnO, FeO, CoO, and NiO. Within the local spin density approximation (LSDA), the energy gaps calculated for MnO and NiO are too small [46] and, even worse, FeO and CoO are predicted to be metallic, whereas experimentally they have been found to be large-gap insulators. While the GW approximation yields an energy gap of NiO in reasonable agreement with experiment [47], the computational cost of this procedure is very high. The SIC-LDA method reproduces quite well the strong localization of the d electrons in transition metal compounds, but the orbital energies obtained by SIC are usually in strong disagreement with experimental results (for transition metal oxides, for example, occupied d bands are approximately $\frac{1}{2}$ Hartree below the oxygen valence band—a separation not seen in spectroscopic data: see, for example, the experimental results in [48]) [49]. An alternative solution to this problem is offered by the LDA+ U method [49, 50], where LDA encompasses the LSDA. In the LDA+ U technique, the electrons are divided into two subsystems which are treated separately: the strongly localized (d or f) electrons and the delocalized s and p electrons. The latter are treated by standard LDA. The on-site interactions among the

strongly localized electrons on each atom, however, are taken into account by a term $\frac{1}{2}U \sum_{i \neq j} n_i n_j$, where n_i are the occupation numbers of the strongly localized orbitals and U is the Coulomb interaction parameter (for details on the first-principles calculation of U , see [51]). At least for localized d or f states, the LDA+ U technique may be viewed as an approximation to the GW approximation [49]. Band gaps, valence band widths and magnetic moments have been calculated with LDA+ U that agree with experiment for a variety of transition metal compounds [49, 52], among other applications.

B3.2.3.2 ALL-ELECTRON DFT METHODS

(A) INTRODUCTION

When the highest accuracy is sought for the electronic and geometric properties of crystals, all the electrons of the atoms in the crystal and the full Coulomb singularity of the nuclear potential must be accounted for. All-electron approaches, which do just that, generally cannot compete with pseudopotential techniques in speed and simplicity of algorithm. However, the latter suffer from severe drawbacks when it comes to the construction of the very pseudopotentials these methods depend upon: even for so-called *ab initio* potentials, the pseudopotentials are far from uniquely determined. Additionally, problems with transferability and the construction of potentials for such elements as the transition metals remain. All-electron techniques can deal with any element and there are no worries about transferability of the potential. However, the accuracy comes at a price: due to the Coulomb singularity of the potential at the nuclear positions, the wavefunctions are highly oscillatory close to the nucleus. For those all-electron methods that use wavefunctions to represent the electrons (a Green's function method, for example, does not), this means that a simple plane wave basis set cannot be used for the expansion of the wavefunctions. To reach convergence of a plane wave $\exp(i\mathbf{k} \cdot \mathbf{r})$ expansion would require a prohibitive number of basis functions. Thus, specialized basis sets have been invented for all-electron calculations.

We now discuss the most important theoretical methods developed thus far: the augmented plane wave (APW) and the Korringa–Kohn–Rostoker (KKR) methods, as well as the linear methods (linear APW (LAPW), the linear muffin-tin orbital [LMTO] and the projector-augmented wave [PAW]) methods.

In the early all-electron techniques, the crystal was separated into spheres around the atoms, so-called ‘muffin-tin’ spheres, and the interstitial region in between. Inside the spheres, the potential was approximated as spherically symmetric, while in the interstitial region it was assumed to be constant. This shape approximation of the potential is reasonable for close-packed crystals such as hexagonally close-packed metals, where the spheres cover a large fraction of the crystal volume. However, in less densely arranged crystals, such as diamond structure semiconductors (see figure (A1.3.4) the muffin-tin approximation leads to large errors. In the diamond and the related zincblende structures, only 34% of the volume is covered by touching muffin-tin spheres (figure (B3.2.3)). For all of the all-electron methods, versions have been developed that are not restricted to shape approximations of the potential. These techniques are referred to as general, or full, potential methods.

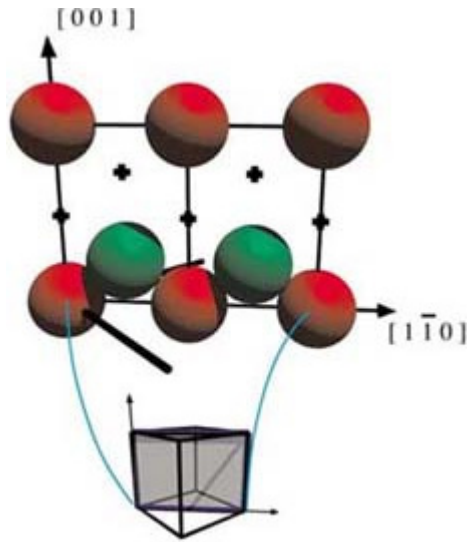


Figure B3.2.3. The muffin-tin spheres in the (110) plane of a zincblende crystal. The nuclei are surrounded by spheres of equal size, covering about 34% of the crystal volume. Unoccupied tetrahedral positions are indicated by crosses. The conventional unit cell is shown at the bottom; the crystal directions are noted.

(B) THE AUGMENTED PLANE WAVE METHOD

The APW technique was proposed by Slater in 1937 [53, 54]. It remains the most accurate of the band structure methods for the muffin-tin approximation of the potential. The wavefunction is expanded in basis functions $\varphi_i(\mathbf{k} + \mathbf{G}_i, E, \mathbf{r})$, the APWs, each of which is identical to the plane wave $\exp(i(\mathbf{k} + \mathbf{G}_i) \cdot \mathbf{r})$ in the interstitial region, where \mathbf{G}_i are the reciprocal lattice vectors (see section A1.3.4). The plane waves are augmented, i.e. they are joined continuously at the surface of the spheres by solutions of the radial Schrödinger equation. This means that in the spherical harmonic expansion of a plane wave around the centre of a muffin-tin sphere, the respective Bessel function inside the sphere is replaced by a solution $\phi_{li}(r, E)$ of the radial Schrödinger equation for a given energy. The radial function matches the Bessel function, $j_l(|\mathbf{k} + \mathbf{G}_i|r)$, value at the sphere boundary and must be regular (non-singular) at the origin.

-12-

With the basis functions $\varphi_i(\mathbf{k} + \mathbf{G}_i, E, \mathbf{r})$, a variational solution is sought to the Kohn–Sham equation, equation (B3.2.4). Since the Hamiltonian matrix elements now depend nonlinearly upon the energy due to the energy-dependent basis functions, the resulting secular equation is solved by finding the roots of the determinant of the $\underline{H}(E) - E\underline{S}(E)$ matrix. (The problem cannot be treated by the eigenvalue routines of linear algebra.)

Numerically, the determination of the roots can be difficult because the determinant's value may change by several orders of magnitude when the energy E is changed by only a few meV. Another difficulty can result at degenerate roots where the value of the determinant does not change sign. Additionally, the secular equation becomes singular when a node of the radial solution falls at the muffin-tin sphere boundary (the so-called 'asymptote problem').

Physically, the APW basis functions are problematic as they are not smooth at the sphere boundary, i.e., they have discontinuous slope. While in a fully converged solution of the secular equation, this discontinuity should disappear, alternative methods have been sought instead. Following a suggestion by Marcus [55] in 1967, the LAPW provided a way to avoid the above-mentioned drawbacks of the APW technique, as we now discuss.

(C) THE LINEAR AUGMENTED PLANE WAVE METHOD

The main disadvantage of the APW technique is that it leads to a nonlinear secular problem because the basis functions depend on the energy. A number of attempts have been made to construct linear versions of the APW approach by introducing energy-independent basis functions in different ways. In 1970, Koelling invented the *alternative* APW [56] and Bross the *modified* APW [57]. In 1975, Andersen constructed the LAPW [58] formalism, which today is the most popular APW-like band structure method. Further extensions of the linear methods appeared in the early 1990s: Singh developed the LAPW *plus localized orbitals* (LAPW+LO [59]) in 1991 and Krasovskii the *extended* LAPW (ELAPW [60]) in 1994. Recently the APW+LO technique has been implemented by Sjöstedt and Nordström [61] according to an idea by Singh. While the LAPW technique is generally used in combination with DFT approaches, it has also been applied based on the LDA+ U [62] and HF theories [63].

The LAPW method, as suggested in 1975 [58, 64], avoids the problem of the energy dependence of the Hamiltonian matrix by introducing energy-independent APW basis functions. Here, too, the APWs are derived from plane waves by augmentation: Bessel functions $j_l(|\mathbf{k} + \mathbf{G}_l|r)$ in the Rayleigh decomposition inside the muffin-tin sphere are replaced by functions $u_l(r)$ derived from the spherical potential, which are *independent* of the energy of the state that is sought and that match the Bessel functions at the sphere radius in value and in slope (see figure (B3.2.4)). The plane wave part of the basis remains the same but the energy-independent APWs allow the energies and the wavefunctions to be determined by solving a standard generalized eigenvalue problem.

-13-

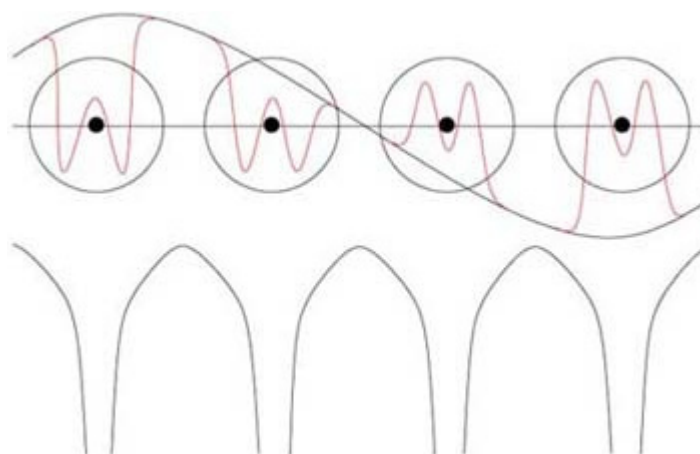


Figure B3.2.4. A schematic illustration of an energy-independent augmented plane wave basis function used in the LAPW method. The black sine function represents the plane wave, the localized oscillations represent the augmentation of the function inside the atomic spheres used for the solution of the Schrödinger equation. The nuclei are represented by filled black circles. In the lower part of the picture, the crystal potential is sketched.

In linearizing the APW problem as it is done in the LAPW method, the variational freedom of the APW basis set is reduced. The reason is that the wavefunction inside the spheres is rigidly coupled to its plane wave expansion in the interstitial region [65]. This means that the method cannot yield an accurate wavefunction even if the eigenvalue is within a few eV of the chosen energy parameters [66]. Flexibility is defined in this context as the possibility to change the wavefunction inside the spheres independently from the wavefunction in the interstitial region. Flexibility can be achieved in the linear band structure methods by adding basis functions localized inside the spheres whose value and slope vanish at the sphere boundary [54, 67, 68]. A ‘flexible’ basis set extending the LAPW with localized functions is preferable to the one used in the pure LAPW technique. Flexible linear methods are the MAPW, the LAPW+LO and the ELAPW, the latter of which provides a necessary degree of flexibility with a minimal number of basis functions [65].

The additional functions increase the matrix dimension slightly and thus the computational effort. However, the increased flexibility of the basis set makes possible a number of extensions of the LAPW method. One is a $\mathbf{k} \cdot \mathbf{p}$ formulation of the ELAPW method [68], which would lead to large errors in the regular LAPW due to its lesser flexibility. The augmented Fourier components (AFC) technique [69] for treating a general potential is based on this. The AFC method is an alternative to the full-potential LAPW (FLAPW) method [70, 71]. (Recently progress has been made in increasing the computational efficiency of the FLAPW method [72].) The AFC method does not have the same demanding convergence criteria as the FLAPW method but yields physically equivalent results [69].

The general potential LAPW techniques are generally acknowledged to represent the state of the art with respect to accuracy in condensed matter electronic-structure calculations (see, for example, [62, 73]). These methods can provide the best possible answer within DFT with regard to energies and wavefunctions.

-14-

(D) THE KORRINGA-KOHN-ROSTOKER TECHNIQUE

The KKR method uses multiple-scattering theory to solve the Kohn-Sham equations [74, 75]. Rather than calculate the wavefunction, modern incarnations calculate the Green's function G . The Green's function is the solution to the equation schematically given by $(H - E)G(E) = -\delta$, where H is the Hamiltonian, E the single-electron energy and δ the delta function $\delta(\mathbf{r} - \mathbf{r}')$. The properties of the system, such as the electron density, the density of states and the total energy can be derived from the Green's function [73]. The crystal is represented as a sum of non-overlapping potentials; in the modern version, there are no shape approximations, i.e. the potentials are space-filling [76]. Within the multiple-scattering formalism, the wavefunction is built up by taking into account the scattering and rescattering of a free-electron wavefunction by scatterers. The scatterers are (generally) the atoms of the crystal and the single-scattering properties (the properties of the isolated scatterer) are derived from the effective, singular potentials of the atoms (given in equation (B3.2.2)). The Green's matrix is then constructed from the knowledge of the scattering properties of the single scatterers and the analytically known Green's function of the free electron. The full-potential KKR method has been shown to have the same level of accuracy as the full-potential LAPW method [73]. The Green's function formulation offers the advantage of easy inclusion of defects in the bulk or clean surfaces. Such calculations start with the Green's function of the periodic crystal and include the perturbation through a Dyson equation [77]. Yussouff states that the difference in speeds between the linear methods and his 'fast' KKR technique is at most a factor of ten, in favour of the former [78]. While the KKR technique has an accuracy comparable to the APW method, it has the disadvantage of not being a linear approach, limiting speed and simplicity.

(E) THE LINEAR MUFFIN-TIN ORBITAL METHOD

The LMTO method [58, 79] can be considered to be the linear version of the KKR technique. According to official LMTO historians, the method has now reached its 'third generation' [79]: the first starting with Andersen in 1975 [58], the second commonly known as TB-LMTO. In the LMTO approach, the wavefunction is expanded in a basis of so-called muffin-tin orbitals. These orbitals are adapted to the potential by constructing them from solutions of the radial Schrödinger equation so as to form a minimal basis set. Interstitial properties are represented by Hankel functions, which means that, in contrast to the LAPW technique, the orbitals are localized in real space. The small basis set makes the method fast computationally, yet at the same time it restricts the accuracy. The localization of the basis functions diminishes the quality of the description of the wavefunction in the interstitial region.

In the commonly used atomic sphere approximation (ASA) [79], the density and the potential of the crystal are approximated as spherically symmetric within overlapping muffin-tin spheres. Additionally, all integrals, such as for the Coulomb potential, are performed only over the spheres. The limits on the accuracy of the method imposed by the ASA can be overcome with the full-potential version of the LMTO (FP-LMTO)

which gives highly accurate total energies [79, 80]. It was found that the FP-LMTO is ‘at least as accurate as, and much faster than,’ pseudopotential plane wave calculations in the determination of structural and dynamic properties of silicon [80]. The FP-LMTO is considerably slower than LMTO-ASA, however, and it has been found that ASA calculations can yield accurate results if the full expansion, rather than only the spherical part, of the charge is used in what is called a full-charge (rather than a full-potential) method and the integrals are performed exactly [73, 79].

The LMTO method is the fastest among the all-electron methods mentioned here due to the small basis size. The accuracy of the general potential technique can be high, but LAPW results remain the ‘gold standard’.

(F) THE PROJECTOR AUGMENTED WAVE TECHNIQUE

The projector augmented-wave (PAW) DFT method was invented by Blöchl to generalize both the pseudopotential and the LAPW DFT techniques [81]. PAW, however, provides all-electron one-particle wavefunctions not accessible with the pseudopotential approach. The central idea of the PAW is to express the all-electron quantities in terms of a pseudo-wavefunction (easily expanded in plane waves) term that describes interstitial contributions well, and one-centre corrections expanded in terms of atom-centred functions, that allow for the recovery of the all-electron quantities. The LAPW method is a special case of the PAW method and the pseudopotential formalism is obtained by an approximation. Comparisons of the PAW method to other all-electron methods show an accuracy similar to the FLAPW results and an efficiency comparable to plane wave pseudopotential calculations [82, 83]. PAW is also formulated to carry out DFT dynamics, where the forces on nuclei and wavefunctions are calculated from the PAW wavefunctions. (Another all-electron DFT molecular dynamics technique using a mixed-basis approach is applied in [84].)

PAW is a recent addition to the all-electron electronic structure methods whose accuracy appears to be similar to that of the general potential LAPW approach. The implementation of the molecular dynamics formalism enables easy structure optimization in this method.

(G) ILLUSTRATIVE EXAMPLES OF THE ELECTRONIC AND OPTICAL PROPERTIES OF MODERN MATERIALS

As an indication of the types of information gleaned from all-electron methods, we focus on one recent approach, the ELAPW method. It has been used to determine the band structure and optical properties over a wide energy range for a variety of crystal structures and chemical compositions ranging from elementary metals [60] to complex oxides [85], layered dichalcogenides [86, 87] and nanoporous semiconductors [88]. The $\mathbf{k} \cdot \mathbf{p}$ formulation has also enabled calculation of the complex band structure of the Al (100) surface [89].

As an illustration of the accuracy of the AFC ELAPW- $\mathbf{k} \cdot \mathbf{p}$ method, we present the dielectric function of GaAs. The dielectric function is a good gauge of the quality of a method, since not only do the energies enter the calculation, but also the wavefunctions *via* the matrix elements of the momentum operator $-i\nabla$. For the calculation of the dielectric function (equation (A1.3.87)) of GaAs, the conduction bands were rigidly shifted so that the highest peak agreed in both experiment and theory, a shift of 0.75 eV. The imaginary part of the dielectric function is shown in [figure B3.2.5](#). Comparing the energy differences between the three peaks, we find that they agree to within 2 meV. For a wider comparison, we plot the results of two more experiments (which only have measured the two peaks at lower photon energy) and several all-electron calculations of the dielectric function of GaAs in [figure B3.2.6](#). The FLAPW results agree almost exactly with the AFC ELAPW values. The discrepancies compared to the experimental results found for the other methods are considerably larger than for the general potential LAPW results, particularly for E_1 .

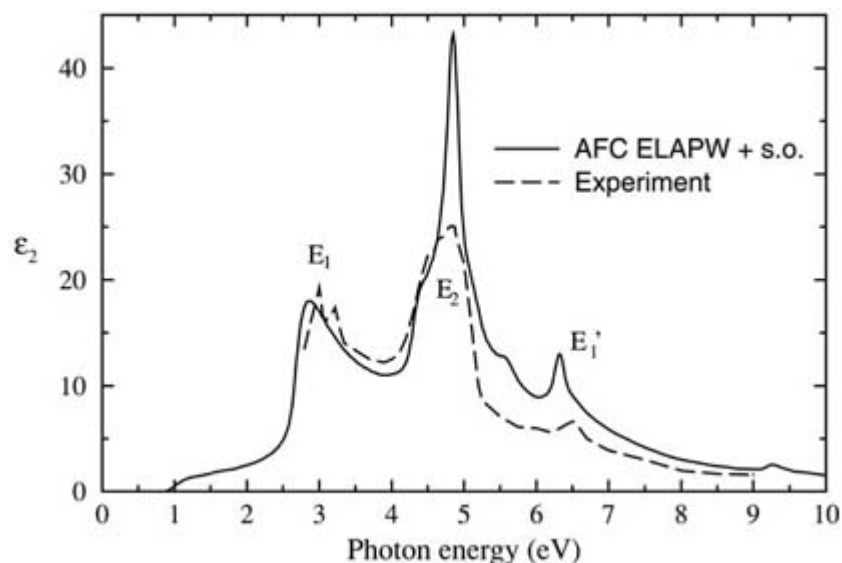


Figure B3.2.5. The imaginary part of the dielectric function of GaAs, according to the AFC ELAPW- $k\cdot p$ method (solid curve) [195] and the experiment (dashed curve) [196]. To correct for the band gap underestimated by the local density approximation, the conduction bands have been shifted so that the E_2 peaks agree in theory and experiment.

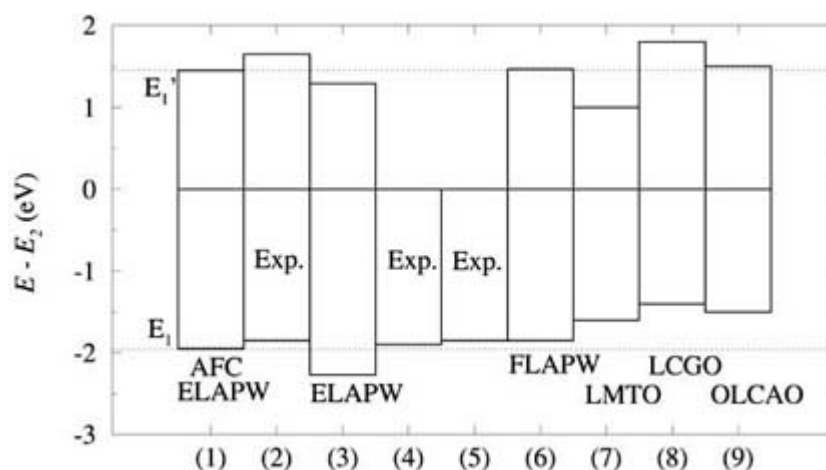


Figure B3.2.6. The energies of the E_1 and E'_1 peaks relative to the E_2 peak of the imaginary part of the dielectric function of GaAs, calculated by self-consistent DFT all-electron methods. These energies do not depend on the gap size. The theoretical methods are noted, as are experimental results obtained by ellipsometry (see [chapter B1.2.6](#)). The lower (upper) histogram gives the energy of peak E_1 (E'_1) relative to E_2 . LCGO designates a linear-combination-of-Gaussian-orbitals method, OLCAO an orthogonalized linear-combination-of-atomic-orbitals approach. Sources: (1) [195], (2) [196], (3) [197], (4) [198], (5) [199], (6) [200], (7) [199], (8) [201], (9) [202].

A recent study of a class of nanoporous materials, the cetineites [88], offers further illustration of the possibilities offered by the modern band structure methods. The crystal is constructed of tubes of 0.7 nm diameter arranged in a two-dimensional hexagonal structure with ‘flattened’ SbSe_3 pyramids arranged between the tubes (see figure B3.2.7). Cetineites are of potential technological interest because, singularly among nanoporous materials, they are semiconductors rather than insulators. In [figure B3.2.8](#), we show the

comparison of the predicted density of states to the ultraviolet photoemission spectrum (PES, see [chapter B1.1](#)). The DOS can explain the two main structures in the PES at about -3 and -12 eV. Their relative intensities agree with those suggested by the DOS curve. Three structures in the DOS at -1 , -6 and -9 eV are not resolved in the PES. This may be due to the selection rules of the photoemission process, not accounted for in the theory, or perhaps due to incomplete angle integration experimentally. The experimental results confirm, in particular, that the number of states is very high close to the valence band maximum. An orbital analysis shows that these states are derived mainly from the p states of the O and Se constituents of the crystal, with the chalcogen dominating near the top of the valence band. Electrons in the Se p states are thus most easily excited into the conduction band. This, together with their high DOS, makes the Se p states located on the pyramids the prime candidates for the initial states of the photoconductivity observed in the cetineites.

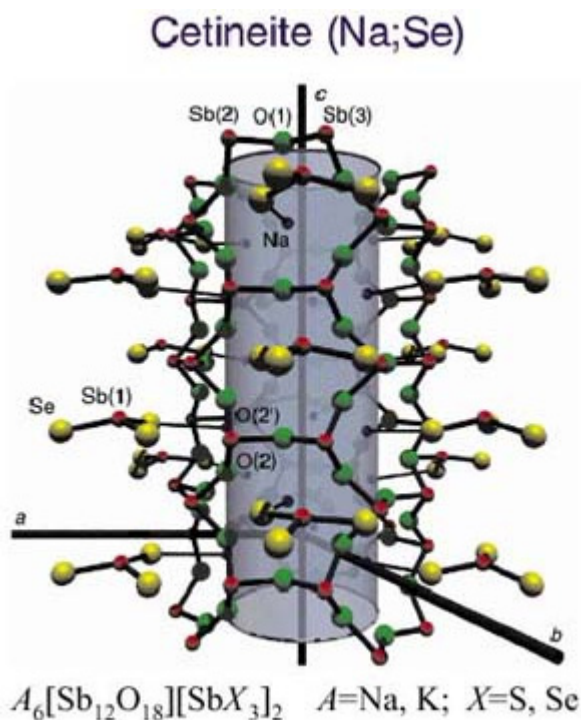


Figure B3.2.7. A perspective view of the cetineite (Na;Se). The height of the figure is three lattice constants c . The shaded tube is included only as a guide to the eye. (From [88].)

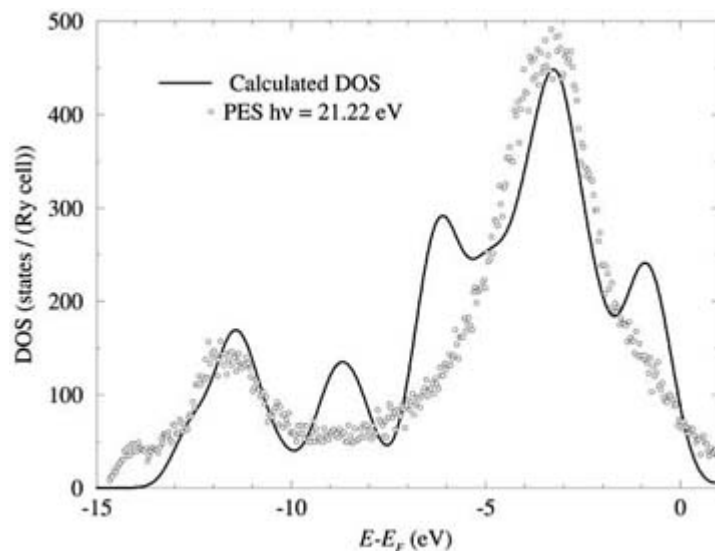


Figure B3.2.8. Comparison of the photoemission spectrum for the cetinite (Na;Se) and the density of states calculated by the AFC ELAPW- k - p method [88].

As another example of properties extracted from all-electron methods, figure B3.2.9 shows the results of a PAW simulation of benzene molecules on a graphite surface. The study aimed to show the extent to which the electronic structure of the molecule is modified by interaction with the surface, and why the images do not reflect the molecular structure. The PAW method was used to determine the structure of the molecule at the surface, the strength of the interaction between the surface and the molecules, and to predict and explain scanning tunnelling microscope (STM) images of the molecule on the surface [90] (the STM is described in section B1.19).



Figure B3.2.9. A benzene molecule on a graphite surface [90]. The geometry and the charge density (indicated by the surfaces of constant density) have been obtained using the PAW method. (Figure by Professor P E Blöchl.)

B3.2.3.3 LINEAR-SCALING ELECTRONIC STRUCTURE METHODS

DFT calculations such as the ones mentioned in chapter A1.3 and section B3.2.3.2 become computationally very expensive when the unit cell of the interesting system becomes large and complex, with certain parts of the computational algorithm typically scaling cubically with system size. A recent objective for treating large systems is to have the computational burden scale no more than linearly with system size. Methods achieving this are called linear-scaling or $O(N)$ (order N) methods, most of which are based on the Kohn–Sham equation (see equation (B3.2.4)), aiming to calculate single-electron wavefunctions, the Kohn–Sham orbitals. These

methods tend to be faster than the conventional Kohn–Sham approach above a few hundred atoms [20, 91, 92 and 93]. Another class of methods is based directly on the DFT of Hohenberg and Kohn [24]. With these techniques one seeks to determine directly the density that minimizes the total energy; they are often referred to as orbital-free methods [94, 95, 96 and 97]. Such orbital-free calculations do not have the bottlenecks present in orbital-based $O(N)$ DFT calculations, such as the need to localize orbitals to achieve linear scaling, orbital orthonormalization, or Brillouin zone sampling. Without such bottlenecks, the calculations become very inexpensive.

Equation (B3.2.3) lists the terms comprising the calculation of the total energy. The term due to the external potential and the Hartree term describing the Coulomb repulsion energy among the electrons already explicitly depend on the density instead of on orbitals. More difficult to evaluate is $G[\rho] = T_s[\rho] + E_{xc}[\rho]$, a functional which is not known exactly. However, over the years a number of high-quality exchange–correlation functionals have been developed for all kinds of systems. Only quite recently have more accurate kinetic energy density functionals (KEDFs) become available [97, 98 and 99] that afford linear-scaling computations.

One current limitation of orbital-free DFT is that since only the total density is calculated, there is no way to identify contributions from electronic states of a certain angular momentum character l . This identification is exploited in non-local pseudopotentials so that electrons of different l character ‘see’ different potentials, considerably improving the quality of these pseudopotentials. The orbital-free methods thus are limited to local pseudopotentials, connecting the quality of their results to the quality of the available local potentials. Good local pseudopotentials are available for the alkali metals, the alkaline earth metals and aluminium [100, 101] and methods exist for obtaining them for other atoms (see section VI.2 of [97]).

The orbital-free method has been used for molecular-dynamics studies of the formation of the self-interstitial defect in Al [102], pressure-induced glass-to-crystal transitions in sodium [103] and ion–electron correlations in liquid metals [101]. Calculations of densities for various Al surfaces have shown excellent agreement between the charge densities as calculated by Kohn–Sham DFT and an orbital-free method using a KEDF with a density-dependent response kernel [99]. The method was used recently to examine the metal–insulator transition in a two-dimensional array of metal quantum dots [104], where the theory showed that minute overlap of the nanoparticle’s wavefunctions is enough to transform the array from an insulator to a metal. As an example of the ease with which large simulations can be performed, figure B3.2.10 shows a plot of the charge density from an orbital-free calculation of a vacancy among 255 Al atoms [98], carried out on a workstation.

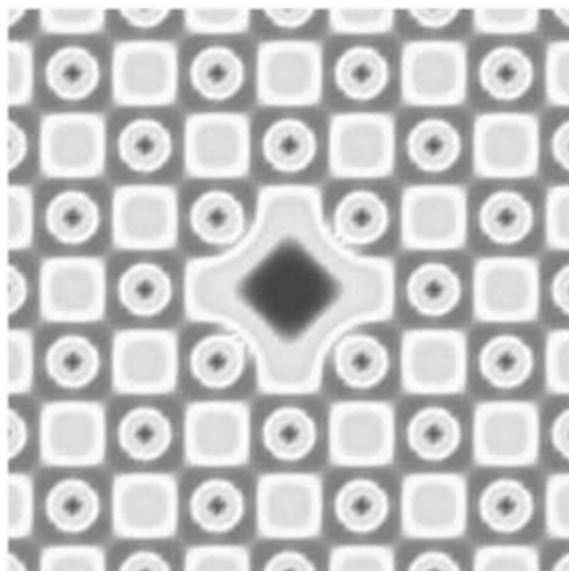


Figure B3.2.10. Contour plot of the electron density obtained by an orbital-free Hohenberg–Kohn technique [98]. The figure shows a vacancy in bulk aluminium in a 256-site cell containing 255 Al atoms and one empty site, the vacancy. Dark areas represent low electron density and light areas represent high electron density. A Kohn–Sham calculation for a cell of this size would be prohibitively expensive. Calculations on smaller cell sizes using both techniques yielded densities that were practically identical.

B3.2.3.4 THE HARTREE–FOCK METHOD IN CRYSTALS

The HF method (discussed in section A1.3.1.2) is an alternative to DFT approaches. It does not include electron correlation effects, i.e. non-classical electron–electron interactions beyond the Coulomb and exchange interactions. The neglect of these terms means that the Coulomb interaction is unscreened, and hence the electron repulsion energy is too large, overestimating ionic character, which leads to band gaps that are too large by a factor of two or more and valence band widths that are too wide by 30–40% [63]. However, the HF results can be used as a well defined starting point for the inclusion of many-particle corrections such as the GW approximation [31, 32] or, with considerably less computational effort, the results can be improved considerably by accounting for the Coulomb hole and screening the exchange interaction using the dielectric function [63, 105].

Ab initio HF programs for crystals have been developed [106, 107] and have been applied to a wide variety of bulk and surface systems [108, 109]. As an example, a periodic HF calculation using pseudopotentials and an LCAO basis predicted binding energies, lattice parameters, bulk moduli and central-zone phonon frequencies of 17 III–V and IV–IV semiconductors. The authors find that ‘...[o]n the whole, the HF LCAO data appear no worse than other *ab initio* results obtained with DF-based Hamiltonians’ [110]. They suggest that the largest part of the errors with respect to experiment is due to correlation effects and to a lesser extent due to the imperfections of the pseudopotentials [110]. More recently, the electronic and magnetic properties of transition metal oxides and halides such as perovskites, which had been a problem earlier, have been investigated with spin-unrestricted HF [111]. In general, the periodic HF method is best suited for the study of highly ionic, large band gap crystals because such systems are the least sensitive to the lack of electron correlation.

B3.2.3.5 QUANTUM MONTE CARLO

QMC techniques provide highly accurate calculations of many-electron systems. In variational QMC (VMC) [112, 113 and 114], the total energy of the many-electron system is calculated as the expectation value of the Hamiltonian. Parameters in a trial wavefunction are optimized so as to find the lowest-energy state (modern methods instead minimize the variance of the local energy $\frac{H\Psi}{\Psi}$ [115]). A Monte Carlo (MC) method is used to perform the multi-dimensional integrations necessary to determine the expectation value,

$$E = \frac{\int |\Psi|^2 \frac{H\Psi}{\Psi} d\tau}{\int |\Psi|^2 d\tau}$$

where ψ is the trial wavefunction and $|\Psi|^2 / \int |\Psi|^2 d\tau$ is a normalized probability distribution. The integration is performed by summing up the local energy at points, corresponding to electron configurations, given by the probability distribution. A random walk algorithm, such as the Metropolis algorithm [116], is used to sample those regions of configuration space more heavily where the probability density is high. The standard Slater–Jastrow trial wavefunction is the product of a Slater determinant of single-electron orbitals and a Jastrow factor, a function which includes the description of two-electron correlation. As an example, the trial wavefunction used for a silicon crystal contained 32 variational parameters whose optimization required the calculation of the local energy for 10 000–20 000 statistically independent electron configurations [117]. In contrast to the DMC technique described below, the accuracy of a VMC calculation depends on the quality of the many-particle wavefunction used [114]. In figure B3.2.11 we show the determination of the lattice constant of GaAs by VMC by minimization of the total energy [118]. This figure illustrates the roughness of the potential energy surface due to statistical errors, which poses a challenge then for the calculation of forces with QMC.

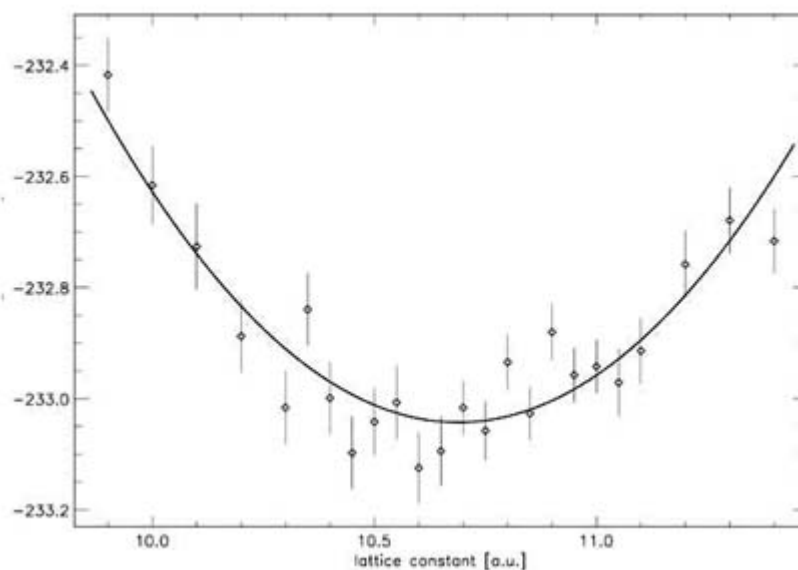


Figure B3.2.11. Total energy *versus* lattice constant of gallium arsenide from a VMC calculation including 256 valence electrons [118]; the curve is a quadratic fit. The error bars reflect the uncertainties of individual values. The experimental lattice constant is 10.68 au, the QMC result is 10.69 (± 0.1) au (Figure by Professor W Schattke).

In the diffusion QMC (DMC) method [114, 119], the evolution of a trial wavefunction (typically wavefunctions of the Slater–Jastrow type, for example, obtained by VMC) proceeds in imaginary time, $\tau = it$, according to the time-dependent Schrödinger equation, which then becomes a diffusion equation. All

components of the wavefunction except for the ground-state wavefunction are damped by the time evolution operator $\exp(-iHt) = \exp(-H\tau)$. The DMC was developed as a simplification of the Green's function MC technique [113]. A particularly well known use of the Green's function MC technique was the determination by Ceperley and Alder of the energy of the uniform electron gas as a function of its density [120]. This $E(\rho)$ was subsequently parametrized by Perdew and Zunger for the commonly used LDA exchange–correlation potential [40]. Usually two approximations are made to make DMC calculations tractable: the fixed-node approximation, in which the nodes, the places where the trial function changes sign, are kept fixed for the solution to enforce the fermion symmetry of the wavefunction and the so-called short-time approximation, whose effect can be made very small [114]. Excited states have been calculated by replacing an orbital in the Slater determinant of the trial wavefunction by a conduction-band orbital [121].

Recently, a method has been proposed to overcome the problems associated with calculating forces in both VMC and DMC [122]. It has been suggested that the use of QMC in the near future to tackle the energetics of systems as challenging as liquid binary iron alloys is not unthinkable [123].

B3.2.3.6 SUMMARY AND COMPARISONS

As we have outlined, a very wide variety of methods are available to calculate the electronic structure of solids. Empirical TB methods (such as discussed in [section B3.2.2](#)) are the least expensive, affording the calculation of unit cells with large numbers (e.g. 10^3) of atoms, or to provide cheap input to subsequent methods, at the price of quantitative accuracy. DFT methods ([section A1.3.5.4](#) and [section B3.2.3.3](#)), on the other hand, are responsible for many of the impressive results obtained in computational materials theory in recent years. The tradeoff for DFT is the opposite: its expense, except in the not-yet-general linear scaling methods, limits it typically to systems with at most a few hundred atoms. Once the $O(N)$ DFT methods become more general (for example, when orbital-free DFT can treat non-metallic systems), then the DFT method will be able routinely to treat systems as large as those treated now with TB.

The diversity of approaches based on HF ([section B3.2.3.4](#)) is small at present compared to the diversity found for DFT. For solids, HF appears to yield results inferior to DFT due to the neglect of electron correlation, but being a genuine many-particle theory it offers the possibility for consistent corrections, in contrast to DFT. Finally, the QMC techniques ([section B3.2.3.4](#)) hold promise for genuine many-particle calculations, yet they are still far from able to offer the same quantities for the same range of materials and geometries as the theories mentioned before. With this wide range of methods now introduced, we will look at their application to chemisorption on solid surfaces.

B3.2.4 QUANTUM STRUCTURAL METHODS FOR SOLID SURFACES

B3.2.4.1 INTRODUCTION

First-principles models of solid surfaces and adsorption and reaction of atoms and molecules on those surfaces range from *ab initio* quantum chemistry (HF; configuration interaction (CI), perturbation theory (PT), etc: for details see [chapter B3.1](#)) on small, finite clusters of atoms to HF or DFT on two-dimensionally infinite slabs. In between these

two extremes lie embedded cluster models, which recognize and attempt to correct the drastic approximation made by using a finite cluster to describe, for example, a metallic conductor whose electronic structure is inherently delocalized or an ionic crystal with long-range Coulomb interactions. Upon chemisorption, the binding of an atom or a molecule to a surface involves significant sharing of electrons in the bond between the

adsorbate and surface atoms and this breaking of the crystal symmetry will induce localization of the electrons. The attractive feature of the embedded cluster idea is that it preserves the strengths of the cluster approach, namely it allows one to describe the very local process of chemisorption to a high degree of accuracy by, for example, quantum chemical methods, while at the same time attempting to account for the presence of the rest of the surface and bulk. Surface reconstruction and molecular adsorption have been studied on a variety of surfaces, including insulators, semiconductors and metals. To illustrate these methods, we will focus on those used to examine adsorption of atoms and molecules on transition metal surfaces. This is not a comprehensive review of each approach; rather, we provide selected examples that demonstrate the range of techniques and applications, and some of the lessons learned.

B3.2.4.2 THE FINITE CLUSTER MODEL

The most straightforward molecular quantum mechanical approach is to treat adsorption on a small, finite cluster of transition metal atoms, ranging from as small as four atoms up to ~40 atoms. Though all-electron calculations can be performed, typically the core electrons of transition metal atoms are replaced by an effective core potential (ECP, the quantum chemistry version of a pseudopotential that accounts approximately for the core–valence electron interaction), while the valence electrons of each metal atom are treated explicitly within a HF, CI, PT, or DFT formalism. Typically, a few atoms in the chemisorption region contain the valence (or all) electrons explicitly, while surrounding atoms tend to be described more crudely with, for example, a one-electron ECP representation, model pseudopotentials or, in the case of ionic crystals, a finite array of point charges. Generally, the structure of the cluster is chosen to be a fixed fragment of the bulk. Examples of this type of approach include the early work of Upton and Goddard [124], who examined adsorption of electronegative and electropositive atoms on a Ni₂₀ cluster designed to mimic various low-index faces of Ni. In this model, only the 4s electrons on each Ni atom were treated explicitly, while the 3d electrons were subsumed into an ECP. They made predictions concerning preferred binding sites, geometries, vibrational frequencies and binding energies. Bagus *et al* [125] published an important comparison study showing that it is more accurate to treat metal atoms directly interacting with an adsorbate at an all-electron level, while it is sufficient to describe the surrounding metal atoms with ECPs. Panas *et al* [126] proposed the idea that a cluster should be ‘bond-prepared’, namely that one should study an electronic state of the finite cluster that has enough singly-occupied orbitals of the correct symmetry to interact with the incoming molecule to form the necessary covalent bonds between the adsorbate and the metal. In one of the first studies of a metal surface reaction, Panas *et al* [127] examined dissociative chemisorption pathways at the multi-reference CI level for O₂ on a Ni₁₃ cluster, generally using ECPs for all but the 4s electrons. Salahub and co-workers [128] used DFT-LDA with a Gaussian basis to examine chemisorption of C, O, H, CO and HCOO on Ni clusters containing up to 16 atoms meant to represent various low-index faces of Ni. Gradient corrections to the LDA scheme improved dramatically the binding energies for hydrogen bound to small Ni clusters, when compared to experimental results for Ni(111) and Ni(100) [129]. Multiple adsorbates were also studied by DFT-LDA: for example, in the case of hydrogen on Pd clusters modelling Pd(110) [130]. Diffusion barriers were also calculated by DFT-LDA for clusters containing up to 13 metal atoms of Pd, Rh, Sn, and Zn [131]. Other examples include HF calculations of K adsorbed on Cu clusters [132], HF and Møller–Plesset second-order PT (MP2) calculations of acetylene on Cu and Pd clusters [133], modified coupled pair functional (CPF) calculations for CO on Cu clusters [134], averaged CPF calculations of hydrogen adsorption on relaxed Cu clusters [135], HF, CASSCF (complete active space self-consistent field) and multireference CI and PT calculations for CO [136] and O [137] on Pt clusters, and spin-polarized DFT of *c*-CH₂N₂ on Pd and Cu tetramers [138] and of K and CO on Pd_{8,14} [139].

The advantage of the finite-cluster model is that one can systematically include high levels of electron correlation; this is to be balanced against the lack of a proper band structure, the presence of edge effects and the fact that it is generally limited to modelling low coverages. Next we outline current strategies for ameliorating some of these difficulties.

B3.2.4.3 FINITE-CLUSTER MODEL IN CONTACT WITH A CLASSICAL BACKGROUND

Several modifications of the finite-cluster model meant to account for the background Fermi sea of electrons and to compensate for the lack of a proper band structure have been developed. They rely on simple approximations of the surface/bulk, usually involving classical electrostatic interactions and usually applied to ionic crystals (see, for example, [140]). Of these, the model invented by Nakatsuji is the primary one that has considered adsorption on metal surfaces [141]. The so-called ‘dipped adcluster model’ [142] considers a small cluster plus an adsorbate as the ‘adcluster’ that is ‘dipped’ onto the Fermi sea of electrons of the bulk metal. A normal HF calculation on the small system is performed, in which electrons are added to or removed from the cluster in each calculation. By comparing the variation in the total energy with respect to the fractional electron transfer, dE/dn , to the work function of the metal, μ , the extent of electron transfer between the adcluster and the bulk metal can be established. Thus, charges on a small cluster are optimized and an image charge correction is also accounted for. In certain cases, integral charges are transferred between the cluster and the ‘surroundings’; then electron correlation calculations, for example CI, can be carried out. This is a purely classical electrostatic approach to accounting for the background electrons in an implicit, rather than explicit, manner. Nakatsuji has used this to study adsorption of ionic adsorbates on metals, and finds that one can describe the polarization of the metal reasonably well. We have worked briefly with this approach [143], but found that there is a problem with extending the method beyond two-dimensional clusters, because of an ambiguity of where to place the image plane. Indeed, Nakatsuji’s examples are always small one- or two-dimensional clusters. It is also likely that the wavefunction for such small clusters (typically ≤ 4 metal atoms) would not adequately represent a true metal surface wavefunction.

A simple, implicit means of describing the metallic band structure [144] was introduced by Rösch, using a Gaussian broadening of the cluster energy levels in order to determine a cluster Fermi level within DFT, originally by the X_α method (a simplified version of DFT-LDA; see section A1.3.3.3). Recent applications of this method have utilized more accurate forms of gradient-corrected spin-polarized DFT to look at adsorption of, for example, acetylene on $\text{Ni}_{14,20}$ clusters [145], CO adsorption on Ni, Pd and Pt clusters of eight or nine atoms [146] and NO adsorption on Ru [147].

B3.2.4.4 SLAB CALCULATIONS

The other extreme of modelling chemisorption is to use a slab described by DFT or HF. The slab is typically taken to be periodic in the directions parallel to the surface and contains a few atomic layers in the direction normal to the surface. For the adatoms not to influence each other, unless that is intended, the unit cell needs to be sufficiently large, parallel to the surface. For computational reasons, it is advantageous in some methods, namely plane wave techniques, to have periodicity in three dimensions. In the supercell geometry, this periodicity is gained by considering slabs which are periodic in the direction perpendicular to the surface but separated from each other by vacuum regions. The vacuum region has to be thick enough so that there is no influence between the surfaces facing each other (the same is true for the slab thickness). For a schematic description of several simulation model geometries, see [figure B3.2.12](#).

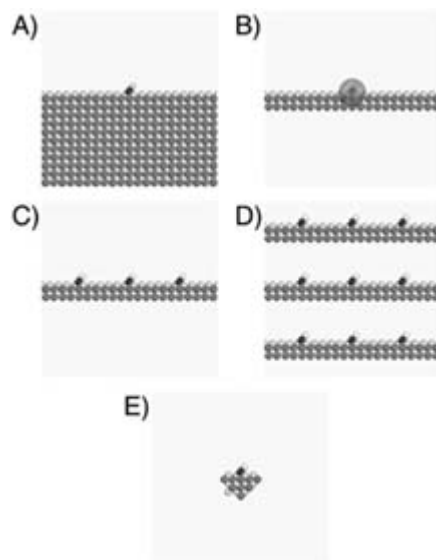


Figure B3.2.12. Schematic illustration of geometries used in the simulation of the chemisorption of a diatomic molecule on a surface (the third dimension is suppressed). The molecule is shown on a surface simulated by (A) a semi-infinite crystal, (B) a slab and an embedding region, (C) a slab with two-dimensional periodicity, (D) a slab in a supercell geometry and (E) a cluster.

Freeman and co-workers developed the FLAPW method (see B3.2.2) during the early 1980s [70, 148]. This was a major advance, because the conventional ‘muffin-tin’ potential was eliminated from their calculation allowing general-shape potentials to be evaluated instead. Freeman’s group first developed this for thin films and then for bulk metals. As mentioned before, the LAPW basis, along with the elimination of any shape approximations in the potential, allows for highly accurate calculations on transition metal surfaces, within the DFT-LDA and the generalized gradient approximation, GGA (see section A1.3.3.3). For the ‘stand-alone’ slab geometry, figure B3.2.12(C) the LAPW basis functions decay exponentially into the vacuum. The numerous interfacial systems examined by Freeman’s group include, for example, CO with K or S coadsorption on Ni(001) [149], adsorption of sulfur alone on Ni(001) [150], Fe monolayers on Ni(111) [151], Ag monolayers on MgO(001) [152], Au-capped Fe monolayers on MgO(001) [153], NO adsorption on Rh, Pd and Pt [154], and Li on Ru(001) [155]. Typical properties predicted are the equilibrium positions, magnetic moments, charge densities and surface densities of states.

More recently, other groups—primarily in Europe—have begun doing pseudopotential plane wave (often gradient-corrected) DFT supercell slab calculations (figure B3.2.12(D)) for chemisorption on metals. The groups of Nørskov [156], Scheffler [157], Baerends [158, 159] and Hafner and Kresse [160, 161] have been the most active. Adsorbate–metal surface systems examined include: alkalis and N₂ on Ru [156], NO on Pd [156], H₂ on Al [156], Cu [156, 158], Pd [157, 158] and sulfur-covered Pd [157], CO oxidation on Ru [157], CO on Ni, Pd and Pt [158], O on Pt [160], and H₂ on Rh, Pd and Ag [161].

An interesting study by te Velde and Baerends [159] compared slab- and cluster-DFT results for CO absorption on Cu(100). They found large oscillations in the chemisorption binding energy of CO to finite copper clusters as a function of cluster size. This suggests that the finite-cluster model (figure B3.2.12(E)) is likely to be inadequate, at least for modelling metal surfaces. By contrast, the slab calculations converge quickly with the number of Cu layers for the CO heat of adsorption and CO–CO distances.

The supercell plane wave DFT approach is periodic in three dimensions, which has some disadvantages: (i) thick vacuum layers are required so the slab does not interact with its images, (ii) for a tractably sized unit cell, only high adsorbate coverages are modelled readily and (iii) one is limited in accuracy by the form of the

exchange–correlation functional chosen. In particular, while DFT, especially using gradient-corrected forms of the exchange–correlation functional (GGA), has proven to be remarkably reliable in many instances, there are a number of examples for chemisorption in which the commonly used GGAs have been shown to fail dramatically (errors in binding energies of 1 eV or greater) [162, 163]. This naturally motivates the next set of approaches, namely the embedded cluster strategy.

B3.2.4.5 EMBEDDED-CLUSTER SCHEMES: CLUSTER IN CLUSTER

Whitten and co-workers developed a metal cluster embedding scheme appropriate for CI calculations during the 1980s [164]. In essence, the method consists of: (i) solving for a HF minimum basis set (one 4s orbital/atom) description of a large cluster (e.g., ~30–90 atoms); (ii) localizing the orbitals *via* exchange energy maximization with atomic basis functions on the periphery; (iii) using these localized orbitals to set up effective Coulomb and exchange operators for the electrons within the cluster to be embedded; (iv) improving the basis set on the atoms comprising the embedded cluster and (v) performing a small CI calculation ($O(10^3)$ configurations) within orbitals localized on the embedded cluster. This strategy provides an approximate way of accounting for nearby electrons outside the embedded cluster itself. Whitten and co-workers have applied it to a variety of adsorbates (H, N, O, C—containing small molecules) on, primarily, Ni surfaces. Duarte and Salahub recently reported a DFT-cluster-in-DFT-cluster variant of Whitten’s embedding, with a couple of twists on the original approach (for example, fractional orbital occupancies and charges, and an extra buffer region) [165]. Earlier, Sellers developed a related scheme for embedding a MP2 cluster within another cluster, where the background was modelled with screened ECPs [166]. Also, Ravenek and Geurts [167] and Fukunishi and Nakatsuji [168] extended the Green’s matrix method of Pisani [169] (who developed it mainly for ionic crystals) to again embed a cluster within a cluster by introducing a semiorthogonal basis and renormalizing the charge on the cluster. It was implemented within the X_α method in the former case [167], and by broadening each discrete energy level to mimic the bulk band structure within HF theory for the cluster in the latter case [168].

Pisani [169] has used the density of states from periodic HF (see B3.2.2.4) slab calculations to describe the host in which the cluster is embedded, where the applications have been primarily to ionic crystals such as LiF. The original calculation to derive the external Coulomb and exchange fields is usually done on a finite cluster and at a low level of *ab initio* theory (typically minimum basis set HF, one electron only per atom treated explicitly).

The main drawback of the cluster-in-cluster methods is that the embedding operators are derived from a wavefunction that does not reflect the proper periodicity of the crystal: a two-dimensionally infinite wavefunction/density with a proper band structure would be preferable. Indeed, Rösch and co-workers pointed out recently a series of problems with such cluster-in-cluster embedding approaches. These include the lack of marked improvement of the results over finite clusters of the same size, problems with the orbital space partitioning such that charge conservation is violated, spurious mixing of virtual orbitals into the density matrix [170], the inherent delocalized nature of metallic orbitals [171], etc.

B3.2.4.6 EMBEDDING OF CLUSTERS IN PERIODIC BACKGROUND

One of the first cluster embedding schemes was put forth by Ellis and co-workers [172]. They were interested in studying transition metal impurities in NiAl alloys, so they considered a TMAI_xNi_y cluster embedded in a periodic self-consistent crystal field appropriate for bulk β' -NiAl. The field was calculated via X_α calculations, as was the cluster itself. The idea was to provide a relatively inexpensive alternative to supercell DFT calculations.

Perhaps the most sophisticated embedding scheme for describing metal surfaces to date is the LDA-based

self-consistent Green's function method for semi-infinite crystals. Inglesfield, Benesh, and co-workers embed the near-surface layers using an embedding potential constructed from the bulk Green's function within an all-electron approach, using an LAPW basis [173]. Scheffler and co-workers developed a similar approach using a Gaussian basis for the valence electrons and pseudopotentials [174]. The formulation of the latter method is somewhat different from Inglesfield's and Benesh's, in that a reference system is chosen for which the Green's function and density are known (typically the bulk metal), and a Δ (Green's function) is solved for in order to get a Δ (embedding potential) and hence a Δ (density). This allows one to solve for the embedding potential locally in a small region around the adsorbate. These methods allow for an economical yet accurate calculation of the embedding density, which yields a trustworthy description of charge transfer and other equilibrium properties, though subject to the accuracy limitations inherent in DFT-LDA.

In the late 1980s, Feibelman developed his Green's function scattering method using LDA with pseudopotentials to describe adsorption on two-dimensionally infinite metal slabs [175], based on earlier work by Williams *et al* [176]. The physical basis for the technique is that the adsorbate may be considered a defect off which the Bloch waves of the perfect substrate scatter. The interaction region is short-range because of screening by the electron gas of the metal. Feibelman has used this technique to study, for example, the chemisorption of an H₂ molecule on Rh(001) [177], S adatoms on Al(331) [178] and Ag adatoms on Pt(111) [179]. Charge densities, relative energies for various adsites and diffusion barriers (the latter in good agreement with experiment) were the typical quantities predicted.

Krüger and Rösch implemented within DFT the Green's matrix approach of Pisani within an approximate periodic slab environment [180]. They were able to successfully extend Pisani's embedding approach to metal surfaces by smoothing out the step function that determines the occupation numbers near the Fermi level. Keys to the numerical success of their method included: (i) symmetric orthogonalization of the Bloch basis to produce a localized set of functions that yielded a balanced distribution of charge in the system and (ii) self-consistent evaluation of the Fermi energy by fixing the charge on the cluster to be neutral. The slab was described with a Slater basis at the DFT-LDA level, while the embedded cluster orbitals were expanded in terms of Gaussian functions at the DFT-LDA level. While some properties exhibited non-monotonic behaviour with increasing cluster size, the charge transfer between the metal surface and the adsorbate seemed to be well described. They concluded that properties are not well converged in this method if the cluster does not contain shells of metal atoms that are at least next-nearest-neighbours to the adsite metal atoms.

Head and Silva used occupation numbers obtained from a periodic HF density matrix for the substrate to define localized orbitals in the chemisorption region, which then defines a cluster subspace on which to carry out HF calculations [181]. Contributions from the surroundings also only come from the bare slab, as in the Green's matrix approach. Increases in computational power and improvements in minimization techniques have made it easier to obtain the electronic properties of adsorbates by supercell slab techniques, leading to the Green's function methods becoming less popular [182].

Cortona embedded a DFT calculation in an orbital-free DFT background for ionic crystals [183], which necessitates evaluation of kinetic energy density functionals (KEDFs). Wesolowski and Warshel [184] had similar ideas to Cortona, except they used a frozen density background to examine a solute in solution and examined the effect of varying the KEDF. Stefanovich and Truong also implemented Cortona's method with a frozen density background and applied it to, for example, water adsorption on NaCl(001) [185].

B3.2.4.7 EMBEDDING EXPLICIT CORRELATION METHODS IN A DFT BACKGROUND

In principle, DFT calculations with an ideal exchange–correlation functional should provide consistently accurate energetics. The catch is, of course, that the exact exchange–correlation functional is not known.

While various GGAs have been remarkably successful, there are notable exceptions [186, 187], including ones specific to surface adsorption mentioned earlier, where the binding-energy errors can be more than an eV [162, 163]. As another example, Louie and Cohen and co-workers found no systematic improvement over the LDA when gradient corrections were included in calculations of Al, Nb and Pd bulk properties, including the cohesive energy [186]. Indeed, the design of exchange–correlation functionals constitutes an active field of research (see, for example, [188]). The lack of completely systematic means to improve these functionals is an unappealing aspect of these calculations.

A first step towards a systematic improvement over DFT in a local region is the method of Aberenkov *et al* [189], who calculated a correlated wavefunction embedded in a DFT host. However, this is achieved using an analytic embedding potential function fitted to DFT results on an indented crystal. One must be cautious using a bare indented crystal to represent the surroundings, since the density at the surface of the indented crystal will have inappropriate Friedel oscillations inside and decay behaviour at the indented surface not present in the real crystal.

We have developed a different first-principles embedding theory that combines DFT with explicit correlation methods. We sought to develop a method for treating bulk or surface phases that is more accurate than current implementations of DFT. The idea is to provide more accurate predictions for local energetics, such as chemisorption binding energies and adsorbate electronic excitation energies. To achieve this, our theory improves upon the DFT description of electron correlation in a local region. This is accomplished by an embedding theory that treats a small region within an accurate quantum chemistry approach [190, 191], which interacts with its surroundings *via* an embedding potential, $v_{\text{embed}}(\mathbf{r})$. This $v_{\text{embed}}(\mathbf{r})$ is derived from a periodic DFT calculation on the total system. It is expressed purely in terms of orbital-free DFT (kinetic and potential energy) interaction terms between the embedded region and its surroundings *à la* Cortona and, in particular, purely in terms of functionals of the total density, ρ_{tot} , and the density of the embedded region, ρ_1 . We thus avoid construction of localized orbitals to describe the electrons in the surrounding environment. This is especially important for metal surfaces, where the extensive \mathbf{k} -point sampling required to get a well converged density makes localization impractical (very expensive). This way of expressing the embedding operator also eliminates problems that occur in other forms of embedding, such as those of matching conditions at the embedding boundary, or spurious charge transfer, since the electrostatic potential and the density are continuous by construction. Its only real disadvantage is that there is an arbitrariness associated with the choice of T_s . Development of optimal T_s functionals is an active area of research in our group [97, 98 and 99].

The self-consistent embedding cycle proceeds as follows. First, a well converged density, ρ_{tot} , is calculated for the extended metal surface in the presence of an adsorbate. This is accomplished within a standard pseudopotential plane wave DFT calculation (see [chapter A1.3](#)). Second, we partition the system into the region of interest (typically the adsorbate and neighbouring metal atoms at or near the surface) and its surroundings (all the other atoms in the periodic unit cell). The embedded region is defined by the integral number of electrons and nuclei within that region but not by

-29-

a particular physical, fixed boundary. This allows for the electron density from the embedded region to expand or contract variationally into the surroundings, thus affording some effective charge polarization to occur as needed.

The electron density, ρ_1 , of the embedded cluster/adsorbate atoms is calculated using quantum chemistry methods (HF, PT, multireference SCF, or CI). The initial step in this iterative procedure sets $v_{\text{embed}}(\mathbf{r})$ to zero, since ρ_1 is needed in order to calculate it. On subsequent iterations, the third step is to use ρ_1 and ρ_{tot} to calculate $v_{\text{embed}}(\mathbf{r})$, then insert it, as a one-electron operator expressed in matrix form in the atomic orbital basis of the adsorbate/cluster, into the quantum chemistry calculation of step two, and then ρ_1 is updated (*via* the wavefunction). We repeatedly update $v_{\text{embed}}(\mathbf{r})$ and then ρ_1 until full self-consistency is achieved, with

fixed ρ_{tot} . In this way, we variationally optimize both the quantum chemistry wavefunction and, implicitly, the density of the surroundings, subject to fixed ρ_{tot} . We tacitly assume that the DFT-slab density for the total system, ρ_{tot} , is in fact a good representation and does not need to be adjusted.

We have shown that our embedding total energies may be written in terms of the total energy obtained in step one (the DFT total energy for the entire system), plus a correction term, that subtracts out the DFT energy in the local region I and adds back in an *ab initio* total energy for that same region,

$$E_{\text{tot}}^{\text{embed}} = E_{\text{tot}}^{\text{DFT}} + (E_{\text{I}}^{\text{ab initio}} - E_{\text{I}}^{\text{DFT}}).$$

Thus, another way to think of the embedding is that the *ab initio* treatment of region I is *correcting* the DFT results in the same region, for the same self-consistent density. We expect, then, that such a treatment should reduce, for example, the famous LDA overbinding problem (LDA bond energies are generally significantly overestimated). We have indeed seen a smooth decrease in the LDA overbinding as a function of increasing electron correlation. We benchmarked the method against nearly exact calculations on a small system and then further corroborated it on experimentally well studied chemisorption systems: CO on transition metal surfaces. Our binding energies are in good agreement with nearly full configuration interaction in the former and experimental adsorbate binding energies in the latter. Very recently, we have demonstrated that excitation energies for adsorbed CO are dramatically improved compared to experiment upon inclusion of the embedding potential [192]. In the future, we hope this method will provide a general means for accurate predictions of the local electronic structure of condensed matter.

B3.2.5 OUTLOOK

Computational solid-state physics and chemistry are vibrant areas of research. The all-electron methods for high-accuracy electronic structure calculations mentioned in [section B3.2.3.2](#) are in active development, and with PAW, an efficient new all-electron method has recently been introduced. Ever more powerful computers enable more detailed predictions on systems of increasing size. At the same time, new, more complex materials require methods that are able to describe their large unit cells and diverse atomic make-up. Here, the new orbital-free DFT method may lead the way. More powerful techniques are also necessary for the accurate treatment of surfaces and their interaction with atoms and, possibly complex, molecules. Combined with recent progress in embedding theory, these developments make possible increasingly sophisticated predictions of the quantum structural properties of solids and solid surfaces.

ACKNOWLEDGMENTS

The authors would like to thank Professor P E Blöchl, Dr H Eckstein, Professor W Schattke, Professor K E Smith and T Strasser for making figures available for this publication. FS thanks Dr E E Krasovskii for introducing him to the LAPW method.

REFERENCES

- [1] Whitten J L and Yang H 1996 Theory of chemisorption and reactions on metal surfaces *Surf. Sci. Rep.* **24** 59–124

- [2] Mehl M J and Papaconstantopoulos D A 1998 Tight-binding parametrization of first-principles results *Topics in Computational Materials Science* ed C Y Fong (Singapore: World Scientific); URL <http://cst-www.nrl.navy.mil/~mehl/review/rev4.html>
- [3] Starrost F, Bornholdt S, Solterbeck C and Schattke W 1996 Band-structure parameters by genetic algorithm *Phys. Rev. B* **53** 12 549; *Phys. Rev. B* **54** 17 226E
Strasser T, Starrost F, Solterbeck C and Schattke W 1997 Valence-band photoemission from GaN(001) and GaAs: GaN surfaces *Phys. Rev. B* **56** 13 326
- [4] Klimeck G, Brown R C, Boykin T B, Salazar-Lazaro C, Cwik T A and Stoica A 2000 Si tight-binding parameters from genetic algorithm fitting *Superlattices Microstruct.* **27** 10
Klimeck G, Brown R C, Boykin T B, Salazar-Lazaro C, Cwik T A and Stoica A 1999 *Preprint* 1006/spmi.1999.0797
- [5] Cohen R E, Mehl M J and Papaconstantopoulos D A 1994 Tight-binding total-energy method for transition and noble metals *Phys. Rev. B* **50** 14 694–7
- [6] Haas H, Wang C Z, Fähnle M, Elsässer C and Ho K M 1998 Environment-dependent tight-binding model for molybdenum *Phys. Rev. B* **57** 1461
- [7] Harrison W A 1989 *Electronic Structure and the Properties of Solids* (New York: Dover)
- [8] Menon M and Subbaswamy K R 1994 Transferable nonorthogonal tight-binding scheme for silicon *Phys. Rev. B* **50** 11 577
- [9] Slater P C and Koster G F 1954 Simplified LCAO method for the periodic potential problem *Phys. Rev.* **94** 1498–524
- [10] Mehl M J and Papaconstantopoulos D A 1996 Applications of a tight-binding total-energy method for transition and noble metals: Elastic constants, vacancies and surfaces of monatomic metals *Phys. Rev. B* **54** 4519
Mazin I I, Papaconstantopoulos D A and Singh D J 2000 Tight-binding Hamiltonians for Sr-filled ruthenates: Application to the gap anisotropy and Hall coefficient in Sr_2RuO_4 *Phys. Rev. B* **61** 5223
- [11] Mercer J L Jr and Chou M Y 1994 Tight-binding model with intra-atomic matrix elements *Phys. Rev. B* **49** 8506
- [12] Watson S C, Carter E A, Walters M K and Madden P A (unpublished)
- [13] Seifert G, Eschrig H and Bieger W 1986 An approximate variation of the LCAO- X_α method *Z. Phys. Chem.* **267** 529
- [14] Horsfield A P 1997 Efficient *ab initio* tight binding *Phys. Rev. B* **56** 6594–602
- [15] Elstner M, Porezag D, Jungnickel G, Elsner J, Haugk M, Frauenheim Th, Suhai S and Seifert G 1998 Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties *Phys. Rev. B* **58** 7260
- [16] Lee S M, Belkhir M A, Zhu X Y, Lee Y H, Huang Y G and Frauenheim Th 2000 Electronic structure of GaN edge dislocations *Phys. Rev. B* **61** 16 033

- [17] Bowler D R, Aoki M, Goringe C M, Horsfield A P and Pettifor D G 1997 A comparison of linear scaling tight-binding methods *Modelling Simulation Mater. Sci.* **5** 199
- [18] Wang C S and Callaway J 1978 BNDPKG. A package of programs for the calculation of electronic energy bands by the LCGO method *Comput. Phys. Commun.* **14** 327
- [19] Voß D, Krüger P, Mazur A and Pollmann J 1999 Atomic and electronic structure of WSe_2 from *ab initio* theory: bulk crystal and thin film systems *Phys. Rev. B* **60** 14 311
- [20] Artacho E, Sánchez-Portal D, Ordejón P, García A and Soler J M 1999 Linear-scaling *ab initio* calculations for large and complex systems *Phys. Status Solidi B* **215** 809
- [21] Hoffmann R 1963 An extended Hückel theory. I. Hydrocarbons *J. Chem. Phys.* **39** 1397
- [22] Henk J, Schattke W, Carstensen H, Manzke R and Skibowski M 1993 Surface-barrier and polarization effects in the photoemission from GaAs(110) *Phys. Rev. B* **47** 2251
- [23] Strasser T, Solterbeck C, Starrost F and Schattke W 1999 Valence-band photoemission from the GaN(0001) surface *Phys. Rev. B* **60** 11 577
- [24] Hohenberg P and Kohn W 1964 Inhomogeneous electron gas *Phys. Rev.* **136** B 864
- [25] Kohn W and Sham L J 1965 Self-consistent equations including exchange and correlation effects *Phys. Rev.* **140** A1133
- [26] Janak J F 1978 Proof that $\partial E/\partial n_i = \epsilon_i$ in density-functional theory *Phys. Rev. B* **18** 7165–8

- [27] Perdew J P, Parr R G, Levy M and Balduz J L Jr 1982 Density-functional theory for fractional particle number: derivative discontinuities of the energy *Phys. Rev. Lett.* **49** 1691–4
 Kleinman L 1997 Significance of the highest occupied Kohn–Sham eigenvalue *Phys. Rev. B* **56** 12 042–5
 Perdew J P and Levy M 1997 Comment on ‘Significance of the highest occupied Kohn–Sham eigenvalue’ *Phys. Rev. B* **56** 16 021–8
 Kleinman L 1997 Reply to ‘Comment on ‘Significance of the highest occupied Kohn–Sham eigenvalue’’ *Phys. Rev. B* **56** 16 029–30
- [28] Bechstedt F 1992 Quasiparticle corrections for energy gaps in semiconductors *Adv. Solid State Phys.* **32** 161
- [29] Fiorentini V and Baldereschi A 1995 Dielectric scaling of the self-energy scissor operator in semiconductors and insulators *Phys. Rev. B* **51** 17 196
- [30] Pulci O, Onida G, Shkrebtii A I, Del Sole R and Adolph B 1997 Plane-wave pseudopotential calculation of the optical properties of GaAs *Phys. Rev. B* **55** 6685
- [31] Hedin L and Lundqvist S 1969 Effects of electron–electron and electron–phonon interactions on the one-electron states of solids *Solid State Phys.* **23** 1
- [32] Hybertsen M S and Louie S G 1985 First-principles theory of quasiparticles: Calculation of band gaps in semiconductors and insulators *Phys. Rev. Lett.* **55** 1418
- [33] Louie S G 1987 Theory of quasiparticle energies and excitation spectra of semiconductors and insulators *Electronic Band Structure and Its Applications (Lecture Notes in Physics vol 283)* ed M Youssouf (Berlin: Springer)
- [34] Rohlfing M, Krüger P and Pollmann J 1997 Quasiparticle calculations of semicore states in Si, Ge, and CdS *Phys. Rev. B* **56** R7065–8
- [35] Godby R W, Schlüter M and Sham L J 1988 Self-energy operators and exchange–correlation potentials in semiconductors *Phys. Rev. B* **37** 10159–75
- [36] Massidda S, Continenza A, Posternak M and Baldereschi A 1997 Quasiparticle energy bands of transition-metal oxides within a model GW scheme *Phys. Rev. B* **55** 13 494–502
- [37] Shirley E L 1998 Many-body effects on bandwidths in ionic, noble gas, and molecular solids *Phys. Rev. B* **58** 9579–83
- [38] Rieger M M and Godby R W 1998 Charge density of semiconductors in the GW approximation *Phys. Rev. B* **58** 1343

- [39] Campillo I, Silkin V M, Pitarke J M, Chulkov E V, Rubio A and Echenique P M 2000 First-principles calculations of hot-electron lifetimes in metals *Phys. Rev. B* **61** 13 484–92
- [40] Perdew J P and Zunger A 1981 Self-interaction correction to density-functional approximations for many-electron systems *Phys. Rev. B* **23** 5048
- [41] Svane A and Gunnarsson O 1990 Transition-metal oxides in the self-interaction-corrected density-functional formalism *Phys. Rev. Lett.* **65** 1148
- [42] Szotek Z, Temmerman W M and Winter H 1993 Application of the self-interaction correction to transition-metal oxides *Phys. Rev. B* **47** 4029
- [43] Svane A, Temmerman W and Szotek Z 1999 Theory of pressure-induced phase transitions in cerium chalcogenides *Phys. Rev. B* **59** 7888
- [44] Vogel D, Krüger P and Pollmann J 1997 Structural and electronic properties of group-III nitrides *Phys. Rev. B* **55** 12 836, and references therein
- [45] Stampfl C, van de Walle C G, Vogel D, Krüger P and Pollmann J 2000 Native defects and impurities in InN: First-principles studies using the local-density approximation and self-interaction and relaxation-corrected pseudopotentials *Phys. Rev. B* **61** R7846–9
- [46] Terakura K, Williams A R, Oguchi T and Kübler J 1984 Transition-metal monoxides: Band or Mott insulators *Phys. Rev. Lett.* **52** 1830
 Terakura K, Oguchi T, Williams A R and Kübler J 1984 Band theory of insulating transition-metal monoxides: Band-structure calculations *Phys. Rev. B* **30** 4734
- [47] Aryasetiawan F and Gunnarsson O 1995 Electronic structure of NiO in the GW approximation *Phys. Rev. Lett.* **74** 3221
- [48] Anisimov V I, Kuiper P and Nordgren J 1994 First-principles calculation of NiO valence spectra in the impurity-Anderson-model approximation *Phys. Rev. B* **50** 8257–65
- [49] Anisimov V I, Aryasetiawan F and Liechtenstein A I 1997 First-principles calculations of the electronic structure and spectra of strongly correlated systems: The LDA+U method *J. Phys.: Condens Matter* **9** 767

- [50] Anisimov V I, Zaanen J and Andersen O K 1991 Band theory and Mott insulators: Hubbard U instead of Stoner I *Phys. Rev. B* **44** 943
- [51] Gunnarsson O, Andersen O K, Jepsen O and Zaanen J 1989 Density-functional calculation of the parameters in the Anderson model: Application to Mn in CdTe *Phys. Rev. B* **39** 1708–22
Anisimov V I and Gunnarsson O 1991 Density-functional calculation of effective Coulomb interactions in metals *Phys. Rev. B* **43** 7570–4
- [52] Kwon S K and Min B I 2000 Unquenched large orbital magnetic moment in NiO *Phys. Rev. B* **62** 73
- [53] Slater J C 1937 Wave functions in a periodic potential *Phys. Rev.* **51** 846
- [54] Singh D J 1994 *Planewaves, Pseudopotentials and the LAPW Method* (Norwell, MA: Kluwer)
- [55] Marcus P M 1967 Variational methods in the computation of energy bands *Int. J. Quantum Chem.* **1 S** 567
- [56] Koelling D D 1970 Alternative augmented-plane-wave technique: theory and application to copper *Phys. Rev. B* **2** 290–8
- [57] Bross H, Bohn G, Meister G, Schubö W and Stöhr H 1970 New version of the modified augmented-plane wave method *Phys. Rev. B* **2** 3098–103
- [58] Andersen O K 1975 Linear methods in band theory *Phys. Rev. B* **12** 3060
- [59] Singh D and Krakauer H 1991 H-point phonon in molybdenum: Superlinearized augmented-plane-wave calculations *Phys. Rev. B* **43** 1441–5
- [60] Krasovskii E E, Yaresko A N and Antonov V N 1994 Theoretical study of ultraviolet photoemission spectra of noble metals *J. Electron Spectrosc. Relat. Phenom.* **68** 157
- [61] Sjöstedt E, Nordström L and Singh D J 2000 An alternative way of linearizing the augmented plane-wave method *Solid State Commun.* **114** 15

- [62] Shick A B, Liechtenstein A I and Pickett W E 1999 Implementation of the LDA+ U method using the full-potential linearized augmented plane-wave basis *Phys. Rev. B* **60** 10 763
- [63] Massidda S, Posternak M and Baldereschi A 1993 Hartree–Fock LAPW approach to the electronic properties of periodic systems *Phys. Rev. B* **48** 5058
- [64] Koelling D D and Arbmán G O 1975 Use of energy derivative of the radial solution in an augmented plane wave method: application to copper *J. Phys. F: Met. Phys.* **5** 2041
- [65] Krasovskii E E 1997 Accuracy and convergence properties of the extended linear augmented-plane-wave method *Phys. Rev. B* **56** 12 866
- [66] Krasovskii E E, Nemoshkalenko V V and Antonov V N 1993 On the accuracy of the wavefunctions calculated by LAPW method *Z. Phys. B* **91** 463
- [67] Singh D 1991 Ground-state properties of lanthanum: treatment of extended-core states *Phys. Rev. B* **43** 6388
- [68] Krasovskii E E and Schattke W 1995 The extended-LAPW-based $k * p$ method for complex bandstructure calculations *Solid State Commun.* **93** 775
- [69] Krasovskii E E, Starrost F and Schattke W 1999 Augmented Fourier components method for constructing the crystal potential in self-consistent band-structure calculations *Phys. Rev. B* **59** 10 504
- [70] Wimmer E, Krakauer H, Weinert M and Freeman A J 1981 Full-potential self-consistent linearized-augmented-plane-wave method for calculating the electronic structure of molecules and surfaces: O_2 molecule *Phys. Rev. B* **24** 864
- [71] Weinert M 1981 Solution of Poisson's equation: beyond Ewald-type methods *J. Math. Phys.* **22** 2433
- [72] Petersen M, Wagner F, Hufnagel L, Scheffler M, Blaha P and Schwarz K 2000 Improving the efficiency of FP-LAPW calculations *Comp. Phys. Commun.* **126** 294–309
- [73] Asato M, Settels A, Hoshino T, Asada T, Blügel S, Zeller R and Dederichs P H 1999 Full-potential KKR calculations for metals and semiconductors *Phys. Rev. B* **60** 5202
- [74] Korringa J 1947 On the calculation of the energy of a Bloch wave in a metal *Physica (Amsterdam)* **13** 392–400
- [75] Kohn W and Rostoker N 1954 Solution of the Schrödinger equation in periodic lattices with an application to metallic lithium *Phys. Rev.* **94** 1111–20
- [76] Drittler B, Weinert M, Zeller R and Dederichs P H 1991 Vacancy formation energies of fcc transition metals calculated

by a full potential Green's function method *Solid State Commun.* **79** 31

- [77] Podloucky R, Zeller R and Dederichs P H 1980 Electronic structure of magnetic impurities calculated from first principles *Phys. Rev. B* **22** 5777
- [78] Yussouff M 1987 Fast self-consistent KKR method *Electronic Band Structure and Its Applications (Lecture Notes in Physics vol 283)* ed M Yussouff (Berlin: Springer) pp 58–76
- [79] Tank R W and Arcangeli C 2000 An introduction to the third-generation LMTO method *Status Solidi B* **217** 89
- [80] Methfessel M, Rodriguez C O and Andersen O K 1989 Fast full-potential calculations with a converged basis of atom-centered linear muffin-tin orbitals: structural and dynamic properties of silicon *Phys. Rev. B* **40** 2009–12
- [81] Blöchl P E 1994 Projector augmented-wave method *Phys. Rev. B* **50** 17 953
- [82] Holzwarth N A W, Matthews G E, Dunning R B, Tackett A R and Zeng Y 1997 Comparison of the projector augmented-wave, pseudopotential and linearized augmented-plane-wave formalisms for density-functional calculations of solids *Phys. Rev. B* **55** 2005
- [83] Alfè D, Kresse G and Gillan M J 2000 Structure and dynamics of liquid iron under Earth's core conditions *Phys. Rev. B* **61** 132
- [84] Ohtsuki T, Ohno K, Shiga K, Kawazoe Y, Maruyama Y and Masumoto K 1998 Insertion of Xe and Kr atoms into C60 and C70 fullerenes and the formation of dimers *Phys. Rev. Lett.* **81** 967–70

-34-

- [85] Krasovska O V, Krasovskii E E and Antonov V N 1995 *Ab initio* calculation of the optical and photoelectron properties of RuO₂ *Phys. Rev. B* **52** 11 825
- [86] Leventi-Peetz A, Krasovskii E E and Schattke W 1995 Dielectric function and local field effects of TiSe₂ *Phys. Rev. B* **51** 17 965
- [87] Traving M, Boehme M, Kipp L, Skibowski M, Starrost F, Krasovskii E E, Perlov A and Schattke W 1997 Electronic structure of WSe₂: a combined photoemission and inverse photoemission study *Phys. Rev. B* **55** 10 392–9
- [88] Starrost F, Krasovskii E E, Schattke W, Jockel J, Simon U, Adelung R and Kipp L 2000 Chalcogenites: electronic, optical, and conduction properties of nanoporous chalcogenoantimonates *Phys. Rev. B* **61** 15 697
- [89] Krasovskii E E and Schattke W 1997 Surface electronic structure with the linear methods of band theory *Phys. Rev. B* **56** 12 874
- [90] Fisher A J and Blöchl P E 1993 Adsorption and scanning-tunneling-microscope imaging of benzene on graphite and MoS₂ *Phys. Rev. Lett.* **70** 3263–6
- [91] Goedecker S 1999 Linear scaling electronic structure methods *Rev. Mod. Phys.* **71** 1085
- [92] Galli G 2000 Large-scale electronic structure calculations using linear scaling methods *Status Solidi B* **217** 231
- [93] Fattebert J-L and Bernholc J 2000 Towards grid-based O(N) density-functional theory methods: optimized nonorthogonal orbitals and multigrid acceleration *Phys. Rev. B* **62** 1713–22
- [94] Wang L-W and Teter M P 1992 Kinetic-energy functional of the electron density *Phys. Rev. B* **45** 13 196–220
- [95] Perrot F 1994 Hydrogen–hydrogen interaction in an electron gas *J. Phys.: Condens Matter* **6** 431–46
- [96] Smargiassi E and Madden P A 1994 Orbital-free kinetic-energy functionals for first-principles molecular dynamics *Phys. Rev. B* **49** 5220–6
- [97] Wang Y A and Carter E A 2000 Orbital-free kinetic-energy density functional theory *Theoretical Methods in Condensed Phase Chemistry (Progress in Theoretical Chemistry and Physics Series)* ed S D Schwartz (Boston: Kluwer) pp 117–84
- [98] Wang Y A, Govind N and Carter E A 1998 Orbital-free kinetic energy functionals for the nearly-free electron gas *Phys. Rev. B* **58** 13 465
Wang Y A, Govind N and Carter E A 1999 *Phys. Rev. B* **60** 17 162E
- [99] Wang Y A, Govind N and Carter E A 1999 Orbital-free kinetic-energy density functionals with a density-dependent kernel *Phys. Rev. B* **60** 16 350
- [100] Watson S, Jesson B J, Carter E A and Madden P A 1998 *Ab initio* pseudopotentials for orbital-free density functional *Europhys. Lett.* **41** 37–42

- [101] Anta J A, Jesson B J and Madden P A 1998 Ion–electron correlations in liquid metals from orbital-free *ab initio* molecular dynamics *Phys. Rev. B* **58** 6124–32
- [102] Jesson B J, Foley M and Madden P A 1997 Thermal properties of the self-interstitial in aluminum: an *ab initio* molecular-dynamics study *Phys. Rev. B* **55** 4941–6
- [103] Aoki M I and Tsumuraya K 1997 *Ab initio* molecular-dynamics study of pressure-induced glass-to-crystal transitions in the sodium system *Phys. Rev. B* **56** 2962–8
- [104] Watson S C and Carter E A 2000 Linear-scaling parallel algorithms for the first principles treatment of metals *Comp. Phys. Commun.* **128** 67–92
- [105] Hedin L 1965 New method for calculating the one-particle Green's function with application to the electron–gas problem *Phys. Rev.* **139** A796
- [106] Pisani C, Dovesi R and Roetti C 1988 *Hartree–Fock Ab Initio Treatment of Crystalline Systems (Lecture Notes in Chemistry, vol 48)* (Berlin: Springer)
- [107] CRYSTAL98 is the current version of the commercial HF program developed at the University of Torino and at Daresbury Laboratory (<http://www.dl.ac.uk/TCS/Software/CRYSTAL/>)

-35-

- [108] Su Y-S, Kaplan T A, Mahanti S D and Harrison J F 1999 Crystal Hartree–Fock calculations for La_2NiO_4 and La_2CuO_4 *Phys. Rev. B* **59** 10 521–9
- [109] Fu L, Yaschenko E, Resca L and Resta R 1999 Hartree–Fock studies of surface properties of BaTiO_3 *Phys. Rev. B* **60** 2697–703
- [110] Causà M, Dovesi R and Roetti C 1991 Pseudopotential Hartree–Fock study of seventeen III-V and IV-IV semiconductors *Phys. Rev. B* **43** 11 937–43
- [111] Chartier A, D'Arco P, Dovesi R and Saunders V R 1999 *Ab initio* Hartree–Fock investigation of the structural, electronic, and magnetic properties of Mn_3O_4 *Phys. Rev. B* **60** 14 042–8, and references therein
- [112] McMillan W L 1965 Ground state of liquid ^4He *Phys. Rev.* **138** A442
- [113] Ceperly D M and Kalos M H 1986 Quantum many-body problems, *Monte Carlo Methods in Statistical Physics (Topics in Current Physics, vol 7)* 2nd edn, ed K Binder (Berlin: Springer) pp 145–94
- [114] Rajagopal G, Needs R J, James A, Kenney S D and Foulkes W M C 1995 Variational and diffusion quantum Monte Carlo calculations at nonzero wave vectors: theory and application to diamond-structure germanium *Phys. Rev. B* **51** 10 591–600
- [115] Umrigar C J, Wilson K G and Wilkins J W 1988 Optimized trial wavefunctions for quantum Monte Carlo calculations *Phys. Rev. Lett.* **60** 1719–22
- [116] Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H and Teller E 1953 Equation of state calculations by fast computing machines *J. Chem. Phys.* **21** 1087
- [117] Kent P R C, Hood R Q, Williamson A J, Needs R J, Foulkes W M C and Rajagopal G 1999 Finite-size errors in quantum many-body simulations of extended systems *Phys. Rev. B* **59** 1917–29
- [118] Eckstein H, Schattke W, Reigrotzki M and Redmer R 1996 Variational quantum Monte Carlo ground state of GaAs *Phys. Rev. B* **54** 5512–15
- [119] Hammond B L, Lester W A and Reynolds P J 1994 *Monte Carlo Methods in Ab Initio Quantum Chemistry* (Singapore: World Scientific)
- [120] Ceperly D M and Alder B J 1980 Ground state of the electron gas by a stochastic method *Phys. Rev. Lett.* **45** 566–9
- [121] Towler M D, Hood R Q and Needs R J 2000 Minimum principles and level splitting in quantum Monte Carlo excitation energies: application to diamond *Phys. Rev. B* **62** 2330–7
- [122] Filippi C and Umrigar C J 2000 Correlated sampling in quantum Monte Carlo: a route to forces *Phys. Rev. B* **61** R16 291
- [123] Alfè D, Gillan M J and Price G D 2000 Constraints on the composition of the Earth's core from *ab initio* calculations *Nature* **405** 172–5
- [124] Upton T H and Goddard W A III 1981 Chemisorption of H, Cl, Na, O, and S atoms on Ni(100) surfaces: a theoretical study using Ni_{20} clusters *Crit. Rev. Solid State Mater. Sci.* **10** 261–96

- [125] Bagus P S, Bauschlicher C W Jr, Nelin C J, Laskowski B C and Seel M 1984 A proposal for the proper use of pseudopotentials in molecular orbital cluster model studies of chemisorption *J. Chem. Phys.* **81** 3594–602
- [126] Panas I, Schüle J, Siegbahn P and Wahlgren U 1988 On the cluster convergence of chemisorption energies *Chem. Phys. Lett.* **149** 265–72
 Siegbahn P E M, Nygren M A and Wahlgren U 1992 *Cluster Models for Surface and Bulk Phenomena* ed G Pacchioni, P S Bagus and F Parmigiani (*NATO ASI Series B: Physics vol 283*) (New York: Plenum) p 267
- [127] Panas I, Siegbahn P and Wahlgren U 1989 The mechanism for the O₂ dissociation on Ni(100) *J. Chem. Phys.* **90** 6791–801
- [128] Fournier R and Salahub D R 1990 Chemisorption and magnetization: A bond order-rigid band model *Surf. Sci.* **238** 330–40
 Ushio J, Papai I, St-Amant A and Salahub D R 1992 Vibrational analysis of formate adsorbed on Ni(110): LCGTO-MCP-LSD study *Surf. Sci.* **262** L134–8

-36-

- [129] Mlynarski P and Salahub D R 1991 Local and nonlocal density functional study of Ni₄ and Ni₅ clusters. Models for the chemisorption of hydrogen on (111) and (100) nickel surfaces *J. Chem. Phys.* **95** 6050–6
- [130] Papai I, Salahub D R and Mijoule D 1990 An LCGTO=MCP-LSD study of the (2 × 1) H-covered Pd(110) surface *Surf. Sci.* **236** 241–9
- [131] Rochefort A, Andzelm J, Russo N and Salahub D R 1990 Chemisorption and diffusion of atomic hydrogen in and on cluster models of Pd, Rh, and bimetallic PdSn, RhSn, and RhZn catalysts *J. Am. Chem. Soc.* **112** 8239–47
- [132] Bagus P S and Pacchioni G 1995 Ionic and covalent electronic states for K adsorbed on Cu₅ and Cu₂₅ cluster models of the Cu(100) surface *J. Chem. Phys.* **102** 879
- [133] Clotet A and Pacchioni G 1996 Acetylene on Cu and Pd(111) surfaces: A comparative theoretical study of bonding mechanism, adsorption sites, and vibrational spectra *Surf. Sci.* **346** 91
- [134] Bauschlicher C W Jr 1994 A theoretical study of CO/Cu(100) *J. Chem. Phys.* **101** 3250
- [135] Triguero L, Wahlgren U, Boussard P and Siegbahn P 1995 Calculations of hydrogen chemisorption energies on optimized Cu clusters *Chem. Phys. Lett.* **237** 550
- [136] Illas F, Zurita S, Marquez A M and Rubio J 1997 On the bonding mechanism of CO to Pt(111) and its effect on the vibrational frequency of chemisorbed CO *Surf. Sci.* **376** 279
- [137] Illas F, Rubio J, Ricart J M and Pacchioni G 1996 The importance of correlation effects on the bonding of atomic oxygen on Pt(111) *J. Chem. Phys.* **105** 7192
- [138] Rochefort A, McBreen P and Salahub D R 1996 Bond selectivity in the dissociative adsorption of c-CH₂N₂ on single crystals: a comparative DFT-LSD investigation for Pd(110) and Cu (110) *Surf. Sci.* **347** 11
- [139] Filali Baba M, Mijoule C, Godbout N and Salahub D R 1994 Coadsorption of K and CO on Pd clusters: a density functional study *Surf. Sci.* **316** 349
- [140] Kantorovich L N 1988 An embedded-molecular-cluster method for calculating the electronic structure of point defects in non-metallic crystals. I. General theory *J. Phys. C: Solid State Phys.* **21** 5041
 Meng J, Pandey R, Vail J M and Kunz A B 1989 Impurity potentials derived from embedded quantum clusters: Ag⁺ and Cu⁺ transport in alkali halides *J. Phys.: Condens Matter* **1** 6049–58
 Grimes R W, Catlow C R A and Stoneham A M 1989 A comparison of defect energies in MgO using Mott–Littleton and quantum mechanical procedures *J. Phys.: Condens Matter* **1** 7367–84
 Zuo J, Pandey R and Kunz A B 1991 Embedded-cluster study of the lithium trapped-hole center in magnesium oxide *Phys. Rev. B* **44** 7187–91
 Zuo J, Pandey R and Kunz A B 1992 Embedded-cluster study of Cu⁺-induced lattice relaxation in alkali halides *Phys. Rev. B* **45** 2709–11
 Visser O, Visscher L, Aerts P J C and Nieuwpoort W C 1992 Molecular open shell configuration interaction calculations using the Dirac–Coulomb Hamiltonian: the f⁶-manifold of an embedded EuO_6^{2-} cluster *J. Chem. Phys.* **96** 2910–19
 Pisani C, Orlando R and Cora F 1992 On the problem of a suitable definition of the cluster in embedded-cluster treatments of defects in crystals *J. Chem. Phys.* **97** 4195–204
 Martin R L, Pacchioni G and Bagus P S 1992 *Cluster Models for Surface and Bulk Phenomena* ed G Pacchioni *et al* (*NATO ASI Series B: Physics vol 283*) (New York: Plenum) p 485
 Martin R L, Pacchioni G and Bagus P S 1992 *Cluster Models for Surface and Bulk Phenomena* ed G Pacchioni *et al*

(NATO ASI Series B: Physics vol 283) (New York: Plenum) p 305

Pisani C 1993 Embedded-cluster techniques for the quantum-mechanical study of surface reactivity *J. Mol. Catal.* **82** 229

Hermann K 1992 *Cluster Models for Surface and Bulk Phenomena* ed G Pacchioni *et al* (NATO ASI Series B: Physics vol 283) (New York: Plenum) p 209

-37-

- [141] Nakatsuji H 1987 Dipped adcluster model for chemisorptions and catalytic reactions on metal surface *J. Chem. Phys.* **87** 4995–5001
Nakatsuji H and Nakai H 1990 Theoretical study on molecular and dissociative chemisorptions of an O₂ molecule on an Ag surface: dipped adcluster model combined with symmetry-adapted cluster-configuration interaction method *Chem. Phys. Lett.* **174** 283–6
Nakatsuji H, Nakai H and Fukunishi Y 1991 Dipped adcluster model for chemisorptions and catalytic reactions on a metal surface: Image force correction and applications to Pd-O₂ adclusters *J. Chem. Phys.* **95** 640–7
Nakatsuji H and Nakai H 1992 Dipped adcluster model study for the end-on chemisorption of O₂ on an Ag surface *Can. J. Chem.* **70** 404–8
Nakatsuji H, Kuwano R, Morita H and Nakai H 1993 Dipped adcluster model and SAC-CI method applied to harpooning, chemical luminescence and electron emission in halogen chemisorption on alkali metal surface *J. Mol. Catal.* **82** 211–28
Zhen-Ming Hu and Nakatsuji H 1999 Adsorption and disproportionation reaction of OH on Ag surfaces: dipped adcluster model study *Surf. Sci.* **425** 296–312
- [142] Nakatsuji H 1997 Dipped adcluster model for chemisorption and catalytic reactions *Prog. Surf. Sci.* **54** 1
- [143] Chang T-M, Martinez T J and Carter E A 1994 unpublished results
- [144] Rösch N, Sandl P, Gorling A and Knappe P 1988 Toward a chemisorption cluster model using the LCGTO-X α method: application to Ni(100)/Na *Int. J. Quantum Chem. Symp.* **22** 275
- [145] Weinelt M, Huber W, Zebisch P, Steinrück H-P, Ulbricht P, Birkenheuer U, Boettger J C and Rösch N 1995 The adsorption of acetylene on Ni(110): an experimental and theoretical study *J. Chem. Phys.* **102** 9709
- [146] Pacchioni G, Chung S-C, Krüger S and Rösch N 1997 Is CO chemisorbed on Pt anomalous compared with Ni and Pd? An example of surface chemistry dominated by relativistic effects *Surf. Sci.* **392** 173
- [147] Staufer M *et al* 1999 Interpretation of x-ray emission spectra: NO adsorbed on Ru(001) *J. Chem. Phys.* **111** 4704–13
- [148] Weinert M, Wimmer E and Freeman A J 1982 Total-energy all-electron density functional method for bulk solids and surfaces *Phys. Rev. B* **26** 4571–8
Jansen H J F and Freeman A J 1984 Total-energy full-potential linearized augmented plane-wave method for bulk solids: electronic and structural properties of tungsten *Phys. Rev. B* **30** 561–9
- [149] Wimmer E, Fu C L and Freeman A J 1985 Catalytic promotion and poisoning: all-electron local-density-functional theory of CO on Ni(001) surfaces coadsorbed with K or S *Phys. Rev. Lett.* **55** 2618–21
- [150] Fu C L and Freeman A J 1989 Covalent bonding of sulfur on Ni(001): S as a prototypical adsorbate catalytic poisoner *Phys. Rev. B* **40** 5359
- [151] Wu R and Freeman A J 1992 Structural and magnetic properties of Fe/Ni(111) *Phys. Rev. B* **45** 7205
- [152] Li C, Wu R, Freeman A J and Fu C L 1993 Energetics, bonding mechanism, and electronic structure of metal–ceramic interfaces: Ag/MgO(001) *Phys. Rev. B* **48** 8317–22
- [153] Wu R and Freeman A J 1994 Magnetism at metal–ceramic interfaces: effects of a Au overlayer on the magnetic properties of Fe/MgO(001) *J. Magn. Magn. Mater.* **137** 127–33
- [154] Mannstadt W and Freeman A J 1997 Dynamical and geometrical aspects of NO chemisorption on transition metals: Rh, Pd, and Pt *Phys. Rev. B* **55** 13 298
- [155] Mannstadt W and Freeman A J 1998 LDA theory of the coverage dependence of the local density of states: Li adsorbed on Ru(001) *Phys. Rev. B* **57** 13 289
- [156] Mortensen J J, Hammer B and Norskov J K 1998 Alkali promotion of N₂ dissociation over Ru(0001) *Phys. Rev. Lett.* **80** 4333
Hammer B and Norskov J K 1997 Adsorbate reorganization at steps: NO on Pd(211) *Phys. Rev. Lett.* **79** 4441
Hammer B, Scheffler M, Jacobsen K W and Norskov J K 1994 Multidimensional potential energy surface for H₂ dissociation over Cu(111) *Phys. Rev. Lett.* **73** 1400
Gundersen K, Jacobsen K W, Norskov J K and Hammer B 1994 The energetics and dynamics of H₂ dissociation on

- [157] Wei C M, Gross A and Scheffler M 1998 *Ab initio* calculation of the potential energy surface for the dissociation of H₂ on the sulfur-covered Pd(100) surface *Phys. Rev. B* **57** 15 572
Tomanek D, Wilke S and Scheffler M 1997 Hydrogen-induced polymorphism of the Pd(110) surface *Phys. Rev. Lett.* **79** 1329
Stampfl C and Scheffler M 1997 Mechanism of efficient carbon monoxide oxidation at Ru(0001) *J. Vac. Sci. Technol. A* **15** 1635
Stampfl C and Scheffler M 1997 Anomalous behavior of Ru for catalytic oxidation: a theoretical study of the catalytic reaction CO+1/2 O₂ to CO₂ *Phys. Rev. Lett.* **78** 1500
Stampfl C and Scheffler M 1996 Theoretical study of O adlayers on Ru(0001) *Phys. Rev. B* **54** 2868
- [158] Philipsen P H T, van Lenthe E, Snijders J G and Baerends E J 1997 Relativistic calculations on the adsorption of CO on the (111) surfaces of Ni, Pd and Pt within the zeroth-order regular approximation *Phys. Rev. B* **56** 13 556
Olsen R A, Philipsen P H T, Baerends E J, Kroes G J and Louvik O M 1997 Direct subsurface adsorption of hydrogen on Pd(111): quantum mechanical calculations on a new two-dimensional potential energy surface *J. Chem. Phys.* **106** 9286
Wiesenecker G, Kroes G J and Baerends E J 1996 An analytical six-dimensional potential energy surface for dissociation of molecular hydrogen on Cu(100) *J. Chem. Phys.* **104** 7344
Philipsen P H T, te Velde G and Baerends E J 1994 The effect of density-gradient corrections for a molecule-surface potential energy surface. Slab calculations on Cu(100)c(2x2)-CO *Chem. Phys. Lett.* **226** 583
- [159] te Velde G and Baerends E J 1993 Slab versus cluster approach for chemisorption studies, CO on Cu(100) *Chem. Phys.* **177** 399
- [160] Feibelman P J, Hafner J and Kresse G 1998 Vibrations of O on stepped Pt(111) *Phys. Rev. B* **58** 2179–84
- [161] Eichler A, Kresse G and Hafner J 1998 *Ab-initio* calculations of the 6D potential energy surfaces for the dissociative adsorption of H₂ on the (100) surfaces of Rh, Pd and Ag *Surf. Sci.* **397** 116–36
- [162] Rösch N 1998 *Lecture Given at the 7th International Symposium on Theoretical Aspects of Heterogeneous Catalysis, Cambridge, 25-28 August*
- [163] Hammer B, Hansen L B and Nørskov J K 1999 Improved adsorption energetics within density functional theory using revised Perdew–Burke–Eberhart functionals *Phys. Rev. B* **59** 7413–21
- [164] Whitten J L and Pakkanen T A 1980 Chemisorption theory for metallic surfaces: Electron localization and the description of surface interactions *Phys. Rev. B* **21** 4357–67
Madhavan P and Whitten J L 1982 Theoretical studies of the chemisorption of hydrogen on copper *J. Chem. Phys.* **77** 2673–83
Cremaschi P and Whitten J L 1987 The effect of hydrogen chemisorption on titanium surface bonding *Theor. Chim. Acta.* **72** 485–96
Whitten J L 1992 *Cluster Models for Surface and Bulk Phenomena* ed G Pacchioni *et al* (NATO ASI Series B: Physics vol 283) (New York: Plenum) p 375
Whitten J L 1993 Theoretical studies of surface reactions: embedded cluster theory *Chem. Phys.* **177** 387–97
- [165] Duarte H A and Salahub D R 1998 Embedded cluster model for chemisorption using density functional calculations: oxygen adsorption on the Al(100) surface *J. Chem. Phys.* **108** 743
- [166] Sellers H 1991 On modeling chemisorption processes with metal cluster systems. II. Model atomic potentials and site specificity of N atom chemisorption on Pd(111) *Chem. Phys. Lett.* **178** 351–7
- [167] Ravenek W and Geurts F M M 1986 Hartree–Fock–Slater–LCAO implementation of the moderately large-embedded-cluster approach to chemisorption. Calculations for hydrogen on lithium (100) *J. Chem. Phys.* **84** 1613–23
- [168] Fukunishi Y and Nakatsuji H 1992 Modifications for *ab initio* calculations of the moderately large-embedded-cluster model. Hydrogen adsorption on a lithium surface *J. Chem. Phys.* **97** 6535–43
- [169] Pisani C 1978 Approach to the embedding problem in chemisorption in a self-consistent-field-molecular-orbital formalism *Phys. Rev. B* **17** 3143
Pisani C, Dovesi R and Nada R 1990 *Ab initio* Hartree–Fock perturbed-cluster treatment of local defects in crystals *J. Chem. Phys.* **92** 7448
Pisani C 1993 Embedded-cluster techniques for the quantum-mechanical study of surface reactivity *J. Mol. Catal.* **82** 229
Casassa S and Pisani C 1995 Atomic-hydrogen interaction with metallic lithium: an *ab initio* embedded-cluster study *Phys. Rev. B* **51** 7805

- [170] Gutdeutsch U, Birkenheuer U, Krüger S and Rösch N 1997 On cluster embedding schemes based on orbital space partitioning *J. Chem. Phys.* **106** 6020
- [171] Gutdeutsch U, Birkenheuer U and Rösch N 1998 A strictly variational procedure for cluster embedding based on the extended subspace approach *J. Chem. Phys.* **109** 2056
- [172] Ellis D E, Benesh G A and Byrom E 1978 Self-consistent embedded-cluster model for magnetic impurities: β' -NiAl *J. Appl. Phys.* **49** 1543
Ellis D E, Benesh G A and Byrom E 1979 Self-consistent embedded-cluster model for magnetic impurities: Fe, Co, and Ni in β' -NiAl *Phys. Rev. B* **20** 1198
- [173] Benesh G A and Inglesfield J E 1984 An embedding approach for surface calculations *J. Phys. C: Solid Phys.* **17** 1595
Inglesfield J E and Benesh G A 1988 Surface electronic structure: embedded self-consistent calculations *Phys. Rev. B* **37** 6682
Aers G C and Inglesfield J E 1989 Electric field and Ag(001) surface electronic structure *Surf. Sci.* **217** 367
Colbourn E A and Inglesfield J E 1991 Effective charges and surface stability of O on Cu(001) *Phys. Rev. Lett.* **66** 2006
Crampin S, van Hoof J B A N, Nekovee M and Inglesfield J E 1992 Full-potential embedding for surfaces and interfaces *J. Phys.: Condens Matter* **4** 1475
Benesh G A and Liyanage L S G 1994 Surface-embedded Green-function method for general surfaces: application to Al(111) *Phys. Rev. B* **49** 17 264
Trioni M I, Brivio G P, Crampin S and Inglesfield J E 1996 Embedding approach to the isolated adsorbate *Phys. Rev. B* **53** 8052–64
- [174] Scheffler M, Droste Ch, Fleszar A, Maca F, Wachutka G and Barzel G 1991 A self-consistent surface-Green-function (SSGF) method *Physica B* **172** 143
Wachutka G, Fleszar A, Maca F and Scheffler M 1992 Self-consistent Green-function method for the calculation of electronic properties of localized defects at surfaces and in the bulk *J. Phys.: Condens Matter* **4** 2831
Bormet J, Neugebauer J and Scheffler M 1994 Chemical trends and bonding mechanisms for isolated adsorbates on Al(111) *Phys. Rev. B* **49** 17 242
Wenzi B, Bormet J and Scheffler M 1995 Green function for crystal surfaces I *Comp. Phys. Commun.* **88** 230
- [175] Feibelman P J 1987 Force and total-energy calculations for a spatially compact adsorbate on an extended, metallic crystal surface *Phys. Rev. B* **35** 2626
- [176] Williams A R, Feibelman P J and Lang N D 1982 Green's-function methods for electronic-structure calculations *Phys. Rev. B* **26** 5433
- [177] Feibelman P J 1991 Orientation dependence of the hydrogen molecule's interaction with Rh(001) *Phys. Rev. Lett.* **67** 461
- [178] Feibelman P J 1994 Sulfur adsorption near a step on Al *Phys. Rev. B* **49** 14 632
- [179] Feibelman P J 1994 Diffusion barrier for a Ag adatom on Pt(111) *Surf. Sci.* **313** L801
- [180] Krüger S and Rösch N 1994 The moderately-large-embedded-cluster method for metal surfaces; a density-functional study of atomic adsorption *J. Phys.: Condens Matter* **6** 8149
Krüger S, Birkenheuer U and Rösch N 1994 Density functional approach to moderately large cluster embedding for infinite metal substrates *J. Electron Spectrosc. Relat. Phenom.* **69** 31
- [181] Head J D and Silva S J 1996 A localized orbitals based embedded cluster procedure for modeling chemisorption on large finite clusters and infinitely extended surfaces *J. Chem. Phys.* **104** 3244
- [182] Brivio G P and Trioni M I 1999 The adiabatic molecule-metal surface interaction: theoretical approaches *Rev. Mod. Phys.* **71** 231–65
- [183] Cortona P 1991 Self-consistently determined properties of solids without band structure calculations *Phys. Rev. B* **44** 8454
Cortona P 1992 Direct determination of self-consistent total energies and charge densities of solids: A study of the cohesive properties of the alkali halides *Phys. Rev. B* **46** 2008
- [184] Wesolowski T A and Warshel A 1993 Frozen density functional approach to *ab initio* calculations of solvated molecules *J. Phys. Chem.* **97** 8050
Wesolowski T A and Warshel A 1994 *Ab initio* free energy perturbation calculations of solvation free energy using the frozen density functional approach *J. Phys. Chem.* **98** 5183
-

- [185] Stefanovich E V and Truong T N 1996 Embedded density functional approach for calculations of adsorption on ionic crystals *J. Chem. Phys.* **104** 2946
- [186] Garcia A, Elsässer C, Zhu J, Louie S G and Cohen M L 1992 Use of gradient-corrected functionals in total-energy calculations for solids *Phys. Rev. B* **46** 9829
- [187] Nachtigall P, Jordan K D, Smith A and Jónsson H 1996 Investigation of the reliability of density functional methods: reaction and activation energies for Si-Si bond cleavage and H₂ elimination from silanes *J. Chem. Phys.* **104** 148
- [188] Perdew J P, Burke K and Ernzerhof M 1996 Generalized gradient approximation made simple *Phys. Rev. Lett.* **77** 3865
 Proynov E I, Sirois S and Salahub D R 1997 Extension of the LAP functional to include parallel spin correlation *Int. J. Quantum Chem.* **64** 427
 Tozer D J, Handy N C and Green W H 1997 Exchange-correlation functionals from *ab initio* electron densities *Chem. Phys. Lett.* **273** 183
 Filatov M and Thiel W 1997 A new gradient-corrected exchange-correlation density functional *Mol. Phys.* **91** 847
 van Voorhis T and Scuseria G E 1998 A novel form for the exchange-correlation energy functional *J. Chem. Phys.* **109** 400
 Zhang Y and Yang W 1998 *Phys. Rev. Lett.* **80** 890
 Tozer D J and Handy N C 1998 The development of new exchange-correlation functionals *J. Chem. Phys.* **108** 2545
- [189] Abarenkov I V, Bulatov V L, Godby R, Heine V, Payne M C, Souchko P V, Titov A V and Tupitsyn I I 1997 Electronic-structure multiconfiguration calculation of a small cluster embedded in a local-density approximation host *Phys. Rev. B* **56** 1743
- [190] Govind N, Wang Y A, da Silva A J R and Carter E A 1998 Accurate *ab initio* energetics of extended systems via explicit correlation embedded in a density functional environment *Chem. Phys. Lett.* **295** 129
- [191] Govind N, Wang Y A and Carter E A 1999 Electronic structure calculations by first principles density-based embedding of explicitly correlated systems *J. Chem. Phys.* **110** 7677
- [192] Kluener T, Wang Y A, Govind N and Carter E A 2000 in preparation
- [193] Rubio A, Corkill J L, Cohen M L, Shirley E L and Louie S G 1993 Quasiparticle band structure of AlN and GaN *Phys. Rev. B* **48** 11 810–16
- [194] Dhesi S S, Stagaescu C B, Smith K E, Doppalapudi D, Singh R and Moustakas T D 1997 Surface and bulk electronic structure of thin-film wurtzite GaN *Phys. Rev. B* **56** 10 271–5
- [195] Starrost F 1999 *PhD Thesis* Christian-Albrechts-Universität Kiel
 Starrost F, Krasovskii E E and Schattke W 1999 unpublished
- [196] Günther O, Janowitz C, Jungk G, Jenichen B, Hey R, Däweritz L and Ploog K 1995 Comparison between the electronic dielectric functions of a GaAs/AlAs superlattice and its bulk components by spectroscopic ellipsometry using core levels *Phys. Rev. B* **52** 2599–609
- [197] Starrost F, Krasovskii E E and Schattke W 1998 An alternative full-potential ELAPW method *Verhandl. DPG (VI)* **33** 741
- [198] Aspnes D E and Studna A A 1983 Dielectric functions and optical parameters of Si, Ge, GaP, GaAs, GaSb, InP, InAs, and InSb from 1.5 to 6.0 eV *Phys. Rev. B* **27** 985–1009
- [199] Logothetidis S, Alouani M, Garriga M and Cardona M 1990 E_2 interband transitions in Al_xGa_{1-x}As alloys *Phys. Rev. B* **41** 2959–65
- [200] Hughes J L P and Sipe J E 1996 Calculation of second-order optical response in semiconductors *Phys. Rev. B* **53** 10 751–63
- [201] Wang C S and Klein B M 1981 First-principles electronic structure of Si, Ge, GaP, GaAs, ZnS and ZnSe. II. Optical properties *Phys. Rev. B* **24** 3417–29
- [202] Huang Ming-Zhu and Ching W Y 1993 Calculation of optical excitations in cubic semiconductors. I. Electronic structure and linear response *Phys. Rev. B* **47** 9449–63

FURTHER READING

Pisani C (ed) 1996 *Quantum-Mechanical Ab-initio Calculation of the Properties of Crystalline Materials (Lecture Notes in Chemistry vol 67)* (Berlin: Springer)

A general introduction.

Dreizler R M and Gross E K U 1990 *Density Functional Theory: an Approach to the Quantum Many-body Problem* (Berlin: Springer)

A monograph on the foundations of density functional theory.

Pisani C, Doves R and Roetti C 1988 *Hartree–Fock Ab Initio Treatment of Crystalline Systems (Lecture Notes in Chemistry vol 48)* (Berlin: Springer)

An introduction to periodic Hartree–Fock.

Nemoshkalenko V V and Antonov V N 1998 *Computational Methods in Solid State Physics* (Amsterdam: Gordon and Breach)

An explicit introduction to the all-electron methods.

Singh D J 1994 *Planewaves, Pseudopotentials and the LAPW Method* (Norwell, MA: Kluwer)

A textbook on plane-wave and LAPW methods.

Whitten J L and Yang H 1996 Theory of chemisorption and reactions on metal surfaces *Surf. Sci. Rep.* **24** 59–124

-1-

B3.3 Statistical mechanical simulations

Michael P Allen

B3.3.1 INTRODUCTION

Computer simulation, at the molecular level, has grown enormously in importance over the last 50 years. Affordable computer chips have historically doubled in power every 18 months, so the computer simulator, regarded as an experimentalist, has the unique advantage of rapidly improving apparatus. With the recent explosion in personal computing, there seems every prospect that this situation will continue, allowing computer simulation to become of even more practical value in fields such as the design of drugs and molecular materials. This provides a stimulus to develop simulation methods, and an industry has grown up marketing the necessary software.

This chapter concentrates on describing molecular simulation methods which have a connection with the statistical mechanical description of condensed matter, and hence relate to theoretical approaches to understanding phenomena such as phase equilibria, rare events, and quantum mechanical effects.

B3.3.1.1 THE AIMS OF SIMULATION

We carry out computer simulations in the hope of understanding bulk, macroscopic properties in terms of the microscopic details of molecular structure and interactions. This serves as a complement to conventional experiments, enabling us to learn something new; something that cannot be found out in other ways.

Computer simulations act as a bridge between microscopic length and time scales and the macroscopic world of the laboratory (see [figure B3.3.1](#)). We provide a guess at the interactions between molecules, and obtain ‘exact’ predictions of bulk properties. The predictions are ‘exact’ in the sense that they can be made as accurate as we like, subject to the limitations imposed by our computer budget. At the same time, the hidden detail behind bulk measurements can be revealed. Examples are the link between the diffusion coefficient and

velocity autocorrelation function (the former easy to measure experimentally, the latter much harder); and the connection between equations of state and structural correlation functions.

Simulations act as a bridge in another sense: between theory and experiment (see [figure B3.3.2](#) . We can test a theory using idealized models, conduct ‘thought experiments’, and clarify what we measure in the laboratory. We may also carry out simulations on the computer that are difficult or impossible in the laboratory (for example, working at extremes of temperature or pressure).

-2-

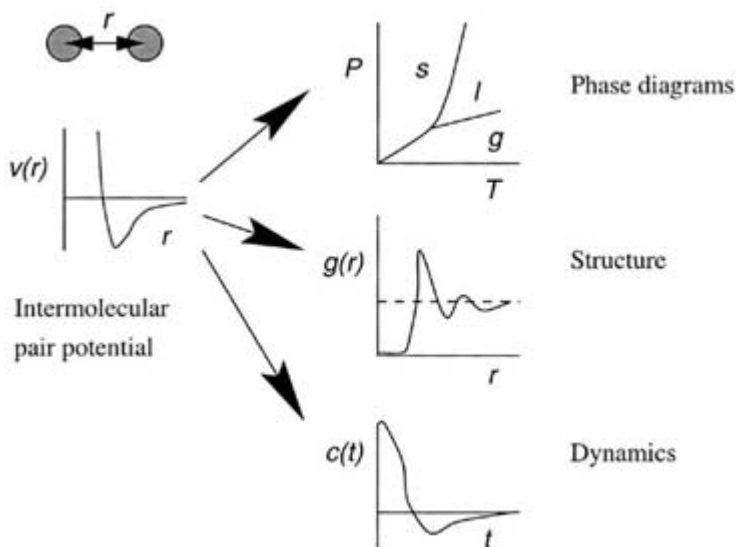


Figure B3.3.1. Simulations as a bridge between the microscopic and the macroscopic. We input details of molecular structure and interactions; we obtain predictions of phase behaviour, structural and time-dependent properties.

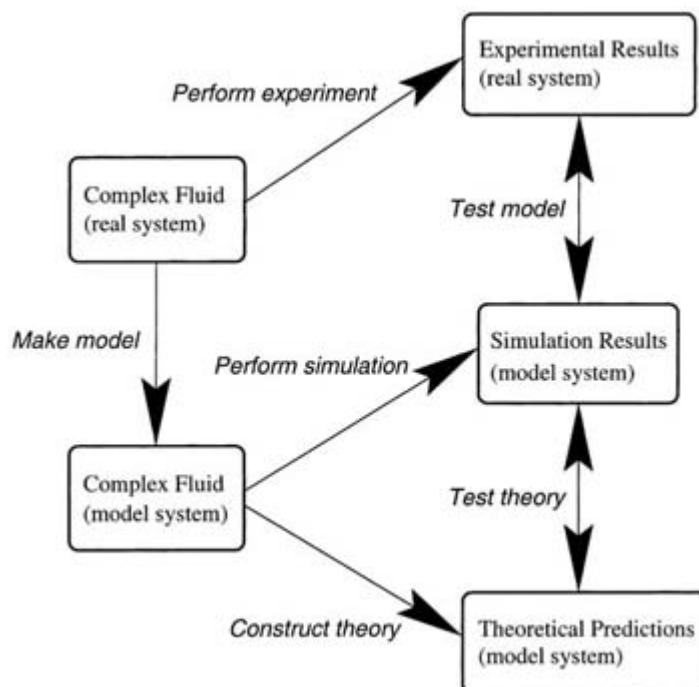


Figure B3.3.2. Simulation as a bridge between theory and experiment. We may test a theory by conducting a

simulation using the same model. We may test the model by comparing with experimental results.

-3-

Ultimately we may want to make direct comparisons with experimental measurements made on specific materials, in which case a good model of molecular interactions is essential. The aim of so-called *ab initio* molecular dynamics is to reduce the amount of fitting and guesswork in this process to a minimum. On the other hand, we may be interested in phenomena of a rather generic nature, or we may simply want to discriminate between good and bad theories. When it comes to aims of this kind, it is not necessary to have a perfectly realistic molecular model; one that contains the essential physics may be quite suitable.

The two main families of simulation technique are molecular dynamics (MD) and Monte Carlo (MC). Additionally, there is a whole range of hybrid techniques which combine features from both MC and MD.

B3.3.1.2 THE TECHNIQUES OF SIMULATION

Molecular dynamics consists of the brute-force solution of Newton's equations of motion. It is necessary to encode in the program the potential energy and force law of interaction between molecules; the equations of motion are solved numerically, by finite difference techniques. The system evolution corresponds closely to what happens in 'real life' and allows us to calculate dynamical properties, as well as thermodynamic and structural functions. For a range of molecular models, packaged routines are available, either commercially or through the academic community.

Monte Carlo can be thought of as a prescription for sampling configurations from a statistical ensemble. The interaction potential energy is coded into the program, and a random walk procedure adopted to go from one state of the system to the next. MC programs can be relatively easy to program; they allow us to calculate thermodynamic and structural properties, but not exact dynamics. It is relatively simple to specify external conditions (constant temperature, pressure etc.) and many tricks may be devised to improve the efficiency of the sampling.

Both MD and MC techniques evolve a finite-sized molecular configuration forward in time, in a step-by-step fashion. (In this context, MC simulation 'time' has to be interpreted liberally, but there is a broad connection between real time and simulation time (see [1, chapter 2].) Common features of MD and MC simulation techniques are that there are limits on the typical timescales and length scales that can be investigated. The consequences of finite size must be considered both in specifying the molecular interactions, and in analysing the results.

B3.3.2 SIMULATION AND STATISTICAL MECHANICS

Here we consider various aspects of statistical mechanics (see also [chapter A2.3](#) and [2, 3]) that have a direct bearing on computer simulation methodology.

B3.3.2.1 SIMULATION TIME AND LENGTH SCALES

Simulation runs are typically short ($t \sim 10^3 - 10^6$ MD or MC steps, corresponding to perhaps a few nanoseconds of real time) compared with the time allowed in laboratory experiments. This means that we need to test whether or not a simulation has reached equilibrium before we can trust the averages calculated in it. Moreover, there is a clear need to subject the simulation averages to a statistical analysis, to make a realistic estimate of the errors.

How long should we run? This depends on the system and the physical properties of interest. Suppose that we are interested in a variable X , defined such that its ensemble average $X = \langle \chi \rangle = 0$. (Here and throughout we use script letters for instantaneous dynamical variables, i.e., functions of coordinates and momenta, to distinguish them from averages and thermodynamic quantities.) A characteristic time, τ , may be defined, over which the correlations $\langle \chi(0)\chi(t) \rangle$ decay towards zero. The simulation run time t_{run} should be significantly longer than τ . The time scales of properties of interest will vary from one system to another; they may not be predictable in advance, and this will have a bearing on the length of simulation required.

Similar considerations apply to the size of system simulated. The samples involved are typically quite small on the laboratory scale. Most fall in the range $N \sim 10^3$ – 10^6 particles, thus imposing a restriction on the length scales of the phenomena that may be investigated, in the nanometre–submicron range. Indeed, in many cases, there is an overriding need to do a system-size analysis of simulation results, to quantify these effects.

How large a simulation do we need? Once more this depends on the system and properties of interest. From a spatial correlation function $\langle \chi(0)\chi(r) \rangle$ relating values computed at different points r apart, we may define a characteristic distance ξ over which the correlation decays. The simulation box size L should be significantly larger than ξ in order not to influence the results.

The ratios, t_{run}/τ and L/ξ , appear in expressions for estimating the errors on simulation-averaged quantities. Roughly speaking, a simulation sample can be regarded as a collection of $\sim(L/\xi)^3$ sub-samples, each making a statistically independent contribution to the average properties. Also, a simulation run may be regarded as a succession of $\sim t_{\text{run}}/\tau$ statistically independent sub-runs. Then, the usual rules for combining independent samples apply, and estimated error bars are inversely proportional to the square root of the run time. For further information see [4, 5, 6 and 7].

Near critical points, special care must be taken, because the inequality $L \gg \xi$ will almost certainly not be satisfied; also, critical slowing down will be observed. In these circumstances a quantitative investigation of finite size effects and correlation times, with some consideration of the appropriate scaling laws, must be undertaken. Examples of this will be seen later; one of the most encouraging developments of recent years has been the establishment of reliable and systematic methods of studying critical phenomena by simulation.

B3.3.2.2 PERIODIC BOUNDARY CONDITIONS

Small sample size means that, unless surface effects are of particular interest, periodic boundary conditions need to be used. Consider 1000 atoms arranged in a $10 \times 10 \times 10$ cube. Nearly half the atoms are on the outer faces, and these will have a large effect on the measured properties. Surrounding the cube with replicas of itself takes care of this problem. Provided the potential range is not too long, we can adopt the *minimum image convention* that each atom interacts with the nearest atom or image in the periodic array. In the course of the simulation, if an atom leaves the basic simulation box, attention can be switched to the incoming image. This is shown in [figure B3.3.3](#). Of course, it is important to bear in mind the imposed artificial periodicity when considering properties which are influenced by long-range correlations. Special attention must be paid to the case where the potential range is not short: for example, for charged and dipolar systems. Methods for handling this are discussed later.

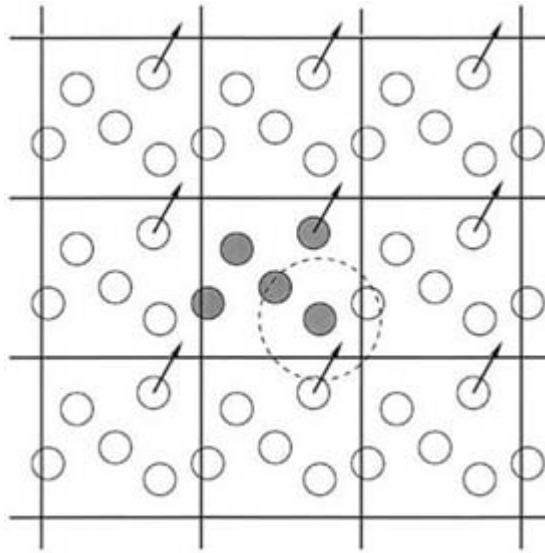


Figure B3.3.3. Periodic boundary conditions. As a particle moves out of the simulation box, an image particle moves in to replace it. In calculating particle interactions within the cutoff range, both real and image neighbours are included.

B3.3.2.3 MOLECULAR INTERACTIONS

Let us denote a ‘state of the system’ by γ . For the purposes of discussion, we shall concentrate on a system composed of atoms, and for this γ represents the complete set of coordinates $\mathbf{r}^{(N)} = (r_1, r_2, \dots, r_N)$ and conjugate momenta $\mathbf{p}^{(N)} = (p_1, p_2, \dots, p_N)$. Then the energy, or Hamiltonian, may be written as a sum of kinetic and potential terms $\mathcal{H} = \mathcal{K} + \mathcal{V}$. For atomic systems, \mathcal{V} is a function of coordinates only and \mathcal{K} may be written as a function of momenta; in molecular systems represented as rigid bodies, or in terms of generalized coordinates, the kinetic energy may also depend on the coordinates [8].

Sticking, for simplicity, with a simple atomic system, the kinetic energy may be written

$$\mathcal{K}(\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_N) = \sum_{i=1}^N \sum_{\alpha=x,y,z} p_{i\alpha}^2 / 2m_i.$$

The potential energy \mathcal{V} is traditionally split into one-body, two-body, three-body . . . terms:

$$\mathcal{V}(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N) = \sum_i v^{(1)}(\mathbf{r}_i) + \sum_i \sum_{j>i} v^{(2)}(\mathbf{r}_i, \mathbf{r}_j) + \sum_i \sum_{j>i} \sum_{k>j} v^{(3)}(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k) + \dots.$$

The $v^{(1)}$ term represents an externally applied potential field or the effects of the container walls; it is usually dropped for fully periodic simulations of bulk systems. Also, it is usual to neglect $v^{(3)}$ and higher terms (which in reality might be of order 10% of the total energy in condensed phases) and concentrate on $v^{(2)}$. For brevity henceforth we will just call this $v(r)$. There is an extensive literature on the way these potentials are determined experimentally, or modelled

theoretically (see, e.g., [9, 10 and 11]). In simulations, it is common to use the simplest models that faithfully represent the essential physics: the hard-sphere, square-well, and Lennard-Jones potentials have the longest

history. The latter has the functional form

$$v^{\text{LJ}}(r) = 4\varepsilon \left\{ \left(\frac{d}{r} \right)^{12} - \left(\frac{d}{r} \right)^6 \right\}$$

with two parameters: d , the diameter, and ε , the well depth. This potential was used, for instance, in the earliest studies of the properties of liquid argon [12, 13]. For molecular systems, we simply build the molecules out of site–site potentials of this, or similar, form (figure B3.3.4). If electrostatic charges are present, we add the appropriate Coulomb potentials

$$v^{\text{Coulomb}}(r) = \frac{Q_1 Q_2}{4\pi \epsilon_0 r}$$

where Q_1, Q_2 are the charges. We may also use rigid-body potentials which depend on centre of mass positions and orientations. An example is the Gay–Berne potential [14]

$$v^{\text{GB}}(r, \hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2) = 4\varepsilon(\hat{\mathbf{r}}, \hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2)[\varrho^{-12} - \varrho^{-6}]$$

with

$$\varrho = \frac{r - d(\hat{\mathbf{r}}, \hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2) + d_0}{d_0}$$

which depends upon the molecular axis vectors $\hat{\mathbf{u}}_1$ and $\hat{\mathbf{u}}_2$, and on the direction $\hat{\mathbf{r}}$ and magnitude r of the centre–centre vector $\mathbf{r} = \mathbf{r}_1 - \mathbf{r}_2$. The parameter d_0 determines the smallest molecular diameter and there are two orientation-dependent quantities in the above shifted Lennard-Jones form: a diameter $d(\hat{\mathbf{r}}, \hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2)$ and an energy $\varepsilon(\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2)$. Each quantity depends in a complicated way (not given here) on parameters characterizing molecular shape and structure. This potential has been extensively used in the study of molecular liquids and liquid crystals [15, 16, 17, 18, 19 and 20].

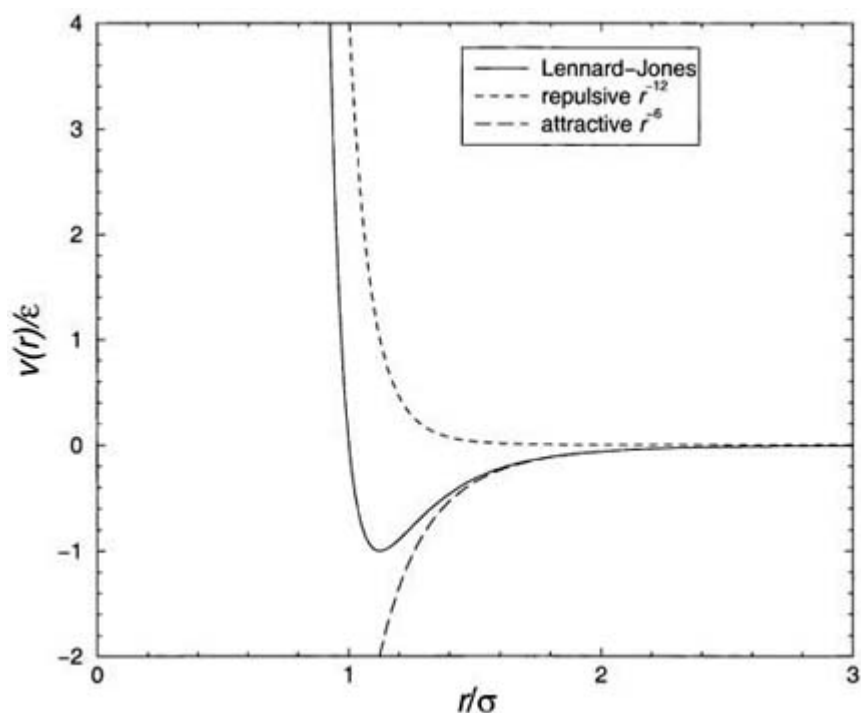


Figure B3.3.4. Lennard-Jones pair potential showing the r^{-12} and r^{-6} contributions.

It is common practice in classical computer simulations *not* to attempt to represent intramolecular bonds by terms in the potential energy function, because these bonds have very high vibration frequencies and should really be treated in a quantum mechanical way rather than in the classical approximation. Instead, the bonds are treated as being constrained to have fixed length, and some straightforward ways have been devised to incorporate these constraints into the dynamics (see later).

For a wide range of physical problems, a lattice spin system provides a useful, if very coarse-grained, description. The great advantage of such an approach is the speed with which such systems may be simulated, especially when a single spin may be taken to represent not just one molecule but a larger region of the physical system. The state γ of such systems may be specified by a set of discrete or continuous spin values $\sigma^{(N)} = (\sigma_1, \sigma_2, \dots, \sigma_N)$, where $\sigma_i = \pm 1$ for the archetypal Ising model, but takes other values for other models. For Ising-like systems, the energy may be written $\mathcal{H} = -J \sum_{\langle i,j \rangle} \sigma_i \sigma_j$, where $\sum_{\langle i,j \rangle}$ indicates a sum over nearest neighbour spins i, j and J is a coupling constant. Again, there is much flexibility in the nature of the energy function. There is an extensive literature on spin simulations [6, 21, 22 and 23] especially in relation to the theory of critical phenomena [24]. Spin models are not restricted to the obvious area of magnetic solids; it has proved possible to include, for instance, polymer liquids in this class [25], allowing the study of otherwise inaccessible behaviour.

B3.3.2.4 SIMULATIONS AND ENSEMBLES

One of the flexibilities of computer simulation is that it is possible to define the thermodynamic conditions corresponding to one of many statistical ensembles, each of which may be most suitable for the purpose of the study. A knowledge of the underlying statistical mechanics is essential in the design of correct simulation methods, and in the analysis of simulation results. Here we describe two of the most common statistical ensembles, but examples of the use of other ensembles will appear later in the chapter.

The microcanonical ensemble corresponds to an isolated system, with specified number of particles N ,

volume V , and energy E . The fundamental thermodynamic potential is the entropy, and it is related to statistical mechanical quantities as follows:

$$(B3.3.1)$$

Here, Ω_{NVE} is the number of states available to the system at given NVE , written as an integral over a thin energy shell; $\delta(\dots)$ is the Dirac delta function. The ensemble average $\langle \chi \rangle_{NVE}$ is defined in terms of the ensemble probability density function $Q_{NVE}(\Gamma)$.

The canonical ensemble corresponds to a system of fixed N and V , able to exchange energy with a thermal bath at temperature T , which represents the effects of the surroundings. The thermodynamic potential is the Helmholtz free energy, and it is related to the partition function Q_{NVT} as follows:

$$\begin{aligned} A &= E - TS = -kT \ln Q_{NVT} \\ Q_{NVT} &= \int d\Gamma e^{-\beta \mathcal{H}(\Gamma)} = \int dE \Omega_{NVE} e^{-\beta E} \\ Q_{NVT}(\Gamma) &= Q_{NVT}^{-1} e^{-\beta \mathcal{H}(\Gamma)} \\ \langle \mathcal{X} \rangle_{NVT} &= \int d\Gamma Q_{NVT}(\Gamma) \mathcal{X}(\Gamma). \end{aligned} \quad (B3.3.2)$$

Here $\beta=1/kT$. In a real system the thermal coupling with surroundings would happen at the surface; in simulations we avoid surface effects by allowing this to occur homogeneously. The state of the surroundings defines the temperature T of the ensemble.

Since $\mathcal{H}=\mathcal{K} + \mathcal{V}$, the canonical ensemble partition function factorizes into ideal gas and excess parts, and as a consequence most averages of interest may be split into corresponding ideal and excess components, which sum to give the total. In MC simulations, we frequently calculate just the excess or configurational parts: in this case, γ consists just of the atomic coordinates, not the momenta, and the appropriate expressions are obtained from equation b3.3.2 by replacing \mathcal{H} by the potential energy \mathcal{V} . The ideal gas contributions are usually easily calculated from exact

expressions, in which the integrations over atomic momenta have been carried out analytically.

B3.3.2.5 AVERAGES AND DISTRIBUTIONS

It is generally well known that, for most averages, differences between ensembles disappear in the thermodynamic limit. However, for finite-sized systems of the kind studied in simulations, it is necessary to consider the differences between ensembles, which will be significant for mean-squared values (fluctuations) and, more generally, for the probability distributions of measured quantities. For example, energy fluctuations in the constant- NVE ensemble are (by definition) zero, whereas in the constant- NVT ensemble they are not. Since these points have a bearing on various aspects of simulation methodology, we expand on them a little here.

It is a standard result in the canonical ensemble that energy fluctuations are related to the heat capacity $C_V = (\partial E / \partial T)_V$:

$$kT^2 C_V = \langle \mathcal{H}^2 \rangle - \langle \mathcal{H} \rangle^2 = \langle \delta \mathcal{H}^2 \rangle.$$

Since C_V and E are both extensive properties ($\propto N$), the root-mean-square energy fluctuations are smaller, by a factor $1/\sqrt{N}$, than typical average energies E . As the system size increases, the relative magnitude of fluctuations decreases, and the thermodynamic limit is achieved.

It is instructive to see this in terms of the canonical ensemble probability distribution function for the energy, $\mathcal{P}_{NVT}(E)$. Referring to [equation B3.3.1](#) and [equation \(B3.3.2\)](#), it is relatively easy to see that

$$\mathcal{P}_{NVT}(E) = \langle \delta(\mathcal{H}(\Gamma) - E) \rangle_{NVT} = \frac{\Omega_{NVE} e^{-E/kT}}{\mathcal{Q}_{NVT}}.$$

The product of a rapidly increasing function of energy, Ω_{NVE} , and a rapidly decreasing function $e^{-E/kT}$, gives a distribution of energies which is very sharply peaked about the most likely value, as shown in figure B3.3.5. A reasonable first approximation to $\mathcal{P}_{NVT}(E)$ is a Gaussian function, centred on this most probable value, with a width determined by C_V .

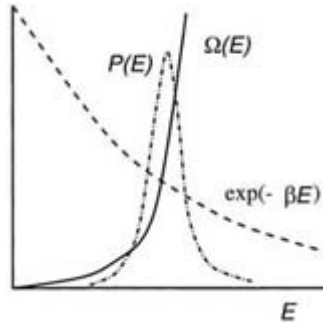


Figure B3.3.5. Energy distributions. The probability density is proportional to the product of the density of states and the Boltzmann factor.

-10-

In principle, these formulae may be used to convert results obtained at one state point into averages appropriate to a neighbouring state point. For any canonical ensemble average

$$\langle \mathcal{X} \rangle_{NVT_1} = \frac{\langle \mathcal{X} e^{-\delta\beta \mathcal{H}} \rangle_{NVT_0}}{\langle e^{-\delta\beta \mathcal{H}} \rangle_{NVT_0}} \quad (\text{B3.3.3})$$

$$\mathcal{P}_{NVT_1}(E) \equiv \langle \delta(\mathcal{H} - E) \rangle_{NVT_1} = \mathcal{P}_{NVT_0}(E) \frac{e^{-\delta\beta E}}{\langle e^{-\delta\beta \mathcal{H}} \rangle_{NVT_0}}. \quad (\text{B3.3.4})$$

where $\delta\beta = \beta_1 - \beta_0$. Choosing $\chi = \delta(\mathcal{H} - E)$ gives a way of re-weighting the energy distribution

Such histogram re-weighting techniques have a long history [26, 27, 28 and 29]. The usefulness of this equation depends sensitively on accurate sampling of energies in the region of interest, at T_1 , which may be far away from the maximum in \mathcal{P} . We have seen how the mean-squared fluctuations in E are related to the heat capacity; higher-order, non-Gaussian terms in $\mathcal{P}_{NVT}(E)$ are also related to thermodynamic derivatives [30, 31], but they are smaller, and so are hard to measure accurately. This limits the extension to nearby state points; a common theme of computer simulation is the devising of techniques to get around this problem, and we shall return to this later. Similar considerations apply to volume distributions in constant-pressure

ensembles, and indeed to other cases of thermodynamically conjugated pairs of variables.

Statistical mechanics may be used to derive practical microscopic formulae for thermodynamic quantities. A well-known example is the virial expression for the pressure, easily derived by scaling the atomic coordinates in the canonical ensemble partition function

$$PV = NkT - \frac{1}{3} \left\langle \sum_i \sum_{j \neq i} w_{ij} \right\rangle.$$

Here we assumed pairwise additivity $V = \sum_i \sum_{j \neq i} v_{ij}$, and defined $w(r) = r(dv(r)/dr)$. Also easily derived in the canonical ensemble is the general virial-like form, where q may be any coordinate or momentum,

$$\left\langle \mathcal{X} \frac{\partial \mathcal{H}}{\partial q} \right\rangle = kT \left\langle \frac{\partial \mathcal{X}}{\partial q} \right\rangle.$$

A well known example of this is obtained by setting $\chi = p_i \alpha$, $\alpha = x, y, z$, any component of momentum, giving the equipartition-of-energy relation

$$\left\langle \sum_{i\alpha} p_{i\alpha}^2 / 2m_i \right\rangle = \frac{3}{2} NkT.$$

This is commonly used to measure the temperature in a MD simulation. Less well known is the hypervirial relation

-11-

obtained by setting $\chi = -(\partial V / \partial r_{i\alpha}) = f_{i\alpha}$, a component of the force:

$$\left\langle \sum_{i\alpha} \left(\frac{\partial \mathcal{V}}{\partial r_{i\alpha}} \right)^2 \right\rangle = kT \left\langle \sum_{i\alpha} \left(\frac{\partial^2 \mathcal{V}}{\partial r_{i\alpha}^2} \right) \right\rangle$$

which relates the Laplacian of the potential with the mean-squared force. Butler et al [32] have suggested using this expression to measure the ‘configurational’ temperature (as a check) in MC simulations; they also provide a derivation of the corresponding expression in the microcanonical ensemble (see also [33]); it is surprising that this useful proposal has only been made so recently.

Finally, by considering increasing the number of particles by one in the canonical ensemble (looking at the excess, non-ideal, part), it is easy to derive the Widom [34] test-particle formula

$$\mu^{\text{ex}} \approx A_{N+1}^{\text{ex}} - A_N^{\text{ex}} = -kT \ln \langle e^{-\beta v_{\text{test}}} \rangle. \quad (\text{B3.3.5})$$

Here we have separated terms in the potential energy which involve the extra ‘test’ particle, $V_{N+1} = V_N + v_{\text{test}}$. The ensemble average here includes an unweighted average over inserted particle coordinates. In practice this means randomly inserting a test particle, many times, and averaging the Boltzmann factor of the associated energy change. More details of free energy calculations will be given later.

B3.3.2.6 TIME DEPENDENCE

A knowledge of time-dependent statistical mechanics is important in three general areas of simulation. First, in recent years there have been significant advances in the understanding of MD algorithms, which have arisen out of an appreciation of the formal operator approach to classical mechanics. Second, an understanding of equilibrium time correlation functions, their link with dynamical properties and especially their connection with transport coefficients, is essential in making contact with experiment. Third, the last decade has seen a rapid development of the use of nonequilibrium MD, with a better understanding of the formal aspects, particularly the link between the dynamical algorithm, dissipation, chaos and fractal geometry. Space does not permit a full description of this here: the interested reader should consult [35, 36] and references therein.

The Liouville equation dictates how the classical statistical mechanical distribution function $\varrho(\mathbf{r}^{(N)}, \mathbf{p}^{(N)}, t)$ evolves in time. (Also, quantum dynamics may be expressed in a formally equivalent way, but we shall concentrate exclusively on classical systems here.) From considerations of standard, Hamiltonian, mechanics [8] and the flow of representative systems in an ensemble through a particular region of phase space, it is easy to derive the Liouville equation

$$\frac{\partial \varrho}{\partial t} = - \left\{ \sum_{i\alpha} \dot{r}_{i\alpha} \frac{\partial}{\partial r_{i\alpha}} + \dot{p}_{i\alpha} \frac{\partial}{\partial p_{i\alpha}} \right\} \varrho \equiv -i\hat{L}\varrho$$

defining the Liouville operator \hat{L} . Compare this equation for ϱ with the time evolution equation for a dynamical variable $\chi(\mathbf{r}^{(N)}, \mathbf{p}^{(N)})$, which comes directly from the chain rule applied to Hamilton's equations

-12-

$$\dot{\chi} = \sum_{i\alpha} \dot{r}_{i\alpha} \frac{\partial \chi}{\partial r_{i\alpha}} + \dot{p}_{i\alpha} \frac{\partial \chi}{\partial p_{i\alpha}} \equiv i\hat{L}\chi.$$

The formal solutions of the time evolution equations are

$$\varrho(t) = e^{-i\hat{L}t} \varrho(0) \quad \text{and} \quad \chi(t) = e^{i\hat{L}t} \chi(0). \quad (\text{B3.3.6})$$

A number of manipulations are possible, once this formalism has been established. There are useful analogies both with the Eulerian and Lagrangian pictures of incompressible fluid flow, and with the Heisenberg and Schrödinger pictures of quantum mechanics [37, chapter 7], [38, chapter 11]. These analogies are particularly useful in formulating the equations of classical response theory [39], linking transport coefficients with both equilibrium and nonequilibrium simulations [35].

The Liouville equation applies to any ensemble, equilibrium or not. Equilibrium means that ϱ should be *stationary*, i.e., that

$$\partial \varrho / \partial t = 0.$$

In other words, if we look at any phase-space volume element, the rate of incoming state points should equal the rate of outflow. This requires that ϱ be a function of the constants of the motion, and especially $\varrho = \varrho(\mathcal{H})$. Equilibrium also implies $d\langle \chi \rangle / dt = 0$ for any χ . The extension of the above equations to nonequilibrium ensembles requires a consideration of entropy production, the method of controlling energy dissipation (thermostatting) and the consequent non-Liouville nature of the time evolution [35].

B3.3.3 MOLECULAR DYNAMICS

The solution of Newton's or Hamilton's equations on the computer

$$\dot{\mathbf{r}}_i = \mathbf{p}_i/m_i \quad \text{and} \quad \dot{\mathbf{p}}_i = \mathbf{f}_i$$

where m_i is the mass of atom i , and \mathbf{f}_i is the total force acting on it, is intrinsically a simple task. Many methods exist to perform step-by-step numerical integration of systems of coupled ordinary differential equations. Characteristics of these equations are: (a) they are 'stiff', i.e., there may be short and long time scales, and the algorithm must cope with both; (b) calculating the forces is expensive, typically involving a sum over pairs of atoms, and should be performed as infrequently as possible.

Also we must bear in mind that the advancement of the coordinates fulfils two functions: (i) accurate calculation of dynamical properties, especially over times as long as typical correlation times τ ; (ii) accurately staying on the constant-energy hypersurface, for much longer times t_{run} . Exact time reversibility is highly desirable (since the original equations

-13-

are exactly reversible). To ensure rapid sampling of phase space, we wish to make the time step as large as possible, consistent with these requirements. For these reasons, simulation algorithms have tended to be of *low order* (i.e., they do not involve storing high derivatives of positions, velocities etc): this allows the time step to be increased as much as possible without jeopardizing energy conservation. It is unrealistic to expect the numerical method to accurately follow the true trajectory for very long times t_{run} . The 'ergodic' and 'mixing' properties of classical trajectories, i.e., the fact that nearby trajectories diverge from each other exponentially quickly, make this impossible to achieve.

All these observations tend to favour the Verlet algorithm in one form or another, and we look closely at this in the following sections. For historical reasons only, we mention the more general class of predictor-corrector methods which have been optimized for classical mechanics simulations, [40, 41]; further details are available elsewhere [7, 42, 43].

B3.3.3.1 THE VERLET ALGORITHM

There are various, essentially equivalent, versions of the Verlet algorithm, including the original method employed by Verlet [13, 44] in his investigations of the properties of the Lennard-Jones fluid, and a 'leapfrog' form [45]. Here we concentrate on the 'velocity Verlet' algorithm [46], which may be written

$$\begin{aligned}\mathbf{r}_i(t + \delta t) &= \mathbf{r}_i(t) + \delta t \mathbf{p}_i(t)/m_i + \frac{1}{2} \delta t^2 \mathbf{f}_i(t)/m_i \\ \mathbf{p}_i(t + \delta t) &= \mathbf{p}_i(t) + \frac{1}{2} \delta t [\mathbf{f}_i(t) + \mathbf{f}_i(t + \delta t)].\end{aligned}$$

This advances the coordinates and momenta over a small time step δt . A piece of pseudo-code illustrates how this works:

```
call force(r,f)
do step = 1, nstep
  r = r + dt*p/m + (0.5*dt**2)*f/m
  p = p + 0.5*dt*f
  call force(r,f)
  p = p + 0.5*dt*f
enddo
```

The forces are calculated from the positions at the start of a simulation. They are used to advance the positions, and ‘half-advance’ the velocities or momenta. The new forces $\mathbf{f}(t+\delta t)$ are calculated, and these are used to complete the momentum update. At the end of the step, positions, momenta, and forces all conveniently refer to the same time point. Moreover, as we shall see shortly there is an interesting theoretical derivation of this version of the algorithm.

Important features of the Verlet algorithm are: (a) it is *exactly* time reversible; (b) it is *low* order in time, hence permitting long time steps; (c) it is easy to program.

-14-

B3.3.3.2 PROPAGATORS AND THE VERLET ALGORITHM

The velocity Verlet algorithm may be derived by considering a standard approximate decomposition of the Liouville operator which preserves reversibility and is *symplectic* (which implies that volume in phase space is conserved). This approach [47] has had several beneficial consequences.

The Liouville operator of equation b3.3.6 may be written [48]

$$e^{i\hat{L}t} = (e^{i\hat{L}\delta t})_{\text{approx}}^P + \mathcal{O}(P\delta t^3)$$

where $\delta t = t/P$ and an approximate propagator, correct at short time steps $\delta t \rightarrow 0$, appears in the parentheses. This is a formal way of stating what we do in MD, when we split a long time period t into a large number P of small time steps δt , using an *approximation* to the true equations of motion over each time step. It turns out that useful approximations arise from splitting \hat{L} into two parts

$$\hat{L} = \hat{L}_p + \hat{L}_r.$$

The following approximation

$$e^{i\hat{L}\delta t} = e^{(i\hat{L}_p + i\hat{L}_r)\delta t} \approx e^{i\hat{L}_p\delta t/2} e^{i\hat{L}_r\delta t} e^{i\hat{L}_p\delta t/2} \quad (\text{B3.3.7})$$

is asymptotically exact in the limit $\delta t \rightarrow 0$. For nonzero δt this is an approximation to $e^{i\hat{L}\delta t}$ because in general \hat{L}_p and \hat{L}_r do not commute, but it is still exactly time reversible. Tuckerman et al [47] set

$$i\hat{L}_p = \sum_{i\alpha} \dot{p}_{i\alpha} \frac{\partial}{\partial p_{i\alpha}} = \sum_{i\alpha} f_{i\alpha} \frac{\partial}{\partial p_{i\alpha}} \quad i\hat{L}_r = \sum_{i\alpha} \dot{r}_{i\alpha} \frac{\partial}{\partial r_{i\alpha}} = \sum_{i\alpha} (p_{i\alpha}/m) \frac{\partial}{\partial r_{i\alpha}}.$$

A straightforward derivation (not reproduced here) shows that the effect of the three successive steps embodied in equation (b3.3.7), with the above choice of operators, is precisely the velocity Verlet algorithm. This approach is particularly useful for generating multiple time-step methods.

B3.3.3.3 MULTIPLE TIME STEPS

An important extension of the MD method allows it to tackle systems with multiple time scales: for example, molecules which have very strong internal springs representing the bonds, while interacting externally through softer potentials, or molecules consisting of both heavy and light atoms. A simple MD algorithm will have to adopt a time step short enough to handle the fast-varying internal motions. Tuckerman et al [47] set

out methods for generating time-reversible Verlet-like algorithms using the Liouville operator formalism described above. Here we suppose that there are two types of force in the system: slow-moving external forces F_i and fast-moving internal forces f_i . The momentum satisfies $\dot{p} = f_i + F_i$. Then we break up the Liouville

-15-

operator $i\hat{L} = i\hat{L}_p + i\hat{\ell}$:

$$i\hat{L}_p = \sum_{i\alpha} F_{i\alpha} \frac{\partial}{\partial p_{i\alpha}} \quad i\hat{\ell} = \sum_{i\alpha} \dot{r}_{i\alpha} \frac{\partial}{\partial r_{i\alpha}} + f_{i\alpha} \frac{\partial}{\partial p_{i\alpha}} \equiv i\hat{\ell}_r + i\hat{\ell}_p.$$

The propagator approximately factorizes

$$e^{i\hat{L}\Delta t} \approx e^{i\hat{L}_p\Delta t/2} e^{i\hat{\ell}\Delta t} e^{i\hat{L}_p\Delta t/2}$$

where Δt represents a long time step. The middle part is then split again, using the conventional separation, and iterating over small time steps $\delta t = \Delta t/P$:

$$e^{i\hat{\ell}\Delta t} \approx [e^{i\frac{1}{2}\delta t\hat{\ell}_r} e^{i\delta t\hat{\ell}_p} e^{i\frac{1}{2}\delta t\hat{\ell}_r}]^P.$$

So the fast-varying forces must be computed many times at short intervals; the slow-varying forces are used just before and just after this stage, and they only need be calculated once per long time step.

This actually translates into a fairly simple algorithm, based closely on the standard velocity Verlet method. Written in a Fortran-like pseudo-code, it is as follows. At the start of the run we calculate both rapidly-varying (f) and slowly-varying (F) forces, then, in the main loop:

```

do STEP = 1, NSTEP
  p = p + 0.5*DT*F
  do step = 1, nstep
    r = r + dt*p/m + (0.5*dt**2)*f/m
    p = p + 0.5*dt*f
    call force(r,f)
    p = p + 0.5*dt*f
  enddo
  call FORCE(r,F)
  p = p + 0.5*DT*F
enddo

```

The entire simulation run consists of `NSTEP` long steps; each step consists of `nstep` shorter sub-steps. `DT` and `dt` are the corresponding time steps, `DT = nstep*dt`.

A particularly fruitful application, which has been incorporated into the computer code ORAC [49], is to split the interatomic force law into a succession of components covering different ranges: the short-range forces change rapidly with time and require a short time step, but advantage can be taken of the much slower time variation of the long-range forces, by using a longer time step and less frequent evaluation for these. Having said this, multiple time step algorithms are still under active study [50], and there is some concern that resonances may occur between the natural frequencies of the system under study, and the various time steps

used in schemes of this kind [51].

-16-

B3.3.3.4 CONSTRAINTS

Although, in principle, multiple time steps provide a method of integrating stiff degrees of freedom, such as intramolecular bonds, the alternative of rigidly constraining bond lengths is still very popular. In classical mechanics, constraints are introduced through the Lagrangian [8] or Hamiltonian [52] formalisms. Given a set of algebraic relations between atomic coordinates, for example, a fixed bond length b between atoms 1 and 2

$$\sigma(\mathbf{r}_1, \mathbf{r}_2) = (\mathbf{r}_1 - \mathbf{r}_2) \cdot (\mathbf{r}_1 - \mathbf{r}_2) - b^2 = 0$$

the constraint force between the atoms will have the following form

$$\mathbf{g}_1 = \lambda \frac{\partial \sigma}{\partial \mathbf{r}_1} \quad \text{and} \quad \mathbf{g}_2 = \lambda \frac{\partial \sigma}{\partial \mathbf{r}_2}$$

and it will appear in the equations of motion along with the normal forces. It is easy to derive an exact expression for the multiplier λ ; for many constraints, a system of equations (one per constraint) is obtained. In practice, since the equations of motion are only solved approximately, the constraints will be increasingly violated as the simulation proceeds. The breakthrough in this area came with the proposal of a scheme, SHAKE, to solve the equations for the constraint forces *approximately* (i.e., to the same level of approximation as the dynamical algorithm) in such a way that the constraints are satisfied exactly at the end of each time step [53, 54]; for a review see [55]. The appropriate version of this scheme for the velocity Verlet algorithm is called RATTLE [56].

It is important to realize that a simulation of a system with rigidly constrained bond lengths is not equivalent to a simulation with, for example, harmonic springs representing the bonds, even within the limit of very strong springs. One obvious point is that the momenta conjugated to the bond coordinates are nonzero and store some kinetic energy in the spring case, while they are zero by definition in the constrained case. A subtle, but crucial, consequence of this is that it has an effect on the distribution function for the other coordinates. If we obtain the configurational distribution function by integrating over the momenta, the difference arises because in one case a set of momenta is set to zero, and not integrated, while in the other an integration is performed, which may lead to an extra term depending on particle coordinates. This is frequently called the ‘metric tensor problem’; it is explained in more detail in [7, 57], and there are well-established ways of determining when the difference is likely to be significant [58] and how to handle it, if necessary [59].

Some people prefer to use the multiple time step approach to handle fast degrees of freedom, while others prefer to use constraints, and there are situations in which both techniques are applicable. Constraints also find an application in the study of rare events, where a system may be studied at the top of a free energy barrier (see later), or for convenience when it is desired to fix a thermodynamic order parameter or ordering direction [17].

B3.3.3.5 NEIGHBOUR LISTS

In the inner loops of MD and MC programs, we consider an atom i and loop over all atoms j to calculate the minimum image separations. If $r_{ij} > r_c$, the potential cutoff, the program skips to the end of the inner loop, avoiding expensive

calculations, and considers the next neighbour. In this method, the time to examine all pair separations is proportional to N^2 ; for every pair, one must compute at least r_{ij}^2 ; this still consumes a lot of time. Some economies result from the use of lists of nearby pairs of atoms.

Verlet [13] suggested a technique for improving the speed of a program by maintaining a list of neighbours. The potential cutoff sphere, of radius r_c , around a particular atom is surrounded by a 'skin', to give a larger sphere of radius r_{list} , as shown in figure B3.3.6. At the first step in a simulation, a list is constructed of all the neighbours of each atom, for which the pair separation is within r_{list} . Over the next few MD time steps, only pairs appearing in the list are checked in the force routine. From time to time the list is reconstructed: it is important to do this before any unlisted pairs have crossed the safety zone and come within interaction range. It is possible to trigger the list reconstruction automatically, if a record is kept of the distance travelled by each atom since the last update. The choice of list cutoff distance r_{list} is a compromise: larger lists will need to be reconstructed less frequently, but will not give as much of a saving on cpu time as smaller lists. This choice can easily be made by experimentation.

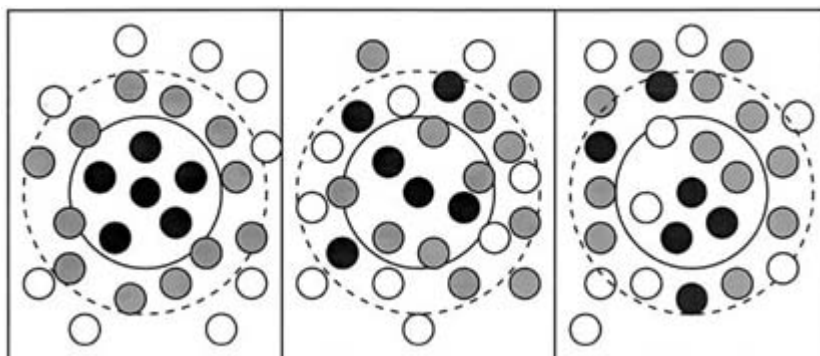


Figure B3.3.6. The Verlet list on its construction, later, and too late. The potential cutoff range, and the list range, are indicated. The list must be reconstructed before particles originally outside the list range have penetrated the potential cutoff sphere.

For larger systems ($N \geq 1000$ or so, depending on the potential range) another technique becomes preferable. The cubic simulation box (extension to noncubic cases is possible) is divided into a regular lattice of $n_c \times n_c \times n_c$ cells; see figure B3.3.7. These cells are chosen so that the side of the cell $r_{cell} = L/n_c$ is greater than the potential cutoff distance r_c . If there is a separate list of atoms in each of those cells, then searching through the neighbours is a rapid process: it is only necessary to look at atoms in the same cell as the atom of interest, and in nearest neighbour cells. The cell structure may be set up and used by the method of linked lists [45, 60]. The first part of the method involves sorting all the atoms into their appropriate cells. This sorting is rapid, and may be performed at every step. Then, within the force routine, pointers are used to scan through the contents of cells, and calculate pair forces. This approach is very efficient for large systems with short-range forces. A certain amount of unnecessary work is done because the search region is cubic, not (as for the Verlet list) spherical.

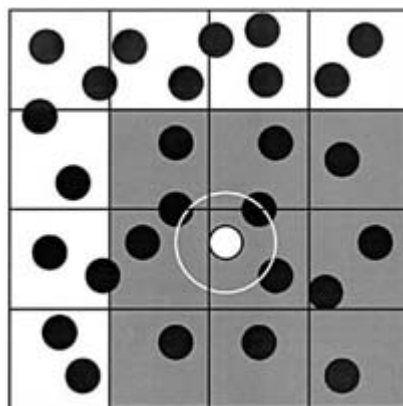


Figure B3.3.7. The cell structure. The potential cutoff range is indicated. In searching for neighbours of an atom, it is only necessary to examine the atom's own cell, and its nearest-neighbour cells.

B3.3.3.6 LONG-RANGE FORCES

Many realistic simulations will involve the Coulomb interaction between charges, which decreases with separation as r^{-1} , and the dipole–dipole interaction, which decreases as r^{-3} . These cannot be treated simply by applying a spherical cutoff: it is essential to consider the effects of the surrounding medium. Two somewhat different techniques have been used in the majority of computer simulations to handle long-range forces: the reaction field method and the Ewald sum.

In the reaction field method, the space surrounding a dipolar molecule is divided into two regions: (i) a cavity, within which electrostatic interactions are summed explicitly, and (ii) a surrounding medium, which is assumed to act like a smooth continuum, and is assigned a dielectric constant ϵ_r . Ideally, this quantity will be equal to the dielectric constant ϵ of the liquid itself, but calculating this, of course, is frequently one of the goals of the simulation, not one of the input parameters. The essence of the reaction field method is to calculate the total dipole moment of the molecules in the cavity, hence obtaining the polarization of the surrounding continuum, and to use this to work out the reaction field on the molecule at the centre. This supplements the direct electrostatic interaction with molecules in the cavity, in the calculation of the total energy. The reaction field method was used by Barker and Watts [61] in early simulations of water and has been discussed by Neumann and Steinhauser [62] and Patey et al [63].

In the Ewald method, a lattice sum is performed over charges within the periodically repeating simulation box. This is a subtle matter, since the sum, for Coulomb potentials, is only conditionally convergent: the result depends on the order of terms. Nonetheless, the procedure has been carefully analysed by de Leeuw et al [64] and Felderhof [65]. To make the summation a practical proposition, a trick is used: each point charge is screened by a surrounding, Gaussian, charge distribution, which makes the interactions short-ranged: these interactions are tackled in real space in the usual way. The contribution of an equal and opposite set of Gaussians is tackled in reciprocal space, using Fourier transforms. The choice of the width of the Gaussians is a parameter which may be varied to optimize the speed of calculation (for a given accuracy): Perram et al [66] have shown that the optimal choice leads to an algorithm whose expense grows as $N^{3/2}$.

When carried out properly, the results of the reaction field method and the Ewald sum are consistent [67]. Recently, the reaction field method has been recommended on grounds of efficiency and ease of programming [68, 69]. The

expense of the Ewald method, particularly as the system size grows, has led to the search for alternative formulations [see, e.g., 70]. Recently, the practical implementation of the Ewald method has been significantly improved. The smooth particle mesh Ewald method, inspired by the approach of [45], employs a mesh and an interpolation scheme to allow evaluation of the reciprocal space sums using fast Fourier transforms [71, 72]. This approach has been incorporated into a standard code [49] and seems very promising for large biomolecular systems. It has to be said that there are still some subtleties involved in the handling of long-range forces [see, e.g., 73] and the reader should consult carefully the references if approaching the simulation of such systems from the beginning.

B3.3.4 MONTE CARLO

It is important to realize that MC simulation does not provide a way of calculating the statistical mechanical partition function: instead, it is a method of sampling configurations from a given statistical ensemble and hence of calculating ensemble averages. A complete sum over states would be impossibly time consuming for systems consisting of more than a few atoms. Applying the trapezoidal rule, for instance, to the configurational part of Q_{NVT} , entails discretizing each atomic coordinate on a fine grid; then the dimensionality of the integral is extremely high, since there are $3N$ such coordinates, so the total number of grid points is astronomically high. The MC integration method is sometimes used to estimate multidimensional integrals by randomly sampling points. This is not feasible here, since a very small proportion of all points would be sampled in a reasonable time, and very few, if any, of these would have a large enough Boltzmann factor to contribute significantly to the partition function. MC simulation differs from such methods by sampling points in a nonuniform way, chosen to favour the important contributions.

B3.3.4.1 IMPORTANCE SAMPLING

MC simulation is a method of concentrating the sampled points in the important regions, namely the regions with high Boltzmann factor $e^{-\beta\mathcal{H}}$: a random walk is devised, moving from one point to the next, with a biasing probability chosen to generate the desired distribution. Unfortunately, a consequence of this approach is that it is no longer possible to estimate the partition function itself, merely *ratios* of sums over states, that is, ensemble averages. Suppose that we have succeeded in selecting states γ with probability proportional to $\varrho(\Gamma) = \exp\{-\beta\mathcal{H}(\Gamma)\}$. Then, if we have conducted N_t ‘observations’ or ‘steps’ in the process, the ensemble average becomes an average over steps

$$\langle \mathcal{X} \rangle_{NVT} = \sum_{\Gamma} \varrho(\Gamma) \mathcal{X}(\Gamma) = \frac{1}{N_t} \sum_{i=1}^{N_t} \mathcal{X}_i.$$

The Boltzmann weight appears *implicitly* in the way the states are chosen. The form of the above equation is like a time average as calculated in MD. The MC method involves designing a stochastic algorithm for stepping from one state of the system to the next, generating a trajectory. This will take the form of a *Markov chain*, specified by transition probabilities which are independent of the prior history of the system.

Write $\varrho(\Gamma) \equiv \varrho_{\Gamma}$ treating it as a component of a (very large) column vector. Consider an *ensemble* of systems all evolving at once. Specify a matrix whose elements $\pi_{\Gamma\leftarrow\gamma}$ give the probability of going to state Γ from state γ , for

every pair of states. The matrix must satisfy $\sum_{\Gamma'} \pi_{\Gamma' \leftarrow \Gamma} = 1$, to conserve probability. At each step, implement jumps with this transition matrix. This generates a *Markov chain* of states, i.e., one in which the transition probabilities do not depend on the history. Feller's theorem [74] tells us that, subject to some reasonable conditions, there exists a limiting (equilibrium) distribution of states and the system will tend towards this limiting distribution. (Recently [75], it has been shown that the Markov condition can be relaxed, and the system will still behave in this way.) A little thought shows that the limiting distribution will satisfy

$$Q_{\Gamma'} = \sum_{\Gamma} \pi_{\Gamma' \leftarrow \Gamma} Q_{\Gamma}$$

which is a matrix eigenvalue equation. The eigenvector is already known: it is the Boltzmann distribution. The MC method is specified by choosing a transition matrix which satisfies this equation. One way of guaranteeing this is to ensure that

$$\pi_{\Gamma \leftarrow \Gamma'} Q_{\Gamma'} = \pi_{\Gamma' \leftarrow \Gamma} Q_{\Gamma}$$

which is usually termed the *microscopic reversibility* condition. An immediate consequence of this is that the ratio of probabilities $Q_{\Gamma' \leftarrow \Gamma} / Q_{\Gamma \leftarrow \Gamma'}$ is equal to the ratio of transition matrix elements $\pi_{\Gamma' \leftarrow \Gamma} / \pi_{\Gamma \leftarrow \Gamma'}$. This relationship is analogous to that relating the equilibrium constant for a chemical reaction to the ratio of forward and backward rate constants.

The most commonly used prescription [76] is

$$\begin{aligned} \pi_{\Gamma' \leftarrow \Gamma} &= \alpha_{\Gamma' \leftarrow \Gamma} \min(1, Q_{\Gamma'} / Q_{\Gamma}) & \Gamma \neq \Gamma' \\ \pi_{\Gamma \leftarrow \Gamma} &= 1 - \sum_{\Gamma' \neq \Gamma} \pi_{\Gamma' \leftarrow \Gamma} & \text{otherwise.} \end{aligned}$$

Here, an underlying matrix, with elements $\alpha_{\Gamma' \leftarrow \Gamma}$, sets the probability of *attempting* a move like $\Gamma' \leftarrow \Gamma$, and the other factor gives the probability of *accepting* such a move. This scheme only requires a knowledge of the ratio $Q_{\Gamma' \leftarrow \Gamma} / Q_{\Gamma \leftarrow \Gamma'}$:

$$\min(1, Q_{\Gamma'} / Q_{\Gamma}) = \min(1, e^{-\beta(\mathcal{H}(\Gamma') - \mathcal{H}(\Gamma))}) = \min(1, e^{-\beta\delta\mathcal{H}}).$$

It does not require knowledge of the factor normalizing the Q , i.e., the partition function. For atomic and molecular systems, the partition function is split into a product of 'ideal' (exactly calculable) and 'excess' terms: the position and momentum distributions also factorize, and we wish to sample

$$Q_{NVT}(r) \propto \exp\{-\beta\mathcal{V}(r)\}.$$

The prescription for accepting or rejecting moves is exactly as written before, but with \mathcal{V} replacing \mathcal{H} . Assuming that the interaction potential is short-ranged, it is not necessary to perform a complete recalculation of \mathcal{V} every time an atom is moved: just the part involving that atom. For a given trial move, this is done twice: once before the attempted move and once after. Some improvement in efficiency may be obtained by using neighbour lists, as described earlier for MD.

Selecting trial moves in an unbiased way typically means (a) choose an atom 'randomly', with equal probability from the complete set; (b) displace it by random amounts in the x , y and z directions, chosen

independently and uniformly from a predefined range (symmetric about the origin). These choices use a random number generator; the quality of the random numbers may be an issue. A key aspect of move selection is that the probabilities for *attempting* forward and reverse moves must be equal, so $\alpha_{\Gamma \square \leftarrow \Gamma} = \alpha_{\Gamma \leftarrow \Gamma \square}$. For the above prescription, it should be evident that this is true, by considering the number of ways of selecting an atom, and the number of positions it might be moved to (assuming a fine discretization of space). In the case of a rigid molecule, move selection will include a procedure for randomly rotating a molecule in an unbiased way. The magnitudes of trial moves are parameters of the method, chosen to give a reasonable acceptance rate, traditionally 50% or so. There is no special reason for this value. Ideally, for every study, one would investigate which choice gives the most efficient sampling of phase space.

The analysis of Manousiouthakis and Deem [75] mentioned above has demonstrated that it is also correct to choose atoms sequentially rather than randomly: it has been tacitly assumed for many years that this violation of the Markovian restriction is acceptable, so a proof of this kind is very welcome.

B3.3.4.2 WEIGHTED AND BIASED SAMPLING

It is useful to write down here the basic formulae for sampling with an additional weight function applied, sometimes called non-Boltzmann or umbrella sampling, and for sampling when the selection of trial moves is done in a biased way, i.e., the α matrix is not symmetrical.

A weight factor $\mathcal{W}(\Gamma)$ may be introduced in the MC sampling algorithm, to generate a modified, or ‘weighted’, distribution,

$$q_{\mathcal{W}} \propto \mathcal{W}(\Gamma) \exp\{-\beta\mathcal{H}(\Gamma)\}.$$

The usual MC procedure is adopted, with trial moves $\Gamma' \leftarrow \Gamma$ selected as usual, but now accepted with probability

$$\min\left(1, \frac{\mathcal{W}(\Gamma')}{\mathcal{W}(\Gamma)} e^{-\beta\delta\mathcal{H}}\right).$$

In the calculation of ensemble averages, we correct for the weighting as follows

$$\langle \mathcal{X} \rangle = \frac{\langle \mathcal{X}/\mathcal{W} \rangle_{\mathcal{W}}}{\langle 1/\mathcal{W} \rangle_{\mathcal{W}}}$$

where $\langle \dots \rangle_{\mathcal{W}}$ represents the weighted simulation averages. This kind of sampling may be useful when the most important states for our purposes are not those which have the highest weights in the canonical ensemble: for example, when we wish to compute a free energy difference between two states. We will return to this later.

Biased move selection means that $\alpha_{\Gamma \square \leftarrow \Gamma} \neq \alpha_{\Gamma \leftarrow \Gamma \square}$. Suppose that we wish, nonetheless, to sample the canonical distribution. To do this, we need to calculate the ratio $\alpha_{\Gamma \square \leftarrow \Gamma} / \alpha_{\Gamma \leftarrow \Gamma \square}$ as well as $q_{\Gamma \square} / q_{\Gamma} = e^{-\beta\delta\mathcal{H}}$. Then we accept the move with probability

$$\min \left(1, \frac{\alpha_{\Gamma \leftarrow \Gamma'}}{\alpha_{\Gamma' \leftarrow \Gamma}} e^{-\beta \delta H} \right).$$

A consideration of the transition probabilities allows us to prove that microscopic reversibility holds, and that canonical ensemble averages are generated. This approach has greatly extended the range of simulations that can be performed. An early example was the preferential sampling of molecules near solutes [77], but more recently, as we shall see, polymer simulations have been greatly accelerated by this method.

B3.3.5 SIMULATION IN DIFFERENT ENSEMBLES

It is very convenient to be able to choose a nonstandard ensemble for a simulation. Generally, it is more straightforward to do this in MC, but MD techniques for various ensembles have been developed. We consider MC implementations first.

B3.3.5.1 MC IN DIFFERENT ENSEMBLES

The isothermal–isobaric ensemble corresponds to a system whose volume and energy can fluctuate, in exchange with its surroundings at specified NPT . The thermodynamic driving force is the Gibbs free energy $G = A + PV$. The configurational distribution function may be written

$$Q_{NPT}(\mathbf{s}^{(N)}, V) \propto V^N e^{-\beta PV} e^{-\beta \mathcal{V}}.$$

Here we have introduced scaled coordinates $\mathbf{s}^{(N)} = L^{-1} \mathbf{r}^{(N)}$ where L is the box length (assumed cubic).

This ensemble is a weighted superposition of NVT ensembles for different volumes. A typical MC sweep consists of N attempted single-particle moves, exactly as for constant- NVT MC, followed by one attempt to scale, homogeneously, the volume of the simulation box, together with the coordinates of all the particles in it. This is accepted or rejected so as to generate the above distribution. One prescription for selecting the volume move is to attempt to change $V \rightarrow V' = V + \delta V$ where δV is uniformly sampled from an interval $[-\delta V_{\max}, \delta V_{\max}]$. The new box length is computed, and all the particle coordinates scaled by an appropriate factor; then the new potential energy is computed. Assuming that the selection of δV is unbiased, the probability ratio to use in the Metropolis prescription is just the ratio of the two ensemble densities, $Q_{NPT}(V')/Q_{NPT}(V)$ and the move is accepted with probability

$$\min(1, e^{-\beta \delta \mathcal{W}}) \quad \text{where} \quad \delta \mathcal{W} = \delta \mathcal{V} + P \delta V - NkT \ln(V'/V).$$

Here $\delta \mathcal{V} = \mathcal{V}' - \mathcal{V}$ and $\delta V = V' - V$. The maximum attempted volume change is chosen to give a reasonable acceptance rate, traditionally 35–50% or so; there is no firm reason for this choice.

The above prescription for selecting volume changes is not unique. It may seem more natural to make random,

uniform, changes in the box length L ; some people prefer to sample V uniformly [78]. These choices are *not*

precisely consistent with the accept/reject procedure described above. They can be regarded as unbiased sampling of a variable different from V (in which case a simple transformation of variables is needed to convert Q_{NPT} into the new form, and additional powers of V will appear in it) or as biased sampling in V -space, in which case a small correction factor (the same extra powers of V) will appear in the accept/reject procedure. An analysis of the forward and backward transition probabilities will give the appropriate acceptance/rejection criterion in each case.

The grand canonical ensemble corresponds to a system whose number of particles and energy can fluctuate, in exchange with its surroundings at specified μ VT. The relevant thermodynamic quantity is the grand potential $\Omega = A - \mu N$. The configurational distribution is conveniently written

$$Q_{\mu VT}(s^{(N)}, N) \propto (N!)^{-1} V^N z^N \exp\{-\beta\mathcal{V}\}.$$

Here again we have introduced scaled coordinates $s^{(N)} = L^{-1}r^{(N)}$ where L is the box length (assumed cubic), $z = \exp\{\beta\mu\}/\Lambda^3$ is the activity, and $\Lambda = h/\Lambda = h/\sqrt{2\pi mkT}$ is the thermal de Broglie wavelength.

This ensemble is a weighted superposition of NVT ensembles with different values, of N . As a rule of thumb, a typical MC sweep consists of N attempted moves, each of which is chosen randomly to be (i) a displacement (handled exactly as in constant- NVT MC); (ii) the creation of a new particle at a randomly selected position; (iii) the destruction of a randomly selected particle from the system. The probabilities for attempting creation and destruction must be equal (for consistency with what follows), but they need not be equal to the probability for attempting displacement (although they often are).

For a creation attempt, a position is chosen uniformly at random within the box, and an attempt made to create a new particle there. The probability ratio for creation is:

$$\frac{Q_{\mu VT}(N+1)}{Q_{\mu VT}(N)} = \frac{zV}{N+1} \exp\{-\beta\delta\mathcal{V}\} \equiv \exp\{-\beta\delta\mathcal{Z}^{\text{create}}\}$$

where $\delta\mathcal{V} = \mathcal{V}' - \mathcal{V}$ is the potential energy change associated with inserting the new particle. In a Metropolis scheme, the creation attempt is accepted with probability $\min(1, \exp\{-\beta\delta\mathcal{Z}^{\text{create}}\})$.

For a destruction attempt, one of the existing N particles is selected at random, and an attempt made to destroy it. The probability ratio to use is

$$\frac{Q_{\mu VT}(N-1)}{Q_{\mu VT}(N)} = \frac{N}{zV} \exp\{-\beta\delta\mathcal{V}\} \equiv \exp\{-\beta\delta\mathcal{Z}^{\text{destroy}}\}$$

where $\delta\mathcal{V}$ is the potential energy change associated with removing the particle. In a Metropolis scheme, the destruction attempt is accepted with probability $\min(1, \exp\{-\beta\delta\mathcal{Z}^{\text{destroy}}\})$. These expressions can be shown to satisfy microscopic reversibility [57].

In a dense system, the acceptance rate of particle creation and deletion moves will decrease, and the number of attempts must be correspondingly increased: eventually, there will come a point at which grand canonical simulations are not practicable, without some tricks to enhance the sampling.

B3.3.5.2 MD IN DIFFERENT ENSEMBLES

In this section we discuss MD methods in the constant- NVT ensemble, and the constant- NPT ensemble.

There are three general approaches to conducting MD at constant temperature rather than constant energy. One method, simple to implement and reliable, is to periodically reselect atomic velocities at random from the Maxwell–Boltzmann distribution [79]. This is rather like an occasional random coupling with a thermal bath. The resampling may be done to individual atoms, or to the entire system; some guidance on the reselection frequency may be found in [79].

A second approach, due originally to Nos'e [80] and reformulated in a useful way by Hoover [81], is to introduce an extra ‘thermal reservoir’ variable into the dynamical equations:

$$\begin{aligned}\dot{\mathbf{r}}_i &= \mathbf{p}_i/m & \dot{\mathbf{p}}_i &= \mathbf{f}_i - \zeta \mathbf{p}_i \\ \dot{\zeta} &= \frac{\sum_{i\alpha} p_{i\alpha}^2/m - gkT}{Q} \equiv v_T^2 \left[\frac{\sum_{i\alpha} p_{i\alpha}^2/m}{gkT} - 1 \right] = v_T^2 \left[\frac{T}{T} - 1 \right].\end{aligned}$$

Here ζ is a friction coefficient which is allowed to vary in time; Q is a thermal inertia parameter, which may be replaced by v_T , a relaxation rate for thermal fluctuations; $g \approx 3N$ is the number of degrees of freedom. T stands for the instantaneous ‘mechanical’ temperature. It may be shown that the distribution function for the ensemble is proportional to $\exp\{-\beta\mathcal{H}'\}$ where $\mathcal{H}' = \mathcal{H} + \frac{1}{2}3NkT\zeta^2/v_T^2$. These equations lead to the following time variation of the system energy $\mathcal{H} = \sum_{i\alpha} p_{i\alpha}^2/2m + \mathcal{V}$, and for the variable \mathcal{H}' :

$$\dot{\mathcal{H}} = \sum_{i\alpha} p_{i\alpha} \dot{p}_{i\alpha}/m - \sum_{i\alpha} f_{i\alpha} \dot{r}_{i\alpha} = -\zeta \sum_{i\alpha} p_{i\alpha}^2/m \quad \dot{\mathcal{H}}' = -3NkT\dot{\zeta}.$$

If $T > T$, i.e., the system is too hot, then the ‘friction coefficient’ ζ will tend to increase; when it is positive the system will begin to cool down. If the system is too cold, the reverse happens, and the friction coefficient may become negative, tending to heat the system up again. In some circumstances, this approach generates non-ergodic behaviour, but this may be ameliorated by the use of chains of thermostat variables [82]. Tobias et al [83] give an example of the use of this scheme in a biomolecular simulation.

As an alternative to sampling the canonical distribution, it is possible to devise equations of motion for which the ‘mechanical’ temperature is constrained to a constant value [84, 85, 86]. The equations of motion are

$$\dot{\mathbf{r}}_i = \mathbf{p}_i/m \quad \dot{\mathbf{p}}_i = \mathbf{f}_i - \zeta \mathbf{p}_i \quad \zeta = \frac{\sum_{i\alpha} f_{i\alpha} p_{i\alpha}}{\sum_{i\alpha} p_{i\alpha}^2}.$$

-25-

Here the friction coefficient ζ is completely determined by the instantaneous values of the coordinates and momenta. It is easy to see that the kinetic energy $\mathcal{K} = \sum_{i\alpha} p_{i\alpha}^2/2m$ is now a constant of the motion:

$$\dot{\mathcal{K}} = \sum_{i\alpha} p_{i\alpha} \dot{p}_{i\alpha}/m_i = \sum_{i\alpha} p_{i\alpha} f_{i\alpha}/m_i - \zeta \sum_{i\alpha} p_{i\alpha}^2/m_i = 0.$$

It is possible to devise extended-system methods [79, 87] and constrained-system methods [88] to simulate the constant- NPT ensemble using MD. The general methodology is similar to that employed for constant-

NVT , and in the course of the simulation the volume V of the simulation box is allowed to vary, according to the new equations of motion. A useful variant allows the simulation box to change shape as well as size [89, 90]. It is also possible to extend the Liouville operator-splitting approach to generate algorithms for MD in these ensembles; examples of explicit, reversible, integrators are given by Martyna et al [91].

B3.3.6 FREE ENERGIES, CHEMICAL POTENTIALS AND WEIGHTED SAMPLING

A major drawback of MD and MC techniques is that they calculate *average* properties. The free energy and entropy functions cannot be expressed as simple averages of functions of the state point γ . They are directly connected to the logarithm of the partition function, and our methods do not give us the partition function itself. Nonetheless, calculating free energies is important, especially when we wish to determine the relative thermodynamic stability of different phases. How can we approach this problem?

B3.3.6.1 FREE ENERGY DIFFERENCES

It is possible to calculate *derivatives* of the free energy directly in a simulation, and thereby determine free energy differences by thermodynamic integration over a range of state points between the state of interest and one for which we know A exactly (the ideal gas, or harmonic crystal for example):

$$(\beta A)_2 - (\beta A)_1 = \int_{\beta_1}^{\beta_2} E \, d\beta \quad \text{or} \quad A_2 - A_1 = - \int_{V_1}^{V_2} P \, dV.$$

This is reliable and fairly accurate, if tedious. It was used, for example, by Hoover [92] to locate the melting parameters for soft-sphere systems. The only point to watch out for is that one should not cross any phase transitions in taking the path from 1 to 2: it must be *reversible*.

-26-

A free energy difference between two systems with energy functions \mathcal{H}_0 and \mathcal{H}_1 , respectively, may be written in a way analogous to [equation b3.3.3](#)

$$A_1 = A_0 - kT \ln \langle \exp\{-\beta \Delta \mathcal{H}\} \rangle_0$$

where $\langle \dots \rangle_0$ is an ensemble average for system 0 and $\Delta \mathcal{H} = \mathcal{H}_1 - \mathcal{H}_0$. In the extreme case $\mathcal{H}_0 \equiv \mathcal{H}_1$, this would give an unweighted estimate of the partition function of system 1, but would be extremely poorly sampled for the reasons discussed in [section b3.3.2](#). Reasonable estimates will result when the two systems do not differ ‘too much’, so that contributing values of the Boltzmann factor are given a significant weight by the sampling over states in system 0. One famous example of this is the test-particle insertion formula, [equation \(B3.3.5\)](#), for estimating the chemical potential, where system 1 contains an additional particle. Another example is where a molecule in the system can be mutated into another. The efficiency of the sampling may depend critically on the direction of the perturbation change: estimating μ^{ex} by particle *removal*, for instance, is formally possible but usually less accurate than particle insertion. Kofke and Cummings [93] have reviewed various approaches in this field and make the general recommendation that the change should take place in the direction of *decreasing entropy*.

B3.3.6.2 HISTOGRAM RE-WEIGHTING

A way of looking at the points raised in the previous section is to compare energy *distributions* in two systems whose free energies we wish to relate. In particular, consider measuring, in a simulation of system 0, the function $\mathcal{P}_0(\Delta E)$, i.e., the probability density per unit ΔE of configurations for which \mathcal{H}_0 and \mathcal{H}_1 differ by the prescribed amount ΔE . The distribution $\mathcal{P}_1(\Delta E)$ may be similarly calculated by simulating system 1. These two functions may be straightforwardly related [94]:

$$\ln \mathcal{P}_1(\Delta E) = \ln \mathcal{P}_0(\Delta E) + \beta \Delta A - \beta \Delta E.$$

Therefore, apart from an unknown constant $\beta \Delta A$, and a known linear term $\beta \Delta E$, these are the same function. Bennett [94] suggested two graphical methods for determining $\beta \Delta A$ from $\mathcal{P}_0(\Delta E)$ and $\mathcal{P}_1(\Delta E)$, which rely on the two distributions, at worst, nearly overlapping (i.e., being measurable, with good statistics, for the same or similar values of ΔE). To broaden the sampling into the wings of the distribution, thereby improving statistics and extending the overlap region, we may use weighted sampling as described in section b3.3.4.2. There are many related approaches, variously called umbrella [95], multicanonical [96] and entropic [97] sampling, simulated tempering [98] and expanded ensembles [99].

Windowing is a special case of umbrella sampling: the weight function is a constant inside a specified region of configuration space, and zero outside. In MC we simply reject moves which would take the system outside the window, and otherwise proceed as usual. This allows us to examine a distribution function, and hence a free energy curve, piece by piece, matching up the resulting curves afterwards. The way to do this combination of histograms has been discussed by Ferrenberg and Swendsen [100], and the statistical errors in histogram re-weighting have been discussed by Swendsen [101] and Ferrenberg et al [102]. Ultimately, this approach leads back to the idea of performing simulations in a (nearly)-*microcanonical* ensemble and relating the results at nearby energies, as we do in thermodynamic integration; as emphasized in the review of Dönweg [103], ‘for any thermodynamic integration procedure there is an equivalent multistage sampling or histogram procedure, and *vice versa*’.

B3.3.6.3 THE CHEMICAL POTENTIAL

The ensemble average in the Widom formula, $\langle\langle \exp\{-\beta v_{\text{test}}\} \rangle\rangle$, is sometimes loosely referred to as the ‘insertion probability’. It becomes very low for dense fluids. For example, for hard spheres, we can use the scaled-particle theory [104] or the Carnahan–Starling equation of state [105] to estimate it (see figure B3.3.8). The insertion probability falls below 10^{-4} , well before the freezing transition at $\eta \approx 0.49$. Similar estimates can be made for the Lennard-Jones fluid. The lower this factor becomes, the poorer the statistics, and the more unreliable will be the estimate of μ^{ex} . The problem is particularly acute for dense molecular fluids where, as a first guess, one could take the overall Boltzmann factor to be the *product* of the individual atomic values.

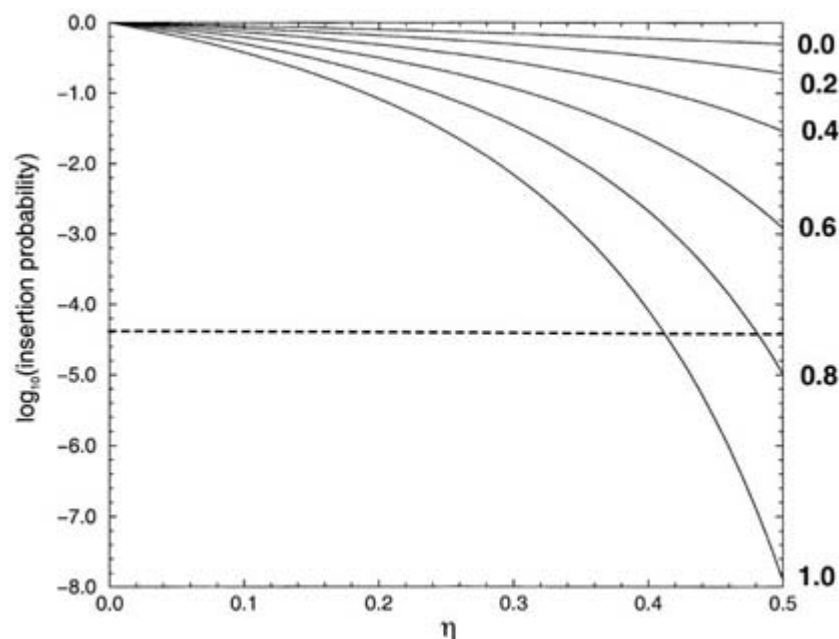


Figure B3.3.8. Insertion probability for hard spheres of various diameters (indicated on the right) in the hard sphere fluid, as a function of packing fraction η , predicted using scaled particle theory. The dashed line is a guide to the lowest acceptable value for chemical potential estimation by the simple Widom method.

A simple method of improving the efficiency of test particle insertion [106, 107, 108 and 109] involves dividing the simulation box into small cubic regions, and identifying those which would make a negligible contribution to the Widom formula, due to overlap with one or more atoms. These cubes are excluded from the sampling, and a correction applied afterwards for the consequent bias.

Another trick is applicable to, say, a two-component mixture, in which one of the species, A, is smaller than the other, B. From figure B3.3.8 for hard spheres, we can see that A need not be *particularly* small in order for the test particle insertion probability to climb to acceptable levels, even when insertion of B would almost always fail. In these circumstances, the chemical potential of A may be determined directly, while that of B is evaluated indirectly, relative to that of A. The related ‘semi-grand’ ensemble has been discussed in some detail by Kofke and Glandt [110].

This naturally leads to the idea of estimating μ^{ex} by gradual insertion [111, 112 and 113]. This can be thought of as a thermodynamic integration pathway, connecting the states with N and $N + 1$ particles via a set of intermediate points, characterized by a parameter λ , $0 \leq \lambda \leq 1$ which determines the degree to which the extra ‘ λ -particle’ is ‘switched on’. A MC scheme is constructed, which (in addition to the usual moves) allows λ to vary either continuously or in predefined discrete jumps. Then the chemical potential is expressed $\mu_{\text{ex}} = \Delta A_{\text{ex}} = A_{\text{ex}}(\lambda = 1) - A_{\text{ex}}(\lambda = 0)$. It is advantageous to apply a weighting function (see section b3.3.4.2) [112, 113], to ensure more or less uniform sampling of the different λ states. Consider the probability histogram $\mathcal{P}(\lambda)$ of the sampled values of λ during the runs. Without an external biasing potential this will be directly related to a Landau free energy

$$\mathcal{P}(\lambda) \propto \exp\{-\beta\mathcal{F}(\lambda)\}$$

where $\mathcal{F}(1) - \mathcal{F}(0) = \Delta A_{\text{ex}}$ is the desired free energy difference; to obtain a uniform distribution, a weight function $\mathcal{W}(\Gamma) \propto \exp\{-\beta\Psi(\lambda)\}$ with a biasing potential $\Psi(\lambda) = -\mathcal{F}(\lambda)$ would be used. This ideal weighting function is not known at the start of the simulation, but an initial guess may be iteratively refined from the measured $\mathcal{P}(\lambda)$ in a series of runs.

It is also advantageous to ensure that the λ -particle samples a wide range of positions in the fluid: this is achieved by attempting large-scale moves to new, randomly-selected positions from time to time, and also frequently attempting exchanges of position with a randomly-selected full-size particle. The former moves will have a high probability of success when λ is small, and the latter when λ is large. Camp et al [114] provide an example of this method in action, for a model of liquid crystals.

B3.3.6.4 FREE ENERGY OF SOLIDS

Early attempts to calculate solid-state free energies, and hence locate the melting transition, introduced the idea of conducting a thermodynamic integration along an artificial pathway [115, 116]. Each atom is artificially restricted to a single cell in space so that, as the density is lowered, the system converts more or less smoothly into a kind of ‘lattice gas’. More recently, Frenkel et al [117] proposed a method in which λ corresponds to switching between the true potential and a harmonic spring, which couples each atom at instantaneous position \mathbf{r}_i to its ideal lattice site $\mathbf{r}_i^{(0)}$:

$$\mathcal{V}(\lambda) = \lambda\mathcal{V}_1 + (1 - \lambda)\mathcal{V}_0.$$

Here \mathcal{V}_1 is the original, many-body potential energy function, while \mathcal{V}_0 is a sum of single-particle spring potentials proportional to $|\mathbf{r}_i - \mathbf{r}_i^{(0)}|^2$. As $\lambda \rightarrow 0$ the system becomes a perfect Einstein crystal, whose free energy is exactly calculable. Recently, a combination of approaches has been used [118]: first the solid is subjected to a set of one-particle spring potentials, and then the influence of the interparticle forces is reduced to zero by expanding the crystal. This method was used to locate the melting transition for a model of nitrogen at $T = 300$ K.

Density functional theory arguments [119, 120] suggest that the springs should be of the correct strength to produce the same mean-squared displacement in the Einstein limit as in the original crystal. More precisely, it is best to switch over to a one-body potential in such a way that the one-body density $\rho(\mathbf{r})$ is the same at all points along the integration path. Such a path is guaranteed not to traverse a first-order phase transition.

B3.3.7 CONFIGURATION-BIASED MC

The biased-sampling approach may be considerably generalized, to allow the construction of MC moves step-by-step, with each step depending on the success or failure of the last. Such a procedure is biased, but it is then possible to correct for the bias (by considering the possible reverse moves). The technique has dramatically speeded up polymer simulations, and is capable of wider application.

The idea may be illustrated by considering first a method for increasing the acceptance rate of moves (but at the expense of trying, and discarding, several other possible moves). Having picked an atom to move, calculate the new trial interaction energy v_t for a range of trial positions $t = 1 \dots k$. Pick the actual attempted move from this set, with a probability proportional to the Boltzmann factor. This biases the move selection, towards high-probability states, but we can calculate the contribution of the bias to $\alpha_{\Gamma \rightarrow \Gamma'}$. Then we must

calculate $\alpha_{\Gamma \leftarrow \Gamma'}$ for the hypothetical reverse move: we do this by selecting $k-1$ possible trial positions around the new position of the atom, plus the place it originally came from, making k in all. The ratio $\alpha_{\Gamma \leftarrow \Gamma'}/\alpha_{\Gamma' \leftarrow \Gamma}$ is used in the accept/reject decision, along with the relevant Boltzmann factors (see [section b3.3.4.2](#)). For $k=1$ this gives the usual Metropolis prescription; for $k \rightarrow \infty$ it is easy to show that the acceptance rate tends to unity, but the method becomes very expensive, since all the work has gone into calculating the biasing factors.

The expense is justified, however, when tackling polymer chains, where reconstruction of an entire chain is expressed as a succession of atomic moves of this kind [[121](#)]. The first atom is placed at random; the second selected nearby (one bond length away), the third placed near the second, and so on. Each placement of an atom is given a greater chance of success by selecting from multiple locations, as just described. Biasing factors are calculated for the whole multi-atom move, forward and reverse, and used as before in the Metropolis prescription. For further details see [[122](#), [123](#), [124](#), [125](#)]. A nice example of this technique is the study [[126](#), [127](#)] of the distribution of linear and branched chain alkanes in zeolites.

B3.3.8 PHASE TRANSITIONS

Here we discuss the exploration of phase diagrams, and the location of phase transitions. See also [[128](#), [129](#), [130](#), [131](#)] and [[22](#), chapters 8–14]. Very roughly we classify phase transitions into two types: first-order and continuous. The fact that we are dealing with a finite-sized system must be borne in mind, in either case.

B3.3.8.1 FIRST-ORDER AND CONTINUOUS TRANSITIONS

At a continuous phase transition, a correlation length ξ (see [section b3.3.2.1](#)) diverges and an order parameter, typically the ensemble average of the corresponding dynamical variable, becomes macroscopically large. The divergence heralding the transition is describable in terms of universal exponent relations. Effects of finite size close to continuous phase transitions are well studied [[24](#), [132](#)]. By contrast, a first-order phase transition is abrupt, as one phase becomes thermodynamically more stable than another; there are no transition precursors. In the thermodynamic limit, there is a step-function discontinuity in most properties, including thermodynamic derivatives of the free energy. Again it is possible to describe the effects of finite size [[132](#), [133](#)].

For both first-order and continuous phase transitions, finite size *shifts* the transition and *rounds* it in some way. The shift for first-order transitions arises, crudely, because the chemical potential, like most other properties, has a finite-size correction $\mu(N) - \mu(\infty) \sim \mathcal{O}(1/N)$. An approximate expression for this was derived by Siepmann et al [[134](#)]. Therefore, the line of intersection of two chemical potential surfaces $\mu_I(T,P)$ and $\mu_{II}(T,P)$ will shift, in general, by an amount $\mathcal{O}(1/N)$. The rounding is expected because the partition function only has singularities (and hence produces discontinuous or divergent properties) in the limit $L \rightarrow \infty$; otherwise, it is analytic, so for finite N the discontinuities must be smoothed out in some way. The shift for continuous transitions arises because the transition happens when $\xi \rightarrow L$ for the finite system, but when $\xi \rightarrow \infty$ in the infinite system. The rounding happens for the same reason as it does for first-order phase transitions: whatever the nature of the divergence in thermodynamic properties (described, typically, by critical exponents) it will be limited by the finite size of the system.

In either case, first-order or continuous, it is useful to consider the probability distribution function for variables averaged over a spatial block of side L ; this may be the complete simulation box (in which case we

must specify the ensemble and boundary conditions) or it may be a sub-system. For purposes of illustration we shall not distinguish these possibilities.

B3.3.8.2 CONTINUOUS PHASE TRANSITIONS

Here we discuss only briefly the simulation of continuous transitions (see [132, 135] and references therein). Suppose that the transition is characterized by a non-vanishing order parameter X and a corresponding divergent correlation length ξ . We shall be interested in the block average value $X_L \equiv \langle \chi \rangle_L$, where the L reminds us of the system size. In a magnetic system, X is the magnetization; in a fluid it might be the density. The basic idea of finite size scaling analysis is that the values of properties of the system are dictated by the ratio ξ/L , and that no other length scales enter the problem, near a critical point. Any property can be written

$$X_L = X \times \Phi(\xi/L)$$

where X is the average value in the infinite system limit and Φ is some scaling function. There will be exponent laws dictating the behaviour of X in the vicinity of the phase transition, and more scaling laws stating how ξ behaves inside the function Φ . We can apply a scaling analysis to the distribution function $\mathcal{P}(X_L)$ [136, 137]. Actually at the critical point, the distribution can be calculated by simulation, or predicted by renormalization group theory [136, 137, 138 and 139]; different universal forms will be seen for different universality classes. Examination of these functions is a powerful way of locating and characterizing critical points, and in the critical region the histogram reweighting method is a particularly useful way of maximizing the information obtained from individual simulations. For example, the prewetting critical point has been shown to lie in the $d = 2$ Ising universality class in this way [139]. A further example is the study of the critical point of the $d = 2$ Lennard-Jones fluid [140, 141]. For this, long runs of order 10^6 – 10^8 sweeps were needed, but the system sizes were relatively small: $N \approx 100$ and 400.

B3.3.8.3 FIRST-ORDER PHASE TRANSITIONS

Consider simulating a system in the canonical ensemble, close to a first-order phase transition. In one phase, $\mathcal{P}_{NVT}(E)$ is essentially a Gaussian centred around a value E_I , while in the other phase the peak is around E_{II} . Far from the transition, one or other of these will apply. Close to the phase transition we will see contributions from both Gaussians, and a double-peaked distribution. The weight of each Gaussian changes as the temperature is varied. Thus, a smooth

-31-

crossover occurs from one branch of the equation of state $E(T) = \langle \mathcal{H} \rangle_{NVT}$ to the other. In the transition region we may expect to see anomalies such as an increased specific heat: the double-peaked distribution is wider than its constituent single-peaked ones, and recall that C_V is linked to $\langle \delta \mathcal{H}^2 \rangle$. The corresponding Landau free energy

$$\mathcal{F}_{NVT}(E) = -kT \ln \mathcal{P}_{NVT}(E)$$

has two minima separated by a barrier. The high-probability, low free energy values correspond to the single phase configurations; the intermediate values are for mixtures of the two phases, with an interfacial free energy penalty.

In the microcanonical ensemble, the signature of a first-order phase transition is the appearance of a ‘van der Waals loop’ in the equation of state, now written as $T(E)$ or $\beta(E)$. The $\beta(E)$ curve switches over from one

branch, phase I, of the equation of state to the other, phase II, tracing out a loop in the transition region. This loop is a finite-size effect, due to the interfacial free energy contributions in the transition region, just mentioned. For a larger system size the loop will flatten out, becoming a horizontal line in the thermodynamic limit, joining the two coexisting energies at the transition temperature; for $N \rightarrow \infty$ the interfacial properties contribute a negligible amount to the total free energy. (Calling it a ‘van der Waals loop’ is therefore misleading: it has no connection with the loop in the approximate van der Waals equation of state for fluids, which in any case is independent of system size.) It is possible to inter-relate the form of this loop, and the double-peaked structure of the energy distribution, with the thermodynamic coexistence conditions [[1](#), [132](#), [142](#), [143](#), [144](#), [145](#)].

The previous discussions translate directly over into pressure–volume variables, if we compare the constant- NVT and constant- NPT ensembles. Double-peaked distributions of volumes are seen near a transition at constant pressure.

Direct coexistence of solid and fluid phases of hard spheres and disks was observed in the early simulations of Wood and Jacobson [[146](#)] and Alder and Wainwright [[147](#), [148](#)]; the appearance of a ‘van der Waals loop’ in the equation of state was explained in some detail shortly afterwards [[142](#), [149](#), [150](#)]. Very detailed analyses of this situation, especially in relation to spin systems, have appeared in recent years [[143](#), [144](#), [145](#), [151](#) and [152](#)]. Histogram reweighting can be useful here [[153](#), [154](#)] and measuring the height of the interfacial free energy barrier as a function of system size has been recommended as a test for first-order behaviour [[155](#)]. A nice example of this approach for a spin model of a liquid crystal, which exhibits a tricky *weak* first-order transition, is the work of Zhang et al [[156](#), [157](#)], following on from earlier work by Fabbri and Zannoni [[158](#)]. An example for a *strong* first-order transition is the study of melting and nucleation barriers in the Lennard-Jones system [[159](#)] and models of metallic systems [[160](#)]. This approach used windowing and biased sampling techniques.

Recently, Orkoulas and Panagiotopoulos [[161](#)] have shown that it is possible to use histogram reweighting and multicanonical simulations, starting with individual simulations near the critical point, to map out the liquid–vapour coexistence curve in a very efficient way.

Simulations in the Gibbs ensemble attempt to combine features of Widom’s test particle method with the direct simulation of two-phase coexistence in a box. The method of Panagiotopoulos et al [[162](#), [163](#)] uses two fully-periodic boxes, I and II.

In the simplest version, a one-component system is simulated at a given temperature T in both boxes; particles in different boxes do not interact directly with each other; however, volume moves and particle creation and deletion

moves are coupled such that the total volume V and the total number of particles N are conserved. With appropriate acceptance/rejection probabilities for the volume exchange and particle exchange moves, together with the usual MC procedure for moving particles around within the two boxes, the thermodynamic conditions for mechanical and chemical equilibrium between the boxes are ensured. A typical MC cycle would consist of: one attempted move per particle in each box; one attempt to exchange volumes between boxes; a predetermined number of attempts to exchange particles. The technique has been reviewed by Panagiotopoulos [[131](#), [164](#), [165](#)] and Smit [[129](#)]. The partition function for the two-box system is simply the usual canonical sum over all possible states, including a sum over all the distributions of particles between the boxes such that $N_I + N_{II} = N$, and an integral over all box volumes such that $V_I + V_{II} = V$. The probability distribution function for the ensemble, and the acceptance and rejection rules for particle and volume

exchanges, are easily derived.

The characteristic feature of the technique is the behaviour of the system if the overall density N/V lies in a two-phase region. For a single simulation box, both phases would appear, with an interface between them; in the Gibbs ensemble, the interface free energy penalty can be avoided by the system arranging to have each phase entirely in its own box. This phase separation happens automatically during the equilibration stage of the simulation.

The great advantages of the technique are its avoidance of interfacial properties, and the semi-automatic way that it converges on the coexisting densities without the need to input chemical potentials or guess equations of state. Unavoidably, it suffers from the same problems as the Widom test-particle method: at high density the particle exchange moves are accepted with very low probability, and special techniques are required to overcome this. It is essential to monitor the success rate of exchanges, and carry out enough of them to ensure that a few percent of molecules are exchanged at each step.

The Gibbs ensemble method has been outstandingly successful in simulating complex fluids and mixtures. For a multicomponent system, it is possible to simulate at constant pressure rather than constant volume, as separation into phases of different compositions is still allowed. The method allows one to study straightforwardly phase equilibria in confined systems such as pores [166]. Configuration-biased MC methods can be used in combination with the Gibbs ensemble. An impressive demonstration of this has been the determination by Siepmann et al [167] and Smit et al [168] of liquid–vapour coexistence curves for *n*-alkane chain molecules as long as 48 atoms.

As we have seen, insertion of small molecules can be dramatically easier than large ones; this leads to a ‘semi-grand’ version of the Gibbs ensemble [131, 169, 170]: the smaller particles are exchanged between the boxes, while moves that interconvert particle species are carried out within the boxes. The ideas seen before for computing the chemical potential by gradual insertion [111, 112, 113] can be naturally generalized to the Gibbs ensemble: at each stage a given molecule may be in an intermediate state of transfer between one box and another. As an example, Escobedo and de Pablo [171] use an expanded Gibbs ensemble for polymers, and Nath et al [172] have computed the liquid–vapour envelopes for long-chain alkanes, using different potential models and comparing with previous work [167, 168].

B3.3.8.5 THERMODYNAMIC METHODS

The alternative to direct simulation of two-phase coexistence is the calculation of free energies or chemical potentials together with solution of the thermodynamic coexistence conditions. Thus, we must solve (say) $\mu_I(P) = \mu_{II}(P)$ at constant T . A reasonable approach [173, 174, 175 and 176] is to conduct constant- NPT simulations, measure μ by test-particle insertion, and also to note that the simulations give the derivative $\partial\mu/\partial P = \langle V \rangle / N$ directly. Thus, conducting

one or two simulations may be enough for a preliminary fit to the equations of state $\mu_I(P)$, $\mu_{II}(P)$ allowing one to home in on the intersection point quite quickly.

Once a point on the coexistence line has been found, one can trace out more of it using the approach of Kofke [177, 178] to numerically integrate the Clapeyron equation

$$\left(\frac{dP}{d\beta}\right)_\sigma = -\frac{\Delta h}{\beta\Delta v}.$$

Here, $\Delta h = h_\beta - h_\alpha$ is the difference in molar enthalpies of the coexisting phases, and Δv is the difference in molar volumes; the suffix σ indicates that the derivative is to be evaluated along the coexistence line.

The method consists of solving the above equation in a standard step-by-step manner, for example using a predictor–corrector algorithm. The right-hand side is calculated by simulating both phases at constant T and P in separate, uncoupled boxes. At intervals, a small change in T (the independent variable) is made in both boxes, and this is accompanied by a change in P (the dependent variable) as dictated by the differential equation solver. The approach relies on a starting point at which the two phases are at thermodynamic equilibrium, $\Delta\mu = 0$; thereafter the Clapeyron equation, if solved accurately, should guarantee that equilibrium is maintained. The method has been applied to the liquid–vapour coexistence curve [177, 178] and to the melting and sublimation curves [179] for the Lennard-Jones system; it was also extended by Agrawal and Kofke [180] to study the melting transition of a large family of soft-sphere systems, showing the emergence of the bcc phase as being stable relative to fcc for high enough softness parameters. Various technical details of this approach have been discussed [178, 179] and possible sources of inaccuracy considered.

An example of the use of this method in a complex situation is the study by Bolhuis and Kofke [181, 182] of the freezing of polydisperse hard spheres (a system in which there is a distribution of atomic diameters). A semi-grand ensemble imposes a distribution of chemical potential differences (for different hard-sphere diameters) on the system: the width of this distribution controls the polydispersity. The same chemical potentials (for all the different species) apply in two coexisting phases. The thermodynamic integration technique may then be used to map out the freezing–melting line in the pressure–polydispersity plane, starting from the monodisperse limit (simple hard spheres). The resulting phase diagram, in volume fraction–polydispersity variables, is shown in figure B3.3.9. An important result is that fractionation between the two phases allows a highly polydisperse fluid to precipitate a solid which is only slightly polydisperse—never higher than 5.7% in terms of the average sphere diameter.

-34-

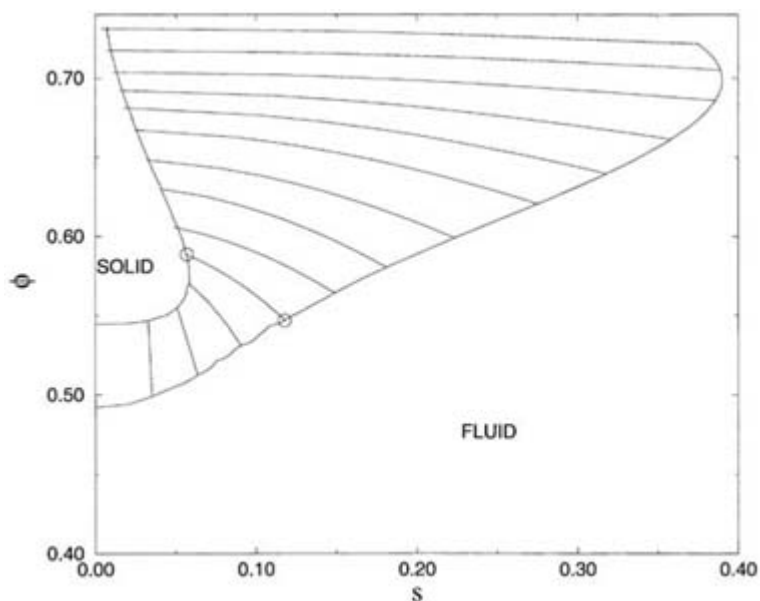


Figure B3.3.9. Phase diagram for polydisperse hard spheres, in the volume fraction (ϕ)–polydispersity (s) plane. Some tie-lines are shown connecting coexisting fluid and solid phases. Thanks are due to D A Kofke and P G Bolhuis for this figure. For further details see [181, 182].

B3.3.8.6 STUDIES OF INTERFACES

Simulation of both bulk phases in a single box, separated by an interface, is closest to what we do in real life. It is necessary to establish a well defined interface, most often a planar one between two phases in ‘slab’ geometry. A large system is required, so that one can characterize the two phases far from the interface, and read off the corresponding bulk properties. Naturally, this is the approach of choice if the interfacial properties (for example, the surface tension) are themselves of interest. The first stage in such a simulation is to prepare bulk samples of each phase, as close to the coexisting densities as possible, in cuboidal periodic boundaries, using boxes whose cross sections match. The two boxes are brought together, to make a single longer box, giving the desired slab arrangement with two planar interfaces. There must then follow a period of equilibration, with mass transfer between the phases if the initial densities were not quite right.

Equilibration of the interface, and the establishment of equilibrium between the two phases, may be very slow. Holcomb et al [183] found that the density profile $\rho(z)$ equilibrated much more quickly than the profiles of normal and transverse pressure, $P_N(z)$ and $P_T(z)$, respectively. The surface tension is proportional to the z -integral of $P_N(z)-P_T(z)$. The bulk liquid in the slab may continue to contribute to this integral, indicating lack of equilibrium, for very long times if the initial liquid density is chosen a little too high or too low. A recent example of this kind of study, is the MD simulation of the liquid–vapour surface of water at temperatures between 316 and 573 K by Alexandre et al [184].

B3.3.9 RARE EVENTS

By definition, rare events happen infrequently, but this does not mean that they happen slowly. Accordingly, molecular simulations may contribute greatly to our understanding of such events, but some special sampling tricks will be needed since the ‘natural’ timespans of simulations are already very short. The simplest example is where the system crosses a free energy barrier from one region to another in phase space, and a single ‘reaction coordinate’ can be identified to characterize the two stable regions and the transition state. The statistical mechanical background to barrier crossing rates, in terms of linear response theory, has been given by Chandler [2, 185]. The transition rate is typically a product of two factors: the equilibrium probability density for finding the system at the top of the barrier, and a dynamical quantity, essentially the inverse of a relaxation time for the system to settle from the barrier into one or other stable region.

We have already discussed weighted sampling methods for exploring regions of high free energy, so the first part of this problem is tractable. The calculation of time-dependent functions, which start from the barrier top, is facilitated by the so-called ‘blue-moon’ ensemble [186] in which a constraint is applied to keep the system exactly on the desired hypersurface. This allows sampling of the starting conditions with good statistics; then the constraint may be released and subsequent dynamics accumulated. (Metric tensor factors associated with the constraint are discussed elsewhere [187, 188].) To compute the time-dependent part of the barrier-crossing rate, special approaches have been developed to suppress transient behaviour and statistical noise [187].

In many cases, it may not be possible to identify a single reaction coordinate. A good example of a free-energy surface depending on two variables is found in the study by ten Wolde and Frenkel [189] of the mechanism of protein crystal nucleation, and the possible influence of fluctuations induced by a nearby metastable critical point. Here, the relevant variables are ‘density’ and ‘crystallinity’, as illustrated in [figure B3.3.10](#). At high levels of supercooling, well away from the hidden critical point, nucleation follows a route whereby a crystalline nucleus forms from the start: this involves a high free energy barrier. Close to the critical point, a different mechanism operates: critical fluctuations encourage formation of a liquid-like nucleus, and only after this has grown to a certain size does a crystal start to form. This pathway involves a

much lower free energy barrier. This work is an example of the insight obtainable from simulations into a previously poorly understood area, namely the art of obtaining good protein crystals for structure determination, as well as illustrating the qualitatively different mechanisms that may operate under different conditions.

-36-

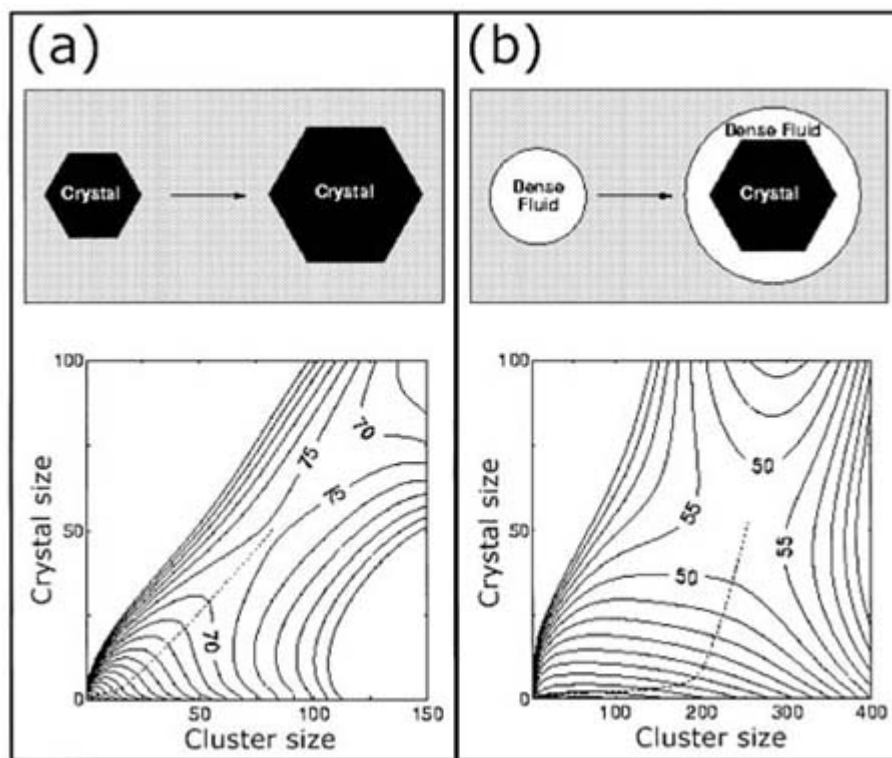


Figure B3.3.10. Contour plots of the free energy landscape associated with crystal nucleation for spherical particles with short-range attractions. The axes represent the number of atoms identifiable as belonging to a high-density cluster, and as being in a crystalline environment, respectively. (a) State point significantly below the metastable critical temperature. The nucleation pathway involves simple growth of a crystalline nucleus. (b) State point at the metastable critical temperature. The nucleation pathway is significantly curved, and the initial nucleus is liquidlike rather than crystalline. Thanks are due to D Frenkel and P R ten Wolde for this figure. For further details see [189].

The more general problem of finding a transition pathway when the relevant reaction coordinates are not obvious, has recently been tackled [190, 191]. The basic idea [192] is to generate chains of states linking the two stable states, through a weighted sampling procedure which makes no assumptions about the mechanism. The method is very general, but inevitably expensive.

B3.3.10 QUANTUM SIMULATION USING PATH INTEGRALS

In this section we look briefly at the problem of including quantum mechanical effects in computer simulations. We shall only examine the simplest technique, which exploits an isomorphism between a quantum system of atoms and a classical system of ring polymers, each of which represents a path integral of the kind discussed in [193]. For more details on work in this area, see [22, 194] and particularly [195, 196, 197].

The coordinate representation of the density matrix, in the canonical ensemble, may be written

-37-

$$\mathcal{Q}(\mathbf{r}^{(N)}, \mathbf{r}^{(N')}) = \mathcal{Q}_{NVT}^{-1}(\mathbf{r}^{(N)}) |e^{-\beta \hat{H}}| \mathbf{r}^{(N')}$$

and correspondingly for the partition function

$$\mathcal{Q}_{NVT} = \int d\mathbf{r}^{(N)} (\mathbf{r}^{(N)} | e^{-\beta \hat{H}} | \mathbf{r}^{(N)}).$$

Here we have adopted a Dirac bracket notation ($\dots | \dots \rangle$) which should be distinguished from the ensemble average $\langle \dots \rangle$. Actually evaluating this is tricky, because the Hamiltonian is the sum of the kinetic and potential energy operators, $\hat{H} = \hat{K} + \hat{V}$, which do not commute. Hence

$$e^{-\beta(\hat{K}+\hat{V})} \neq e^{-\beta\hat{K}} e^{-\beta\hat{V}}.$$

When the exponent is small (e.g., at high temperature), reasonable approximations exist. This problem is attacked in a manner similar to that used to derive expressions for the propagator $e^{i\hat{L}t}$, as a succession of small time step propagators, in [section b3.3.3.2](#): we split the exponential up into smaller pieces. So, we write

$$e^{-\beta\hat{H}} = [e^{-\beta\hat{H}/P}]^P$$

and insert this into the expression for the partition function

$$\mathcal{Q}_{NVT} = \int d\mathbf{r}^{(N)} (\mathbf{r}^{(N)} | e^{-\beta\hat{H}/P} e^{-\beta\hat{H}/P} \dots e^{-\beta\hat{H}/P} | \mathbf{r}^{(N)}).$$

Now we do one of the standard quantum mechanical tricks, inserting the identity operator as a complete sum of states in the coordinate representation:

$$\hat{1} = \int d\mathbf{r}^{(N')} | \mathbf{r}^{(N')} \rangle \langle \mathbf{r}^{(N')} |$$

in between each exponential. This will introduce $P - 1$ additional integrations over coordinates. Each of the contributions $(\mathbf{r}^{(N)} | e^{-\beta\hat{H}/P} | \mathbf{r}^{(N')})$ is an un-normalized, off-diagonal, density matrix $\mathcal{Q}(\mathbf{r}^{(N)}, \mathbf{r}^{(N')})$ evaluated at a temperature a factor P higher than the temperature of the real system. For more background on this approach to quantum mechanics, see [\[193\]](#).

As in the case of the propagator, we shall be applying a symmetrical version of the Trotter formula [\[48\]](#) to the high-temperature density matrix

$$(\mathbf{r}^{(N)} | e^{-\beta\hat{H}/P} | \mathbf{r}^{(N')}) \approx (\mathbf{r}^{(N)} | e^{-\frac{1}{2}\beta\hat{V}/P} e^{-\beta\hat{K}/P} e^{-\frac{1}{2}\beta\hat{V}/P} | \mathbf{r}^{(N')}).$$

The potential energy part is diagonal in the coordinate representation, and we drop the hat indicating an operator henceforth. The kinetic energy part may be evaluated by transforming to the momentum representation and carrying out a Fourier transform. The result is

$$Q_{NVT} \approx \left(\frac{Pm}{2\pi\beta\hbar^2} \right)^{dP/2} \int \dots \int d\mathbf{r}_1^{(N)} \dots d\mathbf{r}_P^{(N)} e^{-\beta\mathcal{V}_{\text{qu}}} e^{-\beta\mathcal{V}_{\text{cl}}}$$

$$\mathcal{V}_{\text{qu}} = -\frac{Pm}{2\beta^2\hbar^2} (|\mathbf{r}_1^{(N)} - \mathbf{r}_2^{(N)}|^2 + |\mathbf{r}_2^{(N)} - \mathbf{r}_3^{(N)}|^2 + \dots + |\mathbf{r}_P^{(N)} - \mathbf{r}_1^{(N)}|^2)$$

$$\mathcal{V}_{\text{cl}} = \frac{1}{P} (\mathcal{V}(\mathbf{r}_1^{(N)}) + \mathcal{V}(\mathbf{r}_2^{(N)}) + \dots + \mathcal{V}(\mathbf{r}_P^{(N)}))$$

This is better understood with a picture: see figure B3.3.11. The discretized path-integral is isomorphic to the classical partition function of a system of N ring polymers each having P atoms. Each atom in a given ring corresponds to a different ‘imaginary time’ point $p = 1 \dots P$. $\mathcal{V}(\mathbf{r}^{(N)})$ represents the interatomic interactions (for example, Lennard-Jones) between the atoms of the real system. This couples together only *correspondingly labelled* atoms, i.e., atoms with the same index p . So, between each pair of the original atoms, there are P such interactions, each one weaker than the true potential by a factor $1/P$. In addition, harmonic quantum ‘springs’ couple together successively indexed atoms within a ring polymer. We may simulate this classical ring polymer system by conventional MC or MD.

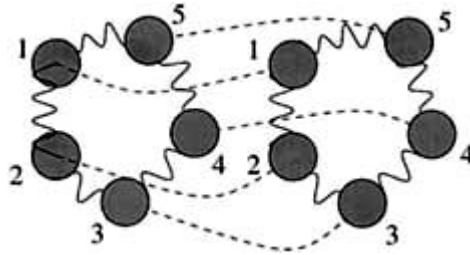


Figure B3.3.11. The classical ring polymer isomorphism, for $N=2$ atoms, using $P=5$ beads. The wavy lines represent quantum ‘spring bonds’ between different imaginary-time representations of the same atom. The dashed lines represent real pair-potential interactions, each diminished by a factor P , between the atoms, linking corresponding imaginary times.

Temperature appears in the partition function in an unusual way. The average energy takes the form

$$E = \frac{3}{2}NPkT + \langle \mathcal{V}_{\text{cl}} \rangle - \langle \mathcal{V}_{\text{qu}} \rangle.$$

As P is increased, the partial cancellation between the kinetic part and the spring part may worsen the statistics on E . This has led to suggestions of alternative ways of estimating E [198]. As P goes up, the springs become stronger, the interactions in \mathcal{V}_{cl} become (individually) weaker, and this leads to sampling problems. In MD, one needs to use multiple time step methods to ensure proper handling of the spring vibrations, and there is a possible physical bottleneck in the transfer of energy between the spring system and the other degrees of freedom which must be handled properly [199]. In MC, one needs to use special methods to sample configuration space efficiently [200, 201].

For fermions (especially) and bosons there are additional problems. Let \hat{P} be one of the $N!$ permutations of particle labels. Then the fermion density matrix ϱ_F has the symmetry

$$\varrho_F(\mathbf{r}^{(N)}, \mathbf{r}^{(N')}) = (-1)^{\hat{P}} \varrho_F(\hat{P}\mathbf{r}^{(N)}, \mathbf{r}^{(N')}) = (-1)^{\hat{P}} \varrho_F(\mathbf{r}^{(N)}, \hat{P}\mathbf{r}^{(N')}).$$

It is possible to relate this to the ‘Boltzmann’ (i.e., distinguishable particle) density matrix $\varrho(\mathbf{r}^{(N)}, \mathbf{r}^{(N')})$ by

$$\varrho_F(\mathbf{r}^{(N)}, \mathbf{r}^{(N')}) = \frac{1}{N!} \sum_{\hat{P}} (-1)^{\hat{P}} \varrho(\hat{P}\mathbf{r}^{(N)}, \mathbf{r}^{(N')}).$$

It is necessary to sum over these permutations in a path integral simulation. (The same sum is needed for bosons, without the sign factor.) For fermions, odd permutations contribute with negative weight. Near-cancelling positive and negative permutations constitute a major practical problem [196].

B3.3.11 CAR–PARRINELLO SIMULATIONS

Car and Parrinello [202] proposed a technique for efficiently solving the Schrödinger equation which has had an enormous impact on materials simulation (for reviews, see [203, 204, 205, 206]). The technique is an *ab initio* one, i.e., free of empirical parameters, and is based on the use of a quantum mechanical orthonormal basis set $\psi^{(n)} \equiv \psi_i(\mathbf{r})$ to describe the electronic degrees of freedom. Specifically, the aim is to obtain the electron density $\rho(\mathbf{r})$; the total energy of the system may then be written as a functional of this density, whose minimization yields the ground state energy [207]. Pseudopotentials [208] represent the effects of the atomic cores on the valence electrons, allowing some economies. The energy functional is written

$$E[\psi^{(n)}, \mathbf{R}^{(N)}] = \sum_i (\psi_i | \hat{\mathcal{K}} + \hat{\mathcal{V}}_{\text{ps}} | \psi_i) + \frac{1}{2} \int d\mathbf{r} d\mathbf{r}' \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \quad (\text{B3.3.8})$$

$$+ E_{\text{xc}}[\rho] + \frac{1}{2} \sum_{i \neq j} \frac{Q_i Q_j}{|\mathbf{R}_i - \mathbf{R}_j|}. \quad (\text{B3.3.9})$$

Here we distinguish between nuclear coordinates \mathbf{R} and electronic coordinates \mathbf{r} ; $\hat{\mathcal{K}}$ is the single-particle kinetic energy operator, and $\hat{\mathcal{V}}_{\text{ps}}$ is the total pseudopotential operator for the interaction between the valence electrons and the combined nucleus + frozen core electrons. The electron–electron and nucleus–nucleus Coulomb interactions are easily recognized, and the remaining term $E_{\text{xc}}[\rho]$ is the electronic exchange and correlation energy functional. This is usually treated in the local density approximation, using ground-state data for the homogeneous electron gas [209]; the most promising improvements seem to be based on the addition of gradient corrections [205, 206].

For each configuration of the nuclei, minimization of the total energy with respect to the electron density yields the instantaneous value of a potential energy function $\mathcal{V}(\mathbf{R}^{(N)})$, and the corresponding forces on the nuclei. In principle,

assuming an adiabatic separation between nuclear and electronic motion, Newton's equations for the nuclei may be solved in the usual way, while the electrons are allowed to evolve according to Schrödinger's equations, remaining on the instantaneous ground-state surface. This turns out to be very inefficient in practice, and the breakthrough came with the suggestion [202] that a classical dynamical evolution of the electronic configuration could be used to stay in the ground state. A Lagrangian, involving both nuclear and electronic degrees of freedom, is written down; the electrons are given fictitious masses, and minimization of the electronic energy may be performed by introducing a friction coefficient.

The Car–Parrinello method has found wide applicability, especially for studying systems in which structure and bonding are inseparable, or for materials under extreme conditions for which empirical potential would be unreliable. Examples are the studies by Alfe and Gillan [210] and de Wijs et al [211] of iron in the Earth's core, at temperatures of several thousand Kelvin and pressures sufficient to compress the metal to about half its normal volume. It was concluded that the liquid iron in the core is not exceptionally viscous (as has been suggested by some seismic measurements) and that dissolved sulphur atoms show no tendency to form clusters or chains (which might have a large effect on viscosity). This is shown in figure B3.3.12. Additionally, the simulations suggest that the solid part of the core has the hcp crystal structure, contrary to that inferred from experiments at lower pressure and temperature.

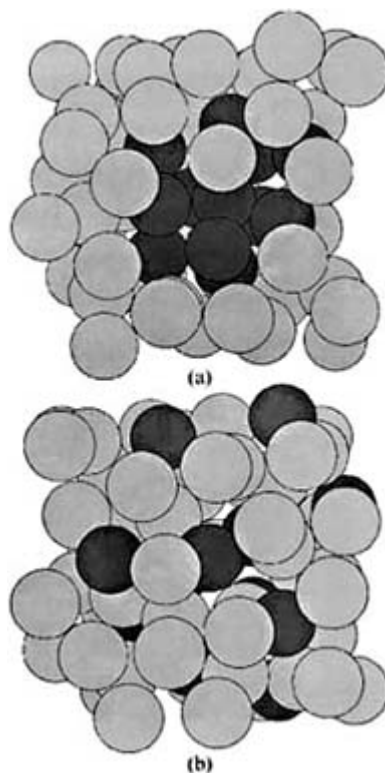


Figure B3.3.12. Sulphur atoms in liquid iron at the Earth's core conditions, simulated by first-principle Car–Parrinello molecular dynamics. (a) Initial conditions, showing a manually-prepared initial cluster of sulphur atoms. (b) A short time later, indicating spontaneous dispersal of the sulphur atoms, which mingle with the surrounding iron atoms. Thanks are due to D Alfe and M J Gillan for this figure. For further details see [210, 211].

A similar approach, in spirit, has been proposed [212] for the study of two-component classical systems, for example polyelectrolytes, which consist of mesoscopic, highly-charged, polyions, and microscopic,

oppositely-charged, counterions. This time, an ‘effective free energy’, depending parametrically on the polyion coordinates \mathbf{R} , arises by integration over the counterion coordinates \mathbf{r} . A unified dynamical scheme, in which the counterions are given fictitious masses, allows the counterion density to adjust adiabatically to the slower motion of the polyions, and hence permits the free energy to be minimized as the system evolves.

B3.3.12 PARALLEL SIMULATIONS

MD programs may be efficiently parallelized, that is, the computational work divided between many processors to result in faster execution. Well established message-passing standards and software make it relatively easy to write portable and efficient codes. The algorithm to advance positions and momenta is trivially handled, with each processor being responsible for a subset of the atoms. The critical considerations are (i) the parallelization of the time-consuming force calculation, and (ii) the overheads associated with communicating information between processors. Two general methodologies seem to be most promising. In the *replicated data* method, all the processors hold copies of all the atomic coordinates; however, in the double loop over pair interactions, each processor deals with a subset of pairs. Some care needs to be taken to balance the load between processors, and the results of the force calculations must be broadcast to all other processors, which may be time-consuming, but perhaps the biggest drawback is the memory requirement of holding copies of all data on all nodes. Nonetheless, this method is easy to program and reasonably efficient for many purposes [213, 214, 215]. In the *domain decomposition* method, the simulation box is split into (usually) cubic regions, and each processor is responsible only for the atoms in a given region; there is some communication of information from neighbouring domains before the force calculation, and also some redistribution of atoms as they move around the system. This approach may be integrated with the link-cell approach of section b3.3.3.5, and is especially efficient for systems with short-range forces [213, 215, 216, 217, 218]. An example of the capabilities of such an approach on a massively parallel supercomputer is the study by Holian and Lomdahl [219] of shock waves in a fcc crystal of 10 million atoms, as illustrated in figure B3.3.13. In this case, a shock is generated by reflecting atoms at a piston face, i.e., imposing a momentum mirror. A system of this size was essential to ensure that the periodic boundaries do not limit the plastic flow induced by the shock: this can be seen in the randomly-spaced plaid pattern of the figure. The wave propagates back about 60 lattice spacings, generates a large number of stacking faults distributed randomly on the four {111} slip systems, and eventually produces a nonplanar propagation front.

-42-

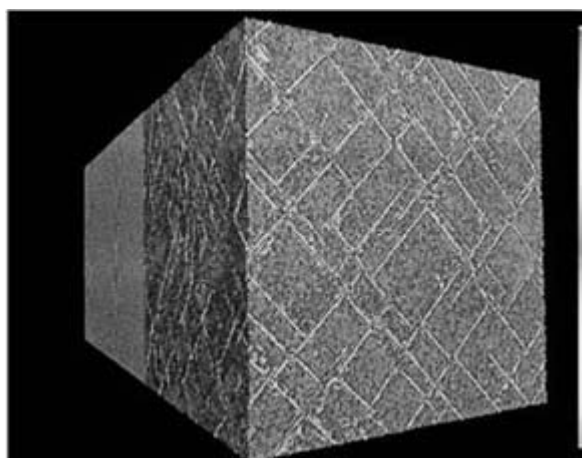


Figure B3.3.13. Intersecting stacking faults in a fcc crystal at the impact plane induced by collision with a momentum mirror for a square cross section of side 100 unit cells. The shock wave has advanced half way to the rear (~250 planes). Atom shading indicates potential energy. Thanks are due to B Holian for this figure.

For further details see [219].

Although this section has concentrated on MD, it should not be forgotten that lattice-based MC codes may be parallelized very efficiently; for more information on parallel simulation methods see [220, 221, 222 and 223] and references therein.

B3.3.13 OUTLOOK

With the rapid development of computer power, and the continual innovation of simulation methods, it is impossible to predict what may be achieved over the next few years, except to say that the outlook is very promising. The areas of rare events, phase equilibria, and quantum simulation continue to be active.

An easily recognizable trend is the increasing application of simulation methods to problems of direct practical benefit to industry [224]. A pointer to this kind of use is provided by the synthesis of a small-pore microporous material, using a structure-directing template molecule designed by computer [225]. The aim is to promote formation of a desired material (here a cobalt aluminophosphate catalyst in the so-called CHA structure) without generating competing microporous phases. The simulation procedure allows molecular entities to be grown from a seed molecule by adding standard fragments, under the control of a cost function to minimize non-bonded overlaps with the surrounding CHA framework. Likely templates are ranked by binding energy, which measures how well they fit in the pore. The result is illustrated in [figure B3.3.14](#). This led to a successful synthetic route: the suggested template molecule forms the desired pure mesoporous material in 4 h at 180°C, without the formation of competing structures which are found with other templates.

-43-

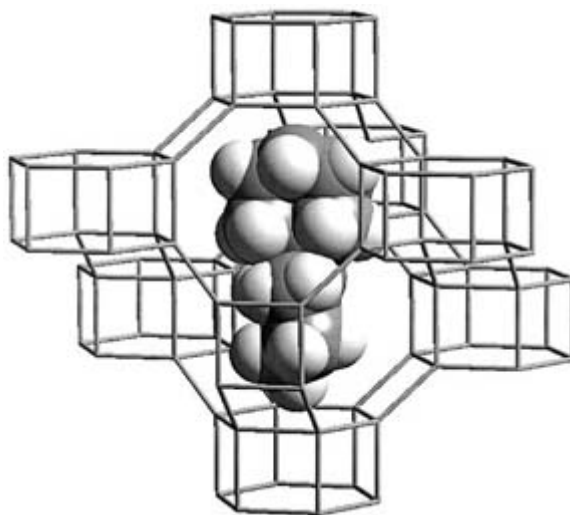


Figure B3.3.14. Template molecule in a zeolite cage. The CHA structure (periodic in the calculation but only a fragment shown here) is drawn by omitting the oxygens which are positioned approximately halfway along the lines shown connecting the tetrahedral silicon atoms. The molecule shown is 4-piperidinopiperidine, which was generated from the dicyclohexane motif suggested by computer. Thanks are due to D W Lewis and C R A Catlow for this figure. For further details see [225].

A further theme is the development of techniques to bridge the length and time scales between truly molecular-scale simulations and more coarse-grained descriptions. Typical examples are dissipative particle dynamics [226] and the lattice-Boltzmann method [227]. Part of the motivation for this is the recognition that

brute-force molecular simulation will always be limited in time scale by achievable chip speeds, even if increased use of parallel computers allows one to tackle larger length scales. Nonetheless, there will always be the need to relate such work to underlying molecular parameters, through statistical mechanics. A more detailed discussion of these techniques would take us beyond the scope of this chapter.

REFERENCES

- [1] Binder K and Heermann D W 1997 *Monte Carlo Simulation in Statistical Physics* 3rd edn, vol 80 *Solid State Sciences* (Berlin: Springer)
 - [2] Chandler D 1987 *Introduction to Modern Statistical Mechanics* (New York: Oxford University Press)
 - [3] Hansen J-P and McDonald I R 1986 *Theory of Simple Liquids* 2nd edn (London: Academic Press)
 - [4] Möller-Krumbhaar H and Binder K 1973 Dynamic properties of the Monte-Carlo method in statistical mechanics *J. Stat. Phys.* **8** 1–24
-
- 44-
- [5] Ferrenberg A M, Landau D P and Binder K 1991 Statistical and systematic errors in Monte-Carlo sampling *J. Stat. Phys.* **63** 867–82
 - [6] Binder K (ed) 1995 *The Monte Carlo Method in Condensed Matter Physics* vol 71 *Topics in Applied Physics* 2nd edn (Berlin: Springer)
 - [7] Allen M P and Tildesley D J 1987 *Computer Simulation of Liquids* (Oxford: Clarendon)
 - [8] Goldstein H 1980 *Classical Mechanics* 2nd edn (Reading, MA: Addison-Wesley)
 - [9] Maitland G C, Rigby M, Smith E B and Wakeham W A 1981 *Intermolecular Forces: Their Origin and Determination* (Oxford: Clarendon)
 - [10] Gray C and Gubbins K E 1984 *Theory of Molecular Fluids* (Oxford: Clarendon)
 - [11] Sprik M 1993 Effective pair potentials and beyond *Computer Simulation in Chemical Physics* vol 397 *NATO ASI Series C* ed M P Allen and D J Tildesley (Dordrecht: Kluwer) pp 211–59
 - [12] Rahman A 1964 Correlations in the motion of liquid argon *Phys. Rev. A* **136** 405–11
 - [13] Verlet L 1967 Computer experiments on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules *Phys. Rev.* **159** 98–103
 - [14] Gay J G and Berne B J 1981 Modification of the overlap potential to mimic a linear site–site potential *J. Chem. Phys.* **74** 3316–19
 - [15] de Miguel E, Rull L F, Chalam M K and Gubbins K E 1991 Liquid crystal phase diagram of the Gay–Berne fluid *Mol. Phys.* **74** 405–24
 - [16] Berardi R, Emerson A P J, Smith W and Zannoni C 1993 Monte Carlo investigations of a Gay–Berne liquid crystal *J. Chem. Soc. Faraday Trans.* **89** 4069–78
 - [17] Allen M P, Warren M A, Wilson M R, Sauron A and William S 1996 Molecular dynamics calculation of elastic constants in Gay–Berne nematic liquid crystals *J. Chem. Phys.* **105** 2850–8
 - [18] Bates M A and Luckhurst G R 1996 Computer simulation studies of anisotropic systems. 26. Monte Carlo investigations of a Gay–Berne discotic at constant pressure *J. Chem. Phys.* **104** 6696–709

- [19] Wilson M R 1997 Molecular dynamics simulations of flexible liquid crystal molecules using a Gay-Berne/Lennard-Jones model *J. Chem. Phys.* **107** 8654–63
- [20] Billeter J and Pelcovits R 1998 Simulations of liquid crystals *Comput. Phys.* **12** 440–8
- [21] Binder K 1992 *The Monte Carlo Method in Condensed Matter Physics* (Berlin: Springer)
- [22] Binder K and Ciccotti G (ed) 1996 *Monte Carlo and Molecular Dynamics of Condensed Matter Systems* vol 49 (Bologna: Italian Physical Society)
- [23] Newman M E J and Barkema G T 1999 *Monte Carlo Methods in Statistical Physics* (Oxford: Clarendon)

-45-

- [24] Binney J J, Dowrick N J, Fisher A J and Newman M E J 1992 *The Theory of Critical Phenomena* (Oxford: Oxford University Press)
- [25] Kremer K 1996 Computer simulation methods for polymer physics *Monte Carlo and Molecular Dynamics of Condensed Matter Systems* vol 49, ed K Binder and G Ciccotti (Bologna: Italian Physical Society) pp 669–723
- [26] Salsburg Z W, Jacobson J D, Fickett W and Wood W W 1959 Application of the Monte Carlo method to the lattice gas model. Two dimensional triangular lattice *J. Chem. Phys.* **30** 65–72
- [27] Chesnut D A and Salsburg Z W 1963 Monte Carlo procedure for statistical mechanical calculation in a grand canonical ensemble of lattice systems *J. Chem. Phys.* **38** 2861–75
- [28] McDonald I R and Singer K 1967 Calculation of thermodynamic properties of liquid argon from Lennard-Jones parameters by a Monte Carlo method *Discuss. Faraday Soc.* **43** 40–9
- [29] Valleau J P and Card D N 1972 Monte Carlo estimation of the free energy by multistage sampling *J. Chem. Phys.* **57** 5457–62
- [30] Rickman J M and Phillpot S R 1991 Temperature dependence of thermodynamic quantities from simulations at a single temperature *Phys. Rev.L* **66** 349–52
- [31] Allen M P 1993 Back to basics *Computer Simulation in Chemical Physics* vol 397 *NATO ASI Series C* ed M P Allen and D J Tildesley (Dordrecht: Kluwer) pp 49–92
- [32] Butler B D, Ayton O, Jepps O G and Evans D J 1998 Configurational temperature: verification of Monte Carlo simulations *J. Chem. Phys.* **109** 6519–22
- [33] Rugh H H 1997 Dynamical approach to temperature *Phys. Rev.L* **78** 772–4
- [34] Widom B 1963 Some topics in the theory of fluids *J. Chem. Phys.* **39** 2808–12
- [35] Evans D J and Morriss G P 1990 *Statistical Mechanics of Nonequilibrium Liquids* (London: Academic)
- [36] Holian B L 1996 The character of the nonequilibrium steady state: beautiful formalism meets ugly reality *Monte Carlo and Molecular Dynamics of Condensed Matter Systems*, vol 49, ed K Binder and G Ciccotti (Bologna: Italian Physical Society) pp 791–822
- [37] Reichl L E 1980 *A Modern Course in Statistical Physics* (Austin, TX: University of Texas Press)
- [38] Friedman H L 1985 *A Course in Statistical Mechanics* (Englewood Cliffs, NJ: Prentice-Hall)
- [39] Holian B L and Evans D J 1985 Classical response theory in the Heisenberg picture *J. Chem. Phys.* **83** 3560–6
- [40] Gear C W 1966 The numerical integration of ordinary differential equations of various orders *ANL* 7126

- [41] Gear C W 1971 *Numerical Initial Value Problems in Ordinary Differential Equations* (Englewood Cliffs, NJ: Prentice-Hall)
-

-46-

- [42] Haile J M 1992 *Molecular Dynamics Simulation: Elementary Methods* (New York: Wiley)
- [43] Rapaport D C 1995 *The Art of Molecular Dynamics Simulation* (Cambridge: Cambridge University Press)
- [44] Verlet L 1968 Computer experiments on classical fluids. II. Equilibrium correlation functions *Phys. Rev.* **165** 201–14
- [45] Hockney R W and Eastwood J W 1988 *Computer Simulations Using Particles* (Bristol: Adam Hilger)
- [46] Swope W C, Andersen H C, Berens P H and Wilson K R 1982 A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: application to small water clusters *J. Chem. Phys.* **76** 637–49
- [47] Tuckerman M, Berne B J and Martyna G J 1992 Reversible multiple time scale molecular-dynamics *J. Chem. Phys.* **97** 1990–2001
- [48] Trotter H F 1959 On the product of semi-groups of operators *Proc. Am. Math. Soc.* **10** 545–51
- [49] Procacci P, Darden T A, Paci E and Marchi M 1997 ORAC: a molecular dynamics program to simulate complex molecular systems with realistic electrostatic interactions *J. Comput. Chem.* **18** 1848–62
- [50] Deuffhard P, Hermans J, Leimkuhler B, Mark A E, Reich S and Skeel R D (ed) 1998 *Computational Molecular Dynamics: Challenges, Methods, Ideas* vol 4 *Lecture Notes in Computational Science and Engineering* (Berlin: Springer)
- [51] Schlick T, Mandziuk M, Skeel R D and Srinivas K 1998 Nonlinear resonance artifacts in molecular dynamics simulations *J. Comput. Phys.* **140** 1–29
- [52] de Leeuw S W, Perram J W and Petersen H G 1990 Hamilton equations for constrained dynamic systems *J. Stat. Phys.* **61** 1203–22
- [53] Ryckaert J-P, Ciccotti G and Berendsen H J C 1977 Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes *J. Comput. Phys.* **23** 327–41
- [54] Ciccotti G, Ferrario M and Ryckaert J-P 1982 Molecular dynamics of rigid systems in cartesian coordinates. A general formulation *Mol. Phys.* **47** 1253–64
- [55] Ciccotti G and Ryckaert J P 1986 Molecular dynamics simulation of rigid molecules *Comput. Phys. Rep.* **4** 345–92
- [56] Andersen H C 1983 RATTLE: a 'velocity' version of the SHAKE algorithm for molecular dynamics calculations *J. Comput. Phys.* **52** 24–34
- [57] Frenkel D and Smit B 1996 *Understanding Molecular Simulation: From Algorithms to Applications* (San Diego: Academic)
- [58] van Gunsteren W F 1980 Constrained dynamics of flexible molecules *Mol. Phys.* **40** 1015–19
- [59] Fixman M 1974 Classical statistical mechanics of constraints: a theorem and application to polymers *Proc. Natl Acad. Sci.* **71** 3050–3
-

-47-

- [60] Knuth D 1973 *The Art of Computer Programming* 2nd edn (Reading, MA: Addison-Wesley)
- [61] Barker J A and Watts R O 1973 Monte Carlo studies of the dielectric properties of water-like models *Mol. Phys.* **26** 789–92
- [62] Neumann M and Steinhauser O 1980 The influence of boundary conditions used in machine simulations on the structure of polar systems *Mol. Phys.* **39** 437–54
- [63] Patey G N, Levesque D and Weis J J 1982 On the theory and computer simulation of dipolar fluids *Mol. Phys.* **45** 733–46
- [64] de Leeuw S W, Perram J W and Smith E R 1980 Simulation of electrostatic systems in periodic boundary conditions. I. Lattice sums and dielectric constant *Phys. Rev.S A* **373** 27–56
- [65] Felderhof B U 1980 Fluctuation theorems for dielectrics with periodic boundary conditions *Physica A* **101** 275–82
- [66] Perram J W, Petersen H G and DeLeeuw S W 1988 An algorithm for the simulation of condensed matter which grows as the $3/2$ power of the number of particles *Mol. Phys.* **65** 875–93
- [67] Neumann M, Steinhauser O and Pawley G S 1984 Consistent calculation of the static and frequency-dependent dielectric constant in computer simulations *Mol. Phys.* **52** 97–113
- [68] Gray C G, Sainger Y S, Joslin C G, Cummings P T and Goldman S 1986 Computer simulation of dipolar fluids. Dependence of the dielectric constant on system size: a comparative study of Ewald sum and reaction field approaches *J. Chem. Phys.* **85** 1502–4
- [69] Gil-Villegas A, McGrother S C and Jackson G 1997 Reaction-field and Ewald summation methods in Monte Carlo simulations of dipolar liquid crystals *Mol. Phys.* **92** 723–34
- [70] Greengard L and Rokhlin V 1987 A fast algorithm for particle simulations *J. Comput. Phys.* **73** 325–48
- [71] Darden T, York D and Pedersen L 1993 Particle mesh Ewald—an $N \cdot \log(N)$ method for Ewald sums in large systems *J. Chem. Phys.* **98** 10089–92
- [72] Essmann U, Perera L, Berkowitz M L, Darden T, Lee H and Pedersen L G 1995 A smooth particle mesh Ewald method *J. Chem. Phys.* **103** 8577–93
- [73] Procacci P, Marchi M and Martyna G J 1998 Electrostatic calculations and multiple time scales in molecular dynamics simulation of flexible molecular systems *J. Chem. Phys.* **108** 8799–803
- [74] Feller W 1957 *An Introduction to Probability Theory and its Applications* 2nd edn, vol 1 (New York: Wiley)
- [75] Manousiouthakis V I and Deem M W 1999 Strict detailed balance is unnecessary in Monte Carlo simulation *J. Chem. Phys.* **110** 2753–6
- [76] Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H and Teller E 1953 Equation of state calculations by fast computing machines *J. Chem. Phys.* **21** 1087–92
- [77] Owicki J C and Scheraga H A 1977 Preferential sampling near solutes in Monte Carlo calculations on dilute solutions *Chem. Phys. Lett.* **47** 600–2

- [78] Eppenga R and Frenkel D 1984 Monte Carlo study of the isotropic and nematic phases of infinitely thin hard platelets *Mol. Phys.* **52** 1303–34
- [79] Andersen H C 1980 Molecular dynamics simulations at constant pressure and/or temperature *J. Chem.*

Phys. **72** 2384–93

- [80] Nosé S 1984 A molecular dynamics method for simulations in the canonical ensemble *Mol. Phys.* **52** 255–68
- [81] Hoover W G 1985 Canonical dynamics: equilibrium phase-space distributions *Phys. Rev. A* **31** 1695–7
- [82] Martyna G J, Klein M L and Tuckerman M 1992 Nosé–Hoover chains: the canonical ensemble via continuous dynamics *J. Chem. Phys.* **97** 2635–43
- [83] Tobias D J, Martyna G J and Klein M L 1993 Molecular dynamics simulations of a protein in the canonical ensemble *J. Phys. Chem.* **97** 12959–66
- [84] Hoover W G, Ladd A J C and Moran B 1982 High strain rate plastic flow studied via nonequilibrium molecular dynamics *Phys. Rev.L* **48** 1818–20
- [85] Ladd A J C and Hoover W G 1983 Plastic-flow in close-packed crystals via non-equilibrium molecular-dynamics *Phys. Rev. B* **28** 1756–62
- [86] Evans D J 1983 Computer experiment for nonlinear thermodynamics of Couette flow *J. Chem. Phys.* **78** 3297–302
- [87] Nosé S 1984 A unified formulation of the constant-temperature molecular dynamics methods *J. Chem. Phys.* **81** 511–19
- [88] Evans D J and Morriss G P 1983 The isothermal isobaric molecular dynamics ensemble *Phys. Lett. A* **98** 433–6
- [89] Parrinello M and Rahman A 1980 Crystal structure and pair potentials: a molecular dynamics study *Phys. Rev.L* **45** 1196–9
- [90] Parrinello M and Rahman A 1981 Polymorphic transitions in single crystals: a new molecular dynamics method *J. Appl. Phys.* **52** 7182–90
- [91] Martyna G J, Tuckerman M, Tobias D J and Klein M L 1996 Explicit reversible integrators for extended systems dynamics *Mol. Phys.* **87** 1117–57
- [92] Hoover W G, Ross M, Johnson K W, Henderson D, Barker J A and Brown B C 1970 Soft sphere equation of state *J. Chem. Phys.* **52** 4931–41
- [93] Kofke D A and Cummings P T 1997 Quantitative comparison and optimization of methods for evaluating the chemical potential by molecular simulation *Mol. Phys.* **92** 973–96
- [94] Bennett C H 1976 Efficient estimation of free energy differences from Monte Carlo data *J. Comput. Phys.* **22** 245–68
- [95] Torrie G M and Valleau J P 1977 Nonphysical sampling distributions in Monte Carlo free energy estimation: umbrella sampling *J. Comput. Phys.* **23** 187–99

- [96] Berg B A and Neuhaus T 1992 Multicanonical ensemble—a new approach to simulate 1st-order phase transitions *Phys. Rev.L* **68** 9–12
- [97] Lee J 1993 New Monte Carlo algorithm—entropic sampling *Phys. Rev.L* **71** 211–14
- [98] Marinari E and Parisi G 1992 Simulated tempering: a new Monte Carlo scheme *Europhys. Lett.* **19** 451–8
- [99] Lyubartsev A P, Martsinovski A A, Shevkunov S V and Vorontsov-Velyaminov P N 1992 New approach to Monte Carlo calculation of the free-energy—method of expanded ensembles *J. Chem. Phys.* **96**

- [100] Ferrenberg A M and Swendsen R H 1989 Optimized Monte Carlo data analysis *Phys. Rev.L* **63** 1195–8
- [101] Swendsen R H 1993 Modern methods of analyzing Monte Carlo computer simulations *Physica A* **194** 53–62
- [102] Ferrenberg A M, Landau D P and Swendsen R H 1995 Statistical errors in histogram reweighting *Phys. Rev. E* **51** 5092–100
- [103] Dònweg B 1996 Simulation of phase transitions: critical phenomena *Monte Carlo and Molecular Dynamics of Condensed Matter Systems* vol 49, ed K Binder and G Ciccotti (Bologna: Italian Physical Society) pp 215–54
- [104] Reiss H, Frisch H L and Lebowitz J L 1959 Statistical mechanics of rigid spheres *J. Chem. Phys.* **31** 369–80
- [105] Carnahan N F and Starling K E 1969 Equation of state for nonattracting rigid spheres *J. Chem. Phys.* **51** 635–6
- [106] Deitrick G L, Scriven L E and Davis H T 1989 Efficient molecular simulation of chemical potentials *J. Chem. Phys.* **90** 2370–85
- [107] Yoon K, Chae D G, Ree T and Ree F H 1981 Computer simulation of a grand canonical ensemble of rodlike molecules *J. Chem. Phys.* **74** 1412–23
- [108] Lee Y S, Chae D G, Ree T and Ree F H 1981 Computer simulations of a continuum system of molecules with a hard-core interaction in the grand canonical ensemble *J. Chem. Phys.* **74** 6881–7
- [109] Swope W C and Andersen H C 1995 A computer simulation method for the calculation of chemical potentials of liquids and solids using the bicanonical ensemble *J. Chem. Phys.* **102** 2851–63
- [110] Kofke D A and Glandt E D 1988 Monte Carlo simulation of multicomponent equilibria in a semigrand canonical ensemble *Mol. Phys.* **64** 1105–31
- [111] Mon K K and Griffiths R B 1985 Chemical potential by gradual insertion of a particle in Monte Carlo simulation *Phys. Rev. A* **31** 956–9
- [112] Nezbeda I and Kolafa J 1991 A new version of the insertion particle method for determining the chemical potential by Monte Carlo simulation *Mol. Simul.* **5** 391–403
- [113] Attard P 1993 Simulation of the chemical potential and the cavity free energy of dense hard-sphere fluids *J. Chem. Phys.* **98** 2225–31

- [114] Camp P J, Mason C P, Allen M P, Khare A A and Kofke D A 1996 The isotropic–nematic transition in uniaxial hard ellipsoid fluids: coexistence data and the approach to the Onsager limit *J. Chem. Phys.* **105** 2837–49
- [115] Hoover W G and Ree F H 1967 Use of computer experiments to locate the melting transition and calculate the entropy in the solid phase *J. Chem. Phys.* **47** 4873–8
- [116] Hoover W G and Ree F H 1968 Melting transition and communal entropy for hard spheres *J. Chem. Phys.* **49** 3609–17
- [117] Frenkel D, Mulder B M and McTague J P 1984 Phase-diagram of a system of hard ellipsoids *Phys. Rev.L* **52** 287–90
- [118] Meijer E J, Frenkel D, LeSar R A and Ladd A J C 1990 Location of melting point at 300 K of nitrogen by Monte Carlo simulation *J. Chem. Phys.* **92** 7570–5

- [119] Lovett R 1995 Can a solid be turned into a gas without passing through a first order phase transition? *Observation, Prediction and Simulation of Phase Transitions in Complex Fluids* vol 460 *NATO ASI Series C* ed M Baus, L F Rull and J-P Ryckaert (Dordrecht: Kluwer) pp 641–54
- [120] Sheu S-Y, Mou C-Y and Lovett R 1995 How a solid can be turned into a gas without passing through a first-order phase transformation *Phys. Rev. E* **51** R3795–8
- [121] Rosenbluth M N and Rosenbluth A W 1995 Monte Carlo calculation of the average extension of molecular chains *J. Chem. Phys.* **23** 356–9
- [122] Harris J and Rice S A 1988 A lattice model of a supported monolayer of amphiphile molecules—Monte Carlo simulations *J. Chem. Phys.* **88** 1298–306
- [123] Siepmann J I and Frenkel D 1992 Configurational bias Monte Carlo—a new sampling scheme for flexible chains *Mol. Phys.* **75** 59–70
- [124] Frenkel D, Mooij G A M and Smit B 1992 Novel scheme to study structural and thermal properties of continuously deformable molecules *J. Phys.: Condens. Matter* **4** 3053–76
- [125] de Pablo J J, Laso M and Suter U W 1992 Simulation of polyethylene above and below the melting point *J. Chem. Phys.* **96** 2394–403
- [126] Smit B, Loyens L D J C and Verbist G L M M 1997 Simulation of adsorption and diffusion of hydrocarbons in zeolites *Faraday Disc. Chem. Soc.* **106** 93–104
- [127] Vlucht T J H, Krishna R and Smit B 1999 Molecular simulations of adsorption isotherms for linear and branched alkanes and their mixtures in silicalite *J. Phys. Chem. B* **103** 1102–18
- [128] Frenkel D 1986 Free-energy computation and first-order phase transitions *Molecular Dynamics Simulation of Statistical Mechanical Systems* ed G Ciccotti and W G Hoover (Amsterdam: North-Holland) pp 151–88
- [129] Smit B 1993 Computer simulations in the Gibbs ensemble *Computer Simulation in Chemical Physics* vol 397 *NATO ASI Series C* ed M P Allen and D J Tildesley (Dordrecht: Kluwer) pp 173–209

- [130] Frenkel D 1995 Numerical techniques to study complex liquids *Observation, Prediction and Simulation of Phase Transitions in Complex Fluids* vol 460 *NATO ASI Series C* ed M Baus, L F Rull and J-P Ryckaert (Dordrecht: Kluwer) pp 357–419
- [131] Panagiotopoulos A Z 1995 Gibbs ensemble techniques *Observation, Prediction and Simulation of Phase Transitions in Complex Fluids* ed M Baus, L F Rull and J-P Ryckaert, vol 460 *NATO ASI Series C* (Dordrecht: Kluwer) pp 463–501
- [132] Privman V (ed) 1990 *Finite Size Scaling and Numerical Simulation of Statistical Systems* (Singapore: World Scientific)
- [133] Binder K 1995 Introduction *The Monte Carlo Method in Condensed Matter Physics* vol 71 *Topics in Applied Physics* ed K Binder (Berlin: Springer) pp 1–22
- [134] Siepmann J I, McDonald I R and Frenkel D 1992 Finite-size corrections to the chemical potential *J. Phys.: Condens. Matter* **4** 679–91
- [135] Cardy J L (ed) *Finite-Size Scaling* vol 2 *Current Physics—Sources and Comments* (Amsterdam: North-Holland)
- [136] Binder K 1981 Finite size scaling analysis of Ising-model block distribution-functions *Z. Phys. B. Condens. Matter.* **43** 119–40
- [137] Bruce A D 1981 Probability density functions for collective coordinates in Ising-like systems *J. Phys. C: Solid State Phys.* **14** 3667–88
- [138] Nicolaides D and Bruce A D 1988 Universal configurational structure in two-dimensional scalar models *J.*

- [139] Nicolaides D and Evans R 1989 Nature of the prewetting critical-point *Phys. Rev.L* **63** 778–81
- [140] Bruce A D and Wilding N B 1992 Scaling fields and universality of the liquid-gas critical point *Phys. Rev.L* **68** 193–6
- [141] Wilding N B and Bruce A D 1992 Density fluctuations and field mixing in the critical fluid *J. Phys.: Condens. Matter* **4** 3087–108
- [142] Wood W W 1968 Monte Carlo studies of simple liquid models *Physics of Simple Liquids* ed H N V Temperley, J S Rowlinson and G S Rushbrooke (Amsterdam: North Holland) chapter 5, pp 115–230
- [143] Milchev A, Binder K and Heermann D W 1986 Fluctuations and lack of self-averaging in the kinetics of domain growth *Z. Phys. B. Condens. Matter.* **63** 521–35
- [144] Challa M S S, Landau D P and Binder K 1986 Finite-size effects at temperature-driven 1st-order transitions *Phys. Rev. B* **34** 1841–52
- [145] Brown F R and Yegulalp A 1991 Microcanonical simulation of 1st-order phase transitions in finite volumes *Phys. Lett. A* **155** 252–6
- [146] Wood W W and Jacobson J D 1957 Preliminary results from a recalculation of the Monte Carlo equation of state of hard spheres *J. Chem. Phys.* **27** 1207–8
-

- [147] Alder B J and Wainwright T E 1957 Phase transition for a hard sphere system *J. Chem. Phys.* **27** 1208–9
- [148] Alder B J and Wainwright T E 1962 Phase transition in elastic disks *Phys. Rev.* **127** 359–61
- [149] Mayer J E and Wood W W 1965 Interfacial tension effects in finite periodic two-dimensional systems *J. Chem. Phys.* **42** 4268–74
- [150] Wood W W 1968 Monte Carlo calculations for hard disks in the isothermal–isobaric ensemble *J. Chem. Phys.* **48** 415–34
- [151] Binder K and Landau D P 1984 Finite size scaling at 1st-order phase transitions *Phys. Rev. B* **30** 1477–85
- [152] Binder K and Heermann D W 1988 *Monte Carlo Simulation in Statistical Physics* vol 80 *Solid State Sciences* (Berlin: Springer)
- [153] Ferrenberg A M and Swendsen R H 1988 New Monte-Carlo technique for studying phase-transitions *Phys. Rev.L* **61** 2635–8
- [154] Ferrenberg A M 1989 Addition *Phys. Rev.L* **63** 1658
- [155] Lee J and Kosterlitz J M 1990 New numerical method to study phase transitions *Phys. Rev.L* **65** 137–40
- [156] Zhang Z, Mouritsen O G and Zuckermann M J 1992 Weak first-order orientational transition in the Lebwohl–Lasher model of liquid crystals *Phys. Rev.L* **69** 2803–6
- [157] Zhang Z, Zuckermann M J and Mouritsen O G 1993 Phase transition and director fluctuations in the 3-dimensional Lebwohl–Lasher model of liquid crystals *Mol. Phys.* **80** 1195–221
- [158] Fabbri U and Zannoni C 1986 A Monte Carlo investigation of the Lebwohl–Lasher lattice model in the vicinity of its orientational phase transition *Mol. Phys.* **58** 763–88
- [159] van Duijneveldt J S and Frenkel D 1992 Computer simulation study of free-energy barriers in crystal nucleation *J. Chem. Phys.* **96** 4655–68
- [160] Lynden-Bell R M, van Duijneveldt J S and Frenkel D 1993 Free-energy changes on freezing and melting ductile metals *Mol. Phys.* **80** 801–14

- [161] Orkoulas G and Panagiotopoulos A Z 1999 Phase behaviour of the restricted primitive model and square-well fluids from Monte Carlo simulations in the grand canonical ensemble *J. Chem. Phys.* **110** 1581–90
- [162] Panagiotopoulos A Z 1987 Direct determination of phase coexistence properties of fluids by Monte Carlo simulation in a new ensemble *Mol. Phys.* **61** 813–26
- [163] Panagiotopoulos A Z, Quirke N, Stapleton M and Tildesley D J 1988 Phase equilibria by simulation in the Gibbs ensemble. Alternative derivation, generalization and application to mixture and membrane equilibria *Mol. Phys.* **63** 527–45
- [164] Panagiotopoulos A Z 1992 Direct determination of fluid phase equilibria by simulation in the Gibbs ensemble: a review *Mol. Simul.* **9** 1–23
-

-53-

- [165] Panagiotopoulos A Z 1994 Molecular simulation of phase equilibria *Supercritical Fluids—Fundamentals for Application NATO ASI Series E* ed E Kiran and J M H Levelt Sengers (Dordrecht: Kluwer)
- [166] Panagiotopoulos A Z 1987 Adsorption and capillary condensation of fluids in cylindrical pores by Monte Carlo simulation in the Gibbs ensemble *Mol. Phys.* **62** 701–19
- [167] Siepmann J I, Karaborni S and Smit B 1993 Simulating the critical behaviour of complex fluids *Nature* **365** 330–2
- [168] Smit B, Karaborni S and Siepmann J I 1995 Computer simulations of vapor–liquid phase equilibria of *n*-alkanes *J. Chem. Phys.* **102** 2126–40
- [169] Panagiotopoulos A Z 1989 Exact calculations of fluid-phase equilibria by Monte Carlo simulation in a new statistical ensemble *Int. J. Thermophys.* **10** 447–57
- [170] de Pablo J J and Prausnitz J M 1989 Phase equilibria for fluid mixtures from Monte Carlo simulation *Fluid Phase Equilibria* **53** 177–89
- [171] Escobedo F A and de Pablo J J 1996 Expanded grand canonical and Gibbs ensemble Monte Carlo simulation of polymers *J. Chem. Phys.* **105** 4391–4
- [172] Nath S K, Escobedo F A and de Pablo J J 1998 On the simulation of vapor–liquid equilibria for alkanes *J. Chem. Phys.* **108** 9905–11
- [173] Möller D and Fischer J 1990 Vapour liquid equilibrium of a pure fluid from test particle method in combination with *NpT* molecular dynamics simulations *Mol. Phys.* **69** 463–73
- [174] Möller D 1992 Correction *Mol. Phys.* **75** 1461–2
- [175] Lotfi A, Vrabec J and Fischer J 1992 Vapour liquid equilibria of the Lennard-Jones fluid from the *NpT* plus test particle method *Mol. Phys.* **76** 1319–33
- [176] Boda D, Liszi J and Szalai I 1995 An extension of the *NpT* plus test particle method for the determination of the vapour-liquid equilibria of pure fluids *Chem. Phys. Lett.* **235** 140–5
- [177] Kofke D A 1993 Gibbs–Duhem integration: a new method for direct evaluation of phase coexistence by molecular simulation *Mol. Phys.* **78** 1331–6
- [178] Kofke D A 1993 Direct evaluation of phase coexistence by molecular simulation via integration along the saturation line *J. Chem. Phys.* **98** 4149–62
- [179] Agrawal R and Kofke D A 1995 Thermodynamic and structural properties of model systems at solid–fluid coexistence. II. Melting and sublimation of the Lennard-Jones system *Mol. Phys.* **85** 43–59
- [180] Agrawal R and Kofke D A 1995 Thermodynamic and structural properties of model systems at solid–fluid coexistence. I. Fcc and bcc soft spheres *Mol. Phys.* **85** 23–42
- [181] Bolhuis P G and Kofke D A 1996 Monte Carlo study of freezing of polydisperse hard spheres *Phys. Rev. E* **54**

-
- [182] Kofke D A and Bolhuis P G 1999 Freezing of polydisperse hard spheres *Phys. Rev. E* **59** 618-22
- [183] Holcomb C D, Clancy P and Zollweg J A 1993 A critical study of the simulation of the liquid-vapour interface of a Lennard-Jones fluid *Mol. Phys.* **78** 437-59
- [184] Alejandre J, Tildesley D J and Chapela G A 1995 Molecular dynamics simulation of the orthobaric densities and surface tension of water *J. Chem. Phys.* **102** 4574-83
- [185] Chandler D 1978 Statistical mechanics of isomerisation dynamics in liquids and the transition state approximation *J. Chem. Phys.* **68** 2959-70
- [186] Carter E A, Ciccotti G, Hynes J T and Kapral R 1989 Constrained reaction coordinate dynamics for the simulation of rare events *Chem. Phys. Lett.* **156** 472-7
- [187] Ruiz-Montero M J, Frenkel D and Brey J J 1997 Efficient schemes to compute diffusive barrier crossing rates *Mol. Phys.* **90** 925-41
- [188] Ciccotti G and Ferrario M 1998 Constrained and nonequilibrium molecular dynamics *Classical and Quantum Dynamics in Condensed Phase Simulations* ed B J Berne, G Ciccotti and D F Coker (Singapore: World Scientific) pp 157-77
- [189] ten Wolde P R and Frenkel D 1997 Enhancement of protein crystal nucleation by critical density fluctuations *Science* **277** 1975-8
- [190] Chandler D 1998 Finding transition pathways: throwing ropes over rough mountain passes, in the dark *Classical and Quantum Dynamics in Condensed Phase Simulations* (Singapore: World Scientific) pp 51-66
- [191] Dellago C, Bolhuis P G, Csajka F S and Chandler D 1998 Transition path sampling and the calculation of rate constants *J. Chem. Phys.* **108** 1964-77
- [192] Pratt L R 1986 A statistical method for identifying transition states in high dimensional problems *J. Chem. Phys.* **85** 5045-8
- [193] Feynman R P and Hibbs A R 1965 *Quantum Mechanics and Path Integrals* (New York: McGraw-Hill)
- [194] Berne B J, Ciccotti G and Coker D F (ed) 1998 *Classical and Quantum Dynamics in Condensed Phase Simulations* (Singapore: World Scientific)
- [195] De Raedt H 1996 Quantum theory *Monte Carlo and Molecular Dynamics of Condensed Matter Systems* ed K Binder and G Ciccotti (Bologna: Italian Physical Society) pp 401-42
- [196] Ceperley D M 1996 Path integral Monte Carlo for fermions *Monte Carlo and Molecular Dynamics of Condensed Matter Systems* vol 49, ed K Binder and E G Ciccotti (Bologna: Italian Physical Society) pp 443-82
- [197] Tuckerman M E and Hughes A 1998 Path integral molecular dynamics: a computational approach to quantum statistical mechanics *Classical and Quantum Dynamics in Condensed Phase Simulations* ed B J Berne, G Ciccotti and D F Coker (Singapore: World Scientific) pp 311-57
- [198] Herman M F, Bruskin E J and Berne B J 1982 On path integral Monte Carlo simulations *J. Chem. Phys.* **76** 5150-5
-

- [199] Tuckerman M, Berne B J, Martyna G J and Klein M L 1993 Efficient molecular dynamics and hybrid Monte Carlo algorithms for path integrals *J. Chem. Phys.* **99** 2796–808
- [200] Sprik M, Klein M L and Chandler D 1985 Staging—a sampling technique for the Monte-Carlo evaluation of path-integrals *Phys. Rev. B* **31** 4234–44
- [201] Berne B J and Thirumalai D 1986 On the simulation of quantum systems—path integral methods *Ann. Rev. Phys. Chem.* **37** 401–24
- [202] Car R and Parrinello M 1985 Unified approach for molecular dynamics and density-functional theory *Phys. Rev. L* **55** 2471–4
- [203] Remler D K and Madden P A 1990 Molecular dynamics without effective potentials via the Car–Parrinello approach *Mol. Phys.* **70** 921–66
- [204] Galli G and Pasquarello A 1993 First-principles molecular dynamics *Computer Simulation in Chemical Physics* vol 397 *NATO ASI Series C* ed M P Allen and D J Tildesley (Dordrecht: Kluwer) pp 261–313
- [205] Car R 1996 Molecular dynamics from first principles *Monte Carlo and Molecular Dynamics of Condensed Matter Systems* vol 49 ed K Binder and G Ciccotti (Bologna: Italian Physical Society) pp 601–34
- [206] Sprik M 1998 Density functional techniques for simulation of chemical reactions *Classical and Quantum Dynamics in Condensed Phase Simulations* ed B J Berne, G Ciccotti and D F Coker (Singapore: World Scientific) pp 285–309
- [207] Hohenberg P C and Kohn W 1964 Inhomogeneous electron gas *Phys. Rev. B* **136** 864–71
- [208] Vanderbilt D 1990 Soft self-consistent pseudopotentials in a generalized eigenvalue formalism *Phys. Rev. B* **41** 7892–5
- [209] Ceperley D M and Alder B J 1980 Ground state of the electron gas by a stochastic method *Phys. Rev. L* **45** 566–9
- [210] Alfe D and Gillan M J 1998 First-principles simulations of liquid Fe–S under Earth’s core conditions *Phys. Rev. B* **58** 8248–56
- [211] de Wijs G A, Kresse G, Vožadlo L, Dobson D, Alfe D, Gillan M J and Price G D 1998 The viscosity of liquid iron at the physical conditions of the Earth’s core *Nature* **392** 805–7
- [212] Löwen H, Hansen J-P and Madden P A 1993 Nonlinear counterion screening in colloidal suspensions *J. Chem. Phys.* **98** 3275–89
- [213] Smith W 1991 Molecular dynamics on hypercube parallel computers *Comput. Phys. Commun.* **62** 229–48
- [214] Smith W 1992 A replicated data molecular dynamics strategy for the parallel Ewald sum *Comput. Phys. Commun.* **67** 392–406
- [215] Wilson M R, Allen M P, Warren M A, Sauron A and Smith W 1997 Replicated data and domain decomposition molecular dynamics techniques for the simulation of anisotropic potentials *J. Comput. Chem.* **18** 478–88

- [216] Rapaport D C 1991 Multi-million particle molecular dynamics II. Design considerations for distributed processing *Comput. Phys. Commun.* **62** 217–28
- [217] Esselink K, Smit B and Hilbers P A J 1993 Efficient parallel implementation of molecular dynamics on a toroidal network. I. Parallelizing strategy *J. Comput. Phys.* **106** 101–7
- [218] Beazley D M and Lomdahl P S 1993 Message-passing multi-cell molecular dynamics on the Connection Machine 5 *Parallel Comput.* **20** 173–95
- [219] Holian B L and Lomdahl P S 1998 Plasticity induced by shock waves in nonequilibrium molecular-dynamics simulations *Science* **280** 2085–8

- [220] Hilbers P A J and Esselink K 1992 Parallel molecular dynamics *Parallel Computing: From Theory to Sound Practice* ed W Joosen and E Milgrom (Amsterdam: IOS Press) pp 288–99
- [221] Hilbers P A J and Esselink K 1993 Parallel computing and molecular dynamics simulations *Computer Simulation in Chemical Physics* vol 397 *NATO ASI Series C* ed M P Allen and D J Tildesley (Dordrecht: Kluwer) pp 473–95
- [222] Heermann D W and Burkitt A N 1995 Parallel algorithms for statistical physics problems *The Monte Carlo Method in Condensed Matter Physics* vol 71 *Topics in Applied Physics* ed K Binder (Berlin: Springer) pp 53–74
- [223] Heermann D W 1996 Parallelization of computational physics problems *Monte Carlo and Molecular Dynamics of Condensed Matter Systems* vol 49, ed K Binder and G Ciccotti (Bologna: Italian Physical Society) pp 887–906
- [224] Gubbins K E and Quirke N 1996 *Molecular Simulation and Industrial Applications* (Reading: Gordon and Breach)
- [225] Lewis D W, Sankar G, Wyles J K, Thomas J M, Catlow C R A and Willock D J 1997 Synthesis of a small-pore microporous material using a computationally designed template *Angew. Chem. Int. Ed. Engl.* **36** 2675–7
- [226] Groot R D and Warren P B 1997 Dissipative particle dynamics: bridging the gap between atomistic and mesoscopic simulation *J. Chem. Phys.* **107** 4423–35
- [227] Chen S and Doolen G D 1998 Lattice Boltzmann method for fluid flows *Ann. Rev. Fluid Mech.* **30** 329–64

FURTHER READING

Allen M P and Tildesley D J 1989 *Computer Simulation of Liquids* (Oxford: Clarendon)

A comprehensive introduction to the field, covering statistical mechanics, basic Monte Carlo, and molecular dynamics methods, plus some advanced techniques, including computer code.

Frenkel D and Smit B 1996 *Understanding Molecular Simulation: From Algorithms to Applications* (San Diego: Academic)

A comprehensive and up-to-date introduction to the ideas of molecular dynamics and Monte Carlo, with statistical mechanical background, advanced techniques and case studies, supported by a Web page for software download.

Binder K and Ciccotti G (ed) 1995 *Monte Carlo and Molecular Dynamics of Condensed Matter Systems: Proc. Euroconference (Como, Italy, 3–28 July 1995)* vol 49 (Bologna: Italian Physical Society)

One of the most comprehensive and up-to-date summer school proceedings, with contributions from many of the world's leading experts. Almost every aspect of molecular simulation is covered.

Berne B J, Ciccotti G and Coker D F (ed) 1998 *Classical and Quantum Dynamics in Condensed Phase Simulations: Proc. Euroconference (Lerici, Italy, 7–18 July, 1997)* (Singapore: World Scientific)

A substantial summer school proceedings, concentrating on modern techniques for studying rare events and quantum mechanical phenomena.

Rapaport D C 1995 *The Art of Molecular Dynamics Simulation* (Cambridge: Cambridge University Press)

A detailed and comprehensive book on molecular dynamics, with a tutorial approach plus many examples of

computer code, supported by a Web page for software download.

Binder K and Heermann D W 1997 *Monte Carlo Simulation in Statistical Physics* vol 80 *Solid State Sciences* 3rd edn (Berlin: Springer)

A compact and readable introduction to Monte Carlo, with examples and exercises, plus useful pointers to the literature on lattice models.

-1-

B3.4 Quantum dynamics and spectroscopys

Sybil M Anderson, Rovshan G Sadygov and Daniel Neuhauser

B3.4.1 INTRODUCTION

The study of quantum effects associated with nuclear motion is a distinct field of chemistry, known as quantum molecular dynamics. This section gives an overview of the methodology of the field; for further reading, consult [[1](#), [2](#), [3](#), [4](#) and [5](#)].

The importance of non-classical behaviour in molecular dynamics has its origins in the inherently quantal nature of atomic motion. The de Broglie wavelengths of atoms are small but non-vanishing. For example, hydrogen has a de Broglie wavelength that can be as high as $\approx 1.0 \text{ \AA}$ at room temperature. Specific quantum effects in atomic and molecular motion include: zero-point motion, most notably for hydrogen which has a zero-point energy of $\sim 10 \text{ kcal mol}^{-1}$ in many of its covalent interactions; interference resonances, which are spikes in reaction or pre-dissociation probabilities associated with quasi-bound states of molecules [[6](#), [7](#), [8](#) and [9](#)]; and tunnelling of nuclei which is important in many catalytic reactions [[10](#), [11](#), [12](#), [13](#) and [14](#)].

In its most fundamental form, quantum molecular dynamics is associated with solving the Schrödinger equation for molecular motion, whether using a single electronic surface (as in the Born–Oppenheimer approximation— [section B3.4.2](#) or with the inclusion of multiple electronic states, which is important when discussing non-adiabatic effects, in which the electronic state is changed [[15](#), [16](#), [17](#), [18](#) and [19](#)].

[Section B3.4.3](#), [Section B3.4.4](#), [Sections B3.4.5](#) describe methods of solving the Schrödinger equation for scattering events. [Sections B3.4.6](#) and [Sections B3.4.7](#) proceed to discuss photo-dissociation and bound states.

As these methods are explored, it is quickly realized that the numerical effort in the theoretical description grows prohibitively large with the number of atoms in a molecule. The difficulty lies in precisely what makes molecular motion fundamentally quasi-classical, i.e. the large molecular masses (relative to the mass of the electron). Consequently, a molecular wavefunction has many oscillations and is difficult to model numerically. There have been many attempts at developing alternate approaches for representing quantum wavefunctions and observables without the use of large grids or basis sets, ranging from approximations to path-integral descriptions. The basics of these approaches are described in [Sections B3.4.8](#). Later, [Sections B3.4.9](#) describes the issues involved in the study of non-adiabatic phenomena.

Finally, [Sections B3.4.10](#) touches on the application of quantum molecular dynamics to a very exciting field: laser interactions with molecules. This field presents, in principle, the opportunity to influence chemistry by lasers rather than to simply observe it.

The scope of this section restricts the discussion. One omitted topic is the collision and interaction of molecules with surfaces (see [[20](#), [21](#)] and [section A3.9](#)). This topic connects quantum molecular dynamics in gas and condensed phases. Depending on the time scales of the interaction of a molecule with a surface, the

reactions are similar to those

-2-

in one phase or the other. If the collision is fast, so that one may neglect the motion of the surface molecules and treat them as frozen, it is effectively a gas-phase reaction. On the other extreme, if the motion of the adsorbate is slow or comparable in time to the motion of surface and subsurface molecules, then the collision problem becomes very similar to the interaction of molecules in condensed phases. The latter is a subject of a separate [Sections C3.5](#).

Another modern and highly exciting topic, omitted here due to lack of space, is the motion of very cold molecules [[22](#), [23](#) and [24](#)], which can have de Broglie wavelengths that are as large or larger than the distances between the molecules. The simplest examples are essentially extensions of floppy van der Waals structures, but at the extreme, when the wavelength is extremely large and there are many molecules per molecular wavelength, one ends up with Bose–Einstein condensates (where the wavefunctions of the molecules coalesce to form one giant coherent molecular function) and even molecular lasers (i.e. lasers where the fundamental particles are atoms or molecules rather than photons [[25](#)]) can be made. [Sections C1.4](#) provides an overview of this new field.

As in any field, it is useful to clarify terminology. Throughout this section an ‘atom’ more specifically refers to its nuclear centre. Also, for most of the section the $\hbar=1$ convention is used. Finally, it should be noted that in the literature the label ‘quantum molecular dynamics’ is also sometimes used for a purely classical description of atomic motion under the potential created by the electronic distribution.

Finally, this section is related to several others, especially [Sections A3.11](#) on formal scattering, which should be carefully consulted.

B3.4.2 QUANTUM MOTION ON A SINGLE ELECTRONIC SURFACE

A corner-stone of a large portion of quantum molecular dynamics is the use of a single electronic surface. Since electrons are much lighter than nuclei, they typically adjust their wavefunction to follow the nuclei [[26](#)]. Specifically, if a collision is started in which the electrons are in their ground state, they typically remain in the ground state. An exception is non-adiabatic processes, which are discussed later in this section.

The single-surface assumption, known also as the Born–Oppenheimer approximation, implies that the nuclei are described by a single wavefunction ($\psi(\mathbf{x},t)$ where \mathbf{x} is a multi-dimensional vector describing the nuclear position). The time-dependent equation for the evolution of the wavefunction is simply

$$i\frac{\partial\psi}{\partial t} = \hat{H}\psi \equiv [\hat{K} + V(\mathbf{x})]\psi \quad (\text{B3.4.1})$$

where \hat{H} is the Hamiltonian governing the motion of the nuclei (or the atomic motion, as typically denoted), \hat{K} is the kinetic term (a sum of terms of the form $-(1/2m_j)(\partial^2/\partial\mathbf{x}_j^2)$ for each atom j) and V is the Born–Oppenheimer potential which is defined as the electronic ground-state energy for nuclear configuration $\bar{\mathbf{x}}$, including the nuclear–nuclear repulsion.

As a word of caution, the Born–Oppenheimer assumption is not universally valid. There are many reactions in which,

for example, non-adiabatic curve-crossing processes occur. In these cases, two electronic potentials—e.g., the ground state and an excited state—are locally equal to one another at some configuration. Curve-crossing effects are highly non-trivial and can affect the nuclear dynamics even at energies which are way below the curve crossing. These points are briefly discussed in [Sections B3.4.9](#).

B3.4.3 SCATTERING

B3.4.3.1 COLLINEAR MOTION

[Equation \(B3.4.1\)](#) is general and applies to both scattering and bound state spectroscopy. Scattering will be considered first. For simplicity, the discussion uses the collinear model for the $\mathbf{A} + \mathbf{BC} \rightarrow \mathbf{AB} + \mathbf{C}$ reaction (i.e. assuming all particles lie on a line). This model is easy to visualize and embodies most elements of three-dimensional (3D) scattering of larger molecules.

After removal of centre-of-mass motion, there are two independent distances which need to be considered for a collinear problem, r_{BA} ($\equiv r_{\text{B}} - r_{\text{A}}$, where r_{B} and r_{A} denote the positions of A and B on the line) and r_{CB} , which is similarly defined. Unfortunately, the kinetic energy is not conveniently described with these coordinates; therefore, alternate systems are used. The most convenient one is reactant Jacobi coordinates (r, R) , where $r = r_{\text{CB}}$, and R is the distance between A and the centre of mass of B and C [27]. In these coordinates, the kinetic energy gets a simple and separable form so that the Schrödinger equation is

$$i \frac{\partial \psi}{\partial t} = \left[-\frac{1}{2M} \frac{\partial^2}{\partial R^2} - \frac{1}{2\mu} \frac{\partial^2}{\partial r^2} + V(R, r) \right] \psi \quad (\text{B3.4.2})$$

where μ is the reduced mass associated with the BC vibration, and M is the mass associated with the A–BC motion.

[Figure B3.4.1](#) shows the potential surface for a simple collinear reaction, $\mathbf{D} + \mathbf{H}_2 \rightarrow \mathbf{HD} + \mathbf{H}$. The most notable aspect of this potential is the angle between the products and reactant arrangement. Below breakup ($\mathbf{D} + \mathbf{H}_2 \rightarrow \mathbf{D} + \mathbf{H} + \mathbf{H}$, which can occur only at several electronvolts above the reaction threshold), the potential surface has three relevant regions: the reactants ($\mathbf{D} + \mathbf{H}_2$) asymptote (arrangement) at large R and small r ; the products ($\mathbf{H} + \mathbf{DH}$) asymptote; and a strong interaction region where all three atoms are closely spaced. The strong-interaction region is extended over a region of $\approx 1 \text{ \AA} \times 1 \text{ \AA}$, containing several oscillations of the full (DH_2) wavefunction. Since the potential is not separable in R and r , upon reaction the particles would exchange vibrational and translational energy (and in the three-dimensional case, also rotational energy).

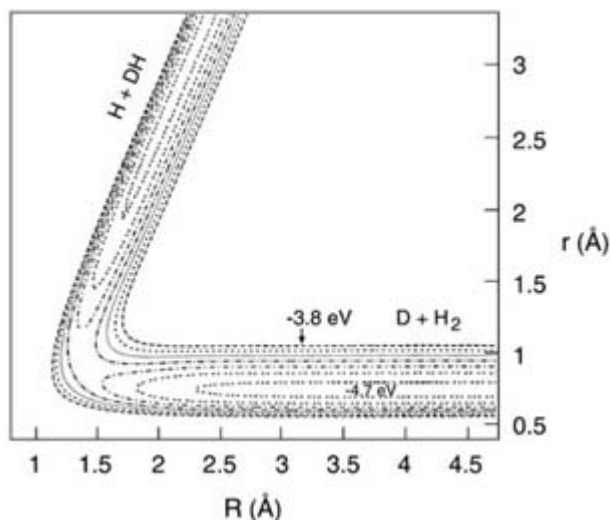


Figure B3.4.1. The potential surface for the collinear $D + H_2 \rightarrow DH + H$ reaction (this potential is the same as for $H + H_2 \rightarrow H_2 + H$, but to make the products and reactants identification clearer the isotopically substituted reaction is used). The $D + H_2$ reactant arrangement and the $DH + H$ product arrangement are denoted. The coordinates are r , the H_2 distance, and R , the distance between the D and the H_2 centre of mass. Distances are measured in angströms; the potential contours shown are 4.7 eV, -4.55 eV, . . ., -3.8 eV. (The potential energy is zero when the particles are far from each other. Only the first few contours are shown.) For reference, the zero-point energy for H_2 is -4.47 eV, i.e. 0.27 eV above the H_2 potential minimum (-4.74 eV); the room-temperature thermal kinetic energy is approximately 0.03 eV. The graph uses the accurate Liu–Seigbahn–Truhlar–Horowitz (LSTH) potential surface [195].

The collinear model does not include bifurcation, i.e. the possibility of *several* product channels which the system can access. A model potential surface for an $A + BC \rightarrow AB + C$, $AC + B$ reaction is shown in figure B3.4.2. Both of these examples will be used in the discussion below.

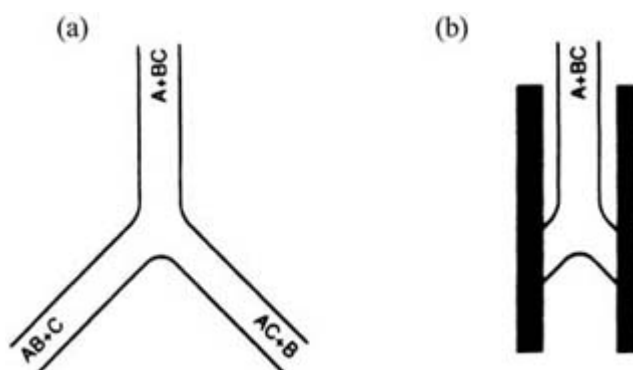


Figure B3.4.2. (a) A schematic potential surface showing bifurcation for a triatomic reactive system. (b) By blocking the products' arrangement with an absorbing potential (shaded area) the reactive system is reduced to one arrangement; this scheme enables calculation of both total reactivities and state-to-state information. Reprinted from [46] with permission.

B3.4.3.2 BOUNDARY CONDITIONS

The eventual goal in scattering calculations is essentially to obtain the scattering matrix, S (see Section A3.11 and equation (B3.4.4) below). The scattering matrix can be obtained by reference to the solution of the time-independent Schrödinger equation, fulfilling

$$(\hat{H} - E)\psi = 0 \quad (\text{B1.4.3})$$

which is the Fourier transform of [equation B3.4.1](#). However, this wavefunction must obey the appropriate boundary conditions. It should have components associated with a single ‘incoming’ channel (an incoming wave associated with the translational $\mathbf{A} + \mathbf{BC}$ motion, e^{-ikR} , multiplied by a wavefunction for BC at a specific initial target state, $\phi_{n_0}(r)$). In addition, there are components associated with all ‘outgoing’ channels. (See [Section A3.11](#) and [\[27\]](#).) For example, for the bifurcating potential case where three asymptotic arrangements are formally possible, the wavefunction (with an index n_0 attached) is

$$\begin{aligned} \psi_{n_0}(\mathbf{x}, E) = & \phi_{n_0}(r) \frac{e^{-ik_{n_0}R}}{\sqrt{k_{n_0}/M}} - \sum_n \phi_n(r) \frac{e^{ik_n R}}{\sqrt{k_n/M}} S_{nn_0} \\ & - \sum_{\bar{n}} \phi_{\bar{n}}(\bar{r}) \frac{e^{ik_{\bar{n}} \bar{R}}}{\sqrt{k_{\bar{n}}/M}} S_{\bar{n}n_0} - \sum_{\bar{\bar{n}}} \phi_{\bar{\bar{n}}}(\bar{\bar{r}}) \frac{e^{ik_{\bar{\bar{n}}} \bar{\bar{R}}}}{\sqrt{k_{\bar{\bar{n}}}/\bar{M}}} S_{\bar{\bar{n}}n_0} \end{aligned} \quad (\text{B3.4.4})$$

where ϕ_n is the n th vibrational state of the BC diatomic and k_n is the translational momentum of A when BC is in the n th state ($k_n^2/2M = E - \epsilon_n$). Quantities with a $(\bar{r}, \bar{R}, \text{etc})$ refer to the product channel $\mathbf{AB} + \mathbf{C}$, so \bar{R} is the distance between $\mathbf{AB} + \mathbf{C}$, etc. In addition, for cases in which the $\mathbf{AC} + \mathbf{B}$ channel is open, a double-bar notation is used ($\bar{\bar{n}}, \bar{\bar{r}}, \text{etc}$). $S_{\bar{n}n_0}$ is the scattering matrix associated with the amplitude of the system to emerge at product channel \bar{n} when it is initially at reactant channel n_0 . The equality sign is in quotes to denote that this relation is only valid in the asymptote where the system is separated into an atom and a diatom.

B3.4.3.3 SCATTERING TECHNIQUES

The presence of the multiple arrangements make molecular scattering very challenging theoretically. After much trial and error, several techniques have been developed. These techniques generally fall into two broad categories:

- methods which aim at treating all possible arrangements simultaneously;
- arrangement-decoupling approaches [\[28, 29, 30 and 31\]](#) where an absorbing potential is used to convert multiple-arrangement problems to inelastic-scattering (or even bound-state-like) problems. In recent years, these approaches have become very powerful for large-scale applications.

B3.4.3.4 ALL-ARRANGEMENT METHODS

(A) WAVEFUNCTION EXPANSION

The conceptually simplest approach to solve for the S -matrix elements is to require the wavefunction to have the form of [equation \(B3.4.4\)](#), supplemented by a bound function which vanishes in the asymptote [\[32, 33, 34 and 35\]](#) This approach is analogous to the full configuration-interaction (CI) expansion in electronic structure calculations, except that now one is expanding the nuclear wavefunction. While successful for intermediate size problems, the resulting matrices are not very sparse because of the use of multiple coordinate systems, so that this type of method is prohibitively expensive for diatom–diatom reactions at high energies.

(B) CLOSE COUPLING

Alternatively, one can use close-coupling methods. These methods are easiest to understand for single arrangement problems (i.e. when both the **AB** + **C** and **AC** + **B** product arrangements are very high in energy so that only the **A** + **BC** reactant arrangement can be accessed). Then one writes

$$\psi_{n_0}(R, r, E) = \sum_n a_{nn_0}(R) \phi_n(r) \quad (\text{B3.4.5})$$

and it is readily shown that the Schrödinger equation can be written as

$$\frac{\partial^2}{\partial R^2} a_{nn_0} = 2M \sum_m (U_{nm}(R) - E \delta_{mn}) a_{mn_0} \quad (\text{B3.4.6})$$

where

$$U_{nm}(R) = \int \phi_n(r) \left(-\frac{1}{2\mu} \frac{\partial^2}{\partial r^2} + V(R, r) \right) \phi_m(r) dr. \quad (\text{B3.4.7})$$

Equation (B3.4.6) is solved by starting at a small value of R , denoted by R_{start} , where the potential is high and the wavefunction is exponentially vanishing, and picking random values for a_{nn_0} and $\delta a_{nn_0}/\delta R$. Then, the equations are propagated towards larger R . Eventually, $\psi(R, r)$ is resolved at a large value of R to yield the S -matrix [36]. In practice, one has to avoid linear dependence between the solutions associated with different initial conditions. This is achieved by simple stabilization approaches.

The close-coupling approach works readily and simply if the reaction is purely ‘inelastic’. The method can also be made to work very simply for a single product arrangement (as in collinear reactions), by using a ‘twisted’ coordinate system, most conveniently reaction path coordinates [37, 38 and 39] as shown in [figure B3.4.3](#).

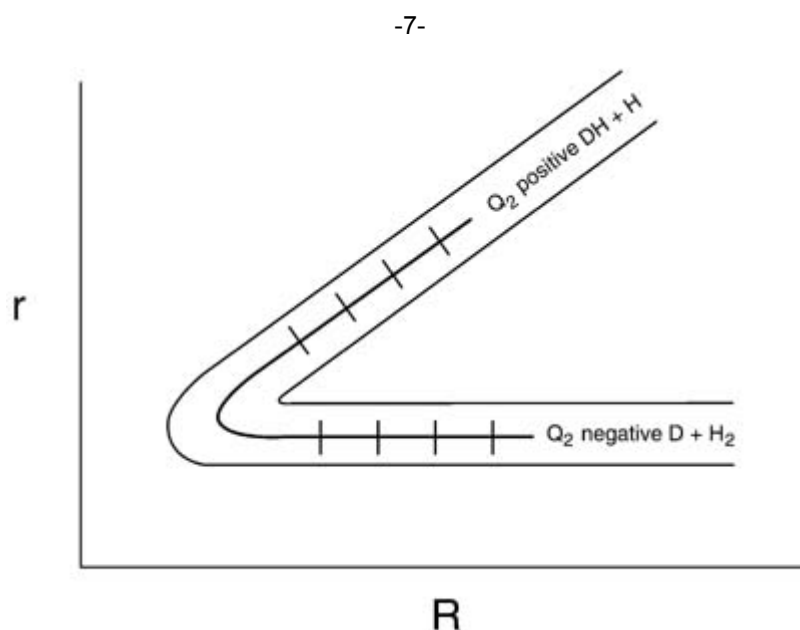


Figure B3.4.3. A schematic figure showing, for the DH_2 collinear system, a reaction-path coordinate Q connecting continuously the reactants and the single products asymptote. Also shown are the cuts denoting the coordinate perpendicular to Q .

The complications which occur with bifurcation, i.e. when more than one product arrangement is accessible, can be solved by various methods. Historically, the first close-coupling approaches for multiple product channels employed fitting procedures [40], where the close-coupling equations are simultaneously propagated from each of the asymptotes inwards and then are fitted together at a dividing surface. This approach has been replaced in recent calculations by two methods. One is based on using absorbing potentials to turn the reactive problem into an inelastic one, as explained later. The other is to use hyperspherical coordinates for carrying out the close-coupling propagation [41, 42, 43, 44 and 45]. The hyperspherical coordinates consist of a single radius ρ , which is zero at the origin (when all nuclei are stuck together) and increases outwards, and a set of angles. For the collinear problem as well as the atom-diatom problem (involving three independent distances) the hyperspherical coordinates are typically just the regular spherical coordinates. Close-coupling propagation starts at $\rho = 0$ and moves outward until a large value of ρ is reached. When the asymptote are reached one fits the wavefunction to have the form of [equation \(B3.4.4\)](#) and thus obtains the scattering matrix.

B3.4.4 ARRANGEMENT DECOUPLING BY ABSORBING POTENTIALS

A simplifying approach to scattering is to eliminate all the product asymptote. This can be done efficiently and rigorously [28, 30, 31, 46] by inserting in the Hamiltonian a negative-imaginary potential (or more generally a complex potential with a negative imaginary term) [30, 47, 48 and 49]. This potential, denoted as $-iV_{\lambda}(R,r)$, acts to ‘chop’ away the product arrangements, while retaining the correct form of the wavefunction (see [figure B3.4.4](#)).

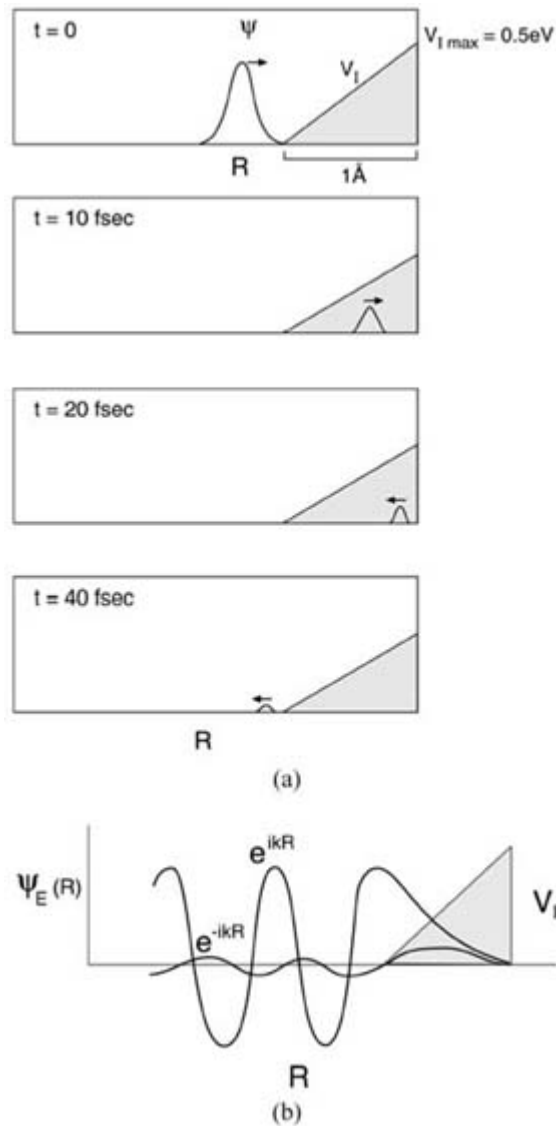


Figure B3.4.4. (a) Schematic evolution in a 1D problem of a wavepacket impinging on an absorbing potential with typical parameters (shaded). The width and magnitude of the absorbing potential must be sufficiently large so that a wavepacket impinging on it would eventually be completely absorbed, with very little reflected. (b) In time-independent language, the absorbing potential forces the wavefunction in the region preceding it to have an outgoing (e^{ikR}) form, with very little reflected (e^{-ikR}) component.

To understand this unique feature of the negative imaginary potential it is easiest to refer to the time-dependent language, discussed later in the section. Heuristically, note that the time-dependent propagator, $e^{-i(\hat{H}-iV_I)t}$ essentially contains a $e^{-V_I t}$ term which decays the wavefunction in regions where V_I is positive. The negative imaginary potential therefore prevents reflection and thus imposes ‘outgoing boundary conditions’ [31] on the wavefunctions to which they are applied.

There is considerable freedom in the choice of absorbing potentials; they are simply required [30] to be sufficiently extended to absorb any wavefunction which impinges on them, while not rising too sharply to avoid reflection from their rising slopes before the wave gets absorbed. This implies that they typically need to extend only over approximately one to two de Broglie wavelengths, which is usually short enough to add

only a negligible overhead to the size of the required grids.

The negative imaginary potentials can be applied in any scattering formalism. In close coupling, they can be implemented to block any product arrangement [31] (see figure B3.4.2) and this thereby converts the reactive problem to an inelastic one; the only cost is the propagation of a complex matrix, a_{nn_0} rather than a real one.

Alternately, absorbing potentials can also be applied to convert scattering to a bound-state-like problem. One method is to write the Schrödinger wavefunction as a sum of two terms $\psi_{n_0}(E) = \chi_{n_0}(E) + \zeta_{n_0}(E)$, where ζ_{n_0} includes the known incoming wave term, and χ includes the unknown outgoing wave part (see figure B3.4.5). The final equation is then [28]

$$(E - (\hat{H} - iV_I))\chi_{n_0} = (\hat{H} - E)\zeta_{n_0} \quad (\text{B3.4.8})$$

where the absorbing potentials are inserted to impose the correct boundary conditions on χ_{n_0} .

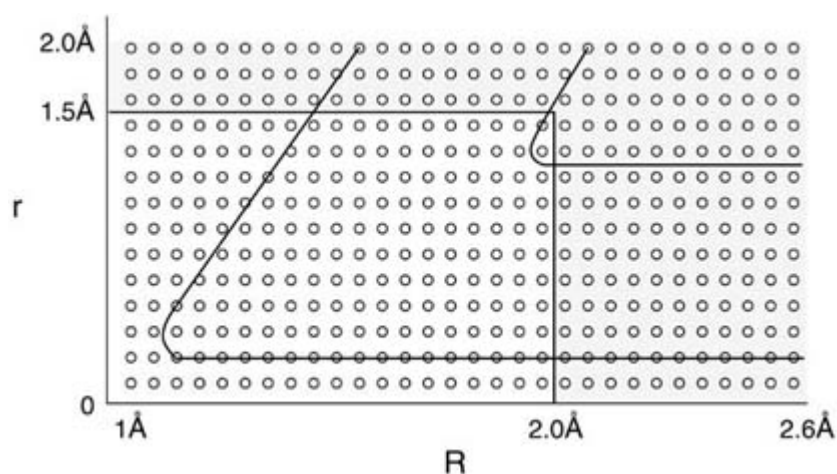


Figure B3.4.5. Schematic plot of a two-dimensional potential surface for $D + H_2$ restricted by an absorbing potential (shaded area). The absorbing potential $iV_I(R, r)$ (shaded region) rises gently outward towards the edges of the grid. In practice, the grid needs to be extended only by $\approx 0.5\text{--}1 \text{ \AA}$, or less for heavier mass systems. The absorbing potential imposes the correct boundary condition on the wavefunction in the inner region. This basic paradigm of a small grid (denoted by dots), used in the strong-interaction region to describe the main part of the wavefunction, applies in several different formulations: time-independent arrangement-decoupling scattering, where the time-independent wavefunction $\chi_{n_0}(R, r, E)$ is placed on this grid (supplemented by a function describing the initial wavefunction); time-dependent scattering (where it is used to describe the non-elastic part of the wavefunction, and the elastic part is

represented on a separate grid); flux-flux studies, and photodissociation. All desired scattering information can be obtained from the information on the wavefunction in the strong-interaction region.

This approach has one key advantage [30]. Although when solving for χ one needs to invert $(E - \hat{H} + iV_I)$ (more precisely calculate the action of the so-called Green's function $(E - \hat{H} + iV_I)^{-1}$ on the initial state), the operator is defined in a finite region so that the wavefunction can be written on a single small grid covering the small-interaction region (as no asymptotic regions are involved). This makes the operation by \hat{H} very rapid

(formally, \hat{H} is then very sparse) so that efficient iterative methods can be used [50, 51, 52, 53 and 54]. It is then possible to handle grid sizes of more than a million points. Thus systems like $\mathbf{AB} + \mathbf{CD}$ rearrangement scattering, with six or more floppy distances (and ≈ 10 points per degree of freedom) are now routinely done with arrangement decoupling approaches based on absorbing potentials. The most widely applied iterative method with absorbing potentials and arrangement decoupling was developed within a time-dependent formulation, and is discussed below.

B3.4.4.1 THE TIME-DEPENDENT METHOD

The approaches discussed so far are generally called time-independent methods, since they start from the time-independent Schrödinger equation, $(\hat{H}-E)\psi_{n_0}$. An alternative is to use the time-dependent Schrödinger equation [28, 29, 50, 55, 56, 57, 58, 59, 60, 61, 62, 63 and 64]. Conceptually, the time-dependent approach is very simple: prepare an initial wavepacket on an appropriate grid; propagate the initial wavepacket for a sufficiently long time; and analyse the results. The approach is efficient, since propagation of a wavepacket is relatively cheap and since results on scattering at many energies are extracted at once. Correct boundary conditions have a very simple meaning in time-dependent approaches: the scattered component of the wavepacket is not returned from the edges of the grid. This is done by adding an absorbing potential to the Hamiltonian, which absorbs any component of the wavefunction that reaches the edge of the grid.

In more detail, in time-dependent approaches an initial wavepacket associated with the separate parts of the colliding system is prepared. The most efficient approach for a grid construction is to use two grids. Thus, the total wavefunction is divided into two parts [30, 65], a simple one which defines the initial wavefunction and a more complicated part, represented (for collinear scattering) on a two-dimensional grid of R, r values, which is used to carry most of the wavefunction (essentially the scattered part) and which is padded with absorbing potentials (see figure B3.4.5). With this approach and with methods which reduce the number of grid points (in a given region) to at most two per oscillation [50, 66, 67], the total number of grid points can be reduced to 150–800 for collinear scattering involving hydrogen with energies of up to 2 eV.

Once the grid (or two grids) are prepared, there are two similar types of approaches to propagate the initial wavefunction forward with time. One approach is split-operator methods, [59] where the short-time propagator is divided into a kinetic and potential parts so that

$$|\psi_{n_0}(t+dt)\rangle \approx \exp\left[-i\frac{(V-iV_I)dt}{2}\right] \exp\left[-i\left(\frac{P^2}{2M} + \frac{p^2}{2\mu}\right)dt\right] \exp\left[-\frac{i(V-iV_I)dt}{2}\right] |\psi_{n_0}(t)\rangle \quad (\text{B 3.4.9})$$

-11-

where a bra-ket notation is used. The action by the potential is trivial, since it is local in coordinate space, and amounts to multiplication of $\psi_{n_0}(R, r, t)$ by $\exp(-iV(R,r)dt/2)$, and by a damping term, $\exp(-V_I(R,r)dt/2)$.

The action by the kinetic term is only slightly more complicated: the coordinate grid wavefunction, $\psi_{n_0}(R, r)$, undergoes a fast Fourier transform (FFT) to convert it to momentum space:

$$\psi_{n_0}(P, p, t) = \sum_{R,r} e^{-i(pr+PR)} \psi_{n_0}(R, r, t). \quad (\text{B 3.4.10})$$

The function is then multiplied by $\exp(-i((P^2/2M) + (p^2/2\mu))dt)$ and then returned to coordinate space by the

inverse of equation (B3.4.10).

The key to this method is thus to act with each operator (exponential of the potential or kinetic term) in the representation (coordinate or momentum grid) in which it is local [50, 66, 67].

An alternative to split operator methods is to use iterative approaches. In these methods, one notes that the wavefunction is formally $|\psi_{n_0}(t)\rangle = \exp(-i\hat{H}t)|\psi_{n_0}\rangle$, and the action of the exponential operator is obtained by repetitive application of H on a function (i.e. on the computer, by repetitive applications of the sparse matrix H on wavefunction vectors). The simplest iterative method is the Taylor expansion of $e^{-i\hat{H}t}|\psi_{n_0}\rangle$ as $\sum_n ((-i)^n t^n / n!) \hat{H}^n |\psi_{n_0}\rangle$. On the computer, this expansion would be performed by acting with \hat{H} on ψ_{n_0} , then acting with \hat{H} on the resulting vector, etc, and adding the contribution to the sum at each stage. The action by \hat{H} on a vector is straightforward, as in the split operator approach: the potential is local, and the kinetic energy is evaluated by Fourier transforming back and forth onto the momentum grid.

The Taylor series by itself is not numerically stable, since the individual terms can be very large even if the result is small, but other polynomials which are highly convergent can be found, e.g. Chebyshev [50, 62, 63 and 64] or Lancosz polynomials [51, 68].

The wavepacket is propagated until a time where it is all scattered and is away from the interaction region. This time is short (typically 10–100 fs) for a direct reaction. However, for some types of systems, e.g. for reactions with wells, the system can be trapped in resonances which are quasi-bound states (see section B3.4.7). There are efficient ways to handle time-dependent scattering even with resonances, by propagating for a short time and then extracting the resonances and adding their contribution [69].

The last stage is the extraction of energy-resolved information, obtained automatically and simultaneously at many energies, by Fourier transforming the wavefunction to produce an energy-resolved state:

$$\psi_{n_0}(R, r, E) = \frac{1}{a_E} \int_{-\infty}^{\infty} \psi_{n_0}(R, r, t) e^{iEt} dt \quad (\text{B 3.4.11})$$

-12-

where a_E is related to the energy content of the initial wavepacket. It is easy to show that $\psi(R, r, E)$ fulfills the time-independent Schrödinger equation. (In practice, ψ is known analytically prior to $t = 0$, so that the wavefunction only needs to be propagated forward in time.) The scattering matrix is then obtained from either of several formulae, all of the form [46, 65, 70, 71, 72, 73 and 74]

$$S_{\bar{n}n_0} = \langle \xi_{\bar{n}} | \psi_{n_0}(E) \rangle \quad (\text{B 3.4.12})$$

where $\xi_{\bar{n}}$ is a simple function which is associated with the final state \bar{n} . These formulae extract long-range scattering information from the wavefunction values in the strong-interaction region. Scattering information can therefore be extracted even when absorbing potentials are used to remove the asymptotic regions.

An interesting side point is that it is possible to recast the time-dependent approach, as described here, in a purely time-independent fashion, since from the equations above it follows that [74]

$$\psi_{n_0}(E) = \text{constant} \frac{1}{E - \hat{H} + iV_I} \psi_{n_0}.$$

The time-dependent approach is thus just one technique for evaluating the action of the Green's function on the initial wavepacket.

B3.4.4.2 LARGE-SCALE APPLICATIONS

Both close-coupling approaches (hyperspherical or with absorbing potentials) and iterative/time-dependent absorbing-potential arrangement-decoupling approaches are readily extended to three-dimensional atom-molecule and molecule-molecule scattering. The wavefunction representation becomes more complicated and includes rotational matrices, but the essence and application of the method remains analogous [58, 65, 75, 76].

Iterative approaches, including time-dependent methods, are especially successful for very large-scale calculations because they generally involve the action of a very localized operator (the Hamiltonian) on a function defined on a grid. The effort increases relatively mildly with the problem size, since it is proportional to the number of points used to describe the wavefunction (and not to the cube of the number of basis sets, as is the case for methods involving matrix diagonalization). Present computational power allows calculations with optimized grids with sizes of 10^5 – 10^7 points or more. This enables efficient simulations of four-body reactions involving six independent distances and up to two overall rotational coordinates. Thus far there have been several four-body reactions reported using this method, including $\text{H}_2 + \text{OH} \rightleftharpoons \text{H}_2\text{O} + \text{H}$ [75, 76, 77 and 78] and $\text{CO} + \text{HO} \rightleftharpoons \text{CO}_2 + \text{H}$ [79, 80], as well as surface reactions [58, 81] (see figure B3.4.6 for an example).

-13-

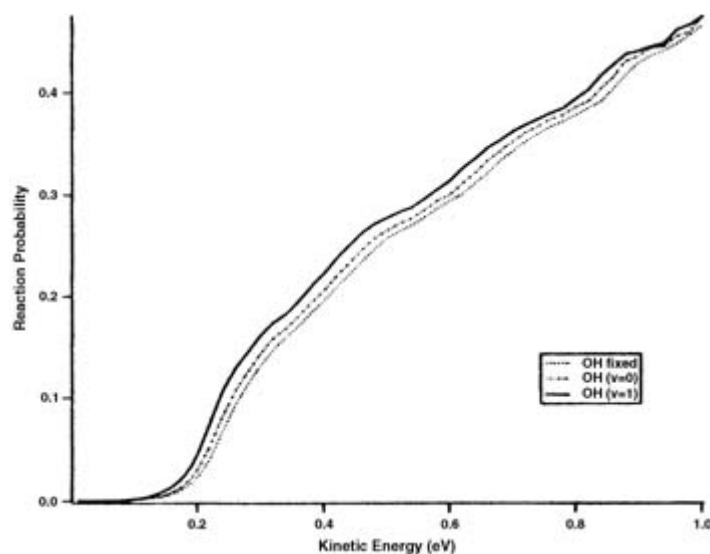


Figure B3.4.6. Reaction probabilities for the initial-state-selected process $\text{H}_2(v=0, j=0) + \text{OH}(v, j=0) \rightarrow \text{H}_2\text{O} + \text{H}$, for zero total angular momentum. Taken from [75] with permission.

B3.4.5 COARSE INFORMATION

The methodology presented so far allows the calculations of state-to-state S -matrix elements. However, often one is not interested in this high-level of detail but prefers instead to find more average information, such as the initial-state selected reaction probability, i.e. the probability of rearrangement given an initial state n_0 . In general, this probability is

$$P_{n_0}(E) = \sum_{\bar{n}} |S_{\bar{n}n_0}(E)|^2. \quad (\text{B 3.4.13})$$

For example, for the collinear reaction A+BC this would be the probability that if initially the diatom BC is in a vibrational state $\phi_{n_0}(r)$, then after the reaction a diatom AB is formed (in *any* product vibrational state). In practice, the initial-state selected probability is easily calculated from the flux of the wavefunction $\psi_{n_0}(R, r, E)$ calculated at the product arrangement (e.g., at a large value of r , the B–C separation).

At times, however, even the information presented by P_{n_0} is too detailed. If one wants to rigorously calculate the thermal rate of rearrangement reactions, the initial vibrational state is not important. The relevant quantity is the sum of the initial-state-selected probabilities

$$N(E) = \sum_{n_0} P_{n_0}(E) = \sum_{n_0 \bar{n}} |S_{\bar{n}n_0}(E)|^2. \quad (\text{B 3.4.14})$$

-14-

$N(E)$ is called the cumulative reaction probability. It is directly related to the thermal reaction rate $k(T)$ by

$$k(T) = \frac{\int N(E) e^{-E/kT} dE}{Q} \quad (\text{B 3.4.15})$$

where T is the temperature and Q is the reactants' partition function.

A major achievement [71, 82, 83, 84, 85, 86, 87 and 88] was the development of a simple quantum ('flux-flux') expression for the cumulative reaction probability, $N(E)$, with the final result [88]

$$N(E) \propto \text{Im Tr}(FGFG^*) \quad (\text{B 3.4.16})$$

where the Green's function is, as mentioned earlier,

$$G(E) = \frac{1}{E - (\hat{H} + iV_I)} \quad (\text{B 3.4.17})$$

and F is the flux operator. In this expression, the trace is evaluated over a small grid region. In principle, the grid has to contain only a small-interaction region, in which the system 'decides' its final arrangement (i.e. with what probability to react). This expression does not refer to the scattering matrix and therefore the asymptotic region does not have to be included in the grid.

The flux-flux expression and its extensions have been used to calculate reaction probabilities for several important reactions, including $\text{H}_2 + \text{O}_2 \rightarrow \text{H} + \text{H}_2\text{O}$, by explicit calculation of the action of G in a grid representation with absorbing potentials. The main power of the flux-flux formula over the long run will be the natural way in which approximations and semi-classical expressions can be inserted into it to treat larger systems.

B3.4.6 PHOTO-DISSOCIATION

The time-dependent approach has the advantage that it is easy to visualize the propagation of a simple wavepacket and make intuitive sense of a large body of chemical phenomena. This is especially powerful in photo-initiated processes. As a result of a photon absorption, the ground-state wavefunction is ‘jumped’ to a higher potential energy surface of a different electronic state and propagates on this new surface. The initial excitation is, by the Frank–Condon principle, essentially vertical (i.e. the nuclear position and momentum do not change, only the electronic state). The subsequent process (see [figure B3.4.7](#)) is the response of the nuclear coordinates to the change in the electronic state.

-15-

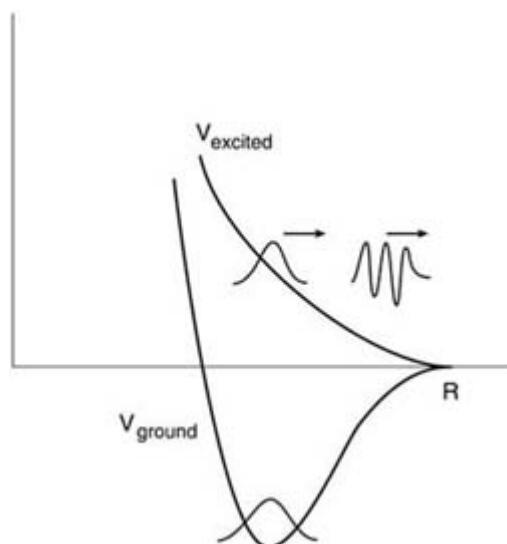


Figure B3.4.7. Schematic example of potential energy curves for photo-absorption for a 1D problem (i.e. for diatomics). On the lower surface the nuclear wavepacket is in the ground state. Once this wavepacket has been excited to the upper surface, which has a different shape, it will propagate. The photoabsorption cross section is obtained by the Fourier transform of the correlation function of the initial wavefunction on the excited surface with the propagated wavepacket.

For two Born–Oppenheimer surfaces (the ground state and a single electronic excited state), the total photo-dissociation cross section for the system to absorb a photon of energy ω , given that it is initially at a state $|\chi\rangle$ with energy E_0 can be shown, by simple application of second-order perturbation theory, to be [89]

$$\sigma(\omega) = \text{constant} \cdot \int e^{i(\omega+E_0)t} c(t) dt \quad (\text{B 3.4.18})$$

where the correlation function is defined as

$$c(t) = \langle \Phi | e^{-i\hat{H}_{\text{exc}}t} | \Phi \rangle \quad (\text{B 3.4.19})$$

$\Phi = \mu\chi$, μ is the dipole moment and χ is the initial vibrational state on the ground surface (with energy E_0). \hat{H}_{exc} is the excited-state potential energy. This expression has a clear physical meaning. Take an initial wavepacket, χ , multiply it by the dipole moment, and use the resulting packet ($\Phi \equiv \mu\chi$) as an initial function so that it is propagated under the excited-state potential ($\Phi(t) \equiv e^{-i\hat{H}_{\text{exc}}t} \Phi$).

equation (B3.4.18) makes a very powerful statement: absorption is only related to the Fourier transform of the

correlation function. All that is needed is to know how the wavepacket propagates in time on the upper surface. Keeping in mind the time–energy uncertainty principle, one can then qualitatively understand the spectral features in

-16-

photo-absorption [89]. Referring to figure B3.4.8 the correlation function starts at 1 at $t = 0$, then undergoes a period of decay over a time scale denoted by T_1 . This initial decay is more rapid if the excited potential surface is steep, slower if it is shallow. This is the shortest time scale in the correlation function, so it corresponds to the broadest feature in the energy spectrum, i.e. the envelope in figure B3.4.9. In figure B3.4.8, the correlation peaks every time the wavepacket returns to the initial placement (denoted by T_2). In the energy picture, the peak spacing is $2\pi/T_2$. The correlation peaks are decreasing in magnitude over time T_3 as the wavepacket either decays to other modes or moves to another region of the potential energy surface. T_3 is therefore the peak's width in the energy spectrum.

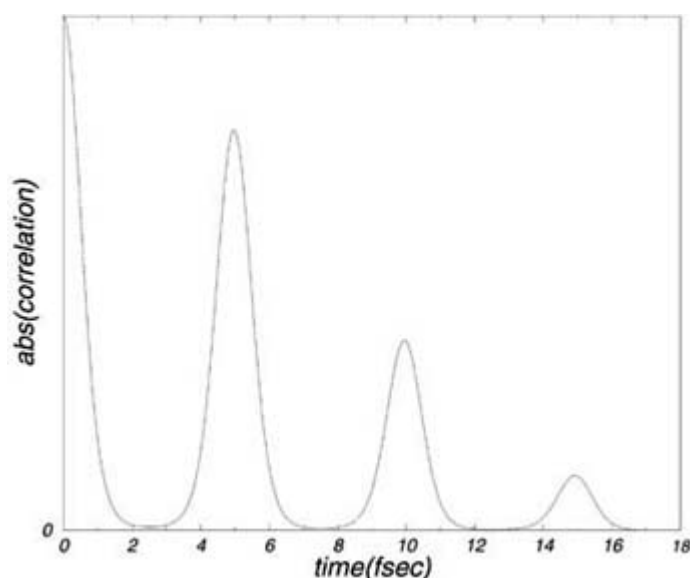


Figure B3.4.8. The correlation function $c(t) = \langle \psi_0 | \psi(t) \rangle$ as a function of time for photodissociation in a collinear (or three-dimensional) polyatomic case. There are three relevant time scales; T_1 , which measures how rapidly the initial wavefunction dephases; T_2 , which measures how long it takes this initial wavefunction to regroup; and T_3 which measures how long the wavefunction takes to ‘leak’ to other degrees of freedom. In practice, photodissociation experiments may yield spectra which are more blurred, if T_1, T_2 and/or T_3 are not well separated.

-17-

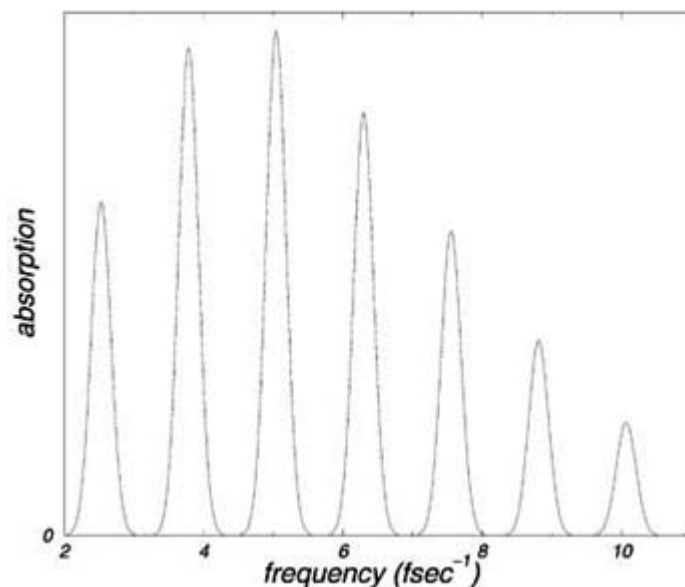


Figure B3.4.9. The Fourier transform of the correlation function, from [figure B3.4.8](#), which gives the absorption spectrum as a function of frequency.

This time-dependent method allows one to nicely connect the theoretical and experimental observations. As mentioned earlier, the correlation function and its generalizations yield the spectra for a large number of other photospectroscopy processes, such as Raman processes [[90](#)], as well as molecular scattering [[73](#), [74](#)].

B3.4.7 BOUND STATES AND RESONANCES—EXTRACTION

B3.4.7.1 RESONANCES—FORMALISM

The quantum dynamics of bound and scattered systems is closely correlated through the concept of resonances which are, heuristically, quasi-bound states in which the system can spend time [[6](#), [7](#), [8](#) and [9](#), [91](#), [92](#), [93](#) and [94](#)]. More formally, resonances are poles of the S -matrix. (See section A3.11.) In a scattering process, the cross section typically exhibits peaks as a function of the scattering energy, exactly at (or near) the energy of the resonances. For example, in a one-dimensional scattering off a double well, the scattering probabilities exhibit sharp peaks when the collision energy matches the energy of the quasi-bound states in the well ([figure B3.4.10](#)). (See [figure B3.4.11](#) for a realistic example.)

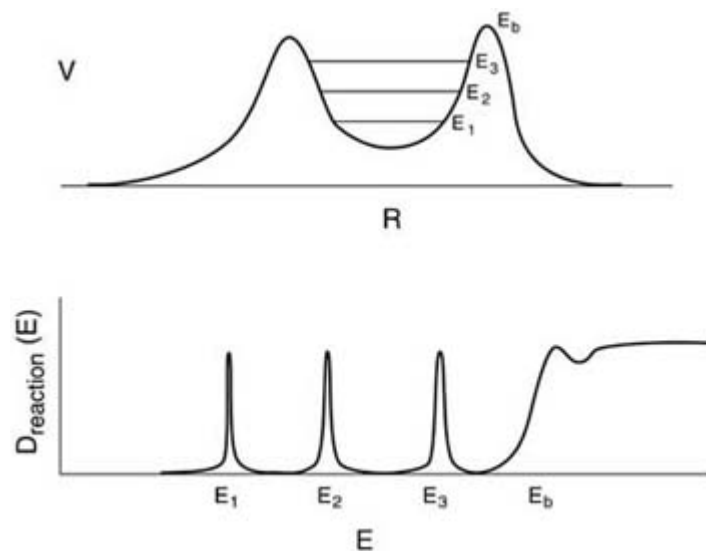


Figure B3.4.10. Schematic figure of a 1D double-well potential surface. The reaction probabilities exhibit peaks whenever the collision energy matches the energy of the resonances, which are here the quasi-bound states in the well (with their energy indicated). Note that the peaks become wider for the higher energy resonances—the high-energy resonance here is less bound and ‘leaks’ more toward the asymptote than do the low-energy ones.

-19-

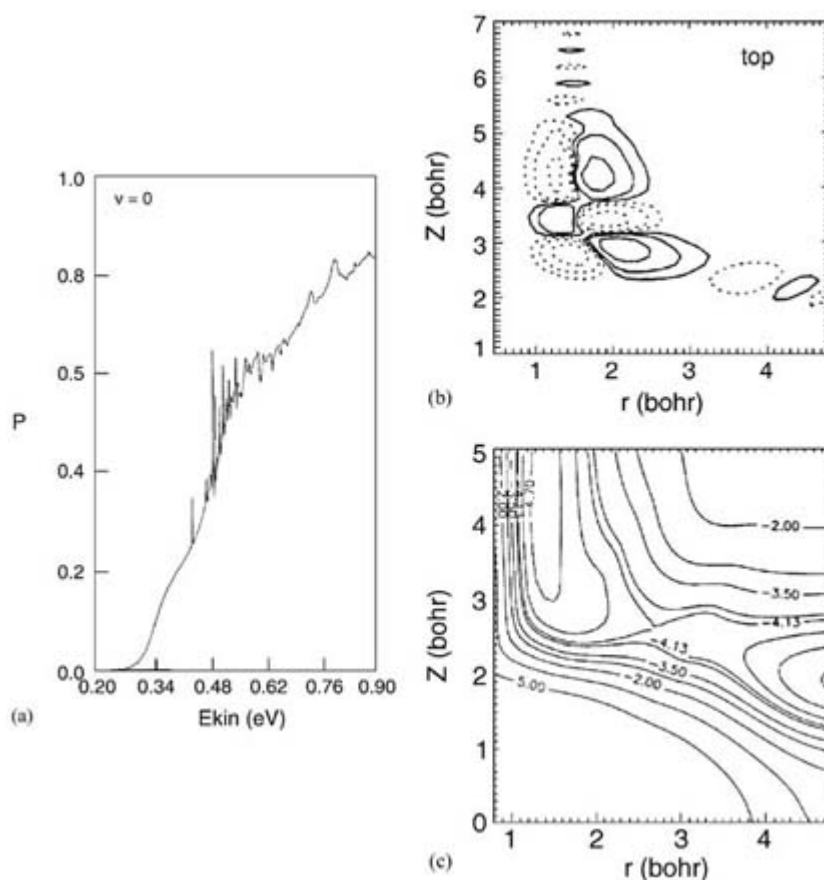


Figure B3.4.11. (a) Reaction probability for a 4D study of the dissociation of incident H_2 on CO. The probability exhibits sharp peaks whenever the energy matches that of a resonance wavefunction. (b) Plot of

the resonance wavefunction associated with one of the peaks, as well as (c) a 2D cut of the potential surface. Note that the resonance wavefunction decays near the end of the grid, due to the use of an absorbing potential, which localizes its effects to the strong-interaction region. Taken from [196], with permission.

The classical counterpart of resonances is periodic orbits [91, 95, 96, 97 and 98]. For example, a purely classical study of the $\mathbf{H}+\mathbf{H}_2$ collinear potential surface reveals that near the transition state for the $\mathbf{H}+\mathbf{H}_2\rightarrow\mathbf{H}_2+\mathbf{H}$ reaction there are several trajectories (in R and r) that are periodic. These trajectories are not stable but they nevertheless affect strongly the quantum dynamics. A study of the resonances in $\mathbf{H}+\mathbf{H}_2$ scattering as well as many other triatomic systems (see, e.g., [99]) reveals that the scattering peaks are closely related to the frequencies of the periodic orbits and the resonance wavefunctions are large in the regions of space where the periodic orbits reside.

Theoretically, resonances are essentially solutions of the Schrödinger equation at complex energies. These specific solutions have the property that they are mainly concentrated in the strong-interaction region and at the asymptote are outgoing waves. For one-dimensional predissociation (figure B3.4.12) where the coordinate is labelled R , the resonance wavefunction is asymptotically (i.e. for large positive R) the following outgoing wave:

-20-

$$\psi_{\text{Res}}(R) \approx a e^{ikR} \quad (\text{B 3.4.20})$$

where a is a constant here and k is the energy-dependent wavevector, $k^2/2m = E$. The existence of such a resonance function seems to be baffling, since graduate level quantum mechanical texts [100] prove that the only bound solutions to the Schrödinger equation are those with real energies and zero flux. Heuristically, the solution to this difficulty is that the formal resonance energies are complex. Thus, k is complex and e^{ikR} blows up when R is very large. Therefore, resonance functions are not bound functions and the regular proofs do not apply to them.

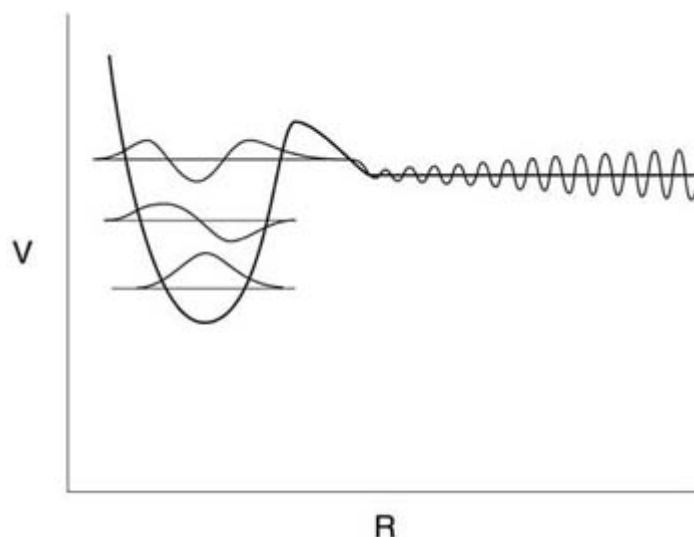


Figure B3.4.12. A schematic 1D vibrational pre-dissociation potential curve (wide full line) with a superimposed plot of the two bound functions and the resonance function. Note that the resonance wavefunction is associated with a complex wavevector and is slowly increasing at very large values of R . In practice this increase is avoided by using absorbing potentials, complex scaling, or stabilization.

The key to practical calculations of resonances is to limit the extent of the grids used for describing the wavefunctions. In the original approach, called ‘stabilization’ [92, 93 and 94], a finite grid or basis set is used and the Hamiltonian is diagonalized on that grid (or for that basis set). Those few eigenvalues which change very little when the grid size is modified are associated with the wavefunction of the resonances. (The resonances are concentrated in the interaction region, so that they are not sensitive to the details of the grid end points.) The main difficulty with this approach is that it necessitates many diagonalizations of the Hamiltonian matrix, one for each grid size.

An alternate and formally very powerful approach to resonance extraction is complex scaling [7, 101, 102, 103, 104, 105, 106 and 107] whereby a new Hamiltonian is solved. In this Hamiltonian, the grid’s multi-dimensional coordinate (e.g., \mathbf{x}) is multiplied by a complex constant α . The kinetic energy gains a constant complex factor ($\partial^2/\partial\mathbf{x}^2 \rightarrow (1/\alpha^2)(\partial^2/\partial\mathbf{x}^2)$), while the potential needs to be evaluated at points with a complex argument $V(\alpha\mathbf{x})$. In a typical calculation, one diagonalizes the resulting complex Hamiltonian for several complex values of α , and the complex resonances are

-21-

those which do not change appreciably with respect to α (analogous to the stabilization approach). Complex scaling has been applied mostly to analytical potentials [7, 101, 102, 103 and 104]; however, it could also be used for numerically– derived potentials [105, 106 and 107].

Finally, the simplest approach to extract resonances is to add to the Hamiltonian an absorbing potential [8, 48, 108, 109], and then look for the complex eigenvalues of the Hamiltonian $\hat{H}-iV_f$. The absorbing potential ensures that the resonance wavefunction has the correct form (is outgoing) in the asymptotic region immediately preceding V_f (see figure B3.4.4). Again, resonance functions are found by varying the parameters (the length and magnitude of V_f).

B3.4.7.2 NUMERICALLY EXTRACTING BOUND STATES AND RESONANCE FUNCTIONS

As explained above, the practical extraction of resonance eigenfunctions and eigenvalues (in complex scaling or with absorbing potentials) amounts to extraction of eigenvalues of a complex Hamiltonian and is thus completely equivalent to extraction of bound states. One method for extracting the complex eigenstates of the Hamiltonian is simply to expand it in terms of a fixed basis set, and diagonalize directly the resulting matrix. This approach works for small- and intermediate-scale problems and/or low-energy eigenstates.

An alternative is to use iterative methods. The simplest iterative technique for calculating bound state or resonances is to pick a random initial wavefunction $\psi_0(\mathbf{x})$ and propagate it forward in time, producing a wavepacket:

$$\psi(\mathbf{x}, t) = e^{-i\hat{H}t} \psi_0(\mathbf{x}). \quad (\text{B 3.4.21})$$

\hat{H} refers here to the complex Hamiltonian, i.e. after complex scaling or inclusion of an absorbing potential; \mathbf{x} is the grid (or basis set) used to represent ψ . Fourier transform ψ with respect to E

$$\psi(\mathbf{x}, E) = \int_0^T e^{iEt} \psi(\mathbf{x}, t) dt \quad (\text{B 3.4.22})$$

where T is a large time. It is clear that the squared norm of $\psi(\mathbf{x}, E)$ (i.e. $\int |\psi(\mathbf{x}, E)|^2 d\mathbf{x}$) has a peak whenever E is

near a resonance or bound-state energy, ϵ_n , since $\psi(x,t)$ has contributions varying as $e^{-i\epsilon_n t}$ from each eigenenergy ϵ_n .

This ‘direct filter’ technique is very powerful [56, 59] in extracting highly excited states, since only the propagation of a wavepacket is required. However, it is inefficient when there are closely-lying eigenvalues (T needs to be larger than the inverse level spacing, $|\epsilon_{n+1} - \epsilon_n|^{-1}$, as a manifestation of Heisenberg’s uncertainty relation) or ϵ_n is a wide resonance (i.e. its imaginary eigenvalue is large in absolute magnitude compared with the level spacing, so that its contribution to the wavepacket, $e^{-i\epsilon_n t}$, is washed out rapidly as a function of time).

To avoid these difficulties an alternate approach, labelled filter diagonalization, was developed [110, 111, 112, 113, 114, 115, 116 and 117]. The approach is powerful for extracting highly excited energies. Mechanistically, filter diagonalization is simple (see figure B3.4.13). The initial wavepacket is propagated for a short time T , and the filtered functions $\psi(x, E)$ are prepared as in equation (B3.4.22) but using a short time T . The filtering is carried out for several

(typically ≈ 50) closely spaced energies which could be way above the ground-state energy. The key is to note that even after a short propagation time, T , the $\psi(x, E)$ functions would only contain contributions from eigenstates within a narrow strip of sampled energies. Thus, the filtered functions $\psi(x, E)$ are an excellent basis for eigenstates and eigenvalues within the filtered energy range. The true eigenvalues within this range are then found by diagonalizing the small (e.g., 50×50) matrix of the Hamiltonian operator within the $\psi(x, E)$ basis. The advantage of filter diagonalization is that it avoids the long propagation times of the pure filter approach as well as the large matrix diagonalizations associated with pure diagonalization.

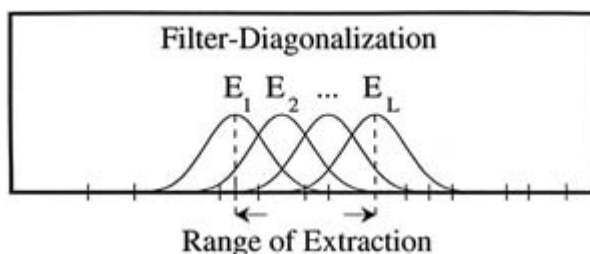


Figure B3.4.13. The basic premise in filter diagonalization is to filter a wavefunction for a short time at several energies, E_1, E_2, \dots , so that, in energy space, the resulting set of several filtered functions (denoted by full bell-shapes) spans the eigenstates (short bars) in the energy range of interest. The short-time filtered wavefunctions can therefore be used to extract the eigenstates at the desired energy range, with a modest cost, since only short-time filter and small-matrix diagonalizations are used.

As a side note, filter diagonalization is also useful in a more general context. It can be shown that it is an efficient approach for extracting frequencies from a short-time segment of a general signal [112, 113 and 114, 118, 119], so that it is not even necessary to use a wavepacket! All one needs is a signal. This feature is very important in semi-classical and path integral simulations discussed below, where all the information is extracted from a time-dependent correlation function, because the quality of the simulations degrades as a function of time (the number of trajectories is typically increased exponentially as the time is increased); therefore, information must be extracted from the shortest time possible.

The formalism outlined in the previous sections is very useful for small systems, but is, as explained, impractical for more than six to ten strongly interacting degrees of freedom. Thus, alternate approaches are required to represent dynamics for large systems. Currently, there are many new approaches developed and tested for this purpose, and these approaches are broadly classified as follows:

- frozen and zero-point approximations;
- mean-field methods and their extensions;

-23-

- Gaussian-wavepacket-based techniques;
- path-integral and semi-classical approaches.

These approaches are generally interwoven, and some of the most exciting developments in chemical dynamics have been associated with their combinations. This section very briefly describes the motivation behind and the application of these techniques.

B3.4.8.1 FROZEN AND ZERO-POINT APPROXIMATIONS

Often a degree of freedom moves very slowly; for example, a heavy-atom coordinate. In that case, a plausible approach is to use a ‘sudden’ approximation, i.e. fix that coordinate and do reduced dimensionality quantum-dynamics simulations on the remaining coordinates. A common application of this technique, in a three-dimensional case, is to fix the angle of approach to the target [120, 121] (see figure B3.4.14).

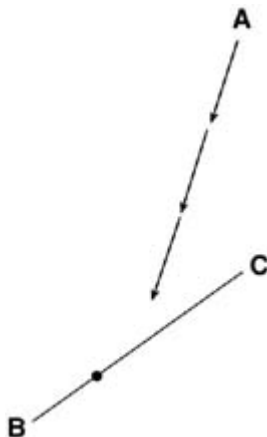


Figure B3.4.14. The infinite-order-sudden approximation for $A + BC \rightarrow AB + C$. In this approximation, the BC molecule does not rotate until reaction occurs.

The sudden approach has been applied widely, starting with atom-diatom calculations and continuing today for diatom-diatom calculations [79, 122, 123]. This approach and a related approximation (coupled states or CS, which involves the neglect of Coriolis coupling terms in three dimensions [120, 121, 124, 125]) are much more powerful when combined with the arrangement decoupling with absorbing potentials approach discussed earlier. The reason is that approximations are typically much easier to formulate and apply, and are more valid, for single-arrangement approaches. For an example of the successful merging of approximations and arrangement decoupling, see [126].

A related and particularly simple approximation is J -shifting [127, 128]. This method is a simple (and

generally useful) trick for calculating reaction probabilities in three dimensions from a single calculation with zero total angular momentum ($J = 0$), by approximating the effect of the non-zero angular momentum as shifting a transition state:

-24-

$$P_J(E) = (2J + 1)P_{J=0}(E - C_J) \quad (\text{B 3.4.23})$$

where C_J is determined from the contribution of angular degrees of freedom to the transition state energy.

A related type of approximation is the reaction path method [[122](#), [123](#), [129](#), [130](#), [131](#) and [132](#)], in which the coordinates are divided into those which are relatively rigid during the reaction and are typically associated with harmonic oscillators, and the remaining coordinates (reaction coordinates) which undergo significant change during the reaction. The contribution of those degrees of freedom which are replaced by harmonic oscillators can be taken, for low-energy reactions, simply by the zero point energy of the harmonic oscillators. More sophisticated treatments, appropriate for higher temperatures, are actively developed now in the area of liquid reaction dynamics where they are used to describe effects of solvents (see section C3.5 for details and further references). In addition, proper inclusion of rotational states of the fragments was recently shown to yield accurate results in a molecular multi-dimensional reaction-path-like approach [[133](#)].

B3.4.8.2 MEAN-FIELD METHODS AND THEIR EXTENSIONS

The mean field technique is one of the most robust and simple methods used to handle larger molecules in gas and liquid environments [[50](#), [134](#), [135](#) and [136](#)]. The basic premise of all mean-field methods is that the full wavefunction represents N very weakly coupled modes (Q_i) and can be approximated as

$$\psi(\{Q_i\}; t) = \prod_{i=1}^N \chi(Q_i, t). \quad (\text{B 3.4.24})$$

The result of this approximation is that each mode is subject to an effective average potential created by all the expectation values of the other modes. Usually the modes are propagated self-consistently. The effective potentials governing the evolution of the mean-field modes will change in time as the system evolves. The advantage of this method is that a multi-dimensional problem is reduced to several one-dimensional problems.

The fundamental disadvantage of the mean-field method is that it does not allow modes to respond in a correlated manner to each other. This problem can be somewhat alleviated by a good definition of the relevant coordinate system [[134](#), [136](#)]. (An extension of mean-field methods that does allow for coupling [[137](#), [138](#) and [139](#)] will be discussed later.)

If one is interested in low-lying eigenvalues or low-energy scattering a CI-like approach can be applied, in which one uses zero-order eigenfunctions of a simple Hamiltonian to expand the wavefunction. Spectroscopic calculations including up to 20 to 30 degrees of freedom have been carried out using such an approach [[140](#), [141](#), [142](#) and [143](#)].

B3.4.8.3 GAUSSIAN-WAVEPACKET BASED TECHNIQUES

Gaussian wavepackets are very special functions which, in a sense, bridge the gap between classical and quantum

descriptions [89, 144]. They are defined as

$$\psi(\mathbf{x}, t) = \text{constant} \times \exp(-(\mathbf{x} - \bar{\mathbf{x}}(t))^2 / 2\sigma^2) \exp(i\bar{\mathbf{p}}(t) \cdot \mathbf{x}) \quad (\text{B 3.4.25})$$

where $\bar{\mathbf{x}}(t)$, σ and $\bar{\mathbf{p}}(t)$ are the average position, width and momentum of the packet, respectively. A Gaussian wavepacket represents the ground state of a harmonic oscillator, shifted in position and momentum. (Gaussian wavepackets are also known as Glauber or ‘coherent’ states—although the latter definition is sometimes applied to more general functions.)

The primary property of Gaussian wavepackets is that on a harmonic potential surface they are solutions of the time-dependent Schrödinger equation. Specifically, if one places a Gaussian wavepacket at $t = 0$ on a harmonic oscillator with an arbitrary average position and momentum, $\bar{\mathbf{x}}(t = 0)$ and $\bar{\mathbf{p}}(t = 0)$, the resulting wavefunction will remain a Gaussian. The average position and momentum change in time exactly like a classical particle would.

From this basic fact, several related approaches have emerged. First, a technique in which one propagates classical trajectories forward in time, and uses a single or multiple sets of Gaussian wavepackets (one for each classical trajectory) as an ansatz for the full wavefunction is introduced. For each Gaussian the position and momentum are specified by the classical trajectory [144]. This technique is already able to account for much of the zero-point energy effects. In addition, interference and tunnelling effects can be partially accounted for by adding several multi-dimensional Gaussians [144, 145, 146 and 147]. The method has been shown recently to be very powerful for non-adiabatic coupling problems (see [section B3.4.9](#)).

Gaussian wavepackets have also found use in a new approach that improves on mean-field techniques [137, 138 and 139, 148]. In this method, the degrees of freedom are divided into those which change strongly during the reaction (the ‘system’ coordinates), which are treated by an explicit wavepacket; the remaining ‘bath’ coordinates are treated by Gaussians. The parameters for the bath Gaussians are dependent on the system state, and this introduces explicitly a correlation between the system and bath modes (the bath responds differently to different parts of the system). Use of this technique in multi-dimensional reactions involving tunnelling has shown it to be significantly superior to mean-field techniques, while requiring modest numerical effort even for multi-dimensional systems.

B3.4.8.4 PATH INTEGRAL AND SEMI-CLASSICAL DYNAMICS

In the 1940s, Feynmann realized that quantum mechanics can be recast in a simple form which is very reminiscent of classical mechanics [149, 150]. This approach, path integrals, has been heavily researched in chemical dynamics since, if properly convergent, it could allow calculations on very large systems. Even earlier, van Vleck [151] postulated a semi-classical approach, i.e. a method which (like the Wentzel–Kramers–Brillouin (WKB) approximation) captures quantum interference effects and is accurate when the masses or energies are large so that \hbar can be considered small. This section briefly describes Feynmann’s approach, semi-classical dynamics, and recent developments and improvements.

For concreteness, assume that we consider several particles described together by a multi-dimensional coordinate \mathbf{x} and assume that the kinetic energy is the usual $(-1/2M)(\partial^2/\partial\mathbf{x}^2)$, where M is a single relevant mass. (Hamiltonians in

quantum dynamics can usually be brought to this form upon rescaling of the coordinates by constants.) The basic ingredient in Feynmann's approach is the correlation function for a general function ψ :

$$\langle \psi | \psi(\tau) \rangle = \langle \psi | e^{-i\hat{H}\tau} | \psi \rangle = \int \psi^*(\mathbf{x}'') G(\mathbf{x}'', \mathbf{x}', \tau) \psi(\mathbf{x}', \tau) d\mathbf{x}' d\mathbf{x}'' \quad (\text{B 3.4.26})$$

where the time-dependent Green's function G is very simple:

$$G(\mathbf{x}'', \mathbf{x}', \tau) = \text{constant} \cdot \sum_{\text{path}} e^{iS(\text{path})/\hbar}. \quad (\text{B 3.4.27})$$

In this section, we insert \hbar explicitly. The sum has the following meaning: for each pair of points \mathbf{x}' and \mathbf{x}'' draw a path $\mathbf{x}(t)$ that starts at \mathbf{x}' ($\mathbf{x}(t=0) = \mathbf{x}'$) and ends at \mathbf{x}'' ($\mathbf{x}(t=\tau) = \mathbf{x}''$). This path need not be a classical path (see figure B3.4.15). Each such path contributes $e^{iS/\hbar}$, where S is the action of the path. S (unrelated to the scattering matrix) is calculated very simply as

$$S = \int \left[\frac{M}{2} \left(\frac{dx}{dt} \right)^2 - V(x(t)) \right] dt \quad (\text{B 3.4.28})$$

where V is the potential. This, in principle, gives a very simple prescription. All one has to do in order to calculate quantum mechanical properties is to sum over 'many' quantum trajectories.

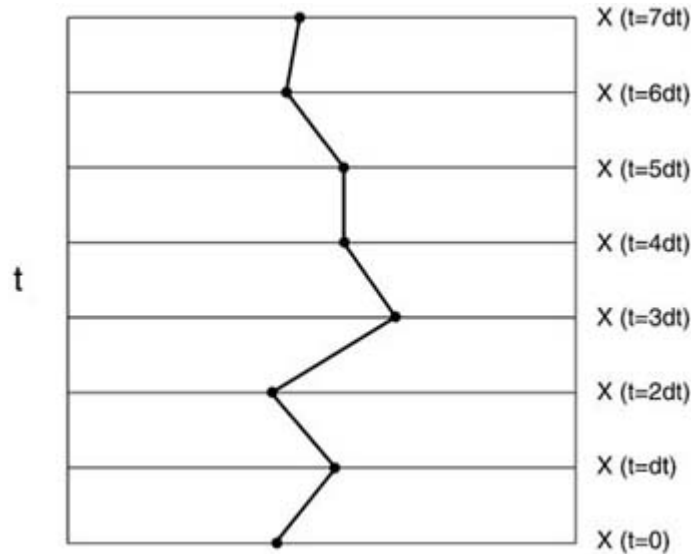


Figure B3.4.15. A possible Feynmann path trajectory for a 1D variable as a function of time. This trajectory carries an oscillating $e^{iS/\hbar}$ component with it, where S is the action of the trajectory. The trajectory is highly fluctuating; its values at each time step ($x(dt), x(2dt), \dots$, etc) are not correlated.

Unfortunately, this simple approach is not plausible numerically. The integral, as presented, will not converge, even for short times. The problem is that even trajectories which are 'wild', i.e. highly fluctuating, contribute,

per path, the same as trajectories that are ‘gentle’. The key to introducing convergence is to realize that highly fluctuating trajectories or trajectories that lie in regions of high potential energy give contributions that are cancelled by those of other nearby trajectories. Thus, a *bundle* of highly fluctuating trajectories gives only a very weak contribution. The only trajectories that give very strong contribution are *classical* trajectories (for example, if there is no potential, the classical trajectory would run with constant velocity from \mathbf{x}' to \mathbf{x}''). This realization is the key to the development of semi-classical approaches, i.e. approaches in which the exact calculation of the time-dependent Green’s function is replaced by the sum of contributions of classical trajectories. The basic semi-classical expression is [151, 152, 153, 154, 155, 156, 157 and 158]

$$\langle \psi | \psi(t) \rangle = \text{constant} \int \sum_{\text{path}} \frac{e^{iS_{\text{path}}/\hbar}}{\sqrt{|\partial \mathbf{p}''/\partial \mathbf{x}'|}} \psi^*(\mathbf{x}'') \psi(\mathbf{x}') d\mathbf{x}'' d\mathbf{x}'. \quad (\text{B 3.4.29})$$

Note the meaning of this expression: for each choice of the initial and final position \mathbf{x}' and \mathbf{x}'' , calculate the classical path that takes you from \mathbf{x}' to \mathbf{x}'' in time t . Specifically, calculate the momentum along the path and the final momentum, \mathbf{p}'' , and find out how \mathbf{p}'' varies with the initial position. This would give, for a multi-dimensional problem, a matrix $\partial \mathbf{p}''/\partial \mathbf{x}'$; whose absolute determinant needs to be inverted.

There is a simple physical explanation to the inverse determinant in equation (B3.4.29). Each classical trajectory has a ‘volume’ of quantum trajectories nearby which have similar phases to it. Beyond that volume, the phases are becoming random. Thus, the larger that volume, the greater contribution would that bundle give to the final semi-classical ‘propagator’. If \mathbf{p}'' varies slowly when \mathbf{x}' is changed (and $|\partial \mathbf{p}''/\partial \mathbf{x}'|$ is small), the action’s phase varies slowly under a change in the end-point position, so the volume of the quantum trajectories that surround the classical trajectory is also large, and a large contribution is expected from the semi-classical propagator.

Expression (B3.4.29) is still not well suited for classical simulations due to several reasons. First, $|\partial \mathbf{p}''/\partial \mathbf{x}'|$ can vanish at specific times, which leads to infinities in the result. (In classical scattering this is related to the existence of ‘scattering rainbows’.) This is easily circumvented by changing integration parameters, from \mathbf{x}'' to \mathbf{p}' (i.e. from the final position to the initial momentum)

$$d\mathbf{x}' d\mathbf{x}'' = \left| \frac{\partial \mathbf{x}''}{\partial \mathbf{p}'} \right| d\mathbf{x}' d\mathbf{p}' \quad (\text{B 3.4.30})$$

leading to [156, 157]

$$\langle \psi | \psi(t) \rangle = \int \left| \frac{\partial \mathbf{p}''}{\partial \mathbf{x}'} \right|^{\frac{1}{2}} \psi^*(\mathbf{x}'') \psi(\mathbf{x}') e^{iS(\mathbf{x}', \mathbf{p}', t)/\hbar} d\mathbf{x}' d\mathbf{p}'. \quad (\text{B 3.4.31})$$

This transform also solves the boundary value problem, i.e. there is no need to find, for an initial position \mathbf{x}' and final position \mathbf{x}'' , the trajectory that connects the two points. Instead, one simply picks the initial momentum and position \mathbf{p}', \mathbf{x}' and calculates the classical trajectories resulting from them at all times. Such methods are generally referred to as initial variable representations (IVR).

Finally, one problem still remains. There are complex terms which need to be associated with the determinant. The complex terms (Maslov indices) have to do with the square root of the determinant, which may be negative, and also appear in the related WKB approximation. They can be calculated, albeit with difficulty

[152, 153, 154, 155, 156, 157 and 158].

Even expression (B3.4.31), although numerically preferable, is not the end of the story as it does not fully account for the fact that nearby classical trajectories (those with similar initial conditions) should be averaged over. One simple methodology for that averaging has been through the division of phase space into parts, each of which is ‘covered’ by a set of Gaussians [159, 160]. This is done by recasting the initial wavefunction as

$$\psi(\mathbf{x}) = \iint d\bar{\mathbf{x}} d\bar{\mathbf{p}} (g_{\bar{\mathbf{x}}\bar{\mathbf{p}}}|\psi\rangle g_{\bar{\mathbf{x}}\bar{\mathbf{p}}}(\mathbf{x})) \quad (\text{B 3.4.32})$$

where $g_{\bar{\mathbf{x}}\bar{\mathbf{p}}}(\mathbf{x})$ is the Gaussian that is centred at point $\bar{\mathbf{x}}$ with momentum $\bar{\mathbf{p}}$

$$g_{\bar{\mathbf{x}}\bar{\mathbf{p}}}(\mathbf{x}) = \text{constant} e^{-(\mathbf{x}-\bar{\mathbf{x}})^2/2\sigma^2} e^{i\bar{\mathbf{p}}\cdot\mathbf{x}/\hbar} \quad (\text{B 3.4.33})$$

and σ is arbitrary. By applying these formulae two times, using the semi-classical approximation and eventually summing again over phases, the following expression results [161, 162 and 163]:

$$\langle\psi|\psi(t)\rangle = \int d\mathbf{x}' d\mathbf{p}' e^{iS(\mathbf{x}',\mathbf{p}',t)/\hbar} F(\mathbf{x}',\mathbf{p}',t). \quad (\text{B 3.4.34})$$

where the exact form of F is given clearly in [163].

The resulting expression is not difficult to numerically propagate.

Numerical applications of the formalism have been very successful lately [162, 163, 164, 165, 166 and 167]. This technique is very powerful for situations where short-time propagation is sufficient. For long-time processes, where the number of required trajectories is large, it is necessary to introduce other ingredients, such as methods for reducing the total propagation time—for example, the filter diagonalization method discussed above which was applied recently to the semi-classical approximation [113, 165], or backward-forward propagation schemes which tend to make the semiclassical integrand much smoother [168].

B3.4.9 NON-ADIABATIC EFFECTS

B3.4.9.1 FORMALISM

The discussion in the previous sections assumed that the electron dynamics is adiabatic, i.e. the electronic wavefunction follows the nuclear dynamics and at every nuclear configuration only the lowest energy (or more generally, for excited states, a single) electronic wavefunction is relevant. This is the Born–Oppenheimer approximation which allows the separation of nuclear and electronic coordinates in the Schrödinger equation.

This assumption breaks down in many molecules, especially upon photo-excitation, since excited states are often close to each other or even cross one another (i.e. have the same electronic energy at a given nuclear position). Thus, the full Schrödinger wavefunction needs to be considered:

$$\Psi(\mathbf{x}, \mathbf{r}_e, t) = \sum_n \psi_n(\mathbf{x}, t) \Phi_n(\mathbf{r}_e; \mathbf{x}) \quad (\text{B 3.4.35})$$

where n is now the index of the electronic states and \mathbf{r}_e is the position of the electron. Φ_n are eigensolutions, for each position of \mathbf{x} , of the electronic part of the Hamiltonian (every part of the electron–nuclear Hamiltonian except for the nuclear kinetic energy). The associated eigenvalues are labelled u_n , and are the adiabatic ground- and excited-state energies. From this expansion there follows an equation for the nuclear wavefunction [15, 16, 17 and 18, 169]. A complication is that the adiabatic electronic states, $\Phi_n(\mathbf{r}_e; \mathbf{x})$, depend themselves on the nuclear coordinate \mathbf{x} . Thus, the nuclear kinetic-energy terms in the Schrödinger equation, which have derivatives in them ($\partial^2/\partial\mathbf{x}^2$), also operate on Φ_n . The resulting time-dependent Schrödinger equation is then straightforwardly shown to be

$$-\frac{1}{2M} \frac{\partial^2}{\partial\mathbf{x}^2} \psi_n + \sum_m \frac{\tau_{nm}^1}{2M} \frac{\partial}{\partial\mathbf{x}} \psi_m + \sum_m \frac{\tau_{nm}^2}{2M} \frac{\partial^2}{\partial\mathbf{x}^2} \psi_m + u_n \psi_n = i \frac{\partial}{\partial t} \psi_n \quad (\text{B 3.4.36})$$

where the matrices τ^1 and τ^2 are

$$\tau_{nm}^1(\mathbf{x}) = \int \Phi_n(\mathbf{r}_e; \mathbf{x}) \frac{\partial}{\partial\mathbf{x}} \Phi_m(\mathbf{r}_e; \mathbf{x}) d\mathbf{r}_e \quad (\text{B 3.4.37})$$

$$\tau_{nm}^2(\mathbf{x}) = \int \Phi_n(\mathbf{r}_e; \mathbf{x}) \frac{\partial^2}{\partial^2\mathbf{x}} \Phi_m(\mathbf{r}_e; \mathbf{x}) d\mathbf{r}_e. \quad (\text{B 3.4.38})$$

The effect of τ^2 is usually negligible, due to the $1/M$ factor. However, τ^1 is fundamentally important, yet mathematically difficult to treat. Specifically, it can be shown that τ_{nm}^1 is proportional to the inverse of the energy difference between

electronic states n and m so its value can become infinite. Several ways to simplify the equation have been developed, starting with the work of Baer [15, 16, 17 and 18, 26, 169, 170]. The mathematical theory is too intricate to discuss here in detail; it is sufficient to say that one does a ‘gauge’ transform, i.e. a transformation from adiabatic wavefunctions (the ψ_n) to a new set of ‘diabatic functions’ for which the derivative coupling (i.e. the τ^1 coefficient of $\partial/\partial\mathbf{x}$) is minimized or completely absent. Instead, a new part of the potential appears as a coupling of the electronic states (i.e. an off-diagonal potential). This coupling potential introduces a new ‘phase’ problem. The difficulty is that the diabatic functions (i.e. the functions which are obtained after the transformation) are defined in terms of a linear combination (sometimes complex) of the adiabatic functions. [169] These linear combinations are non-unique. Consider the generic example of two crossing potential surfaces in two (nuclear) dimensions (see figure B3.4.16). When the nuclei move around the potential contours, the linear combination changes. Upon return to the same starting point, the phase of the states need not be the same! Thus the nuclear diabatic functions are not uniquely defined. This is a very important effect (called the molecular phase or Berry phase [15, 16 and 17, 19, 26, 171]), since it would appear even at low energies, way below the energies in which the conical interaction appears, i.e. way below the energies of the excited states! This phenomenon has been recently shown to be important in scattering of $\mathbf{H}+\mathbf{H}_2$ and its isotopic analogy [19]. (A simple interpretation of the molecular phase phenomenon in a $\mathbf{H}+\mathbf{H}_2$ type reaction system is that the full wavefunction is symmetric under exchange of any of the two hydrogen atoms and it is antisymmetric under electron–electron exchange, so that the nuclear part of it should be antisymmetric under a change of any pair of the three nuclei. This modifies the Schrödinger equation for the atomic motion.) In

addition, it was shown in a model system that they can affect state-to-state transition probabilities in a reactive system [172].

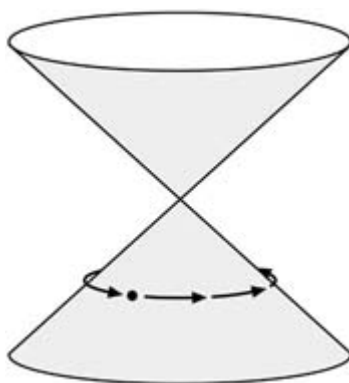


Figure B3.4.16. A generic example of crossing 2D potential surfaces. Note that, upon rotating around the conic intersection point, the phase of the wavefunction need not return to its original value.

The molecular phase effects are especially important when the system has some type of symmetry. Nevertheless, the typical treatment of non-adiabatic effects ignores the adiabatic phase, although, as cautioned, this is a problematic step.

In the remainder of this section, we will follow this simplifying (and problematic) assumption, and postulate that, upon the adiabatic to diabatic transformation, the Schrödinger equation has the form:

$$i \frac{\partial \psi_n}{\partial t} = -\frac{1}{2M} \frac{\partial^2}{\partial x^2} \psi_n + \sum_m V_{mn} \psi_m \quad (\text{B3.4.39})$$

where V_{nm} is called the diabatic potential matrix. (Note that if we were to use a finite set of molecular valence orbitals, $\Phi_n(\mathbf{r}_e)$, that do not depend on the nuclear orbitals, and expand the Schrödinger wavefunction in terms of Φ_n , we would automatically obtain this equation. Such a basis would be automatically diabatic.)

To see physically the problem of motion of wavepackets in a non-diagonal diabatic potential, we plot in figure B3.4.17 a set of two adiabatic potentials and their diabatic counterparts for a 1D problem, for example, vibrations in a diatom (as in metal–metal complexes). As figure B3.4.17 shows, if a wavepacket is started away from the ‘crossing’ point, it would slide towards this crossing point (where $V_{11} = V_{22}$) where it would branch; a part of it would continue on the same *adiabatic* state (i.e. shift to a different diabatic state) and the other part would ‘jump’ to a different adiabatic state.

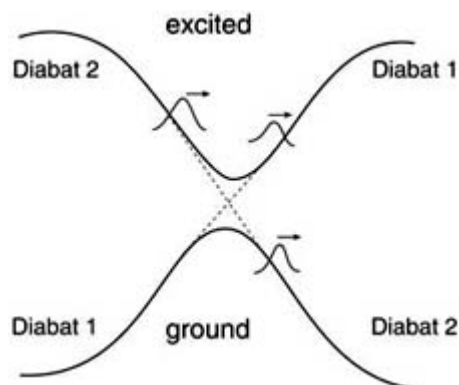


Figure B3.4.17. When a wavepacket comes to a crossing point, it will split into two parts (schematic Gaussians). One will remain on the same adiabat (different diabat) and the other will hop to the other adiabat (same diabat). The adiabatic curves are shown by full lines and denoted by ‘ground’ and ‘excited’; the diabatic curves are shown by dashed lines and denoted 1, 2.

The problem of branching of the wavepacket at crossing points is very old and has been treated separately by Landau and by Zener [15, 173, 174]. The model problem they considered has the following diabatic coupling matrix:

$$V = \begin{pmatrix} -F(R - R_0) & \Gamma \\ \Gamma & F(R - R_0) \end{pmatrix} \quad (\text{B3.4.40})$$

where $2F$ is the difference in slope of the potentials (i.e. the difference in force felt in each state), Γ is the coupling element and R_0 is the crossing point. Landau and Zener showed that, in such a case, the probability for the wavefunction to transfer from the higher adiabatic level to the lower one (i.e. to remain on the same diabat) is

$$\rho_{\text{non-adiabatic}} = \exp\left(\frac{-\pi|\Gamma|^2}{v|F|}\right) \quad (\text{B3.4.41})$$

-32-

while the probability to remain on the same adiabatic level is

$$\rho_{\text{adiabatic}} = 1 - \rho_{\text{non-adiabatic}} \quad (\text{B3.4.42})$$

where v is the velocity of the wavepacket at the crossing point. Note two things about this formula: the steeper the difference is in the potentials, the higher the probability of a non-adiabatic transfer; in addition, if the mass is large (i.e. the velocity is low), then the motion is adiabatic ($\rho_{\text{adiabatic}} \approx 1$).

B3.4.9.2 NUMERICAL APPROACHES FOR SIMULATING NON-ADIABATIC PROCESSES

The simplest approach to simulating non-adiabatic dynamics is by surface hopping [175, 176]. In its simplest form, the approach is as follows. One carries out classical simulations of the nuclear motion on a specific adiabatic electronic state (ground or excited) and at any given instant checks whether the diabatic potential associated with that electronic state is intersecting the diabatic potential on another electronic state. If it is, then a decision is made as to whether a ‘jump’ to the other adiabatic electronic state should be performed,

based on the values for $\rho_{\text{adiabatic}}$ and $\rho_{\text{non-adiabatic}}$ (when $\rho_{\text{non-adiabatic}}$ is close to 1, a jump to the other electronic state is made with a high probability so that the particle remains on the same *diabatic* potential). If a jump is made, the particle continues its motion along the new adiabatic potential surfaces, with the same instantaneous position and momentum.

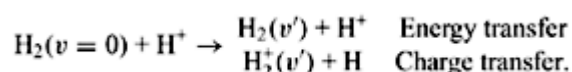
This approach is very simple and powerful. It has been used in numerous studies (for references see [176, 177]) and generally captures the essentials of the adiabatic versus non-adiabatic branching. It is especially useful in circumstances where the nuclear motion is essentially classical (i.e. zero point motion and tunnelling can be ignored).

This basic hopping model has a major disadvantage, however, as it fails to say much about the phases of the wavefunction. Consider a case where the wavefunction visits a region where surface hopping occurs, so a part of it hops, and at some later time it re-visits this region and again a part of it undergoes hopping. These two parts would interfere together and the interference may be constructive or destructive, but the hopping model does not specify this information.

To remedy this difficulty, several approaches have been developed. In some methods, the phase of the wavefunction is specified after hopping [178]. In other approaches, one expands the nuclear wavefunction in terms of a limited number of basis-set functions and works out the quantum dynamical probability for jumping. For example, the quantum dynamical basis functions could be a set of Gaussian wavepackets which move forward in time [147]. This approach is very powerful for short and intermediate time processes, where the number of required Gaussians is not too large.

The ultimate approach to simulate non-adiabatic effects is through the use of a full Schrödinger wavefunction for both the nuclei and the electrons, using the adiabatic–diabatic transformation methods discussed above. The whole machinery of approaches to solving the Schrödinger wavefunction for adiabatic problems can be used, except that the size of the wavefunction is now essentially doubled (for problems involving two-electronic states, to account for both states). The first application of these methods for molecular dynamical problems was for the charge-transfer system

-33-



Here quantum-mechanical vibrational state-to-state differential cross sections were calculated for a translational energy of $E_{\text{tr}} = 20$ eV and compared with experiments, with very good agreement between experiment and theory. In another application of this approach, state-selected integral cross sections were calculated for the $(\text{Ar} + \text{H}_2)^+$ system. Reactive (exchange), charge-transfer and spin transitions processes were treated simultaneously in one single calculation, and they compared very well with experiments [179, 180].

Finally, semi-classical approaches to non-adiabatic dynamics have also been formulated and successfully applied [167, 181]. In an especially transparent version of these approaches [167], one employs a mathematical ‘trick’ which converts the non-adiabatic surfaces to a set of coupled oscillators; the number of oscillators is the same as the number of electronic states. This method is also quite accurate, except that the number of required trajectories grows with time, as in any semi-classical approach.

B3.4.10 CONTROLLING MOLECULAR MOTION

The preceding sections were concerned with the description of molecular motion. An ambitious goal is to proceed further and influence molecular motion. This lofty goal has been at the centrepiece of quantum dynamics in the past decade and is still under intense investigation [182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193 and 194]. Here we will only describe some general concepts and schemes.

The basic Hamiltonian describing the motion of atoms and molecules under a strong laser is simple in the dipole approximation,

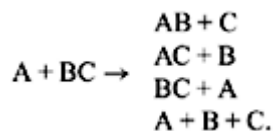
$$\hat{H} = \hat{H}_0 - \mu \cdot E(t) \quad (\text{B3.4.43})$$

where $E = E(t)$ is the time-dependent electric field at the molecule, while \hat{H}_0 is the Hamiltonian of the static system; μ is the transition dipole operator, which typically connects different electronic states.

There are several different possible goals in controlling molecular dynamics. One goal can be the localization of excitations to a specific bond in a molecule, and the molecule could be broken along that bond [188]. Alternatively, one can try to transfer the molecule completely to an excited electronic state [186]. Another is the control of alignment (so that a molecule would point in a certain direction) [189, 190]. Still another goal would be the control of branching ratios; for example, in a reaction of an atom with a diatom, $A + BC$, one may want to control the branching into

-34-

products [182, 183 and 184]:



Finally, one may want to control the emission of light from molecules.

The conceptually simplest approach towards controlling systems by laser field is by ‘teaching’ the field [188, 191, 192 and 193]. Typically, the field is experimentally prepared as, for example, a sum of Gaussian pulses with variable height and positions. Each experiment gives an outcome which can be quantified. Consider, for example, an $A + BC$ reaction where the possible products are $AB + C$ and $AC + B$; if the $AB + C$ product is preferred one would seek to optimize the branching ratio

$$P_{\text{branch}} \equiv \frac{P(A + BC \rightarrow AB + C)}{P(A + BC \rightarrow AB + C, AC + B)}. \quad (\text{B3.4.44})$$

In a purely experimental (non-theory) approach [188, 191, 192 and 193] the branching ratio can be controlled by repeating the experiment many times, each with a randomly chosen set of pulse magnitudes and start times. One can repeat the experiment, varying the electrical field somewhat each time until the best outcome is achieved. This approach maybe the most appropriate one for large systems where little is known about the underlying dynamics and it has recently been demonstrated to work very well on dissecting large molecules [188].

Closely related to these ‘experimental’ approaches are optimal control procedures, in which one simulates

theoretically the effects of the electric field on the system, and then modifies the electric field to give the best objective, i.e. a desired output (in this case: a high branching ratio) [182, 183]. The optimal control algorithm can be recast in a very powerful mathematical form which makes the calculation converge rapidly to give an excellent field for any objective, if it is possible to simulate the system motion theoretically.

A different set of approaches uses simple physical properties to control the system [184]. To demonstrate this type of problem, consider an even simpler branching problem where, upon excitation, two possible degenerate products are simultaneously produced. An example would be to photo-dissociate a diatom AB and produce different states of the system: one state labelled $A + B^*$, in which B is electronically excited and A is receding away slowly; and another state, labelled $A + B$, in which B is in the ground state and A is receding rapidly (so that the total energy is in both cases equal). The simplest method of controlling the $A + B^*$ versus $A + B$ production rate would be to mix two different pathways for obtaining $A + B$ and $A + B^*$; for example, mixing a field of frequency ω with a phase-lagged third harmonic of a field which is three times lower in frequency (see figure B3.4.18):

$$E_1 \cos \omega t + E_3 \cos \left(\frac{\omega t}{3} + \phi \right). \quad (\text{B3.4.45})$$

-35-

Other possible choices are to use two pairs of frequencies which together have the same energies. The key point is that quantum interference between the two pathways can be used to control the branching ratio. This coherent-control approach is very general and can be used in virtually any branch of molecular dynamics, including scattering and photo-dissociation.

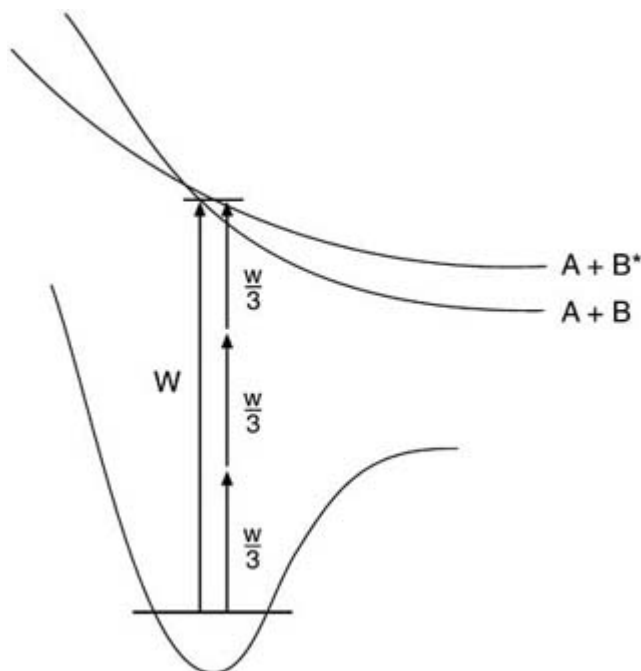


Figure B3.4.18. A schematic use of coherent control in $AB \rightarrow A + B, A + B^*$ dissociation: use of a single high-frequency photon (ω) or three low-intensity ($\omega/3$) photons would lead to emerging wavefunctions in both arrangements. However, by properly combining the amplitudes and phases of the single- and three-photon paths, the wavefunction would emerge in a single channel.

REFERENCES

- [1] Levine R D and Bernstein R B 1974 *Molecular Reaction Dynamics* (New York: Oxford University Press)
 - [2] Zhang J Z H 1999 *Theory and Application of Quantum Molecular Dynamics* (River Edge, NJ: World Scientific)
 - [3] Baer M (ed) 1985 *Theory of Chemical Reaction Dynamics* vols I–IV (Boca Raton, FL: CRC Press)
 - [4] Wyatt R E and Zhang J Z H (eds) 1996 *Dynamics of Molecules and Chemical Reactions* (New York: Dekker)
 - [5] Connor J N L (ed) 1999 Chemical reaction theory *Phys. Chem. Chem. Phys.* **1**(6) (special issue)
 - [6] Truhlar D G (ed) 1984 *Resonances* (Washington, DC: ACS)
-

-36-

- [7] Moiseyev N 1998 Quantum theory of resonances: calculating energies, widths and cross-sections by complex scaling *Phys. Rep.* **302** 212
- [8] Bowman J M 1998 Resonances: bridge between spectroscopy and dynamics *J. Phys. Chem. A* **102** 3006
- [9] Levine R D and Wu S F 1971 Resonances in reactive collisions: computational study of the $\text{H} + \text{H}_2$ collision *Chem. Phys. Lett.* **11** 557
- [10] Truong T N 1997 Thermal rates of hydrogen exchange of methane with zeolite: a direct *ab initio* dynamics study on the importance of quantum tunneling effects *J. Phys. Chem. B* **101** 2750
- [11] Asscher M, Haase G and Kosloff R 1990 Tunneling mechanism for the dissociative chemisorption of N_2 on metal surfaces *Vacuum* **41** 269
- [12] Antoniou D and Schwartz S D 1998 Activated chemistry in the presence of a strongly symmetrically coupled vibration *J. Chem. Phys.* **108** 3620
- [13] Gamarnik A, Johnson B A and Garcia-Garibay M A 1998 Effect of solvents on the photoenolization of omicron-methylanthrone at low temperatures. Evidence for H-atom tunneling from nonequilibrating triplets *J. Phys. Chem. A* **102** 5491
- [14] Jortner J and Pullman B (eds) 1986 *Tunneling* (The Netherlands: Reidel)
- [15] Baer M 1985 The theory of electronic non-adiabatic transitions in chemical reactions *Theory of Chemical Reaction Dynamics* vol II, ed M Baer (Boca Raton, FL: CRC Press) p 281
- [16] Baer M 1975 Adiabatic and diabatic representations for atom–molecule collisions: treatment of the collinear arrangement *Chem. Phys. Lett.* **35** 112
- [17] Mead C A and Truhlar D G 1982 Conditions for the definition of a strictly diabatic electronic basis for molecular systems *J. Chem. Phys.* **77** 6090
- [18] Sadygov R G and Yarkony D R 1998 On the adiabatic to diabatic states transformation in the presence of a conical intersection: a most diabatic basis from the solution to a Poisson's equation. I *J. Chem. Phys.* **109** 20
- [19] Kuppermann A and Wu Y S M 1993 The geometric phase effect shows up in chemical reactions *Chem. Phys. Lett.* **205** 577
- [20] Jackson B 1994 Quantum and semiclassical calculations of gas surface energy transfer and sticking *Comput. Phys. Commun.* **80** 119
- [21] Baer R and Kosloff R 1997 Quantum dissipative dynamics of adsorbates near metal surfaces: a surrogate Hamiltonian theory applied to hydrogen on nickel *J. Chem. Phys.* **106** 8862
- [22] Chu S 1998 The manipulation of neutral particles *Rev. Mod. Phys.* **70** 685
- [23] Cohen-Tannoudji C N 1998 Manipulating atoms with photons *Rev. Mod. Phys.* **70** 707
- [24] Phillips W D 1998 Laser cooling and trapping of neutral atoms *Rev. Mod. Phys.* **70** 721
- [25] Hagley E W, Deng L, Kozuma M, Wen J, Helmerson K, Rolston S L and Phillips W D 1999 A well-collimated quasi-continuous atom laser *Science* **283** 1706
- [26] Yarkony D R (ed) 1995 *Modern Electronic Structure Theory* (River Edge, NJ: World Scientific)
- [27] Baer M 1985 The general theory of reactive scattering: the differential equations approach *Theory of*

- [28] Neuhauser D and Baer M 1989 The time dependent Schrödinger equation: application of absorbing boundary conditions *J. Chem. Phys.* **90** 4351
- [29] Neuhauser D and Baer M 1989 The application of wavepackets to reactive atom–diatom systems: a new approach *J. Chem. Phys.* **91** 4651
- [30] Neuhauser D and Baer M 1990 A new accurate (time independent) method for treating three–dimensional reactive collisions: the application of optical potentials and projection operators *J. Chem. Phys.* **92** 3419
- [31] Neuhauser D, Baer M and Kouri D J 1990 The application of optical potentials for reactive scattering: a case study *J. Chem. Phys.* **93** 2499
- [32] Miller W H 1969 Coupled equations and the minimum principle for collisions of an atom and a diatomic molecule, including rearrangements *J. Chem. Phys.* **50** 407
- [33] Zhang J Z H and Miller W H 1989 Quantum reactive scattering via the S–matrix version of the Kohn variational principle—differential and integral cross sections for $D + H_2 \rightarrow HD + H$ *J. Chem. Phys.* **91** 1528
- [34] Manolopoulos D E, Dmello M and Wyatt R E 1989 Quantum reactive scattering via the log derivative version of the Kohn variational principle—general theory for bimolecular chemical reactions *J. Chem. Phys.* **91** 6096
- [35] Truhlar D G, Schwenke D W and Kouri D J 1990 Quantum dynamics of chemical reactions by converged algebraic variational calculations *J. Phys. Chem.* **94** 7346
- [36] Stechel E B, Walker R B and Light J C 1978 R-matrix solution of coupled equations for inelastic scattering *J. Chem. Phys.* **69** 3518
- [37] Marcus R A 1966 On the analytical mechanics of chemical reactions. Quantum mechanics of linear collisions *J. Chem. Phys.* **45** 4500
- [38] Hofacker G L 1963 Quanten chemischer reaktionen *Naturforschung* **189** 607
- [39] Walker R B, Stechel E B and Light J C 1978 Accurate H_3 dynamics on an accurate H_3 potential surface *J. Chem. Phys.* **69** 2922
- [40] Schatz G C and Kuppermann A 1975 Quantum mechanical reactive scattering: an accurate three-dimensional calculation *J. Chem. Phys.* **62** 2502
- [41] Aquilanti V and Cavalli S 1997 The quantum-mechanical Hamiltonian for tetraatomic systems in symmetric hyperspherical coordinates *J. Chem. Soc. Faraday Trans.* **93** 801
- [42] Bacic Z, Kress J D, Parker G A and Pack R T 1990 Quantum reactive scattering in 3 dimensions using hyperspherical (APH) coordinates .4. discrete variable representation (DVR) basis functions and the analysis of accurate results for $F + H_2$ *J. Chem. Phys.* **92** 2344
- [43] Pogrebnya S K, Echave J and Clary D C 1997 Quantum theory of four-atom reactions using arrangement channel hyperspherical coordinates: Formulation and application to $OH + H_2 \leftrightarrow H_2O + H$ *J. Chem. Phys.* **107** 8975
- [44] Launay J M and Ledourneuf M 1990 Quantum–mechanical calculation of integral cross sections for the reaction $F + H_2(v = 0, j = 0) \rightarrow FH(v', j') + H$ by the hyperspherical method *Chem. Phys. Lett.* **169** 473
- [45] Kuppermann A 1996 Reactive scattering with row-orthonormal hyperspherical coordinates. I. Transformation properties and Hamiltonian for triatomic systems *J. Phys. Chem.* **100** 2621
-

- [46] Neuhauser D 1990 State-to-state reactive probabilities from single-arrangement propagation with absorbing

potentials *J. Chem. Phys.* **93** 7836

- [47] Kosloff R and Kosloff D 1986 Absorbing boundaries for wave propagation problems *J. Comput. Phys.* **63** 363
- [48] Jolicard G, Leforestier C and Austin E J 1988 Resonance states using the optical potential model. Study of Feshbach resonances and broad shape resonances *J. Chem. Phys.* **88** 1026
- [49] D'Mello M, Duneczky C and Wyatt R E 1988 Recursive generation of individual S-matrix elements: application to the collinear H + H₂ reaction *Chem. Phys. Lett.* **148** 169
- [50] Kosloff R 1988 Time-dependent quantum-mechanical methods for molecular dynamics *J. Phys. Chem.* **92** 2087
- [51] Moiseyev N, Friesner R A and Wyatt R E 1986 Natural expansion of vibrational wave functions: RRGM with residue algebra *J. Chem. Phys.* **85** 331
- [52] Manthe U, Seideman T and Miller W H 1993 Full-dimensional quantum mechanical calculation of the rate constant for the H + H₂ O → H₂ + OH reaction *J. Chem. Phys.* **99** 10 078
- [53] Edlund A and Peskin U 1998 A parallel Green's operator for multidimensional quantum scattering calculations *Int. J. Quantum Chem.* **69** 167
- [54] Peskin U, Miller W H and Edlund A 1995 Quantum time evolution in time-dependent fields and time-independent reactive-scattering calculations via an efficient Fourier grid preconditioner *J. Chem. Phys.* **103** 10 030
- [55] McCullough E A and Wyatt R E 1971 Dynamics of the collinear H + H₂ reaction. I. Probability density and flux *J. Chem. Phys.* **54** 3578
- [56] De-Leon N, Davis M J and Heller E J 1984 Quantum manifestations of classical resonance zones *J. Chem. Phys.* **80** 794
- [57] Jackson B and Metiu H 1985 An examination of the use of wave packets for the calculation of atom diffraction by surfaces *J. Chem. Phys.* **83** 1952
- [58] Mowrey R C and Kouri D J 1987 Application of the close coupling wave packet method to long lived resonance states in molecule-surface scattering *J. Chem. Phys.* **86** 6140
- [59] Feit M D and Fleck J A 1983 Solution of the Schrödinger equation by a spectral method II. Vibrational energy levels of triatomic molecules *J. Chem. Phys.* **78** 301
- [60] Neuhauser D, Judson R S, Kouri D J, Adelman D E, Shafer N S, Kliner D A and Zare R N 1992 State-to-state rates for the D + H₂(v = 1, j = 1) → HD(v', j') + H reaction: predictions and measurements *Science* **257** 522
- [61] Neuhauser D, Baer M, Judson R S and Kouri D J 1989 Time-dependent three-dimensional body frame quantal wavepacket treatment of the atomic hydrogen + molecular hydrogen exchange reaction on the Liu-Siegbahn-Truhlar-Horowitz (LSTH) surface *J. Chem. Phys.* **90** 5882
- [62] Gray S K and Balint-Kurti G G 1998 Quantum dynamics with real wave packets, including application to three-dimensional (J = 0)D + H₂ → HD + H reactive scattering *J. Chem. Phys.* **108** 950
- [63] Kroes G J and Neuhauser D 1996 Performance of a time-independent scattering wave packet technique using real operators and wave functions *J. Chem. Phys.* **105** 8690
- [64] Mandelshtam V A and Taylor H S 1995 A simple recursion polynomial expansion of the Green's function with absorbing boundary conditions. Application to the reactive scattering *J. Chem. Phys.* **102**
-

- [65] Neuhauser D, Judson R S, Baer M and Kouri D J 1997 State-to-state time-dependent wavepacket approach to reactive scattering: State-resolved cross-sections for D + H₂(v = 1, j = 1, m) → H + DH(v', j'), *J. Chem. Soc. Faraday Trans.* **93** 727
- [66] Bacic Z and Light J C 1986 Highly excited vibrational levels of floppy triatomic molecules—a discrete variable representation—distributed Gaussian-basis approach *J. Chem. Phys.* **85** 4594

- [67] Colbert D T and Miller W H 1992 A novel discrete variable representation for quantum mechanical reactive scattering via the S-matrix Kohn method *J. Chem. Phys.* **96** 1982
- [68] Leforestier C *et al* 1991 A comparison of different propagation schemes for the time dependent Schrödinger equation *J. Comput. Phys.* **94** 59
- [69] McCormack D A, Kroes G J and Neuhauser D 1998 Resonance affected scattering: Comparison of two hybrid methods involving filter diagonalization and the Lanczos method *J. Chem. Phys.* **109** 5177
- [70] Neuhauser D 1992 Reactive scattering with absorbing potentials in general coordinate systems *Chem. Phys. Lett.* **200** 173
- [71] Seideman T and Miller W H 1992 Quantum mechanical reaction probabilities via a discrete variable representation-absorbing boundary condition Green function *J. Chem. Phys.* **97** 2499
- [72] Balint-Kurti G G, Dixon R N and Marston C C 1990 The Fourier grid Hamiltonian method for bound state eigenvalues and eigenfunctions *J. Chem. Soc. Faraday Trans.* **86** 1741
- [73] Tannor D J and Weeks D E 1993 Wave packet correlation function formulation of scattering theory—the quantum analog of classical S-matrix theory *J. Chem. Phys.* **98** 3884
- [74] Kouri D J, Huang Y, Zhu W and Hoffman D K 1994 Variational principles for the time-independent wave-packet-Schrödinger and wave-packet-Lippmann-Schwinger equations *J. Chem. Phys.* **100**
- [75] Neuhauser D 1994 Fully quantal initial-state-selected reaction probabilities ($J = 0$) for a four-atom system— $\text{H}_2(v = 0, 1, j = 0) + \text{OH}(v = 0, 1, j = 0) \rightarrow \text{H} + \text{H}_2\text{O}$ *J. Chem. Phys.* **100** 9272
- [76] Zhang D H and Zhang J Z H 1994 Full-dimensional time-dependent treatment for diatom-diatom reactions—the $\text{H}_2 + \text{OH}$ reaction *J. Chem. Phys.* **101** 1146
- [77] Zhu W, Dai J Q, Zhang J Z H and Zhang D H 1996 State-to-state time-dependent quantum calculation for reaction $\text{H}_2 + \text{OH} \rightarrow \text{H} + \text{H}_2\text{O}$ in six dimensions *J. Chem. Phys.* **105** 4881
- [78] Zhang D H and Light J C 1996 Quantum state-to-state reaction probabilities for the $\text{H} + \text{H}_2\text{O} \rightarrow \text{H}_2 + \text{OH}$ reaction in six dimensions *J. Chem. Phys.* **105** 1291
- [79] Goldfield E M, Gray S K and Schatz G C 1995 Quantum dynamics of a planar model for the complex forming $\text{OH} + \text{CO} \rightarrow \text{H} + \text{CO}_2$ reaction *J. Chem. Phys.* **102** 8807
- [80] Zhang D H and Zhang J Z H 1995 Quantum calculations of reaction probabilities for $\text{HO} + \text{CO} \rightarrow \text{H} + \text{CO}_2$ and bound states of HOCO *J. Chem. Phys.* **103** 6512
- [81] McCormack D A, Kroes G J, Olsen R A, Baerends E J and Mowrey R C 1999 Rotational effects on vibrational excitation of H_2 on $\text{Cu}(100)$ *Phys. Rev. Lett.* **82** 1410

- [82] Toba M, Kubo R and Saito N 1992 *Statistical Physics I. Equilibrium Statistical Mechanics* (New York: Springer)
- [83] Yamamoto T 1960 Quantum statistical mechanical theory of the rate of exchange chemical reactions in the gas phase *J. Chem. Phys.* **33** 281
- [84] Miller W H 1974 Quantum mechanical transition state theory and a new semiclassical model for reaction rate constants *J. Chem. Phys.* **61** 1823
- [85] Miller W H 1975 Semiclassical limit of quantum mechanical transition state theory for nonseparable systems *J. Chem. Phys.* **62** 1899
- [86] Miller W H, Schwartz S D and Tromp J W 1983 Quantum mechanical rate constants for bimolecular reactions *J. Chem. Phys.* **79** 4889
- [87] Pollak E and Liao J L 1998 A new quantum transition state theory *J. Chem. Phys.* **108** 2733
- [88] Seideman T and Miller W H 1992 Calculation of the cumulative reaction probability via a discrete variable representation with absorbing boundary conditions *J. Chem. Phys.* **96** 4412

- [89] Heller E J 1975 Time-dependent approach to semiclassical dynamics *J. Chem. Phys.* **62** 1544
- [90] Lee S-Y and Heller E J 1979 Time-dependent theory of Raman scattering *J. Chem. Phys.* **71** 4777
- [91] Main J, Mandelshtam V A, Wunner G and Taylor H S 1998 Harmonic inversion as a general method for periodic orbit quantization *Nonlinearity* **11** 1015
- [92] Hazi A U and Taylor H S 1970 Stabilization method of calculating resonance energies: model problem *Phys. Rev. A* **1** 1109
- [93] Mandelshtam V A, Ravuri T R and Taylor H S 1994 The stabilization theory of scattering *J. Chem. Phys.* **101** 8792
- [94] Mandelshtam V A, Taylor H S, Jung C, Bowen H F and Kouri D J 1995 Extraction of dynamics from the resonance structure of H + H₂ spectra *J. Chem. Phys.* **102** 7988
- [95] Pollak E 1985 Periodic orbits and the theory of reactive scattering *Theory of Chemical Reaction Dynamics* vol III, ed M Baer (Boca Raton, FL: CRC Press)
- [96] Schinke R, Weide K, Heumann B and Engel V 1991 Diffuse structures and periodic orbits in the photodissociation of small polyatomic molecules *Faraday Discuss. Chem. Soc.* **91** 31
- [97] Ezra G S 1996 Periodic orbit analysis of molecular vibrational spectra—spectral patterns and dynamical bifurcations in Fermi resonant systems *J. Chem. Phys.* **104** 26
- [98] Kellman M E 1997 Nonrigid systems in chemistry: a unified view *Int. J. Quantum Chem.* **65** 399
- [99] Sadeghi R and Skodje R T 1995 Barriers, thresholds and resonances—spectral quantization of the transition state for the collinear D + H₂ reaction *J. Chem. Phys.* **102** 193
- [100] Messiah A 1961 *Quantum Mechanics* (Amsterdam: North-Holland)
- [101] Reinhardt W P 1982 Complex coordinates in the theory of atomic and molecular structure and dynamics *Ann. Rev. Phys. Chem.* **35** 223
- [102] Chu S I 1991 Complex quasivibrational energy formalism for intense-field multiphoton and above-threshold dissociation—complex-scaling Fourier-grid Hamiltonian method *J. Chem. Phys.* **94** 7901
- [103] Lipkin N, Lefebvre R and Moiseyev N 1992 Resonances by complex nonsimilarity transformations of the Hamiltonian *Phys. Rev. A* **45** 4553

- [104] Moiseyev N, Certain P R and Weinhold F 1978 Resonance properties of complex-rotated Hamiltonians *Molec. Phys.* **36** 1613
- [105] Mandelshtam V A and Moiseyev N 1996 Complex scaling of *ab initio* molecular potential surfaces *J. Chem. Phys.* **104** 6192
- [106] Leforestier C and Museth K 1998 Response to ‘Comment on “On the direct complex scaling of matrix elements expressed in a discrete variable representation: application to molecular resonances” ’ *J. Chem. Phys.* **109** 1204
- [107] Museth K and Leforestier C 1996 On the direct complex scaling of matrix elements expressed in a discrete variable representation—application to molecular resonances *J. Chem. Phys.* **104** 7008
- [108] Leforestier C and Wyatt R E 1983 Optical potential for laser induced dissociation *J. Chem. Phys.* **78** 2334
- [109] Riss U V and Meyer H D 1998 The transformative complex absorbing potential method: a bridge between complex absorbing potentials and smooth exterior scaling *J. Phys. B: At. Mol. Opt. Phys.* **31** 2279
- [110] Neuhauser D 1990 Bound state eigenfunctions from wave packets—time → energy resolution *J. Chem. Phys.* **93** 2611
- [111] Neuhauser D 1994 Circumventing the Heisenberg principle—a rigorous demonstration of filter-diagonalization on a LiCN model *J. Chem. Phys.* **100** 5076
- [112] Wall M R and Neuhauser D 1995 Extraction, through filter-diagonalization, of general quantum eigenvalues or classical normal mode frequencies from a small number of residues or a short-time segment of a signal. I. Theory, and application to a quantum-dynamics model *J. Chem. Phys.* **102** 8011

- [113] Pang J W, Dieckmann T, Feigon J and Neuhauser D 1998 Extraction of spectral information from a short-time signal using filter-diagonalization: recent developments and applications to semiclassical reaction dynamics and nuclear magnetic resonance signals *J. Chem. Phys.* **108** 8360
- [114] Wall M R, Dieckmann T, Feigon J and Neuhauser D 1998 Two-dimensional filter-diagonalization: spectral inversion of 2D NMR time-correlation signals including degeneracies *Chem. Phys. Lett.* **291** 465
- [115] Mandelshtam V A and Taylor H S 1997 Spectral analysis of time correlation function for a dissipative dynamical system using filter diagonalization: application to calculation of unimolecular decay rates *Phys. Rev. Lett.* **78** 3274
- [116] Beck M H and Meyer H D 1998 Extracting accurate bound-state spectra from approximate wave packet propagation using the filter-diagonalization method *J. Chem. Phys.* **109** 3730
- [117] Chen R Q and Guo H 1996 A general and efficient filter diagonalization method without time propagation *J. Chem. Phys.* **105** 1311
- [118] Narevicius E, Neuhauser D, Korsch H J and Moiseyev M 1997 Resonances from short time complex-scaled cross-correlation probability amplitudes by the filter-diagonalization method *Chem. Phys. Lett.* **276** 250
- [119] Main J, Mandelshtam V A and Taylor H S 1997 High resolution quantum recurrence spectra: beyond the uncertainty principle *Phys. Rev. Lett.* **78** 4351
- [120] Parker G A and Pack R T 1978 Rotationally and vibrationally inelastic scattering in the rotational IOS approximation. Ultra-simple calculation of total (differential, integral and transport) cross sections for nonspherical molecules *J. Chem. Phys.* **68** 1585
- [121] Kouri D J 1977 *Atom-Molecule Collision Theory: A Guide for the Experimentalist* ed R B Bernstein (New York: Plenum)

-42-

- [122] Wang D and Bowman J M 1992 Reduced dimensionality quantum calculations of mode specificity in $\text{OH} + \text{H}_2 \rightarrow \text{H}_2\text{O} + \text{H}$ *J. Chem. Phys.* **96** 8906
- [123] Clary D C 1994 Four-atom reaction dynamics *J. Phys. Chem.* **98** 10 678
- [124] Park T J and Light J C 1989 Accurate quantum thermal rate constants for the three-dimensional $\text{H} + \text{H}_2$ reaction *J. Chem. Phys.* **91** 974
- [125] Baer M, Loesch H J, Werner H J and Last I 1994 Integral and differential cross sections for the $\text{Li} + \text{HF}$ to $\text{LiF} + \text{H}$ process. A comparison between J_z quantum mechanical and experimental results *Chem. Phys. Lett.* **219** 372
- [126] Baer M, Faubel M, Martinez-Haya B, Rusin L Y, Tappe U and Toennies J P 1998 A study of state-to-state differential state cross-sections for the: $\text{F} + \text{D}_2(\nu_i = 0, j_i) \rightarrow \text{DF}(\nu_f, j_f) + \text{D}$ reactions: a detailed comparison between experimental and three dimensional quantum mechanical results *J. Chem. Phys.* **108** 9694
- [127] Cho S W, Wagner A F, Gazdy B and Bowman J M 1992 Theoretical studies of the reactivity and spectroscopy of $\text{H} + \text{CO}$ to or from HCO . I. Stabilization and scattering studies of resonances for $J = 0$ on the Harding *ab initio* surface *J. Chem. Phys.* **96** 2812
- [128] Zhang D H and Zhang J Z H 1994 Accurate quantum calculations for $\text{H}_2 + \text{OH} \rightarrow \text{H}_2\text{O} + \text{H}$ —reaction probabilities, cross sections and rate constants *J. Chem. Phys.* **100** 2697
- [129] Heidrich D (ed) 1995 *The Reaction Path in Chemistry: Current Approaches and Perspectives* (Boston: Kluwer Academic)
- [130] Miller W H, Handy N C and Adams J E 1980 Reaction path Hamiltonian for polyatomic molecules *J. Chem. Phys.* **72** 99
- [131] Fast P L and Truhlar D G 1998 Variational reaction path algorithm *J. Chem. Phys.* **109** 3721
- [132] Billing G D 1992 Quantum classical reaction-path model for chemical reactions *Chem. Phys.* **161** 245
- [133] Zhang J Z H 1999 The semirigid vibrating rotor target model for quantum polyatomic reaction dynamics *J.*

- [134] Gerber R B, Buch V and Ratner M A 1982 Simplified time-dependent self consistent field approximation for intramolecular dynamics *Chem. Phys. Lett.* **91** 173
- [135] Peskin U and Steinberg M 1998 A temperature-dependent Schrödinger equation based on a time-dependent self consistent field approximation *J. Chem. Phys.* **109** 704
- [136] Hammes-Schiffer S 1998 Quantum dynamics of multiple modes for reactions in complex systems *Faraday Discuss. Chem. Soc.* **110** 391
- [137] Diz A, Deumens E and Ohrn Y 1990 Quantum electron-nuclear dynamics *Chem. Phys. Lett.* **166** 203
- [138] Anderson S M, Zink J I and Neuhauser D 1998 A simple and accurate approximation for a coupled system-bath: locally propagating Gaussians *Chem. Phys. Lett.* **291** 387
- [139] Anderson S M, Park T J and Neuhauser D 1999 Locally propagating Gaussians: flexible vs. frozen widths *Phys. Chem. Chem. Phys.* **1** 1343
- [140] Carter S and Bowman J M 1998 The adiabatic rotation approximation of rovibrational energies of many-mode systems: description and tests of the method *J. Chem. Phys.* **108** 4397
-

- [141] Kosloff R and Hammerich A D 1991 Nonadiabatic reactive routes and the applicability of multi configuration time dependent self consistent field approximations *Faraday Discuss. Chem. Soc.* **91** 239–47
- [142] Manthe U 1996 A time-dependent discrete variable representation for multiconfigurational Hartree methods *J. Chem. Phys.* **105** 2646
- [143] McCoy A B and Siebert E L 1996 Canonical Van Vleck perturbation theory and its applications to studies of highly vibrationally excited states of polyatomic molecules *Dynamics of Molecules and Chemical Reactions* ed R E Wyatt and J Z H Zhang (New York: Dekker) p 151
- [144] Heller E J 1981 Frozen Gaussians: a very simple semiclassical approximation *J. Chem. Phys.* **75** 2923
- [145] Blake N P and Metiu H 1995 Efficient adsorption line shape calculations for an electron coupled to many quantum degrees of freedom. applications to an electron solvated in dry sodalites and halo-sodalites *J. Chem. Phys.* **103** 4455
- [146] Markovic N and Billing G D 1997 Semi-classical treatment of chemical reactions: extension to 3D wave packets *Chem. Phys.* **224** 53
- [147] Martinez T J, BenNun M and Levine R D 1996 Multi-electronic-state molecular dynamics—a wavefunction approach with applications *J. Phys. Chem.* **100** 7884
- [148] Makri N 1990 Time-dependent self-consistent field approximation with explicit 2-body correlations *Chem. Phys. Lett.* **169** 541
- [149] Feynman R P and Hibbs A R 1965 *Quantum Mechanics and Path Integrals* (New York: McGraw-Hill)
- [150] Schulman L S 1981 *Techniques and Applications of Path Integration* (New York: Wiley)
- [151] van Vleck J H 1928 The correspondence principle in statistical interpretation of quantum mechanics *Proc. Natl. Acad. Sci.* **14** 178
- [152] Pechukas P 1969 Time-dependent semiclassical scattering theory. I. Potential scattering *Phys. Rev.* **181** 166
- [153] Miller W H 1974 Classical-limit quantum mechanics and the theory of molecular collisions *Adv. Chem. Phys.* **25** 69
- [154] Child M S 1991 *Semiclassical Mechanics with Molecular Applications* (Oxford: Clarendon)
- [155] Campolieti G and Brumer P 1994 Semiclassical propagation: phase indices and the initial-value formalism *Phys. Rev. A* **50** 997
- [156] Miller W H 1991 Comments on: Semiclassical time evolution without root searches *J. Chem. Phys.* **95** 9428
- [157] Heller E J 1991 Reply to Comments on: Semiclassical time evolution without root searches *J. Chem. Phys.* **95** 9431

- [158] Gutzwiller M C 1967 Phase-integral approximation in momentum space and the bound states of an atom *J. Math. Phys.* **8** 1979
- [159] Huber D, Ling S, Imre D G and Heller E J 1989 Hybrid mechanics *J. Chem. Phys.* **90** 7317
- [160] Herman M F and Kluk E 1984 A semiclassical justification for the use of non-spreading wavepackets in dynamics calculations *Chem. Phys.* **91** 27
- [161] Herman M F, Kluk E and Davis H L 1986 Comparison of the propagation of semiclassical frozen Gaussian wave functions with quantum propagation for a highly excited anharmonic oscillator *J. Chem. Phys.* **84** 326
-

-44-

- [162] Kay K G 1994 Semiclassical propagation for multidimensional systems by an initial value method *J. Chem. Phys.* **101** 2250
- [163] Walton A R and Manolopoulos D E 1996 A new semiclassical initial value method for Franck-Condon spectra *Mol. Phys.* **87** 961
- [164] Sun X, Wang H B and Miller W H 1998 Semiclassical theory of electronically nonadiabatic dynamics: Results of a linearized approximation to the initial value representation *J. Chem. Phys.* **109** 7064
- [165] Grossmann F, Mandelshtam V A, Taylor H S and Briggs H S 1997 Harmonic inversion of semiclassical short time signals *Chem. Phys. Lett.* **279** 355
- [166] Thompson K and Makri N 1999 Influence functionals with semiclassical propagators in combined forward-backward time *J. Chem. Phys.* **110** 1343
- [167] Stock G and Thoss M 1997 Semiclassical description of nonadiabatic quantum dynamics *Phys. Rev. Lett.* **78** 578
- [168] Thompson K and Makri N 1999 Rigorous forward-backward semiclassical formulation of many-body dynamics *Phys. Rev. E* **59** 4729
- [169] Baer M, Yahalom A and Engelman R 1998 Time-dependent and time-independent approaches to study effects of degenerate electronic states *J. Chem. Phys.* **109** 6550
- [170] Mead C A and Truhlar D G 1979 On the determination of Born-Oppenheimer nuclear motion wave functions including complications due to conical intersections and identical nuclei *J. Chem. Phys.* **70** 2284
- [171] Berry M 1990 Anticipations of the geometric phase *Physics Today* **43** 34
- [172] Baer M, Charutz D M, Kosloff R and Baer M 1996 A study of conical intersection effects on scattering processes—the validity of adiabatic single-surface approximations within a quasi-Jahn-Teller model *J. Chem. Phys.* **105** 9141
- [173] Landau L D 1932 Zur theorie der Energieübertragung bei Stößen. II *Phys. Zts. Sowjet.* **2** 46
- [174] Zener C 1932 Nonadiabatic crossing of energy levels *Proc. R. Soc. A* **137** 696
- [175] Tully J C and Preston R K 1971 Trajectory surface hopping approach to nonadiabatic molecular collisions: the reaction of H⁺ with D₂ *J. Chem. Phys.* **55** 562
- [176] Hammes-Schiffer S and Tully J C 1995 Nonadiabatic transition state theory and multiple potential energy surfaces molecular dynamics of infrequent events *J. Chem. Phys.* **103** 8528
- [177] Niv M Y, Krylov A I and Gerber R B 1997 Photodissociation, electronic relaxation and recombination of HCl in Ar-n(HCl) clusters—non-adiabatic molecular dynamics simulations *Faraday Discuss. Chem. Soc.* **108** 243-54
- [178] Zhu C and Nakamura H 1994 Theory of nonadiabatic transition for general curved potentials I. *J. Chem. Phys.* **101** 10 630
- [179] Baer M 1992 *State Selected and State-to-State Ion-Molecule Reaction Dynamics. Part II: Theory* ed M Baer and C Y Ng (New York: Wiley)
- [180] Baer M, Niedner-Schattner G and Toennies J P 1989 A 3-dimensional quantum mechanical study of vibrationally resolved charged transfer processes in H⁺+ H₂ at E_{CM} = 20 eV *J. Chem. Phys.* **91** 4169
- [181] Meyer H D and Miller W H 1979 A classical analog for electronic degrees of freedom in nonadiabatic

- [182] Tannor D J, Kosloff R and Rice S A 1986 Coherent pulse sequence induced control of selectivity of reactions: exact quantum mechanical calculations *J. Chem. Phys.* **85** 5805
- [183] Neuhauser D and Rabitz H 1993 Paradigms and algorithms for controlling molecular motion *Acc. Chem. Res.* **26** 496
- [184] Seideman T, Shapiro M and Brumer P 1989 Coherence chemistry—controlling chemical reactions with lasers *Chem. Phys.* **90** 7132
- [185] Smith T J and Cina J A 1996 Toward preresonant impulsive Raman preparation of large amplitude *J. Chem. Phys.* **104** 1272
- [186] Krause J L, Messina M, Wilson K R and Yan Y J 1995 Quantum control of molecular dynamics—the strong response regime *J. Phys. Chem.* **99** 13 736
- [187] Meyer S and Engel V 1997 Vibrational revivals and the control of photochemical reactions *J. Phys. Chem. A* **101** 7749
- [188] Assion A, Baumert T, Bergt M, Brixner T, Kiefer B, Seyfried V, Strehle M and Gerber G 1998 Control of chemical reactions by feedback-optimized phase-shaped femtosecond laser pulses *Science* **282** 919
- [189] Althorpe S C and Seideman T 1999 Molecular alignment from femtosecond time-resolved photoelectron angular distributions: nonperturbative calculations on NO *J. Chem. Phys.* **110** 147
- [190] Friedrich B and Herschbach D 1996 Alignment enhanced spectra of molecules in intense non-resonant laser fields *Chem. Phys. Lett.* **262** 41
- [191] Judson R S and Rabitz H 1992 Teaching lasers to control molecules *Phys. Rev. Lett.* **68** 1500
- [192] Bardeen C J, Yakovlev V V, Wilson K R, Carpenter S D, Weber P M and Warren W S 1997 Feedback quantum control of molecular electronic population transfer *Chem. Phys. Lett.* **280**
- [193] Zare R N 1998 Laser control of chemical reactions *Science* **279** 1875
- [194] Bigwood R, Gruebele M, Leitner D M and Wolynes P G 1998 The vibrational energy flow transition in organic molecules: theory meets experiment *Proc. Natl Acad. Sci.* **95** 5960
- [195] Truhlar D G and Horowitz C J 1978 Functional representation of Liu and Siegbahn's accurate *ab initio* potential energy calculations for H + H₂ *J. Chem. Phys.* **68** 2466
- [196] Kroes G J, Wiesenekker G, Baerends E J, Mowrey R C and Neuhauser D 1996 Dissociative chemisorption of H₂ on Cu(100)—a four-dimensional study of the effect of parallel translational motion on the reaction dynamics *J. Chem. Phys.* **105** 5979

B 3.5 Optimization and reaction path algorithms

Peter Pulay and Jon Baker

B 3.5.1 INTRODUCTION

A quantum mechanical treatment of molecular systems usually starts with the Born–Oppenheimer approximation, i.e., the separation of the electronic and nuclear degrees of freedom. This is a very good approximation for well separated electronic states. The expectation value of the total energy in this case is a function of the nuclear coordinates and the parameters in the electronic wavefunction, e.g., orbital coefficients. The wavefunction parameters are most often determined by the variation theorem: the electronic energy is made stationary (in the most important ground-state case it is minimized) with respect to them. The

optimized energy, calculated as a function of the nuclear coordinates, is known as the potential energy surface (PES). Finding its (local) minimum gives the calculated equilibrium structure. The latter, although experimentally not directly accessible, is perhaps the most satisfactory definition of molecular geometry, and serves as the best starting point for a treatment of molecular vibrations. A large part of the total effort expended in quantum chemical calculations is spent in optimizing either electronic parameters or molecular geometries; the latter task is dominant in empirical force field calculations. The optimization of electronic wavefunctions is usually treated separately from the optimization of molecular geometries; however, there are enough similarities between the two problems to discuss them together.

Both the electronic and the geometry optimization problem, particularly the latter, may have more than one solution. For small, rigid molecules, the approximate molecular geometry is chemically obvious, and the presence of multiple minima is not a serious concern. For large, flexible molecules, however, finding the absolute minimum, or a complete set of low-lying equilibrium structures, is only a partially solved problem. This topic will be discussed in the last section of this chapter. The rest of the article deals with local optimization, i.e., finding a minimum from a reasonably close starting point. We will also discuss the determination of other stationary points—most importantly saddle points—constrained optimization, and reaction paths. Several reviews have been published on geometry optimization [1, 2]. The optimization of SCF-type wavefunctions is often highly nonlinear, particularly for the multiconfigurational case, and this has received most attention [3, 4].

The most important consideration affecting the choice of the method for locating minima, or stationary points in general, is the availability of analytical derivatives of the object function, in our case the energy. Zeroth-order (energy only) methods can be used for a few variables but are notoriously inefficient for a larger number of degrees of freedom. First-order methods, which use both energy and gradient (first-derivative) information, are particularly useful in quantum chemistry because the extra effort needed to evaluate all first derivatives is usually comparable to the calculation of the energy itself and may be less, particularly for the electronic degrees of freedom [5]. Second-order methods, which use second derivatives, further improve the convergence of the optimization process. However, calculating second derivatives tends to be much more expensive than calculating the gradient, and full second-order methods are usually cost efficient only when first-order methods have severe convergence problems. Derivatives higher than the second have been used occasionally, but they are not generally available and are expensive to calculate. Consequently, this article will mainly concentrate on first- and second-order methods.

-2-

The electronic energy W in the Born–Oppenheimer approximation can be written as $W = W(\mathbf{q}, \mathbf{p})$, where \mathbf{q} is the vector of nuclear coordinates and the vector \mathbf{p} contains the parameters of the electronic wavefunction. The latter are usually orbital coefficients, configuration amplitudes and occasionally nonlinear basis function parameters, e.g., atomic orbital positions and exponents. The electronic coordinates have been integrated out and do not appear in W . Optimizing the electronic parameters leaves a function depending on the nuclear coordinates only, $E = E(\mathbf{q})$. We will assume that both $W(\mathbf{q}, \mathbf{p})$ and $E(\mathbf{q})$ and their first derivatives are continuous functions of the variables q_i and p_j .

B 3.5.2 OVERVIEW OF TECHNIQUES FOR LOCAL OPTIMIZATION

B3.5.2.1 CHARACTERIZATION OF STATIONARY POINTS

In this section, we will discuss general optimization methods. Our example is the geometry optimization problem, i.e., the minimization of $E(\mathbf{q})$. However, the results apply to electronic optimization as well. There are a number of useful monographs on the minimization of continuous, differentiable functions in many variables [6, 7].

For a point on the potential energy surface to be a stationary point, all its first derivatives, $\partial E/\partial q_i$, must vanish, and thus the whole gradient vector $\mathbf{g} = \{\partial E/\partial q_i, i = 1, n\}$ should be zero. The character of a stationary point, i.e., whether it is a local minimum, maximum or saddle point, can be determined by examining the second derivatives. Expanding the energy change in the neighbourhood of a stationary point \mathbf{q}^0 in a power series in terms of displacement coordinates from the stationary point, $\delta q_i = q_i - q_i^0$, gives

$$E(\mathbf{q}) - E(\mathbf{q}^0) = \sum_i (\partial E/\partial q_i)|_{\mathbf{q}^0} \delta q_i + 1/2 \sum_{i,j} (\partial^2 E/\partial q_i \partial q_j)|_{\mathbf{q}^0} \delta q_i \delta q_j + \text{higher terms.} \quad (\text{B3.5.1})$$

As \mathbf{q}^0 is a stationary point, the linear terms $(\partial E/\partial q_i)|_{\mathbf{q}^0} \delta q_i$ in equation B3.5.1 vanish, and higher-order terms do not have to be considered for local characterization of stationary points, because lower-order terms (quadratics in this case) always dominate for sufficiently small displacements. Introducing the force constant, or Hessian matrix, i.e., the matrix of second derivatives \mathbf{H} , $H_{ij} = (\partial^2 E/\partial q_i \partial q_j)|_{\mathbf{q}^0}$, the above equation can be written in a convenient matrix notation as

$$\Delta E = E(\mathbf{q}) - E(\mathbf{q}^0) = \frac{1}{2} \delta \mathbf{q}^\dagger \mathbf{H} \delta \mathbf{q}. \quad (\text{B3.5.2})$$

Let us express the displacement coordinates as linear combinations of a set of new coordinates \mathbf{y} : $\delta \mathbf{q} = \mathbf{U}\mathbf{y}$; then $\Delta E = \mathbf{y}^\dagger \mathbf{U}^\dagger \mathbf{H} \mathbf{U} \mathbf{y}$. \mathbf{U} can be an arbitrary non-singular matrix, and thus can be chosen to diagonalize the symmetric matrix \mathbf{H} : $\mathbf{U}^\dagger \mathbf{H} \mathbf{U} = \Lambda$, where the diagonal matrix Λ contains the (real) eigenvalues of \mathbf{H} . In this form, the energy change from the stationary point is simply $\Delta E = \frac{1}{2} \sum_i \Lambda_i y_i^2$. It is clear now that a sufficient condition for a minimum is that all eigenvalues of \mathbf{H} be positive, i.e., \mathbf{H} must be a positive definite matrix. Otherwise choosing $y_i \neq 0$, all other $y_j = 0$, where Λ_i is a negative eigenvalue, will decrease the energy, i.e., the stationary point cannot be a minimum. Zero eigenvalues of the Hessian (inflection points) need not be considered because their probability in the general case is

-3-

vanishingly small. Stationary points with only one negative Hessian eigenvalue are called first-order saddle points. They have considerable importance as transition states in chemical reactions. The energy difference between a transition state and the reactant(s) is the barrier corresponding to the reaction path passing through that transition state. Stationary points with two or more negative eigenvalues are far less important, as in this case there is always a reaction path with lower barrier which determines the reaction probability (however, in symmetrical systems higher-order saddle points may be preferred reference geometries[8]).

The simplest smooth function which has a local minimum is a quadratic. Such a function has only one, easily determinable stationary point. It is thus not surprising that most optimization methods try to model the unknown function with a local quadratic approximation, in the form of [equation \(B3.5.1\)](#).

B3.5.2.2 ENERGY-ONLY METHODS

As noted in the introduction, energy-only methods are generally much less efficient than gradient-based techniques. The simplex method [9] (not identical with the similarly named method used in linear programming) was used quite widely before the introduction of analytical energy gradients. The intuitively most obvious method is a sequential optimization of the variables (sequential univariate search). As the optimization of one variable affects the minimum of the others, the whole cycle has to be repeated after all variables have been optimized. A one-dimensional minimization is usually carried out by finding the

minimum of a parabola fitted to three points obtained by varying one of the variables (keeping the others constant), changing values so as to bracket the minimum, and zeroing in on the minimum by diminishing the step size [6]. Generalized to any vector direction on the surface, this is called a line search. The convergence rate of the sequential univariate search can be exceedingly slow if the variables are strongly coupled and, thus, this method is not recommended. A better alternative is to convert gradient methods, covered in the next section, to energy-only methods by calculating the gradients numerically. One of the most widely used energy-only methods is the modified Fletcher–Powell method described by Schlegel [1]; perhaps better is the numerical version of Baker’s Eigenvector Following (EF) algorithm (see later) [10]. In spite of these ingenious algorithms, the general consensus among researchers in the field is that energy-only methods are simply not cost effective for systems with more than a few degrees of freedom.

B3.5.2.3 GRADIENT METHODS

All efficient optimization methods require the gradient vector, i.e., the first derivatives of the function to be optimized. As the quantum mechanical energy as a function of nuclear coordinates is the result of an iterative procedure, and ordinary first-order perturbation theory is inapplicable in the usual case where the basis functions move with the nuclei, this is not a trivial problem. The introduction of analytical energy derivatives (forces on the atoms in the context of geometry optimization) of the SCF energy [11], and later generalizations to more complex wavefunctions (for reviews see, e.g., [5, 12]) improved the efficiency of geometry optimizations by one or two orders of magnitude, depending on the molecular size, making possible structure optimization for large polyatomic molecules.

(A) NEWTON’S METHOD

Most gradient optimization methods rely on a quadratic model of the potential surface. The minimum condition for the

-4-

quadratic energy expression

$$E(\mathbf{q}) = E(\mathbf{q}^0) + \delta\mathbf{q}^\dagger \mathbf{g} + \frac{1}{2} \delta\mathbf{q}^\dagger \mathbf{H} \delta\mathbf{q}$$

using the symmetry of \mathbf{H} , leads to

$$\mathbf{g} + \mathbf{H} \delta\mathbf{q} = \mathbf{0} \text{ and } \delta\mathbf{q} = -\mathbf{H}^{-1} \mathbf{g}. \quad (\text{B3.5.3})$$

Here \mathbf{g} is the gradient vector at \mathbf{q}^0 , $\mathbf{g}_i = (\partial E / \partial q_i)|_{\mathbf{q}^0}$. The minimizer $\mathbf{q}^0 + \delta\mathbf{q}$ is exact on a quadratic surface and requires only the solution of a linear system of equations. For a nonlinear surface, this method has to be iterated. Near the solution, the iterative procedure is quadratically convergent. In practice, this means that three or four iterations usually suffice for locating the minimum to high accuracy. This is the basic Newton (or Newton–Raphson) method. Despite its rapid convergence for nearly quadratic surfaces, it suffers from a number of shortcomings and in its original form is seldom used for geometry optimization. It has some importance for difficult cases of wavefunction optimization. The principal defect of Newton’s method is that it requires the Hessian (second-derivative) matrix at every iteration. Second derivatives are typically much more expensive than gradients in quantum chemistry applications. Another problem is that far from the minimum the Hessian may not be positive definite. The energy change in a Newton–Raphson cycle is $\delta\mathbf{q}^\dagger \mathbf{g} = -\frac{1}{2} \mathbf{g}^\dagger \mathbf{H}^{-1} \mathbf{g}$ in the quadratic approximation. Thus the energy does not necessarily decrease, even for small steps, unless \mathbf{H} (and obviously its inverse) is positive definite.

(B) SIMPLE RELAXATION

Both defects of the Newton method can be eliminated by replacing the exact inverse Hessian \mathbf{H}^{-1} by a (fixed) positive definite approximation to it, \mathbf{F} . This method is known as simple relaxation. In both geometry and wavefunction optimization, it is usually possible to construct a fairly good approximate Hessian. For geometry optimization, this can be based on the molecular connectivity and transferability of potential parameters, or on previous low-level calculations. For wavefunction optimization, a guess based on orbital energy differences is often reasonably accurate. Far from the minimum, approximate Hessian methods using positive definite matrices are preferable to the Newton method, as they have the descent property, i.e., the energy decreases for sufficiently small steps. However, they lack the quadratic terminal convergence rate of the Newton method. Instead, the residual error vector (the distance from the accurate minimum) is given by $\mathbf{r}^{(n)} = (\mathbf{I} - \mathbf{FH})^n \mathbf{r}^{(0)}$ on a quadratic surface. Here \mathbf{I} is the unit matrix and $^{(n)}$ denotes the n th cycle. The ultimate convergence rate is governed by the magnitude of the largest eigenvalue of the matrix $(\mathbf{I} - \mathbf{FH})$. This will be small if \mathbf{F} is a good approximation to \mathbf{H}^{-1} . To show this, we introduce new variables, through a linear transformation of the old ones, which diagonalize \mathbf{FH} . Using these coordinates, the k th component of the residue in step n is $\lambda_k^n \mathbf{r}_k^{(0)}$ where λ_k is the k th eigenvalue of $(\mathbf{I} - \mathbf{FH})$. This explains a common property of simple relaxation: it usually shows good initial convergence but slows down later as the surviving components of the residuum take on directions in which the Hessian is poorly estimated. If one of the eigenvalues of $(\mathbf{I} - \mathbf{F}^{-1}\mathbf{H})$ exceeds 1 in absolute magnitude then simple relaxation without a line search will ultimately diverge.

If there is no approximate Hessian available, then the unit matrix is frequently used, i.e., a step is made along the gradient. This is the steepest descent method. The unit matrix is arbitrary and has no invariance properties, and thus the

-5-

resulting step may be made arbitrarily large or small by scaling the coordinates. Therefore, steepest descent methods require a line search for a minimum along the direction of the gradient vector. Line searches are often recommended in general optimization texts. However, they tend to be less efficient in quantum chemistry, as the evaluation of the gradient vector costs roughly the same as the calculation of the energy for a wide range of methods, and supplies much more information. Nevertheless, they may be necessary for strongly non-quadratic functions (or, what is essentially the same thing, at points far from the minimum). A good compromise which requires no additional energy evaluations was suggested by Schlegel [13]: a polynomial is fitted to the energies at two points, and to the gradients projected on the line connecting them, and its minimum is located. The polynomial can be cubic or, as recommended by Schlegel, a special quartic with only one minimum. If a line search is used, the energy of simple relaxation and steepest descent steps should always decrease, and they should ultimately converge. However, convergence may be very slow if \mathbf{F} is a poor approximation to the inverse Hessian \mathbf{H}^{-1} , as is usually the situation in the steepest descent method. In this case, illustrated in figure B3.5.1 the optimization takes a zigzag path converging slowly to the minimum.

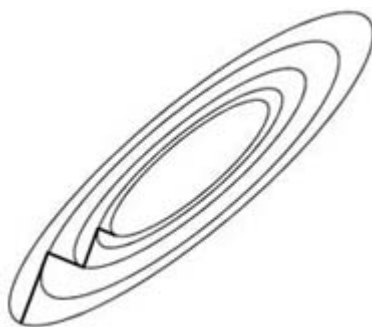


Figure B3.5.1. Contour line representation of a quadratic surface and part of a steepest descent path zigzagging toward the minimum.

In simple relaxation (the fixed approximate Hessian method), the step does not depend on the iteration history. More sophisticated optimization techniques use information gathered during previous steps to improve the estimate of the minimizer, usually by invoking a quadratic model of the energy surface. These methods can be divided into two classes: variable metric methods and interpolation methods.

(C) VARIABLE METRIC METHODS

In these methods, also known as quasi-Newton methods, the approximate Hessian is improved (updated) based on the results in previous steps. For the exact Hessian and a quadratic surface, the quasi-Newton equation $\Delta \mathbf{g}^{(n)} = \mathbf{H} \Delta \mathbf{q}^{(n)}$ and its analogue $\mathbf{H}^{-1} \Delta \mathbf{g}^{(n)} = \Delta \mathbf{q}^{(n)}$ must hold (where $\Delta \mathbf{g}^{(n)} = \mathbf{g}^{(n+1)} - \mathbf{g}^{(n)}$, and similarly for $\Delta \mathbf{q}^{(n)}$). These equations, which have only n components, are obviously insufficient to determine the $n(n+1)/2$ independent components of the Hessian or its inverse. Therefore, the updating is arbitrary to a certain extent. It is desirable to have an updating scheme that converges to the exact Hessian for a quadratic function, preserves the quasi-Newton conditions obtained in previous steps, and—for minimization—keeps the Hessian positive definite. Updating can be performed on either \mathbf{F} or its inverse, the approximate Hessian. In the former case repeated matrix inversion can be avoided. All updates use dyadic products, usually built from $\Delta \mathbf{q}^{(n)}$ and $\mathbf{F} \Delta \mathbf{g}^{(n)}$. Fletcher [6] gives a detailed description of various update techniques. The most important update formulae are the Murtagh–Sargent update [14]:

-6-

$$\mathbf{F}^{(n+1)} = \mathbf{F}^{(n)} + (\Delta \mathbf{q}^{(n)} - \mathbf{F}^{(n)} \Delta \mathbf{g}^{(n)}) (\Delta \mathbf{q}^{(n)} - \mathbf{F}^{(n)} \Delta \mathbf{g}^{(n)})^T / \{ (\Delta \mathbf{q}^{(n)} - \mathbf{F}^{(n)} \Delta \mathbf{g}^{(n)})^T \Delta \mathbf{g}^{(n)} \}$$

the Davidon–Fletcher–Powell (DFP) update [15]:

and the Broyden–Fletcher–Goldfarb–Shanno (BFGS) update [6]:

$$\mathbf{F}^{(n+1)} = \mathbf{F}^{(n)} + s(1 + s \Delta \mathbf{g}^{(n)T} \mathbf{F}^{(n)} \Delta \mathbf{g}^{(n)}) \Delta \mathbf{q}^{(n)} \Delta \mathbf{q}^{(n)T} - s(\Delta \mathbf{q}^{(n)} \Delta \mathbf{g}^{(n)T} \mathbf{F}^{(n)} + \mathbf{F}^{(n)} \Delta \mathbf{g}^{(n)} \Delta \mathbf{q}^{(n)T})$$

where $s = 1/(\Delta \mathbf{q}^{(n)T} \Delta \mathbf{g}^{(n)})$. A linear combination of these updates is also possible. Both the DFP and BFGS updates preserve positive definite \mathbf{F} matrices provided $\Delta \mathbf{q}^{(n)T} \Delta \mathbf{g}^{(n)} > 0$; current opinion is that the latter is the best update to use for general minimization.

For transition state searches, none of the above updates is particularly appropriate as a positive definite Hessian is not desired. A more useful update in this case is the Powell update [16]:

$$\mathbf{H}^{(n+1)} = \mathbf{H}^{(n)} + \{ \mathbf{v} \Delta \mathbf{q}^{(n)T} + \Delta \mathbf{q}^{(n)} \mathbf{v}^T - (\mathbf{v}^T \Delta \mathbf{q}^{(n)T}) (\Delta \mathbf{q}^{(n)} \Delta \mathbf{q}^{(n)T}) / t \} / t$$

where $\mathbf{v} = \Delta \mathbf{g}^{(n)} - \mathbf{H}^{(n)} \Delta \mathbf{q}^{(n)}$ and $t = (\Delta \mathbf{q}^{(n)T} \Delta \mathbf{q}^{(n)})$. The Powell update allows the signature of the Hessian, i.e., the number of negative eigenvalues, to change, which is necessary if the region of the potential energy surface is inappropriate for the stationary point being sought. Perhaps the best Hessian update for transition state searches is a linear combination of the Powell and Murtagh–Sargent updates proposed by Bofill [17, 18].

For a very large number of variables, the question of storing the approximate Hessian or inverse Hessian \mathbf{F} becomes important. Wavefunction optimization problems can have a very large number of variables, a million or more. Geometry optimization at the force field level can also have thousands of degrees of freedom. In these cases, the initial inverse Hessian is always taken to be diagonal or sparse, and it is best to store the

upgrade vectors and associated scalars and generate the inverse Hessian *in situ*, rather than store the full updated inverse Hessian itself.

A more general update method, widely used in the Gaussian suite of programs [19], is due to Schlegel [13]. In this method, the Hessian in the n -dimensional subspace spanned by taking differences between the current $\mathbf{q}^{(n)}$ and previous geometries $\mathbf{q}^{(n-1)}, \dots, \mathbf{q}^{(0)}$ is calculated numerically. This is possible (although not terribly accurate), as the $n \Delta \mathbf{q}$ and $n \Delta \mathbf{g}$ values suffice for the calculation of an n -dimensional Hessian by forward differences. The Hessian in the small subspace is then projected back to the full space. A line search along the new correction vector is avoided by using the constrained quartic interpolation scheme described above.

(D) INTERPOLATION METHODS

For a quadratic surface, the gradient vector is a linear function of the coordinates. An alternative way of using

-7-

information gathered during the optimization is to interpolate among the coordinate vectors obtained in the preceding cycles. The basic interpolation method is the preconditioned conjugate gradient (CG) method [20]. Although usually formulated in a different way, it is equivalent to first making a simple relaxation step using an approximate inverse Hessian \mathbf{F} , called the preconditioner, and replacing the calculated displacement by a linear combination of the current and all previous coordinate displacement vectors $\Delta \mathbf{q}^{(i)}$:

$$\Delta \mathbf{q}^{(n+1)} = -\mathbf{F} \mathbf{g}^{(n+1)} + \sum_{i=1}^n \beta_i \Delta \mathbf{q}^{(i)}.$$

The coefficients β_i are chosen so that, on a quadratic surface, the interpolated gradient becomes orthogonal to all $\Delta \mathbf{q}^{(i)}$. This condition is equivalent to minimizing the energy in the space spanned by the displacement vectors. In the quadratic case, a further simplification can be made as it can be shown that all β_i with the exception of β_n vanish. The latter is given by

$$\beta_n = \mathbf{g}^{(n+1)\text{T}} \mathbf{F} \mathbf{g}^{(n+1)} / (\mathbf{g}^{(n)\text{T}} \Delta \mathbf{q}^{(n)})$$

although there are several forms which are equivalent for a quadratic function but not in general [6], e.g., the Polak–Ribiere form [21]. The gradient at the (interpolated) new point is not recalculated but is itself interpolated. For very large problems, the conjugate gradient method has the advantage that it needs to store only a few vectors (the preconditioner is usually diagonal or sparse, and must be positive definite).

A similar method, direct inversion in the iterative subspace (DIIS) [22, 23] tries to minimize the norm of the error vector (in most cases the gradient) by interpolating in the subspace spanned by the previous vectors. Unlike the CG method, DIIS is able to converge to saddle points. DIIS is now the standard method for the SCF optimization problem. It is also useful for geometry optimization [24]. It does not have the conjugate property and therefore requires the storage of previous coordinate and gradient vectors (in practice, usually restricted to about 20 or fewer). However, not using the CG property, which is valid for quadratic surfaces only, probably adds to the stability of the method.

To derive the DIIS equations, let us consider a linear combination of coordinate vectors $\mathbf{q}^{(1)}, \dots, \mathbf{q}^{(n)}$, $\mathbf{q} = \sum_i^n c_i \mathbf{q}^{(i)}$. On a quadratic surface, the gradient (or any linear function of the gradient) is an analogous linear combination if the coefficients sum to unity:

$$\mathbf{g} = \sum_i^n c_i \mathbf{g}^{(i)}. \quad (\text{B3.5.4})$$

Minimizing the square of the gradient vector under the condition $\sum_i^n c_i = 1$ yields the following linear system of equations

-8-

$$\begin{pmatrix} B_{11} & \dots & B_{1n} & -1 \\ \vdots & & \vdots & \vdots \\ B_{n1} & \dots & B_{nn} & -1 \\ -1 & \dots & -1 & 0 \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_n \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -1 \end{pmatrix}$$

where $B_{ij} = \mathbf{g}^{(i)\text{T}} \mathbf{g}^{(j)}$. Due to the wide dynamic range of the B_{ij} coefficients, it is best to normalize the diagonal elements of this equation to 1. This procedure yields an extrapolated geometry and [gradient B3.5.4](#). The next step is calculated by relaxing the extrapolated gradient with the approximate inverse Hessian and adding it to the extrapolated geometry. Hessian updating can be combined with DIIS, but the method works well even with a static Hessian. Any linear function of the gradient can be used instead of the gradient. This changes the weighting of the error somewhat.

B3.5.2.4 SECOND-ORDER METHODS

As mentioned in the introduction, full second-order methods (e.g., the Newton method) are usually not cost efficient, particularly for geometry optimization, due to the high cost of Hessian evaluation. For wavefunction optimization, the explicit evaluation of the full Hessian is not practicable due to the large number of degrees of freedom. Second-order methods can still be utilized by using direct methods, i.e., finding the solution of the Newton–Raphson equation $\mathbf{H}\delta\mathbf{q} = -\mathbf{g}$ without explicitly constructing and inverting \mathbf{H} . In such a case, second-order methods are competitive, and in difficult cases superior to first-order methods in the quadratic region, i.e., close to the minimum. Optimization of transition states also frequently requires the explicit evaluation of the Hessian or a submatrix of it.

B3.5.2.5 DAMPING METHODS

Particularly in the early stages of an optimization, when the gradient is large and Hessian information is inaccurate, the computed step size may be too great; using such large steps may lead to divergence or convergence to unwanted minima. Methods which incorporate a line search are usually immune to this problem; however, as discussed above, line searches are inefficient in quantum chemistry because the evaluation of the full gradient vector can often take *less* time than the evaluation of a single energy.

The simplest way to deal with large, potentially disastrous steps is to limit the step size, either its maximum component or its norm. For geometry optimization, 0.2 to 0.3 Å or rad appears to be a reasonable value for a maximum single component; 0.3 rad is also appropriate for the maximum orbital rotation component in wavefunction optimization. A better method than simply scaling the displacement is to use a trust radius. The idea behind the trust radius is to restrict the step taken so that it lies in the local region of the energy surface where the truncation of the original power series expansion ([equation \(B3.5.1\)](#)) to quadratic terms only is valid. The neighbourhood about the current point where quadratic behaviour holds is called the trust region.

If the computed step size exceeds the trust radius, t , its direction is reoptimized under the condition that $|\Delta\mathbf{q}| = t$, i.e., the Lagrangian

$$E(\mathbf{q}, d) = E(\mathbf{q}^0) + \Delta\mathbf{q}^T \mathbf{g} + \frac{1}{2} \Delta\mathbf{q}^T \mathbf{H} \Delta\mathbf{q} + \frac{1}{2} d (|\Delta\mathbf{q}|^2 - t^2)$$

-9-

is minimized. The solution is $\Delta\mathbf{q} = -(\mathbf{H} + d\mathbf{I})^{-1} \mathbf{g}$ where the positive denominator shift d is a complex function of t and is usually determined iteratively from the condition $|\Delta\mathbf{q}| = t$. The trust radius can be adjusted based on the accuracy of the energy difference predicted by the quadratic model, compared with the actual energy difference [2, 6]. One problem with the trust radius method, and other methods which limit $|\Delta\mathbf{q}|$, is that they do not scale properly with the system size: they are not ‘size consistent’. For instance, if n identical, non-interacting molecules are optimized simultaneously, the maximum displacement norm for each decreases like $n^{-1/2}$. For this reason, it is perhaps best to limit the maximum component, rather than the norm of the displacement [6].

An alternative, and closely related, approach is the augmented Hessian method [25]. The basic idea is to interpolate between the steepest descent method far from the minimum, and the Newton–Raphson method close to the minimum. This is done by adding to the Hessian a constant shift matrix which depends on the magnitude of the gradient. Far from the solution the gradient is large and, consequently, so is the shift d . One can, e.g., choose d to be proportional to the expected energy lowering, $d = -\alpha^2 \mathbf{g}^T \Delta\mathbf{q}$ (note that $\mathbf{g}^T \Delta\mathbf{q}$ is negative for minimization and thus d is positive), and solve the damped equation

$$(\mathbf{H} + d\mathbf{I})\Delta\mathbf{q} = -\mathbf{g}.$$

This is equivalent to finding the lowest eigenvalue λ (which is always negative and approaches zero at convergence) of the generalized eigenvalue equation

$$\begin{pmatrix} \mathbf{H} & \mathbf{g} \\ \mathbf{g}^T & 0 \end{pmatrix} \begin{pmatrix} \Delta\mathbf{q} \\ 1 \end{pmatrix} = \lambda \begin{pmatrix} \alpha^2 \mathbf{I} \\ \mathbf{0}^T \end{pmatrix} \begin{pmatrix} \mathbf{0} & \Delta\mathbf{q} \\ 1 & 1 \end{pmatrix}. \quad (\text{B3.5.5})$$

Equation B3.5.5 is, in turn, equivalent to the minimum condition on the rational function

$$E(\mathbf{q}) = E(\mathbf{q}^0) + (\mathbf{g}^T \Delta\mathbf{q} + \frac{1}{2} \Delta\mathbf{q}^T \mathbf{H} \Delta\mathbf{q}) / (1 + \alpha^2 \Delta\mathbf{q}^T \Delta\mathbf{q}).$$

For $\alpha = 0$, minimization of this expression yields the Newton–Raphson formula for $\Delta\mathbf{q}$. For large values of α , $\Delta\mathbf{q}$ becomes asymptotically $\alpha^{-1} \mathbf{g}/|\mathbf{g}|$, i.e., the steepest descent formula with a step length $1/\alpha$. The augmented Hessian method is closely related to eigenvector (mode) following, discussed in [section B3.5.5.2](#). The main difference between rational function and trust radius optimizations is that, in the latter, the level shift is applied only if the calculated step exceeds a threshold, while in the former it is imposed smoothly and is automatically reduced to zero as convergence is approached.

B 3.5.3 THE OPTIMIZATION OF WAVEFUNCTIONS

The basic self-consistent field (SCF) procedure, i.e., repeated diagonalization of the Fock matrix [26], can be viewed, if sufficiently converged, as local optimization with a fixed, approximate Hessian, i.e., as simple relaxation. To show this, let us consider the closed-shell case and restrict ourselves to real orbitals. The SCF orbital coefficients are not the

best set of coordinates to work with because, being constrained to be orthonormal, they are not independent. A better set of parameters can be chosen as the above-diagonal elements of an antisymmetric matrix \mathbf{K} , used to build an orthogonal matrix \mathbf{U} by $\mathbf{U} = \exp(\mathbf{K})$ [27] or, alternatively $\mathbf{U} = 2(\mathbf{I} - \frac{1}{2}\mathbf{K})^{-1} - \mathbf{I}$ (see, e.g., [23]). \mathbf{U} describes a generalized rotation between the orbitals. We start with an orthonormal set of orbitals \mathbf{C}_0 , defined as $\varphi^T = \chi^T \mathbf{C}_0$, where φ^T and χ^T are row vectors of molecular orbitals and atomic basis functions, respectively. All possible orthonormal orbital sets can be expressed as $\mathbf{C} = \mathbf{C}_0 \mathbf{U}$. Not all elements of \mathbf{K} are relevant. Rotations between virtual orbitals can obviously be omitted, and rotations between two occupied orbitals have no effect on the energy because the determinantal wavefunction is invariant against such rotations. Therefore, only rotations between occupied and virtual orbitals, i.e., the elements K_{ia} , $i \leq n$, $a > n$ are needed if, as usual, the n occupied orbitals φ_i precede the virtual ones φ_a .

The gradient and second derivative components of the SCF energy can be expressed for both kinds of parametrization (see [28]) as

$$\frac{1}{4} \partial E / \partial K_{ia} = F_{ia} \quad (\text{B3.5.6})$$

$$\frac{1}{4} \partial^2 E / \partial K_{ia} \partial K_{jb} = F_{ab} \delta_{ij} - F_{ij} \delta_{ab} + 4(ia|jb) - (ij|ab) - (ib|ja) \quad (\text{B3.5.7})$$

where, e.g., $(ij|ab)$ is a two-electron integral in the usual Mulliken notation. In a typical SCF iteration near convergence, the Fock matrix is nearly diagonal, and the orbital rotation parameter corresponding to a small occupied-virtual (Brillouin-violating) element F_{ia} is, from first-order perturbation theory, $K_{ia} = F_{ia} / (\epsilon_a - \epsilon_i)$. Comparing this with (B3.5.7) and noting that in a canonical orbital basis $F_{ii} = \epsilon_i$, $F_{aa} = \epsilon_a$ and the off-diagonal elements of \mathbf{F} are zero, it is clear that the ordinary SCF iteration is equivalent to neglecting the two-electron integrals in the electronic Hessian, equation (B3.5.7). This explains the observation that straight SCF iteration frequently slows down as the SCF procedure progresses, cf. [B3.5.2.3](#).

For ordinary SCF problems, interpolation methods are particularly suitable, as they require the storage of only a limited amount of information. The standard method for closed-shell or simple open-shell problems is DIIS (direct inversion in the iterative subspace) [22, 23]. Equation (B3.5.6) is not appropriate for the error vector because each error vector is expressed in a different basis. Transforming the occupied-virtual block of the orbital Fock matrix to a common basis, e.g., to the atomic orbital basis, yields the commutator $\mathbf{SDF} - \mathbf{FDS}$, arranged as a vector, for the gradient [22, 23]. DIIS usually converges well for closed shell systems from a reasonable starting wavefunction. Several modifications of this method have been proposed [29, 30]. For unrestricted (UHF) wavefunctions, DIIS is widely used but it is less appropriate. As it is a gradient norm minimization technique, it has a tendency to converge to the closest stationary point. For an even number of electrons, the closed-shell wavefunction is a formal solution of the UHF equations, and DIIS, unless started close to the expected minimum, may converge uphill to the closed-shell solution. The preconditioned conjugate gradient method (the preconditioner being the SCF approximation to the Hessian) is probably more appropriate in this case. Similarly, in density functional calculations with a plane wave basis set, the basis set is often huge, and the conjugate gradient method, with its limited storage requirement, is preferable [31, 32].

Due to the large number of variables in wavefunction optimization problems, it may appear that full second-order methods are impractical. For example, the storage of the Hessian for a modest closed-shell wavefunction with 500

basis functions and 200 electrons requires more than $(100 \times 400)^2/2 = 8 \times 10^8$ words. However, as shown by Bacskay [28] for closed- and open-shell SCF, and Lengsfeld and Liu for MC-SCF [33], using techniques analogous to direct configuration interaction [34], the solution of the linear system of equations (B3.5.3) can be accomplished iteratively, each micro-iteration (to be distinguished from the SCF iterations, which are called macro-iterations) taking about the same effort as an SCF cycle. For closed- and open-shell SCF, the resulting doubly iterative algorithm is comparable in efficiency with DIIS [35] but it is more complex, and thus less widely used.

The situation is different for the multi-configurational SCF (MC-SCF) case. Although DIIS has been used successfully for simpler cases [36], the strong coupling between orbital rotations and configuration interaction (CI) coefficients mandates the use of second-order or approximate second-order methods (see the reviews [3, 4] and references therein). As the signature of the Hessian is frequently incorrect, the augmented Hessian (rational function) method, which forces a step in the right direction, is generally employed. Perhaps the most efficient method is that proposed by Werner and Meyer [37] and further expanded by Werner and Knowles [38]. In this method, an approximate MC-SCF energy expression is defined, which is accurate to second order in terms of the orbital coefficients **C** and not the unitary parameters **K**. Through this change of variables, the effect of orthonormality and the periodicity of the orbitals as functions of the orbital rotations are taken into account correctly, resulting in a large increase of the radius of convergence of the Newton–Raphson method.

B 3.5.4 OPTIMIZATION OF MOLECULAR GEOMETRIES

This section discusses techniques specific to the optimization of molecular geometries.

There are four main factors that influence the rate of convergence of molecular structure optimizations: (1) the initial guess geometry; (2) the optimization algorithm; (3) the quality of the Hessian matrix and (4) the coordinate system. The first of these is obvious; the closer the starting geometry is to the final converged geometry, the fewer optimization cycles it should take to get there. Optimization algorithms will not be discussed here, as with a reasonable starting geometry and Hessian, most standard methods (see section B3.5.2) perform well. The choice of algorithm is, however, much more crucial for transition states, and one method, the Eigenvector Following algorithm, will be described in the section dealing with transition state optimization. The third point can also be dealt with briefly. Most current optimization algorithms use approximate second-derivative (Hessian) information with updating to help predict the next step. Assuming that the surface can be adequately modelled by a quadratic function, the more reliable the initial Hessian information and the updating is, the better will be the predicted step and the fewer cycles it should take to converge. Lower-level calculations, force fields and simple universal force constant [39, 40 and 41] formulae can be employed to generate the initial Hessian. The fourth factor, the coordinates used to carry out the optimization, is now recognized as being vitally important and it is the choice of coordinates that is largely responsible for the efficiency of modern geometry optimization algorithms.

B3.5.4.1 THE COORDINATE SYSTEM

As noted above, the coordinate system is now recognized as being of fundamental importance for efficient geometry optimization; indeed, most of the major advances in this area in the last ten years or so have been due to a better choice of coordinates. This topic is seldom discussed in the mathematical literature, as it is in general not possible to choose simple and efficient new coordinates for an abstract optimization problem. A nonlinear molecule with N atoms and no

symmetry needs $3N - 6$ internal coordinates to specify its geometry. Unless symmetry or other constraints fix the values of some coordinates, it is not possible to omit coordinates. However, it is possible, and sometimes

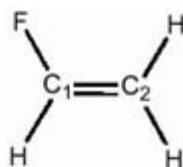
useful to use more. The coordinates are then not independent, a situation called redundancy.

The two key factors that make a good set of coordinates are to minimize the degree of coupling between them, and make the potential energy surface more quadratic. In general, the less coupling the better, as variation of one particular coordinate will then have minimal impact on the other coordinates. Coupling manifests itself primarily as relatively large mixed partial derivative terms between different coordinates. For example, a strong harmonic coupling between two different coordinates, i and j , results in a large off-diagonal element, H_{ij} , in the Hessian matrix. Cubic and higher-order couplings are even more deleterious for optimization, as they cannot be eliminated by a linear transformation.

Cartesian coordinates are an obvious choice as they can be defined for all systems, and gradients and second derivatives are calculated directly in Cartesians. Unfortunately, they normally make a poor coordinate set for optimization as they are fairly heavily coupled, their only advantage being their simplicity and completely general nature. If the quadratic model holds, i.e., for very small displacements, and if the gradient and Hessian are properly transformed, Cartesians are equivalent to any other coordinate set [42]. Of course, the source of good Hessian data is often a force field expressed in valence coordinates, so the latter are used implicitly. A further minor inconvenience of Cartesians is that the Hessian is singular, due to the presence of translational degrees of freedom (interestingly, rotations cause singularity of the force constant matrix—and zero vibrational frequencies—only at stationary points, a problem first discussed in [12] and rediscovered many times since). Cartesian optimization is used almost exclusively in molecular mechanics, despite its inefficiency, as it requires no transformation of the coordinates and derivatives. While the computational effort required by these transformations is negligible in *ab initio* work, it becomes significant in force field methods with energy and gradient computation being so rapid. Recent work promises to improve the efficiency of methods using valence internal coordinates [43, 44 and 45].

Z-matrix coordinates are widely used to define molecular geometries. A Z matrix specifies the molecular geometry in a treelike manner, by connecting each new atom in the system to those that have been defined previously. The first three atoms in the Z matrix are unique, with the first atom at the origin, the second lying on the Z axis (connected to the first by a single stretch) and the third lying in the XZ plane (connected to either the first or second atom via a stretch and defining a bend with the unconnected atom). Each new atom after the third is defined with respect to atoms previously defined in the Z matrix using, for example, one stretch, one bend and one torsion. An example of a typical Z matrix (for fluoroethylene) is shown in [figure B3.5.2](#).

-13-



C1						
C2	C1	L1				
F	C1	L2	C2	A1		
H	C1	L3	C2	A2	F	180.0
H	C2	L4	C1	A3	F	180.0
H	C2	L5	C1	A4	F	0.0
L1		1.3				
L2		1.365				
L3		1.08				
L4		1.08				
L5		1.08				
A1		122.5				
A2		118.0				
A3		120.0				
A4		120.0				

Figure B3.5.2. Example Z matrix for fluoroethylene. Notation: for example, line 4 of the Z matrix means that a H atom is bonded to carbon atom C1 with bond length L3 (ångströms), making an angle with carbon atom C2 of A3 (degrees) and a dihedral angle with the fluorine atom of 180.0° . All parameters given lettered variable names (L1, A1 etc) will be optimized; the dihedral angles are given explicitly as these are fixed by symmetry (the molecule is planar). Simple constraints can be imposed by removing parameters from the optimization list.

Initially, the Z matrix was utilized simply as a means of geometry input. It was subsequently found that optimization was generally more efficient in Z -matrix coordinates than in Cartesians, especially for acyclic systems. This is not always the case, and care must be taken in constructing a suitable Z matrix. A short discussion on good Z -matrix construction strategy is given by Schlegel [39].

The first *ab initio* gradient geometry optimizations were performed in what are now called natural internal coordinates [46], although they were formally defined only later [47]. These coordinates are derived from vibrational spectroscopy and are appropriate for covalent (mainly organic) molecules. They include all individual bond stretching coordinates, but only non-redundant linear combinations of bond angles and torsions as deformational coordinates. Suitable linear combinations of bends and torsions (the two are considered separately) are selected using group theoretical arguments based on approximate local symmetry. The major advantage of natural internal coordinates in geometry optimization is that they significantly reduce the coupling, both harmonic *and* anharmonic, between the various coordinates. Compared to natural internals, Z -matrix coordinates arbitrarily omit some angles and torsions—to prevent redundancy—and this can induce strong anharmonic coupling between the coordinates, especially with a poorly constructed Z matrix. Successful minimizations can be carried out in natural internals with only an approximate (e.g., diagonal) Hessian provided at the starting geometry but a good starting Hessian is still needed for a transition state search. Using a suitable set of internal coordinates can reduce the number of cycles required to converge compared to the corresponding Cartesian optimization by an order of magnitude or more, depending on the system.

-14-

Despite their clear advantages, natural internals have only become popular relatively recently, principally because in early programs they had to be user defined, a tedious and error-prone procedure for large molecules. This situation changed with the development of algorithms capable of generating natural internals automatically from input Cartesians [48, 49]. For minimization, natural internals and their successors have become the coordinates of choice [50, 51].

However, there are some disadvantages to natural internal coordinates. Their automatic construction proceeds by an exhaustive topological analysis involving thousands of lines of code and, for molecules with a complex structure, e.g., multiply fused rings and cages, the algorithm may be unable to generate a suitable non-redundant set of coordinates. Additionally, *more* coordinates than the $3N - 6$ (where N is the number of atoms) required may be generated. The redundancies can be removed by eliminating some coordinates, but this is arbitrary and may negatively influence convergence.

Various methods have been suggested for dealing with redundant coordinates. The normal coordinate optimization method [52] can use a force field defined in redundant coordinates, but is restricted to rectilinear coordinates. A general force field, expressed in redundant coordinates, can be transformed to a non-redundant set [13]. The current method of choice is to carry out the optimization directly in the redundant coordinate space [53] using the concept of a generalized inverse. If the total number of internal coordinates, including redundancies, is $n > 3N - 6$, then one constructs and diagonalizes the $n \times n$ matrix $\mathbf{G} = \mathbf{B}\mathbf{B}^T$ where \mathbf{B} is the first-order transformation matrix from Cartesians to internal coordinates, $\Delta\mathbf{q} = \mathbf{B}\Delta\mathbf{x}$. Diagonalization of \mathbf{G} results in two sets of eigenvectors; a set of $m = 3N - 6$ eigenvectors with eigenvalues $\lambda > 0$, and a set of $n - m$ eigenvectors with eigenvalues $\lambda = 0$ (to numerical precision). The eigenvalue equation for \mathbf{G} can be written

$$\mathbf{G}(\mathbf{UR}) = (\mathbf{UR}) \begin{pmatrix} \Lambda & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \quad (\text{B3.5.8})$$

and the generalized inverse of \mathbf{G} , \mathbf{G}^- , involves inverting the non-zero eigenvalues only and back-transforming

$$\mathbf{G}^- = \mathbf{U}\Lambda^{-1}\mathbf{U}^T.$$

In this way the optimization can be cast in terms of the original coordinate set, including the redundancies. Exactly the same transformations between Cartesian and internal coordinate quantities hold as for the non-redundant case (see the next section), but with the generalized inverse replacing the regular inverse.

The redundant optimization scheme [53] can be applied to natural internal coordinates, which are sometimes redundant for polycyclic and cage compounds. It can also be applied directly to the underlying primitives. This has the disadvantage that the coordinate space is larger, and contains many redundancies, but it is simpler to implement than a full natural internal coordinate scheme and can handle essentially any molecule, regardless of the topology, thus avoiding any failure in the generating algorithm. The well known Gaussian *ab initio* program [19] now uses, as a default, this type of algorithm [54].

-15-

As originally implemented, the redundant optimization scheme involved solution of equation (B3.5.8) at the beginning of *every* optimization cycle, which may be expensive in semiempirical or force field methods. A scheme which involves a *single* diagonalization was introduced by Baker *et al* [55]. Diagonalization of the \mathbf{G} matrix partitions the original coordinate space into two subspaces, a redundant subspace spanned by the set of vectors in \mathbf{R} and a non-redundant subspace spanned by \mathbf{U} . Since \mathbf{R} is redundant, it can be discarded and the set of vectors in \mathbf{U} defines a complete, non-redundant coordinate set which can be retained throughout the entire optimization. Unlike natural internals, which are linear combinations of just a few of the primitives localized in small regions of the molecule, each vector in \mathbf{U} is potentially a linear combination of *all* of the primitives and is delocalized over the entire molecule; they are known as *delocalized internal coordinates* [55]. Despite their apparent complexity, delocalized internals perform in practice as well as natural internals.

We present in [table B3.5.1](#) a comparison of Cartesian, *Z*-matrix and delocalized internal coordinate optimizations, using the semiempirical PM3 method [56] on ten typical medium-sized organic molecules. All optimizations were started with a unit Hessian and used the EF algorithm (see later) [57] with a BFGS Hessian update [6] to compute the optimization step; the only difference is in the coordinate system. The final column shows our best results using an initial non-unit Hessian matrix, diagonal in the space of *primitive* internals, with diagonal force constants estimated using the recipe of Schlegel [39]. The results, in terms of the number of cycles required for convergence, clearly show the advantages of using a good set of coordinates combined with a reliable estimate of the corresponding force constants.

-16-

Table B3.5.1 Number of cycles to converge for geometry optimizations of some typical organic molecules using Cartesian, *Z*-matrix and delocalized internal coordinates^a.

Molecule	Formula	Symmetry	Cycles to converge			
			CART ^b	ZMAT ^b	INT ^b	INT ^c

azulene	C ₁₀ H ₈	C _{2v}	24	13	16	10
lumazine	C ₆ H ₄ N ₄ O ₂	C _s	26	18	14	8
dichloropropane (<i>gauche</i>)	C ₃ H ₆ Cl ₂	C ₁	48	24	33	7
3,3,3-trifluoro-2-methyl propene	C ₄ H ₅ F ₃	C _s	24	10	13	7
cyanomethyl methyl ether	C ₃ H ₅ NO	C ₁	45	25	33	9
salicylic acid	C ₇ H ₆ O ₃	C _s	32	45 ^d	13	10
isoxanthopterin	C ₆ H ₅ N ₅ O ₂	C _s	36	18	12	9
pyrroloquinoline quinone anion (3-)	C ₁₄ H ₃ N ₂ O ₈	C ₁	167 ^e	F ^f	109	36
2,5-bis-(4-aminophenyl)-1,3,4-oxadiazol	C ₁₄ H ₁₂ N ₄ O	C _{2v}	27	F ^f	14	7
permethyl-nonasilane (<i>gauche</i> conformer)	Si ₉ (CH ₃) ₂₀	C ₁	355 ^e	113 ^g	213	47

^a Calculations using the semiempirical PM3 method with standard convergence criteria of 0.0003 au on the maximum component of the gradient vector and *either* an energy change from the previous cycle of $< 10^{-6}$ hartree *or* a maximum predicted displacement for the next step of < 0.0003 au.

^b Started with a unit Hessian matrix.

^c Started with a Hessian diagonal in the space of *primitive* internals using the recipe of Schlegel [39].

^d Poor *Z* matrix.

^e Converged prematurely with too high energy due to small energy changes between steps.

^f *Z* matrix generated using Cartesian \rightarrow *Z*-matrix conversion program. Severe converge problems with energy oscillation; halted after 90 cycles with energy higher than at starting geometry.

^g Acyclic system; good *Z* matrix.

B3.5.4.2 TRANSFORMATION BETWEEN COORDINATE SYSTEMS

This section deals with the transformation of coordinates and forces [11, 47] between different coordinate systems. In particular, we will consider the transformation between Cartesian coordinates, in which the geometry is ultimately specified and the forces are calculated, and internal coordinates which allow efficient optimization.

(A) TRANSFORMATION OF FIRST AND SECOND DERIVATIVES

Let us consider the energy expanded through second order in two sets of displacement coordinates $\Delta\mathbf{x}$ and $\Delta\mathbf{q}$. The two coordinate systems are related by

-17-

$$\Delta q_i = \sum_a B_{ia} \Delta x_a + \frac{1}{2} \sum_{a,b} C_{ab}^i \Delta x_a \Delta x_b + \dots \quad (\text{B3.5.9})$$

The potential energy is given as

$$\begin{aligned}
E &= E_0 + \sum_i \Delta q_i g_i + \frac{1}{2} \sum_{i,j} H_{ij} \Delta q_i \Delta q_j + \dots \\
&= E_0 + \sum_a \Delta x_a f_a + \frac{1}{2} \sum_{a,b} K_{ab} \Delta x_a \Delta x_b + \dots
\end{aligned}
\tag{B3.5.10}$$

or, in matrix notation,

$$E = E_0 + \Delta \mathbf{q}^T \mathbf{g} + \frac{1}{2} \Delta \mathbf{q}^T \mathbf{H} \Delta \mathbf{q} + \dots = E_0 + \Delta \mathbf{x}^T \mathbf{f} + \frac{1}{2} \Delta \mathbf{x}^T \mathbf{K} \Delta \mathbf{x} + \dots$$

where \mathbf{g} and \mathbf{H} are the gradient and the force constant matrix, respectively, in internal coordinates, and \mathbf{f} and \mathbf{K} are the same in Cartesians. Substituting (B3.5.9) into equation (B3.5.10) and equating equal powers one obtains

$$\begin{aligned}
\mathbf{f} &= \mathbf{B}^T \mathbf{g} \\
\mathbf{K} &= \mathbf{B}^T \mathbf{H} \mathbf{B} + \sum_i \mathbf{C}^i g_i.
\end{aligned}$$

If the two coordinate systems are connected by a non-singular transformation then, defining $\mathbf{A} = (\mathbf{B}^T)^{-1}$, the more important inverse transformations are given by

$$\begin{aligned}
\mathbf{g} &= \mathbf{A} \mathbf{f} \\
\mathbf{H} &= \mathbf{A} \mathbf{K} \mathbf{A}^T - \sum_i g_i \mathbf{A} \mathbf{C}^i \mathbf{A}^T.
\end{aligned}
\tag{B3.5.11}$$

Thus the transformation matrix for the gradient is the inverse transpose of that for the coordinates. In the case of transformation from Cartesian displacement coordinates ($\Delta \mathbf{x}$) to internal coordinates ($\Delta \mathbf{q}$), the transformation is singular because the internal coordinates do not specify the six translational and rotational degrees of freedom. One could augment the internal coordinate set by the latter but a simpler approach is to use the generalized inverse [58]

$$\mathbf{A} = (\mathbf{B} \mathbf{M} \mathbf{B}^T)^{-1} \mathbf{B} \mathbf{M}$$

where \mathbf{M} is any non-singular $3N \times 3N$ matrix, in the simplest case the $3N$ -dimensional unit matrix.

The second term in equation B3.5.11 deserves comment. This term shows that Hessian (second-derivative) matrices

in different coordinate systems are not related simply by a similarity transformation, except at stationary points. In particular, its signature (number of negative eigenvalues) does not have to be the same in different coordinate systems. Properly chosen internal coordinates tend to make the Hessian positive definite, and are probably one of the reasons why internal coordinates are preferable for molecular geometry optimization.

When working with any coordinate system other than Cartesians, it is necessary to transform finite displacements between Cartesian and internal coordinates. Transformation from Cartesians to internals is seldom a problem as the latter are usually geometrically defined. However, to transform a geometry displacement from internal coordinates to Cartesians usually requires the solution of a system of coupled nonlinear equations. These can be solved by iterating the first-order step [47]

$$\Delta \mathbf{x} = \mathbf{A}^T \Delta \mathbf{q}$$

where $\Delta \mathbf{q}$ is the difference between the current internal coordinates and their desired values, calculated to full accuracy from the Cartesians. If $\Delta \mathbf{q}$ is large, it may be better to proceed in stages, converging $\Delta \mathbf{x}$ roughly between each stage.

B3.5.4.3 CONSTRAINED OPTIMIZATION

Constrained optimization refers to optimizations in which one or more variables (usually some internal parameter such as a bond distance or angle) are kept fixed. The best way to deal with constraints is by elimination, i.e., simply remove the constrained variable from the optimization space. Internal constraints have typically been handled in quantum chemistry by using Z matrices; if a Z matrix can be constructed which contains all the desired constraints as individual Z -matrix variables, then it is straightforward to carry out a constrained optimization by elimination.

The situation is more complicated in molecular mechanics optimizations, which use Cartesian coordinates. Internal constraints are now relatively complicated, nonlinear functions of the coordinates, e.g., a distance constraint between atoms i and j in the system is $R_{ij} = \sqrt{\{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2\}} = R_0$, and this cannot be handled by simple elimination. There are two main approaches if elimination is not possible, penalty functions and Lagrange multipliers.

The general constrained optimization problem can be considered as minimizing a function of n variables $F(\mathbf{x})$, subject to a series of m constraints of the form $C_i(\mathbf{x}) = 0$. In the penalty function method, additional terms of the form $\frac{1}{2}\sigma_i C_i(\mathbf{x})^2$, $\sigma_i > 0$, are formally added to the original function, thus

$$F(\mathbf{x})^{\text{pen}} = F(\mathbf{x}) + \sum \frac{1}{2}\sigma_i C_i(\mathbf{x})^2$$

with the summation over all m constraints. If the constraint is satisfied, the additional term is zero, but if not then the value of the function increases in proportion to the square of the deviation, i.e., the additional term penalizes any geometries that do not satisfy the constraint. In practice, the value of the function is left unaltered and what is done is to modify the gradient according to

-19-

$$\partial F(\mathbf{x})^{\text{pen}} / \partial x_j = \partial F(\mathbf{x}) / \partial x_j + \sum \sigma_i \partial C_i(\mathbf{x}) / \partial x_j.$$

Exactly the same types of step as for an unconstrained optimization can then be taken, using the modified as opposed to the regular gradient.

The performance of the penalty function algorithm is heavily influenced by the value chosen for σ_i . The larger the value of σ_i , the better the constraints are satisfied but the slower the rate of convergence. Optimizations with very high values of σ_i encounter severe convergence problems. However, the method is very general and

easy to apply.

A better approach is the method of Lagrange multipliers. This introduces the Lagrangian function [59]

$$L(\mathbf{x}, \lambda) = F(\mathbf{x}) - \sum \lambda_i C_i(\mathbf{x})$$

which replaces the function $F(\mathbf{x})$ in the unconstrained case. Here the λ_i are the so-called Lagrange (or unknown) multipliers, one for each constraint. Differentiating with respect to \mathbf{x} and λ gives

$$\partial F(\mathbf{x})^{\text{pen}} / \partial x_j = \partial F(\mathbf{x}) / \partial x_j + \sum \sigma_i \partial C_i(\mathbf{x}) / \partial x_j.$$

and

$$\partial L(\mathbf{x}, \lambda) / \partial \lambda_i = -C_i(\mathbf{x}).$$

At a stationary point of the Lagrangian function, we have $\nabla L = \mathbf{0}$, i.e., all $\partial L / \partial x_j = 0$ and all $\partial L / \partial \lambda_i = 0$. This latter condition means that all $C_i(\mathbf{x}) = 0$ and so all constraints are satisfied. Hence finding a set of values (\mathbf{x}, λ) for which $\nabla L = \mathbf{0}$ gives a solution to the constrained optimization problem in exactly the same way as finding an \mathbf{x} for which $\nabla F = \mathbf{0}$ gives a solution to the corresponding unconstrained problem.

A major difference between the penalty function and Lagrange multiplier methods is that in the latter the unknown multipliers are part of the optimization space and are treated essentially as additional variables. The Lagrange multiplier method usually converges significantly faster than the penalty function method and has the further advantage that constraints are satisfied essentially exactly. Note that, in both methods, constraints do not need to be satisfied in the starting geometry, but are instead satisfied at convergence. An efficient algorithm (within the context of Cartesian optimization) for imposing constraints in Cartesian coordinates, which incorporates both penalty functions and Lagrange multipliers, was presented by Baker in 1992 [60], with further improvements in the following year [61].

By combining the Lagrange multiplier method with the highly efficient delocalized internal coordinates, a very powerful algorithm for constrained optimization has been developed [62]. Given that delocalized internal coordinates are potentially linear combinations of *all* possible primitive stretches, bends and torsions in the system, cf. *Z*-matrix coordinates which are *individual* primitives, it would seem very difficult to impose any constraints at all; however, as

shown in the original reference [55], any desired internal constraint can be imposed using a relatively simple Schmidt orthogonalization procedure. By projecting unit vectors (with unit components corresponding to particular primitives) onto the non-redundant subspace \mathbf{U} (see equation (B3.5.8)) and Schmidt orthogonalizing the resultant vectors against all other vectors in \mathbf{U} , it is possible to isolate individual primitives in a consistent manner into *single* vectors. By removing these vectors from the optimization space, optimizations can be carried out in which the primitives involved retain their initial values throughout the optimization. The resulting algorithm has all the advantages of redundant internal coordinate optimizations in terms of efficiency, combined with the advantage of the Lagrange multiplier method that desired constraints do not have to be satisfied in the starting geometry. As constraints do become satisfied, they can simply be eliminated from the optimization space (this cannot be done with Cartesian optimizations due to the form of the constraints). By starting with an appropriate constraint vector, it is possible to impose constraints on linear

combinations of variables rather than on individual primitives. It is also possible to perform constrained transition state searches. For more details see [55] and [62].

There are alternative methods for imposing constraints, but they are less satisfactory than those discussed above. One commonly used alternative is to use projection techniques [12, 63]. In this approach, components in directions that would result in motion that would violate the constraints are projected out of the gradient vector and Hessian matrix before calculating the next step. Unfortunately, while projection works fine for constraints that are linear in the coordinates (the standard method for imposing the Eckart conditions is by projection), nonlinear constraints have to be linearized, and consequently ‘feasibility corrections’ [63] must be applied to prevent deviations from the desired constraints increasing as the optimization progresses. The interesting method of Taylor and Simons [64] for combining linearized geometrical constraints with mode following also suffers from this drawback.

Another way of attempting to constrain variables without eliminating them from the optimization space is to set the appropriate force constants to very large values. This is what is currently done in Schlegel’s redundant internal coordinate algorithm to prevent motion in the redundant subspace [54]. Perhaps a better method is to set the corresponding rows and columns in the inverse Hessian to zero. This method must begin with a geometry that satisfies the constraints.

B 3.5.5 OPTIMIZATION OF TRANSITION STATES

Searching for transition states poses additional difficulties compared to minimization. The first problem is the starting geometry. There is a host of structural information that can be called upon to provide a good estimate of the likely geometry of a local minimum. Far less knowledge is available about transition state geometries. Second is the structure of the Hessian. In the region of a transition structure, the Hessian must have one, and only one, negative eigenvalue. Unlike the situation for a minimum search, there is no simple and cheap method for guessing a reasonable starting Hessian with the appropriate eigenvalue structure. Even if you calculate an exact initial Hessian, if your starting geometry is poor, its eigenstructure will probably be inappropriate. One thing that is often done for transition states is to calculate a few rows and columns of the Hessian—those corresponding to variables in the ‘active site’ where most of the geometrical changes are expected—by finite difference on the gradient, and guess diagonal Hessian matrix elements for the rest of the system in the same way one would do for a minimum search. The starting Hessian will then hopefully have an appropriate negative eigenvalue.

-21-

In addition to the problem of generating a starting Hessian with the correct signature, there are problems in retaining it. The Hessian updates commonly used for minimization generally retain positive definiteness (see section B3.5.2.3); there are no such guarantees for retaining a negative eigenvalue during a transition state search. Once the desired region of the potential energy surface (PES) has been reached, quasi-Newton techniques can be used to refine the geometry; however they must be able to correct for the occasional bad update which may destroy the Hessian eigenstructure. Transition state searches can thus be separated into two parts; first find the correct region on the PES and then home in onto the transition state. Many of the methods described below for locating approximate transition states have the advantage that they require no second-derivative information.

B3.5.5.1 LOCATING THE CORRECT NEIGHBOURHOOD

(A) COORDINATE DRIVING

A commonly used approach is coordinate driving. Here an appropriate internal coordinate, or a linear combination of coordinates, is chosen as a reaction coordinate. At various intervals along this coordinate, between its value in the reactants and in the products, all the other variables are optimized. This then defines a minimum energy path. The energy maximum on this path can be shown to be the transition state geometry. Usually, however, the maximum on the path is located only approximately. Coordinate driving involves several minimizations in $(n - 1)$ variables; consequently it is quite expensive. Moreover, its success depends on a good definition of the reaction coordinate; it should be roughly parallel with the true reaction path. If, at any point along the path, the reaction coordinate becomes nearly perpendicular to the reaction path, the latter may become discontinuous. The minimum energy path defined in this way has little physical significance, as different choices of reaction coordinate can produce different pathways.

(B) SYNCHRONOUS TRANSIT

Another approach requiring less intuition is the synchronous transit method [65]. Here the path between reactants and products is interpolated linearly between the reactant and the product. The interpolation can be carried out in Cartesians, internal coordinates or, perhaps best, in terms of distance coordinates [66]; the results depend somewhat on the interpolation method. A maximum is first found along this linear synchronous transit path. This is followed by alternate minimization along directions *orthogonal* to the original direction, combined with maximum searches along a parabolic path (the quadratic synchronous transit) joining the reactant, the product and the current estimate of the transition state. For very curved reaction paths the quadratic synchronous transit path may be a poor approximation, and the reaction path may have to be approximated piecewise. A similar algorithm, which involves minimization in a space *conjugate* to the maximum search direction, was developed by Bell and Crighton [67]. This last reference also contains a good discussion of various transition-state search strategies.

Both of the above methods can be considered as approximations to the Fukui reaction path, discussed later in this article. The maximum of the Fukui reaction path also yields the transition state, although usually at significantly more expense than coordinate driving or the synchronous transit method. A more modern development of these ideas has been given by Ionova and Carter [68].

(C) WALKING UPHILL

These algorithms try to walk up to transition states from minima, usually along the shallowest path, i.e., along the

-22-

eigenvector of the Hessian which has the lowest eigenvalue [69, 70 and 71]. They are more important when the transition state has already been located approximately and will be discussed in the next section.

(D) BRACKETING THE TRANSITION STATE

These methods try to bracket the transition state from both the reactant and the product side [72, 73]. For example, in the method of Dewar *et al* [73], two structures, one in the reactant valley and one in the product valley, are optimized simultaneously. The lower-energy structure is moved to reduce the distance separating the two structures by a small amount, e.g. by 10%, and its structure is reoptimized under the constraint that the distance is fixed. This process is repeated until the distance between the two structures is sufficiently small.

(E) SEAM CROSSING

Here the transition state is approximated by the lowest crossing point on the seam intersecting the *diabatic* (non-interacting) potential energy surfaces of the reactant and product. The method was originally developed

for excited state surfaces [74], and has subsequently been used to locate approximate transition states [75, 76].

B3.5.5.2 REFINING THE TRANSITION STATE

It is usually not efficient to use the methods described above to refine the transition state to full accuracy. Starting from a qualitatively correct region on the potential surface, in particular one where the Hessian has the right signature, efficient gradient optimization techniques, with minor modifications, are usually able to zero in on the transition state quickly.

(A) DIRECT INVERSION IN THE ITERATIVE SUBSPACE

One of the methods which is appropriate is DIIS applied to geometry optimization [24]; being a gradient norm minimization method, it will converge to any stationary point. This is, however, one of its problems as it may converge to the wrong point. The convergence radius of augmented Hessian type methods is larger. One of these methods, the eigenvector following (EF) method [57], is generally useful for both transition states and minima and will be described here as an example of a modern optimization algorithm.

(B) EIGENVECTOR FOLLOWING

The EF algorithm [57] is based on the work of Cerjan and Miller [69] and, in particular, Simons and coworkers [70, 71]. It is closely related to the augmented Hessian (rational function) approach [25]. We have seen in [section B3.5.2.5](#) that this is equivalent to adding a constant level shift (damping factor) to the diagonal elements of the approximate Hessian \mathbf{H} . An appropriate level shift effectively makes the Hessian positive definite, suitable for minimization.

Although a single shift parameter can also be used to find a transition state, the eigenvector following algorithm utilizes *two* level shifts: one for the Hessian (transition state) mode along which the energy is to be maximized and the other for modes for which it is minimized. In terms of a diagonal Hessian representation, transforming the gradient appropriately, we have the two eigenvalue equations

-23-

$$\begin{pmatrix} b_2 & \dots & 0 & g'_2 \\ & \ddots & & \\ 0 & & b_n & g'_n \\ g'_2 & & g'_n & 0 \end{pmatrix} \begin{pmatrix} \Delta q_2 \\ \vdots \\ \Delta q_n \\ 1 \end{pmatrix} = \lambda_n \begin{pmatrix} \Delta q_2 \\ \vdots \\ \Delta q_n \\ 1 \end{pmatrix} \quad (\text{B3.5.12})$$

and

$$\begin{pmatrix} b_1 & g'_1 \\ g'_1 & 0 \end{pmatrix} \begin{pmatrix} \Delta q_1 \\ 1 \end{pmatrix} = \lambda_p \begin{pmatrix} \Delta q_1 \\ 1 \end{pmatrix}. \quad (\text{B3.5.13})$$

In these two equations, the b_i are the eigenvalues of \mathbf{H} ($b_1 < b_2 < \dots < b_n$), \mathbf{g}' is the gradient vector transformed to the basis of the eigenvectors \mathbf{U} of \mathbf{H} : $\mathbf{g}' = \mathbf{U}^T \mathbf{g}$, and we have (arbitrarily) set the factor α in [equation \(B3.5.5\)](#) to unity. Note that λ_n is the lowest eigenvalue of equation (B3.5.12) (it is always negative and approaches zero at convergence), while λ_p is the highest eigenvalue of equation (B3.5.13) (it is always positive and again approaches zero at convergence). Once suitable values of λ_p and λ_n have been determined, the final step is given by

$$\Delta \mathbf{q} = -\mathbf{g}'_1 \mathbf{u}_1 / (b_1 - \lambda_p) - \sum \mathbf{g}'_i \mathbf{u}_i / (b_i - \lambda_n) \quad (i = 2, \dots, n)$$

where it is assumed that we are maximizing along the lowest Hessian mode \mathbf{u}_1 , and minimizing along all the others. This holds *regardless* of the Hessian eigenvalue structure (unlike the Newton–Raphson step), and so the algorithm can handle Hessian matrices with the wrong signature.

It is also possible to maximize along modes other than the lowest and, in this way perhaps, locate transition states for alternative rearrangements/dissociations from the same initial starting point. For maximization along the k th mode (instead of the lowest), b_1 would be replaced by b_k , and the summation would now exclude the k th mode but include the lowest. Since what was originally the k th mode is the mode along which the negative eigenvalue is required, then this mode will eventually become the lowest mode at some stage of the optimization. To ensure that the original mode is being followed smoothly from one cycle to the next, the mode that is actually followed is the one with the greatest overlap with the mode followed on the previous cycle. This procedure is known as mode following. For more details and some examples, see [57]. Mode following can work well for small systems, but for larger, flexible molecules there are usually a number of soft modes which lead to transition states for conformational rearrangements and not to the more interesting reaction saddle points. Moreover, each eigenvector can be followed in two opposite directions and frequently only one leads to a reaction.

Although it was originally developed for locating transition states, the EF algorithm is also efficient for minimization and usually performs as well as or better than the standard quasi-Newton algorithm. In this case, a single shift parameter is used, and the method is essentially identical to the augmented Hessian method.

B 3.5.6 SIMULTANEOUS OPTIMIZATION OF GEOMETRIES AND WAVEFUNCTIONS

So far, we have considered the optimization of wavefunction and geometry parameters separately. In view of the much shorter timescale and higher energy associated with the former, this is reasonable. However, additional savings can be potentially obtained by optimizing the wavefunction and the geometry simultaneously. This was first proposed for density functional methods [77] and later for traditional quantum chemistry techniques [78]. With the large increase of computing speed compared to disk input/output speed, direct techniques [79] were generally adopted. In direct methods, the large disparity between calculating the gradients of the molecular energy with respect to electronic parameters (the Fock matrix in SCF theory) and nuclear coordinates disappeared; gradients are now only a few times more expensive than a Fock matrix evaluation, making simultaneous wavefunction-geometry optimization much more attractive. In spite of this, such methods are not yet widely used, except in the crude form of relaxing the SCF convergence criteria if the geometry parameters are far from convergence.

The molecular dynamics method introduced by Car and Parrinello [80], though not strictly an optimization method, has many features in common with simultaneous optimization of the wavefunction and geometry. In this method, the electronic wavefunction and energy are close to, but not identical with, the Born–Oppenheimer energy. The basic idea is to consider the electronic degrees of freedom as dynamical variables, along with the nuclear coordinates. The Lagrangian contains the kinetic energy of the nuclei, the potential energy as a function of both the nuclear and electronic degrees of freedom and a fictitious kinetic energy term which is the square of the time derivative of the electronic wavefunction multiplied by a small mass. The inertia of this fictitious electronic mass causes the wavefunction to deviate slightly from the Born–Oppenheimer surface. The Car–Parrinello method is most efficient for plane wave basis sets, as the

calculation of the nuclear gradient is very inexpensive in this method, but it has also been introduced into SCF theory [81, 82].

B 3.5.7 REACTION PATH ALGORITHMS

The reaction path is defined by Fukui [83] as the line $\mathbf{q}(s)$ leading down from a transition state along the steepest descent direction

$$\partial \mathbf{q}(s) / \partial s = -\mathbf{g}(s) / |\mathbf{g}(s)|. \quad (\text{B3.5.14})$$

Here s is the path length, $ds = (dq_1^2 + \dots + dq_n^2)^{1/2}$. The reaction path is, unfortunately, dependent on the coordinate system. This should perhaps be emphasized more than is generally the case. Scaling one coordinate by a factor $\alpha > 1$ increases the coordinate value but decreases the corresponding gradient component, so that if the reaction path was antiparallel to the gradient before the scaling it will not be so after scaling. For qualitative studies of chemical reactions, there is little to recommend one particular reaction path over another. However, for dynamical studies, the intrinsic reaction coordinate (IRC) [83], defined as the path length along the reaction path in mass-weighted Cartesian coordinates, $\xi_i = m_i^{1/2} x_i$, has advantages over other definitions (for example the kinetic energy matrix is the unit

-25-

matrix). Here m_i is the atomic mass corresponding to the Cartesian coordinate x_i , making the reaction path isotope dependent. A major difficulty with reaction paths is that to decide whether a given point is on the path the whole path must be constructed; local information (energy, gradient, force constants etc.) is insufficient.

B3.5.7.1 FOLLOWING THE REACTION PATH DOWNHILL

The most widely used methods try to follow the gradient downhill starting from a transition state. At the transition state itself, the gradient vanishes and the first step must be made along the imaginary eigenvector of the Hessian in the proper coordinates, i.e., mass-weighted Cartesians for the IRC path. As pointed out by Schlegel [1, 2], (B3.5.14) is a stiff differential equation and its integration by simply making small downhill steps along the gradient, a method equivalent to Euler's method, requires very small steps and consequently much effort. Otherwise, the calculated reaction path diverges from the true one, at first slowly and then more rapidly. To deal with this problem requires either constrained minimization steps at each point on the path, or alternatively second-order (both gradient and Hessian) information. This increases the cost of the individual steps but allows much larger steps to be taken.

The method of Ishida *et al* [84] includes a minimization in the direction in which the path curves, i.e. along $(\mathbf{g}/|\mathbf{g}| - \mathbf{g}'/|\mathbf{g}'|)$, where \mathbf{g} and \mathbf{g}' are the gradient at the beginning and the end of an Euler step. This technique, called the stabilized Euler method, performs much better than the simple Euler method but may become numerically unstable for very small steps. Several other methods, based on higher-order integrators for differential equations, have been proposed [85, 86].

Page *et al* [87] use a local quadratic model for the surface. This requires the Hessian, but once it is available, the reaction path can be inexpensively determined for a quadratic (or even higher-order) analytical surface (see also [88]). Gonzales and Schlegel [89, 90] approximate the reaction path by an arc of a circle. They first make a step along the gradient of length half the current stepsize to an intermediate point. From this, they make another half step so that the energy is minimized, subject to the stepsize constraint. The wavefunction

and the gradient need not be evaluated at the intermediate point. This method is implemented in the Gaussian series of programs [19] and is widely used. It does not need the exact Hessian, but a good estimate should be available so that the many local optimizations converge rapidly. An advantage of this method is that it yields the curvature of the reaction path at the transition state correctly.

B3.5.7.2 APPROACHING THE REACTION PATH FROM THE SIDE

These methods, which probably deserve more attention than they have received to date, simultaneously optimize the positions of a number of points along the reaction path. The method of Elber and Karplus [91] was developed to find transition states. It furnishes, however, an approximation to the reaction path. In this method, a number (typically 10–20) equidistant points are chosen along an approximate reaction path connecting two stationary points **a** and **b**, and the average of their energies is minimized under the constraint that their spacing remains equal. This is obviously a numerical quadrature of the integral $S^{-1} \int_a^b E(q(s))$ where S is the path length between the points **a** and **b**. The Euler equation to this variation problem yields the condition for the reaction path, [equation \(B3.5.14\)](#). A similar method has been proposed by Stachó and Bán [92].

B3.5.7.3 BIFURCATION OF THE REACTION PATH AND VALLEY–RIDGE INFLECTION POINTS

As shown by Valtazanous and Ruedenberg [93], steepest descent paths (e.g., the Fukui intrinsic reaction coordinate)

-26-

can bifurcate, i.e., split in two, only at stationary points. Thus, the intuitive notion of a reaction path forking, e.g., upon ascent in a valley to two different transition states, or, starting down from a transition state to two different minima, is impossible. This should be regarded as an inherent limitation of the standard definition of a reaction path, not as a physical impossibility. Such cases, in which a valley floor is gradually transformed into a ridge are, in fact, quite common. Mathematically, they are characterized by fact that one of the eigenvalues of the Hessian in the subspace perpendicular to the path changes from positive to negative. The point at which the eigenvalue is zero is called the valley–ridge inflection point. The reaction path, started at a stationary point, will run directly along the ridge and thus becomes non-physical past a valley–ridge inflection point. The actual reaction, of course, will not follow the reaction path in this case, not even qualitatively. Steepest descent paths started a little away from the reaction path will veer away from the latter after passing the valley–ridge inflection point. Baker and Gill [94] have devised a method for locating valley–ridge inflection points (which they call branching points). The reader is reminded, however, that the signature of the Hessian at non-stationary points depends strongly on the coordinate system. Thus, the location of a valley–ridge inflection point may be quite different in Cartesians or mass-weighted Cartesians than in internal coordinates. In particular, the Hessian in Cartesian coordinates may have spurious negative eigenvalues corresponding to rotational coordinates.

B 3.5.8 GLOBAL OPTIMIZATION

For our purposes, global optimization refers to the location of the *lowest* minimum on a given potential energy surface. As mentioned in the introduction, this is currently only a partially solved problem. The number of conformational minima, e.g., for a large protein, increases enormously with the size of the system, and the only way to be *absolutely* sure that the lowest-energy structure has been found is to do an exhaustive search of the entire energy surface; for large molecules this is essentially impossible. Even if the lowest-energy structure were successfully located, this would likely have only limited chemical significance, as there would be many structures energetically close to the global minimum (within a kcal or so) which would need to be

considered for an accurate treatment of the thermodynamics. It is almost a certainty (though the authors are unaware of a formal proof) that finding the global minimum on molecular potential energy surfaces is computationally NP complete, and thus scales factorially with the size of the problem. Such problems are generally regarded as insoluble (however, this does not exclude their solution in a given case).

With systematic PES searches excluded, random (stochastic) methods have become the most common techniques for global minimization. The two most popular methods are simulated annealing [95] and genetic algorithms [96]. The former method derives its name from the annealing process in condensed matter physics in which a solid is melted in a bath and the temperature is then slowly decreased; the particles are expected to settle into their lowest-energy states, provided the initial temperature is sufficiently high and the cooling rate is sufficiently low. In practical optimizations, cooling is represented by local minimizations and heating by random jumps, i.e., random displacements of some or all of the atoms. After a ‘sufficient number’ of local minimization/random jump cycles, the procedure is terminated with the lowest-energy structure found so far taken as the global minimum.

The genetic algorithm method takes its name from the trading of genetic information in chromosomes between parents to produce an offspring. A random population of individuals (geometrical structures for the system in question) is created, and local minimizations are performed on each individual. Selected structural components (genes) from mostly the lowest-energy individuals are allowed to exchange, producing a new set of individuals for the next round of local

-27-

minimizations. After a sufficiently large number of rounds, the global minimum should be located.

Both of these global optimization methods require a very large number of essentially full local optimizations and, consequently, are normally restricted to moderate-sized systems described using mechanics force fields. A somewhat different approach has been developed by Piela and coworkers [97], utilizing the diffusion or heat conduction equation. In this method, a surface containing multiple local minima is smoothly deformed in such a way that wells on the surface gradually disappear, with shallower wells vanishing faster than deeper, lower-energy wells. Eventually a surface will be derived which has just one minimum, related to the lowest-energy, global minimum on the original surface. By carefully reversing the procedure, keeping track of the minimum as it evolves, one is (hopefully) led back to the global minimum as the original surface is reformed.

Other deterministic methods for global optimization have also been developed (see, e.g., [98]).

REFERENCES

- [1] Schlegel H B 1987 Optimization of equilibrium geometries and transition structures *Adv. Chem. Phys.* **67** 249
- [2] Schlegel H B 1995 Geometry optimization on potential energy surfaces *Modern Electronic Structure Theory* ed D Yarkony (Singapore: World Scientific) pp 459–500
- [3] Werner H-J 1987 Matrix-formulated direct multiconfigurational self-consistent field and multireference configuration interaction methods *Adv. Chem. Phys.* **69** 1
- [4] Shepard R 1987 The multiconfiguration self-consistent field method *Adv. Chem. Phys.* **69** 63
- [5] Pulay P 1987 Analytical derivative methods in quantum chemistry *Adv. Chem. Phys.* **69** 241
- [6] Fletcher R 1981 *Practical Methods of Optimization: Vol 1—Unconstrained Optimization* (New York: Wiley)
- [7] Dennis J E and Schnabel R B 1983 *Numerical Methods for Unconstrained Optimization and Non-linear Equations* (Englewood Cliffs, NJ: Prentice-Hall)

- [8] Miller W H 1983 Symmetry-adapted transition-state theory and a unified treatment of multiple transition states *J. Phys. Chem.* **87** 21
- [9] Spendley W, Hext G R and Himsworth F R 1962 Sequential application of simplex designs in optimization and evolutionary operation *Technometrics* **4** 441
- [10] Baker J 1987 An algorithm for geometry optimization without analytical gradients *J. Comput. Chem.* **8** 563
- [11] Pulay P 1969 *Ab initio* calculation of force constants and equilibrium geometries in polyatomic molecules. I. Theory *Mol. Phys. J.* **17** 197
- [12] Pulay P 1977 Direct use of the gradients for investigating molecular energy surfaces *Applications of Electronic Structure Theory* ed H F Schaefer III (New York: Plenum) p 153
- [13] Schlegel H B 1982 Optimization of equilibrium geometries and transition states *J. Comput. Chem.* **3** 214
- [14] Murtagh B A and Sargent R W 1970 Computational experience with quadratically convergent minimisation methods *Comput. J.* **13** 185
- [15] Fletcher R and Powell M D 1963 A rapidly convergent descent method for minimization *Comput. J.* **6** 163
-

-28-

- [16] Powell M J D 1971 Recent advances in unconstrained optimization *Math. Prog.* **1** 26
- [17] Bofill J M 1994 Updated Hessian matrix and the restricted step method for locating transition structures *J. Comput. Chem.* **15** 1
- [18] Baker J and Chan F 1996 The location of transition states: a comparison of Cartesian, Z-matrix, and natural internal coordinates *J. Comput. Chem.* **17** 888
- [19] Frisch M J *et al* 1995 *Gaussian 94* revision C.3, Gaussian (Pittsburgh, PA)
- [20] Fletcher R and Reeves C M 1964 Function minimization by conjugate gradients *Comput. J.* **7** 149
- [21] Polak E 1971 *Computational Methods in Optimization: a Unified Approach* (New York: Academic)
- [22] Pulay P 1980 Convergence acceleration in iterative sequences: the case of SCF iteration *Chem. Phys. Lett.* **73** 393
- [23] Pulay P 1982 Improved SCF convergence acceleration *J. Comput. Chem.* **3** 556
- [24] Császár P and Pulay P 1984 Geometry optimization by direct inversion in the iterative subspace *J. Mol. Struct. (Theochem)* **114** 31
- [25] Lengsfeld B H III 1980 General second-order MC-SCF theory: a density matrix directed algorithm *J. Chem. Phys.* **73** 382
- [26] Szabo A and Ostlund N S 1982 *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory* (New York: Macmillan)
- [27] Levy B 1969 Multi-configuration self-consistent wavefunctions for formaldehyde *Chem. Phys. Lett.* **4** 17
- [28] Bacskay G B 1981 A quadratically convergent Hartree–Fock (QC-SCF) method. Applications to the closed-shell case *Chem. Phys.* **61** 385
- [29] Sellers H L 1991 ADEM-DIOS, an SCF convergence algorithm for difficult cases *Chem. Phys. Lett.* **180** 461
- [30] Ionova I V and Carter E A 1995 Orbital-based direct inversion in the iterative subspace for the generalized valence bond method *J. Chem. Phys.* **102** 1251
- [31] tich I, Car R, Parrinello M and Baroni S 1989 Conjugate gradient minimization of the energy functional: a new method for electronic structure calculation *Phys. Rev. B* **39** 4997
- [32] Payne M C, Teter M P, Allan D C, Arias T A and Joanopoulos J D 1992 Iterative minimization techniques for *ab initio* total energy calculations: molecular dynamics and conjugate gradient *Rev. Mod. Phys.* **64** 1045
- [33] Lengsfeld B H III and Liu B 1981 A second-order MCSCF method for large CI expansions *J. Chem. Phys.* **75** 478

- [34] Roos B 1972 A new method for large-scale CI calculations *Chem. Phys. Lett.* **15** 153
- [35] Chaban G, Schmidt M W and Gordon M S 1997 Approximate second order methods for orbital optimization of SCF and MCSCF wavefunctions *Theor. Chim. Acta* **97** 88
- [36] Hamilton T P and Pulay P 1986 Direct inversion in the iterative subspace (DIIS) optimization of open-shell, excited-state and small multiconfigurational SCF wavefunctions *J. Chem. Phys.* **84** 5728
- [37] Werner H-J and Meyer W 1981 A quadratically convergent MCSCF method for the simultaneous optimization of several states {*J. Chem. Phys.*} **74** 5794
- [38] Werner H-J and Knowles P 1985 A second order multiconfiguration SCF procedure with optimum convergence *J. Chem. Phys.* **82** 5053
- [39] Schlegel H B 1984 Estimating the Hessian for gradient-type geometry optimizations *Theor. Chim. Acta* **66** 333
-

-29-

- [40] Fischer T H and Almlöf J 1992 General methods for geometry and wavefunction optimization *J. Phys. Chem.* **96** 9768
- [41] Lindh R, Bernhardsson A, Karlström G and Malmqvist P-Å 1995 On the use of a Hessian model function in molecular geometry optimizations *Chem. Phys. Lett.* **241** 423
- [42] Baker J and Hehre W J 1991 Geometry optimization in Cartesian coordinates: The end of the Z-matrix? *J. Comput. Chem.* **12** 606
- [43] Paizs B, Fogarasi G and Pulay P 1998 An efficient direct method for geometry optimization of large molecules *J. Chem. Phys.* **109** 6571
- [44] Farkas Ö and Schlegel H B 1998 Methods for geometry optimization in large molecules. I. An $O(N^2)$ algorithm for solving systems of linear equations for the transformation of coordinates and forces *J. Chem. Phys.* **109** 7100
- [45] Baker J, Kinghorn D and Pulay P 1999 Geometry optimization in delocalized internal coordinates: An efficient quadratically scaling algorithm for large molecules *J. Chem. Phys.* **110** 4986
- [46] Pulay P and Meyer W 1971 *Ab initio* calculation of the force field of ethylene *J. Mol. Spectrosc.* **40** 59
- [47] Pulay P, Fogarasi G, Pang F and Boggs J E 1979 Systematic *ab initio* gradient calculation of molecular geometries, force constants and dipole moment derivatives *J. Am. Chem. Soc.* **101** 2550
- [48] Fogarasi G, Zhou X, Taylor P W and Pulay P 1992 The calculation of *ab initio* molecular geometries: efficient optimization by natural internal coordinates and empirical correction by offset forces *J. Am. Chem. Soc.* **114** 8191
- [49] Pye C C and Poirier R A 1998 Graphical approach for defining natural internal coordinates *J. Comput. Chem.* **19** 504
- [50] Baker J 1993 Techniques for geometry optimization: a comparison of Cartesian and natural internal coordinates *J. Comput. Chem.* **14** 1085
- [51] Eckert F, Pulay P and Werner H-J 1997 *Ab initio* geometry optimization for large molecules *J. Comput. Chem.* **18** 1473
- [52] Sellers H L, Klimkowski V J and Schäfer L 1978 Normal coordinate *ab initio* force relaxation *Chem. Phys. Lett.* **58** 541
- [53] Pulay P and Fogarasi G 1992 Geometry optimization in redundant internal coordinates *J. Chem. Phys.* **96** 2856
- [54] Peng C, Ayala P Y, Schlegel H B and Frisch M J 1996 Using redundant internal coordinates to optimize equilibrium geometries and transition states *J. Comput. Chem.* **17** 49
- [55] Baker J, Kessi A and Delley B 1996 The generation and use of delocalized internal coordinates in geometry optimization *J. Chem. Phys.* **105** 192
- [56] Stewart J J P 1989 Optimization of parameters for semiempirical wavefunctions *J. Comput. Chem.* **10**

- [57] Baker J 1986 An algorithm for the location of transition states *J. Comput. Chem.* **7** 385
- [58] Crawford B Jr and Fletcher W H 1951 The determination of normal coordinates *J. Chem. Phys.* **19** 141
- [59] Fletcher R 1981 *Practical Methods of Optimization: vol. 2—Constrained Optimization* (New York: Wiley)
- [60] Baker J 1992 Geometry optimization in Cartesian coordinates: constrained optimization *J. Comput. Chem.* **13** 240
- [61] Baker J and Bergeron D 1993 Constrained optimization in Cartesian coordinates *J. Comput. Chem.* **14** 1339
- [62] Baker J 1997 Constrained optimization in delocalized internal coordinates *J. Comput. Chem.* **18** 1079
- [63] Lu D-H, Zhao M and Truhlar D G 1991 Projection operator method for geometry optimization with constraints *J. Comput. Chem.* **12** 376
-

- [64] Taylor H and Simons J 1985 Imposition of geometrical constraints on potential energy walking procedures *J. Phys. Chem.* **89** 684
- [65] Halgren T A and Lipscomb W N 1977 The synchronous transit method for determining reaction pathways and locating molecular transition states *Chem. Phys. Lett.* **49** 225
- [66] Ehrenson S 1974 Analysis of least motion paths for molecular deformations *J. Am. Chem. Soc.* **96** 3778
- [67] Bell S and Crighton J 1984 Locating transition states *J. Chem. Phys.* **80** 2464
- [68] Ionova I V and Carter E A 1993 Ridge method for finding saddle points on potential energy surfaces *J. Chem. Phys.* **98** 6377
- [69] Cerjan C J and Miller W H 1981 On finding transition states *J. Chem. Phys.* **75** 2800
- [70] Simons J, Jørgensen P, Taylor H and Ozment J 1983 Walking on potential energy surfaces *J. Phys. Chem.* **87** 2745
- [71] Banerjee A, Adams N, Simons J and Shepard R 1985 Search for stationary points on surfaces *J. Phys. Chem.* **89** 52
- [72] Müller K and Brown L D 1979 Location of saddle points and minimum energy paths by a constrained simplex optimization procedure *Theor. Chim. Acta* **53** 75
- [73] Dewar M J S, Healy E F and Stewart J J P 1984 Location of transition states in reaction mechanisms *J. Chem. Soc. Faraday Trans. II* **80** 227
- [74] Koga N and Morokuma K 1985 Determination of the lowest energy point on the crossing seam between two potential surfaces using the energy gradient *Chem. Phys. Lett.* **119** 371
- [75] McDouall J J W, Robb M A and Bernardi F 1986 An efficient algorithm for the approximate location of transition structures in a diabatic surface formalism *Chem. Phys. Lett.* **129** 595
- [76] Jensen F 1994 Transition structure modeling by intersecting potential energy surfaces *J. Comput. Chem.* **15** 1199
- [77] Bendt P and Zunger A 1982 New approach for solving the density functional self-consistent field problem *Phys. Rev. B* **26** 3114
- [78] Head-Gordon M and Pople J A 1988 Optimization of wavefunction and geometry in the finite basis Hartree–Fock method *J. Phys. Chem.* **92** 3063
- [79] Almlöf J 1995 Direct methods in electronic structure theory *Modern Electronic Structure Theory* ed D Yarkony (Singapore: World Scientific) pp 110–51
- [80] Car R and Parrinello M 1985 Unified approach for molecular dynamics and density functional theory *Phys. Rev. Lett.* **55** 2471
- [81] Field M J 1991 Constrained optimization of *ab initio* and semiempirical Hartree–Fock wavefunctions using

direct minimization or simulated annealing *J. Phys. Chem.* **95** 5104

- [82] Hartke B and Carter E A 1992 Spin eigenstate-dependent Hartree–Fock molecular dynamics *Chem. Phys. Lett.* **189** 358
- [83] Fukui K 1970 A formulation of the reaction coordinate *J. Phys. Chem.* **74** 4161
- [84] Ishida K, Morokuma K and Komornicki A 1977 The intrinsic reaction coordinate. An *ab initio* calculation for $\text{HCN} \rightarrow \text{HNC}$ and $\text{H}^- + \text{CH}_4 \rightarrow \text{CH}_3 + \text{H}^-$ *J. Chem. Phys.* **66** 2153
- [85] Baldrige K K, Gordon M S, Steckler R and Truhlar D G 1989 *Ab initio* reaction paths and direct dynamics calculations *J. Phys. Chem.* **93** 5107
-

-31-

- [86] Melissas V S, Truhlar D G and Garrett B C 1992 Optimized calculations of reaction paths and reaction-path functions for chemical reactions *J. Chem. Phys.* **96** 5758
- [87] Page M, Doubleday C and McIver J W Jr 1990 Following steepest descent reaction paths. The use of higher energy derivatives with *ab initio* electronic structure methods *J. Chem. Phys.* **93** 5634 and references therein
- [88] Sun J-Q and Ruedenberg K 1993 Quadratic steepest descent on potential energy surfaces. I. Basic formalism and quantitative assessment *J. Chem. Phys.* **99** 5257
- [89] Gonzales C and Schlegel H B 1991 Improved algorithms for reaction path following: higher-order implicit algorithms *J. Chem. Phys.* **95** 5853
- [90] Schlegel H B 1994 Some thoughts on reaction-path following *J. Chem. Soc. Faraday Trans.* **90** 1569
- [91] Elber R and Karplus M 1987 A method for determining reaction paths in large molecules: application to myoglobin *Chem. Phys. Lett.* **139** 375
- [92] Stachó L L and Bán M I 1992 A global strategy for determining reaction paths *Theor. Chim. Acta* **83** 433
- [93] Valtazanov P and Ruedenberg K 1986 Bifurcations and transition states *Theor. Chim. Acta* **69** 281
- [94] Baker J and Gill P M W 1988 An algorithm for the location of branching points on reaction paths *J. Comput. Chem.* **9** 465
- [95] Kirkpatrick S, Gelatt C D Jr and Vecchi M P 1983 Optimization by simulated annealing *Science* **220** 671
- [96] Golding D E 1989 *Genetic Algorithms in Search, Optimization and Machine Learning* (Reading, MA: Addison Wesley)
- [97] Piela L, Kostrowicki J and Scheraga H A 1989 The multiple-minima problem in the conformational analysis of molecules. Deformation of the potential energy hypersurface by the diffusion equation method *J. Phys. Chem.* **93** 3339
- [98] Floudas C and Pardalos P M 1991 *Recent Advances in Global Optimization* (Princeton, NJ: Princeton University Press)
-

FURTHER READING

Fletcher R 1981 *Practical Methods of Optimization: Vol 1—Unconstrained Optimization; Vol. 2—Constrained Optimization* (New York: Wiley)

A classic in the field, very readable, highly recommended, full of practical advice.

Polak E 1997 *Optimization: Algorithms and Consistent Approximations* (New York: Springer)

A complete and mathematically precise treatment of the subject. Includes topics which are not usually

discussed in introductory texts. Complete with exercises. Best suited for the mathematically inclined reader.

Dennis J E and Schnabel R B 1983 *Numerical Methods for Unconstrained Optimization and Non-linear Equations* (Englewood Cliffs, NJ: Prentice-Hall)

A very pedagogical, highly readable introduction to quasi-Newton optimization methods. It includes a modular system of algorithms in pseudo-code which should be easy to translate to popular programming languages like C or Fortran.

-32-

Schlegel H B 1995 Geometry optimization on potential energy surfaces *Modern Electronic Structure Theory* ed D Yarkony (Singapore: World Scientific) pp 459–500

An excellent, up-to-date treatise on geometry optimization and reaction path algorithms for *ab initio* quantum chemical calculations, including practical aspects.

Werner H-J 1987 Matrix-formulated direct multiconfigurational self-consistent field and multireference configuration interaction methods *Adv. Chem. Phys.* **69** 1

A lucid and carefully written exposition of this difficult subject from one of the authors of the highly acclaimed MOLPRO suite of programs. It contains examples and plenty of physical insight.

Shepard R 1987 The multiconfiguration self-consistent field method *Adv. Chem. Phys.* **69** 63

A very detailed, pedagogical treatment of the subject, including much of the mathematical background and a nearly complete list of references prior to 1987.

Pulay P 1995 Analytical derivative techniques and the calculation of vibrational spectra *Modern Electronic Structure Theory* ed D Yarkony (Singapore: World Scientific) pp 1191–240

A concise introduction to the calculation of analytical derivatives in quantum chemistry, with applications to simulating vibrational spectra.

-1-

B3.6 Mesoscopic and continuum models

Marcus Müller

B3.6.1 INTRODUCTION

Many systems in physical chemistry exhibit structure on length scales that greatly exceed the atomic dimensions. Systems containing surfactants—detergents or milk, for instance—often consist of droplets of one component dissolved in another phase. The size of these droplets exceeds the extension of the molecular constituents by far. Very generally, mesoscopic and continuum models describe the properties of materials on length scales larger than the atomic dimensions by incorporating the details of the underlying atomic structure only in terms of a reduced number of effective variables. In this very broad sense, the Navier–Stokes equation

[1], which describes the motion of a fluid via a density, energy, and velocity field and elasticity theory [2], and which describes solids in terms of stress and displacement fields, also belongs to this class of model. In both approaches, the subject of the model is not the properties of individual atoms (e.g., their position or quantum state) but rather their average properties (like the density or velocity) in a small coarse-graining volume. Usually the coarse graining is not performed explicitly, but it is understood that the averaging volume is large enough to result in a continuous spatial variation of the variables of the mesoscopic model and yet still be smaller than the characteristic length scale of the phenomena under consideration.

In the following entry we shall restrict ourselves to discussing mesoscopic and continuum models for complex fluids in chemical physics. The wide span of time and length scales in these materials is illustrated in [figure B3.6.1](#) for a blend of two polymers. On the atomistic scale each polymer consists of chemical repeat units joined together to form the chain molecule. The length scale is set by the distance between the atoms along the backbone of the polymer, typically in the range of 1–2 Å. The vibrations of the atoms occur on the timescale of picoseconds. In a dense melt, the flexible chain molecules adopt a random-walk-like conformation. The ‘step length’ of the random walk, or persistence length b , is typically of the order of a few nanometres. Since several thousands of repeat units form a polymer, the overall size of a single molecule, as specified by its radius of gyration, exceeds the persistence length by 1–3 orders of magnitude. On this range of length scales the structure of the polymer is self-similar. If the two components of the blend are not miscible, as it is generally the case, one species forms droplets that are dispersed in a matrix of the other species. The size of the droplets is in the micrometre range. On even larger length scales (say 1 mm) the material appears homogeneous. Clearly the properties on the mesoscopic length scale are important for application properties. A decrease of the droplet size or even the formation of a connected morphology (i.e. a microemulsion) improves the mechanical properties of the composite material. A similar span of time and length scales is encountered in many other systems (e.g., mixtures of oil, water and surfactant or glassy materials) and this behaviour is rather typical for complex fluids.

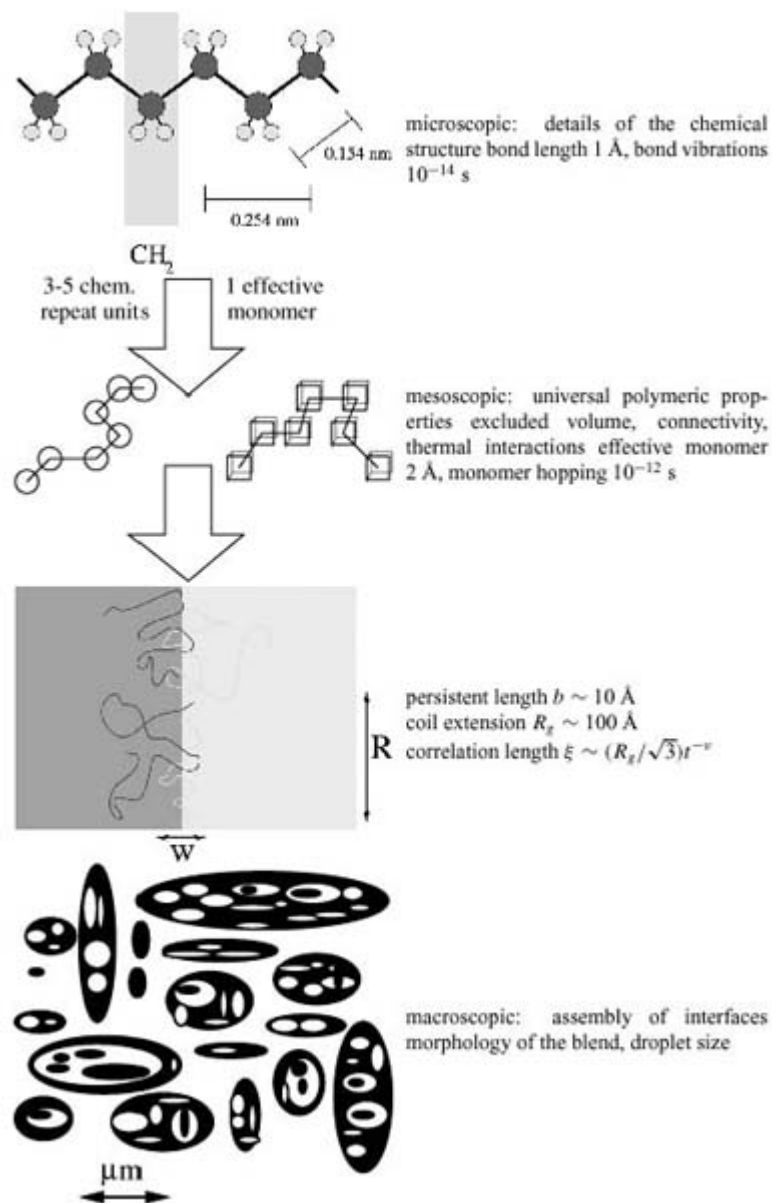


Figure B3.6.1. Illustration of the wide span of length scale in a binary polymer blend. (See the text for further explanation.)

A unified model that describes the structure from the atomistic length scale up to macroscopic properties is not analytically tractable. Even state-of-the-art supercomputers cannot cope with such a broad spread of time and length scales in numerical simulations. Today's largest simulated systems in thermal equilibrium comprise about 10^7 particles and, hence, span about 2–3 decades in length scales. With the increase of computing power and progress in simulation methodology, simulating larger and larger system sizes will become feasible, but computer modelling from atomic to macroscopic scales in the framework of a single, unified model is not feasible at present or in the near future.

Another caveat for the modelling from the atomistic level up to the macroscopic level is the requirement of sufficiently accurate interaction potentials. Minor inaccuracies in calculations on small length scales can give rise to pronounced effects on the mesoscopic scale. Consider, for instance the self-assembly of amphiphilic molecules (see section B3.6.3) into a spatially ordered structure. The free-energy difference between the different morphologies can be as small as $10^{-4}kT$ per molecule. The *ab initio* prediction of such a small free-

energy difference is certainly a formidable task.

Mesoscopic and continuum models do not attempt to describe large-scale phenomena starting from the smallest atomic length scale, but rather incorporate the local structure via a small number of effective parameters. Mesoscopic models lump a small number of atoms into an effective particle. These particles interact via coarse-grained interactions. By this coarse-graining procedure much of the atomistic detail is lost, and only those interactions pertinent to the phenomena on the mesoscopic length scales are retained. Even if the interactions on the microscopic scale are extremely complex (e.g., hydrophobic interactions [3] in lipid water mixtures), they can often be captured by simple expressions on the mesoscopic length scale. Coarse-grained models thus yield valuable insights into the structure on large length scales. For specific examples the effective interactions are derived by eliminating the degrees of freedom on the smallest (atomistic) length scales, retaining only those on larger length scales; for some systems (e.g., polymer chains in the gas phase) this coarse-graining procedure has a formal justification due to the self-similar structure on a large range of length scales; for other systems the mapping between the atomistic/microscopic level and the mesoscopic description is rather a concept than a practicable procedure. In this latter case, the application of mesoscopic models rests on the observation that different systems (e.g., diblock copolymers and lipid water mixtures) share a common behaviour on mesoscopic scales. Universal mesoscopic behaviour that does not depend on the details on the atomistic level in a qualitative way is the subject of mesoscopic models.

Continuum models go one step further and drop the notion of particles altogether. Two classes of models shall be discussed: field theoretical models that describe the equilibrium properties in terms of spatially varying fields of mesoscopic quantities (e.g., density or composition of a mixture) and effective interface models that describe the state of the system only in terms of the position of interfaces. Sometimes these models can be derived from a mesoscopic model (e.g., the Edwards Hamiltonian for polymeric systems) but often the Hamiltonians are based on general symmetry considerations (e.g., Landau–Ginzburg models). These models are well suited to examine the generic universal features of mesoscopic behaviour.

Mesoscopic and continuum models bridge the gap between atomistic realistic simulations and the description on the macroscopic level, (e.g., elasticity theory). The objectives of mesoscopic models are twofold. On the one hand they help identify interactions that are necessary to bring about the phenomena on a mesoscopic scale (e.g., phase separation or self-assembly) and they aid in investigating the dependence of the mesoscopic behaviour on the effective interactions. This information also yields some qualitative insight into how the microscopic parameters influence mesoscopic behaviour (e.g., the dependence of the structure in a self-assembled system on the architecture/shape of the amphiphilic molecules). On the other hand, this class of models elucidates universal behaviour on the mesoscopic scale (e.g., identifying various morphologies into which systems can self-assemble, the relation between confinement and phase behaviour, or the consequences of fluctuations) and establishes a relation between behaviour on large length scales and experimentally accessible (mesoscopic) quantities (e.g., Flory–Huggins parameter, interfacial tension, or bending rigidity of membranes).

The hierarchy of models is complemented by a variety of methods and techniques. Mesoscopic models that incorporate some fluid-like packing (e.g., spring–bead models for polymer solutions) are investigated by Monte Carlo

simulations, molecular dynamics or density functional techniques. Lattice models are studied by Monte Carlo simulations. The larger the span of length scales considered, the larger the computational effort required. Models without pronounced packing effects (e.g., the Edwards Hamiltonian) are investigated by self-consistent field techniques. Continuum models are often analytically tractable, at least in the mean field approximation, and simple analytical expressions for various quantities (e.g., interfacial tension between two immiscible polymers) can be obtained in some limiting cases. The effect of fluctuations has been assessed by computer simulations, transfer matrix calculations and renormalization group techniques.

At the heart of mesoscopic and continuum models lies the question: Which degrees of freedom are to be retained as relevant and which can be ignored? The answer depends on the specific problem. By comparing different models the degree of universality and the relevance of interactions can be gauged. This yields much insight into the mechanisms which underly the phenomena. Mesoscopic and continuum models make contact with chemical models on the atomistic level as well as with the macroscopic descriptions. Effort is being made to incorporate more chemical realism into the models as well as to extend them to larger length scales.

In the following we shall describe various applications of mesoscopic models to complex fluids. The examples extend from applications that are quite close to the atomistic level (e.g., coarse-grained polymer models) to highly idealized models (e.g., effective interface Hamiltonians or Ginzburg–Landau models). Moreover, we restrict ourselves mainly to the description of thermodynamic equilibrium. The remainder of this entry is organized as follows. In section B3.6.2 we discuss applications of coarse-grained models to systems involving homopolymers. Mesoscopic models for the description of self-repelling chains, polymer solutions, polymer melts and binary blends are introduced. From these models, more coarse-grained descriptions can be derived in terms of Ginzburg–Landau expansions or effective interface Hamiltonians. [Section B3.6.3](#) then considers amphiphilic molecules. Their co-operative behaviour on the supramolecular level has been explored in the framework of models with various degrees of detail. Chain models retain the salient features of the amphiphile’s architecture while lattice models or continuum models yield a description in terms of a spatially varying concentration. On even larger scales, the statistical mechanics of interfaces has been investigated via random interface models. This article closes with a brief look at the application of mesoscopic and continuum models to dynamical phenomena.

B3.6.2 POLYMERIC SYSTEMS

B3.6.2.1 POLYMER SOLUTIONS

Coarse-grained models have a longstanding history in polymer science. Long-chain molecules share many common mesoscopic characteristics which are independent of the atomistic structure of the chemical repeat units [4, 5 and 6]. The self-similar structure [7, 8, 9 and 10] on large length scales is only characterized by a single length scale, the chain extension R .

The important interactions in polymer solutions are the connectivity of the segments along the chain molecules and interactions between segments. The solvent molecules are often not treated explicitly, but their effect is incorporated into the effective interactions between polymer segments. A good solvent corresponds to an effective repulsion between segments and the polymer chains adopt a swollen configuration. A bad solvent gives rise to an attraction between the polymer segments and leads to a collapse.

The observation of the universality and self-similarity of the large-length-scale properties has a theoretical basis. In 1972 de Gennes [11] related the structure of a polymer chain in a good solvent to a field theory of a n component vector model in the limit $n \rightarrow 0$. This class of models (see the entry on phase transitions and critical phenomena; A2.5) exhibits a continuous phase transition and the properties close to this critical point have been investigated extensively with renormalization group calculations [8, 9 and 10]. The inverse chain length plays the role of the distance from the critical point of the $n = 0$ component vector model. As in the theory of critical phenomena, the behaviour in the vicinity of this critical point (i.e. $1/N \ll 1$) is governed by a universal scaling behaviour that is brought about by only a few relevant interactions. The relation between the behaviour of polymer chains in the limit of $N \rightarrow \infty$ and the critical behaviour justifies the use of highly coarse-grained models that incorporate only two relevant interactions: connectivity along the chain and binary segmental interactions.

Lattice models of polymer solutions are a particularly simple and computationally efficient realization, and therefore have attracted abiding interest [12]. In simple lattice models, a small group of atomistic repeat units is represented by a site on a simple cubic lattice. Segments along a polymer occupy neighbouring lattice sites and multiple occupation of lattice sites is forbidden (excluded volume). The latter constraint corresponds to the repulsive binary interaction under good solvent conditions. Isolated chains on the lattice adopt configurations of self-avoiding walks. The polymer's end-to-end distance R scales with the chain length like $R \sim N^\nu$. The exponent $\nu = 0.588$ has been calculated using renormalization group techniques [9, 10], enumeration techniques for short chain lengths and Monte Carlo simulations [13].

The application of lattice models to study the behaviour of multi-chain systems (i.e. dilute and semi-dilute solutions, and dense melts) is straightforward in principle. The equilibration of dense multi-chain systems is, however, a challenging problem for computer simulations, and simple lattice models have been a testing bed for many algorithms. Some methods are tailored to isolated chains or very dilute systems (e.g., the pivot algorithm [13] or the construction of a chain via the pruned-enriched Rosenbluth method [14]); other methods provide an effective relaxation of the overall chain dimensions in dense systems (e.g., configurational bias Monte Carlo [15, 16] or the recoil growth algorithm [17]).

Though these simple lattice models reproduce the universal features of polymer solutions, it is difficult to incorporate details of the chain architecture. The simple lattice model allows only for two bond angles which makes the investigation of orientational effects prone to lattice artefacts. Moreover the particles in real fluids arrange to form neighbouring shells. This local packing structure of the fluid does not affect the universal scaling behaviour but it is pertinent to the relation between the coarse-grained effective interactions and the underlying microscopic potentials. Since the vacancies on the lattice and the polymer segments have the same size, packing effects in the density correlation function are largely absent. More sophisticated lattice models (e.g., the bond fluctuation model [18]), in which monomers are represented by extended objects (e.g., a whole unit cube) on the lattice, have been explored. These models exhibit packing effects and a large number of bond angles while still retaining the computational advantages of lattice models. They also allow for a diffusive dynamics of the polymers on the lattice which consists of random local displacements of the monomers. Moreover, the bond vectors can be chosen such that the excluded volume constraint prevents bonds from crossing through each other in the course of these local displacements. This non-crossability takes account of topological effects which are important for the dynamical properties of linear chains [19] and influence the conformational statistics of ring polymers [20, 21 and 22] (e.g., in order to avoid topological interactions rings collapse in a concentrated solution).

Off-lattice models enjoy a growing popularity. Again, a particle corresponds to a small number of atomistic repeat units

along the backbone of the polymer. Off-lattice models allow simulations at constant pressure or the calculation of the pressure via the virial expression. This yields direct access to the pVT behaviour. By modelling polymers as a sequence of tangent hard spheres in continuous space, computer simulators have investigated the equation of state in polymer solutions and the detailed packing structure of polymer solutions in contact with a hard wall. This class of model is particularly suited for comparing the results to analytical theories (e.g., Wertheim's theory [23] or density functional approaches [24, 25 and 26]) because of the existence of elaborated analytical descriptions for the corresponding hard-sphere monomer fluid.

Hard-sphere models lack a characteristic energy scale and, hence, only entropic packing effects can be investigated. A more realistic modelling has to take hard-core-like repulsion at small distances and an attractive interaction at intermediate distances into account. In non-polar liquids the attraction is of the van der Waals type and decays with the sixth power of the interparticle distance r . It can be modelled in the form of a Lennard-Jones potential $V_{LJ}(r)$ between segments

$$V_{\text{LJ}}(\mathbf{r}) = 4\epsilon \left\{ \left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right\} \quad (\text{B3.6.1})$$

where the exponent of the first, repulsive term is chosen for computational convenience. The Lennard-Jones radius σ sets the microscopic length scale and ϵ sets the energy scale. In many simulational applications the potential is truncated and shifted so as to yield a continuous, finite ranged potential. This does not alter the qualitative behaviour but shifts the temperature and density of the liquid–vapour critical point. The Lennard-Jones particles are tied together to form chain molecules. The constraint of fixed bond length or an harmonic bonding potential has been employed. Another popular choice is the FENE potential [27, 28]. It takes the form

$$V_{\text{FENE}}(r) = -\frac{k}{2} R_0^2 \ln \left(1 - \frac{r^2}{R_0^2} \right) \quad \text{with} \quad R_0 = 1.5\sigma. \quad (\text{B3.6.2})$$

The parameter k tunes the stiffness of the potential. It is chosen such that the repulsive part of the Lennard-Jones potential makes a crossing of bonds highly improbable (e.g., $k = 30$). This off-lattice model has a rather realistic equation of state and reproduces many experimental features of polymer solutions. Due to the attractive interactions the model exhibits a liquid–vapour coexistence, and an isolated chain undergoes a transition from a self-avoiding walk at high temperatures to a collapsed globule at low temperatures. Since all interactions are continuous, the model is tractable by Monte Carlo simulations as well as by molecular dynamics. Generalizations of the Lennard-Jones potential to anisotropic pair interactions are available: e.g., the Gay–Berne potential [29]. This latter potential has been employed to study non-spherical particles that possibly form liquid crystalline phases.

In the limit that the number of effective particles along the polymer diverges but the contour length and chain dimensions are held constant, one obtains the Edwards model of a polymer solution [9, 30]. Polymers are represented by random walks that interact via zero-ranged binary interactions of strength v . The partition function of an isolated chain is given by

-7-

$$\mathcal{Z} = \int \mathcal{D}[\mathbf{r}(t)] \exp \left(-\frac{3}{2b^2} \int_0^N dt \left(\frac{d\mathbf{r}}{dt} \right)^2 \right) \exp \left(- \int d\mathbf{r} d\mathbf{r}' \hat{\rho}(\mathbf{r}) v \delta(\mathbf{r} - \mathbf{r}') \hat{\rho}(\mathbf{r}') \right) \quad (\text{B3.6.3})$$

where the density field $\hat{\rho}$ is related to the configuration $\mathbf{r}(t)$ of the polymer via

$$\hat{\rho}(\mathbf{r}') = \int_0^N dt \delta(\mathbf{r}' - \mathbf{r}(t)). \quad (\text{B3.6.4})$$

The path integral \mathcal{D} sums over all polymer conformations $\mathbf{r}(t)$, where $0 \leq t \leq N$ denotes the contour parameter along the polymer. The second term represents the connectivity along the molecule and b denotes the persistence length (i.e. the ‘step length’ of the random walk). In the absence of the third term, the partition function describes a Gaussian chain with the end-to-end distance $\mathbf{R} = b\sqrt{N}$ (Gaussian chain model). This is the only length scale in the problem. Very much like in quantum mechanics, the path integral results in a diffusion equation (i.e. the polymer analogue of Schrödinger’s equation for the propagator) for the probability of finding a chain’s segment after t steps along the chain at position \mathbf{r} in space. The third term describes the interactions: if segments t and t' are located at the same position they interact with the strength v .

This model is very popular for analytical calculations and generalizations to multi-chain systems are straightforward. Properties of polymer solutions have been obtained via renormalization group techniques [8, 9 and 10]. Similar to the simple lattice model the Edwards model includes only the chain connectivity and binary segmental interactions: the detailed structure of the underlying fluid is omitted. For $\nu = 0$ the self-similar Gaussian statistics persist on all length scales and there is no rod-like behaviour on smaller scales. Generalizations of thread-like models to stiff polymers and orientational interactions, however have been explored. The most popular one is the wormlike chain model, in which the second term in the Edwards Hamiltonian (B3.6.3) is replaced by [31]

$$\exp\left(-\frac{\eta}{2}\int_0^N dt\left(\frac{du}{dt}\right)^2\right) \quad (\text{B3.6.5})$$

where $u(t) = dr/dt$ denotes the tangent vector along the path with unit norm. The parameter η controls the local stiffness of the path. On small distances along the chain the tangent vectors u are highly correlated and the stiffness parameter η controls the decay of orientational correlations along the chain's contour: $\langle u(t)u(t') \rangle = \exp(-|t-t'|/\eta)$. On large length scales, however, the Gaussian behaviour is recovered and the end-to-end distance is given by $R^2 = 2\eta N$.

B3.6.2.2 POLYMER BLENDS

The above models can be generalized to multicomponent systems by modifying the segmental interactions. Most applications deal with a binary blend in a common solvent. The excluded volume interaction between segments limits density fluctuations in a dense polymer liquid. Therefore many models of dense multicomponent systems neglect the finite compressibility and model the excluded volume interactions by enforcing a uniform segment density. In 1941, Flory [32] and Huggins [33] employed a simple lattice model to calculate the phase diagram for a dense binary polymer blend in mean field approximation. The two polymer species—denoted A and B—are modelled as walks on a lattice,

-8-

and the binary interactions of strengths ϵ_{AA} , ϵ_{AB} and ϵ_{BB} act between neighbours on the lattice. Since all lattice sites are occupied, only the difference of the segmental interactions (the Flory–Huggins parameter)

$$\chi_{FH} = \frac{z}{kT} \left(\epsilon_{AB} - \frac{\epsilon_{AA} + \epsilon_{BB}}{2} \right) \quad (\text{B3.6.6})$$

determines the phase diagram. Here, $z = 6$ denotes the coordination number of the simple cubic lattice. Typical experimental values of the Flory–Huggins parameter χ are in the range 10^{-2} – 10^{-5} for partially compatible blends while the individual interactions between the segments ϵ_{ij} ($i, j = A, B$) are of the order of $k_B T$. This illustrates that the phase behaviour is governed by a delicate cancellation of interactions. Starting from a model with atomistic details and performing an *ab initio* calculation of the packing structure and the effective segmental interactions in a binary blend would require extremely accurate interatomic potentials as input and a very high numerical quality of the calculation. Therefore, predicting the value of the Flory–Huggins parameter on an *ab initio* basis is virtually impossible. The concept of describing the effective incompatibility of two polymer species by a single mesoscopic parameter χ_{FH} has proven remarkably successful, however. When χ_{FH} is used as an adjustable parameter the mean field theory of Flory and Huggins is quite successful in describing many experimental observations. The values of the χ_{FH} parameter of various pairs of polymers have been extracted from a comparison between theory and experiment and are compiled in, for example, [34]. In the framework of the mean field theory, the excess free energy of mixing per segment

takes a particularly simple form:

$$\frac{\Delta F}{\rho V k T} = \frac{\phi_A}{N_A} \ln \phi_A + \frac{\phi_B}{N_B} \ln \phi_B + \chi_{FH} \phi_A \phi_B \quad (\text{B3.6.7})$$

where N_A and N_B denote the chain length of the two polymer species and ϕ_A and ϕ_B denote the relative amount of A or B segments, respectively. $\phi_A + \phi_B = 1$. The first term represents the translational entropy of mixing. Due to the connectivity of the segments it is reduced by a factor $1/N_A$ or $1/N_B$, respectively. The second term describes the repulsion between unlike segments. The chain conformations are assumed to be independent of the composition. Therefore the conformational entropy does not give any contribution to the free energy of mixing in the Flory–Huggins treatment. Most notably, the theory rationalizes the fact that long macromolecules tend to demix, because a small repulsion is sufficient to far outweigh the entropy of mixing, which is reduced by the factor $1/N$. This expression for the excess free energy of mixing also forms the basis for self-consistent field models of spatially inhomogeneous systems.

In order to gain qualitative insight into how to relate the Flory–Huggins parameter to the architectural properties of the components, mesoscopic models with various degrees of structural detail have been investigated. Complex lattice models allow monomeric units to occupy more than one lattice site. In the lattice cluster model of Freed and co-workers [35] the effect of explicit monomer structure has been explored. The partition function of the model is expressed in a systematic double expansion with respect to the inverse temperature and the inverse coordination number of the underlying lattice. To zero order the approach recovers the results of the original Flory–Huggins theory. Higher-order terms account for geometric packing on the monomer scale and non-random mixing effects. This approach has been successful in predicting various subtle influences of the monomer architecture, including the occurrence of entropic contributions to the Flory–Huggins parameter.

-9-

Similar questions can be addressed by the P-RISM (polymer reference interaction site model) theory of Curro and Schweizer [36]. This integral equation theory generalizes the Ornstein–Zernike equation to polymeric systems in order to account for the fluid-like packing structure. Details of the molecular architecture enter via the single-chain structure factor. The P-RISM approach yields a detailed description of the phase behaviour and the local structure and has been applied to models with various degrees of structural detail. In the limit that the chains are modelled as infinitely thin Gaussian paths the results are very similar to the Flory–Huggins theory. The theory has been applied to fairly realistic chain models taking the experimentally measured single-chain structure factors as input. More recently, this approach has been applied self-consistently to calculate the change of the molecular conformation upon blending.

The bond fluctuation model [37] and off-lattice [38, 39] models have been used to investigate the binary polymer blends within Monte Carlo simulations. Attention has focused on rather different topics: (i) Monte Carlo simulations appropriately account for the effect of composition fluctuations. They are important in the vicinity of the critical temperature of the unmixing transition. When the chain length is increased, this fluctuation-dominated region shrinks and one observes a crossover between the 3D Ising universality class and mean field critical behaviour [40]. (ii) The relation between the polymer architecture and the Flory–Huggins parameter has been explored in simulations. Disparities in the architecture on the scale of the coarse-grained monomers (e.g., different local stiffness of the chains or different monomer shapes) alter the packing structure and give rise to enthalpic and entropic contributions to the Flory–Huggins parameter [37, 39]. When comparing experimental data to the predictions of the mean field theory, deviations from the simple proportionality $\chi_{FH} \sim 1/T$ of the Flory–Huggins parameter are rather the rule than an exception. (iii) Monte Carlo simulations reveal that the chains in the minority phase shrink. By reducing their size, they increase the local density of their own monomers and reduce the number of unfavourable contacts with the opposite species. The latter effect is, however, not captured in simple mean field theories. (iv) Off-lattice models have

been employed to study binary blends at constant pressure and to explore the effect of compressibility on the miscibility behaviour [38, 39].

These coarse-grained approaches investigate the generic behaviour and the qualitative dependence on the chain architecture. Again it should be pointed out that these simulations and analytical methods cannot predict the absolute value of the Flory–Huggins parameter of a specific pair of polymers. However by a careful choice of the coarse-grained model, they help in identifying relevant parameters for the miscibility on a coarse-grained scale.

B3.6.2.3 SELF-CONSISTENT FIELD APPROACH AND GINZBURG–LANDAU MODELS

Long polymers tend to demix and the properties of the interfaces between the coexisting phases have attracted longstanding interest. Using the Gaussian chain model, Helfand and Tagami [41] investigated the interfacial properties in the self-consistent field theory. Within the mean field approximation the problem of interacting polymers is formulated in terms of a single-chain problem in an effective, external field. This effective, external field replaces the interactions with the surrounding polymers in the binary A/B blend. The effective field

$$w_A(\mathbf{r}) = \xi(\mathbf{r})\{\phi_A(\mathbf{r}) + \phi_B(\mathbf{r}) - 1\} + \chi\phi_B(\mathbf{r}) \quad (\text{B3.6.8})$$

acts on a monomer of species A at position \mathbf{r} with ϕ_A and ϕ_B denoting the local composition of the blend. A similar equation holds for w_B . The first term enforces the incompressibility; the factor ξ is adjusted to comply with the constraint $\phi_A(\mathbf{r}) + \phi_B(\mathbf{r}) = 1$ everywhere. The second term describes the repulsion between different species parameterized by χ . The local composition, in turn, depends on the fields and is obtained as the Boltzmann average of

-10-

isolated A and B chains in the fields w_A and w_B , respectively. For Gaussian chains, described by the first part of the Hamiltonian (B3.6.3), this leads to a diffusion equation in the external potential w_A and w_B . Since the fields depend on the local composition the equations have to be solved self-consistently.

Helfand and Tagami calculated the composition profiles across the interface and determined the interfacial tension. In general, the self-consistent field equations have to be solved numerically. Different schemes in real space [42], on lattices [43] and in Fourier representation [44] have been devised. There are, however, two interesting limits in which simple analytical expressions for the interfacial width and the interfacial tension can be obtained. The limit in which the width of the interface is much smaller than the extension of the polymer and yet larger than the persistence length b is called the strong segregation limit. It corresponds to the range $1 \gg \chi \gg 1/N$ of incompatibility. This strong segregation limit is only accessible for long chain lengths N and corresponds to truly polymeric behaviour. The interfacial width w and tension γ are described by the simple forms

$$w = b/\sqrt{6\chi} \quad \text{and} \quad \gamma = \rho b\sqrt{\chi/6} \quad (\text{B3.6.9})$$

where ρ denotes the monomer density. The leading corrections to the strong segregation behaviour are of the order $1/\chi N$ and have been the subject of much investigation [45, 46 and 47]. Of course, the Gaussian chain model cannot describe the structure on length scales smaller than or comparable to the persistence length b of the polymer. This restricts the application of this mesoscopic model to the range $\chi \ll 1$.

The binary polymer blend exhibits a second-order unmixing transition. Close to the critical temperature the

concentration of the two coexisting phases does not differ very much and the characteristic length scale of composition fluctuations ξ or the interfacial width w are large compared with the size R of the polymer coil. In this weak segregation limit polymer blends behave very similarly to mixtures of small molecules in the vicinity of the critical point. The difference between the composition of the coexisting phases and the composition of the mixture at the critical point defines the order parameter m of the unmixing transition (see the entry on phase transitions and critical phenomena; A2.5). It increases with a universal power law upon cooling the system below the critical temperature T_c :

$$m \sim t^\beta \quad \text{with} \quad t = \frac{T_c - T}{T_c} > 0. \quad (\text{B3.6.10})$$

β is the critical exponent and t denotes the reduced distance from the critical temperature. In the vicinity of the critical point, the free energy can be expanded in terms of powers and gradients of the local order parameter $m(\mathbf{r}) = \phi_A(\mathbf{r}) - \phi_B(\mathbf{r})$:

$$\frac{F[m(\mathbf{r})]}{\rho k T} \sim \int d^3 \mathbf{r} \left\{ f(m) + \frac{l^2}{2} (\nabla m)^2 \right\} \quad \text{with} \quad f(m) = -\frac{t}{2} m^2 + \frac{1}{12} m^4. \quad (\text{B3.6.11})$$

This form is called a Ginzburg–Landau expansion. The first term $f(m)$ corresponds to the free energy of a homogeneous (bulk-like) system and determines the phase behaviour. For $t > 0$ the function f exhibits two minima at $m = \pm\sqrt{3t}$. This value corresponds to the composition difference of the two coexisting phases. The second contribution specifies the cost of an inhomogeneous order parameter profile. l sets the typical length scale.

-11-

The general form of the expansion is dictated by very general symmetry considerations; the specific coefficients for the example of a polymer blend can be derived from the self-consistent field theory. For a binary blend this yields $l^2 = b^2 N / 18$. In mixtures of small molecules the coefficient is determined by the range of the interactions; in polymeric systems the coefficient is associated with the conformational entropy. It is the shape of the extended molecule and its deformation at a spatial inhomogeneity that gives rise to the free energy cost.

Ginzburg–Landau models constitute a widely used example of continuum models. This class of continuum models describes the generic behaviour of all binary mixtures close to the unmixing transition. The properties of the specific model enter only via the coefficients of the expansion which set the energy scale and length scale. Extensions to different transitions (e.g., first-order transitions or microemulsion) are available (see also [section B3.6.3](#)). This approach, however, does not incorporate any structural detail of the underlying systems and hence becomes quantitatively inaccurate at lower temperatures, where the coexisting phases differ more strongly in their composition or the characteristic length scale (i.e. the correlation length ξ) becomes comparable to the size of the molecules.

Within this continuum approach Cahn and Hilliard [48] have studied the universal properties of interfaces. While their elegant scheme is applicable to arbitrary free-energy functionals with a square gradient form we illustrate it here for the important special case of the Ginzburg–Landau form. For an ideally planar interface the profile depends only on the distance z from the interfacial plane. In mean field approximation, the profile $m(z)$ minimizes the free-energy functional (B3.6.11). This yields the Euler–Lagrange equation

$$\frac{\delta \mathcal{F}}{\delta m} = 0 \quad \Rightarrow \quad -tm + \frac{1}{3} m^3 - l^2 \frac{d^2 m}{dz^2} = 0 \quad (\text{B3.6.12})$$

which, in turn, is solved by a simple function

$$m = \pm\sqrt{3}l^{1/2} \tanh\left(\frac{z}{w}\right) \quad \text{with} \quad w = \frac{\sqrt{2}l}{l^{1/2}}. \quad (\text{B3.6.13})$$

In the vicinity of the critical point (i.e. $|t| \ll 1$) the interfacial width w is much larger than the microscopic length scale l and the Landau–Ginzburg expansion is applicable.

Both Monte Carlo simulations of lattice models [49, 50] and spring–bead models [51] have been employed to study interfaces in polymeric systems. The simulations yield insight into the local properties of the polymeric fluid. Unlike in the Landau–Ginzburg expansion, the notion of polymers is retained and the orientation of the extended molecules at the interface or the enrichment of end segments have been studied. Moreover, the simulations incorporate fluctuations, which are ignored in the mean field approximation. In the vicinity of the critical temperature composition fluctuations are important. The mean field treatment overestimates the critical point and the binodals are flatter in the simulations which exhibit 3D Ising critical behaviour ($\beta = 0.324$) than in the mean field case ($\beta = 1/2$). The importance of composition fluctuations can be gauged by the Ginzburg criterion [52]: The neglect of fluctuations is justified when the order parameter fluctuations in one ‘correlation volume’ of size ξ^3 are small compared with the order parameter itself. In the case of a symmetric binary polymer blend this condition yields

-12-

$$\frac{\chi - \chi_c}{\chi_c} \gg \frac{N^2}{\rho^2 R^6} \sim \frac{1}{N}. \quad (\text{B3.6.14})$$

Ultimately, in the vicinity of the critical point, composition fluctuations are important, but the region in which these fluctuations dominate the behaviour decreases with the chain length N . Qualitatively, the behaviour can be understood as follows: long-chain molecules do not fill space and strongly interdigitate; the number of other chains in the volume of a reference chain increases like \sqrt{N} with chain length. This large number of interaction partners results in a strong suppression of fluctuations in the interactions on the level of a whole molecule and, hence, replacing the interactions by a non-fluctuating mean field is a good approximation.

Another important difference between the mean field treatment and the simulations or experiments are fluctuations of the local interfacial position. While the mean field treatment assumes a perfectly flat, planar interface right from the outset, the local interfacial position fluctuates in experiments and simulations. A typical snapshot of the local interface position, as obtained from a Monte Carlo simulation of a binary polymer blend is depicted in [figure B3.6.2](#). On not too small length scales the local position of the interface is smooth and without bubbles or overhangs. The system configuration can be described by two ingredients: the position $u(\mathbf{r}_{\parallel})$ of the centre of the interface as a function of the lateral coordinates \mathbf{r}_{\parallel} and the local structure described by profiles across the interface. The latter quantities depend only on the coordinate normal to the interface. In many applications the coupling between the long-wavelength fluctuations of the local interfacial position u and the intrinsic profile is neglected. In this case the intrinsic profiles describe the variation of quantities across an ideally planar interface. The apparent interfacial profile $p_{\text{app}}(z)$, which is averaged over fluctuations of the local interfacial position in experiments or simulations, can be approximated by a convolution of the intrinsic profile $p_{\text{int}}(z)$ and the distribution $P(u)$ of the local interface position [53]

$$p_{\text{app}}(z) = \int du P(u) p_{\text{int}}(z - u) \quad (\text{B3.6.15})$$

If one is only interested in the properties of the interface on scales much larger than the width of the intrinsic profiles, the interface can be approximated by an infinitely thin sheet and the properties of the intrinsic profiles can be cast into a few effective parameters. Using only the local position of the interface, effective interface Hamiltonians describe the statistical mechanics of fluctuating interfaces and membranes.

-13-

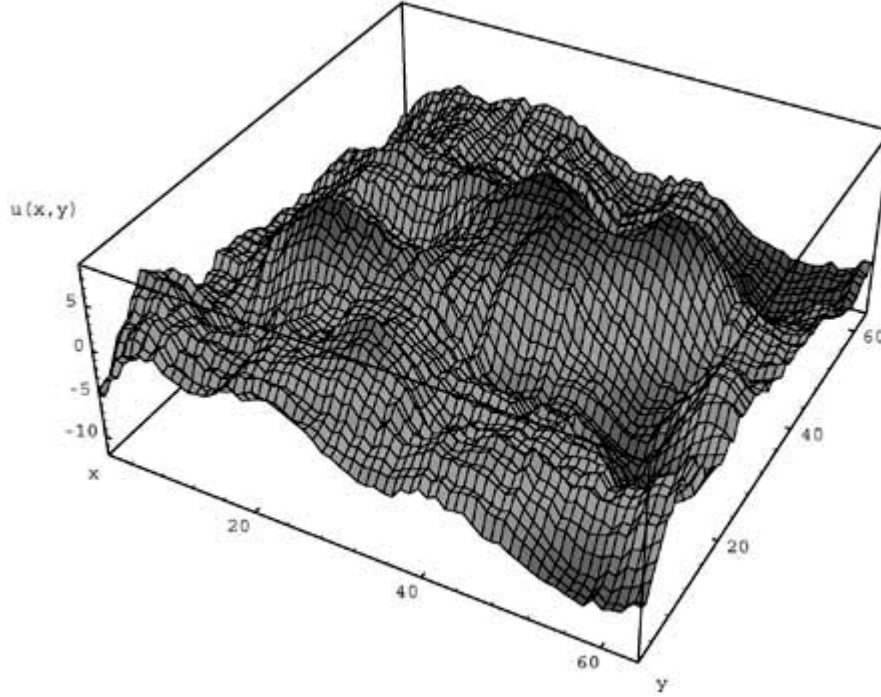


Figure B3.6.2. Local interface position in a binary polymer blend. After averaging the interfacial profile over small lateral patches, the interface can be described by a single-valued function $u(\mathbf{r}_{\parallel})$. (Monge representation). Thermal fluctuations of the local interface position are clearly visible. From Werner et al [49].

B3.6.2.4 EFFECTIVE INTERFACE HAMILTONIANS

The fluctuations of the local interfacial position increase the effective area. This increase in area is associated with an increase of free energy \mathcal{H} which is proportional to the interfacial tension γ . The free energy of a specific interface configuration $u(\mathbf{r}_{\parallel})$ can be described by the capillary wave Hamiltonian:

$$\mathcal{H}[u(\mathbf{r}_{\parallel})] = \gamma \int dx dy \left\{ \sqrt{1 + \left(\frac{du}{dx}\right)^2} \sqrt{1 + \left(\frac{du}{dy}\right)^2} - 1 \right\} \approx \frac{\gamma}{2} \int d^2 \mathbf{r}_{\parallel} (\nabla u)^2. \quad (\text{B3.6.16})$$

The functional $\mathcal{H}[u]$ can be diagonalized via a Fourier transformation with respect to the lateral coordinates \mathbf{r}_{\parallel} . This results in

$$\mathcal{H}[u_q] = \frac{\gamma}{2} \sum_q q^2 |u_q|^2. \quad (\text{B3.6.17})$$

In this Fourier representation the Hamiltonian is quadratic and the equipartition theorem yields for the thermal

fluctuations: $\langle u^2(\mathbf{q}) \rangle = k_B T / \gamma q^2$. This spectrum corresponds to a Gaussian distribution of the local interface position:

$$P(u) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{u^2}{2s^2}\right) \quad \text{with} \quad (B3.6.18)$$

$$s^2 = \frac{1}{4\pi^2} \int d^2 q_{\parallel} \langle u^2(q_{\parallel}) \rangle = \frac{kT}{2\pi\gamma} \ln\left(\frac{q_{\max}}{q_{\min}}\right)$$

where a short and long wave length scale cut-off q_{\max} and q_{\min} have to be introduced to avoid the divergence at $q \rightarrow \infty$ and $q \rightarrow 0$.

The interfacial fluctuations broaden laterally averaged profiles. Within the convolution approximation (B3.6.15) one obtains a profile with the shape of the erfc function [49]:

$$w_{\text{app}}^2 = w_{\text{int}}^2 + \frac{kT}{4\gamma} \ln\left(\frac{q_{\max}}{q_{\min}}\right). \quad (B3.6.19)$$

Thus, the apparent interfacial width w_{app} , which is measured in simulations or experiments, is larger than the intrinsic width w_{int} and depends via the wavevector cut-offs on the geometry considered. This can actually be used to measure the interfacial tension in computer simulations.

For a free interface the cut-off at large length scales is determined by the lateral patch size on which the interface is observed. In simulations this is set by the size of the simulation cell. In scattering experiments (e.g., neutron reflectivity) it is associated with the lateral coherence length of the beam. If the coexisting phases differ in density, gravitation will give rise to a large-scale cut-off for capillary waves [54]

$$q_{\max} = \sqrt{\frac{g\Delta\rho}{\gamma}} \quad (B3.6.20)$$

where $\Delta\rho$ is the density difference and g the gravitational constant. Similarly, interactions with boundaries (e.g. van der Waals forces) limit fluctuations and give rise to a large length scale cut-off. In this case the cut-off depends on the distance between the interface and the wall and the cut-off imparts a dependence of the apparent interfacial width on the distance between the wall and the interface.

On short length scales the coarse-grained description breaks down, because the fluctuations which build up the (smooth) intrinsic profile and the fluctuations of the local interface position are strongly coupled and cannot be distinguished. The effective interface Hamiltonian can describe the properties only on length scales large compared with the width $1/q_{\max} \sim w$ of the intrinsic profile. The absolute value of the cut-off is difficult to determine: the apparent profiles are experimentally accessible, but in order to use equation (B3.6.19) the width of an hypothetically flat interface without fluctuations has to be known. Polymer blends are suitable candidates for investigating this problem. Since the self-consistent field theory gives an accurate description of the interface profile except for

fluctuations, it yields a quantitative description of the intrinsic, ideally flat, profile. The comparison with

Monte Carlo simulations, which include fluctuations, then yields q_{max} . Simulations of a coarse-grained polymer blend by Werner *et al* find $q_{max} = 1.65/w_{int}$ [49] in the strong segregation limit, in rather good agreement with the value $q_{max} = 2/w_{int}$ suggested by analytical theory [55].

An important application of effective interface Hamiltonians are wetting phenomena. If a binary mixture is confined, the wall of the container will favour one component of the mixture, say A. This component forms an enrichment layer at the wall, while the B component is expelled from the wall region. Rather than describing the detailed composition profile at the wall, the effective interface Hamiltonian specifies the system configuration solely by the distance between the A-rich enrichment layer at the wall and the B component further away. This coarse graining concept is sketched in [figure B3.6.3](#). The profile is distorted in the vicinity of the wall and this gives rise to a short-range effective interaction between the wall and the interface. The length scale of the interaction is set by the characteristic length scale of the (free) interface profile. Dispersion forces give rise to an additional long-ranged effective interaction, which decays like a power-law with the distance l . The effective interfacial Hamiltonian takes the form:

$$\mathcal{H}[l(\mathbf{r}_1)] = \int d^2\mathbf{r}_1 \left\{ \frac{\gamma(l)}{2} (\nabla l)^2 + g(l) \right\} \quad (\text{B3.6.21})$$

where $g(l)$ denotes the effective interaction between the wall and a portion of the interface at a distance l . The first term corresponds to the capillary wave Hamiltonian. In general, the coefficient $\gamma(l)$ in front of the square gradient depends on the distance between the wall and the interface [56], because the intrinsic profile is distorted in the presence of the wall. Only for large distances $l \rightarrow \infty$ does the effective interfacial tension $\gamma(l)$ tend to its macroscopic value. The second term describes the effective interface potential between the wall and the interface. Depending on the shape of the interface potential $g(l)$ different situations are encountered [57]: if the effective potential exhibits a minimum in the vicinity of the wall, the interface is bound to the wall. This corresponds to a microscopically thin layer of the preferred component A at the wall. One says: A does not wet the wall. If $g(l)$ has a minimum at infinite distance $l \rightarrow \infty$ there is a macroscopically thick layer of A at the wall: component A wets the wall. The transition between both states is the wetting transition, which can be continuous (i.e. the thickness of the enrichment layer diverges upon approaching the transition temperature) or (most often) discontinuous (i.e. the thickness of the layer jumps from a microscopic value to a macroscopic one).

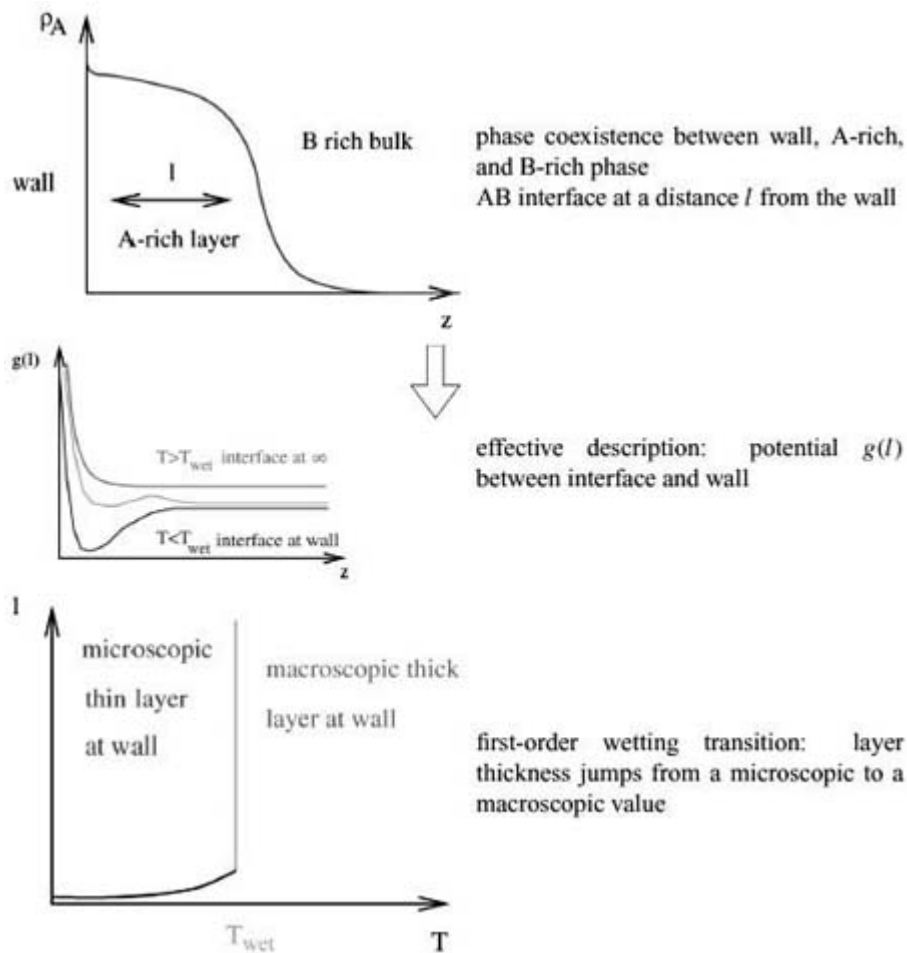


Figure B3.6.3. Sketch of the coarse-grained description of a binary blend in contact with a wall. (a) Composition profile at the wall. (b) Effective interaction $g(l)$ between the interface and the wall. The different potentials correspond to complete wetting, a first-order wetting transition and the non-wet state (from above to below). In case of a second-order transition there is no double-well structure close to the transition, but $g(l)$ exhibits a single minimum which moves to larger distances as the wetting transition temperature is approached from below. (c) Temperature dependence of the thickness l of the enrichment layer at the wall. The jump of the layer thickness indicates a first-order wetting transition. In the case of a continuous transition the layer thickness would diverge continuously upon approaching T_{wet} from below.

In the mean field considerations above, we have assumed a perfectly flat interface such that the first term in the Hamiltonian (B3.6.21) is ineffective. In fact, however, fluctuations of the local interface position are important, and its consequences have been studied extensively [57, 58].

B3.6.3 AMPHIPHILIC MODELS

Another important class of materials which can be successfully described by mesoscopic and continuum models are amphiphilic systems. Amphiphilic molecules consist of two distinct entities that like different environments. Lipid molecules, for instance, comprise a polar head that likes an aqueous environment and one or two hydrocarbon tails that are strongly hydrophobic. Since the two entities are chemically joined together they cannot separate into macroscopically large phases. If these amphiphiles are added to a binary mixture (say, water and oil) they greatly promote the dispersion of one component into the other. At low amphiphile

concentrations the molecules enrich at the interface so as to place their different ends into the corresponding phases. This displaces the water and the oil from the oil/water interface and greatly reduces the interfacial tension. At larger concentration of amphiphiles the molecules self-assemble into complex morphologies. These might either be isotropic (i.e. a microemulsion) or possess liquid crystalline order. The spatial structure is selected by a balance to minimize the contacts between the different entities and to fill space. Some of the possible morphologies are displayed in figure B3.6.4. Analogous morphologies are encountered in polymeric systems involving block copolymers.



Figure B3.6.4. Illustration of three structured phases in a mixture of amphiphile and water. (a) Lamellar phase: the hydrophilic heads shield the hydrophobic tails from the water by forming a bilayer. The amphiphilic heads of different bilayers face each other and are separated by a thin water layer. (b) Hexagonal phase: the amphiphiles assemble into a rod-like structure where the tails are shielded in the interior from the water and the heads are on the outside. The rods arrange on a hexagonal lattice. (c) Cubic phase: amphiphilic micelles with a hydrophobic centre order on a BCC lattice.

The relation between the architecture of the molecules and the spatial morphology into which they assemble has attracted longstanding interest because of their importance in daily life. Lipid molecules are important constituents of the cell membrane. Amphiphilic molecules are of major importance for technological applications (e.g., in detergents and the food industry).

The large length scale on which the self-assembly occurs and the universality of the morphologies borne out in experiments on a large variety of different systems make mesoscopic and continuum models suitable tools for investigating the underlying universal mechanism. Experiments suggest that many of the generic features can be captured by the amphiphilicity of the molecules. The models that have been employed can be broadly divided into models that aim at correlating the molecular architecture with the morphology and those models which investigate the generic phase behaviour and the influence of fluctuations.

B3.6.3.1 CHAIN MODELS

The architecture of the lipid molecules or the diblock copolymers results in the typical amphiphilic properties, like surface activity and self-assembly. On the most qualitative level, understanding of the self-assembly in lipid systems [3] is provided by a characterization of the molecules as a simple geometrical object (‘wedge’) parameterized by its volume, the maximum chain length and the area per head group. The different phases result from simple geometric packing considerations. Similar arguments on the balance between chain stretching and interfacial tension yield the qualitative features of the phase diagrams in systems containing diblock copolymers [59].

Chain models capture the basic elements of the amphiphilic behaviour by retaining details of the molecular architecture. Ben-Shaul *et al* [60] and others [61] explored the organization of the hydrophobic portion in lipid micelles and bilayers by retaining the conformational statistics of the hydrocarbon tail within the RIS (rotational isomeric state) model [4, 5] while representing the hydrophilic/hydrophobic interface merely by an

effective tension. By invoking a mean field approximation and calculating the properties of the tails by an enumeration of a large sample of conformations, they investigated the packing effects inside the hydrocarbon core for various detailed chain architectures. This mean field technique has been extended, for example, to include a modelling of the hydrophilic head and to study the self-assembly of lipids in aqueous solutions [62] or to investigate the absorption of proteins at surfaces covered with a polymer brush [63].

Many simulation approaches use a coarse-grained description of the amphiphiles by representing them via short-chain molecules on a lattice. The lattice is there only for computational convenience but is assumed to play no role otherwise. Typically, the number of lattice sites to model the amphiphiles is small and does not exceed 32. Each site is conceived as a small number of atomistic units along the amphiphilic molecule. A particularly popular model has been suggested by Larson [64]. There are two types of sites: hydrophilic and hydrophobic. Hydrophobic sites correspond to the oil or the hydrocarbon tail of the amphiphiles; hydrophilic sites represent the polar head of the amphiphiles or water. Oil and water are modelled as single-site entities. There is a short-range repulsion between unlike segments. The phase diagram of ternary oil/amphiphiles/water and binary amphiphile/water mixtures has been investigated by Monte Carlo simulations. Many phases observed in experiments (disordered, lamellar, hexagonal and even the gyroid phase) can be obtained as a function of temperature, composition and architecture of the amphiphile. Of course, special care has to be devoted to the study of finite-size effects. Typically, only a small number of unit cells of the spatially periodic structure fit into a simulation cell. If the size of the simulation box is close to a multiple of the unit cell size the stability of the phase might be greatly enhanced; if the size of the simulation box is incompatible with the spatially periodic structure the morphology is strongly distorted and its stability reduced. Very similar effects occur in nature if a spatially periodic structure is confined into a thin film.

A multitude of different variants of this model has been investigated using Monte Carlo simulations (see, for example [65]). The studies aim at correlating the phase behaviour with the molecular architecture and revealing the local structure of the aggregates. This type of model has also proven useful for studying rather complex structures (e.g., vesicles or pores in bilayers).

For structures with a high curvature (e.g., small micelles) or situations where orientational interactions become important (e.g., the gel phase of a membrane) lattice-based models might be inappropriate. Off-lattice models for amphiphiles, which are quite similar to their counterparts in polymeric systems, have been used to study the self-assembly into micelles [66], or to explore the phase behaviour of Langmuir monolayers [67] and bilayers. In those systems, various phases with a nematic ordering of the hydrophobic tails occur.

Since the amphiphilic nature is essential for the phase behaviour, systems of small molecules (e.g., lipid water mixtures) and polymeric systems (e.g., homopolymer copolymer blends) share many common features. Within the mean field approximation, the phase behaviour of block-copolymer models can conveniently be explored in the framework of the Gaussian chain model. The investigation of the self-assembly into various complex phases takes advantage of a Fourier decomposition of the spatially varying densities. The phase diagrams for pure diblock copolymers, binary blends of diblock copolymers and binary and ternary solutions have been investigated [44, 48]. These calculations reveal a rich variety of different morphologies as a function of the incompatibility, architecture and amount of homopolymer 'solvent'. In binary and ternary solutions, highly swollen phases are found in which the periodicity of the structure far exceeds the radius of gyration. An example of the possible phases in a ternary blend of two homopolymers and a symmetric diblock copolymer is presented in figure B3.6.5. At a fixed incompatibility one finds a complex phase diagram, including disordered homopolymer-rich phases, a symmetric lamellar phase L and asymmetric swollen lamellar phases L_A and L_B , which accommodate different amounts of the homopolymer components. Note that very similar phase diagrams are found in ternary oil water amphiphile mixtures. In the self-consistent field calculations not only the phase behaviour but also effective properties of the internal interfaces (e.g., the interfacial tension of bending moduli) are accessible [69]. The latter information might serve as input to

effective interface Hamiltonians.

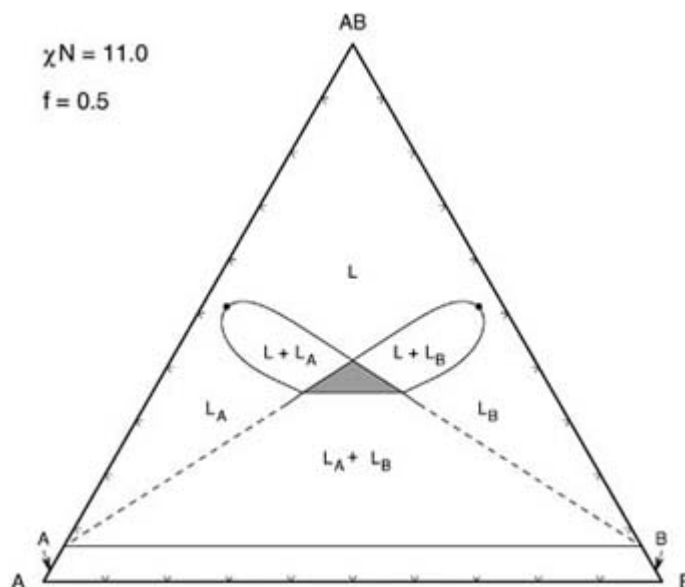


Figure B3.6.5. Phase diagram of a ternary polymer blend consisting of two homopolymers, A and B, and a symmetric AB diblock copolymer as calculated by self-consistent field theory. All species have the same chain length N and the figure displays a cut through the phase prism at $\chi N = 11$ (which corresponds to weak segregation). The phase diagram contains two homopolymer-rich phases A and B, a symmetric lamellar phase L and asymmetric lamellar phases, which are rich in the A component L_A or rich in the B component L_B , respectively. From Janert and Schick [68].

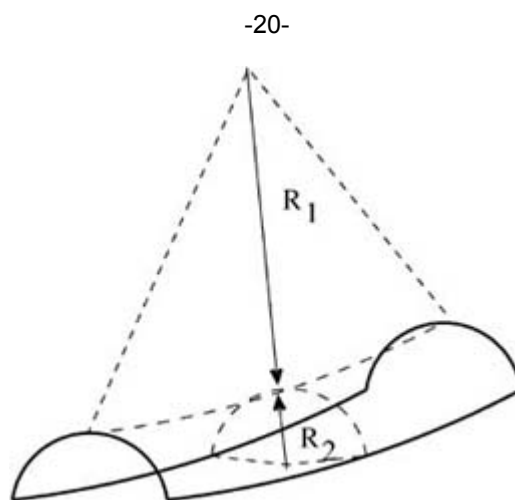


Figure B3.6.6. Illustration of the two principal radii of curvature for a membrane.

In general, the self-consistent field calculations are accurate for long polymers; for short chains however, fluctuations become important. Self-consistent field calculations and simulations of polymeric models show for example, that the bending rigidity of copolymer-laden interfaces might be quite small for short chain lengths. In a region where the self-consistent field theory predicts a highly swollen lamellar phase, the lamellar order is unstable with respect to interfacial fluctuations and a microemulsion forms [70]. Polymeric microemulsions have been observed in simulations [71] and experiments [72], but they are much more common in small molecular amphiphilic systems. Similarly, fluctuations easily destroy the body-centred cubic arrangement of micelles found in the self-consistent field theory and lead to formation of a micellar solution.

These chain models are well suited to investigate the dependence of the phase behaviour on the molecular architecture and to explore the local properties (e.g., enrichment of amphiphiles at interfaces, molecular conformations at interfaces). In order to investigate the effect of fluctuations on large length scales or the shapes of vesicles, more coarse-grained descriptions have to be explored.

B3.6.3.2 LATTICE MODELS

A further step in coarse graining is accomplished by representing the amphiphiles not as chain molecules but as single site/bond entities on a lattice. The characteristic architecture of the amphiphile—the hydrophilic head and hydrophobic tail—is lost in this representation. Instead, the interaction between the different lattice sites, which represent the oil, the water and the amphiphile, have to be carefully constructed in order to bring about the amphiphilic behaviour.

As early as 1969, Wheeler and Widom [73] formulated a simple lattice model to describe ternary mixtures. The bonds between lattice sites are conceived as particles. A bond between two positive spins corresponds to water, a bond between two negative spins corresponds to oil and a bond connecting opposite spins is identified with an amphiphile. The contact between hydrophilic and hydrophobic units is made infinitely repulsive; hence each lattice site is occupied by either hydrophilic or hydrophobic units. These two states of a site are described by a spin variable s_i , which can take the values +1 and -1. Obviously, oil/water interfaces are always completely covered by amphiphilic molecules. The Hamiltonian of this Widom model takes the form

$$\mathcal{H} = -h \sum_i s_i - J \sum_{\langle ij \rangle} s_i s_j - 2M \sum_{\langle\langle ij \rangle\rangle} s_i s_j - M \sum_{\langle\langle\langle ij \rangle\rangle\rangle} s_i s_j \quad (\text{B3.6.22})$$

where $\langle ij \rangle$, $\langle\langle ij \rangle\rangle$, and $\langle\langle\langle ij \rangle\rangle\rangle$ denote nearest, next-nearest and fourth-nearest neighbours on the lattice, respectively. The first two terms correspond to the Ising Hamiltonian. h acts as a chemical potential which favours positive spins, while J controls the incompatibility between water and oil. If only these two terms were present the model would describe a simple binary mixture. The additional terms have to be incorporated to bring about the amphiphilic properties. The case of negative M has been much investigated. In this case, the third term imparts some kind of bending rigidity to the oil/water interface, while the fourth term favours sequences of the form $(\dots + + - - + + - - \dots)$. This leads to the formation of lamellar phases which are not directly tied to the lattice spacing.

Slightly more complex models treat the water, the amphiphile and the oil as three distinct variables corresponding to the spin variables $S = +1, 0$, and -1 . The most general Hamiltonian with nearest-neighbour interactions has the form

$$\mathcal{H} = - \sum_{\langle ij \rangle} \{J S_i S_j + K S_i^2 S_j^2 + C(S_i^2 S_j + S_j^2 S_i)\} - \sum_i \{H S_i - \Delta S_i^2\}. \quad (\text{B3.6.23})$$

This Blume–Emery–Griffiths (BEG) model [74] has been studied both by mean field calculations as well as by simulations. There is no pronounced difference between the amphiphile molecules $S = 0$, the oil or the water. Indeed, the model was first suggested in a quite different context. An extension of the model by Schick and Shih [75] includes an additional interaction of the form

$$(\text{B3.6.24})$$

$$\Delta\mathcal{H} = -L \sum_{(ijk)} S_i(1 - S_j^2)S_k$$

where $(i j k)$ denotes three sites in a line. For negative values of L the term favours local conformations in which the amphiphile sits between the water and the oil. The model exhibits an oil-rich phase, a water-rich phase, a lamellar phase and a disordered phase, which exists between the lamellar phase and the oil–water coexistence. The disordered phase consists of water and oil domains separated by amphiphile sheets. It is homogeneous on large length scales, but shows—for certain parameter regions—oscillating structure on smaller length scales. These two length scales are associated with the structure of a microemulsion. The period of the oscillations characterize the local domain size in the microemulsion, while this nearly liquid-crystalline order ‘dephases’ on larger length scales. This defines the persistence length ξ . The latter length scale is mesoscopic (i.e. of the order of 100 Å), while the former is roughly the size of the molecules.

Lattice models have been studied in mean field approximation, by transfer matrix methods and Monte Carlo simulations. Much interest has focused on the occurrence of a microemulsion. Its location in the phase diagram between the oil-rich and the water-rich phases, its structure and its wetting properties have been explored [76]. Lattice models reproduce the reduction of the surface tension upon adsorption of the amphiphiles and the progression of phase equilibria upon increasing the amphiphile concentration. Spatially periodic (lamellar) phases are also describable by lattice models. However, the structure of the lattice can interfere with the properties of the periodic structures.

-22-

B3.6.3.3 CONTINUUM MODELS

An even coarser description is attempted in Ginzburg–Landau-type models. These continuum models describe the system configuration in terms of one or several, continuous order parameter fields. These fields are thought to describe the spatial variation of the composition. Similar to spin models, the amphiphilic properties are incorporated into the Hamiltonian by construction. The Hamiltonians are motivated by fundamental symmetry and stability criteria and offer a unified view on the general features of self-assembly. The universal, generic behaviour—the possible morphologies and effects of fluctuations, for instance—rather than the description of a specific material is the subject of these models.

An important example is the one-order-parameter model invented by Gompper and Schick [77], which describes a ternary mixture in terms of the density difference ϕ between water and oil:

$$\mathcal{F}[\phi(\mathbf{r})] = \int d\mathbf{r} \{ f(\phi) + g(\phi)|\nabla\phi|^2 + c|\Delta\phi|^2 \}. \quad (\text{B3.6.25})$$

The first two terms resemble the Ginzburg–Landau Hamiltonian for the polymeric systems. $f(\phi)$ describes the bulk free energy and there is a gradient square term to account for the free-energy costs of a spatially varying order parameter profile. In principle, the functions f , g and c of this Ginzburg–Landau expansion can be derived from a more microscopic model (e.g., the lattice models of the previous section). Such a derivation serves to relate the input parameters of the Ginzburg–Landau theories to microscopic parameters (e.g., length of the amphiphile) of the underlying model. The coefficients f , g and c are also related to the scattering intensity and, hence, some guidance from the experiment is available on how to choose them. Though the amphiphiles do not occur explicitly in the description, they determine the density dependence of the functions f and g . In order to model three-phase coexistence between an oil-rich phase, a water-rich phase and a microemulsion with roughly equal amounts of oil and water, the function f has to exhibit three minima. The amphiphiles decrease the free-energy cost of interfaces. This is modelled by a negative value of g in some

intermediate composition range. This favours the formation of interfaces and, therefore, the third term (with $c > 0$) is required to ensure thermodynamic stability.

By virtue of their simple structure, some properties of continuum models can be solved analytically in a mean field approximation. The phase behaviour interfacial properties and the wetting properties have been explored. The effect of fluctuations is investigated in Monte Carlo simulations as well as non-equilibrium phenomena (e.g., phase separation kinetics). Extensions of this one-order-parameter model are described in the review by Gompper and Schick [76]. A very interesting feature of these models is that effective quantities of the interface—like the interfacial tension and the bending moduli—can be expressed as a functional of the order parameter profiles across an interface [78]. These quantities can then be used as input for an even more coarse-grained description.

B3.6.3.4 RANDOM INTERFACE MODELS

Most characteristics of amphiphilic systems are associated with the alteration of the interfacial structure by the amphiphile. Addition of amphiphiles might reduce the free-energy costs by a dramatic factor (up to 10^{-2} dyn cm^{-1} in the oil/water/amphiphile mixture). Adding amphiphiles to a solution or a mixture often leads to the formation of a microemulsion or spatially ordered phases. In many aspects these systems can be conceived as an assembly of internal interfaces. The interfaces might separate oil and water in a ternary mixture or they might be amphiphilic bilayers in

-23-

binary solutions. Random interface models study the large-scale structure of amphiphilic systems by describing the configuration of the local, instantaneous interfacial position.

The effective free energy of the system of interfaces takes the general form [79, 80 and 81]

$$\mathcal{H} = \int dS \{ \gamma + \lambda_s H + 2\kappa H^2 + \bar{\kappa} K \} \quad (\text{B3.6.26})$$

where dS denotes the surface element, H the local mean curvature and K the local Gaussian curvature of the interface. The latter two quantities are related to the two principal radii of curvature via $2H = 1/R_1 + 1/R_2$ and $K = 1/R_1 R_2$. The interfacial tension γ controls the area of the interface. λ_s describes a spontaneous curvature of the interface, which is related to an asymmetry of the interface. This might occur even in a bilayer, when it is composed of an amphiphilic mixture and the two sheets have different compositions. The coefficients κ and $\bar{\kappa}$ characterize the bending rigidity and the saddle-splay modulus respectively. If the interface is closed, the Gauss–Bonnet theorem relates $\int dS K = 2\pi \chi_E$ to the Euler characteristic χ_E . Since this quantity is a topological invariant, the last term in the Hamiltonian can be omitted if the topology of the interface does not change (e.g., in the case of a vesicle).

One can regard the Hamiltonian (B3.6.26) above as a phenomenological expansion in terms of the two invariants K and H of the surface. To establish the connection to the effective interface Hamiltonian (b3.6.16) it is instructive to consider the limit of an almost flat interface. Then, the local interface position u can be expressed as a single-valued function of the two lateral parameters $u(\mathbf{r}_{\parallel})$. In this Monge representation the interface Hamiltonian can be written as

(B3.6.27)

$$\mathcal{H} = \int d^2r_{\parallel} \left\{ \frac{\gamma}{2} |\nabla u|^2 + \frac{\kappa}{2} |\Delta u|^2 \right\}.$$

Among the different problems which have been tackled with random interface Hamiltonians are the following. (i) The phase diagram of the random interface Hamiltonian has been explored by Huse and Leibler [82]. The phase diagram comprises a droplet phase, in which the minority component is dissolved into the matrix of the majority component, disordered phases and lamellar phases. (ii) Much interest has focused on the role of fluctuations. In the presence of a wall or another interface, the fluctuations of the local interface position are restricted. This gives rise to an entropic repulsion between the fluctuating interface and confining boundaries (Helfrich interaction [83]). (iii) In order to avoid the free-energy cost of a rim, membranes close up to form vesicles. The shapes of vesicles as a function of the bending rigidity and the pressure difference between the vesicle's interior and the outside have been mapped out [84].

B3.6.4 APPLICATIONS TO DYNAMIC PHENOMENA

Though this entry has focused on equilibrium properties, mesoscopic and continuum models in chemical physics can also describe non-equilibrium phenomena, and we shall mention some techniques briefly.

Mesoscopic models can often be treated by molecular dynamics simulations. This method generates a realistic

-24-

(Hamiltonian) trajectory in the phase space of the model from which information about the equilibrium dynamics can readily be extracted. The application to non-equilibrium phenomena (e.g., the kinetics of phase separation) is, in principle, straightforward.

Exploring the hydrodynamic behaviour of complex fluids with conventional molecular dynamics models poses a challenge to computational resources, because the hydrodynamic behaviour appears only on large time and length scales. Coarse-grained models [85, 86 and 87] have been explored in which a particle does not correspond to a molecule or a small number of atoms but rather to a fluid element. These 'fluid particles' [86] interact with an extremely soft potential, which does not diverge as two effective particles approach each other (see the Lennard-Jones potential (B3.6.1)) but increases only linearly with the interparticle distance. This very soft repulsive potential allows for very large time steps in a molecular dynamics simulation. As the particles correspond to coarse-grained fluid elements they do not conserve energy when they collide. This provides a motivation for a dissipative friction force and a random force. The strength of the friction and the noise are related by a fluctuation-dissipation theorem, which ensures that the equilibrium distribution corresponds to the canonical ensemble. Unlike the standard implementation of noise and friction forces in molecular dynamics schemes, noise and friction in the dissipative particle dynamics do not act on the velocity of a single particle but on pairs of particles. In this way momentum is conserved. The macroscopic behaviour is not diffusive but hydrodynamic (note, however, that the energy is not conserved and there is no transport equation of the energy). This method promises to be an efficient way to study dynamic effects on the mesoscopic scale of complex fluids. An application of dissipative particle dynamics to a binary homopolymer blend is described in [87].

Monte Carlo schemes generate a stochastic trajectory through phase space (see the entry about statistical mechanical simulations; B3.3). If the Monte Carlo moves resemble the configurational changes in a realistic dynamics (e.g., the conformations evolve via small displacements of particles) some dynamical information can be gained. Since there is no momentum in Monte Carlo simulations the dynamics is diffusive. However, many Monte Carlo algorithms employ moves that involve rather large changes in the system conformation (e.g., deletion of a molecule and subsequent insertion at a random position). These 'unphysical' moves are

extremely efficient in propagating the system through configuration space, but they do not allow for a dynamic interpretation of the trajectory.

A lattice scheme which does capture hydrodynamic behaviour is the lattice Boltzmann method [88, 89, 90 and 91]. This method has been devised as an effective numerical technique of computational fluid dynamics. The basic variables are the time-dependent probability distributions $f_\alpha(\mathbf{x}, t)$ of a velocity class α on a lattice site \mathbf{x} . This probability distribution is then updated in discrete time steps using a deterministic local rule. A careful choice of the lattice and the set of velocity vectors minimizes the effects of lattice anisotropy. This scheme has recently been applied to study the formation of lamellar phases in amphiphilic systems [92, 93].

Analytic techniques often use a time-dependent generalization of Landau–Ginzburg free-energy functionals. The different universal dynamic behaviours have been classified by Hohenberg and Halperin [94]. In the simple example of a binary fluid (model B) the concentration difference can be used as an order parameter m . A gradient in the local chemical potential $\mu(\mathbf{r}) = \delta F / \delta m(\mathbf{r})$ gives rise to a current \mathbf{j}

$$\mathbf{j} = -\Lambda \nabla \frac{\delta F}{\delta m(\mathbf{r})} \quad (\text{B3.6.28})$$

-25-

which strives to minimize the free energy. The kinetic coefficient Λ denotes a phenomenological constant which sets the time scale. In complex fluids (e.g., polymer blends) the relation between the gradient of the chemical potential and the current is non-local and the kinetic coefficient has to be generalized [95]. If the order parameter is conserved (e.g., in the demixing of a binary mixture) the change of the order parameter and the current are related by the continuity equation

$$\frac{\partial m(\mathbf{r}, t)}{\partial t} = \nabla \cdot \mathbf{j} = -\nabla \cdot \Lambda \nabla \frac{\delta F}{\delta m(\mathbf{r})}. \quad (\text{B3.6.29})$$

This time development of the order parameter is completely deterministic; when the equilibrium $\mu(\mathbf{r}) = \text{const}$ is reached the dynamics comes to rest. Noise can be added to capture the effect of thermal fluctuations. This leads to a Langevin dynamics for the order parameter.

Time-dependent Ginzburg–Landau models can be generalized to models with or without conserved order parameters. Also, the effect of additional conservation laws (for example, the inclusion hydrodynamic effects) has been explored. More complicated forms of the free-energy functional can be used to incorporate more details of the systems and alleviate the restriction to small order parameters inherent in the Ginzburg–Landau expansion. Shi and Noolandi [96] have used the free energy functional of the self-consistent field theory to explore fluctuations in spatially structured phases of diblock copolymers. A similar free-energy functional is employed by Fraaije and co-workers [97] to study the kinetics of self-assembly in amphiphilic systems. Extensions of the time-dependent Ginzburg–Landau equation to a formal scheme for the time evolution of non-equilibrium systems in terms of a set of coarse-grained variables have been explored [98].

REFERENCES

- [1] Landau L D and Lifshitz E M 1959 *Fluid Mechanics (Course of Theoretical Physics vol 6)* (Oxford: Pergamon)

- [2] Landau L D and Lifshitz E M 1970 *Theory of Elasticity (Course of Theoretical Physics vol 7)* (Oxford: Pergamon)
- [3] Israelachvili J 1991 *Intermolecular and Surface Forces* 2nd edn (New York: Academic)
- [4] Flory P J 1953 *Principles of Polymer Chemistry* (Ithaca, NY: Cornell University Press)
- [5] Flory P J 1969 *Statistical Mechanics of Chain Molecules* (New York: Wiley–Interscience)
- [6] Grosberg A Y and Khokhlov A R 1994 *Statistical Physics of Macromolecules (AIP Series in Polymers and Complex Materials)* (New York: AIP)
- [7] de Gennes P G 1979 *Scaling Concepts in Polymer Physics* (Ithaca, NY: Cornell University Press)
- [8] Freed K F 1987 *Renormalization Group Theory of Macromolecules* (New York: Wiley–Interscience)
- [9] des Cloizeaux J and Jannink G 1990 *Polymers in Solution: Their Modelling and Structure* (Oxford: Oxford Science Publications)
- [10] Schäfer L 1999 *Excluded Volume Effects in Polymer Solutions* (Berlin: Springer)
- [11] de Gennes P G 1972 Exponents for the excluded volume problem as derived by the Wilson method *Phys. Lett. A* **38** 339

-26-

- [12] Kremer K and Binder K 1988 Monte Carlo simulations of lattice models for macromolecules *Comp. Phys. Rep.* **7** 259
- [13] Sokal A D 1995 *Monte Carlo and Molecular Dynamics Simulations in Polymer Science* ed K Binder (New York: Oxford University Press) ch 3
- [14] Grassberger P 1997 Pruned-enriched Rosenbluth method: simulations of theta polymers of chain length up to 1,000,000 *Phys. Rev. E* **56** 3682
- [15] Frenkel D, Mooij G C A M and Smit B 1992 Novel scheme to study structural and thermal properties of continuously deformable molecules *J. Phys.: Condens. Matter* **4** 3053
- [16] Laso M, dePablo J J and Suter U W 1992 Simulation of phase equilibria for chain molecules *J Chem. Phys.* **97** 2817
- [17] Consta S, Wilding N B, Frenkel D and Alexandrowicz Z 1999 Recoil growth: an efficient simulation method for multi-polymer systems *J Chem. Phys.* **110** 3220
- [18] Carmesin I and Kremer K 1988 The bond fluctuation method—a new effective algorithm for the dynamics of polymers in all spatial dimensions *Macromolecules* **21** 2819
- [19] Doi M and Edwards S F 1986 *The Theory of Polymer Dynamics* (Oxford: Clarendon)
- [20] Deutsch J M and Cates M E 1986 Conjectures on the statistics of ring polymers *J. Physique* **47** 2121
- [21] Khokhlov A R and Nechaev S K 1985 Polymer chain in an array of obstacles *Phys. Lett. A* **112** 156
- [22] Müller M, Wittmer J P and Cates M E 1996 Topological effects in ring polymers: a computer simulation study *Phys. Rev. E* **53** 5063
- [23] Wertheim M S 1987 Thermodynamic perturbation theory of polymerization *J Chem. Phys.* **87** 7323
- [24] Yethiraj A and Woodward C E 1995 Monte Carlo density functional theory of nonuniform polymer melts *J Chem. Phys.* **102** 5499
- [25] Kierlik E and Rosinberg M L 1993 Perturbation density functional theory for polyatomic fluids III: application to hard chain molecules in slitlike pores *J Chem. Phys.* **100** 1716
- [26] Sen S, Cohen J M, McCoy J D and Curro J G 1994 The structure of a rotational isomeric state alkane melt near a hard wall *J Chem. Phys.* **101** 9010
- [27] Kremer K and Grest G S 1990 Dynamics of entangled linear polymer melts: a molecular-dynamics simulation *J Chem. Phys.* **92** 5057
- [28] Bishop M, Ceperley D, Frisch H L and Kalos M H 1980 Investigation of static properties of model bulk polymer fluids *J Chem. Phys.* **72** 3228
- [29] Gay J G and Berne B J 1981 Modification of the overlap potential to mimic a linear site–site potential *J Chem. Phys.* **74** 3316
- [30] Edwards S F 1966 The theory of polymer solutions at intermediate concentration *Proc. Phys. Soc.* **88** 265
- [31] Saito N, Takahashi K and Yunoli Y 1967 The statistical mechanical theory of stiff chains *J Phys. Soc. Japan* **22** 219
- [32] Flory P J 1941 *J Chem. Phys.* **9** 660

- [33] Huggins M L 1941 *J Chem. Phys.* **9** 440
- [34] Orwoll R A and Arnold P A 1996 Polymer–solvent interaction parameter χ *Physical Properties of Polymers, Handbook* ed J E Mark (Woodbury, NY: AIP) ch 14
- [35] Foreman K W and Freed K F 1998 Lattice cluster theory of multicomponent polymer systems: chain semiflexibility and specific interactions *Adv. Chem. Phys.* **103** 335
- [36] Schweizer K S and Curro J G 1997 Integral equation theories of the structure, thermodynamics and phase transitions of polymer fluids *Adv. Chem. Phys.* **98** 1
- [37] Müller M 1999 Miscibility behavior and single chain properties in polymer blends: a bond fluctuation model study *Macromol. Theory Simul.* **8** 343

-27-

- [38] Escobedo F A and de Pablo J J 1999 On the scaling of the critical solution temperature of binary polymer blends with chain length *Macromolecules* **32** 900
- [39] Taylor-Maranas J K, Debenedetti P G, Graessley W W and Kumar S K 1997 Compressibility effects in neutron scattering by polymer blends *Macromolecules* **30** 6943
- [40] Deutsch H-P and Binder K 1993 Mean-field to Ising crossover in the critical behavior of polymer mixtures—a finite size scaling analysis of Monte Carlo simulations *J. Physique II* **3** 1049
- [41] Helfand E and Tagami Y 1972 Theory of the interface between immiscible polymers *J. Polym. Sci. Polym. Lett.* **9** 741
Helfand E and Tagami Y 1972 *J. Chem. Phys.* **56** 3592
- [42] Hong K M and Noolandi J 1981 Theory of inhomogeneous multicomponent polymer systems *Macromolecules* **14** 727
- [43] Flerer G J, Cohen Stuart M A, Scheutjens J M H M, Cosgrove T and Vincent B 1993 *Polymers at Interfaces* (London: Chapman and Hall)
- [44] Matsen M W and Schick M 1994 Stable and unstable phases of a diblock copolymer melt *Phys. Rev. Lett.* **72** 2660
- [45] Broseta D, Fredrickson G H, Helfand E and Leibler L 1990 Molecular-weight effects and polydispersity effects at polymer–polymer interfaces *Macromolecules* **23** 132
- [46] Helfand E, Bhattacharjee S M and Fredrickson G H 1989 Molecular weight dependence of the polymer interfacial tension and concentration profile *J. Chem. Phys.* **91** 7200
- [47] Semenov A N 1996 Theory of long-range interactions in polymer systems *J. Physique II* **6** 1759
- [48] Cahn J W and Hilliard J E 1958 Free energy of a nonuniform system: I. Interfacial free energy *J. Chem. Phys.* **28** 258
- [49] Werner A, Schmid F, Müller M and Binder K 1997 Anomalous size-dependence of interfacial profiles between coexisting phases of polymer mixtures in thin film geometry: a Monte-Carlo study *J. Chem. Phys.* **107** 8175
- [50] Werner A, Schmid F, Müller M and Binder K 1999 Intrinsic profiles and capillary waves at homopolymer interfaces: a Monte Carlo Study *Phys. Rev. E* **59** 728
- [51] Lacasse M D, Grest G S and Levine A J 1998 Capillary–wave and chain length effects at polymer/polymer interfaces *Phys. Rev. Lett.* **80** 309
- [52] Ginzburg V L 1960 *Sov. Phys. Solid State* **1** 1824
deGennes P G 1977 Qualitative features of polymer demixtion *J. Physique Lett.* **38** 441
- [53] Jasnow D 1984 Critical phenomena at interfaces *Rep. Prog. Phys.* **47** 1059
- [54] Rowlinson J S and Widom B 1982 *Molecular Theory of Capillarity* (Oxford: Clarendon)
- [55] Semenov A N 1994 Scattering of statistical structure of polymer–polymer interfaces *Macromolecules* **27** 2732
- [56] Jin A J and Fisher M E 1993 Effective interface Hamiltonians for short-range critical wetting *Phys. Rev.* **47** 7365
- [57] Schick M 1990 Introduction to wetting phenomena *Les Houches Lectures: Liquid at interfaces* ed J Charvolin, J F Joanny and J Zinn-Justin (Amsterdam: Elsevier)
- [58] Dietrich S 1988 *Phase Transitions and Critical Phenomena* vol 12, ed C Domb and J Lebowitz (London: Academic)
- [59] Semenov A N 1985 Contribution to the theory of microphase layering in block-copolymer melts *Sov. Phys.–JETP* **61** 733
- [60] Ben-Shaul A, Szleifer I and Gelbart W M 1985 Chain organization and thermodynamics in micelles and bilayers: I. Theory *J. Chem. Phys.* **83** 3597

- [61] Gruen D W R 1984 A model for the chains in amphiphilic aggregates: I. Comparison with a molecular dynamics simulation of a bilayer *J. Phys. Chem.* **89** 645
- [62] Müller M and Schick M 1998 Calculation of the phase behavior of lipids *Phys. Rev. E* **57** 6973
-

-28-

- [63] Szleifer I and Carignano M A 1996 Tethered polymer layers *Adv. Chem. Phys.* **94** 165
- [64] Larson R G 1996 Monte Carlo simulations of the phase behavior of surfactant solutions *J. Physique II* **6** 1441
- [65] Liverpool T B and Bernardes A T 1995 Monte Carlo simulation of the formation of layered structures and membranes by amphiphiles *J. Physique II* **5** 1003
Liverpool T B and Bernardes A T 1995 Monte Carlo simulation of the formation of layered structures and membranes by amphiphiles *J. Physique II* **5** 1457
- [66] Smit B, Schlijper A G, Rupert L A M and van Os N M 1994 *J. Chem. Phys.* **94** 6933
Karaborni S, Esselink K, Hilbers P A J, Smit B, Karthäuser J, van Os N M and Zana R 1994 Simulating the self-assembly of Gemini (dimeric) surfactants *Science* **266** 254
- [67] Stadler C and Schmid F 1999 Phase behavior of grafted chain molecules: influence of head size and chain length *J. Chem. Phys.* **110** 9697
- [68] Janert P K and Schick M 1997 Phase behavior of ternary homopolymer/diblock blends: microphase unbinding in the symmetric system *Macromolecules* **30** 3916
- [69] Matsen M W 1999 Elastic properties of a diblock copolymer monolayer and their relevance to bicontinuous microemulsion *J. Chem. Phys.* **110** 4658
- [70] de Gennes P G and Taupin C 1982 Microemulsion and the flexibility of oil/water interfaces *J. Phys. Chem.* **86** 2294
- [71] Müller M and Schick M 1996 Bulk and interfacial thermodynamics of a symmetric, ternary homopolymer-copolymer mixture: a Monte Carlo study *J. Chem. Phys.* **105** 8885
- [72] Fredrickson G H and Bates F S 1997 Design of bicontinuous polymeric microemulsions *J. Polym. Sci. B* **35** 2775
- [73] Wheeler J C and Widom B 1968 *J. Am. Chem. Soc.* **90** 3064
- [74] Blume M, Emry V and Griffiths R B 1971 Ising model for the λ transition and phase separation in He³-He⁴ mixtures *Phys. Rev. A* **4** 1071
- [75] Schick M and Shih W-H 1987 Simple microscopic model of a microemulsion *Phys. Rev. Lett.* **59** 1205
- [76] Gompper G and Schick M 1994 Self-assembling amphiphilic systems *Phase Transitions and Critical Phenomena* vol 16, ed C Domb and J Lebowitz (New York: Academic)
- [77] Gompper G and Schick M 1990 Correlation between structural and interfacial properties in amphiphilic systems *Phys. Rev. Lett.* **65** 1116
- [78] Gompper G and Zschocke S 1991 Elastic properties of interfaces in a Ginzburg-Landau theory of swollen micelles, droplet crystals and lamellar phases *Euro. Phys. Lett.* **16** 731
- [79] Canham P B 1970 The minimum energy of bending as a possible explanation of the biconcave shape of the human red blood cell *J. Theoret. Biol.* **26** 61
- [80] Helfrich W 1973 Elastic properties of lipid bilayers: theory and possible experiments *Z. Naturf. c* **28** 693
- [81] Evans E A 1974 Bending resistance and chemically induced moments in membrane bilayers *Biophys. J.* **14** 923
- [82] Huse D A and Leibler S 1988 Phase behavior of an ensemble of nonintersecting random fluid films *J. Physique* **49** 605
- [83] Helfrich W 1977 Steric interaction of fluid membranes in multilayer systems *Z. Naturf. a* **33** 305
- [84] Gompper G and Kroll D M 1995 Phase diagram and scaling behavior of fluid vesicles *Phys. Rev. E* **51** 514
- [85] Hoogerbrugge P J and Koelman J M V A 1992 Simulating microscopic hydrodynamic phenomena with dissipative particle dynamics *Euro. Phys. Lett.* **19** 155
-

-29-

- [86] Espanol P and Warren P 1995 Statistical mechanics of dissipative particles dynamics *Euro. Phys. Lett.* **30** 191
Espanol P 1996 Dissipative particle dynamics for a harmonic chain: a first-principles derivation *Phys. Rev. B* **53** 1572
- [87] Groot R D and Warren P B 1997 Dissipative particles dynamics: bridging the gap between atomistic and mesoscopic simulation *J. Chem. Phys.* **107** 4423
- [88] McNamara G R and Zanetti G 1998 Use of the Boltzmann equation to simulate lattice–gas automata *Phys. Rev. Lett.* **61** 2332
- [89] Quian Y H, D’Humieres D and Lallemand P 1992 Lattice BGK models for Navier–Stokes equation *Euro. Phys. Lett.* **17** 479
- [90] Doolen G D (ed.) 1990 *Lattice Gas Methods for Partial Differential Equations* (Redwood City, CA: Addison-Wesley)
- [91] Doolen G D (ed.) 1991 *Lattice Gas Methods: Theory, Applications, and Hardware* (Cambridge, MA: MIT)
- [92] Gonnella G, Orlandini E and Yeomans J M 1997 Spinodal decomposition to a lamellar phase: effect of hydrodynamic flow *Phys. Rev. Lett.* **78** 1695
- [93] Theissen O, Gompper G and Kroll D M 1998 Lattice–Boltzmann model of amphiphilic systems *Euro. Phys. Lett.* **42** 419
- [94] Hohenberg P C and Halperin B I 1977 Theory of dynamic critical phenomena *Rev. Mod. Phys.* **49** 435
- [95] Binder K 1983 Collective diffusion, nucleation, and spinodal decomposition in polymer mixtures *J. Chem. Phys.* **79** 6387
- [96] Shi A C, Noolandi J and Desai R C 1996 Theory of anisotropic fluctuations in ordered block copolymer phases *Macromolecules* **29** 6487
- [97] Fraaije J G E M 1993 Dynamic density functional theory for micro-phase separation kinetics of block copolymer melts *J. Chem. Phys.* **99** 9202
Fraaije J G E M 1994 *J. Chem. Phys.* **100** 6984 (erratum)
- [98] Öttinger H C 1997 General projection operator formalism for the dynamics and thermodynamics of complex fluids *Phys. Rev. E* **57** 1416

FURTHER READING

- Grosberg A Y and Khokhlov A R 1995 *Statistical Physics of Macromolecules (AIP Series in Polymers and Complex Materials)* (New York: AIP)
- Schäfer L 1999 *Excluded Volume Effects in Polymer Solutions* (Berlin: Springer)
- Fleer G J, Cohen Stuart M A, Scheutjens J M H M, Cosgrove T and Vincent B 1993 *Polymers at Interfaces* (London: Chapman and Hall)
- Rowlinson J S and Widom B 1982 *Molecular Theory of Capillarity* (Oxford: Clarendon)
- Dietrich S 1988 *Phase Transitions and Critical Phenomena* vol 12, ed C Domb and J Lebowitz (London: Academic)
- Gompper G and Schick M 1994 Self-assembling amphiphilic systems *Phase Transitions and Critical Phenomena* vol 16, ed C Domb and J L Lebowitz (New York: Academic)

C1.1 Clusters

Lai-Sheng Wang

C1.1.1 CLUSTERS

C1.1.1.1 INTRODUCTION

As we divide and subdivide a piece of bulk crystal, its properties will not change dramatically until we reach the nanometre scale. As particle size approaches molecular dimensions, all properties of a material change. Particles consisting of a few to a few thousand atoms are called *clusters*. These particles often show unique electronic, magnetic and chemical properties with dramatic size and shape dependence. The field of cluster research involves the elucidation of these unique size-dependent properties and how they evolve from the molecular to the bulk as more and more atoms are added. A wide variety of clusters have been made and investigated, including metals, semiconductors, ionic solids, noble gases and small molecules. The discovery of C₆₀ and fullerenes [1], as a result of studying carbon clusters (see [chapter C1.2](#)), represents the best yield from cluster research. Curl, Kroto and Smalley were awarded the 1996 Nobel Prize in Chemistry for this remarkable discovery.

Clusters are intermediates bridging the properties of the atoms and the bulk. They can be viewed as novel molecules, but different from ordinary molecules, in that they can have various compositions and multiple shapes. Bare clusters are usually quite reactive and unstable against aggregation and have to be studied in vacuum or inert matrices. Interest in clusters comes from a wide range of fields. Clusters are used as models to investigate surface and bulk properties [2]. Since most catalysts are dispersed metal particles [3], isolated clusters provide ideal systems to understand catalytic mechanisms. The versatility of their shapes and compositions make clusters novel molecular systems to extend our concept of chemical bonding, structure and dynamics. Stable clusters or passivated clusters can be used as building blocks for new materials or new electronic devices [4] and this aspect has now led to a whole new direction of research into nanoparticles and quantum dots (see [chapter C2.17](#)). As the size of electronic devices approaches ever smaller dimensions [5], the new chemical and physical properties of clusters will be relevant to the future of the electronics industry.

Cluster research is a very interdisciplinary activity. Techniques and concepts from several other fields have been applied to clusters, such as atomic and condensed matter physics, chemistry, materials science, surface science and even nuclear physics. While the dividing line between clusters and nanoparticles is by no means well defined, typically, nanoparticles refer to species which are passivated and made in bulk form. In contrast, clusters refer to unstable species which are made and studied in the gas phase. Research into the latter is discussed in the current chapter.

C1.1.2 TECHNIQUES FOR CLUSTER GENERATION AND DETECTION IN THE GAS PHASE

C1.1.2.1 CLUSTER GENERATION IN THE GAS PHASE

The formation of clusters in the gas phase involves condensation of the vapour of the constituents, with the exception of the electrospray source [6], where ion-solvent clusters are produced directly from a liquid solution. For rare gas or molecular clusters, supersonic beams are used to initiate cluster formation. For nonvolatile materials, the vapours can be produced in one of several ways including laser vaporization, thermal evaporation and sputtering.

SUPERSONIC EXPANSION SOURCE

Supersonic expansion is an indispensable tool in modern chemical physics and physical chemistry [7]. It is an effective technique to produce weakly bonded clusters from gaseous species. Supersonic expansion of a gas sample through a small orifice cools the gas sample adiabatically to very low temperatures. Cluster growth is initiated through three-body collisions. A number of parameters (nozzle size, shape and backing pressure) can be varied to produce cold clusters and to tune cluster size distributions. Clusters with very low rotational and vibrational temperatures (a few kelvins) can be produced using the seeded beam technique, where a small amount of condensing gas is seeded in a helium beam to promote cluster formation and cooling. Clusters of rare gases and other small molecules are all produced and studied using the supersonic beam technique.

LASER VAPORIZATION SUPERSONIC CLUSTER SOURCE

Laser vaporization is one of the most popular and powerful techniques to produce metal and semiconductor clusters in the gas phase. [Figure C1.1.1](#) shows a schematic of a generic laser vaporization supersonic cluster beam source, first developed by Smalley and coworkers [8]. In this technique, an intense pulsed laser beam is focused onto a target. The rapid electronic to vibrational energy transfer allows the laser beam to heat the radiated spot to up to $\approx 10\,000$ K, producing a plasma with both neutral and charged atomic species. A pulsed high pressure carrier gas (usually helium) is delivered in coincidence with the laser pulse, and the rapid cooling due to the carrier gas initiates the cluster growth. The nascent clusters are entrained in the carrier gas and undergo a supersonic expansion to be further cooled. Both neutral and charged clusters can be produced. The laser vaporization technique is very versatile and it can produce clusters from any metal and semiconductor elements in the periodic table. Mixed clusters can be produced either by using an alloy target or adding a reactive gas in the carrier gas. A two-laser vaporization source has also been used to produce alloy clusters [9, 10].

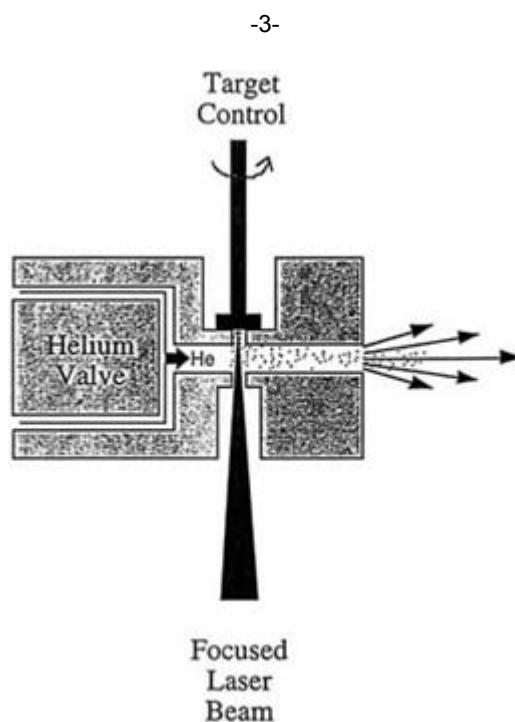


Figure C1.1.1. Schematic of a typical laser vaporization supersonic metal cluster source using a pulsed laser and a pulsed helium carrier gas.

THERMAL EVAPORATION SOURCE

The thermal evaporation source was the earliest used to produce metal clusters in the gas phase [11, 12 and 13], mostly for clusters of the alkalis and other low melting point materials. In this technique, a bulk sample is simply

heated in an oven to produce the atomic vapour. The vapour is entrained in a low-pressure gas flow where nucleation and cluster growth take place. Clusters of sodium atoms with more than 20 000 atoms have been made with this source [14]. A high-pressure carrier gas can also be used to produce a supersonic beam of clusters with the thermal evaporation source.

SPUTTERING SOURCE

Sputtering of a target surface with energetic particles can produce clusters. The energetic particles are typically ion beams of the rare gases or cesium. Clusters are lifted from the target surface by the energetic impact. Cluster sizes produced in this technique are generally limited and the cluster temperatures are high [15]. Intense continuous beams of cluster ions can be produced by sputtering and have been used for size-selected cluster deposition [16]. A related technique is a cold cathode discharge in a flowing rare gas. The discharge ionizes the rare gas, which then sputter metal atoms off the target (cathode). Cluster ions are formed through aggregation in the flowing gas stream. Continuous beams of small cluster ions can be effectively produced with this technique [17].

-4-

ELECTROSPRAY SOURCE

Electrospray was originally invented as a soft ionization technique for biological mass spectrometry [6, 18]. In electrospray, a liquid solution containing the molecules of interest is sprayed through a syringe needle under high voltage. Highly charged droplets produced in this fashion are broken down and desolvated to produce the ions of interest. Electrospray can also be used effectively to produce ion-solvent clusters and weakly bonded complexes [19]. Electrospray is a fairly new technique and is beginning to be used by physical chemists to produce novel gas phase clusters and complexes [20, 21].

C1.1.2.2 CLUSTER DETECTION IN THE GAS PHASE

The whole arsenal of physical chemistry methods has been utilized to investigate clusters. The development of cluster research is driven largely by new techniques to generate clusters and by new experimental tools to probe them. Mass spectrometry is the most useful tool in gas-phase cluster research because the first information one wants to know is the cluster's mass and size. All gas-phase investigations of clusters rely on mass spectrometry one way or the other. Since more stable clusters tend to be more abundant, a mass distribution of clusters contains valuable information about cluster stabilities, revealing 'magic numbers'—clusters with significantly higher abundance than their neighbours. Some of the most important discoveries of cluster science were based on the mass distribution of clusters, for example, the shell structure of free electron metal clusters [22] and C_{60} [1]. There are several ways to perform mass spectrometry, all using charged particles and measuring the mass/charge ratios of clusters. Mass separation of neutral clusters is still a challenging task [23]. The general assumption is that charged clusters also reflect the stability or distribution of the neutral clusters, although that is not always the case.

The most popular mass spectrometric technique is the 'time-of-flight' method [24], in which cluster mass information is obtained by measuring the flight times of a cluster ion beam in a given distance. The time-of-flight method is particularly suitable with pulsed laser vaporization cluster sources, and has high efficiency because the whole mass range is measured for a given laser shot. The time-of-flight technique has moderate mass resolution, but high resolution can be achieved by using a reflection [25, 26]. Ion cyclotron resonance mass spectrometry is another powerful technique used in cluster research [27]. Ions are confined and stored in a three-dimensional trap formed by a strong uniform magnetic field (\mathbf{B}) in the x - y directions and an electrostatic potential well in the z direction. The ion cyclotron frequency is $\omega_c = q\mathbf{B}/m$, so that highly accurate masses (m/q) can be obtained by measuring the cyclotron frequencies. Chemical reaction and fragmentation experiments are routinely performed with the stored ions [28, 29].

There are other techniques for mass separation such as the quadrupole mass filter and Wien filter. Another mass spectrometry technique is based on ion chromatography, which is also capable of measuring the shapes of clusters [30, 31]. In this method, cluster ions of a given mass are injected into a drift tube with well-defined entrance and exit slits and filled with an inert gas. The clusters drift through this tube under a weak electric potential. Since the

cluster mobility depends on collision cross sections with the inert gas, different isomers of a given cluster size are spatially separated in the drift tube. Structural information can be obtained for clusters whose isomers exhibit significantly different shapes, such as carbon and silicon clusters [32].

C1.1.3 METAL CLUSTERS

Metal clusters are bonded by strong covalent or metallic bonds. Clusters of the low melting point metallic elements are produced using the thermal evaporation technique. With the laser vaporization technique, metal clusters from all the metallic elements in the periodic table can be made. Simple metal clusters include those main group elements whose cluster properties are dominated by the delocalized nature of their valence electrons. In contrast to the simple metal clusters, transition-metal clusters are extremely complicated. Because of the unfilled d orbitals, transition-metal clusters possess a high density of electronic states. Transition-metal clusters possess both metallic and covalent bonding characters and exhibit interesting chemical, magnetic and electronic properties. Studies of transition-metal clusters are directly relevant to heterogeneous catalysis, surface science, metal cluster chemistry and metal–metal bonding in inorganic chemistry [33]. Although accurate theoretical descriptions of transition-metal clusters still pose a tremendous challenge, improved experimental and theoretical techniques are expected to make significant progress in the investigation of transition-metal clusters.

C1.1.3.1 SIMPLE METAL CLUSTERS AND THE ELECTRON SHELL MODEL

The simple metal clusters are among the earliest cluster species experimentally investigated [34]. They include those clusters from elements of the main groups IA–IIIA. Clusters of the IB and IIB elements with filled d-shells can also be categorized as simple metal clusters because their properties are largely dominated by the free electron nature of the valence s electrons. The relative ease of their formation through the thermal evaporation source and their relatively simple electronic structure made many experimental and theoretical investigations possible [34, 35 and 36]. In 1984, Knight and coworkers first observed from mass spectra of sodium clusters that clusters with 8, 20, 40, 58 and 92 atoms are more abundant than other clusters [22], as shown in [figure C1.1.2](#). They explained this observation in terms of a one-electron shell model [3], in which the valence electrons of the constituent atoms are completely delocalized within the volume of the cluster.

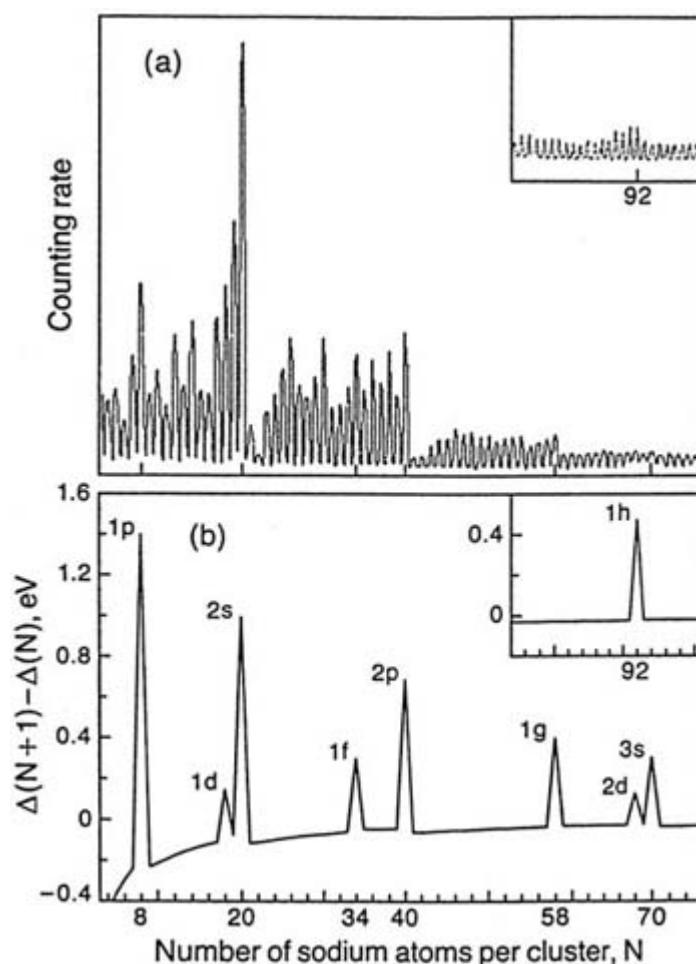


Figure C1.1.2. (a) Mass spectrum of sodium clusters (Na_N), $N = 4-75$. The inset corresponds to $N = 75-100$. Note the more abundant clusters at $N = 8, 20, 40, 58$, and 92 . (b) Calculated relative electronic stability, $\Delta(N+1) - \Delta(N)$ versus N using the spherical electron shell model. The closed shell orbitals are labelled, which correspond to the more abundant clusters observed in the mass spectrum. Knight W D, Clemenger K, de Heer W A, Saunders W A, Chou M Y and Cohen M L 1984 *Phys. Rev. Lett.* **52** 2141, figure 1.

In the shell model [37, 38], the jellium approximation is used to replace the positive cores with a uniform background potential. The valence electrons are treated as a quantized Fermi gas moving in the jellium potential and bounded by the cluster surface. The effective one-electron wavefunctions of a spherically symmetric potential are characterized by a main quantum number, n , and an angular momentum quantum l with degeneracy $2(2l+1)$, including spin. A closed shell system is obtained when all the levels for a given l are occupied with valence electrons. A closed shell system exhibits an enhanced stability because there exists a large energy gap to the next empty level (see figure C1.1.2(b)). The ordering of the levels depends on the shape of the potential. It turns out that the potential form for the metal clusters is analogous to that used in nuclear physics with similar shell structures. Therefore, the quantum numbers in

metal cluster physics follow the nuclear physics convention and there is no restriction on the angular quantum number l (unlike the filling of angular momentum shells in atomic physics underlying the periodic table of the elements). The nuclear shell model was developed in the 1940s to explain particularly stable nuclei. Thus concepts and ideas have been borrowed from nuclear physics to cluster physics and there is an interesting cross fertilization of two very different fields [39]. Besides the idea of shell closing, collective excitations [40, 41 and 42], fission [43, 44 and 45], and scattering [46]—concepts familiar in nuclear physics—have also been studied for the simple metal clusters.

The spherical shell model can only account for the major shell closings. For open shell clusters, ellipsoidal distortions occur [47], leading to subshell closings which account for the fine structures in [figure C1.1.2\(a\)](#). The electron shell model is one of the most successful models emerging from cluster physics. The electron shell effects are observed in many physical properties of the simple metal clusters, including their ionization potentials, electron affinities, polarizabilities and collective excitations [34].

C1.1.3.2 TRANSITION-METAL CLUSTERS: CHEMISTRY

The microscopic understanding of the chemical reactivity of surfaces is of fundamental interest in chemical physics and important for heterogeneous catalysis. Cluster science provides a new approach for the study of the microscopic mechanisms of surface chemical reactivity [48]. Surfaces of small clusters possess a very rich variation of chemisorption sites and are ideal models for bulk surfaces. Chemical reactivity of many transition-metal clusters has been investigated [49]. Transition-metal clusters are produced using laser vaporization, and the chemical reactivity studies are carried out typically in a flow tube reactor in which the clusters interact with a reactant gas at a given temperature and pressure for a fixed period of time. Reaction products are measured at various pressures or temperatures and reaction rates are derived. It has been found that the reactivity of small transition-metal clusters with simple molecules such as H₂ and NH₃ can vary dramatically with cluster size and structure [48, 49, 50, 51 and 52].

[Figure C1.1.3](#) shows a plot of the chemical reactivity of small Fe, Co and Ni clusters with H₂ as a function of size (full curves) [53]. The reactivity changes by several orders of magnitudes simply by changing the cluster size by one atom. Both geometrical and electronic arguments have been put forth to explain such reactivity changes. It is found that the reactivity correlates with the difference between the ionization potential (IP) and the electron affinity (EA) for a given cluster, corrected by a Coulomb energy, e^2/R , where R is the radius of the cluster ([figure C1.1.3](#)), dashed lines). This observation is interpreted using a model in which the probability of H₂ chemisorption is proportional to the magnitude of an entrance channel barrier caused by Pauli repulsion between H₂ and the cluster. This barrier is assumed to be proportional to the energy gap between the highest occupied and lowest unoccupied orbitals, characterized by the difference between the IP and EA of the cluster. [Figure C1.1.3](#) demonstrates the importance of the cluster electronic structures on the cluster reactivity.

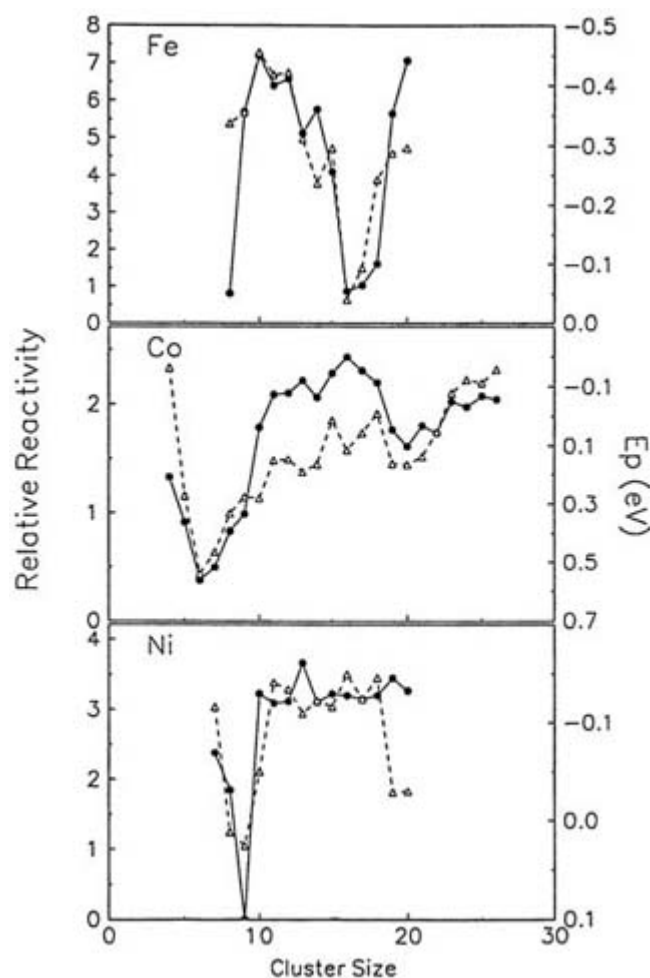


Figure C1.1.3. Relative reactivity of transition-metal clusters with H_2 (full curves, log scale) and the promotion energy $EP(N) = IP(N) - EA(N) - e^2/R(N)$, where IP, EA, and R represent the cluster ionization potential, electron affinity, and radius, respectively. The top figure is for Fe_N ($N = 8-20$), the middle figure is for Co_N ($N = 4-26$), and the lower figure is for Ni_N ($N = 7-20$). Conceicao J, Laaksonen R T, Wang L S, Guo T, Nordlander P and Smalley R E 1995 *Phys. Rev. B* **51** 4668, figure 3.

Cluster chemisorption experiments have been used extensively to probe the geometric structure of transition-metal clusters by Riley and coworkers [54, 55 and 56]. Since most of the atoms of a cluster are on the cluster surface, the available surface sites contain information about the underlying cluster structure. By measuring the maximum uptake of molecules, one can gain an insight into the cluster packing geometry. These studies have been mainly focused on clusters of Fe, Co, Ni and Cu and have found that icosahedral packing is the dominating structural feature of these clusters [54, 55 and 56]. There is ample evidence that multiple structural isomers exist for many of these clusters. For Cu clusters, evidence is provided for both electron shell behaviour and icosahedral geometrical structure [56].

The reactivity of size-selected transition-metal cluster ions has been studied with various types of mass spectrometric techniques [15]. Fourier-transform ion cyclotron resonance (FT-ICR) is a particularly powerful technique in which a cluster ion can be stored and cooled before experimentation. Thus, multiple reaction steps can be followed in FT-ICR, in addition to its high sensitivity and mass resolution. Many chemical reaction studies of transition-metal clusters with simple reactants and hydrocarbons have been carried out using FT-ICR [49, 57, 58]. A special reactive channel is cluster fragmentation induced either by photoabsorption or collisions with an inert gas. Measuring cluster fragmentation pathways and energetics provides information about both the cluster structures and bonding energies. Strong size-dependent bonding energies are found for small transition-metal cluster ions, and the bonding energy approaches the bulk cohesive energy smoothly for large clusters [59, 60].

C1.1.3.3 TRANSITION-METAL CLUSTERS: ELECTRONIC STRUCTURE

The diverse chemical and physical properties of the transition-metal clusters derive from their rich electronic structure. Thus probing the electronic structure of transition-metal clusters is of special interest. Transition metal dimers have been extensively studied by resonance two-photon ionization (R2PI) spectroscopy [61, 62, 63 and 64]. However, the high density of low-lying electronic states, characteristic of the transition-metal clusters, prevents the R2PI technique being used for larger clusters. Single photon photofragmentation spectroscopy of clusters bound with rare gas atoms has been used to probe the electronic structure of larger transition-metal clusters [65, 66]. Photoionization experiments have been used extensively to measure the IPs of transition-metal clusters [67, 68]. Recently, ZEKE (zero kinetic energy) spectroscopy has been applied to small neutral transition-metal clusters [69].

A more powerful experimental technique to probe the electronic structure of transition-metal clusters is size-selected anion photoelectron spectroscopy (PES) [70, 71, 72, 73, 74, 75 and 76]. In PES experiments, a size-selected anion cluster is photodetached by a fixed wavelength photon and the kinetic energies of the photoemitted electrons are measured. PES experiments provide direct measure of the electron affinity and electronic energy levels of neutral clusters. This technique has been used to study many types of clusters over a large cluster size range and can probe how the electronic structures of transition-metal clusters evolve from molecular to bulk [77, 78, 79, 80 and 81]. Research has focused on the 3d transition-metal clusters, for which there have also been many theoretical studies [82, 83, 84, 85, 86, 87, 88 and 89]. It is found that the electronic structure of the small transition-metal clusters is molecular in nature, with discrete electronic states. However, the electronic structure of the transition-metal clusters approaches that of the bulk rapidly. Figure C1.1.4 shows that the electronic structure of vanadium clusters with 65 atoms is already very similar to that of bulk vanadium [90]. Other 3d transition-metal clusters also show bulk-like electronic structures in similar size range [78].

-10-

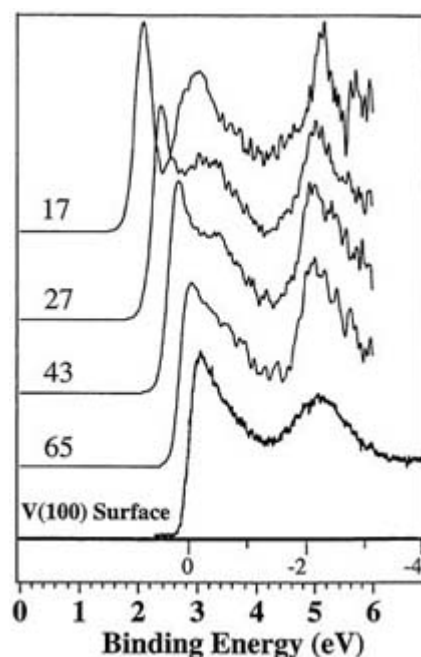


Figure C1.1.4. Photoelectron spectra of V_N^- ($N = 17, 27, 43,$ and 65) at 6.42 eV photon energy, compared to the bulk photoelectron spectrum of V(100) surface at 21.21 eV photon energy. The cluster spectra reveal the appearance of bulk features at V_{17} and how the cluster spectral features evolve toward the bulk. The bulk spectrum is referenced to the Fermi level. Wu H, Desai S R and Wang L S 1996 *Phys. Rev. Lett.* **77** 2436, figure 2.

C1.1.3.4 TRANSITION-METAL CLUSTERS: MAGNETISM

One of the interesting aspects of transition-metal clusters is their novel magnetic properties [91, 92, 93 and 94].

Although most transition-metal atoms have unpaired d-electrons and are magnetic, very few bulk transition-metal crystals are magnetic. Therefore, it is of great interest to understand how the magnetic properties of transition metals develop (diminish) as cluster size increases. The magnetic properties of transition-metal clusters have been investigated using the Stern–Gerlach molecular beam deflection method. Magnetic properties of clusters of the three bulk ferromagnetic materials, Fe, Co and Ni have been extensively studied [95, 96, 97, 98 and 99]. These clusters are found to be superparamagnetic with strong size-dependent magnetic moments. Figure C1.1.5 shows the measured magnetic moments of small Ni clusters as a function of size [99]. The dramatic size dependence of the cluster magnetic moments is interpreted to be due to a surface enhancement: the minima correspond to clusters with closed geometrical shells and maxima to clusters with relatively open structures. Small clusters generally possess much higher moments than the bulk materials, and the moments approach bulk values in the size range of about 500 atoms. Magnetism has also been detected in clusters of those elements whose bulk crystals are nonmagnetic [100].

-11-

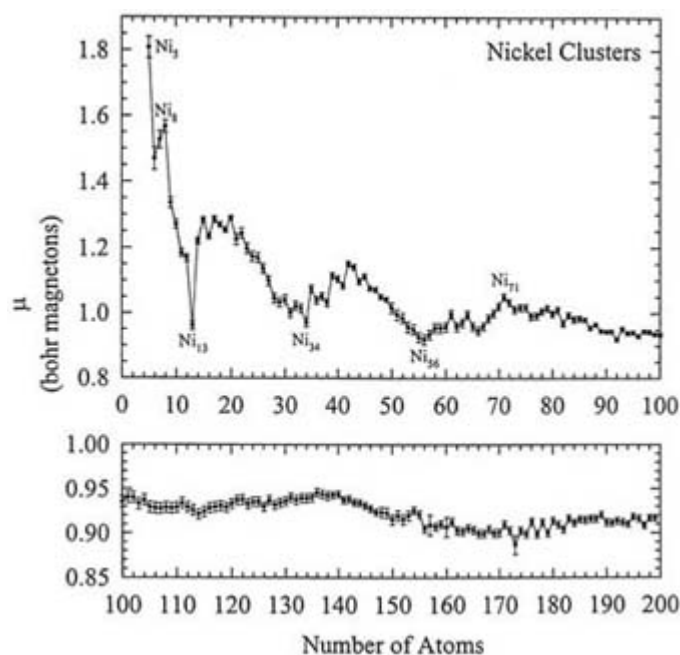


Figure C1.1.5. Nickel cluster magnetic moment per atom (μ) as a function of cluster size, at temperatures between 73 and 198 K. Apsel S E, Emmert J W, Deng J and Bloomfield L A 1996 *Phys. Rev. Lett.* **76** 1441, figure 1.

C1.1.4 SEMICONDUCTOR CLUSTERS

Since silicon is the most important semiconductor material, clusters of silicon have been most extensively studied, both theoretically and experimentally. The electronic structure [101, 102, 103 and 104], geometrical structure [105, 106, 107, 108, 109 and 110] and chemical reactivity [111] of silicon clusters have been investigated. The structures of small silicon clusters assume three-dimensional structures different from both that of the bulk crystal and that of its group IV neighbour, carbon. Ion mobility experiments have been very effective in providing experimental structural information for silicon clusters, and confirm that many structural isomers exist for silicon clusters because of their strong covalent bonding and relatively open structures [106, 110]. Ion mobility results show that silicon clusters up to ~27 atoms follow a prolate growth sequence, resulting in geometries with an aspect ratio of ~3 [106]. Larger clusters appear to assume more spherical geometries. The structures of medium-sized silicon clusters with 12–26 atoms have been studied recently by theoretical calculations using density functional theory in combination with ion mobility experiments [110]. Figure C1.1.6 shows the calculated structures of silicon clusters containing 12–20 atoms. The clusters with less than 18 atoms can be visualized as stacked Si_9 tricapped trigonal prisms, whereas global minima of Si_{19} and Si_{20} assume more spherical structures.

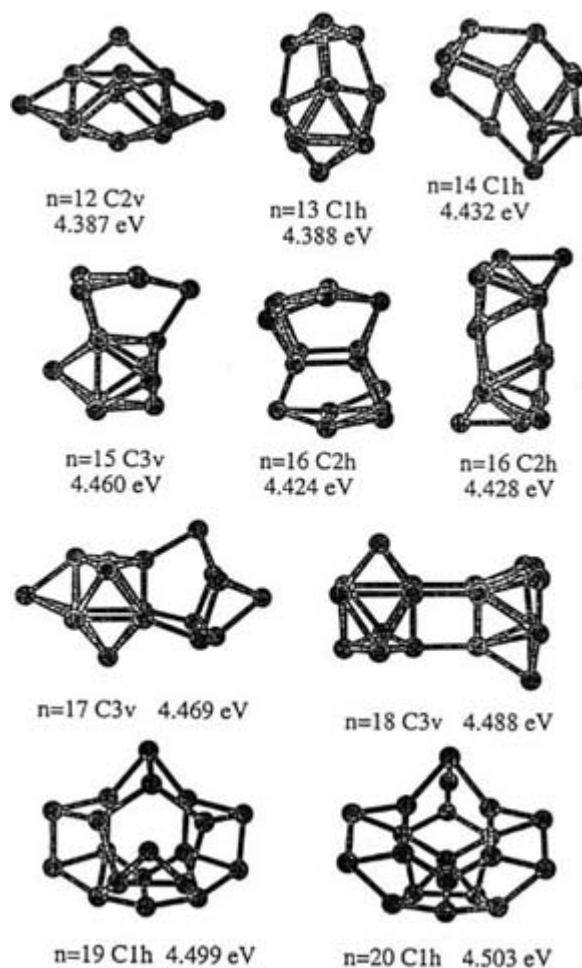


Figure C1.1.6. Minimum energy structures for neutral Si_n clusters ($n = 12\text{--}20$) calculated using density functional theory with the local density approximation. Cohesive energies per atom are indicated. Note the two nearly degenerate structures of Si_{16} . Ho K M, Shvartsburg A A, Pan B, Lu Z Y, Wang C Z, Wachter J G, Fye J L and Jarrold M F 1998 *Nature* **392** 582, figure 2.

Other semiconductor clusters have also been studied, such as germanium clusters [101, 112] and mixed clusters of the III–V semiconductors GaAs [113, 114 and 115] and InP [116, 117]. Of particular interest is the evolution and emergence of the energy band gap in these clusters. Infrared and visible absorption spectroscopy has been performed on indium phosphide clusters (In_xP_y) with $x + y$ up to 14 [116]. An optical-gap-like feature with an onset close to the band gap of bulk crystalline indium phosphide is already observed for the even clusters in this size range. This is surprising, because according to the model of quantum confinement these tiny clusters are expected to have band gaps much larger than that of the bulk crystal. Photoelectron spectroscopy experiments on size selected gallium arsenide cluster anions (Ga_xAs_y^-) showed that the electron affinity of the neutral clusters with $x + y$ around 50 already approached that of the bulk [113], quite different from the behaviour of metal clusters. Even though the electronic structure of

transition-metal clusters already approaches that of the bulk in a similar size regime, their electron affinity is still smaller than that of the bulk by more than 1 eV [77, 78, 79, 80 and 81]. These observations indicate the importance of charge localization in semiconductor clusters [118]. For bulk GaAs and InP, surface reconstruction creates shallow traps for conduction-band electrons. The wavefunction for a trapped electron is localized and is not subjected to the same quantum confinement effects as the delocalized orbital of a conduction-band electron. Thus,

quantum confinement effects are expected to be less important in clusters, where charge localization dominates, as should be the case for small clusters of GaAs and InP. The nature of charge localization suggests that such molecular-sized clusters may be ideal models for studying the surface behaviours of bulk semiconductors.

C1.1.5 IONIC CLUSTERS AND MIXED CLUSTERS

C1.1.5.1 IONIC CLUSTERS

Ionic clusters, such as alkali halide clusters, were among the earliest cluster species experimentally investigated [119]. The binding in ionic clusters is dominated by classical electrostatic effects, and simple interaction potentials can give fairly accurate descriptions of these clusters. One characteristic of the ionic clusters is that they mimic the structures of the bulk ionic crystals even at relatively small sizes, making ionic clusters attractive targets for both experimental and theoretical investigations. Nonstoichiometric alkaline halide clusters or clusters with excess electrons have been used as models to study bulk defects [120, 121 and 122]. Recent investigations have found facile structural transformation in alkaline halide clusters with rather low activation energies [123, 124].

Oxide clusters are another class of important ionic clusters because of the important roles that oxide materials play in both chemical catalysis and advanced materials applications. Oxide clusters of the main group I–III elements are dominated by the electrostatic interactions [125, 126 and 127]. Oxide clusters of the transition metals become more complicated with both ionic and covalent characters [128]. Oxide clusters of the late main group elements, such as silicon, are more dominated by covalent bonding. Oxide clusters are relatively less well characterized. Chemical reactivity of a number of transition-metal oxide clusters has been studied in a fast flow reactor with laser vaporization [129]. Antimony and bismuth oxide clusters have been recently produced, and magic number clusters characteristic of bulk compositions are observed [130, 131]. Photoelectron spectroscopy of size-selected anions has been carried out on a number of oxide cluster series [126, 132, 133 and 134]. The electronic structure evolution from that of a bare cluster to that of an oxide is monitored as the cluster is oxidized step-by-step by oxygen. [Figure C1.1.7](#) shows the structures of a series of Si_3O_y ($y = 1-6$) clusters [134], which can be viewed as a sequential oxidation of a Si_3 cluster. The local Si–O bonding structure from Si_3O to Si_3O_3 mimic that of the initial oxidation of a silicon surface, whereas the larger clusters (Si_3O_4 to Si_3O_6) with a SiO_4 unit begins to mimic that of bulk silicon oxide.

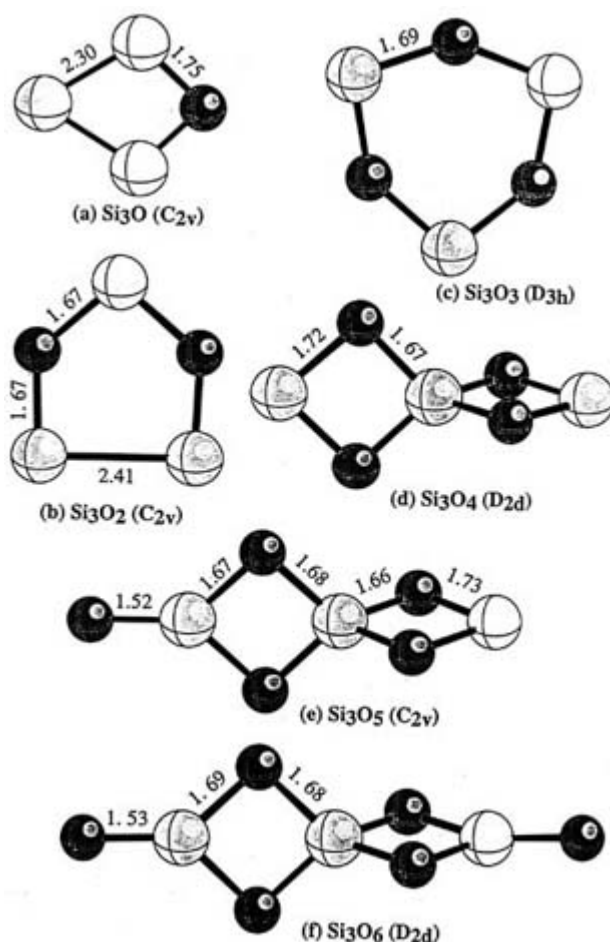


Figure C1.1.7. MP2/6-311 + G* optimized structures of the Si_3O_y ($y = 1-6$) clusters. All bond lengths are in Å. Note that for $y = 1-4$, all the O atoms are bridge bonded to two Si. Wang L S, Nicholas J B, Dupuis M, Wu H and Colson S D 1997 *Phys. Rev. Lett.* **78** 4450, figure 2.

C1.1.5.2 CARBIDE CLUSTERS

Metal-carbide clusters are relevant to the formation of both endohedral fullerenes and carbon nanotubes [135]. There also exists a class of apparently stable metal-carbide cluster ions, $\text{M}_8\text{C}_{12}^+$ ($\text{M} = \text{Ti}, \text{V}, \text{Cr}, \text{Zr}$ and Hf), called metallocarbohedrenes (met-car), first discovered by Castleman and coworkers [136]. The formation mechanisms of these novel clusters and nanostructures are still not elucidated [137]. Understanding the chemical bonding and structures of small metal-carbide clusters provides important insight into their growth mechanisms and can help design more efficient techniques for their bulk synthesis. For example, annealing of LaC_n^+ clusters in the gas phase converts them into endohedral fullerenes, and suggests that the La atom acts as a nucleation centre and the carbon rings arrange themselves around the La atom to form the final products [138]. Met-cars have not been isolated in bulk form and their structures have not been determined despite extensive experimental and theoretical investigations [137]. Many

questions still remain about the structure and bonding of metal-carbide clusters and the met-cars. Detailed characterization of the small carbide clusters will be key to understanding the formation of met-cars and endohedral fullerenes, as well as the catalytic effects in carbon nanotube growth. Mixed clusters in general provide interesting gas phase systems because not only their size, but also their composition, can be systematically varied. Clusters with tailored chemical or physical properties may be designed with the mixed clusters.

C1.1.6 RARE-GAS CLUSTERS AND OTHER WEAKLY BONDED MOLECULAR CLUSTERS

Rare-gas clusters can be produced easily using supersonic expansion. They are attractive to study theoretically because the interaction potentials are relatively simple and dominated by the van der Waals interactions. The Lennard–Jones pair potential describes the structures of the rare-gas clusters well and predicts magic clusters with icosahedral structures [139, 140]. The first five icosahedral clusters occur at 13, 55, 147, 309 and 561 atoms and are observed in experiments of Ar, Kr and Xe clusters [141]. Small helium clusters are difficult to produce because of the extremely weak interactions between helium atoms. Due to the large zero-point energy, bulk helium is a quantum fluid and does not solidify under standard pressure. Large helium clusters, which are liquid-like, have been produced and studied by Toennies and coworkers [142]. Recent experiments have provided evidence of superfluidity in ^4He clusters for as few as 60 atoms [143]. Helium clusters provide an ultracold environment, which can be used as a matrix to trap other molecules. Currently there is considerable interest in using large helium clusters as ‘nanocryostats’ to trap metal clusters and other molecular species for high-resolution spectroscopy investigations [144, 145].

Molecular clusters are weakly bound aggregates of stable molecules. Such clusters can be produced easily using supersonic expansion, and have been extensively studied by both electronic and vibrational spectroscopy [146, 147]. Hydrogen-bonded clusters are an important class of molecular clusters, among which small water clusters have received a considerable amount of attention [148, 149]. Solvated cluster ions have also been produced and studied [150, 151]. These solvated clusters provide ideal model systems to obtain microscopic information about solvation effect and its influence on chemical reactions.

C1.1.7 OUTLOOK

Gas-phase investigations of clusters over the past two decades have provided major advances of our fundamental understanding of these microscopic species and how their physical and chemical properties change with size. With new and continuously improved experimental and theoretical techniques, more discoveries and deeper understanding can be expected. Clusters provide the flexibility in size, shape, and composition in making new molecular species, which will enrich our concept of chemical structure and bonding. With new species that do not follow the classical valency, new chemical bonding theories and ideas will need to be developed to predict their physical and chemical properties. To understand the molecular details of the evolution from small clusters to nanocrystals will continue to be a challenge to both experimental and theoretical investigations. The diverse topics and systems afforded by clusters will make this field continue to be exciting and challenging for the new millennium to come.

REFERENCES

- [1] Kroto H W, Heath J R, O'Brien S C, Curl R F and Smalley R E 1985 C_{60} : Buckminsterfullerene *Nature* **318** 162
- [2] Muetterties E L, Rhodin T N, Band E, Brucher C F and Pretzer W R 1979 Clusters and surfaces *Chem. Rev.* **79** 91
- [3] Ertl G and Freund H J 1999 Catalysis and surface science *Phys. Today* **52** 41
- [4] Andres R P, Averback R S, Brown W L, Brus L E, Goddard W A III, Kaldor A, Louie S G, Moscovits M, Percy P S, Riley S J, Siegel R W, Spaepen F and Wang Y Research opportunity on clusters and cluster-assembled materials—a department of energy, council on materials science panel report *J. Mater. Res.* **4** 704
- [5] Keyes R W 1991 Limits and challenges in electronics *Contemp. Phys.* **32** 403

- [6] Yamashita M and Fenn J B 1984 Electrospray ion source. Another variation on the free-jet theme *J. Phys. C: Solid State Phys.* **88** 4451
- [7] Scoles G (ed) 1988 *Atomic and Molecular Beam Methods* vol 1 & 2 (Oxford: OUP)
- [8] Dietz T G, Duncan M A, Powers D E and Smalley R E 1981 Laser production of supersonic metal cluster beams *J. Chem. Phys.* **74** 6511
- [9] Hihara T, Pokrant S and Becker J A 1998 Magnetic moments and chemical bonding in isolated Bi_NCo_N clusters *Chem. Phys. Lett.* **294** 357
- [10] Nonose S, Sone Y, Onodera K and Kaya K 1990 Structure and reactivity of bimetallic Co_nV_m Clusters *J. Phys. C: Solid State Phys.* **94** 2744
- [11] Herrmann A, Leutwyler S, Schumacher E and Woste L 1978 On metal-atom clusters IV. Photoionization thresholds and multiphoton ionization spectra of alkali-metal molecules *Hel. Chim. Acta* **61** 453
- [12] Sattler K, Muhlbach J and Recknagel E 1980 Generation of metal clusters containing 2 to 500 atoms *Phys. Rev. Lett.* **45** 821
- [13] Martin T P and Schaber H 1985 Mass spectra of Si, Ge, and Sn clusters *J. Chem. Phys.* **83** 855
- [14] Martin T P, Bergmann T, Gohlich H and Lange T 1990 Observation of electronic shells and shells of atoms in large Na clusters *Chem. Phys. Lett.* **172** 209
- [15] Parent D C and Anderson S L 1992 Chemistry of metal and semimetal cluster ions *Chem. Rev.* **92** 1541
- [16] Fayet P, Granzer F, Hegenbart G, Moisar E, Pischel B and Woste L 1985 Latent-image generation by deposition of monodisperse silver clusters *Phys. Rev. Lett.* **55** 3002
- [17] Leopold D G, Ho J and Lineberger W C 1987 Photoelectron spectroscopy of mass-selected metal cluster anions. I. Cu_n^- , $n = 1-10$ *J. Chem. Phys.* **86** 1715

-17-

- [18] Fenn J B, Mann M, Meng C K, Wong S F and Whitehouse C M 1989 Electrospray ionization for mass spectrometry of large biomolecules *Science* **246** 64
- [19] Blades A T, Jayaweera P, Ikonomou M G and Kebarle P 1990 Studies of alkaline earth and transition metal M^{2+} gas phase ion chemistry *J. Chem. Phys.* **92** 5900
- [20] Spence T G, Trotter B T, Burns T D and Posey L A 1998 Metal-to-ligand charge transfer in the gas phase cluster limit *J. Phys. Chem. A* **102** 6101
- [21] Ding C F, Wang X B and Wang L S 1998 Photoelectron spectroscopy of doubly charged anions: intramolecular Coulomb repulsion and solvent stabilization *J. Phys. Chem. A* **102** 8633
- [22] Knight W D, Clemenger K, de Heer W A, Saunders W A, Chou M Y and Cohen M L 1984 Electron shell structure and abundances of sodium clusters *Phys. Rev. Lett.* **52** 2141
- [23] Buck U, Gu X, Lauenstein C and Rudolph A 1988 Infrared photodissociation spectra of size-selected $(\text{CH}_3\text{OH})_n$ clusters from $n = 2$ to 8 *J. Phys. Chem.* **92** 5561
- [24] Wiley W C and McLaren I H 1955 Time-of-flight mass spectrometer with improved resolution *Rev. Sci. Instrum.* **26** 1150
- [25] Karataev V I, Mamyryn B A and Shmikk D V 1972 New method for focusing ion bunches in time-of-flight mass spectrometers *Sov. Phys.-Tech. Phys.* **16** 1177
- [26] Bergmann T, Martin T P and Schaber H 1989 High-resolution time-of-flight mass spectrometer *Rev. Sci. Instrum.* **60** 792

- [27] Comisarow M B and Marshall A G 1974 Fourier transform ion cyclotron resonance spectroscopy *Chem. Phys. Lett.* **25** 282
- [28] Reents W D Jr, Mandich M L and Bondybey V E 1986 Reaction of anionic and cationic silicon clusters with tungsten hexafluoride studied by Fourier transform ion cyclotron resonance mass spectrometry *Chem. Phys. Lett.* **131** 12
- [29] Alford J M, Williams P E, Trevor D E and Smalley R E 1986 Metal cluster ion cyclotron resonance: combining supersonic metal cluster beam technology with FT-ICR *Int. J. Mass Spectrom. Ion Process.* **72** 33
- [30] von Helden G, Hsu M T, Kemper P R and Bowers M T 1991 Structures of carbon cluster ions from 3 to 60 atoms: linears to rings to fullerenes *J. Chem. Phys.* **95** 3835
- [31] Jarrold M F and Bower J E 1992 Mobilities of silicon cluster ions: the reactivity of silicon sausages and spheres *J. Chem. Phys.* **96** 9180
- [32] Jarrold M F 1995 Drift tube studies of atomic clusters *J. Phys. C: Solid State Phys.* **99** 11
- [33] Gonzalez-Moraga G 1993 *Cluster Chemistry* (Berlin: Springer)
- [34] de Heer W A 1993 The physics of simple metal clusters: experimental aspects and simple models *Rev. Mod. Phys.* **65** 611
- [35] Kappes M M 1988 Experimental studies of gas-phase main-group metal clusters *Chem. Rev.* **88** 369
- [36] Bonacic-Koutecky V, Fantucci P and Koutecky J 1991 Quantum chemistry of small clusters of elements of group Ia, Ib, and IIa: fundamental concepts, predictions, and interpretation of experiments *Chem. Rev.* **91** 1035
-

- [37] Cohen M L, Chou M Y, Knight W D and de Heer W A 1987 Physics of metal clusters *J. Phys. Chem.* **91** 3141
- [38] Ekardt W 1984 Work function of small metal particles: self-consistent spherical jellium-background model *Phys. Rev. B* **29** 1558
- [39] Schmidt R, Lutz H O and Dreizler R (eds) 1992 *Nuclear Physics Concepts in the Study of Atomic Cluster Physics* (Berlin: Springer)
- [40] Pollack S, Wang C R C and Kappes M M 1991 On the optical response of Na₂₀ and its relation to computational prediction *J. Chem. Phys.* **94** 2496
- [41] Tiggesbaumker J, Koller L, Lutz H O and Meiwes-Broer K H 1992 Giant resonances in silver-cluster photofragmentation *Chem. Phys. Lett.* **190** 42
- [42] Schlipper R, Kusche R, von Issendorff B and Haberland H 1998 Multiple excitation and lifetime of the sodium cluster plasmon resonance *Phys. Rev. Lett.* **80** 1194
- [43] Brechignac C, Cahuzac P, Kebaili N and Leygnier J 1998 Temperature effects in the Coulombic fission of strontium clusters *Phys. Rev. Lett.* **81** 4612
- [44] Yannouleas C and Landman U 1995 Barriers and deformation in fission of charged metal clusters *J. Phys. Chem.* **99** 14577
- [45] Vieira A and Fiolhais C 1998 Shell effects on fission barriers of metallic clusters: a systematic description *Phys. Rev. B* **57** 7352
- [46] Kresin V V, Tikhonov G, Kasperovich V, Wong K and Brockhaus P 1998 Long-range van der Waals forces between alkali clusters and atoms *J. Chem. Phys.* **108** 6660
- [47] Clemenger K 1985 Ellipsoidal shell structure in free-electron metal clusters *Phys. Rev. B* **32** 1359

- [48] Smalley R E 1985 Supersonic cluster beams: an alternative approach to surface science *Comparison of Ab Initio Quantum Chemistry with Experiment for Small Molecules* ed R J Bartlett (Boston: Reidel)
- [49] Knickelbein M B 1999 Reactions of metal cluster *Ann. Rev. Phys. Chem.* **50** 79
- [50] Richtsmeier S C, Parks E K, Liu K, Polo L G and Riley S J 1985 Gas phase reaction of iron clusters with hydrogen. I. Kinetics *J. Chem. Phys.* **82** 3659
- [51] Trevor D J, Whetten R L, Cox D M and Kaldor A 1985 Gas-phase platinum cluster reactions with benzene and several hexanes: evidence of extensive dehydrogenation and size-dependent chemisorption *J. Am. Chem. Soc.* **107** 518
- [52] Mitchell S A, Lian L, Rayner D M and Hackett P A 1995 Reaction of molybdenum clusters with molecular nitrogen *J. Chem. Phys.* **103** 5539
- [53] Conceicao J, Laaksonen R T, Wang L S, Guo T, Nordlander P and Smalley R E 1995 Photoelectron spectroscopy of transition metal clusters: correlation of valence electronic structure to reactivity *Phys. Rev. B* **51** 4668
- [54] Parks E K, Weiller B H, Bechthold P S, Hoffman W F, Nieman G C, Pobo L G and Riley S J 1988 Chemical probes of metal cluster structure: reactions of iron clusters with hydrogen, ammonia and water *J. Chem. Phys.* **88** 1622
- [55] Parks E K, Winter B J, Klots T D and Riley S J 1992 Evidence for polyicosahedral structure in ammoniated iron, cobalt and nickel clusters *J. Chem. Phys.* **96** 8267

-19-

- [56] Winter B J, Parks E K and Riley S J 1991 Copper clusters: the interplay between electronic and geometrical structure *J. Chem. Phys.* **94** 8618
- [57] Alford J M, Weiss F D, Laaksonen R T and Smalley R E 1986 Dissociative chemisorption of H₂ on niobium cluster ions. A supersonic cluster beam FT-ICR experiment *J. Phys. Chem.* **90** 4480
- [58] Berg C, Beyer M, Achatz U, Joos S, Niedner-Schatteburg G and Bondybey V 1998 Effect of charge upon metal cluster chemistry: reactions of Nb_n and Rh_n anions and cations with benzene *J. Chem. Phys.* **108** 5398
- [59] Lain L, Su C X and Armentrout P B 1992 Collision-induced dissociation of T_n⁺ (n = 2–22) with Xe: bond energies, geometric structures, and dissociation pathways *J. Chem. Phys.* **97** 4084
- [60] Su C X and Armentrout P B 1993 Collision-induced dissociation of C_n⁺ (n = 2–21) with Xe: bond energies, dissociation pathways, and structures *J. Chem. Phys.* **99** 6506
- [61] Morse M D 1986 Clusters of transition-metal atoms *Chem. Rev.* **86** 1049
- [62] James A M, Kowalczyk P, Langlois E, Campbell M D, Ogawa A and Simard B 1994 Resonant two photon ionization spectroscopy of the molecules V₂, VNb, and Nb₂ *J. Chem. Phys.* **101** 4485
- [63] Behm J M and Morse M D 1994 Spectroscopy of jet-cooled AlMn and trends in the electronic structure of the 3d transition metal aluminides *J. Chem. Phys.* **101** 6500
- [64] Arrington C A, Morse M D and Doverstal M 1995 Spectroscopy of mixed early–late transition metal diatomics: ScNi, YPd, and ZrCo *J. Chem. Phys.* **102** 1895
- [65] Knickelbein M B and Menezes W J C 1992 Optical response of small niobium clusters *Phys. Rev. Lett.* **69** 1046
- [66] Collings B A, Athanassenas K, Lacombe D, Rayner D M and Hackett P A 1994 Optical absorption spectra of Au₇, Au₉, Au₁₁ and Au₁₃, and their cations: gold clusters with 6–13 s-electrons *J. Chem. Phys.* **101** 3506
- [67] Yang S and Knickelbein M B 1990 Photoionization studies of transition metal clusters: ionization potentials for Fe_n

and Co_n *J. Chem. Phys.* **93** 1533

- [68] Koretsky G M and Knickelbein M B 1997 Photoionization studies of manganese clusters: ionization potentials for Mn_7 to Mn_{64} *J. Chem. Phys.* **106** 9810
- [69] Yang D S, Zgierski M Z, Rayner D M, Hackett P A, Martine A, Salahub D R, Roy P N and Carrington T Jr 1995 The structure of Nb_3O and Nb_3O^+ determined by pulsed field ionization-zero electron kinetic energy photoelectron spectroscopy and density functional theory *J. Chem. Phys.* **103** 5335
- [70] Leopold D G and Lineberger W C 1986 A study of the low-lying electronic states of Fe_2 and Co_2 by negative ion photoelectron spectroscopy *J. Chem. Phys.* **81** 51
- [71] McHugh K M, Eaton J G, Lee G H, Sarkas H W, Kidder L H, Snodgrass J T, Manaa M R and Bowen K H 1989 Photoelectron spectra of the alkali metal cluster anions: $\text{Na}_n^-(n = 2-5)$, $\text{K}_n^-(n=2-7)$, $\text{Rb}_n^-(n=2, 3)$ and $\text{Cs}_n^-(n=2, 3)$ *J. Chem. Phys.* **91** 3792

-20-

- [72] Cheshnovsky O, Yang S H, Pettiette C L, Craycraft M J and Smalley R E 1987 Magnetic time-of-flight photoelectron spectrometer for mass-selected negative cluster ions *Rev. Sci. Instrum.* **58** 2131
- [73] Gantefor G, Meiwes-Broer K H and Lutz H O 1988 Photodetachment spectroscopy of cold aluminum cluster anions *Phys. Rev. A* **37** 2716
- [74] Kitsopoulos T N, Chick C J, Zhao Y and Neumark D M 1991 Study of the low-lying electronic states of Si_2 and Si_2^- using negative ion photodetachment techniques *J. Chem. Phys.* **95** 1441
- [75] Handschuh H, Gantefor G and Eberhardt W 1995 Vibrational spectroscopy of clusters using a magnetic bottle electron spectrometer *Rev. Sci. Instrum.* **66** 3838
- [76] Wang L S, Cheng H S and Fan J 1995 Photoelectron spectroscopy of size-selected transition metal clusters: Fe_n^- , $n = 3-24$ *J. Chem. Phys.* **102** 9480
- [77] Cheshnovsky O, Taylor K J, Conceicao J and Smalley R E 1990 Ultraviolet photoelectron spectra of mass-selected copper clusters: evolution of the 3d band *Phys. Rev. Lett.* **64** 1785
- [78] Wang L S and Wu H 1998 Probing the electronic structure of transition metal clusters from molecular to bulk-like using photoelectron spectroscopy *Cluster Materials, Advances in Metal and Semiconductor Clusters* vol 4, ed M A Duncan (Greenwich: JAI Press) p 299
- [79] Gantefor G and Eberhardt W 1996 Localization of 3d and 4d electrons in small clusters: the 'roots' of magnetism *Phys. Rev. Lett.* **76** 4975
- [80] Wu H, Desai S R and Wang L S 1996 Electronic structure of small titanium clusters: emergence and evolution of the 3d band *Phys. Rev. Lett.* **76** 212
- [81] Iseda M, Nishio T, Han S Y, Yoshida H, Terasaki A and Kondow T 1997 Electronic structure of vanadium cluster anions as studied by photoelectron spectroscopy *J. Chem. Phys.* **106** 2182
- [82] Castro M, Jamorski C and Salahub D R 1997 Structure, bonding, and magnetism of small Fe_n , Co_n and Ni_n clusters, $n=5$ *Chem. Phys. Lett.* **271** 133
- [83] Massobrio C, Pasquarello A and Corso A D 1998 Structural and electronic properties of small Cu_n clusters using generalized-gradient approximations within density functional theory *J. Chem. Phys.* **109** 6626

- [84] Gronbeck H and Rosen A 1997 Geometric and electronic properties of small vanadium clusters: a density functional study *J. Chem. Phys.* **107** 10 620
- [85] Lee K, Callaway J and Dhar S 1984 Electronic structure of small iron clusters *Phys. Rev. B* **30** 1724
- [86] Pastor G M, Dorantes-Davila J and Bennemann K H 1989 Size and structural dependence of the magnetic properties of small 3d-transition metal clusters *Phys. Rev. B* **40** 7642
- [87] Cheng H S and Wang L S 1996 Dimer growth, structure transition and antiferromagnetic ordering in small chromium clusters *Phys. Rev. Lett.* **77** 51
- [88] Nayak S K, Khanna S N, Rao B K and Jena P 1997 Physics of nickel clusters: energetics and equilibrium geometries *J. Phys. Chem. A* **101** 1072
-

-21-

- [89] Wetzel T L and DePristo A E 1996 Structures and energetics of Ni₂₄-Ni₅₅ clusters *J. Chem. Phys.* **105** 572
- [90] Wu H, Desai S R and Wang L S 1996 Evolution of the electronic structure of small vanadium clusters from molecular to bulk-like *Phys. Rev. Lett.* **77** 2436
- [91] Shi J, Gider S, Babcock K and Awschalom D D 1996 Magnetic clusters in molecular beams, metals, and semiconductors *Science* **271** 937
- [92] Liu F, Khanna S N and Jena P 1991 Magnetism in small vanadium clusters *Phys. Rev. B* **43** 8179
- [93] Pastor G M, Dorantes-Davila J, Pick S and Dreysse H 1995 Magnetic anisotropy of 3d transition-metal clusters *Phys. Rev. Lett.* **75** 326
- [94] Viitala E, Merikoski J, Manninen M and Timonen J 1997 Antiferromagnetic order and frustration in small clusters *Phys. Rev. B* **55** 11 541
- [95] de Heer W A, Milani P and Chatelain A 1990 Spin relaxation in small free iron clusters *Phys. Rev. Lett.* **65** 488
- [96] Bucher J P, Douglass D C and Bloomfield L A 1991 Magnetic properties of free cobalt clusters *Phys. Rev. Lett.* **66** 3052
- [97] Billas I M L, Becker J A, Chatelain A and de Heer W A 1993 Magnetic moments of iron clusters with 25 to 700 atoms and their dependence on temperature *Phys. Rev. Lett.* **71** 4067
- [98] Billas I M L, Chatelain A and de Heer W A 1994 Magnetism from the atom to the bulk in iron, cobalt, and nickel clusters *Science* **265** 1682
- [99] Apsel S E, Emmert J W, Deng J and Bloomfield L A 1996 Surface-enhanced magnetism in nickel clusters *Phys. Rev. Lett.* **76** 1441
- [100] Cox A J, Louderback J G and Bloomfield L A 1993 Experimental observation of magnetism in rhodium clusters *Phys. Rev. Lett.* **71** 923
- [101] Cheshnovsky O, Yang S H, Pettiette C L, Craycraft M J, Liu Y and Smalley R E 1987 Ultraviolet photoelectron spectroscopy of semiconductor clusters: silicon and germanium *Chem. Phys. Lett.* **138** 119
- [102] Ogut S, Chelikowsky J R and Louie S G 1997 Quantum confinement and optical gaps in Si nanocrystals *Phys. Rev. Lett.* **79** 1770
- [103] Patterson C H and Messmer R P 1990 Bonding and structure in silicon clusters: a valence-bond interpretation *Phys. Rev. B* **42** 7530
- [104] Grossman J C and Mitas L 1995 Quantum Monte Carlo determination of electronic and structural properties of Si_n clusters ($n \leq 20$) *Phys. Rev. Lett.* **74** 1323

- [105] Bloomfield L A, Freeman R R and Brown W L 1985 Photofragmentation of mass-resolved $\text{Si}_n^+_{1-12}$ clusters *Phys. Rev. Lett.* **54** 2246
- [106] Jarrold M F and Constant V A 1991 Silicon cluster ions: evidence for a structural transition *Phys. Rev. Lett.* **67** 2994
- [107] Sieck A, Porezag D, Frauenheim T, Pederson M R and Jackson K 1997 Structure and vibrational spectra of low-energy silicon clusters *Phys. Rev. A* **56** 4890
-

-22-

- [108] Rohlfing C M and Raghavachari K 1990 A theoretical study of small silicon clusters using an effective core potential *Chem. Phys. Lett.* **167** 559
- [109] Rothlisberger U, Andreoni W and Parrinello M 1994 Structure of nanoscale silicon clusters *Phys. Rev. Lett.* **72** 665
- [110] Ho K M, Shvartsburg A A, Pan B, Lu Z Y, Wang C Z, Wacker J G, Fye J L and Jarrold M F 1998 Structures of medium-sized silicon clusters *Nature* **392** 582
- [111] Jarrold M F 1991 Nanosurface chemistry on size-selected silicon clusters *Science* **252** 1085
- [112] Hunter J M, Fye J L, Jarrold M F and Bower J E 1994 Structural transitions in size-selected germanium cluster ions *Phys. Rev. Lett.* **73** 2063
- [113] Jin C, Taylor K J, Conceicao J and Smalley R E 1990 Ultraviolet photoelectron spectra of gallium arsenide clusters *Chem. Phys. Lett.* **175** 17
- [114] Lou L, Nordlander P and Smalley R E 1992 Electronic structure of small GaAs clusters. II *J. Chem. Phys.* **97** 1858
- [115] Schlecht S, Schafer R, Woenckhaus J and Becker J A 1995 Electric dipole polarizabilities of isolated gallium arsenide clusters *Chem. Phys. Lett.* **246** 315
- [116] Rinnen K D, Kolenbrander K D, DeSantolo A M and Mandich M L 1992 Direct infrared and visible absorption spectroscopy of stoichiometric and nonstoichiometric clusters of indium phosphide *J. Chem. Phys.* **96** 4088
- [117] Xu C, de Beer E, Arnold D W, Arnold C C and Neumark D M 1994 Anion photoelectron spectroscopy of small indium phosphide clusters (In_xP_y^- ; $x, y = 1-4$) *J. Chem. Phys.* **101** 5406
- [118] Gratzel M 1991 All surface and no bulk *Nature* **349** 740
- [119] Martin T P 1983 Alkaline halide clusters and microcrystals *Phys. Rep.* **95** 167
- [120] Honea E C, Homer M L, Labastie P and Whetten R L 1989 Localization of an excess electron in sodium halide clusters *Phys. Rev. Lett.* **63** 394
- [121] Hakkinen H, Barnett R N and Landman U 1995 Energetics, structure, and excess electrons in small sodium-chloride clusters *Chem. Phys. Lett.* **232** 79
- [122] Yu N, Xia P, Bloomfield L A and Fowler M 1995 Structure and electron localization of anionic NaCl clusters with excess electrons *J. Chem. Phys.* **102** 4965
- [123] Hudgins R R, Dugourd P, Tenenbaum J M and Jarrold M F 1997 Structural transitions in sodium chloride nanocrystals *Phys. Rev. Lett.* **78** 4213
- [124] Fatemi F K, Fatemi D J and Bloomfield L A 1996 Thermal isomerization in isolated cesium-halide clusters *Phys. Rev. Lett.* **77** 4895
- [125] Boutou V, Lebeault M A, Allouche A R, Bordas C, Paulig F, Viallon J and Chevaleyre J 1998 Structural transition in barium suboxide clusters *Phys. Rev. Lett.* **80** 2817

- [126] Wu H, Li X, Wang X B, Ding C F and Wang L S 1998 Al_3O_x ($x = 0-5$) clusters: sequential oxidation, metal-to-oxide transformation, and photoisomerization *J. Chem. Phys.* **109** 449
-

-23-

- [127] Malliavin M J and Coudray C 1997 Ab initio calculations on $(\text{MgO})_n$, $(\text{CaO})_n$, and $(\text{NaCl})_n$ clusters ($n = 1-6$) *J. Chem. Phys.* **106** 2323
- [128] Veliah S, Xiang K H, Pandey R, Recio J M and Newsam J M 1998 Density functional study of chromium oxide clusters: structure, bonding, vibrations, and stability *Phys. Rev. B* **102** 1126
- [129] Bell R C, Zemski K A, Kerns K P, Deng H T and Castleman A W Jr 1998 Reactivities and collision-induced dissociation of vanadium oxide cluster cations *J. Phys. Chem. A* **102** 1733
- [130] France M R, Buchanan J W, Robinson J C, Pullins S H, Tucker J T, King R B and Duncan M A 1997 Antimony and bismuth oxide clusters: growth and decomposition of new magic number clusters *J. Phys. Chem. A* **101** 6214
- [131] Kaiser B, Bernhardt T M, Kinne M, Rademann K and Heidenreich A 1999 Formation, stability, and structures of antimony oxide cluster ions *J. Chem. Phys.* **110** 1437
- [132] Wang L S 2000 Photodetachment photoelectron spectroscopy of transition metal oxide species *Photoionization and Photodetachment Advanced Series in Physical Chemistry* **10**, ed C Y Ng (Singapore: World Scientific)
- [133] Wang L S, Wu H and Desai S R 1996 Sequential oxygen atom chemisorption on surfaces of small iron clusters *Phys. Rev. Lett.* **76** 4853
- [134] Wang L S, Nicholas J B, Dupuis M, Wu H and Colson S D 1997 Si_3O_x ($x = 1-6$): models for oxidation of silicon surfaces and defect sites in bulk oxide materials *Phys. Rev. Lett.* **78** 4450
- [135] Hafner J H, Bronikowski M J, Azamian B R, Nikolaev P, Rinzler A G, Colbert A T, Smith K A and Smalley R E 1998 Catalytic growth of single-wall carbon nanotubes from metal particles *Chem. Phys. Lett.* **296** 195
- [136] Guo B C, Kerns K P and Castleman A W Jr 1992 Tl_8C_{12} -metallo-carbohedrenes: a new class of molecular clusters? *Science* **255** 1411
- [137] Duncan M A 1997 Synthesis and characterization of metal-carbide clusters in the gas phase *J. Cluster Sci.* **8** 239
- [138] Clemmer D E, Shelimov K B and Jarrold M F 1994 Gas-phase self-assembly of endohedral metallofullerenes *Nature* **367** 718
- [139] Northby J A 1987 Structure and binding of Lennard-Jones clusters: $13 \leq N \leq 147$ *J. Chem. Phys.* **87** 6166
- [140] Berry R S 1993 Potential surfaces and dynamics: what clusters tell us *Chem. Rev.* **93** 2379
- [141] Miehle W, Kandler O, Leisner T and Echt O 1989 Mass spectrometric evidence for icosahedral structure in large rare gas clusters: Ar, Kr, Xe *J. Chem. Phys.* **91** 5940
- [142] Toennies J P 1990 Helium clusters *The Chemical Physics of Atomic and Molecular Clusters* ed G Scoles (Amsterdam: North-Holland) p 597
- [143] Grebenev S, Toennies J P and Vilesov A F 1998 Superfluidity within a small helium-4 cluster: the microscopic andronikashvili experiment *Science* **279** 2083
- [144] Bartelt A, Close J D, Federmann F, Quaas N and Toennies J P 1996 Cold metal clusters: helium droplets as a nanoscale cryostat *Phys. Rev. Lett.* **77** 3525
-

-24-

- [145] Higgins J, Ernst W E, Callegari C, Reho J, Lehmann K K, Scoles G and Gutowski M 1996 Spin polarized alkali clusters: observation of quartet states of the sodium trimer *Phys. Rev. Lett.* **77** 4532
- [146] Castleman A W Jr and Bowen K H Jr 1996 Clusters: structure, energetics, and dynamics of intermediate states of matter *J. Phys. Chem.* **100** 12 911
- [147] Bacic Z and Miller R E 1996 Molecular clusters: structure and dynamics of weakly bound systems *J. Phys. Chem.* **100** 12 945
- [148] Xantheas S S 1995 *Ab initio* studies of cyclic water clusters $(\text{H}_2\text{O})_n$, $n = 1-6$. III. Comparison of density functional with MP2 results *J. Chem. Phys.* **102** 4505
- [149] Liu K, Cruzan J D and Saykally R J 1996 Water clusters *Science* **271** 929
- [150] Kim J, Becker I, Cheshnovsky O and Johnson M A 1998 Photoelectron spectroscopy of the 'missing' hydrated electron clusters $(\text{H}_2\text{O})_n^-$, $n = 3, 5, 8, 9$: isomers and continuity with the dominant clusters $n = 6, 7$ and ≥ 11 *Chem. Phys. Lett.* **297** 90
- [151] Sanov A, Nandi S and Lineberger W C 1998 Transient solvent dynamics and incoherent control of photodissociation pathways in I_2^- cluster ions *J. Chem. Phys.* **108** 5155
-

FURTHER READING

Duncan M A (ed) 1993–1998 *Advances in Metal and Semiconductor Clusters* vol I–IV (Greenwich: JAI)

Scoles G (ed) 1990 *The Chemical Physics of Atomic and Molecular Clusters* (Amsterdam: North-Holland)

Jena P, Khanna S N and Rao B K (eds) 1992 *Physics and Chemistry of Finite Systems: From Clusters to Crystals* (Boston: Kluwer)

Haberland H (ed) 1994 *Clusters of Atoms and Molecules* (Berlin: Springer)

Schmidt R, Lutz H O and Dreizler R (eds) 1992 *Nuclear Physics Concepts in the Study of Atomic Cluster Physics* (Berlin: Springer)

Scoles G (ed) 1988 *Atomic and Molecular Beam Methods* vol 1 (Oxford: Oxford University Press)

C1.2 Fullerenes

Dirk M Guldi

INTRODUCTION

The scope of the following article is to survey the physical and chemical properties of the third modification of carbon, namely [60]fullerene and its higher analogues. The enthusiasm that was triggered by these spherical carbon allotropes resulted in an epidemic-like number of publications in the early to mid-1990s. In more recent years the field of fullerene chemistry is, however, dominated by the organic functionalization of the highly reactive fullerene

core, yielding literally thousands of fullerene derivatives with new and, in part, even more intriguing properties than pristine fullerenes. This still growing field is the subject of a new number of excellent review articles and books. The beginning of the current review deals with the fullerenes' structural and electronic configurations, followed by a detailed description of their undoped/doped thin films and a discussion of fullerene-based polymers and Langmuir–Blodgett films. In addition, the properties of fullerenes in condensed media, ranging from electrochemical redox reactions and photoexcited states to electron transfer processes, are elucidated. This account will end with a final section regarding metal incorporated endohedral complexes.

C1.2.1 STRUCTURE

The initial report regarding the existence and characterization of [60]fullerenes (figure C1.2.1) by Kroto *et al* is an important landmark for the chemistry and physics of fullerenes [1]. The importance of this discovery was acknowledged with the Nobel Prize in 1996. It took, however, a few more years until Krätschmer *et al* reported a method describing the arc discharge of carbon rods with the prospect of synthesizing large quantities of fullerene materials (figure C1.2.2) [2]. In parallel, the laser evaporation cluster beam technique has been employed by Smalley and coworkers to vaporize graphite in a helium atmosphere and, thus, to mimic the appropriate fullerene nucleation conditions [3]. With gram quantities at hand, scientists began to investigate the unique chemical and physical properties of this spherical carbon allotrope.

The fundamental concept proposed for the composition of three-dimensional fullerene structures is the introduction of five-membered (i.e. pentagon) rings, which are primarily responsible for the curvature [4]. They function like defects in a graphite structure and lead to nonplanarity of the π -electronic structure. However, the strain energy will only be minimized when the pentagons are as far apart as possible. This 'isolated pentagon' principle [5] has best been achieved in [60]fullerene, which consists of 12 regularly implanted five-membered (i.e. pentagon) and 20 six-membered (i.e. hexagon) rings and, therefore, differs most markedly from two-dimensional carbon structures (i.e. graphite). As a direct result of the 12 pentagonal faces, [60]fullerene shows, in contrast to graphitic sheets, an anisotropic electron distribution. In [60]fullerene, the pentagons are most evenly distributed, but not as far apart as possible.

-2-

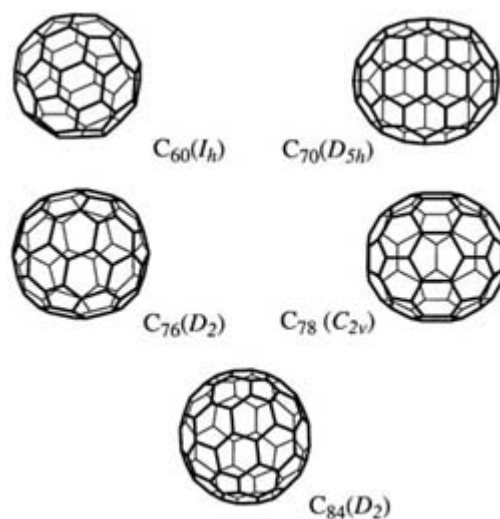


Figure C1.2.1. Structure of [60] fullerene (I_h), [70] fullerene (D_{5h}), [76] fullerene (D_2), [78] fullerene (C_{2v}) and [84] fullerene (D_2).

More important, the surface curvature of the carbon network exerts a profound impact on the reactivity of the fullerene core [6, 7]. In this context, the most striking consequence emerges from the pyramidalization of the individual carbon atoms. Influenced by the curvature, the sp^2 hybrids which exist in truly two-dimensional planar

carbon networks and hydrocarbons adopt an $sp^{2.278}$ hybridization with p orbitals that possess an s character of 0.085. Accordingly, the exterior surface is much more reactive than planar analogues, and is comparable to those of electron deficient polyolefins. This, in turn, rationalizes the high reactivity of the fullerene core towards photolytically and radiolytically generated carbon- and heteroatomic-centred radicals and also other neutral or ionic species [8]. The interior, in contrast, is shown to be practically inert [9]. Despite these surface related effects, the $sp^{2.278}$ character of the carbon atoms is expected to have a stabilizing effect on carbon-centred radical ions as well as carbanions and carbocations.

The first fullerene to be characterized was the I_h [60] fullerene, which was originally identified by its four-band IR absorption spectrum [2]. The proposed cage-like structure of [60] fullerene, with a diameter of 7.1 Å, was unequivocally confirmed by the detection of a single ^{13}C NMR resonance, stemming from the equivalency of all the carbon atoms in this molecule [10]. X-ray crystal structures reveal two different types of C–C bond, i.e. short 6–6 bonds, with a high double bond character and long 6–5 bonds, possessing low double bond character [3]. In contrast, the ^{13}C NMR spectrum of [70] fullerene exhibits five lines, again in perfect agreement with a closed sphere. Also ionization experiments helped to characterize the fullerene structure [11, 12 and 13]. The C_{60}^{++} carbon clusters, produced upon ionization, have large internal energies and cool via the sequential emission of C_2 molecules. The latter route arises predominantly from a combination of the high stability of the even-membered C_2 clusters and their relatively high binding energy of ~ 3.6 eV. This makes the C_2 loss mechanism energetically more favourable than separating two individual carbon atoms.

-3-

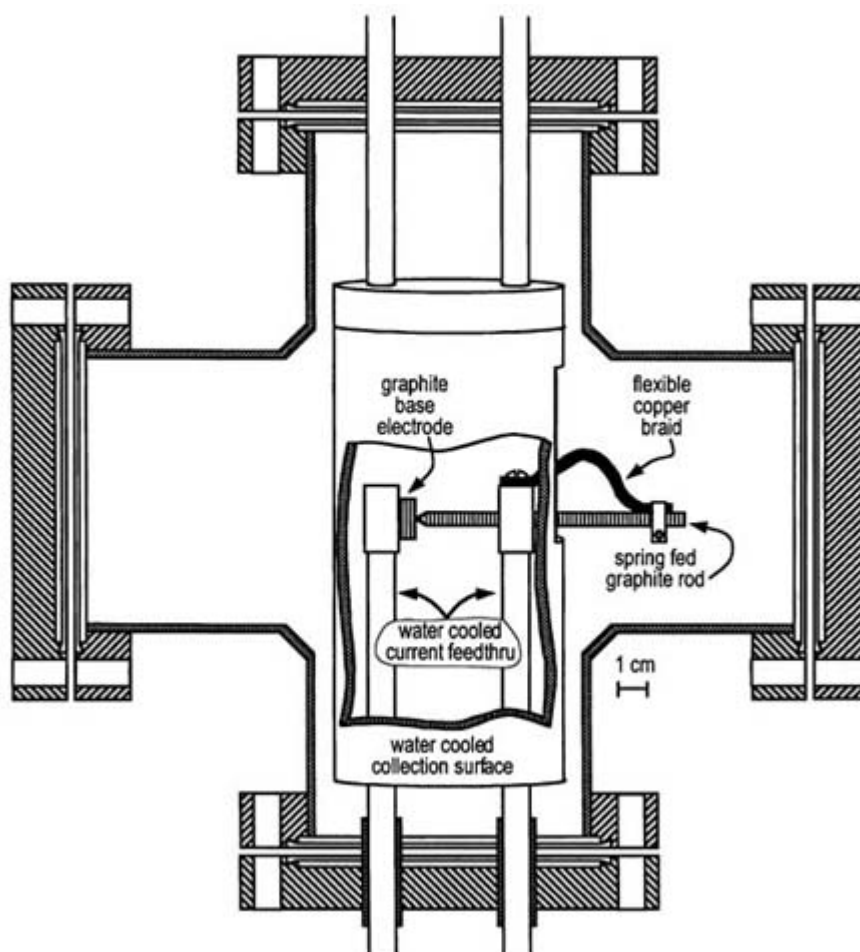


Figure C1.2.2. Diagram of the apparatus used to produce fullerenes from graphite rods.

In addition to the most abundant fullerene, namely [60]fullerene, a number of higher fullerenes have also been isolated and characterized, including [70] (point group D_{5h}), chiral [76] (point group D_2), the D_3 and C_{2v} isomers

of [78] and an equilibrated mixture of [84]fullerene of D_2 and D_{2d} symmetry (see [figure C1.2.1](#)) [14, 15, 16, 17 and 18]. Large crystalline quantities of these higher fullerenes are scarce, while there is a greater complexity associated with the lower symmetry of the molecule. In addition to their relatively small synthetic amounts, the presence of more than a single isomer which satisfies the isolated-pentagon rule results in further complications with respect to separation of these isomeric mixtures.

C1.2.2 CRYSTAL STRUCTURE

Below 90 K, [60]fullerene freezes into an orientational glass in which it adopts a simple cubic structure [19]. This low temperature structure can be traced to the anisotropic electronic structure. Alignment of the electron rich regions of

-4-

one molecule over the electron deficient regions of its neighbouring molecule optimizes the electrostatic contribution to the predominantly van der Waals intermolecular bonding and, in turn, governs the overall stability of this glass phase. Above 90 K, the structure is a primitive cubic structure (space group T_{6h} or $Pa3$) [20]. In this temperature regime, molecular motions are no longer restricted and, accordingly, the molecules start to move freely between two distinct nearly degenerate orientations, differing in energy by ~ 11.4 meV. In principle, this phase of residual rotational motion is followed by a first-order phase transition at 261 K yielding a face-centred cubic (fcc) structure $Fm3m$, characterized by rapid isotropic reorientational motion of the molecules [21]. In this phase transition, from an orientationally ordered to an orientationally disordered phase, a competition dominates between an entropy gain by rotation and an energy gain by intermolecular attraction.

[70]fullerene, with a D_{5h} symmetry, on the other hand, crystallizes in two phases, namely, centred cubic (ccp, [figure C1.2.3](#)) and hexagonal close packed structures (hcp, [figure C1.2.4](#)), which differ, in essence, only in their stacking sequence. Heating experiments on the pure hcp phase (ABAB), by means of dilatometry, in the 200–400 K temperature window give rise to two phase transitions: first to a deformed hcp and secondly to a monoclinic structure. Structural studies on a ccp crystal revealed a transition to an fcc ($Fm3m$) at high temperature and, upon cooling, a phase transition to a rhombohedral phase ($R3m$) (ABCABC) [22, 23, 24 and 25].

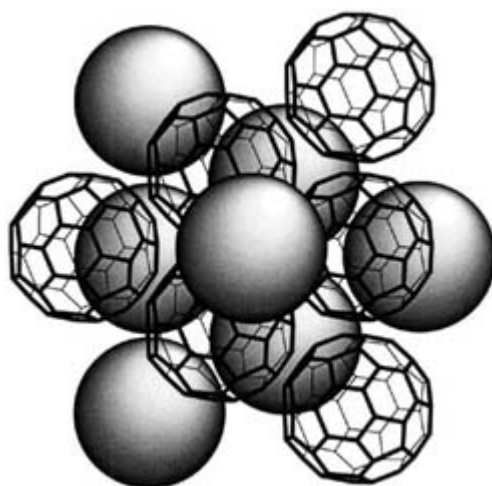


Figure C1.2.3. Cubic close packing (ABC) of [60]fullerene.

-5-

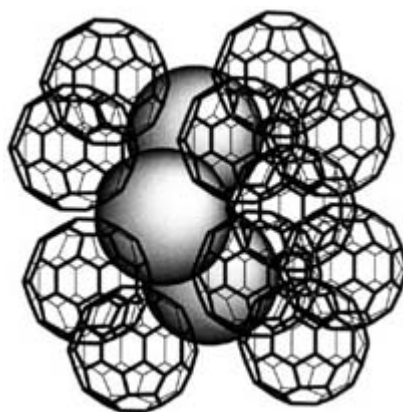


Figure C1.2.4. Hexagonal packing (ABA) of [60]fullerene.

C1.2.3 ELECTRONIC CONFIGURATION

[60]fullerene has a truncated-icosahedral form, with a point group symmetry I_h which allows a degeneracy as high as five. The 30 filled $p\pi$ orbitals host 60 π electrons, in a structural pattern closely resembling that of free particles on the surface of a sphere and, in turn, evoke an equal net atomic charge distribution on each carbon [26]. All 60 carbon atoms have equivalent symmetry, but the bonds fall into two sets, namely, hexagon–pentagon and hexagon–hexagon edges. The 60 Hückel molecular orbitals give rise to the reducible representation: $2A_g + 3T_{1g} + 4T_{2g} + 6G_g + 8H_g + 1A_u + 4T_{1u} + 5T_{2u} + 6G_u + 7H_u$. Only the A_g and the H_g modes are Raman active while the T_{1u} modes are solely IR active [27].

In essence, the 60 MOs split into 30 bonding and 30 antibonding π molecular orbitals with the h_u and t_{1u} broadening into the valence and conduction bands of the solid, respectively [28]. Because of the presence of both pentagonal and hexagonal rings in the fullerene cage, there are six t_{1u} band electrons in addition to the more common σ and π electrons [26]. For example, graphite consists only of hexagons and, hence, only the σ and π electrons are present. Molecular calculations regarding the electronic configuration have determined that this threefold degenerate LUMO (t_{1u}) is separated by ~ 1.8 eV from a lower lying fivefold degenerate HOMO (h_u) and from a higher lying threefold degenerate LUMO + 1 (t_{1g}) (figure C1.2.5) [29, 30]. The moderate optical energy gap not only underlines the remarkable electron accepting features of these carbon spheres, but it emphasizes, in combination with the optical conductivity, the semiconducting characteristics of solid fullerene, comparable to hydrogenated amorphous semiconductor silicon (a-Si:H).

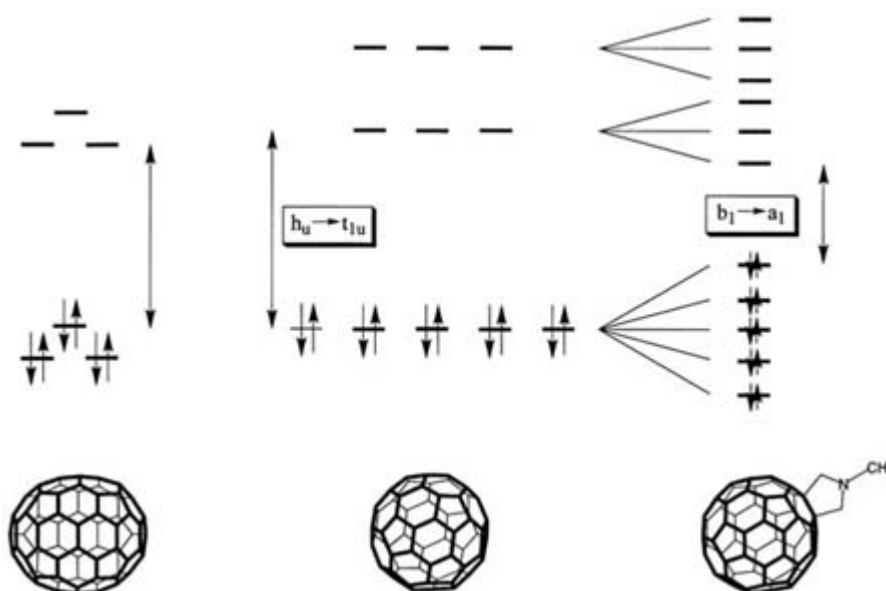


Figure C1.2.5. Illustration of the π orbital energy levels in [60]fullerene, [70]fullerene and monofunctionalized pyrrolidino[60]fullerene [26].

In light of oxidative processes, the high degree of resonance stabilization that arises from the maximally occupied HOMO (10 electrons), makes it an extremely difficult task to remove an electron from the HOMO level [31]. Thus, [60]fullerene can be considered mostly an electronegative entity which is much more easily reduced than oxidized.

The electronic configuration of higher fullerenes, e.g. [70] (figure C1.2.5) [76], [78] and [84]fullerene, is in essence similar to that known for [60]fullerene with, however, one fundamental difference. Their HOMO–LUMO energy gap decreases gradually with increasing number of carbon atoms [32]. On the other hand, recent calculations regarding the smallest so far isolated fullerene species, namely [36]fullerene, indicate also a substantially reduced energy gap of ~ 0.2 eV [33]. From the even stronger constrained curvature, relative to [60]fullerene, stems the fundamental consequence that the carbon atoms in [36]fullerene become so reactive that rapid polymerization occurs, preventing a systematic detailed investigation of any other properties of [36]fullerene.

The most important classes of functionalized [60]fullerene derivatives, e.g. methanofullerenes [34], pyrrolidinofullerenes [35], Diels–Alder adducts [34] and aziridinofullerene [36], all give rise to a cancellation of the fivefold degeneration of their HOMO and threefold degeneration of their LUMO levels (figure C1.2.5). This stems in a first order approximation from a perturbation of the fullerene's π -electron system in combination with a partial loss of the delocalization.

C1.2.4 THIN FILMS

The growth of a well ordered fullerene monolayer, by means of molecular beam epitaxy, has been used for the controlled nucleation of single crystalline thin films. The quality and stability of molecular thin films has been shown

to depend strongly on the interaction between the molecules and the substrates chosen for their growth [37, 38]. It is important to note that fullerenes give rise to much stronger intermolecular interactions relative to conventional organic molecules. Thus, their utilization for molecular thin films, by means of organic molecular beam epitaxy, has been vigorously investigated. In essence, parameters such as the character and also the strength of the interaction between the fullerene core and the substrate determine the morphology and regularity of the film growth. In the case of strong interactions the mobility of the chemisorbed fullerene molecule is limited and

successive deposition leads to the growth of polycrystalline grains with typically small diameters. In clear contrast, substrates such as {001} KBr [39, 40], {0001} MoS₂ [41], {0001} GaSe [41] {111} CaF [42], {111} Si [42, 43], {111} GeS [44], GaAs [45], freshly cleaved mica [46, 47] and layered materials [38], induce sufficiently strong van der Waals interaction between the fullerene molecules and, in turn, overpower the interaction between the substrate and individual fullerene molecules. This leads to a high effective surface mobility of the physisorbed fullerene molecules for fairly large grains. Predominantly a fcc structure with a series of close-packed planes {111} oriented with respect to the substrate plane, or a hcp structure, with {0001} close-packed planes, is found. Also good single-domain epitaxy structure was reported on Au(110), Ag(110) and Ni(110), while multi-domain growth predominates on Cu(111), Au(111), Ag(111) and Pt(111). Charge transfer into the fullerene's LUMO is deemed to be the dominant effect that leads to a strong interaction with the substrate and reduces the effective surface mobility, as has been observed for Au{110} and a variety of metals including Ag, Mg, Cr and Bi [48,49,50,51 and 52].

Consequently, deposition of the first monolayer is the most important factor, determining the growth of the subsequent layers and, consequently, the crystallinity of the resulting multilayered films. General information regarding the strength of the fullerene–substrate bond can be derived from the thermal stability of the fullerene layer. This has been impressively documented via the observation of the principally different crystallinity of [60] fullerene films on semiconductor surfaces such as Si. Depending on the Si surface, being either hydrophobic (passivated) and hydrophilic (non-passivated), films were crystalline with a fcc structure and a noticeable {111} texture, or of amorphous nature, respectively. The amorphous character has been ascribed to the fullerene's interaction with the hydrophilic substrate, where [60]fullerene is mobile and diffuses freely even at temperatures as low as 100 K [37].

C1.2.5 DOPING OF FULLERENES AND SUPERCONDUCTIVITY

The crystal structure of [60]fullerene reveals characteristics of a fcc packing with the molecules located at the lattice points and four fullerene molecules per unit cell [26]. According to the fcc lattice constant, $a = 14.15 \text{ \AA}$ the intramolecular centre-to-centre distance ($a/\sqrt{2}$) is exactly 10.01 \AA . The close packing of the nanometre-sized fullerenes creates two types of interstitial site, sufficiently large to host small-sized molecules or atoms without distorting the crystal [53]. In particular, one octahedral site and two tetrahedral sites per [60]fullerene are present with radii of 2.06 \AA and 1.12 \AA , respectively. It was shown that the respective tetrahedral and octahedral vacancies of the fullerene crystal may be filled with a wide variety of dopants [54,55,56 and 57]. These range from various alkali (Li, Na, K, Rb, Cs), earth alkali metals (Ca, Sr, Ba) and rare-earth metals (Yb, Sm, Eu) to organic donors, such as ferrocene and tetrakis(dimethylamino) ethylene. Metal diffusion in fullerene films and single crystals or vapour transport diffusion are the most widely applied methodologies for fabricating [60]fullerene–metal intercalation composites. On the other hand, intercalation of compounds that have low diffusion coefficients in C₆₀ necessitates sublimation in a UHV chamber.

The occupation of each tetrahedral and octahedral site in these regularly oriented arrays of cavities by, for example, alkali atoms results in the transfer of a single electron to the fullerene's conduction band (t_{1u}) [58]. Consequently,

via stoichiometric alkali metal doping, intercalation structures were fabricated ranging from A₁C₆₀ and A₃C₆₀ to terminally reacted A₆C₆₀ phases and A₁₂C₆₀/A'₆C₆₀ (A = Li and A' = Sr, Ba; full occupation of the LUMO (t_{1u}) and LUMO + 1 (t_{1g}) levels) (figure C1.2.6) [54, 55, 56 and 57]. It was shown that the stability of the alkali A₃C₆₀ and A₆C₆₀ composite structures is mainly governed by the Madelung potential.

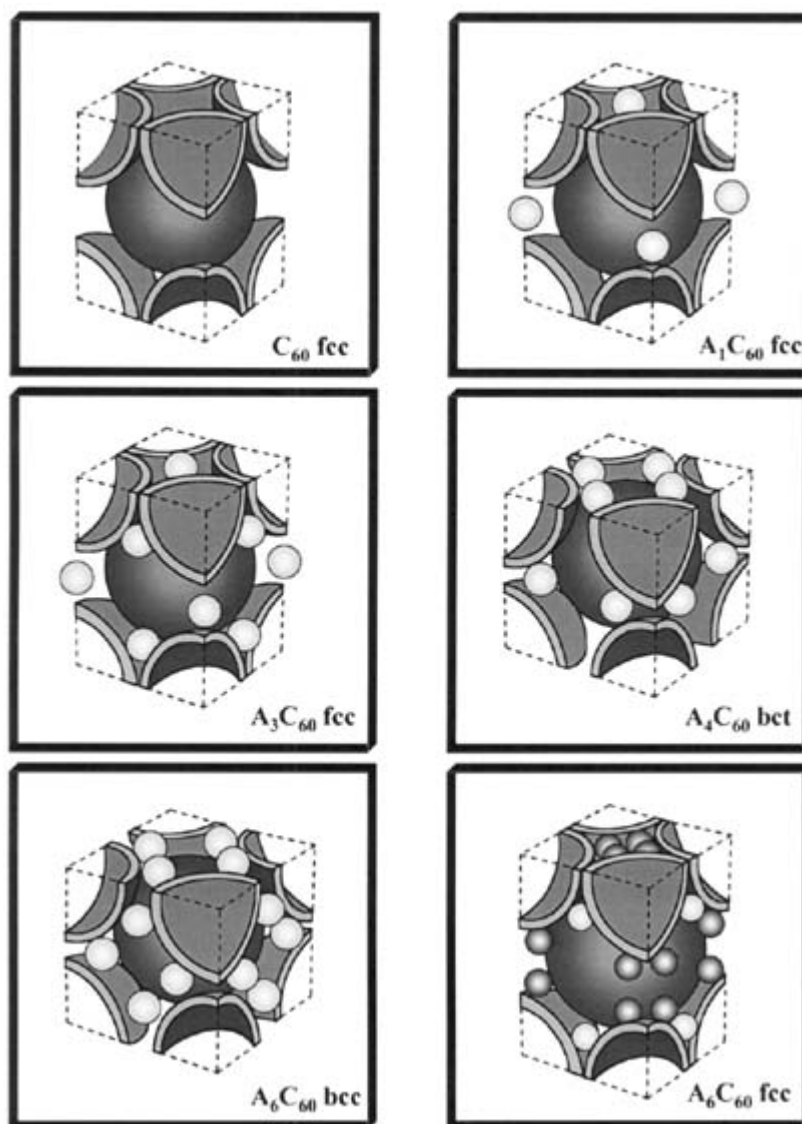


Figure C1.2.6. Summary of fcc [60]fullerene structure and alkali-intercalation composites of [60]fullerene.

-9-

The half-filled molecular conduction band (t_{1u}) in A_3C_{60} gives rise to a maximized density of states at the Fermi level. It is interesting to note that this metallic phase becomes a superconductor at low temperatures, while the A_6C_{60} phase proves to be mainly insulating (figure C1.2.6). The resulting A_3C_{60} compounds are typically ionic $[A^+]_3[C_{60}^{3-}]$ and form fcc lattices which positions the icosahedral fullerene cores in sites of local cubic symmetry with a lattice parameter ($a = 14.24 \text{ \AA}$) that is only slightly expanded from that of fcc dopant free [60]fullerene [54].

Raman scattering is the key technique for probing the doping process in fullerites [59]. Specifically, the position of the $A_g(2)$ pinch mode, which is a characteristic fingerprint for the charge transfer in these alkali doped systems, has been employed with great success to identify the various doped phases. Fundamental experiments have been performed with the scope to identify the nature of the superconductivity in these classes of materials including interpretation of the phase diagram as a function of composition, pressure and magnetic field, structure determination, magnetic susceptibility and, finally, NMR relaxation measurements in the normal state [60]. Specifically, band structure calculations on A_3C_{60} composites indicate that the charge transfer is nearly complete and that the electrons are used to half fill the conduction band [54]. Consequently, a simple charge transfer concept, from the cations to the LUMO of the fullerene, yielding a metallic state, has been proposed for a qualitative rationalization of the electronic properties.

In pristine [60] fullerene, the t_{1u} band is completely empty while, in contrast, the A_6C_{60} phase (bcc lattice) has a completely filled conduction band. In the intermediately doped A_4C_{60} phase (bct lattice), the density of states at the Fermi level is, however, nearly zero. These considerations are consistent with the absence of high temperature superconductivity in [60]fullerene, A_4C_{60} and A_6C_{60} . In conclusion, the superconducting behaviour strongly depends on the concentration of conduction band electrons, on the lattice constant and the degree of orientational order, yielding composites which display T_C values between 2 and 40 K. The highest T_C values that are reported are those of K- (33 K), Rb- (33 K) and Cs- (40 K, stabilized under hydrostatic pressure) doped A_3C_{60} composites (figure C1.2.7). Their properties may be best understood on the basis of a high average phonon frequency in combination with weak intermolecular interactions and strongly scattering intramolecular modes.

Also, novel magnetic properties have been reported in mixed fullerene composites, in which the fullerene is limited to a single negative charge. For instance, the tetrakis(dimethylamino) ethylene/[60]fullerene salt, namely, $[TDAE^+][C_{60}^-]$, has been described as a soft ferromagnet with a Curie temperature of 16 K [61, 62].

-10-

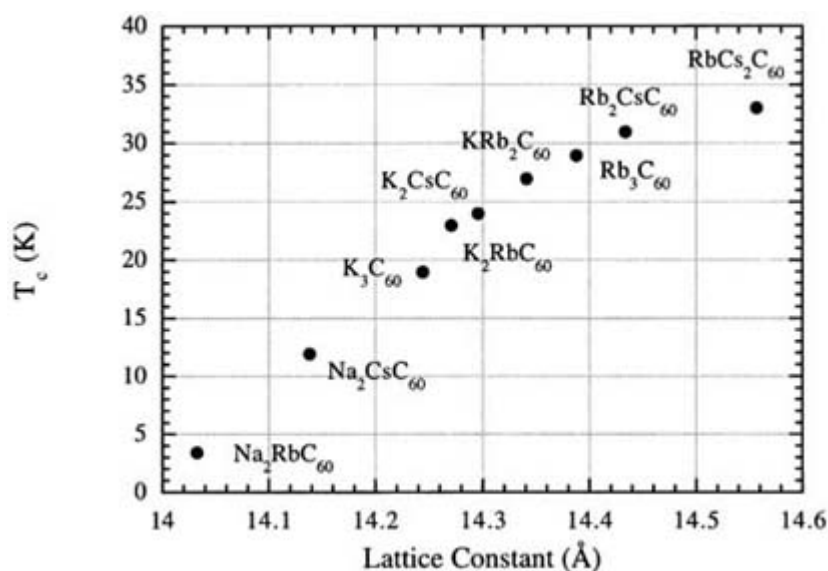


Figure C1.2.7. Superconducting transition temperature plotted as a function of the a lattice parameter for a variety of A_3C_{60} phases [55].

C1.2.6 FULLERENE POLYMERS

Several studies have demonstrated the successful incorporation of [60]fullerene into polymeric structures by following two general concepts: (i) in-chain addition, so called pearl necklace type polymers or (ii) on-chain addition pendant polymers. Pendant copolymers emerge predominantly from the controlled mono- and multiple functionalization of the fullerene core with different amine-, azide-, ethylene propylene terpolymer, polystyrene, poly(oxyethylene) and poly(oxypropylene) precursors [63,64,65,66,67 and 68]. On the other hand, $(-C_{60}Pd-)_n$ polymers of the pearl necklace type were formed via the periodic linkage of [60]fullerene and Pd monomer units after their initial reaction with the p -xylylene diradical [69,70 and 71].

An alternative approach envisages the stimulating idea to produce an all-carbon fullerene polymer in which adjacent fullerenes are linked by covalent bonds and align in well characterized one-, two- and three-dimensional arrays. Polymerization of [60]fullerene, with the selective formation of covalent bonds, occurs upon treatment under pressure and relatively high temperatures, or upon photopolymerization in the absence of a triplet quencher, such as molecular oxygen, using an Ar^+ ion laser at intensities of 50 mW mm^{-2} [72,73,74,75,76,77 and 78]. The synthesis of at least three polymer phases is reported with characteristics ranging from those of rhombohedral and

orthorhombic to tetragonal phases (figure C1.2.8). Typically, in these oligomer and polymer composites the interfullerene C–C linkage evolves from a [2+2] cycloaddition between neighbouring fullerene cores, which results in the formation of four-membered carbon rings (D_{2h} symmetry), fusing together adjacent molecules. Photopolymerization is hindered below the ordering transition of [60]fullerene as the probability of short carbon–carbon bonds approaching intermolecular contact diminishes.

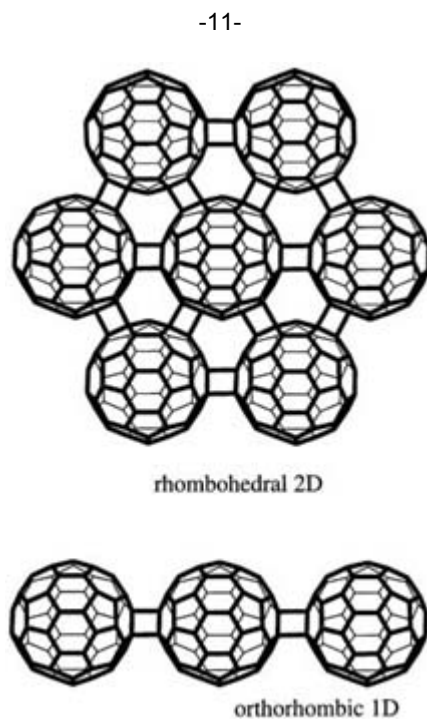


Figure C1.2.8. Orthorhombic 1D and rhombohedral 2D structure of polymerized [60]fullerene.

The C–C linkage in the polymeric [60]fullerene composite is highly unstable and, in turn, the reversible [2+2] phototransformation leads to an almost quantitative recovery of the crystalline fullerene. In contrast the similarly conducted illumination of [70]fullerene films results in an irreversible and randomly occurring photodimerization. The important aspect which underlines the markedly different reactivity of the [60]fullerene polymer material relative to, for example, the analogous [36]fullerene composites, is the reversible transformation of the former back to the initial fcc phase.

C1.2.7 LANGMUIR–BLODGETT FILMS

Well ordered two-dimensional monolayered films are of great interest because of the valuable insights they provide regarding molecule interactions and their potential application to important technologies related to coatings and surface modifications. The optical and electrical properties of [60]fullerene films are strongly affected by the deposition conditions, by impurities or disordered structures. Thus, the exploration of these properties requires the controlled incorporation of fullerenes in well defined two-dimensional arrays and three-dimensional networks. Extensive efforts have been undertaken, ranging from modifying the deposition conditions to applying amphiphilic host molecules, with the scope to generate stable and well ordered monolayered fullerene films [79,80,81,82,83,84,85,86,87,88,89,90,91 and 92]. The strong π – π interaction and the resulting tendency to form aggregates precludes, however, formation of stable monolayers and Langmuir–Blodgett films at the air–water interface and solid substrates, respectively. In essence, the currently available data suggest three promising approaches to overcome these fundamental difficulties: (i) amphiphilic functionalization of the hydrophobic fullerene core via covalent attachment of hydrophilic groups [93,94,95,96,97,98,99,100,101,102,103 and 104], (ii) reducing the hydrophobic surface via controlled multiple

functionalization [105,106 and 107] or (iii) self-assembly via electrostatic attractions of oppositely charged species [108].

The controlled functionalization of the fullerene core, at one or more positions by addends of different hydrophobicities, has been demonstrated to provide a viable alternative to control the supermolecular structures formed upon their spreading on the water surface. It was unequivocally demonstrated that choosing an adequate hydrophobic–hydrophilic balance is a fundamental aspect with regard to forming stable and monolayered fullerene composites. In particular, functionalization with hydrophilic addends, such as cryptates, triethyleneglycol monomethyl ether, benzocrowns, N-acetyl pyrrolidine derivatives, carboxylic acid groups and oxygen ($C_{60}O$), leads to the important promotion of the amphiphilic character of the fullerene core. A simple picture helps to rationalize the success of this concept: The hydrophilic or water-soluble head groups enhance the interaction with the aqueous subphase and, in turn, allow a two-dimensional fixation of the fullerene core at the air–water interface.

C1.2.8 ELECTROCHEMISTRY

One aspect that reflects the electronic configuration of fullerenes relates to the electrochemically induced reduction and oxidation processes in solution. In good agreement with the threefold degenerate LUMO, the redox chemistry of [60]fullerene, investigated primarily with cyclic voltammetry and Osteryoung square wave voltammetry, unravels six reversible, one-electron reduction steps with potentials that are equally separated from each other. The separation between any two successive reduction steps is $\sim 450 \pm 50$ mV. The low reduction potential (only -0.44 V *versus* SCE) of the process, that corresponds to the generation of the π -radical anion [31,109,110,111 and 112], deserves special attention.

In contrast to the relative ease of reduction, oxidation of fullerenes requires more severe conditions [113, 114]. Not only does the resonance stabilization raise the level of the corresponding oxidation potential (1.26 V *versus* Fc/Fc^+), but also the reversibility of the underlying redox process is affected [115].

This behaviour also stands for functionalized [60]fullerene derivatives, with, however, a few striking differences. The most obvious parameter is the negative shift of the reduction potentials, which typically amounts to ~ 100 mV. Secondly, the separation of the corresponding reduction potentials is clearly different. While the first two reduction steps follow closely the trend noted for pristine [60]fullerene, the remaining four steps display an enhanced separation. This has, again, a good resemblance to the HOMO–LUMO calculations, namely, a cancellation of the degeneration for functionalized [60]fullerenes [31, 116, 117].

The electrochemical features of the next higher fullerene, namely, [70]fullerene, resemble the prediction of a doubly degenerate LUMO and a LUMO + 1 which are separated by a small energy gap. Specifically, six reversible one-electron reduction steps are noticed with, however, a larger splitting between the fourth and fifth reduction waves. It is important to note that the first reduction potential is less negative than that of [60]fullerene [31]. Parallel to the shift that the reduction of higher fullerenes shows, oxidation of the latter is also made easier (D_{5h} [70] fullerene: $+1.20$ V *versus* Fc/Fc^+). The underlying HOMO LUMO gap in D_{5h} [70]fullerene (2.22 V) is, therefore, markedly decreased relative to [60]fullerene (2.32 V). This trend is further extended in D_2 [76]fullerene (1.64 V) C_{2v} [78]fullerene (1.72 V) and D_2/D_{2d} [84]fullerene (1.6 V) [32, 118]. In conclusion, higher fullerenes are better electron accepting and, at the same time, better electron donating materials relative to their smaller cousin, [60] fullerene.

Thin films of fullerenes, which were deposited on an electrode surface via, for example, drop coating, were largely heterogeneous, due to the entrapping of solvent molecules into their domains. Consequently, their electrochemical behaviour displayed different degrees of reversibility and stability depending on the time of electrolysis and the

number of consecutive redox cycles scanned. Langmuir–Blodgett films of pristine [60]fullerene, upon electrochemical reduction, formed insoluble films which stem from the immobilization of charge compensating counter-cations into the film. The large separation between the cathodic and anodic waves, indicative of a high degree of irreversibility, has been attributed to structural rearrangements upon the reduction and reoxidation process and documents the high disordering of the fullerene cores in LB films [91, 119, 120, 121, 122, 123 and 124].

C1.2.9 SOLUBILITY

The quasi-aromatic structure of fullerenes affects the solubility of these hydrophobic moieties. Typical representatives are nonpolar organic solvents, such as toluene, benzene and chlorinated hydrocarbons [25]. In toluene, benzene and o-xylene the solubility of [60]fullerene exhibits, surprisingly, a negative temperature dependence, along with a maximum of solubility around 280 K. On the other hand, polar solvents including alcohols and aqueous systems are of impractical use for investigating the physical and chemical properties of fullerenes [126, 127]. For example, in polar solutions the hydrophobic fullerene core aggregates spontaneously, yielding clusters with indefinite aggregate sizes and unknown properties that vary, in part, quite significantly from those of true fullerene monomers [128, 129]. This fullerene clustering has been monitored directly through dynamic light scattering and gel exclusion chromatography. An elegant route to overcome, in particular, the water-insolubility of fullerenes, is their incorporation into water-soluble superstructures. In this context, cyclodextrins [130, 131], calixarenes [132], various micellar [133, 134, 135 and 136] and vesicular host structures [137, 138, 139 and 140] were successfully utilized and the resulting complexes were studied under the aspect of photo-induced cytotoxicity and, if ^{14}C radiolabelled, as potential biochemical tracers [141].

C1.2.10 PHOTOEXCITED STATES

Another interesting physical feature relates to the chromophoric character of fullerenes. Based on the symmetry prohibitions, solutions of [60]fullerene absorb predominantly in the UV region, with distinct maxima at 220, 260 and 330 nm. In contrast to extinction coefficients on the order of $10^5 \text{ M}^{-1} \text{ cm}^{-1}$ at these wavelengths, the visible region shows only relatively weak transitions (λ_{max} at 536 nm; $\epsilon = 710 \text{ M}^{-1} \text{ cm}^{-1}$) [142].

Similar to the fullerene ground state the singlet and triplet excited state properties of the carbon network are best discussed with respect to the three-dimensional symmetry. Surprisingly, the singlet excited state gives rise to a low emission fluorescence quantum yield (Φ_{FLU}) of 1.0×10^{-4} [143]. Despite the highly constrained carbon network, the low Φ value relates to the combination of a short lifetime (1.8 ns) [144], a quantitative intersystem crossing ($\Phi_{\text{ISC}} = 1$) [145] and, finally, the symmetry forbidden nature of the lowest energy transition. To the same extent, also the phosphorescence quantum yield ($\Phi_{\text{PHO}} < 10^{-6}$) [146] is strongly impacted by the spherical structure.

Concerning transient absorption, laser or light excitation throughout the UV–visible region leads to the generation of the singlet excited state. The latter gives rise to a characteristic singlet–singlet absorption, maximizing around 920 nm [144], whose lowest vibrational state has an energy of 1.99 eV. Once formed, the singlet excited state undergoes

a rapid and, more importantly, a quantitative intersystem crossing to the energetically lower lying triplet excited state. The lowest triplet excited state has an energy of $\sim 1.57 \text{ eV}$ [147, 148]. In the case of [60]fullerene, this intersystem crossing takes place with a lifetime of 1.8 ns, governed by a large spin–orbit coupling which makes this process much faster than those known for two-dimensional rigid hydrocarbons. On the other hand, the fast ISC rate and the weak fluorescence provide the means for a triplet quantum yield (Φ_{TRIPLET}) close to unity.

The triplet–triplet absorption spectrum reveals, similar to the singlet–singlet features, a maximum in the near-IR

region around 750 nm [149]. In the absence of alternative deactivation processes, such as triplet–triplet annihilation and also ground-state quenching, the triplet lifetime amounts to ~100 μ s and is, again, much shorter than the triplet lifetime of comparable planar hydrocarbons [150, 151]. In this context, it is interesting to note that the highly constrained carbon network, particularly that of [60]fullerene, prohibits any vibrational motion, C–C bond elongation, or even changes of the dipole moment that may accelerate the deactivation of the singlet or triplet excited states. This argument is further substantiated by a small Stokes effect [152], which correlates to the energetic adaptation of the excited state to a new solvent environment, and also insignificant resonance Raman shifts [153] upon reduction of the fullerene core.

In aerated or oxygen saturated solutions, the fullerene triplet lifetime suffers a marked reduction [147, 148]. Luminescence studies (at 1365 nm) helped to identify singlet oxygen ($^1\text{O}_2$) as a product evolving from a bimolecular, Dexter-type energy transfer reaction. Particularly promising is the quantum yield for the singlet oxygen formation, which is near unity [147, 148]. In other words, the fullerene triplet excited state is quantitatively converted into this biologically important oxygen species. Corresponding experiments with functionalized [60] fullerene derivatives that display sufficient water solubility in the absence of a host structure, such as $\text{C}_{60}[\text{C}(\text{COO}^-)_2]_2$, $\text{C}_{60}[\text{C}(\text{COO}^-)_2]_3$, $\text{C}_{60}[(\text{CH}_2)_4\text{SO}_3\text{Na}]_6$ and $\text{C}_{60}(\text{OH})_{18}$, corroborated the efficient formation of singlet oxygen also in aqueous solutions [154, 155]. This, of course, evoked a tremendous interest to probe fullerenes as a potential agent for photodynamic therapy. Encouraging results stem from the strong cytotoxicity to L929 upon visible light irradiation as a result of superoxide production [141].

There are, indeed, many biological implications that have been triggered by the advent of fullerenes. They range from potential inhibition of HIV-1 protease, synthesis of drugs for photodynamic therapy and free radical scavenging (antioxidants), to participation in photo-induced DNA scission processes [156, 157, 158, 159, 160, 161, 162 and 163]. These examples unequivocally demonstrate the particular importance of water-soluble fullerenes and are summarized in a few excellent reviews [141, 175].

Another application is optical limiting [164, 165]. This is performed by materials whose transmittance strongly and quickly drops as the intensity of a laser pulse traversing them increases beyond a saturation level. The reverse saturable absorption mechanism is assumed to be the major parameter that determines the optical limiting. This mechanism plays an active role when an excited state, that is efficiently populated by optical excitation, has an absorption cross-section larger than that of the ground state at the excitation wavelength. In this light, the weak broad absorption of [60]fullerene throughout the UV–visible spectrum is beneficial to excite the ground state. Most importantly, both excited states, e.g. the singlet and triplet state, display cross sections significantly larger than that of the singlet ground state all over the accessible visible and near-IR spectrum. Consequently, the reverse saturable absorption that occurs with samples of pristine [60]fullerenes and functionalized [60]fullerene derivatives during the picosecond time regime and also on the nano-/microsecond time scales is attributed to emerging from the lowest singlet and triplet excited state, respectively. In particular, sol–gel films of [60]fullerene and some derivatives exhibit a marked enhancement in the red spectral region.

C1.2.11 π -RADICAL ANIONS

The most important species among the reduced fullerenes (π -radical anions to hexa-anions) are the one-electron reduced forms. In general, various techniques were employed for their characterization, ranging from transient absorption spectroscopy [166, 167 and 168] to transient electron spin resonance spectroscopy [111, 112, 167, 169, 170, 171 and 172]. The absorption features of [60], [70], [76], [78] and [84]fullerene π -radical anions, which lie predominantly in the near-IR region, are unmistakably confirmed [173]. For example, the [60]fullerene π -radical anion shows a narrow band around 1080 nm which serves as a diagnostic probe for the identification of this one-electron reduced species and, furthermore, allows an accurate analysis of inter- and intramolecular ET dynamics in [60]fullerene containing systems. This cannot be concluded for the interpretation of the ESR signals, which are still subject to a controversially conducted discussion favouring either a narrow or, alternatively, a broad ESR feature [111, 112, 167, 169, 170, 171 and 172]. A recent hypothesis proposes, in essence, a narrow ESR line for the π -radical anion which, upon rapid dimerization, undergoes a significant line broadening [169, 170]. It remains,

however, to be shown by different techniques such as, for example, time-resolved pulse radiolysis coupled with an ESR detector which assumption can be trusted.

C1.2.12 ELECTRON TRANSFER REACTIONS

The combination of a high degree of electron delocalization within the fullerene's π -system and their effective sizes prompts the application of this carbon material as new electron accepting moieties (figure C1.2.9). More importantly, the total reorganization energy upon reduction has been shown to be relatively small [174]. Hence, fullerenes became very appealing spheres for inter- and intramolecular electron transfer processes under the aspect of energy conversion and energy storage [175, 176 and 177].

A specific case that attracted a lot of attention encompasses the intermolecular electron transfer from a series of arene radical cations to [76] and [78]fullerene [178]. The high degree of charge and energy delocalization within the fullerene moiety is expected to exert an effect in the desired direction as it minimizes vibrational differences between the reaction partners in the ground and transition state. The corresponding relation between the rate constant and the thermodynamic driving force follows the features of a parabola, i.e. the electron transfer rates increase only to a maximal value before they decrease noticeably at higher driving forces. This kinetic study made use of the unequally sized reaction partners, namely, a large-sized electron donor and small-sized electron acceptor couple, elevating the diffusion-controlled limit. Furthermore, the relatively low reorganization energy is clearly beneficial for the possibility to establish a Marcus-inverted behaviour which facilitates reaching the maximum of the exothermic electron transfer process at lower $-\Delta G$ and, in turn, reaching the inverted region at lower energy. This example is one of the rare cases that establish a 'Marcus-inverted' region in a bimolecular electron transfer, beside those reports on the geminate recombination of photolytically generated radical pairs.

-16-

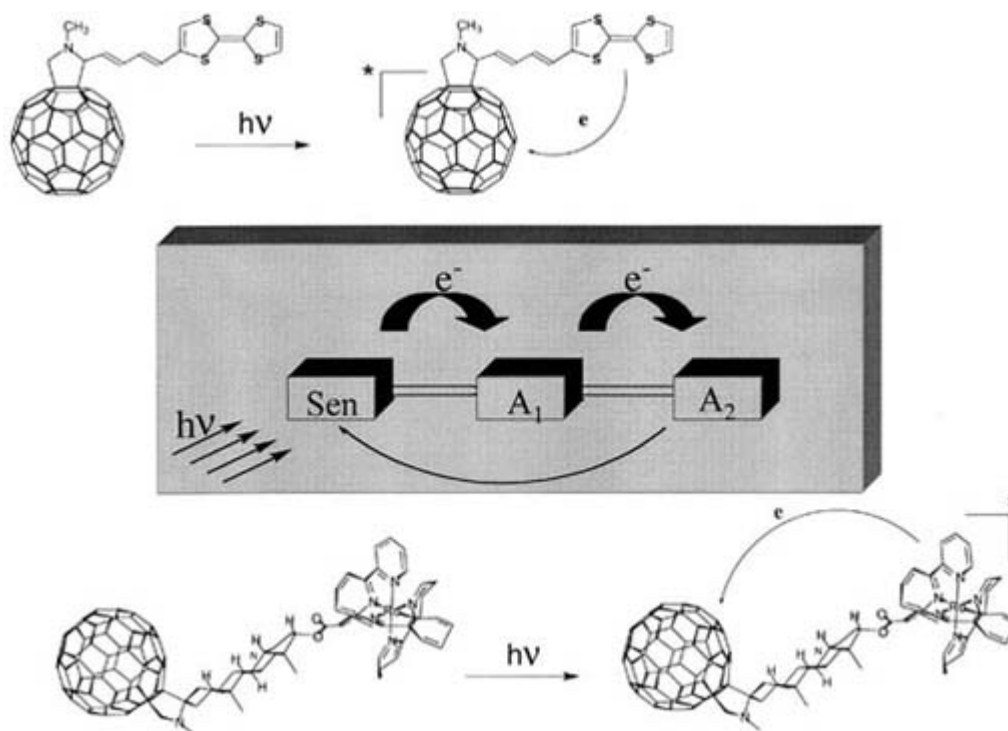


Figure C1.2.9. Schematic representation of photoinduced electron transfer events in fullerene based donor-acceptor arrays (i) from a TTF donor moiety to a singlet excited fullerene and (ii) from a ruthenium excited MLCT state to the ground state fullerene.

The redox properties of pristine fullerenes and monofunctionalized fullerene derivatives in their ground and excited states have drawn much attention for the design of devices such as molecular switches, receptors, photoconductors and photoactive dyads [175, 176 and 177]. These applications are generally based on the implication of fullerenes as a multifunctional electron storage moiety. The excellent electron accepting properties of fullerenes, together with their low reorganization energy, makes [60]fullerene and its derivatives good candidates for building blocks of systems employable for solar energy conversion, batteries and photovoltaics. The concept of linking fullerenes to a number of interesting electro- or photoactive species offers new opportunities in the preparation of materials with building blocks having highly symmetrical and coordinating geometries.

-17-

The chemically simplest case encompasses the covalent attachment of addends that lack any visible absorption and are as such redox inactive, but may indirectly influence electron or energy transfer reactions. Of increasing complexity are arrays carrying electroactive addends, e.g. in *N,N* dimethylaniline (DMA) [179, 180], ferrocene (Fc) [181] or tetrathiofulvalene (TTF) (figure C1.2.10) [182]. In these multicomponent supermolecules, the fullerene core is implemented as a photosensitizer that sequentially accepts an electron from the adjacent electroactive moiety. Accordingly, these systems can be classified as electroactive, but due to their insignificant visible absorption characteristics, photoinactive dyads. Considering the moderate absorption features of [60]fullerene in the visible region, functionalization with antenna molecules, such as metalloporphyrins [183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193 and 194] or MLCT transition metal complexes [195, 196 and 197], have developed as important objectives to promote the visible absorption characteristics of the resulting dyads and, most importantly, to improve the light harvesting efficiency of the fullerene core (figure C1.2.10). As a direct consequence, the role of the fullerene is significantly changed. Under these circumstances fullerenes operate exclusively as either electron or energy acceptor moieties. For details, the reader is directed to a series of excellent review articles, which appeared during recent years [175, 176 and 177].

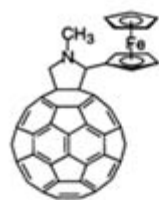
C1.2.13 ENDOHEDRAL FULLERENES

Finally, endohedral fullerenes are discussed. They attracted considerable attention for their potential use as superconductors, organic ferromagnets and magnetic resonance imaging agents (MRI). The enthusiasm that has arisen is based, in part, on the fact that the carbon network of each fullerene surrounds a large empty space which, in turn, renders it capable of encapsulating atomic particles. Furthermore, these novel materials created the stimulating possibility to fine-tune the fullerene's physical and chemical properties via systematic substitution of the embedded metal species. In general, two approaches are pursued to incorporate the metal into the fullerene's interior. The first one implies the synergetic utilization of the arc discharge method of carbon rods in the presence of metal carbides [198, 199]. Thus, the metal is present during the genesis of the fullerene network and can be scavenged by the closing sphere. In contrast to this approach, the second alternative involves the chemically induced opening of the carbon network, stuffing of the vacant interior with metals, and, in the last step, the subsequent re-closing of the open sphere [200]. It should be emphasized that the latter concept is a very challenging endeavour from the standpoint of synthesis. In fact, so far only the first route has led to isolable yields of endohedral fullerenes.

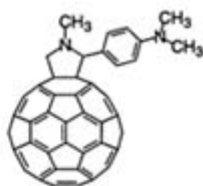
Metallofullerenes are commonly found with [74], [76], [80] and [82] fullerene and span composites that have a single ($M@C_{82}$), two ($M_2@C_{82}$) or even three metal atoms ($M_3@C_{82}$) encapsulated. The first type of metallofullerene extracted from fullerene soot was lanthanum fullerene $La@C_{82}$ followed a short time later by the detection of scandium fullerene $Sc@C_{82}$ and yttrium fullerene $Y@C_{82}$ [201, 202 and 203]. These have been completed by essentially all alkali metals, alkali-earth metals, noble gases and rare-earth metals (figure C1.2.11) [204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219 and 220].

-18-

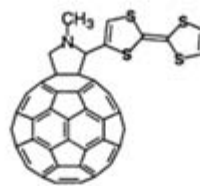
Electroactive Addends



Prato et al., 1993

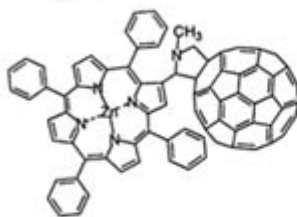


Williams et al., 1995

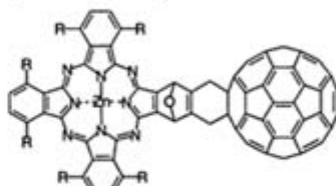


Prato et al., 1995

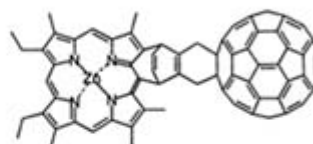
Porphyrin & Phthalocyanine Dyads



Reed et al., 1995

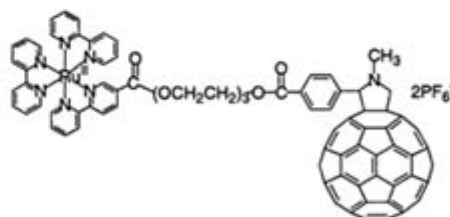


Hirsch et al., 1995



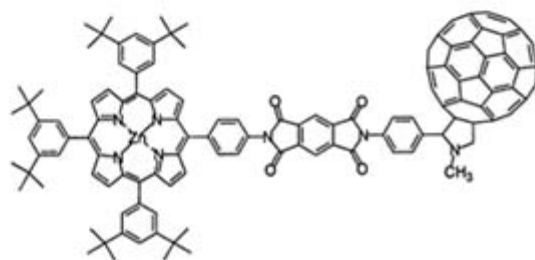
Gust et al., 1994

Ruthenium MLCT Dyad



Prato, Maggini et al., 1994

Porphyrin Triad



Sakata et al., 1998

Figure C1.2.10. Representative examples of fullerene based donor–bridge–acceptor dyads and triads.

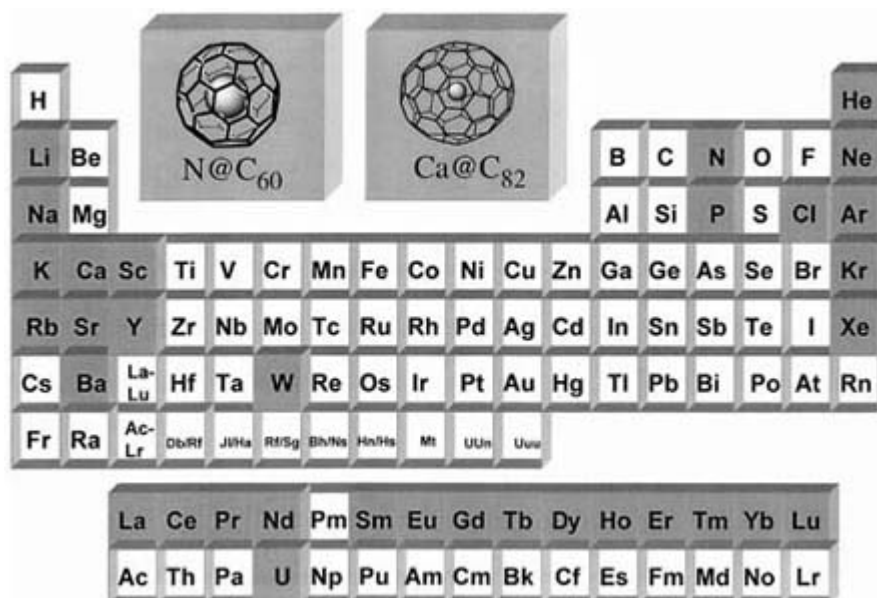


Figure C1.2.11. Synthesized and isolated endohedral fullerenes.

There has been a long dispute on the endo- or exohedral nature of such metallofullerenes. Experimental evidence including scanning tunnelling microscopy, extended x-ray absorption fine structure and transmission electron microscopy confirmed unequivocally the endohedral structure of these fullerene composites [204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217 and 218]. On the other hand, ESR has been proven to be the method of choice to investigate the electronic state [210, 221, 222 and 223]. Spectral evidence from the latter technique demonstrates that the encaged metal atom transfers a significant amount of charge to the carbon cage.

The most recent success in light of stuffing the fullerene's interior and using it as a carrier is the stabilization of atomic nitrogen and phosphorus inside [60]fullerene [9, 224]. Particularly, the remarkable stability of the $N@C_{60}$ system led the researchers to postulate a Faraday-cage-like property of the carbon network. EPR and ENDOR experiments, in combination with the stability of the system in air, clearly demonstrate that the nitrogen is incorporated into the fullerene interior and that, furthermore, it sustains completely its atomic ground state configuration [9]. The spherical symmetry of $N@C_{60}$, as, for example, derived from the absence of anisotropic hyperfine interaction, is interpreted as a strong indicator for the fixation of the nitrogen in the centre of [60] fullerene. In addition, the centre position is substantiated by potential energy calculations suggesting that the highly reactive nitrogen atom is trapped and shielded from the surroundings.

The noble-gas fullerene compounds have no chemical bond between the gas atom and the carbon atoms, yet they are also extremely stable, since the gas atom simply cannot escape from the fullerene cage. In this light, the recently introduced ^3He NMR spectroscopy of endohedral $\text{He}@C_{60}$ is bound to become a major experimental tool to study the structure and reactions of fullerenes [225, 226].

C1.2.14 CONCLUDING REMARKS

Recently, considerable advances have been made in the production of a different class of nanoscopic carbon structures, namely, carbon nanotubes, which stimulated fundamental research exploring the structure–property relationship of these materials [227]. In their simplest form carbon nanotubes are composed of only a single cylindrical graphene shell with a central hollow internal cavity. These structurally uniform cylinders are invariably sealed at both ends by bended carbon caps, which contain both five- and six-membered rings similar to the structures of fullerenes. Based on their similarity with highly graphitized carbonaceous materials, nanotubes have low chemical reactivity. Therefore, the chemistry of carbon nanotubes is mainly focused on opening reactions at its

caps to enable filling of the hollow cavity with electron conducting material. In contrast, the tubewalls are practically nonreactive. These materials are promising candidates for future applications ranging from catalysis to separation and storage technology to electronics and accordingly warrant appropriate attention.

REFERENCES

- [1] Kroto H W, Heath J R, O'Brien S C, Curl R F and Smalley R E 1985 C₆₀ Buckminsterfullerene *Nature* **318** 162–3
 - [2] Krätschmer W, Lamb L D, Fostiropoulos K and Huffman D R 1990 Solid C₆₀; a new form of carbon *Nature* **347** 354–8
 - [3] Smalley R E 1992 Self-assembly of the fullerenes *Accounts.Chem. Res.* **25** 98–105
 - [4] Kroto H W, Allaf A W and Balm S P 1991 C₆₀: Buckminster fullerene *Chem. Rev.* **91** 1213–35
 - [5] Kroto H W 1987 The stability of the fullerenes C_n, with n =24, 28, 32, 36, 50, 60, 70 *Nature* **329** 529–31
 - [6] Haddon R C 1988 π -electrons in three dimensions *Accounts Chem. Res.* **21** 243–9
 - [7] Haddon R C 1993 Chemistry of the fullerenes: manifestation of strain in a class of continuous aromatic molecules *Science* **261** 1545–50
 - [8] Morton J R, Negri F and Preston K F 1998 Addition of free radicals to C₆₀ *Accounts. Chem. Res.* **31** 63–9
 - [9] Mauser H, van Eikema Hommes N J R, Clark T, Hirsch A, Pietzak B, Weidinger A and Dunsch L 1997 Stabilization of atomic nitrogen inside C₆₀ *Angew. Chem. Int. Edn. Engl.* **36** 2835–8
 - [10] Taylor R, Hare J P, Abdul-Sada A K and Kroto H W 1990 Isolation, separation and characterization of the fullerenes C₆₀ and C₇₀, the 3rd form of carbon *J. Chem. Soc., Chem. Commun.* 1423–5
 - [11] Meijer G and Bethune D S 1990 Laser deposition of carbon clusters on surfaces—a new approach to the study of fullerenes *J. Chem. Phys.* **93** 7800–2
 - [12] Meijer G and Bethune D S 1990 Mass spectroscopic confirmation of the presence of C₆₀ in laboratory-produced carbon dust *Chem. Phys. Lett.* **175** 1–2
-

- [13] Campbell E B and Hertel I V 1992 Molecular beam studies of fullerenes *Carbon* **8** 1157–65
- [14] Diederich F, Ettl R, Rubin Y, Whetten R L, Beck R, Alvarez M, Anz S, Sensharma D, Wudl F, Khemani K C and Koch A 1991 The higher fullerenes: isolation and characterization of C₇₀, C₈₄, C₉₀, C₉₄, and C₇₀ and an oxide of D_{5h}-C₇₀ *Science* **252** 548–51
- [15] Hawkins J M, Nambu M and Meyer A 1994 Resolution and configurational stability of the chiral fullerenes C₇₆, C₇₈, and C₈₄. A limit for the activation energy of the Stone-Wales transformation *J. Am. Chem. Soc.* **116** 7642–5
- [16] Ettl R, Chao I, Diederich F and Whetten R L 1991 Isolation of C₇₆, a chiral (D₂) allotrope of carbon *Nature* **353** 149–53
- [17]

- Diederich F, Whetten R L, Thilgen C, Ettl R, Chao I and Alvarez M M 1991 Fullerene isomerism— isolation of C_{2v} - C_{78} and D_3 - C_{78} *Science* **254** 1768–70
- [18] Diederich F and Whetten R L 1992 Beyond C_{60} : the higher fullerenes *Accounts. Chem. Res.* **25** 119
- [19] David W I F, Ibberson R M, Dennis T J S, Hare J P and Prassides K 1992 Structural phase transitions in the fullerene C_{60} *Europhys. Lett.* **18** 219
- [20] David W I F, Ibberson R M, Matthewman J C, Prassides K, Dennis T J S, Hare J P, Kroto H W, Taylor R and Walton D R M 1991 Crystal structure and bonding of ordered C_{60} *Nature* **353** 147
- [21] Heiney P A, Fischer J, McGhie A R, Romanow W J, Denenstien A M, McCauley J P, Smith A B III and Cox D E 1991 Orientational ordering transition in solid C_{60} *Phys. Rev. Lett.* **67** 1468
- [22] Verheijen M A, Meekes H, Meijer G, Bennema P, De Boer J L, van Smaalen S, van Tendeloo G, Amelinckx S, Muto S and van Landuyt J 1992 The structure of different phases of pure C_{70} crystals *Chem. Phys.* **166** 287
- [23] van Tendeloo G, Amelinckx S, De Boer J L, van Smaalen S, Verheijen M A and Meijer G 1993 Structural phase transitions in C_{70} *Europhys. Lett.* **21** 329
- [24] Vaughan G B M, Heiney P A, Fischer J E, Luzzi D E, Ricketts-Foot D A, McGhie A R, Hui Y W, Smith A L, Cox D E, Romanow W J, Allen B H, Coustel N, McCauley J P and Smith A B III 1991 Orientational disorder in solvent-free solid C_{70} *Science* **254** 1350
- [25] Heiney P A 1993 Structure, dynamics and ordering transition of solid C_{60} *J. Phys. Chem. Solids* **53** 1333
- [26] Haddon R C, Brus L E and Raghavachari K 1986 Rehybridization and π -orbital alignment: the key to the existence of spheroidal carbon clusters *Chem. Phys. Lett.* **131** 165
- [27] Prassides K, Kroto H W, Taylor R, Walton D R M, David W I F, Tomkinson J, Haddon R C, Rosseinsky M J and Murphy D W 1992 Fullerenes and fullerites in the solid state: neutron scattering studies *Carbon* **8** 1277–86
- [28] Kelly M K, Etchegoin P, Fuch D, Krätschmer W and Fostiropoulos K 1992 Optical transitions of C_{60} films in the visible and ultraviolet from spectroscopic ellipsometry *Phys. Rev. B* **46** 4963–8
- [29] Reber C, Yee L, McKierman J, Zink J I, Williams R S, Tong W M, Ohlberg D A, Whetten R L and Diederich F 1991 Luminescence and absorption spectra of C_{60} films *J. Phys. Chem.* **95** 2127
-
- 22-
- [30] Saito A Y, Shinohara H, Kato M, Nagashima H, Ohkohchi M and Ando Y 1992 Electric conductivity and band gap of solid C_{60} under high pressure *Chem. Phys. Lett.* **186** 236
- [31] Echegoyen L and Echegoyen L E 1998 Electrochemistry of fullerenes and their derivatives *Accounts. Chem. Res.* **31** 593–601
- [32] Yang Y, Arias F, Echegoyen L, Chibante L P F, Flanagan S, Robertson A and Wilson L J 1995 Reversible fullerene electrochemistry: correlation with the HOMO–LUMO energy difference for C_{60} , C_{70} , C_{76} , C_{78} and C_{84} *J. Am. Chem. Soc.* **117** 7801–4
- [33] Piskoti C, Yarger J and Zettl A 1998 C_{36} , a new carbon solid *Nature* **393** 771–4
- [34] Hirsch A 1994 *The Chemistry of the Fullerenes* (Stuttgart: Thieme)

- [35] Prato M and Maggini M 1998 Fulleropyrrolidines: a family of full-fledged fullerene derivatives *Accounts. Chem. Res.* **31** 519–26
- [36] Hummelen J C, Knight B, Pavlovich J, Gonzalez R and Wudl F 1995 Isolation of the heterofullerene $C_{59}N$ as its dimer $(C_{59}N)_2$ *Science* **269** 1554–6
- [37] Hebard A F, Zhou O, Zhong Q, Fleming R M and Haddon R C 1995 C_{60} films on surface-treated silicon: recipes for amorphous and crystalline growth *Thin Solid Films* **257** 147–53
- [38] Tanigaki K, Kuroshima S and Ebbesen T W 1995 Crystal growth and structure of fullerene thin films *Thin Solid Films* **257** 154–65
- [39] Zhao W B, Zhang X D, Luo K J, Chen J, Ye Z Y, Zhang J L, Li C Y, Yin D L, Gu Z N, Zhou X H and Jin Z X 1993 Growth and structure of C_{60} thin films on NaCl, glass and mica substrates *Thin Solid Films* **232** 149–53
- [40] Ichihashi T, Tanigaki K, Ebbesen T W, Kuroshima S and Iijima S 1992 Structures of C_{60} thin films fabricated on alkali halide substrates by organic MBE *Chem. Phys. Lett.* **190** 179–83
- [41] Sakurai M, Tada H, Saiki K, Koma A, Funasaka H and Kishimoto Y 1993 Epitaxial growth of C_{60} and C_{70} films on GaSe (0001) and MoS_2 (0001) surfaces *Chem. Phys. Lett.* **208** 425–30
- [42] Koma A 1992 Van Der Waals epitaxy—a new epitaxial growth method for a highly lattice mismatched system *Thin Solid Films* **216** 72–6
- [43] Li Y Z, Chander M, Patrin J C, Weaver J H, Chibante L P F and Smalley R E 1992 Adsorption of individual C_{60} molecules on Si (111) *Phys. Rev. B* **45** 13 837–40
- [44] Dura J A, Pippenger P M, Halas N J, Xiong X Z, Chow P C and Moss S C 1993 Epitaxial integration of single crystal C_{60} *Appl. Phys. Lett.* **63** 3443–5
- [45] Li Y Z, Chander M, Patrin J C, Weaver J H, Chibante L P F and Smalley R E 1991 Order and disorder in C_{60} and K_xC_{60} multilayers, direct imaging with scanning tunneling microscopy *Science* **253** 429–33
- [46] Fartash A 1994 Growth and microstructure of interfacially oriented large crystalline grain C_{60} sheets *Appl. Phys. Lett.* **64** 1877–9
- [47] Fischer J E, Werwa E and Heiney P A 1993 Pseudo epitaxial C_{60} films prepared by a hot wall method *Appl. Phys. A* **56** 193–6

- [48] Altman E I and Colton R J 1993 Determination of the orientation of C_{60} adsorbed on Au(111) and Ag(111) *Phys. Rev. B* **48** 18 244–9
- [49] Altman E I and Colton R J 1993 The interaction of C_{60} with noble metal surfaces *Surf. Sci.* **295** 13–33
- [50] Hashizume T, Motai K, Wang X D, Shinohara H, Saito Y, Maruyama Y, Ohno K, Kawazoe Y, Nishina Y, Pickering H W, Kuk Y and Sakurai T 1993 Intramolecular structures of C_{60} molecules adsorbed on the Cu(111) surface *Phys. Rev. Lett.* **71** 2959–62
- [51] Joachim C, Gimzewski J K, Schlittler R R and Chavy C 1995 Electronic transparency of a single C_{60} molecule *Phys. Rev. Lett.* **74** 2102–5

- [52] Gimzewski J K, Modesti S and Schlittler R R 1994 Cooperative self-assembly of Au atoms and C₆₀ on Au(110) surfaces *Phys Rev. Lett.* **72** 1036–9
- [53] Stephens P W, Mihaly L, Lee P L, Whetten R L, Huang S-M, Kaner R, Diederich F and Holczer K 1991 Structure of single phase superconducting K₃C₆₀ *Nature* **351** 632
- [54] Holczer K and Whetten R L 1992 Superconducting and normal state properties of the A₃C₆₀ compounds *Carbon* **8** 1261–76
- [55] Rao C N R and Seshadri R 1994 Phase transitions, superconductivity and ferromagnetism in fullerene systems *MRS Bulletin* **12** 28–30
- [56] Rosseinsky M J 1995 Fullerene intercalation chemistry *J. Mater. Chem.* **5** 1497
- [57] Weaver J H 1992 Fullerenes and fullerides: photoemission and scanning tunneling microscopy studies *Accounts. Chem. Res.* **25** 143–9
- [58] Weaver J H 1992 Electronic structures of C₆₀, C₇₀ and the fullerides—photoemission and inverse photoemission studies *J. Phys. Chem. Solids* **53** 1433
- [59] Bethune D S, Meijer G, Tang W C and Rosen H J 1990 The vibrational Raman spectra of purified solid films of C₆₀ and C₇₀ *Chem. Phys. Lett.* **174** 219
- [60] Fischer J, Heiney P A and Smith A B III 1992 Solid-state chemistry of fullerene-based materials *Accounts. Chem. Res.* **25** 112
- [61] Stephens P W, Cox D, Lauher J W, Mihaly L, Wiley J B, Allemand P M, Hirsch A, Holczer K, Li Q, Thompson J D and Wudl F 1992 Lattice structure of the fullerene ferromagnet TDAE-C₆₀ *Nature* **355** 331
- [62] Allemand P M, Khemani K C, Koch A, Wudl F, Holczer K, Donovan S, Gruner G and Thompson J D 1991 Organic molecular soft ferromagnetism in a fullerene C₆₀ *Science* **253** 301
- [63] Chiang L Y, Wang L Y, Tseng S M, Wu J S and Heieh K H 1994 Fullerenol derived urethane-connected polyether dendritic polymers *J. Chem. Soc., Chem. Commun.* 2675–6
- [64] Amato I 1991 Doing chemistry in the round *Science* **254** 30–1
- [65] Suzuki T, Li Q, Khemani K C, Wudl F and Almarsson Ö 1992 Synthesis of meta-phenylene-phenylenebis(phenylfulleroids) and *p*-phenylenebis(phenylfulleroids)—2-pearl selections of pearl necklace polymers *J. Am. Chem. Soc.* **114** 7300–1

- [66] Suzuki T, Li Q, Khemani K C and Wudl F 1992 Dihydrofulleroid H₂C₆₁—synthesis and properties of the parent fulleroid *J. Am. Chem. Soc.* **114** 7301–2
- [67] Geckeler K E and Hirsch A 1993 Polymer-bound C₆₀ *J. Am. Chem. Soc.* **115** 3850–1
- [68] Benincori T, Brenna E, Sannicolo F, Trimarco L, Zotti G and Sozzani P 1996 The first 'charm bracelet' conjugated polymer: an electroconducting polythiophene with covalently bound fullerene moieties *Angew. Chem. Int. Edn. Engl.* **35** 648–51
- [69] Nagashima H, Kato Y, Satoh H, Kamegashima N, Itoh K, Oi K and Saito Y 1996 Thermal study of a silylmethylated fullerene leading to preparation of its vacuum deposited thin film *Chem. Lett.* 519–20
- [70] Loy D A and Assink R A 1992 Synthesis of a C₆₀-*p*-xylylene copolymer *J. Am. Chem. Soc.* **114** 3977–8

- [71] Ma B, Lawson G E, Bunker C E, Kitaygorodskiy A and Sun Y-P 1995 Fullerene-based macromolecules from photochemical reactions of [60]fullerene and triethylamine *Chem. Phys. Lett.* **247** 51–6
- [72] Zhou P, Dong Z H, Rao A M and Eklund P C 1993 Reaction mechanism for the photopolymerization of solid fullerene C₆₀ *Chem. Phys. Lett.* **211** 337–40
- [73] Wang Y, Holden J M, Dong Z H, Bi X X and Eklund P C 1993 Photodimerization kinetics in solid C₆₀ films *Chem. Phys. Lett.* **211** 341–5
- [74] Wang Y, Holden J M, Bi X X and Eklund P C 1994 Thermal decomposition of polymeric C₆₀ *Chem. Phys. Lett.* **217** 413–17
- [75] Eklund P C, Rao A M, Zhou P, Wang Y and Holden J M 1995 Photochemical transformations of C₆₀ and C₇₀ films *Thin Solid Films* **257** 185–203
- [76] Rao A M, Zhou P, Wang K-A, Hager G T, Holden J M, Wang Y, Lee W-T, Bi X-X, Eklund P C, Cornett D S, Duncan M A and Amster I J 1993 Photoinduced polymerization of solid C₆₀ films *Science* **259** 955
- [77] Iwasa Y, Arima T, Fleming R M, Siegrist T, Zhou O, Haddon R C, Rothberg L J, Lyons K B, Carter H L, Hebard A F, Tycko R, Dabbagh G, Krajewski J J, Thomas G A and Yagi T 1994 New phases of C₆₀ synthesized at high pressure *Science* **264** 1570
- [78] Stephens P W, Bortel G, Faigel G, Tegze M, Janossy A, Pekker S, Oszlanyi G and Forro L 1994 Polymeric fullerene chains in RbC₆₀ and KC₆₀ *Nature* **370** 636
- [79] Wang P, Shamsuzzoha M, Lee W-J, Wu X-L and Metzger R M 1993 Superconductivity in Langmuir–Blodgett multilayers of C₆₀ doped with potassium *Synth. Met.* **55** 3104–9
- [80] Williams G, Pearson C, Bryce M R and Petty M C 1992 Langmuir–Blodgett films of C₆₀ *Thin Solid Films* **209** 150–2
- [81] Williams G, Soi A, Hirsch A, Bryce M R and Petty M C 1993 Langmuir–Blodgett films of 1-*t*-butyl-9-hydrofullerene-60 *Thin Solid Films* **230** 73–7
- [82] Milliken J, Dominguez D D, Nelson H H and Barger W R 1992 Incorporation of C₆₀ in Langmuir–Blodgett films *Chem. Mater.* **4** 252–4
- [83] Nakamura T, Tachibana H, Yumara M, Matsumoto M, Azumi R, Tanaka M and Kawabata Y 1992 Formation of Langmuir–Blodgett films of a fullerene *Langmuir* **8** 4–6

- [84] Iwahashi M, Kikuchi K, Achiba Y, Ikemoto I, Araki T, Mochida T, Yokoi S-I, Tanaka A and Iriyama K 1992 Morphological study of thin-film systems of pure fullerene (C₆₀) and some other amphiphilic compounds on the electron microscopic scale *Langmuir* **8** 2980–4
- [85] Castillo R, Ramos S and Ruiz-Garcia J 1996 Direct observation of Langmuir films of C₆₀ and C₇₀ using Brewster angle microscopy *J. Phys. Chem.* **100** 15 235–41
- [86] Back R and Lennox R B 1992 C₆₀ and C₇₀ at the air-water interface *J. Phys. Chem.* **96** 8149–52
- [87] Wang P, Shamsuzzoha M, Wu X-L, Lee W-J and Metzger R M 1992 Order and disorder in C₆₀ Langmuir–Blodgett films: direct imaging by scanning tunneling microscopy and high-resolution transmission electron

microscopy *J. Phys. Chem.* **96** 9025–8

- [88] Wang P, Maruyama Y and Metzger R M 1996 Superconductivity of C₆₀ Langmuir–Blodgett films doped with potassium: low-field signal and electron spin resonance study *Langmuir* **12** 3932–7
- [89] Nakamura T, Tachibana H, Yumara M, Matsumoto M and Tagaki W 1993 Structure and physical properties of Langmuir–Blodgett films of C₆₀ with amphiphilic matrix molecules *Synth. Met.* **55** 3131–6
- [90] Bulhoes L O S, Obeng Y S and Bard A J 1993 Langmuir–Blodgett and electrochemical studies of fullerene films *Chem. Mater.* **5** 110–14
- [91] Jehoulet C, Obeng Y S, Kim Y-T, Zhou F and Bard A J 1992 Electrochemistry and Langmuir through studies of C₆₀ and C₇₀ films *J. Am. Chem. Soc.* **114** 4237–47
- [92] Xiao Y, Yao Z, Jin D, Yan F and Xue Q 1993 Mixed Langmuir–Blodgett films of C₆₀/AA *J. Phys. Chem.* **97** 7072–4
- [93] Goldenberg L M, Williams G, Bryce M R, Monkman A P, Petty M C, Hirsch A and Soi A 1993 Electrochemical studies on Langmuir–Blodgett films of 1-*tert*-butyl-1,9-dihydrofullerene-60 *J. Chem. Soc., Chem. Commun.* 1310–12
- [94] Zhou D, Gan L, Luo C, Tan H, Huang C, Yao G, Zhao X, Liu Z, Xia X and Zhang B 1996 Langmuir–Blodgett films and photophysical properties of a C₆₀-sarcosine methyl ester derivative *J. Phys. Chem.* **100** 3150–6
- [95] Jonas U, Cardullo F, Belik P, Diederich F, Gügel A, Harth E, Herrmann A, Isaacs L, Müllen K, Ringsdorf H, Thilgen C, Uhlmann P, Vasella A, Waldraff C A A and Walter M 1995 Synthesis of a fullerene[60] cryptate and systematic Langmuir–Blodgett and thin-film investigations of amphiphilic fullerene derivatives *Chem. Eur. J.* **1** 243–51
- [96] Guldi D M, Tian Y, Fendler J H, Hungerbühler H and Asmus K-D 1995 Stable monolayers and Langmuir–Blodgett films of functionalized fullerenes *J. Phys. Chem.* **99** 17 673–6
- [97] Hawker C J, Saville P M and White J W 1994 The synthesis and characterization of a self-assembling amphiphilic fullerene *J. Org. Chem.* **59** 3503–5
- [98] Isaacs L, Ehrlig A and Diederich F 1993 Improved purification of C₆₀ and formation of σ - and π -homoaromatic methano-bridged fullerenes by reaction with alkyl diazoacetates *Helv. Chim. Acta* **76** 1231–50
- [99] Leigh D A, Moody A E, Wade F A, King T A, West D and Bahra G S 1995 Second harmonic generation from Langmuir–Blodgett films of fullerene-aza-crown ethers and their potassium ion complexes *Langmuir* **11** 2334–6
- [100] Diederich F, Jonas U, Gramlich V, Herrmann A, Ringsdorf H and Thilgen C 1993 Synthesis of a fullerene derivative of benzo[18]crown-6 by Diels–Alder reaction: complexation ability, amphiphilic properties, and x-ray crystal structure of a dimethoxy-1,9-(methano[1, 2]benzomethano)fullerene[60] benzene clathrate *Helv. Chim. Acta* **76** 2445–53

- [101] Maggini M, Karlsson A, Pasimeni L, Scorrano G, Prato M and Valli L 1994 Synthesis of N-acylated fulleropyrrolidines: new materials for the preparation of Langmuir–Blodgett films containing fullerenes *Tetrahedron Lett.* **35** 2985–8
- [102] Matsumoto M, Tachibana H, Azumi R, Tanaka M, Nakamura T, Yunome G, Abe M, Yamago S and

- Nakamura E 1995 Langmuir–Blodgett film of amphiphilic C₆₀ carboxylic acid *Langmuir* **11** 660–5
- [103] Patel H M, Didymus J M, Wong K K W, Hirsch A, Skiebe A, Lamparth I and Mann S 1996 Fullerenes: interaction of divalent metal ions with Langmuir monolayers and multilayers in mono-substituted C₆₀-malonic acid *J. Chem. Soc., Chem. Commun.* 611–2
- [104] Maliszewskij N C, Heiney P A, Jones D R, Strongin R M, Cichy M A and Smith A B III 1993 Langmuir films of C₆₀, C₆₀O, and C₆₁H₂ *Langmuir* **9** 1439–41
- [105] Tian Y, Fendler J H, Hungerbühler H, Guldi D M and Asmus K-D 1999 Effects of hydrophobic–hydrophilic balance and stereochemistry on the supramolecular assembly of functionalized fullerenes *Supramol. Sci. C* **7** 67–73
- [106] Guldi D M, Tian Y, Fendler J H, Hungerbühler H and Asmus K-D 1996 Compression-dependent structural changes of functionalized fullerene monolayers *J. Phys. Chem.* **100** 2753–8
- [107] Nierengarten J-F, Schall C, Nicoud J-F, Heinrich B and Guillon D 1998 Amphiphilic cyclic fullerene bisadducts: synthesis and Langmuir films at the air–water interface *Tetrahedron Lett.* **39** 5747–50
- [108] Mirkin C A and Caldwell W B 1996 Thin film, fullerene-based materials *Tetrahedron* **52** 5113–30
- [109] Boulas P L, Gomez-Kaifer M and Echegoyen L 1998 Electrochemistry of supramolecular systems *Angew. Chem. Int. Edn. Engl.* **37** 216–47
- [110] Xie Q, Perez-Cordero E and Echegoyen L 1992 Electrochemical detection of and : enhanced stability of fullerides in solution *J. Am. Chem. Soc.* **114** 3978–80
- [111] Allemant P-M, Koch A, Wudl F, Rubin Y, Diederich F, Alvarez M M, Anz S J and Whetten R L 1991 Two different fullerenes have the same cyclic voltammetry *J. Am. Chem. Soc.* **113** 1051–2
- [112] Dubois D, Kadish K M, Flanagan S, Haufler R E, Chibante L P F and Wilson L J 1991 Spectroelectrochemical study of the C₆₀ and C₇₀ fullerenes and their mono-, di-, tri-, and tetraanions *J. Am. Chem. Soc.* **113** 4364–6
- [113] Bolskar R D, Mathur R S and Reed C A 1996 Synthesis and isolation of a fullerene carbocation *J. Am. Chem. Soc.* **118** 13 093–4
- [114] Bausch J W, Prakash G K S, Olah G A, Tse D S, Lorents D C, Bae Y K and Malhotra R 1991 Considered novel aromatic systems. 11. Diamagnetic polyanions of the C₆₀ and C₇₀ fullerenes. preparation, 13-C and 7-Li NMR spectroscopic observation, and alkylation with methyl iodide to polymethylated fullerenes *J. Am. Chem. Soc.* **113** 3205–6
- [115] Xie Q, Arias F and Echegoyen L 1993 Electrochemically-reversible, single-electron oxidation of C₆₀ and C₇₀ *J. Am. Chem. Soc.* **115** 9818–19
- [116] Suzuki T, Maruyama Y, Akasaka T, Ando W, Kobayashi K and Nagase S 1994 Redox properties of organofullerenes *J. Am. Chem. Soc.* **116** 1359–63
- [117] Arias F, Echegoyen L, Wilson S R, Lu Q Y and Lu Q 1995 Methanofullerenes and methanofulleroids have different electrochemical behavior at negative potentials *J. Am. Chem. Soc.* **117** 1422–7
-
- [118] Azamar-Barrios J A, Munoz E P and Penicaud A 1997 Electrochemical generation of the higher fullerene radicals C₇₆⁻, C₇₈⁻ and C₈₄⁻ under oxygen- and moisture-free conditions and their observation by EPR *J. Chem. Soc., Faraday Trans.* **93** 3119–23

- [119] Miller B, Rosamilia J M, Dabbagh G, Tycko R, Haddon R C, Muller A J, Wilson W, Murphy D W and Hebard A F 1991 Photoelectrochemical behavior of C_{60} films *J. Am. Chem. Soc.* **113** 6291–3
- [120] Zhang Y, Edens G and Weaver M J 1991 Potential-dependent surface Raman spectroscopy of Buckminsterfullerene films on gold: vibrational characteristics of anionic versus neutral C_{60} *J. Am. Chem. Soc.* **113** 9395–7
- [121] Goldenberg L M 1994 Electrochemical properties of Langmuir–Blodgett films *J. Electroanal. Chem.* **379** 3–19
- [122] Chlistunoff J, Cliffel D and Bard A J 1995 Electrochemistry of fullerene films *Thin Solid Films* **257** 166–84
- [123] Seger L, Wen L-Q and Schlenoff J B 1991 Prospects for using C_{60} and C_{70} in lithium batteries *J. Electrochem. Soc.* **138** L81–L83
- [124] Compton R G, Spackman R A, Wellington R G, Green M L H and Turner J 1992 A C_{60} modified electrode. Electrochemical formation of tetra-butylammonium salts of C_{60} anions *J. Electroanal. Chem. Interfacial Electrochem.* **327** 337–41
- [125] Ruoff R S, Tse D S, Malhotra R and Lorents D C 1993 Solubility of C_{60} in a variety of solvents *J. Phys. Chem.* **97** 3379–83
- [126] Guldi D M, Hungerbühler H and Asmus K-D 1995 Unusual redox behavior of a water soluble malonic acid derivative of C_{60} : evidence for possible cluster formation *J. Phys. Chem.* **99** 13 487–93
- [127] Guldi D M, Hungerbühler H and Asmus K-D 1997 Radiolytic reduction of a water-soluble fullerene cluster *J. Phys. Chem. A* **101** 1783–6
- [128] Bezmelnitsin V N, Eletsii A V and Stepanov E V 1994 Cluster origin of fullerene solubility *J. Phys. Chem.* **98** 6665–7
- [129] Andriesvsky G V, Klochkov V K, Karyakina E L and Mchedlov-Petrossyan N O 1999 Studies of aqueous colloidal solutions of fullerene C_{60} by electron microscopy *Chem. Phys. Lett.* **300** 392–6
- [130] Sundahl M, Andersson T, Nilsson K, Wennerstrom O and Westman G 1993 Clusters of C_{60} -fullerene in a water solution containing γ -cyclodextrin: a photophysical study *Synth. Met.* **55** 3252–7
- [131] Andersson T, Nilsson K, Sundahl M, Westman G and Wennerström O 1992 C_{60} embedded in γ -cyclodextrin: a water-soluble fullerene *J. Chem. Soc., Chem. Commun.* 604–6
- [132] Williams R M and Verhoeven J W 1992 Supramolecular encapsulation of C_{60} in a water soluble calixarene: a core–shell charge transfer complex *Recl. Trav. Chim. Pays-Bas* **111** 531–2
- [133] Guldi D M 1997 Capped fullerenes: stabilization of water-soluble fullerene monomers as studied by flash photolysis and pulse radiolysis *J. Phys. Chem. A* **101** 3895–900
- [134] Beeby A, Eastoe J and Crooks E R 1996 Remarkable stability of C_{60} in micelles *Chem. Commun.* 901–2
- [135] Eastoe J, Crooks E R, Beeby A and Heenan R K 1995 Structure and Photophysics in C_{60} -micellar solutions *Chem. Phys. Lett.* **245** 571–7
-

- [136] Yamakoshi Y N, Yagami T, Fukuhara K, Sueyoshi S and Miyata N 1994 Solubilization of fullerenes into water with polyvinylpyrrolidone applicable to biological tests *J. Chem. Soc., Chem. Commun.* 517–18

- [137] Niu S and Mauzerall D 1996 Fast and efficient charge transport across a lipid bilayer is electronically mediated by C₇₀ fullerene aggregates *J. Am. Chem. Soc.* **118** 5791–5
- [138] Hwang K C and Mauzerall D C 1993 Photoinduced electron transport across a lipid bilayer mediated by C₇₀ *Nature* **361** 138–40
- [139] Hwang K C and Mauzerall D C 1992 Vectorial electron transfer from an interfacial photoexcited porphyrin to ground-state C₆₀ and C₇₀ and from ascorbate to triplet C₆₀ and C₇₀ in a lipid bilayer *J. Am. Chem. Soc.* **114** 9705–6
- [140] Bensasson R V, Bienvenue E, Dellinger M, Leach S and Seta P 1994 C₆₀ in model biological systems. A visible–UV absorption study of solvent-dependent parameters and solute aggregation *J. Phys. Chem.* **98** 3492–5000
- [141] Jensen A W, Wilson S R and Schuster D I 1996 Biological applications of fullerenes—a review *Bioorg. Med. Chem.* **4** 767–79
- [142] Leach S, Vervloet M, Despres A, Brcheret E, Hare P, Dennis T J S, Kroto H W, Taylor R and Walton D R M 1992 Electronic spectra and transitions of the fullerene C₆₀ *Chem. Phys.* **160** 451–66
- [143] Sun Y-P, Wang P and Hamilton N B 1993 Fluorescence spectra and quantum yields of Buckminsterfullerene (C₆₀) in room-temperature solutions. No excitation wavelength dependence *J. Am. Chem. Soc.* **115** 6378–81
- [144] Ebbesen T W, Tanigaki K and Kuroshima S 1991 Excited-state properties of C₆₀ *Chem. Phys. Lett.* **181** 501–4
- [145] Tanigaki K, Ebbesen T W and Kuroshima S 1991 Picosecond and nanosecond studies of the excited state properties of C₇₀ *Chem. Phys. Lett.* **185** 189–92
- [146] Zeng Y, Biczok L and Linschitz H 1992 External heavy atom induced phosphorescence emission of fullerenes: the energy of triplet C₆₀ *J. Phys. Chem.* **96** 5237–9
- [147] Arbogast J W, Darmanyan A P, Foote C S, Rubin Y, Diederich F N, Alvarez M M, Anz S J and Whetten R L 1991 Photophysical properties of C₆₀ *J. Phys. Chem.* **95** 11–12
- [148] Arbogast J S and Foote C S 1991 Photophysical properties of C₇₀ *J. Am. Chem. Soc.* **113** 8886–9
- [149] Guldi D M, Huie R E, Neta P, Hungerbühler H and Asmus K-D 1994 Excitation of C₆₀, solubilized in water by Triton X-100 and γ -cyclodextrin, and subsequent charge separation via reductive quenching *Chem. Phys. Lett.* **223** 511–16
- [150] Ausman K D, Benedetto A F, Samuelsand D A and Weisman R B 1998 *Recent Advances in the Chemistry of Fullerenes and Related Materials (The Electrochemical Society Proceedings Series)* vol 6, ed K M Kadish and R S Ruoff (Pennington, NJ: The Electrochemical Society) p 281
- [151] Goudsmit G H and Paul H 1993 Time-resolved EPR investigation of triplet state C₆₀. Triplet–triplet annihilation, CIDEP, and quenching by nitroxide radicals *Chem. Phys. Lett.* **208** 73–8
- [152] Guldi D M and Asmus K-D 1997 Photophysical properties of mono- and multiply-functionalized fullerene derivatives *J. Phys. Chem. A* **101** 1472–81

- [153] McGlashen M L, Blackwood M E and Spiro T G 1993 Resonance Raman spectroelectrochemistry of the C₆₀ radical anion *J. Am. Chem. Soc.* **115** 2074–5

- [154] Guldi D M, Hungerbühler H and Asmus K-D 1999 Inhibition of cluster phenomena in truly water soluble fullerene derivatives: bimolecular electron and energy transfer processes *J. Phys. Chem. A* **103** 1444–53
- [155] Li C, Si J, Yang M, Wang R and Zhang L 1995 Excited-state nonlinear absorption in multi-energy-level molecular systems *Phys. Rev. A* **51** 569–75
- [156] Boutorine A S, Tokuyama H, Takasugi M, Isobe H, Nakamura E and Helene C 1994 Fullerene–oligonucleotide conjugates—photoinduced sequence-specific DNA cleavage *Angew. Chem.* **106** 2526–9
- [157] Irie K, Nakamura Y, Ohigashi H, Tokuyama H, Yamago S and Nakamura E 1996 Photocytotoxicity of water-soluble fullerene derivatives *Biosci. Biotech. Biochem.* **60** 1359–61
- [158] Tokuyama H, Yamago S, Nakamura E, Shiraki T and Sugiura Y 1993 Photoinduced biochemical activity of fullerene carboxylic acid *J. Am. Chem. Soc.* **115** 7918–9
- [159] Sijbesma R, Srdanov G, Wudl F, Castoro J A, Wilkins C, Friedman S H, DeCamp D L and Kenyon G L 1993 Synthesis of a fullerene derivative for the inhibition of HIV enzymes *J. Am. Chem. Soc.* **115** 6510–12
- [160] Friedman S H, DeCamp D L, Sijbesma R, Srdanov G and Wudl F 1993 Inhibition of HIV-1 protease by fullerene derivatives: model building studies and experimental verification *J. Am. Chem. Soc.* **115** 6506–9
- [161] Chiang L Y, Lu F-J and Lin J-T 1995 Free radical scavenging activity of water-soluble fullerenols *J. Chem. Soc., Chem. Commun.* 1283–4
- [162] Yamakoshi Y, Sueyoshi S, Fukuhara K and Miyata N 1998 $\cdot\text{OH}$ and $\text{O}_2^{\cdot-}$ generation in aqueous C_{60} and C_{70} solutions by photoirradiation: an ESR study *J. Am. Chem. Soc.* **120** 12 363–4
- [163] Bernstein R, Prat F and Foote C S 1999 On the mechanism of DNA cleavage by fullerenes *J. Am. Chem. Soc.* **121** 464–5
- [164] Tutt L W and Kost A 1992 Optical limiting performance of C_{60} and C_{70} solutions *Nature* **356** 225
- [165] Klimov V, Smilowitz L, Wang H, Grigorova M, Robinson J M, Koskela A, Mattes B R, Wudl F and Mc Branch D W 1997 Femtosecond to nanosecond dynamics in fullerenes: implications for excited-state optical nonlinearities *Res. Chem. Intermed.* **23** 587–600
- [166] Baumgarten M, Gügel A and Ghergel L 1993 EPR and optical absorption spectra of reduced Buckminsterfullerene *Adv. Mater.* **5** 458–61
- [167] Kato T, Kodama T, Shida T, Nakagawa T, Matsui Y, Suzuki S, Shiromaru H, Yamauchi K and Achiba Y 1991 Electronic absorption spectra of the radical anions and cations of fullerenes: C_{60} and C_{70} *Chem. Phys. Lett.* **180** 446–50
- [168] Guldi D M, Hungerbühler H, Janata E and Asmus K-D 1993 Radical-induced redox and addition reactions with C_{60} studied by pulse radiolysis *J. Chem. Soc., Chem. Commun.* **6** 84
- [169] Stasko A, Brezova V, Biskupic S, Dinse K-P, Schweitzer P and Baumgarten M 1995 EPR study of fullerene radicals generated in photosensitized TiO_2 suspensions *J. Phys. Chem.* **99** 8782–9
- [170] Stasko A, Brezova V, Rapta P, Biskupic D, Dinse K-P and Gügel A 1997 Anion radical of C_{60} fullerenes. An EPR study *Res. Chem. Intermed.* **23** 453–78

- [171] Schell S A J, Mehran F, Eaton G R, Eaton S S, Viehbeck A, O'Toole T R and Brown C A 1992 Electron spin relaxation times of $\text{C}_{60}^{\cdot-}$ in solution *Chem. Phys. Lett.* **195** 225–32
- [172] Khaled M M, Carlin R T, Trulove P C, Eaton G R and Eaton S S 1994 Electrochemical generation and

EPR studies of C_{60}^- , C_{60}^{2-} , C_{60}^{3-} *J. Phys. Chem.* **98** 3465–74

- [173] Guldi D M, Liu D and Kamat P V 1997 Excited state and reduced and oxidized forms of $C_{76}(D_2)$ and $C_{78}(C_{2v})$ *J. Phys. Chem. A* **101** 6195–201
- [174] Imahori H, Hagiwara K, Akiyama T, Aoki M, Taniguchi S, Okada T, Shirakawa M and Sakata Y 1996 The small reorganization energy of C_{60} in electron transfer *Chem. Phys. Lett.* **263** 545–50
- [175] Imahori H and Sakata Y 1997 Donor-linked fullerenes: photoinduced electron transfer and its potential application *Adv. Mater.* **9** 537–46
- [176] Prato M 1997 [60]fullerene chemistry for materials science applications *J. Mater. Chem.* **7** 1097–109
- [177] Martí n N, Sánchez L, Illescas B and P´erez I 1998 C_{60} -based electroactive organofullerenes *Chem. Rev.* **98** 2527
- [178] Guldi D M and Asmus K-D 1997 Electron transfer from $C_{76}(D_2)$ and $C_{78}(C_{2v})$ to radical cations of various arenes: evidence for the Marcus inverted region *J. Am. Chem. Soc.* **119** 5744–5
- [179] Williams R M, Zwier J M and Verhoeven J W 1995 Photoinduced intramolecular electron transfer in a bridged C_{60} (acceptor)-aniline (donor) system. Photophysical properties of the first 'active' fullerene diad *J. Am. Chem. Soc.* **117** 4093–9
- [180] Williams R M, Koeberg M, Lawson J M, An Y-Z, Rubin Y, Paddon-Row M N and Verhoeven J W 1996 Photoinduced electron transfer to C_{60} across extended 3- and 11 σ -bond hydrocarbon bridges: creation of a long-lived charge-separated state *J. Org. Chem.* **61** 5055–62
- [181] Guldi D M, Maggini M, Scorrano G and Prato M 1997 Intramolecular electron transfer in fullerene/ferrocene based donor-bridge-acceptor dyads *J. Am. Chem. Soc.* **119** 974–80
- [182] Llacay J, Veciana J, Vidal-Gancedo J, Bourdelande J L, Gonzalez-Moreno R and Rovira C 1998 Persistent and transient open-shell species derived from C_{60} -TTF cyclohexane-fused dyads *J. Org. Chem.* **63** 5201–10
- [183] Dietel E, Hirsch A, Zhou J and Rieker A 1998 Synthesis and electrochemical investigations of molecular architectures involving C_{60} and tetraphenylporphyrin as building blocks *J. Chem. Soc., Perkin Trans. 2* 1357–64
- [184] Baran P S, Monaco R R, Khan A U, Schuster D I and Wilson S R 1997 Synthesis and cation-mediated electronic interactions of two novel classes of porphyrin-fullerene hybrids *J. Am. Chem. Soc.* **119** 8363–4
- [185] Higashida S, Imahori H, Kaneda T and Sakata Y 1998 Synthesis and photophysical behavior of porphyrins with two C_{60} units *Chem. Lett.* 605–6
- [186] Imahori H, Hagiwara K, Akiyama T, Taniguchi S, Okada T and Sakata Y 1995 Synthesis and photophysical property of porphyrin-linked fullerene *Chem. Lett.* 265–6
- [187] Akiyama T, Imahori H, Ajawakom A and Sakata Y 1996 Synthesis and self assembly of porphyrin-linked fullerene on gold surface using S–Au linkage *Chem. Lett.* 907–8

- [188] Bell T D M, Smith T A, Ghiggino K P, Ranasinghe M G, Shephard M J and Paddon-Row M N 1997 Long-lived photoinduced charge separation in a bridged C_{60} -porphyrin dyad *Chem. Phys. Lett.* **268** 223–8

- [189] Imahori H, Hagiwara K, Aoki M, Akiyama T, Taniguchi S, Okada T, Shirakawa M and Sakata Y 1996 Linkage and solvent dependence of photoinduced electron transfer in porphyrin-C₆₀ dyads *J. Am. Chem. Soc.* **118** 11 771–82
- [190] Kuciauskas D, Lin S, Seely G R, Moore A L, Moore T A, Gust D, Drovetskaya T, Reed C A and Boyd P D W 1996 Energy and photoinduced electron transfer in porphyrin-fullerene dyads *J. Phys. Chem.* **100** 15 926–32
- [191] Liddell P A, Sumida J P, Macpherson A N, Noss L, Seely G R, Clark K N, Moore A L, Moore T A and Gust D 1994 Preparation and photophysical studies of porphyrin-C₆₀ dyads *Photochem. Photobiol.* **60** 537–41
- [192] Liddell P A, Kuciauskas D, Sumida J P, Nash B, Nguyen D, Moore A L, Moore T A and Gust D 1997 Photoinduced charge separation and charge recombination to a triplet state in a carotene-porphyrin-fullerene triad *J. Am. Chem. Soc.* **119** 1400–5
- [193] Shephard M J and Paddon-Row M N 1996 Conformational analysis of C₆₀ ball and chain molecules: a molecular orbital study *Aust. J. Chem.* **49** 395–403
- [194] Carbonera D, DiValentin M, Corvaja C, Agostini G, Giacometti G, Liddell P A, Kuciauskas D, Moore A L, Moore T A and Gust D 1998 EPR investigation of photoinduced radical pair formation and decay to a triplet state in a carotene-porphyrin-fullerene triad *J. Am. Chem. Soc.* **120** 4398–405
- [195] Maggini M, Dono A, Scorrano G and Prato M 1995 Synthesis of a [60]fullerene derivative covalently linked to a ruthenium (II) tris(bipyridine) complex *J. Chem. Soc., Chem. Commun.* 845–6
- [196] Armspach D, Constable E C, Diederich F, Housecroft C E and Nierengarten J-F 1996 Bucky-ligands: fullerene-substituted oligopyridines for metallosupramolecular chemistry *J. Chem. Soc., Chem. Commun.* 2009–10
- [197] Armspach D, Constable E C, Diederich F, Housecroft C E and Nierengarten J-F 1998 Bucky ligands: synthesis, ruthenium(II) complexes, and electrochemical properties *Chem. Eur. J.* **4** 723–33
- [198] Edelman F T 1995 Filled buckyballs—recent developments from the endohedral metallofullerenes of lanthanides *Angew. Chem. Int. Edn. Engl.* **34** 981–5
- [199] Nagase S, Kobayashi K and Akasaka T 1996 Endohedral metallofullerenes: new spherical cage molecules with interesting properties *Bull. Chem. Soc. Japan* **69** 2131–42
- [200] Rubin Y 1997 Organic approaches to endohedral metallofullerenes: cracking open or zipping up carbon shells *Chem. Eur. J.* **3** 1009–16
- [201] Chai Y, Guo C, Jin R E, Hauffer R E, Chibante L P F, Fure J, Wang L, Alford J M and Smalley R E 1991 Fullerenes with metals inside *J. Phys. Chem.* **95** 7564
- [202] Yannoni C S, Hoinks M, deVries H M S, Bethune D S, Salem J R, Crowder M S and Johnson R D 1992 Scandium clusters in fullerene cages *Science* **256** 1191
- [203] Weaver J H, Chai Y, Kroll G H, Ohno T R, Hauffer R E, Guo T, Alford J M, Conceicao J, Chibante L P F, Jain A, Palmer G and Smalley R E 1992 XPS probes of carbon caged metals *Chem. Phys. Lett.* **190** 460
- [204] Kubozono Y, Hiraoka K, Tababayashi Y, Nakai T, Ohta T, Maeda H, Ishida H, Kashino S, Emura S, Ukita S and Sogabe T 1996 Enrichment of Ce@C₆₀ by HPLC technique *Chem. Lett.* 1061–2

- [205] Kubozono Y, Maeda H, Takabayashi Y, Hiraoka K, Nakai T, Kashino S, Emura S, Ukita S and Sogabe T 1996 Extractions of Y@C₆₀, Ba@C₆₀, La@C₆₀, Ce@C₆₀, Pr@C₆₀, Nd@C₆₀ and Gd@C₆₀ with aniline *J. Am. Chem. Soc.* **118** 6998–9
- [206] Shinohara H, Inakuma M, Hayashi N, Sato H, Saito Y, Kato T and Bandow S 1994 Spectroscopic properties of isolated Sc₃@C₈₂ metallofullerene *J. Phys. Chem.* **98** 8597–9

- [207] Shinohara H, Yamaguchi H, Hayashi N, Sato H, Ohno M, Ando Y and Saito Y 1993 Isolation and spectroscopic properties of $\text{Sc}_2@C_{74}$, $\text{Sc}_2@C_{82}$ and $\text{Sc}_2@C_{84}$ *J. Phys. Chem.* **97** 4259–61
- [208] Cagle D W, Kennel S J, Mirzadeh S, Alford J M and Wilson L J 1999 In vivo studies of fullerene-based materials using endohedral metallofullerene radiotracers *Proc. Natl Acad. Sci. USA* **96** 5182–7
- [209] Suzuki T, Maruyama Y, Kato T, Kikuchi K, Achiba Y, Kobayashi K and Nagase S 1995 Electrochemistry and ab-initio study of the dimetallofullerene $\text{La}_2@C_{80}$ *Angew. Chem. Int. Edn. Engl.* **34** 1094–6
- [210] Kikuchi K, Suzuki S, Nakao N, Nakahara N, Wakabayashi T, Shiromaru H, Saito K, Ikemoto I and Achiba Y 1993 Isolation and characterization of the metallofullerene $\text{La}@C_{82}$ *Chem. Phys. Lett.* **216** 67–71
- [211] Kikuchi K, Nakao Y, Suzuki S, Achiba Y, Suzuki T and Maruyama Y 1994 Characterization of the isolated $\text{Y}@C_{82}$ *J. Am. Chem. Soc.* **116** 9367–8
- [212] Akasaka T, Nagase S, Kobayashi K, Suzuki T, Kato K, Yamamoto K, Funasaka H and Takahashi T 1995 Exohedral derivatization of an endohedral metallofullerene $\text{Gd}@C_{82}$ *J. Chem. Soc., Chem. Commun.* 1343–4
- [213] Kirbach W and Dunsch L 1996 The existence of stable $\text{Tm}@C_{82}$ isomers *Angew. Chem. Int. Edn. Engl.* **35** 2380–3
- [214] Tellgmann R C, Krawez N, Lin S-H, Campbell E E B and Hertel I V 1996 Endohedral fullerene production *Nature* **382** 407
- [215] Takata M, Umeda B, Nishibori E, Sakata M, Saito Y, Ohno M and Shinohara H 1995 Confirmation by x-ray diffraction of the endohedral nature of the metallofullerene $\text{Y}@C_{82}$ *Nature* **377** 46
- [216] Akasaka T, Kato T, Kobayashi K, Nagase S, Yamamoto K, Funasaka H and Takahashi T 1995 Exohedral adducts of $\text{La}@C_{82}$ *Nature* **374** 600
- [217] Campbell E E B, Couris S, Fanti M, Kououmas E, Krawez N and Zerbetto F 1999 Third-order susceptibility of $\text{Li}@C_{82}$ *Adv. Mater.* **11** 405–8
- [218] Akasaka T, Nagase S, Kobayashi K, Suzuki T K, Kikuchi K, Achiba Y, Yamamoto K, Funasaka H and Takahashi T 1995 Synthesis of the first adducts of the dimetallofullerene $\text{La}_2@C_{80}$ and $\text{Sc}_2@C_{84}$ by addition of a disilirane *Angew. Chem. Int. Edn. Engl.* **34** 2139
- [219] Xu Z, Nakane T and Shinohara H 1996 Production and isolation of $\text{Ca}@C_{82}$ (I–IV) and $\text{Ca}@C_{84}$ (I,II) metallofullerenes *J. Am. Chem. Soc.* **118** 11 309
- [220] Wan T S M, Zhang H-W, Nakane T, Xu Z, Inakuma M, Shinohara H, Kobayashi K and Nagase S 1998 Production, isolation and electronic properties of missing fullerenes: $\text{Ca}@C_{72}$ and $\text{Ca}@C_{74}$ *J. Am. Chem. Soc.* **120** 6806–7
- [221] RübSam M, Schweitzer P and Dinse K-P 1996 Rotational dynamics of metallo-endofullerenes *J. Phys. Chem.* **100** 19 310
-
- [222] Johnson R D, de Vries M S, Salem J, Bethune D S and Yannoni C S 1992 Electron-paramagnetic resonance studies of lanthanum containing C_{82} *Nature* **355** 239
- [223] Saito Y, Yokoyama S, Inakuma M and Shinohara H 1996 An ESR study of the formation of $\text{La}@C_{82}$ isomers in arc synthesis *Chem. Phys. Lett.* **250** 80

- [224] Murphy T A, Pawlik T, Weidinger A, Höhne M, Alcalá R and Spaeth J M 1996 Observation of atomlike nitrogen in nitrogen-implanted solid C_{60} *Phys. Rev. Lett.* **77** 1075
- [225] Saunders M, Jimenez-Vazquez H A, Cross R J and Poreda R J 1993 Stable compounds of helium and neon at the cost of C_{60} and Ne at the cost of C_{70} *Science* **259** 1428
- [226] Shabtai E, Weitz A, Haddon R C, Hoffman R E, Rabinovitz M, Khong A, Cross R J, Saunders M, Cheng P-C and Scott L 1998 ^3He NMR of $\text{He}@C_{60}^{5-}$ and $\text{He}@C_{70}^{5-}$. New records for the most shielded and the most deshielded ^3He inside a fullerene *J. Am. Chem. Soc.* **120** 6389–93
- [227] Ebbesen T W 1997 *Carbon Nanotubes—Preparation and Properties* (Boca Raton, FL: Chemical Rubber Company)
-

FURTHER READING

- Diederich F and Thilgen C 1996 Covalent fullerene chemistry *Science* **271** 317–23
- Echegoyen L and Echegoyen L E 1998 Electrochemistry of fullerenes and their derivatives *Accounts Chem. Res.* **31** 593–601
- Hirsch A 1994 *The Chemistry of the Fullerenes* (Stuttgart: Thieme)
- Hirsch A (ed) 1999 *Fullerenes and Related Structures (Topics in Current Chemistry 199)* (Berlin: Springer)
- Imahori H and Sakata Y 1997 Donor-linked fullerenes: photoinduced electron transfer and its potential application *Adv. Mater.* **9** 537–46
- Jensen A W, Wilson S R and Schuster D I 1996 Biological applications of fullerenes—a review *Bioorg. Med. Chem.* **4** 767–79
- Martín N, Sánchez L, Illescas B and Pérez I 1998 C_{60} -based electroactive organofullerenes *Chem. Rev.* **98** 2527
- Prato M 1997 [60]fullerene chemistry for materials science applications *J. Mater. Chem.* **7** 1097–109
- Prato M and Maggini M 1998 Fulleropyrrolidines: a family of full-fledged fullerene derivatives *Accounts Chem. Res.* **31** 519–26
- Rosseinsky M J 1995 Fullerene intercalation chemistry *J. Mater. Chem.* **5** 1497
-

-1-

C1.3 Van der Waals molecules

Jeremy M Hutson

C1.3.1 INTRODUCTION

The attractive forces between pairs of atoms or molecules are almost always strong enough to support bound vibrational states. The resulting molecular complexes, or *Van der Waals molecules*, are very weakly bound, and are easily destroyed by collisions with other molecules. They exist in small but significant concentrations in gases and gas mixtures: for example, in Ar gas at 120 K and 1 bar, about 0.4% of the atoms are present as bound dimers [1]. The dimer concentrations decrease with increasing temperature, but are larger for systems with stronger attractive forces, such as most systems containing polar molecules.

Van der Waals complexes can be observed spectroscopically by a variety of different techniques, including microwave, infrared and ultraviolet/visible spectroscopy. Their existence is perhaps the simplest and most direct demonstration that there are attractive forces between stable molecules. Indeed the spectroscopic properties of Van der Waals complexes provide one of the most detailed sources of information available on intermolecular forces, especially in the region around the potential minimum. The measured rotational constants of Van der Waals complexes provide information on intermolecular distances and orientations, and the frequencies of bending and stretching vibrations provide information on how easily the complex can be distorted from its equilibrium conformation. In favourable cases, the whole of the potential well can be mapped out from spectroscopic data.

Studies of Van der Waals complexes have provided a wealth of information on the properties of weak non-chemical bonds. They have allowed the determination of complete potential energy surfaces for small systems, and have thrown considerable light on the nature of the hydrogen bond. Studies of larger clusters have begun to provide information of relevance to the liquid state, and to explain the behaviour of hydrogen-bonded networks. Studies of clusters containing reactive species are now starting to throw new light on the dynamics of chemical reactions. All these examples will be discussed in more detail below.

C1.3.2 TYPES OF SPECTROSCOPY

The spectroscopic signatures of Van der Waals complexes were first observed by Vodar and co-workers in the late 1950s [2], as broad features in the ‘missing Q-branch’ regions of the spectra of hydrogen halides and their mixtures at high pressures. Rank *et al* [3] subsequently observed irregular but fairly sharp features between the monomer vibration–rotation lines at lower pressures. The lines were attributed to dimers because their intensity increased quadratically with the gas pressure. Many attempts were made to obtain resolved spectra that could be analysed reliably. However, the basic problem at the time was that, at the pressures needed to obtain substantial concentrations of dimers, there was so much pressure broadening that the monomer lines swamped the spectra of the dimers. As the pressure was reduced, the dimer lines became somewhat sharper and the accessible region between the monomer lines increased, but the dimer spectra soon became too weak to observe above the noise.

-2-

Greater success was achieved for complexes formed from homonuclear diatoms such as H_2 and N_2 , for which monomer infrared transitions are forbidden by dipole selection rules. Complexes containing such molecules do have an infrared spectrum, because the quadrupole moment of the homonuclear monomer creates an electric field, which induces a dipole moment in the attached atom or molecule. In particular, rotationally resolved spectra were obtained for several complexes formed from H_2 and rare gases [4, 5]; these complexes have unusually large rotational constants because of the presence of H_2 . Spectra were also obtained for analogous complexes containing D_2 and HD. They were used to obtain a succession of intermolecular potentials for H_2 –Ar, H_2 –Kr and H_2 –Xe, of steadily increasing sophistication and accuracy [6, 7 and 8]. It was possible to determine not only the radial dependence of the potential, but also its dependence on intermolecular angle θ (anisotropy) and on the monomer bond length r .

A major breakthrough in the spectroscopy of Van der Waals complexes came in 1972, when Dyke, Howard and Klemperer [9] succeeded in measuring the microwave spectrum of $(HF)_2$ in a molecular beam. Such experiments are described in detail in the chapter on *Jet Spectroscopy*. In a typical molecular beam spectroscopy experiment, gas at around 1 bar pressure is expanded into a vacuum through a nozzle of aperture around 50 μm . Under these conditions, the gas molecules undergo many collisions during the expansion, and the collisions equalize the velocities of the different molecules. This is often referred to as a *supersonic* expansion. At the end of the expansion, nearly all of the random thermal energy of the gas molecules has been converted into ordered translational motion of the beam; all the molecules have almost the same velocity, and the relative velocities are very low. The beam itself, beyond the expansion region, is a nearly collision-free environment. The low relative velocities correspond to very low effective translational temperatures: temperatures of 1–10 K are common in molecular beam spectroscopy experiments. The populations of rotational and vibrational levels do not relax as fast as translation; rotational and vibrational distributions are sometimes characterized by higher temperatures, or may

not follow a Boltzmann distribution at all. In most cases, the rotational distribution is nearly as cold as the translational distribution, while the vibrational temperature is somewhere between the translational temperature and the source temperature. In any case, the effective temperature of the beam is low enough for large concentrations of complexes to be formed, and the complexes are mostly in their ground vibrational state.

C1.3.2.1 MICROWAVE SPECTROSCOPY

Klemperers' early microwave experiments on Van der Waals complexes used molecular beam electric resonance (MBER) spectroscopy, which relies on the fact that a beam of molecules with permanent dipole moments can be focused by an inhomogeneous electric field. Because of the Stark effect, the energy of a dipolar molecule in an electric field depends on the field strength. A MBER spectrometer usually uses a quadrupolar field; the beam is passed down the centre of a set of four parallel rods with alternating positive and negative voltages. This arrangement gives no electric field along the central axis, but the field increases linearly away from the axis. For a molecule in a state with a positive quadratic Stark effect, the energy thus increases quadratically with displacement from the centre; this creates a linear restoring force, so that a beam of such molecules is focused by the rods.

A MBER spectrometer is shown schematically in [figure C1.3.1](#). The technique relies on using two inhomogeneous electric fields, the A and B fields, to focus the beam. Since the Stark effect is different for different rotational states, the A and B fields can be set up so that a particular rotational state (with a positive Stark effect) is focused onto the detector. In MBER spectroscopy, the molecular beam is irradiated with microwave or radiofrequency radiation in the

-3-

central region (C field); any molecule which undergoes a rotational transition to a state with a negative Stark effect will not be focused by the B field, and will not reach the detector. Thus a transition is detected by monitoring the molecular beam flux as a function of the irradiation frequency. A mass spectrometer is usually used as the beam detector, allowing straightforward discrimination between monomers and Van der Waals complexes.

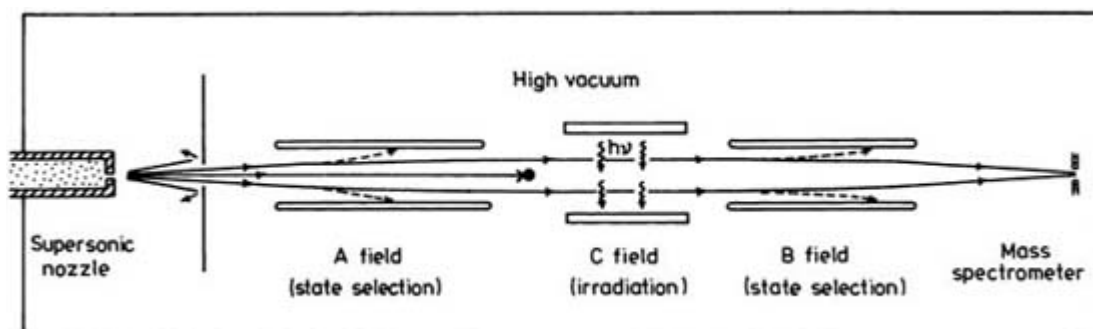


Figure C1.3.1. Schematic diagram of a molecular beam electric resonance spectrometer. (Taken from [60].)

An enormous range of complexes has been observed by MBER spectroscopy. Almost any pair of volatile compounds can be expanded through a nozzle and made to form complexes. In spectroscopic experiments, it is common to use around 1% of the sample gases in a buffer gas such as Ar, in order to minimize the formation of trimers and larger complexes. The restrictions on the technique are that the constituent molecules must be volatile and that the complex must have a dipole moment. In addition, for very large complexes, even the temperature of a molecular beam is high enough that a large number of rotational states are populated. The sensitivity of MBER relies on being able to deplete the beam intensity significantly by removing the molecules in a single rotational state. Thus, if the rotational partition function is very high, MBER may not be sensitive enough for spectra to be detected.

A few Van der Waals complexes have been observed using the analogous technique of molecular beam *magnetic* resonance, in which the molecules are focused using a magnetic rather than an electric field.

An alternative approach to obtaining microwave spectroscopy is Fourier transform microwave (FTMW) spectroscopy in a molecular beam [10]. This may be considered as the microwave analogue of Fourier transform NMR spectroscopy. The molecular beam passes into a Fabry–Perot cavity, where it is subjected to a short microwave pulse (of a few milliseconds duration). This creates a macroscopic polarization of the molecules. After the microwave pulse, the time-domain signal due to coherent emission by the polarized molecules is detected and Fourier transformed to obtain the microwave spectrum.

Microwave studies in molecular beams are usually limited to studying the ground vibrational state of the complex. For complexes made up of two molecules (as opposed to atoms), the intermolecular vibrations are usually of relatively low amplitude (though there are some notable exceptions to this, such as the ammonia dimer). Under these circumstances, the methods of classical microwave spectroscopy can be used to determine the structure of the complex. The principal quantities obtained from a microwave spectrum are the rotational constants of the complex, which are conventionally designated A , B and C in decreasing order of magnitude: there is one rotational constant B for a linear complex, two constants (A and B or B and C) for a complex that is a symmetric top and three constants (A , B and C) for an

-4-

asymmetric top. The rotational constants are related to the moments of inertia of the complex. For a rigid complex, the rotational constants are simply

$$A = \frac{h^2}{2I_A} \quad B = \frac{h^2}{2I_B} \quad C = \frac{h^2}{2I_C} \quad (\text{C1.3.1})$$

where I_A , I_B and I_C are the moments of inertia of the complex about its three principal axes. If the structures of the free monomers are assumed to be unchanged in the complex, then the number of coordinates required to define a ‘structure’ varies from one (for an atom–atom complex) to six (for a complex formed from two nonlinear molecules) as shown in figure C1.3.2. The rotational constants obtained for a single isotopic species are thus not usually enough to determine the structure of the complex, and it is usual to measure spectra for several isotopically substituted species.

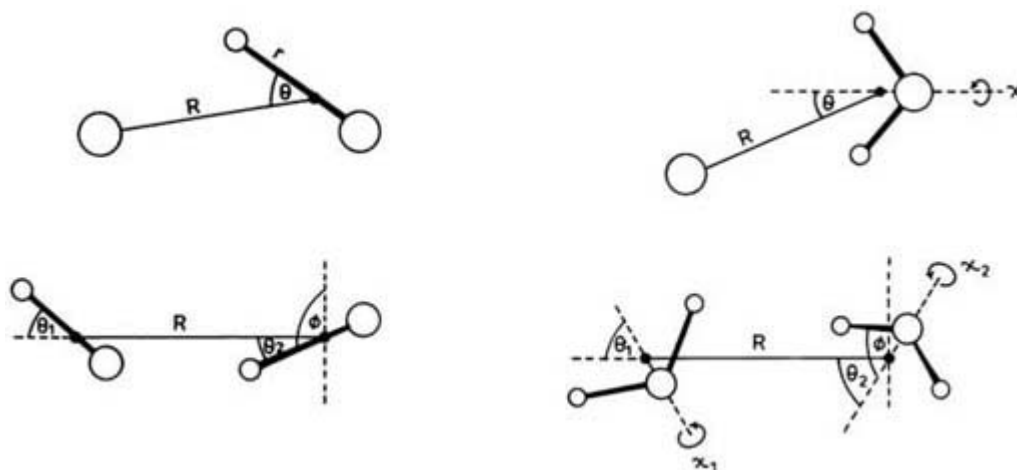


Figure C1.3.2. Coordinate systems used for intermolecular potential energy surfaces. (Taken from [60].)

The use of isotopic substitution to determine structures relies on the assumption that different isotopomers have the same structure. This is not nearly as reliable for Van der Waals complexes as for chemically bound molecules. In particular, substituting D for H in a hydride complex can often change the amplitudes of bending vibrations substantially; under such circumstances, the idea that the complex has a single ‘structure’ is no longer appropriate and it is necessary to think instead of motion on the complete potential energy surface; a well defined equilibrium structure may still exist, but knowledge of it does not constitute an adequate description of the complex.

There are other important properties that can be measured from microwave and radiofrequency spectra of complexes. In particular, the dipole moments and nuclear quadrupole coupling constants of complexes may contain useful information on the structure or potential energy surface. This is most easily seen in the case of the dipole moment. The dipole moment of the complex is a vector, which may have components along all the principal inertial axes.

-5-

Measurements of Stark splittings in microwave and radiofrequency spectra allow these components to be determined. The main contribution to the dipole moment of the complex arises from the permanent dipole moment vectors of the monomers, which project along the axes of the complex according to simple trigonometry (cosines). Thus, measurements of the dipole moment convey information about the orientation of the monomers in the complex. It is of course necessary to take account of effects due to induced dipole moments and to consider whether the effects of vibrational averaging are important.

The argument for nuclear quadrupole coupling constants is similar, except that they are second-rank tensors rather than vectors and so different angular functions (involving second Legendre polynomials, or squares of cosines) are involved in the projections. Nuclear quadrupole coupling constants have the advantage that induced effects are usually very small, so that they contain uncontaminated angular information.

When a complex rotates, it stretches slightly and may also undergo other small structural changes. These changes are reflected in centrifugal distortion constants, which can be extracted from microwave (or other) spectra. The centrifugal distortion constants contain useful information on intermolecular forces, because they measure how easily the complex is distorted; in essence, they reflect the force constants for the bending and stretching vibrations.

It is also possible to measure microwave spectra of some more strongly bound Van der Waals complexes in a gas cell rather than a molecular beam. Indeed, the first microwave studies on molecular clusters were of this type, on carboxylic acid dimers [11]. The resolution that can be achieved is not as high as in a molecular beam, but bulk gas studies have the advantage that vibrational satellites, due to pure rotational transitions in complexes with intermolecular bending and stretching modes excited, can often be identified. The frequencies of the vibrational satellites contain information on how the vibrationally averaged structure changes in the excited states, while their intensities allow the vibrational frequencies to be estimated.

C1.3.2.2 INFRARED SPECTROSCOPY

As described above, classical infrared spectroscopy using grating spectrometers and gas cells provided some valuable information in the early days of cluster spectroscopy, but is of limited scope. However, the advent of tunable infrared lasers in the 1980s opened up the field and made rotationally resolved infrared spectra accessible for a wide range of species. As for microwave spectroscopy, tunable infrared laser spectroscopy has been applied both in gas cells and in molecular beams. In a gas cell, the increased sensitivity of laser spectroscopy makes it possible to work at much lower pressures, so that strong monomer absorptions are less troublesome.

The intermolecular bending and stretching vibrations of Van der Waals complexes typically have wave-numbers between 20 cm^{-1} and 200 cm^{-1} . At the temperatures used in gas cells, usually between 77 K and 300 K, many complexes are in excited vibrational states. In addition, since most complexes (other than those of He and H_2) have rotational constants that are 0.1 cm^{-1} or less, very long rotational progressions (up to $J = 100$) are often observed. Spectra obtained in gas cells can thus be very congested and quite difficult to assign reliably. Nevertheless, the large number of excited states produces a very rich spectrum, and it is sometimes possible to characterize excited states that would not otherwise be accessible.

Infrared spectroscopy can also be carried out in molecular beams. The primary advantages of beam spectroscopy are that it dispenses almost entirely with monomer absorptions that overlap regions of interest, and that the complexes are

internally cold, so that the spectrum is very much simplified. The latter advantage is illustrated in figure C1.3.3 which compares infrared spectra of Ar-HF obtained in a gas cell and in a molecular beam, using a tunable difference frequency laser as the light source in each case.

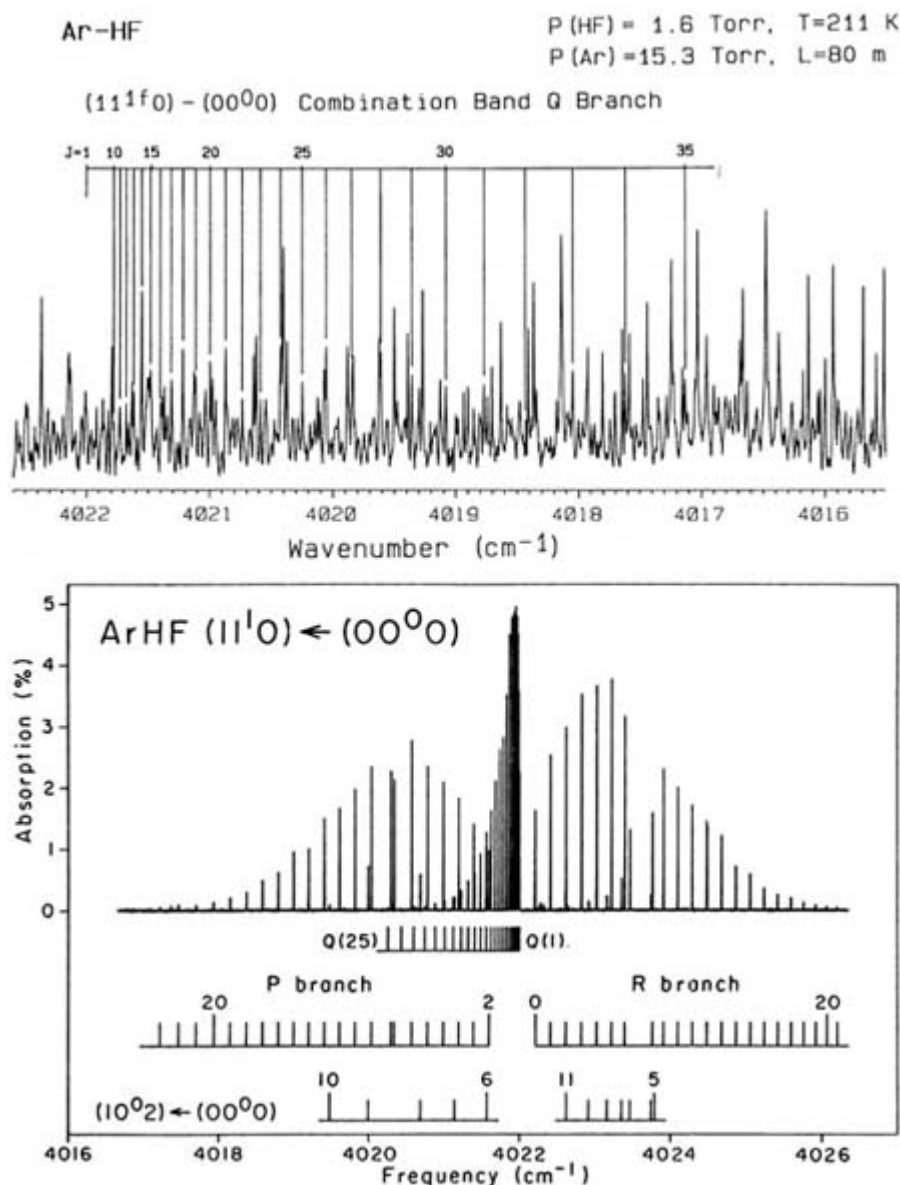


Figure C1.3.3. Comparison between infrared spectra for the p bend combination band of Ar-HF obtained in the gas phase and in a slit jet. (a) The gas-phase spectrum (Taken from [36]). (b) The slit jet spectrum (Taken from [61]).

The earliest molecular beam infrared experiments on Van der Waals complexes used photodissociation spectroscopy: a molecular beam is irradiated with a tunable infrared laser and the molecular beam intensity is measured as a function of

laser frequency [12]. The original experiments did not use state selection, and monitored the beam intensity using a bolometer, which measures the total energy that is deposited by the molecules that reach it. Since complexes in vibrationally excited states usually undergo fast vibrational predissociation, and the recoil velocities are such that the fragments are ejected from the beam and do not reach the detector, transitions involving complexes are detected

as a *decrease* in the energy flux reaching the bolometer. This allows complexes to be distinguished from monomers, which reach the detector undissociated, and produce a positive signal because the energy absorbed from the photon is deposited on the bolometer. Photodissociation experiments have proved to be a very rich source of information, not only on energy levels but also on predissociation dynamics, because the fragments that leave the beam can be detected and their internal states inferred.

One problem with molecular beam techniques is that, although the concentration of complexes is high, the available path length is very low. For some time this precluded the observation of spectra in molecular beams by the conventional spectroscopic approach of monitoring the attenuation in the laser beam intensity caused by absorption. In order to overcome this, Nesbitt and co-workers [13] developed techniques using molecular beams expanded through a slit rather than a circular hole. This provides much longer path lengths, and makes it possible to carry out direct absorption experiments, monitoring the depletion in laser beam intensity (as in a normal spectrometer) rather than the molecular beam intensity.

Most infrared spectroscopy of complexes is carried out in the mid-infrared, which is the region in which the monomers usually absorb infrared radiation. Van der Waals complexes can absorb mid-infrared radiation either with or without simultaneous excitation of intermolecular bending and stretching vibrations. The mid-infrared bands that contain the most information about intermolecular forces are *combination bands*, in which the intermolecular vibrations are excited. Such spectra map out the vibrational and rotational energy levels associated with monomers in excited vibrational states and, thus, provide information on interaction potentials involving excited monomers, which may be slightly different from those for ground-state molecules.

It is thus of great interest to carry out experiments that excite the intermolecular bending and stretching vibrations directly, without exciting the monomers as well. These transitions lie deep in the far infrared, typically in the 20–200 cm^{-1} region, and this has traditionally been a very difficult region for spectroscopy. Fourier transform spectroscopy in gas cells can be applied to obtain rotationally resolved spectra of complexes in this region [14], and it is also possible to measure far-infrared spectra in molecular beams. Early work [15, 16] used laser Stark spectroscopy, in which the molecules were tuned into resonance with the laser by applying a Stark field. However, this was superseded by methods using tunable far-infrared lasers, based on nonlinear mixing of fixed-frequency molecular lasers with microwave radiation [17]; these make it possible to obtain zero-field far-infrared spectra of complexes, which are much easier to interpret than laser Stark spectra.

Mid-infrared combination bands and far-infrared spectra of Van der Waals complexes map out the pattern of energy levels associated with intermolecular bending and stretching vibrations. The principal quantities that can be observed are vibrational frequencies and rotational constants, though once again subsidiary quantities such as centrifugal distortion constants, dipole moments and nuclear quadrupole coupling constants may sometimes be extracted. In addition, observation of line broadening due to predissociation can sometimes provide very direct measurements of binding energies (and hence of the depths of potential wells).

For chemically bound molecules, it is usual to analyse the vibrational energy levels in terms of *normal modes*: a non-linear (or linear) molecule with N atoms has $3N - 6$ (or $3N - 5$) vibrational degrees of freedom. There is a

-8-

fundamental frequency ν_i associated with each degree of freedom i . The vibrational energy levels may be labelled with a quantum number v_i for each normal mode and, to a first approximation, the energy is given by the harmonic expression

$$E_v = \sum_{i=1}^{3N-6(\text{or } 5)} h\nu_i (v_i + \frac{1}{2}). \quad (\text{C1.3.2})$$

For quantitative work, it is necessary to include anharmonic corrections and coupling between the normal modes, but the general picture suffices to handle the lower vibrational levels of most near-rigid molecules.

Unfortunately, the normal-mode picture is quite inadequate for many Van der Waals complexes, even in their lowest few levels. This may be illustrated by considering the bending levels of Ar–HCl, which are shown in [figure C1.3.4](#). The equilibrium geometry of Ar–HCl is linear, so it has one doubly degenerate bending mode. In a normal-mode picture, the vibrational levels would be labelled by a bending quantum number ν and a vibrational angular momentum l (or K). The normal-mode energy level pattern is shown on the right-hand side of [figure C1.3.4](#). The ground state has $\nu^K = 0^0$, while the first excited state has $\nu^K = 1^{\pm 1}$. The $\nu = 2$ excited states lie at about twice the energy of the $1^{\pm 1}$ states and are split into components with $\nu^K = 2^0$ and $2^{\pm 2}$. The actual levels of the complex clearly have a quite different pattern; in particular, the lowest excited state with $K = 0$ lies *below* the lowest $K = \pm 1$ level, not above it. The pattern observed in the complex is in fact much closer to that expected for an HCl molecule undergoing nearly free internal rotation in the complex, shown in the centre of [figure C1.3.4](#). In the free-rotor limit, the HCl energy levels are simply $bj(j + 1)$, where b is the rotational constant of HCl and j is the rotational quantum number for HCl internal rotation (not overall rotation of the complex). In the presence of potential anisotropy, j becomes quantized along the intermolecular axis with a projection quantum number K , which can take integer values up to $\pm j$. In a complex such as Ar–HCl, the splittings between the different K levels for a given value of j are substantial, but the general picture of hindered internal rotation is considerably closer to the truth than the normal-mode picture of low-amplitude vibrations about an equilibrium geometry. A full analysis of the energy level pattern can provide very detailed information on the anisotropic potential energy surface [18].

-9-

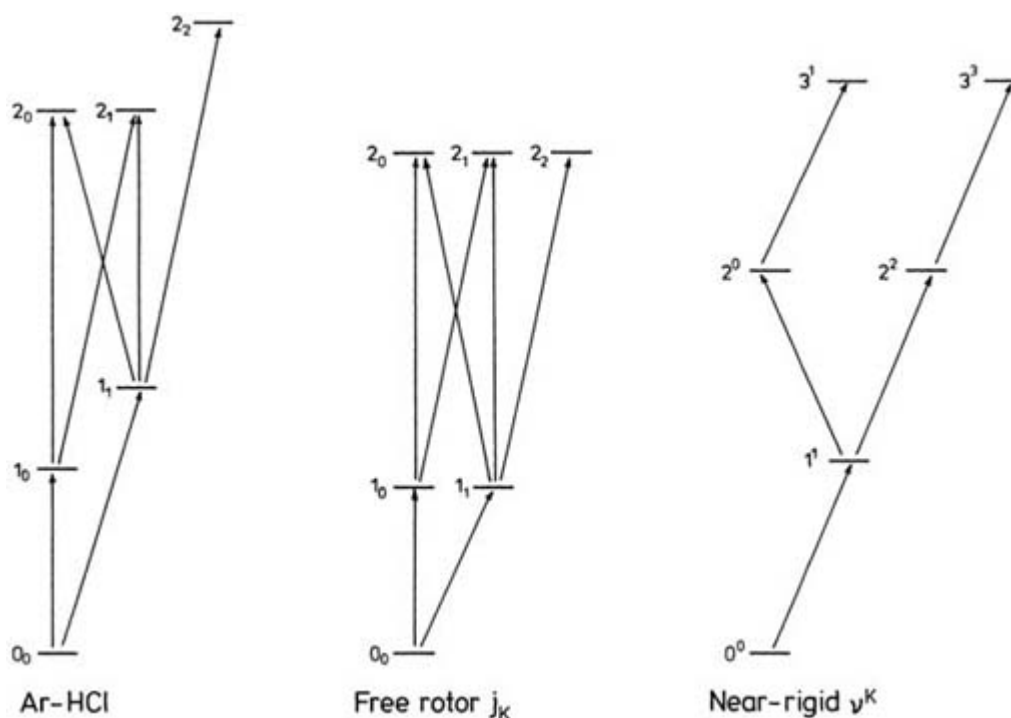


Figure C1.3.4. The real pattern of intermolecular bending energy levels for Ar–HCl (left) compared with the pattern expected for a free internal rotor (centre) and a near-rigid bender (right). The allowed transitions are shown in each case. (Taken from [19].)

The quantum numbers that are appropriate to describe the vibrational levels of a quasilinear complex such as Ar–HCl are thus the monomer vibrational quantum number ν , an intermolecular stretching quantum number n and two quantum numbers j and K to describe the hindered rotational motion. For more rigid complexes, it becomes appropriate to replace j and K with normal-mode vibrational quantum numbers, though there is an awkward intermediate regime in which neither description is satisfactory: see [3] for a discussion of the transition between the two cases. In addition, there is always a quantum number J for the total angular momentum (excluding nuclear spin). The total parity (symmetry under space-fixed inversion of all coordinates) is also a conserved quantity that is spectroscopically important.

C1.3.2.3 PHOTODISSOCIATION AND PREDISSOCIATION

The binding energy of a Van der Waals complex is usually considerably less than the energy of a mid-infrared photon. Accordingly, Van der Waals complexes containing vibrationally excited monomers usually have enough energy to dissociate, by converting the vibrational energy into relative translational energy of the fragments. This process is referred to as *vibrational predissociation*.

For complexes such as Ar-H₂, Ar-HF and Ar-HCl, vibrational predissociation is a very slow process and does not cause appreciable broadening of the lines in the infrared spectrum. Indeed, for Ar-HF, Huang *et al* [20] showed that

-10-

vibrationally excited molecules survive for at least 0.3 ms, and reach a bolometric beam detector undissociated. However, for heavier monomers and more strongly bound complexes, vibrational predissociation can be much faster and lead to observable line broadening in the spectrum. The predissociation lifetime τ and linewidth γ (full width at half maximum, in energy units) are related by

$$\tau = \frac{\hbar}{\gamma}. \quad (\text{C1.3.3})$$

Spectra of two different bands of (HF)₂, showing the difference between predissociated and undissociated spectra [21], are shown in figure C1.3.5

The occurrence of predissociation opens up a new family of observable quantities. It is possible to measure not only linewidths or lifetimes, but also the internal state distributions of the fragments. All these quantities are sensitive to the intermolecular potential and can be used to test or refine proposed potential surfaces.

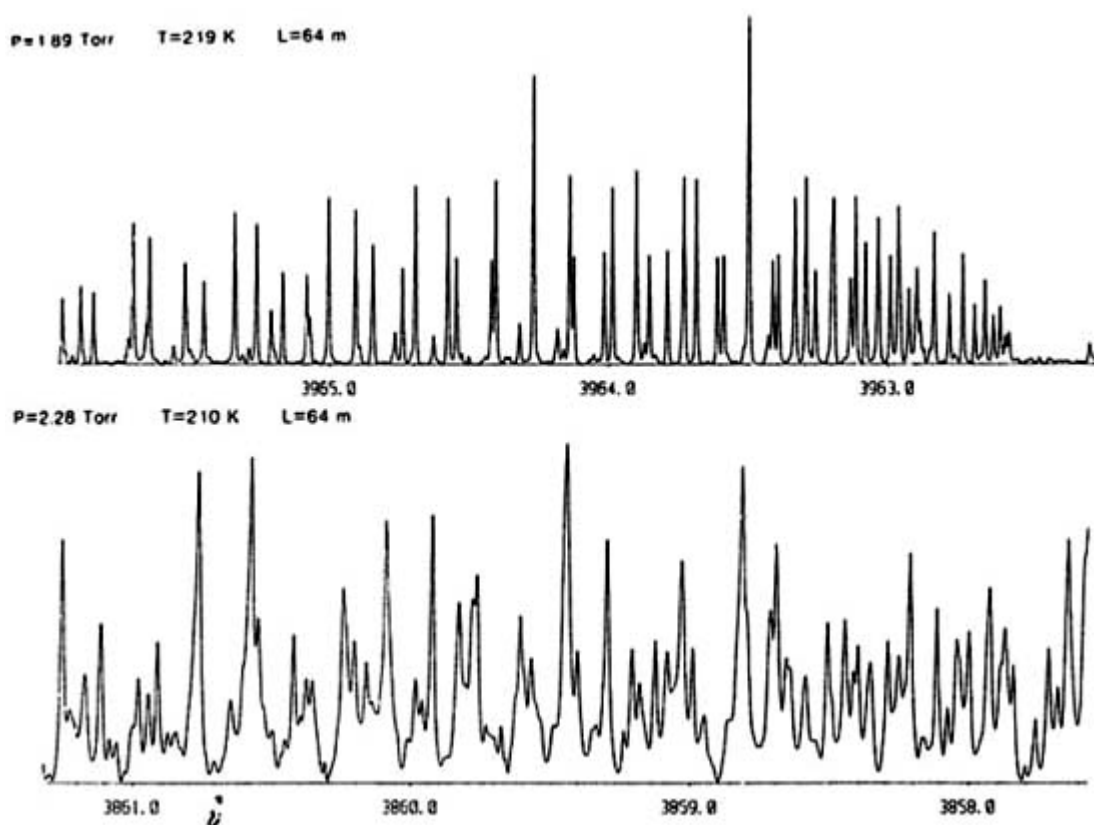


Figure C1.3.5. Spectra of two different infrared bands of HF dimer, corresponding to excitation of the 'bound' (lower panel) and 'free' (upper panel) HF monomers in the complex. Note the additional line width for the 'bound' HF, caused by vibrational predissociation with a lifetime of about 0.8 ns. (Taken from [21].)

-11-

C1.3.2.4 VISIBLE AND ULTRAVIOLET SPECTROSCOPY

Tunable visible and ultraviolet lasers were available well before tunable infrared and far-infrared lasers. There are many complexes that contain monomers with visible and near-UV spectra. The earliest experiments to give detailed dynamical information on complexes were in fact those of Smalley *et al* [22], who observed laser-induced fluorescence (LIF) spectra of He-I₂ complexes. They excited the complex in the I₂ B ← X band, and were able to produce excited-state complexes containing B-state I₂ in a wide range of vibrational states. From line widths and dispersed fluorescence spectra, they were able to study the rates and pathways of dissociation. Such work was subsequently extended to many other systems, including the rare gas-Cl₂ systems, and has given quite detailed information on potential energy surfaces [23].

The homonuclear rare gas pairs are of special interest as models for intermolecular forces, but they are quite difficult to study spectroscopically. They have no microwave or infrared spectrum. However, their vibration-rotation energy levels can be determined from their electronic absorption spectra, which lie in the vacuum ultraviolet (VUV) region of the spectrum. In the most recent work, Herman *et al* [24] have measured vibrational and rotational frequencies to great precision. In the case of Ar-Ar, the results have been incorporated into a multiproperty analysis by Aziz [25] to develop a highly accurate pair potential.

C1.3.2.5 OTHER SPECTROSCOPIC METHODS

Far-infrared and mid-infrared spectroscopy usually provide the most detailed picture of the vibration-rotation energy levels in the ground electronic state. However, they are not always possible and other spectroscopic methods are also important.

It is often difficult to observe direct infrared transitions from the ground state to highly excited intermolecular bending and stretching states, because the spectroscopic intensities are very low. One technique that circumvents this difficulty is stimulated-emission pumping (SEP) spectroscopy. Molecules (or complexes) are first promoted to an excited electronic state by a pump laser, and then emission is stimulated by a second tunable laser (the dump laser) at slightly lower frequency. Meanwhile, the population of the excited state is monitored in some way (perhaps by observing the spontaneous emission signal). When the dump laser is resonant with a transition back down to an excited vibrational level of the ground electronic state, it causes an observable dip in the population of the excited state. SEP spectroscopy has been applied with great success to map out the energy levels of the open-shell complex Ar-OH in its ground electronic state [26], and the resulting spectra have been used to obtain potential energy surfaces for the interaction [27].

C1.3.3 EXAMPLES

C1.3.3.1 BINARY COMPLEXES: AR-HCL AND AR-HF

The Ar-HCl and Ar-HF Van der Waals complexes were among the first to be detected experimentally, by the observation of weak peaks lying between the vibration-rotation lines of HCl and HF in mixtures with rare gases as

-12-

described above. The first measurements were made at a time when very little was known about intermolecular forces involving molecules and it was clear from the beginning that, if high-resolution spectra could be measured, they would contain an immense amount of valuable information. It was also clear that atom–molecule complexes would be much easier to treat theoretically than molecule–molecule complexes. Accordingly, when MBER spectroscopy (see above) was developed in the early 1970s, Ar–HCl [28] and Ar–HF [29] were among the first systems studied.

The early MBER spectra were of two types: first, pure rotational (microwave) transitions and, secondly, radiofrequency transitions between different hyperfine levels; the latter were observable only for Ar–HCl, because both ^{35}Cl and ^{37}Cl have nuclear quadrupole moments while ^{19}F has none. In addition, the rotational levels of both complexes could be split by applying an electric field (Stark effect) and the dipole moment could be determined from the size of the splittings. It turned out that Ar–HCl and Ar–HF were ‘pseudo-linear’ molecules: their rotational energy levels could be interpreted entirely in terms of a single rotational constant B and centrifugal distortion constant D_J . Nevertheless, the spectra contained signatures that showed that they were *not* rigid linear species: the dipole moments of the complexes were only about two-thirds of those of the HCl and HF monomers, and the nuclear quadrupole coupling constants of Ar–HCl were only about one-third of those of the corresponding HCl isotopomers. Both these fractions would have been close to one for a rigid linear species.

The Ar–HCl and Ar–HF complexes became prototypes for the study of intermolecular forces. Holmgren *et al* [30] produced an empirical potential energy surface for Ar–HCl fitted to the microwave and radiofrequency spectra, showing that the equilibrium geometry is linear, Ar–HCl, with a well depth of around 175 cm^{-1} . However, it turned out that this surface gave a poor account of properties that were sensitive to the shape of the repulsive wall, such as the pressure broadening of HCl vibration–rotation lines in the gas phase [31]. This is a general problem with intermolecular potentials determined solely from the spectra of complexes; the spectra do not contain enough information to determine the shape of the repulsive wall. Conversely, the potential energy surface determined from the pressure broadening [31] gave a very poor account of the spectra of the complexes [32].

Hutson and Howard [33] combined the Van der Waals spectra with pressure-broadening data and virial coefficients to produce an improved surface. However, even then, they showed that the microwave spectra determined the potential quite accurately between $\theta = 0$ and 90° , around the absolute minimum, but could not determine whether there was a secondary minimum around $\theta = 180^\circ$, the linear Ar–ClH structure. This again illustrates a general limitation: microwave spectra do not usually sample geometries far from equilibrium. Hutson and Howard [34] suggested that the best way to probe the Ar–ClH region would be to measure mid-infrared or far-infrared spectra, exciting intermolecular bending bands, and gave predictions for such spectra for potential surfaces with and without secondary minima.

Laser techniques capable of obtaining rotationally resolved infrared spectra of complexes came along fairly quickly [12]. For Ar–HCl, Marshall *et al* [15] and Ray *et al* [16] developed laser Stark methods for measuring far-infrared spectra in a molecular beam with fixed-frequency lasers. The laser Stark results were extended by Busarow *et al* [17], who used a tunable far-infrared laser for the first time. High-resolution mid-infrared spectra for Ar–HCl and Ar–HF were measured in both gas cells [35, 36] and slit jets [13, 37] using tunable difference-frequency lasers. Hutson [38] used the new spectra to develop an improved potential energy surface for Ar–HCl, establishing definitively that there is a secondary minimum around the linear Ar–ClH geometry, about 30 cm^{-1} shallower than the primary minimum at the Ar–HCl geometry.

Over the next few years, both the mid-infrared and the far-infrared spectra for Ar–HF and Ar–HCl were extended to numerous other bands and to other isotopic species (most importantly those containing deuterium). In 1992, Hutson [18, 39] combined all the available spectroscopic data to produce definitive potential energy surfaces that included both the angle dependence and the dependence on the HF/HCl monomer vibrational quantum number v (actually through the mass-reduced vibrational quantum number [40], $\eta = (v + \frac{1}{2})/\sqrt{\mu_{\text{HX}}}$). The resulting Ar–HF potential (designated H6(4,3,2)) is shown in figure C1.3.6. The potentials have since been used to calculate

numerous properties, including pressure-broadening coefficients, inelastic scattering integral cross sections, differential cross sections and the spectra of HF and HCl in rare gas clusters and matrices. They have proved to be remarkably reliable. They have also been used extensively as testing grounds for new *ab initio* electronic structure methods.

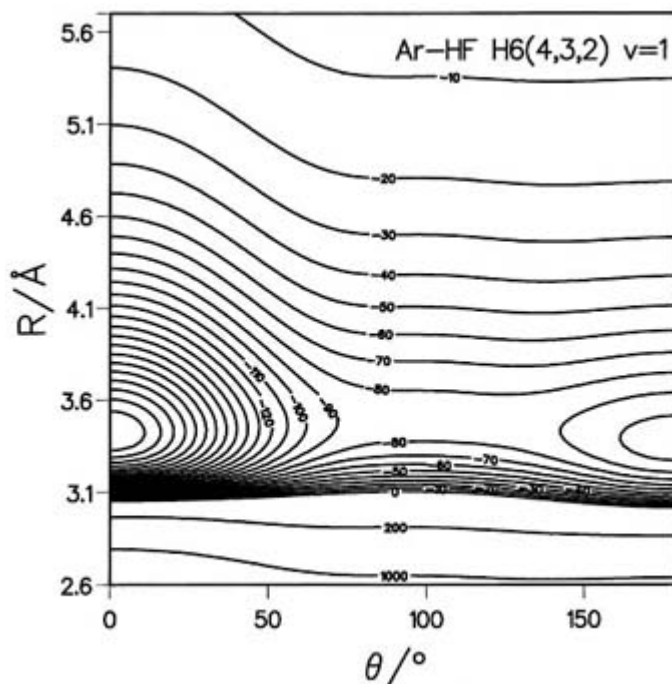


Figure C1.3.6. Contour plot of the H6(4,3,2) potential for Ar--HF, which was fitted principally to data from far-infrared and mid-infrared spectroscopy of the Ar--HF Van der Waals complex. The contours are labelled in cm^{-1} . (Taken from [39].)

C1.3.3.2 LARGER CLUSTERS: $(\text{H}_2\text{O})_N$

The most important molecular interactions of all are those that take place in liquid water. For many years, chemists have worked to model liquid water, using molecular dynamics and Monte Carlo simulations. Until relatively recently, however, all such work was done using ‘effective potentials’ [41], designed to reproduce the condensed-phase properties but with no serious claim to represent the true interactions between a pair of water molecules.

The advent of cluster spectroscopy offered the opportunity to place studies of liquid water and aqueous solutions on a

much firmer footing, by learning first about the water dimer and the true water–water pair potential and then exploring how the interactions change for larger clusters.

There has been considerable progress in this direction. The water dimer has been the subject of intense spectroscopic study, especially by far-infrared vibration–rotation–tunnelling spectroscopy (FIR-VRTS) [42]. Many different bands have been observed, involving intermolecular bending and stretching vibrations and tunnelling motions. The potential energy surface is six-dimensional (one distance and five angles) even when intramolecular vibrations of the water monomer are neglected. Because of this, developing a purely empirical potential surface from the spectroscopic observations is a difficult task. Nevertheless, Fellers *et al* [43] have used the spectra to fit the parameters of a functional form due to Millot and Stone [44].

Larger water clusters, including the trimer, tetramer, pentamer and hexamer, have also been studied spectroscopically. The equilibrium geometries of some of them are shown in [figure C1.3.7](#). A characteristic feature of the water clusters, and of many others, is that they have large numbers of equivalent minima on their potential energy surfaces, with relatively low barriers that allow tunnelling motions between the different minima. There are often low-lying subsidiary minima as well, which also affect the spectroscopy. Assigning and interpreting the spectra is a complex task involving permutation-inversion symmetry as well as sophisticated potential energy surfaces and many-dimensional bound-state calculations. The behaviour of the clusters cannot be explained in terms of small excursions from an equilibrium structure: it really becomes necessary to talk of ‘spectroscopy beyond structure’.

-15-

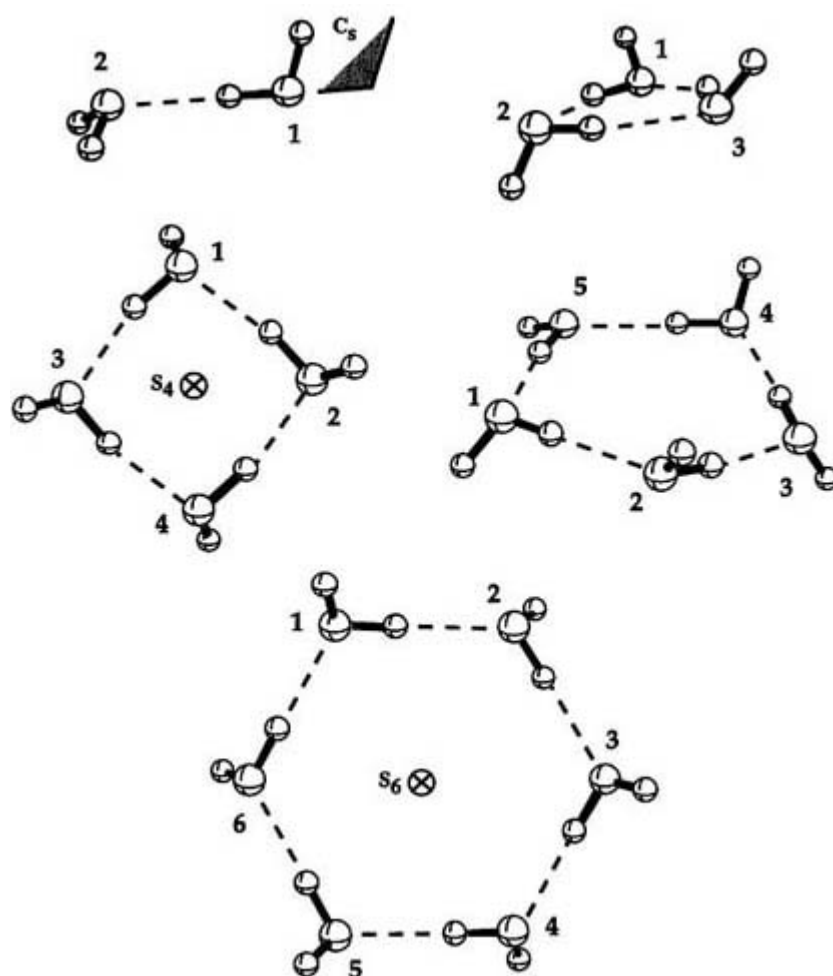


Figure C1.3.7. Equilibrium geometries of some water clusters from *ab initio* calculations. (Taken from [62].)

The equilibrium geometry of the water trimer is chiral, with fast tunnelling between the enantiomers [45]. There are actually 48 right-handed and 48 left-handed forms, all with the same energy. The tetramer is cyclic, but non-chiral, and with many fewer equivalent minima [46]. The pentamer is also cyclic [47], but the hexamer has a cage structure [48]. Higher clusters have also been studied, though at lower resolution, by infrared spectroscopy on beams of mass-selected clusters [49]. The octamer is particularly interesting, because it is believed to have dynamical cubic symmetry. Many other spectroscopic techniques have been used as well.

The ultimate reason for studying water clusters is of course to understand the interactions in bulk water (though clusters are interesting in their own right, too, because finite-size systems can have special properties). There has been

a vast amount of work on simulating the water clusters, both classically and quantumly. Quantum simulations are particularly challenging, because of the high dimensionality of the systems. Nevertheless, calculations on the lowest vibrational states of systems as large as water octamer have been carried out using quantum Monte Carlo methods.

The intermolecular forces between water molecules are strongly non-additive. It is not realistic to expect any pair potential to reproduce the properties of both the water dimer and the larger clusters, let alone liquid water. There has therefore been a great deal of work on developing potential models with explicit pairwise-additive and non-additive parts [44, 50, 51]. It appears that, when this is done, the energy of the larger clusters and ice has a non-additive contribution of about 30%.

An important area that has yet to be fully explored is the effect of the flexibility of water molecules. The intermolecular forces in water are large enough to cause significant distortions from the gas-phase monomer geometry. In addition, the flexibility is crucial in any description of vibrational excitation in water.

C1.3.3.3 REACTIVE SPECIES: H_2-OH

One of the motivations for studying Van der Waals complexes and clusters is that they are floppy systems with similarities to the transition states of chemical reactions. This can be taken one stage further by studying clusters that actually are precursors for chemical reactions, and can be broken up to make more than one set of products. A good example of this is H_2-OH , which can in principle dissociate to form either $H_2 + OH$ or $H_2O + H$. Indeed, dissociation to $H_2O + H$ is energetically favoured: the reaction $H_2 + OH \rightarrow H_2O + H$ is exothermic by about 5000 cm^{-1} , and plays a key role in combustion. It has been extensively studied both in the gas phase and in crossed molecular beams. The only reason that the H_2-OH complex can be observed at all is that there is a barrier to reaction of more than 2000 cm^{-1} , and the transition state has a quite different geometry from the complex.

OH is an open-shell molecule with a $^2\Pi$ ground state. In the complex, the Π state splits into two states of A' and A'' symmetry. *Ab initio* calculations [52] gave a well depth of 188 cm^{-1} (relative to free $OH + H_2$) for the A'' state, in a symmetric $O-H-H_2$ geometry. The first excited state, by contrast, has a well at least 2300 cm^{-1} deep. The first spectroscopic observations of H_2-OH [53], using laser-induced fluorescence, were carried out in parallel with bound-state calculations [54]. The early experiments showed broad peaks because the levels of the excited electronic state are very short-lived; they confirmed the existence of bound H_2-OH complexes, but provided only limited information on the interactions between ground-state molecules. Nevertheless, they opened the way to much higher-resolution studies using infrared overtone pumping to excite the OH vibration [55] and stimulated Raman spectroscopy to excite the H_2 vibration in the complex [56].

The vibrationally excited states of H_2-OH have enough energy to decay either to H_2 and OH or to cross the barrier to reaction. Time-dependent experiments have been carried out to monitor the non-reactive decay (to $H_2 + OH$), which occurs on a timescale of microseconds for H_2-OH but nanoseconds for D_2-OH [57, 58]. Analogous experiments have also been carried out for complexes in which the H_2 vibration is excited [59]. The reactive decay products have not yet been detected, but it is probably only a matter of time. Even if it proves impossible for H_2-OH , there are plenty of other 'pre-reactive' complexes that can be produced. There is little doubt that the spectroscopy of such species will be a rich source of information on reactive potential energy surfaces in the fairly near future.

REFERENCES

- [1] Stogryn D E and Hirschfelder J O 1959 Contribution of bound, metastable and free molecules to the second virial coefficient and some properties of double molecules *J. Chem. Phys.* **31** 1531–5
- [2] Vodar B 1960 Spectra of compressed gases and molecular interactions *Proc. R. Soc. A* **255** 44–55
- [3] Rank D H, Rao B S and Wiggins T A 1963 Absorption spectra of hydrogen halide–rare gas mixtures *J. Chem. Phys.* **37** 2511–15
- [4] McKellar A R W and Welsh H L 1971 Anisotropic intermolecular force effects in spectra of H₂- and D₂-rare gas complexes *J. Chem. Phys.* **55** 595–609
- [5] McKellar A R W 1982 Infrared spectra of hydrogen–rare gas Van der Waals molecules *Faraday Discuss. Chem. Soc.* **73** 89–108
- [6] Le Roy R J and Van Kranendonk J 1974 Anisotropic intermolecular potentials from an analysis of spectra of H₂- and D₂-inert gas complexes *J. Chem. Phys.* **61** 4750–69
- [7] Le Roy R J and Carley J S 1980 Spectroscopy and potential energy surfaces of Van der Waals molecules *Adv. Chem. Phys.* **42** 353–420
- [8] Le Roy R J and Hutson J M 1987 Improved potential energy surfaces for the interaction of H₂ with Ar, Kr and Xe *J. Chem. Phys.* **86** 837–53
- [9] Dyke T R, Howard B J and Klemperer W 1972 Radiofrequency and microwave spectrum of the hydrogen fluoride dimer: a nonrigid molecule *J. Chem. Phys.* **56** 2442–54
- [10] Balle T J, Campbell E J, Keenan M R and Flygare W H 1980 A new method for observing the rotational spectra of weak molecular complexes: KrHCl *J. Chem. Phys.* **72** 922–32
- [11] Costain C C and Srivastava G P 1961 Study of hydrogen bonding: microwave spectra of CF₃COOH–HCOOH *J. Chem. Phys.* **35** 1903–4
- [12] Gough T E, Miller R E and Scoles G 1981 Infrared spectra and vibrational predissociation of (CO₂)_n clusters using laser-molecular beam techniques *J. Phys. Chem.* **85** 4041–6
- [13] Lovejoy C M, Schuder M D and Nesbitt D J 1986 High resolution IR laser spectroscopy of Van der Waals complexes in slit supersonic jets: observation and analysis of ν_1 , $\nu_1 + \nu_2$ and $\nu_1 + 2\nu_3$ in ArHF *J. Chem. Phys.* **85** 4890–902
- [14] McKellar A R W 1994 Long-path equilibrium IR spectra of weakly bound complexes at low temperatures *Faraday Discuss. Chem. Soc.* **97** 69–80
- [15] Marshall M D, Charo A, Leung H O and Klemperer W 1985 Characterization of the lowest-lying Π bending state of Ar–HCl by far infrared laser-Stark spectroscopy and molecular beam electric resonance *J. Chem. Phys.* **83** 4924–33
- [16] Ray D, Robinson R L, Gwo D H and Saykally R H 1986 Vibrational spectroscopy of Van der Waals bonds: measurement of the perpendicular bend of ArHCl by intracavity far infrared laser spectroscopy of a supersonic jet *J. Chem. Phys.* **84** 1171–80

- [17] Busarow K L, Blake G A, Laughlin K B, Cohen R C, Lee Y T and Saykally R J 1987 Tunable far-infrared laser spectroscopy in a planar supersonic jet: the Σ bending vibration of Ar–H³⁵Cl *Chem. Phys. Lett.* **141** 289–91
- [18] Hutson J M 1992 Vibrational dependence of the anisotropic intermolecular potential of Ar–HCl *J. Chem. Phys.* **96** 4237–47
- [19] Hutson J M 1991 An introduction to the dynamics of Van der Waals molecules *Adv. Mol. Vibrat. Coll. Dyn.* **1A** 1–45
- [20] Huang Z S, Jucks K W and Miller R E 1986 The argon–hydrogen fluoride binary complex: an example of a long lived

metastable system *J. Chem. Phys.* **85** 6905–9

- [21] Pine A S, Lafferty W J and Howard B J 1984 Vibrational predissociation, tunneling and rotational saturation in the HF and DF dimers *J. Chem. Phys.* **81** 2939–50
- [22] Smalley R E, Levy D H and Wharton L 1976 The fluorescence excitation spectrum of the HeI₂ Van der Waals complex *J. Chem. Phys.* **64** 3266–76
- [23] Rohrbacher A, Williams J and Janda K C 1999 Rare gas–dihalogen potential energy surfaces *PCCP* **1** 5263–76
- [24] Herman P R, LaRocque P E and Stoicheff B P 1988 Vacuum ultraviolet laser spectroscopy 5: rovibrational spectra of Ar₂ and constants of the ground and excited states *J. Chem. Phys.* **89** 4535–49
- [25] Aziz R A 1993 A highly accurate interatomic potential for argon *J. Chem. Phys.* **99** 4518–25
- [26] Berry M T, Loomis R A, Giancarlo L C and Lester M I 1991 Stimulated emission pumping of intermolecular vibrations in OH–Ar (*X*²I) *J. Chem. Phys.* **96** 7890–903
- [27] Dubernet M-L and Hutson J M 1993 Potential energy surfaces for Ar–OH (*X*²I) obtained by fitting to high-resolution spectroscopy *J. Chem. Phys.* **99** 7477–86
- [28] Novick S E, Davies P, Harris S J and Klemperer W 1973 Determination of the structure of ArHCl *J. Chem. Phys.* **59** 2273–9
- [29] Harris S J, Novick S E and Klemperer W 1974 Determination of the structure of ArHF *J. Chem. Phys.* **60** 3208–9
- [30] Holmgren S L, Waldman M and Klemperer W 1978 Internal dynamics of Van der Waals complexes II: Determination of a potential energy surface for ArHCl *J. Chem. Phys.* **69** 1661–9
- [31] Kircz J G, van der Peijl G J Q and van der Elsken J 1978 Determination of potential energy surfaces of Ar–HCl and Kr–HCl from rotational line-broadening data *J. Chem. Phys.* **69** 4606–16
- [32] Hutson J M and Howard B J 1980 Spectroscopic properties and potential surfaces for atom–diatom Van der Waals complexes *Mol. Phys.* **41** 1123–41
- [33] Hutson J M and Howard B J 1981 The intermolecular potential energy surface of Ar–HCl *Mol. Phys.* **43** 493–516
- [34] Hutson J M and Howard B J 1982 Anisotropic intermolecular forces II: Rare gas–hydrogen chloride systems *Mol. Phys.* **45** 769–90
- [35] Howard B J and Pine A S 1985 Rotational predissociation and libration in the infrared spectrum of Ar–HCl *Chem. Phys. Lett.* **1222** 1–8

- [36] Fraser G T and Pine A S 1986 Van der Waals potentials from the infrared spectra of rare gas–HF complexes *J. Chem. Phys.* **85** 2502–15
- [37] Lovejoy C M and Nesbitt D J 1988 Infrared-active combination bands in ArHCl *Chem. Phys. Lett.* **146** 582–8
- [38] Hutson J M 1988 The intermolecular potential of Ar–HCl: determination from high-resolution spectroscopy *J. Chem. Phys.* **89** 4550–7
- [39] Hutson J M 1992 Vibrational dependence of the anisotropic intermolecular potential of Ar–HF *J. Chem. Phys.* **96** 6752–67
- [40] Stwalley W C 1975 Mass-reduced quantum numbers: application to the isotopic mercury hydrides *J. Chem. Phys.* **63** 3062–80
- [41] Jorgenson W L, Chandrasekhar J, Madura J D, Impey R W and Klein M L 1983 Comparison of simple potential functions for simulating liquid water *J. Chem. Phys.* **79** 926–35
- [42] Busarow K L, Cohen R C, Blake G A, Laughlin K B, Lee Y T and Saykally R J 1989 Measurement of the perpendicular

rotation-tunnelling spectrum of the water dimer by tunable far-infrared laser spectroscopy in a planar supersonic jet *J. Chem. Phys.* **90** 3937–43

- [43] Fellers R S, Leforestier C, Braly L B, Brown M G and Saykally R J 1999 Spectroscopic determination of the water pair potential *Science* **284** 945–8
- [44] Millot C and Stone A J 1992 Towards an accurate intermolecular potential for water *Mol. Phys.* **77** 439–62
- [45] Pugliano N and Saykally R J 1992 Measurement of quantum tunnelling between chiral isomers of the cyclic water trimer *Science* **257** 1936–40
- [46] Cruzan J D, Braly L B, Liu K, Brown M G, Loeser J G and Saykally R J 1996 Quantifying hydrogen bond cooperatively in water: VRT spectroscopy of the water tetramer *Science* **271** 59–62
- [47] Liu K, Brown M G, Cruzan J D and Saykally R J 1996 Vibration–rotation tunnelling spectra of the water pentamer *Science* **271** 62–4
- [48] Liu K, Brown M G and Saykally R J 1997 Terahertz laser vibration rotation tunnelling spectroscopy and dipole moment of a cage form of the water hexamer *J. Phys. Chem. A* **101** 8995–9010
- [49] Buck U, Ettischer I, Melzer M, Buch V and Sadlej J 1998 Structure and spectra of three-dimensional (H₂O)_n clusters, n = 8, 9, 10 *Phys. Rev. Lett.* **80** 2578–81
- [50] Burnham C J, Li J C, Xantheas S S and Leslie M 1999 The parametrization of a Thole-type all-atom polarizable water model from first principles and its application to the study of water clusters (n = 2–21) and the phonon spectrum of ice Ih *J. Chem. Phys.* **110** 4566–81
- [51] Milet A, Moszynski R, Wormer P E S and van der Avoird A 1999 Hydrogen bonding in water clusters: pair and many-body interactions from symmetry-adapted perturbation theory *J. Phys. Chem. A* **103** 6811–19
- [52] Miller S M, Clary D C, Kliesch A and Werner H J 1994 Rotationally inelastic and bound-state dynamics of H₂–OH (X²Π) *Mol. Phys.* **83** 405–28

-20-

- [53] Loomis R A and Lester M I 1995 Stabilization of reactants in a weakly bound complex: OH–H₂ and OH–D₂ *J. Chem. Phys.* **103** 4371–4
- [54] Hernandez R and Clary D C 1995 Electronic spectra of the OH(A²Σ⁺)–H₂ and OH(A₂Σ₊)–D₂ complexes *Chem. Phys. Lett.* **244** 421–6
- [55] Anderson D T, Schwartz R L and Todd M W and Lester M I 1998 Infrared spectroscopy and time-resolved dynamics of the ortho-H₂–OH entrance channel complex *J. Chem. Phys.* **109** 3461–73
- [56] Wheeler M D, Todd M W, Anderson D T and Lester M I 1999 Stimulated Raman excitation of the ortho-H₂–OH entrance channel complex *J. Chem. Phys.* **110** 6732–42
- [57] Krause P J, Clary D C, Anderson D T, Todd M W, Schwartz R L and Lester M I 1998 Time-resolved dissociation of the H₂–OH entrance channel complex *Chem. Phys. Lett.* **294** 518–22
- [58] Hossenlopp J M, Anderson D T, Todd M W and Lester M I 1998 State-to-state inelastic scattering from vibrationally excited OH–H₂ complexes *J. Chem. Phys.* **109** 10 707–18
- [59] Wheeler M D, Anderson D T, Todd M W, Lester M I, Krause O J and Clary D C 1999 Mode-selective decay dynamics of the ortho-H₂–OH complex: experiment and theory *Mol. Phys.* **97** 151–8
- [60] Hutson J M 1990 Intermolecular forces from the spectroscopy of Van der Waals complexes *Ann. Rev. Phys. Chem.* **41** 123–54

- [61] Lovejoy C M and Nesbitt D J 1989 Intramolecular dynamics of Van der Waals molecules: an extended infrared study of Ar–HF *J. Chem. Phys.* **91** 2790–807
- [62] Xantheas S S 1994 *Ab initio* studies of cyclic water clusters (H₂O)ⁿ, n=1–6. 2. Analysis of many-body interactions *J. Chem. Phys.* **100** 7523–34
-

FURTHER READING

Dyke T R 1984 Microwave and radiofrequency spectra of hydrogen-bonded complexes in the vapor phase *Topics in Current Chemistry* **120** 85–113

Hutson J M 1990 Intermolecular forces from the spectroscopy of Van der Waals molecules *Ann. Rev. Phys. Chem.* **41** 123–54

Huston J M 1991 An introduction to the dynamics of Van der Waals molecules *Adv. Mol. Vibrat. Coll. Dyn.* **1A** 1–45

Nesbitt D J 1994 High-resolution direct infrared laser absorption spectroscopy in slit supersonic jets: intermolecular forces and unimolecular vibrational dynamics in clusters *Ann. Rev. Phys. Chem.* **45** 367–99

van der Avoird A, Wormer P E S and Moszynski R 1994 From intermolecular potential to the spectra of Van der Waals molecules and vice versa *Chem. Rev.* **94** 1931–74

-21-

Stone A J 1996 *The Theory of Intermolecular Forces* (Oxford: Oxford University Press)

Bačić Z and Miller R E 1996 Molecular clusters: structure and dynamics of weakly bound systems *J. Phys. Chem.* **100** 12 945–59

Liu K, Cruzan J D and Saykally R J 1996 Water clusters *Science* **271** 929–33

Loomis R A and Lester M I 1997 OH–H₂ entrance channel complexes *Ann. Rev. Phys. Chem.* **48** 643–73

-1-

C1.4 Atom traps and studies of ultracold systems

John Weiner

C1.4.1 INTRODUCTION

Just over a decade ago the first successful experiments and theory demonstrating that light could be used to cool and confine atoms to sub-milliKelvin temperatures [1, 2] opened several exciting new chapters in chemical physics and in atomic, molecular and optical (AMO) physics. Atom optics and interferometry [5, 6], holography [7], optical lattices [8] and Bose–Einstein condensation in dilute gases all exemplify startling new physics where atoms cooled with light play a pivotal role. The nature of the interactions between these cold particles has become the subject of intensive study not only because of their importance to these new areas of AMO physics but also because their investigation has led to new insights into how association spectroscopy of the colliding species can lead to precision measurements of atomic and molecular parameters and how radiation fields can manipulate the outcome

of a collision itself. As a general orientation [figure C1.4.1](#) shows how a typical atomic de Broglie wavelength varies with temperature and where various physical phenomena are situated along the scale. With de Broglie wavelengths of the order of a few thousandths of a nanometre, conventional gas-phase chemistry can usually be interpreted as the interaction of classical nuclear point particles moving along potential surfaces defined by their associated electronic charge distribution. At one time liquid helium was thought to define a regime of cryogenic physics, but it is clear from [figure C1.4.1](#) that optical and evaporative cooling have created ‘cryogenic’ environments below liquid helium by many orders of magnitude. At the level of Doppler cooling and optical molasses the de Broglie wavelength becomes comparable to or longer than the chemical bond, approaching the length of the cooling optical light wave. Here we can expect wave and relativistic effects such as resonances, interferences and interaction retardation to become important. Following Suominen [9], we will term the Doppler cooling and optical molasses temperature range, roughly between 1 mK and 1 μ K, the regime of cold collisions. Most collision phenomena at this level are studied in the presence of one or more light fields used to confine the atoms and to probe their interactions. Excited quasimolecular states often play an important role. Below about 1 μ K, where evaporative cooling and Bose–Einstein condensation (BEC) become the focus of attention, the de Broglie wavelength grows to a scale comparable to the mean distance separating atoms in a dilute gas; quantum degenerate states of the atomic ensemble begin to appear. In this regime ground-state collisions only take place through radial (not angular) motion and are characterized by a phase shift, or scattering length, of the ground-state wave function. Since the atomic translational energy now lies below the kinetic energy transferred to an atom by recoil from a scattered photon, cooling with light fields can be of no further use; and collisions occurring in a temperature range from 1 μ K to 0 must be ground-state interactions.

-2-

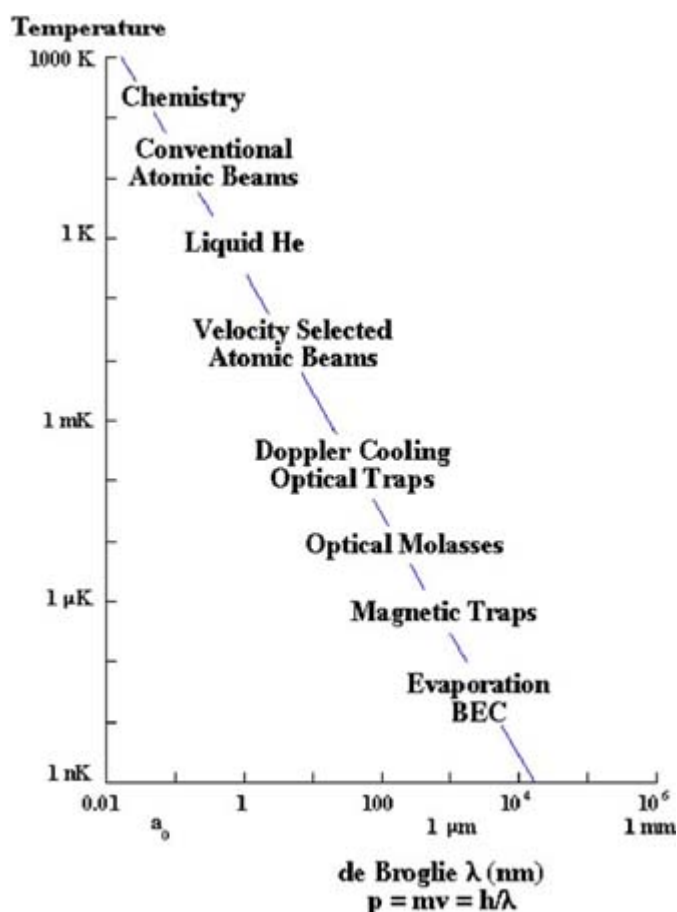


Figure C1.4.1. The situation of various physical phenomena along a scale of temperature plotted against de Broglie wavelength.

C1.4.2 THE PHYSICS OF NEUTRAL-ATOM TRAPS

C1.4.2.1 ATOM TRAPS

(A) LIGHT FORCES AND DOPPLER-LIMIT COOLING

It is well known that a light beam carries momentum and that the scattering of light by an object produces a force. This property of light was first demonstrated by Frisch [10] through the observation of a very small transverse deflection (3×10^{-5} rad) in a sodium atomic beam exposed to light from a resonance lamp. With the invention of the laser, it became easier to observe effects of this kind because the strength of the force is greatly enhanced by the use of intense and highly directional light fields, as demonstrated by Ashkin [11] with the manipulation of transparent dielectric spheres suspended in water. Although the results of Frisch and Ashkin rekindled interest in using light forces

-3-

to control the motion of neutral atoms, the basic groundwork for the understanding of light forces acting on atoms was not laid out before the end of the decade of the 1970s. Unambiguous experimental demonstration of atom cooling and trapping was not accomplished before the mid-80s. In this section we discuss some fundamental aspects of light forces and schemes employed to cool and trap neutral atoms.

The light force exerted on an atom can be of two types: a dissipative, *spontaneous force* and a conservative, *dipole force*. The spontaneous force arises from the impulse experienced by an atom when it absorbs or emits a quantum of photon momentum. When an atom scatters light, the resonant scattering cross section can be written as $\sigma_0 = \lambda_0^2/2\pi$ where λ_0 is the on-resonant wavelength. In the optical region of the electromagnetic spectrum the wavelengths of light are of the order of several hundreds of nanometres, so resonant scattering cross sections become quite large, $\sim 10^{-9}$ cm². Each photon absorbed transfers a quantum of momentum $\hbar k$ to the atom in the direction of propagation (\hbar is the Planck constant divided by 2π , and $k = 2\pi / \lambda$ is the magnitude of the wave vector associated with the optical field). The spontaneous emission following the absorption occurs in random directions; and, over many absorption–emission cycles, it averages to zero. As a result, the *net* spontaneous force acts on the atom in the direction of the light propagation, as shown schematically in the diagram of figure C1.4.2. The saturated rate of photon scattering by spontaneous emission (the reciprocal of the excited-state lifetime) fixes the upper limit to the force magnitude. This force is sometimes called *radiation pressure*.

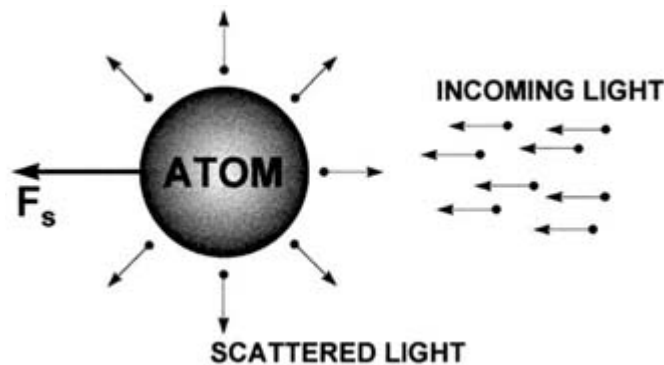


Figure C1.4.2. Spontaneous emission following absorption occurs in random directions, but absorption from a light beam occurs along only one direction.

The dipole force can be readily understood by considering the light as a classical wave. It is simply the time-averaged force arising from the interaction of the transition dipole, induced by the oscillating electric field of the light, with the gradient of the electric field amplitude. Focusing the light beam controls the magnitude of this gradient and detuning the optical frequency below or above the atomic transition controls the sign of the force acting on the atom. Tuning the light below resonance attracts the atom to the centre of the light beam while tuning above resonance repels it. The dipole force is a stimulated process in which no net exchange of energy between the

field and the atom takes place, but photons are absorbed from one mode and reappear by stimulated emission in another. Momentum conservation requires that the change of photon propagation direction from initial to final mode imparts a net recoil to the atom. Unlike the spontaneous force, there is in principle no upper limit to the magnitude of the dipole force since it is a function only of the field gradient and detuning.

-4-

We can bring these qualitative remarks into focus by considering the amplitude, phase and frequency of a classical field interacting with an atomic transition dipole in a two-level atom. A detailed development of the following results is beyond the scope of the present article, but can be found elsewhere [12, 13]. The usual approach is semiclassical and consists in treating the atom as a two-level quantum system and the radiation as a classical electromagnetic field [14]. A full quantum approach can also be employed [15], but it will not be discussed here. What follows immediately is sometimes called the Doppler cooling model. It turns out that atoms with hyperfine structure in the ground state can be cooled below the Doppler limit predicted by this model; and, to explain this unexpected sub-Doppler cooling, models involving interaction between a slowly moving atom and the polarization gradient of a standing wave have been invoked. We will sketch briefly in the next section the physics of these polarization gradient cooling mechanisms.

The basic expression for the interaction *energy* is

$$U = -\boldsymbol{\mu}\mathbf{E} \quad (\text{C1.4.1a})$$

where $\boldsymbol{\mu}$ is the transition dipole and \mathbf{E} is the electric field of the light. The *force* is then the negative of the spatial gradient of the potential,

$$\mathbf{F} = -\nabla_{\mathbf{R}}U = \boldsymbol{\mu}\nabla_{\mathbf{R}}\mathbf{E} \quad (\text{C1.4.2})$$

where we have set $\nabla_{\mathbf{R}}\boldsymbol{\mu}$ equal to zero because there is no spatial variation of the dipole over the length scale of the optical field. The optical-cycle average of the force is expressed as

$$\langle \mathbf{F} \rangle = \langle \boldsymbol{\mu}\nabla_{\mathbf{R}}\mathbf{E} \rangle = \boldsymbol{\mu}[(\nabla_{\mathbf{R}}E_0)u - (E_0\nabla_{\mathbf{R}}(k_{\mathbf{R}}R))v] \quad (\text{C1.4.3})$$

where u and v arise from the steady-state solutions of the optical Bloch equations,

$$u = \frac{\Omega}{2} \frac{\Delta\omega_{\mathbf{L}}}{(\Delta\omega_{\mathbf{L}})^2 + (\Gamma/2)^2 + \Omega^2/2} \quad (\text{C1.4.4})$$

and

$$v = \frac{\Omega}{2} \frac{\Gamma/2}{(\Delta\omega_{\mathbf{L}})^2 + (\Gamma/2)^2 + \Omega^2/2}. \quad (\text{C1.4.5})$$

In equations (C1.4.4) and (C1.4.5) $\Delta\omega_{\mathbf{L}} = \omega - \omega_{\mathbf{R}}$ is the detuning of the optical field from the atomic transition frequency $\omega_{\mathbf{R}}$, Ω is the natural width of the atomic transition and ω is termed the Rabi frequency and reflects the

-5-

strength of the coupling between field and atom,

$$\Omega = -\frac{\mu E_0}{\hbar}. \quad (\text{C1.4.6})$$

In writing [equation \(C1.4.3\)](#) we have made use of the fact that the time-average dipole has in-phase and in-quadrature components,

$$\langle \boldsymbol{\mu} \rangle = 2\mu(u \cos \omega_L t - v \sin \omega_L t) \quad (\text{C1.4.7})$$

and the electric field of the light is given by the classical expression,

$$E = E_0[\cos(\omega t - k_L R)]. \quad (\text{C1.4.8})$$

The time-averaged force, [equation \(C1.4.3\)](#), consists of two terms: the first term is proportional to the gradient of the electric field amplitude; the second term is proportional to the gradient of the phase. Substituting [equation \(C1.4.4\)](#) and [equation \(C1.4.5\)](#) into [equation \(C1.4.3\)](#), we have for the two terms,

$$\begin{aligned} \langle \mathbf{F} \rangle = & \mu(\nabla_{\mathbf{R}} E_0) \frac{\Omega}{2} \left[\frac{\Delta\omega_L}{(\Delta\omega_L)^2 + (\Gamma/2)^2 + \Omega^2/2} \right] \\ & - \mu(E_0 \nabla_{\mathbf{R}}(-k_L R)) \frac{\Omega}{2} \left[\frac{\Gamma/2}{(\Delta\omega_L)^2 + (\Gamma/2)^2 + \Omega^2/2} \right]. \end{aligned} \quad (\text{C1.4.9})$$

The first term is the dipole force, sometimes called the trapping force, F_T , because it is a conservative force and can be integrated to define a trapping potential for the atom:

$$F_T = \mu(\nabla_{\mathbf{R}} E_0) \frac{\Omega}{2} \left[\frac{\Delta\omega_L}{(\Delta\omega_L)^2 + (\Gamma/2)^2 + \Omega^2/2} \right] \quad (\text{C1.4.10})$$

and

$$U_T = - \int F_T dR = \frac{\hbar \Delta\omega_L}{2} \ln \left[1 + \frac{\Omega^2/2}{(\Delta\omega_L)^2 + (\Gamma/2)^2} \right]. \quad (\text{C1.4.11})$$

-6-

The second term is the spontaneous force, sometimes called the cooling force, F_C , because it is a dissipative force and can be used to cool atoms,

$$F_C = \mu E_0 k_L \frac{\Omega}{2} \left[\frac{\Gamma/2}{(\Delta\omega_L)^2 + (\Gamma/2)^2 + \Omega^2/2} \right]. \quad (\text{C1.4.12})$$

Note that in [equation \(C1.4.10\)](#), the line-shape function in square brackets is dispersive and changes sign as the detuning $\Delta\omega_L$ changes sign from negative (red detuning) to positive (blue detuning). In [equation \(C1.4.12\)](#), the

line-shape function is absorptive, peaks at zero detuning and exhibits a Lorentzian profile. These two equations can be recast to bring out more of their physical content. The dipole force can be expressed as

$$F_T = -\frac{1}{2\Omega^2} \nabla \Omega^2 \hbar \Delta\omega_L \left[\frac{s}{1+s} \right] \quad (\text{C1.4.13a})$$

where s , the *saturation parameter*, is defined to be

$$s = \frac{\Omega^2/2}{(\Delta\omega_L)^2 + (\Gamma/2)^2}. \quad (\text{C1.4.14})$$

In equation (C1.4.14) the saturation parameter essentially defines a criterion to compare the time required for stimulated and spontaneous processes. If $s \ll 1$ then spontaneous coupling of the atom to the vacuum modes of the field is fast compared to the stimulated Rabi coupling and the field is considered weak. If $s \gg 1$ then the Rabi oscillation is fast compared to spontaneous emission and the field is said to be strong. Setting s equal to unity defines the saturation condition

$$\Omega_{\text{sat}} = \sqrt{2} \left(\frac{\Gamma}{2} \right) \quad (\text{C1.4.15})$$

and, as can be seen from the line-shape factor in equation (C1.4.12), the resonance line width is power broadened by a factor of $\sqrt{2}$. With the help of the definition of the Rabi frequency, [equation \(C1.4.6\)](#), and the light beam intensity,

$$I = \frac{1}{2} \epsilon_0 c E_0^2 \quad (\text{C1.4.16})$$

-7-

[equation \(C1.4.13a\)](#) can be written in terms of the gradient of the light intensity, the saturation parameter and the detuning,

$$F_T = -\frac{1}{4I} (\nabla I) \hbar \Delta\omega_L \left[\frac{s}{1+s} \right]. \quad (\text{C1.4.17})$$

Note that negative $\Delta\omega_L$ (red detuning) produces a force attracting the atom to the intensity maximum while positive $\Delta\omega_L$ (blue detuning) repels the atom away from the intensity maximum. The spontaneous force or cooling force can also be written in terms of the saturation parameter and the spontaneous emission rate,

$$F_C = \frac{\hbar k_L \Gamma}{2} \left[\frac{s}{1+s} \right] \quad (\text{C1.4.18})$$

which shows that this force is simply the rate of absorption and reemission of momentum quanta $\hbar k_L$ carried by a photon in the light beam. Note that as s increases beyond unity, F_C approaches $\frac{\hbar k_L \Gamma}{2}$, the maximum photon scattering rate. Furthermore, from the previous definitions of I, Ω and Ω_{sat} , we can write

$$(\text{C1.4.19})$$

$$\frac{I}{I_{\text{sat}}} = \frac{\Omega^2}{\Gamma^2/2}$$

and

$$F_C = \frac{\hbar k_L \Gamma}{2} \left[\frac{I/I_{\text{sat}}}{(\frac{2\Delta\omega_L}{\Gamma})^2 + I/I_{\text{sat}} + 1} \right]. \quad (\text{C1.4.20})$$

Now if we consider the atom moving in the $+z$ direction with velocity v_z and counterpropagating to the light wave detuned from resonance by $\Delta\omega_L$, the *net* detuning will be

$$\Delta\omega = \Delta\omega_L + k_L v_z \quad (\text{C1.4.21})$$

where the term $k_L v_z$ is the Doppler shift. The force F_- acting on the atom will be in the direction opposite to its motion. In general,

$$F_{\pm} = \pm \frac{\hbar k_L \Gamma}{2} \left[\frac{I/I_{\text{sat}}}{(\frac{2(\Delta\omega_L \mp k v_z)}{\Gamma})^2 + I/I_{\text{sat}} + 1} \right] \quad (\text{C1.4.22})$$

-8-

Suppose we have two fields propagating in the $\pm z$ directions and we take the net force $F = F_+ + F_-$. If $k v_z$ is small compared to Γ and $\Delta\omega_L$, then we find

$$F \simeq 4\hbar k \frac{I}{I_{\text{sat}}} \frac{k v_z (2\Delta\omega_L / \Gamma)}{[1 + (2\Delta\omega_L / \Gamma)^2]^2}. \quad (\text{C1.4.23})$$

This expression shows that if the detuning $\Delta\omega_L$ is negative (i.e. red detuned from resonance), then the cooling force will oppose the motion and be proportional to the atomic velocity. The one-dimensional motion of the atom, subject to an opposing force proportional to its velocity, is described by a damped harmonic oscillator. The Doppler damping or friction coefficient is the proportionality factor,

$$\alpha_d = -4\hbar k^2 \frac{I}{I_{\text{sat}}} \frac{(2\Delta\omega_L / \Gamma)}{[1 + (2\Delta\omega_L / \Gamma)^2]^2} \quad (\text{C1.4.24})$$

and the characteristic time to damp the kinetic energy of the atom of mass m to $1/e$ of its initial value is

$$\tau = \frac{m}{2\alpha_d}. \quad (\text{C1.4.25})$$

However, the atom will not cool indefinitely. At some point the Doppler cooling rate will be balanced by the heating rate coming from the momentum fluctuations of the atom absorbing and re-emitting photons. Setting these two rates equal and associating the one-dimensional kinetic energy with $\frac{1}{2}k_B T$, we find

$$k_B T = \frac{\hbar \Gamma}{4} \frac{1 + (2\Delta\omega_L/\Gamma)^2}{2|\Delta\omega_L|/\Gamma}. \quad (\text{C1.4.26})$$

This expression shows that T is a function of the laser detuning, and the minimum temperature is obtained when $\Delta\omega_L = -\frac{\Gamma}{2}$. At this detuning,

$$k_B T = \hbar \frac{\Gamma}{2} \quad (\text{C1.4.27})$$

which is called the Doppler-cooling limit. This limit is typically, for alkali atoms, on the order of a few hundred microKelvin. For example the Doppler cooling limit for Na is $T = 240$ microKelvin. In the early years of cooling and trapping, prior to 1988, the Doppler limit was thought to be a real physical barrier. Then the experimental measurements of Lett *et al* [16] showed that in fact Na atoms could be cooled well below the Doppler limit. Although the physics of this sub-Doppler cooling in three dimensions is still not fully understood, the essential role played by the hyperfine structure of the ground state has been worked out in one-dimensional models which we describe here.

-9-

(B) SUB-DOPPLER COOLING

Two principal mechanisms which cool atoms to temperatures below the Doppler limit rely on spatial polarization gradients of the light field through which the atoms move [17]. These two mechanisms, however, invoke very different physics, and are distinguished by the spatial polarization dependence of the light field. Two parameters, the friction coefficient and the velocity capture range, determine the significance of these cooling processes. In this section we compare expressions for these quantities in the sub-Doppler regime to those found in the conventional Doppler cooling model of one-dimensional optical molasses.

In the first case two counterpropagating light waves with orthogonal polarization form a standing wave. This arrangement is familiarly called the ‘lin–perp–lin’ configuration. [Figure C1.4.3](#) shows what happens. We see from the figure that if we take as a starting point a position where the light polarization is linear ϵ_1 , it evolves from linear to circular over a distance of $\lambda/8$ (σ_-). Then over the next $\lambda/8$ interval the polarization again changes to linear but in the direction orthogonal to the first (ϵ_2). Then from $\lambda/4$ to $3\lambda/8$ the polarization again becomes circular but in the sense opposite (σ_+) to the circular polarization at $\lambda/8$, and finally after a distance of $\lambda/2$ the polarization is again linear but antiparallel to (ϵ_1). Over the same half-wavelength distance of the polarization period, atom–field coupling produces a periodic energy (or light) shift in the hyperfine levels of the atomic ground state. To illustrate the cooling mechanism we assume the simplest case, a $J_g = \frac{1}{2} \rightarrow J_e = \frac{3}{2}$ transition. As shown in [figure C1.4.4](#) the atom moving through the region of z around $\lambda/8$, where the polarization is primarily σ_- , will have its population pumped mostly into $J_g = -\frac{1}{2}$. Furthermore the Clebsch–Gordan coefficients controlling the transition dipole coupling to $J_e = \frac{3}{2}$ impose that the $J_g = -\frac{1}{2}$ level couples to σ_- light three times more strongly than does the $J_g = +\frac{1}{2}$ level. The difference in coupling strength leads to the light shift splitting between the two ground states shown in [figure C1.4.4](#). As the atom continues to move to $+z$, the relative coupling strengths are reversed around $3\lambda/8$ where the polarization is essentially σ_+ . Thus the relative energy levels of the two hyperfine ground states oscillate ‘out of phase’ as the atom moves through the standing wave. The key idea is that the optical pumping rate, always redistributing population to the lower-lying hyperfine level, lags the light shifts experienced by the two atom ground-state components as the atom moves through the field. The result is a ‘Sisyphus effect’ where an atom cycles through a period in which it spends most of its time climbing a potential hill, converting kinetic energy to potential energy, subsequently dissipating the accumulated potential energy into the empty modes of the radiation field and simultaneously transferring population back to the lower lying of the two ground-state levels. [Figure C1.4.5](#) illustrates the optical pumping phase lag. In order for this cooling mechanism to work, the optical pumping time, controlled by the light intensity, must be less than the light-shift time, controlled essentially by the velocity of the atom. Since the atom is moving slowly, having been previously cooled by the Doppler mechanism, the light

field must be weak in order to slow the optical pumping rate so that it lags the light-shift modulation rate. This physical picture combines the conservative optical dipole force, whose space integral gives rise to the potential hills and valleys over which the atom moves and the irreversible energy dissipation of spontaneous emission required to achieve cooling. We can make the discussion more precise and obtain simple expressions for the friction coefficient and velocity capture by establishing some definitions. As in the Doppler cooling model we define the friction coefficient $\alpha_{|p|}$ to be the proportionality constant between the force F and the atomic velocity v .

-10-

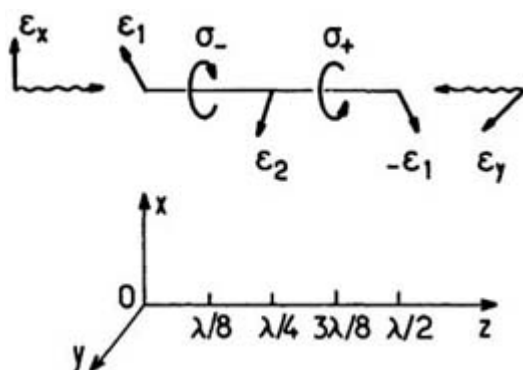


Figure C1.4.3. Schematic diagram of the ‘lin–perp–lin’ configuration showing spatial dependence of the polarization in the standing-wave field (after [17]).

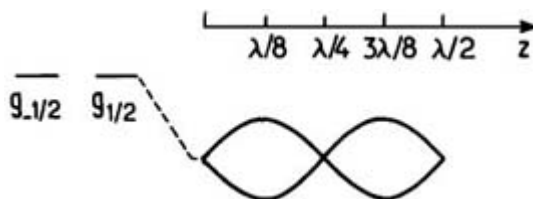


Figure C1.4.4. Schematic diagram showing how the two $\pm\frac{1}{2}$ levels of the ground state couple to the spatially varying polarization of the ‘lin–perp–lin’ standing wave light field (after [17]).

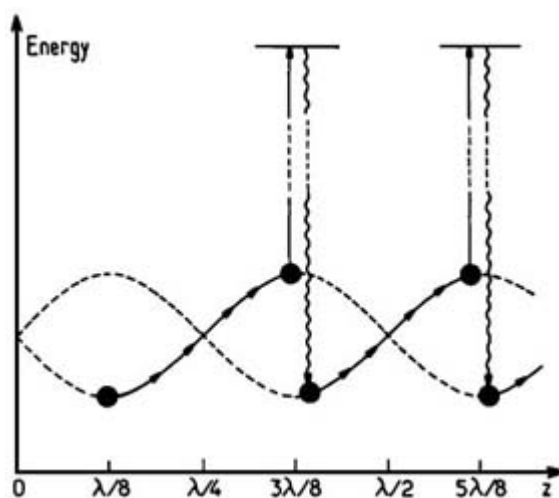


Figure C1.4.5. Population modulation as the atom moves through the standing wave in the ‘lin–perp–lin’ one dimensional optical molasses. The population lags the light shift such that kinetic is converted to potential energy then dissipated into the empty modes of the radiation field by spontaneous emission (after [17]).

$$F = -\alpha_{|p|} v. \quad (C1.4.28)$$

We assume that the light field is detuned to the red of the $J_g \rightarrow J_e$ atomic resonance frequency,

$$\Delta_L = \omega_L - \omega_R \quad (C1.4.29)$$

and term the light shifts of the $J_g = \pm \frac{1}{2}$ levels Δ_{\pm} respectively. At the position $z = \lambda/8$, $\Delta_- = 3\Delta_+$ and at $z = 3\lambda/8$, $\Delta_+ = 3\Delta_-$. Since the applied field is red detuned, all Δ have negative values. Now in order for the cooling mechanism to be effective the optical pumping time τ_p should be comparable to the time required for the atom with velocity v to travel from the bottom to the top of a potential hill, $\frac{\lambda/4}{v}$,

$$\tau_p \simeq \frac{\lambda/4}{v} \quad (C1.4.30)$$

or

$$\Gamma' \simeq kv \quad (C1.4.31)$$

where $\Gamma' = 1/\tau_p$ and $\lambda/4 \simeq 1/k$, with $k = \frac{2\pi}{\lambda}$ the magnitude of the optical wave vector. Now the amount of energy W dissipated in one cycle of hill climbing and spontaneous emission is essentially the average energy splitting of the two light-shifted ground states, between, say, $z = \lambda/8$ and $3\lambda/8$ or $W \simeq -\hbar\Delta$. Therefore the *rate* of energy dissipation is

$$\frac{dW}{dt} = -\Gamma' \hbar \Delta. \quad (C1.4.32)$$

However in general the time-dependent energy change of a system can be always be expressed as $\frac{dW}{dt} = Fv$ so in this one-dimensional model and taking into account equation (C1.4.28) we can write

$$\frac{dW}{dt} = -\alpha_{|p|} v^2 = -\Gamma' \hbar \Delta \quad (C1.4.33)$$

so that

$$\alpha_{|p|} = -\frac{k\hbar\Delta}{v} = -\frac{k^2\hbar\Delta}{\Gamma'}. \quad (C1.4.34)$$

Note that since $\Delta < 0$, α_{lp1} is a positive quantity. Note also that at far detunings ($\Delta_L \gg \Gamma$) [equation \(C1.4.11\)](#) shows that $\Delta = \frac{\Omega^2}{4\Delta_L}$. It is also true that for light shifts large compared to the natural linewidth ($\Delta \gg \Gamma$), $\Gamma/\Gamma' = \frac{\Delta_L^2}{\Omega^2}$ so the sub-Doppler friction coefficient can also be written

$$\alpha_{\text{lp1}} = -\frac{k^2 \hbar \Delta_L}{4\Gamma'} \quad (\text{C1.4.35})$$

Equation (C1.4.35) yields two remarkable predictions: first, that the sub-Doppler friction coefficient can be a big number compared to α_d since at far detuning Δ_L/Γ is a big number; and second, that α_{lp1} is independent of the applied field intensity. This last result contrasts sharply with the Doppler friction coefficient which is proportional to field intensity up to saturation (see [equation \(C1.4.24\)](#)). However, even though α_{lp1} looks impressive, the range of atomic velocities over which it can operate are restricted by the condition that $\Gamma' \simeq kv$. The ratio of the capture velocities for Doppler versus sub-Doppler cooling is therefore only $v_{\text{lp1}}/v_d \simeq \frac{4\Delta}{\Delta_L}$. [Figure C1.4.6](#) illustrates graphically the comparison between the Doppler and the ‘lin-perp-lin’ sub-Doppler cooling mechanism. The dramatic difference in capture range is evident from the figure. Note also that the slopes of the curves give the friction coefficients for the two regimes and that, within the narrow velocity capture range of its action, the slope of the sub-Doppler mechanism is markedly steeper.

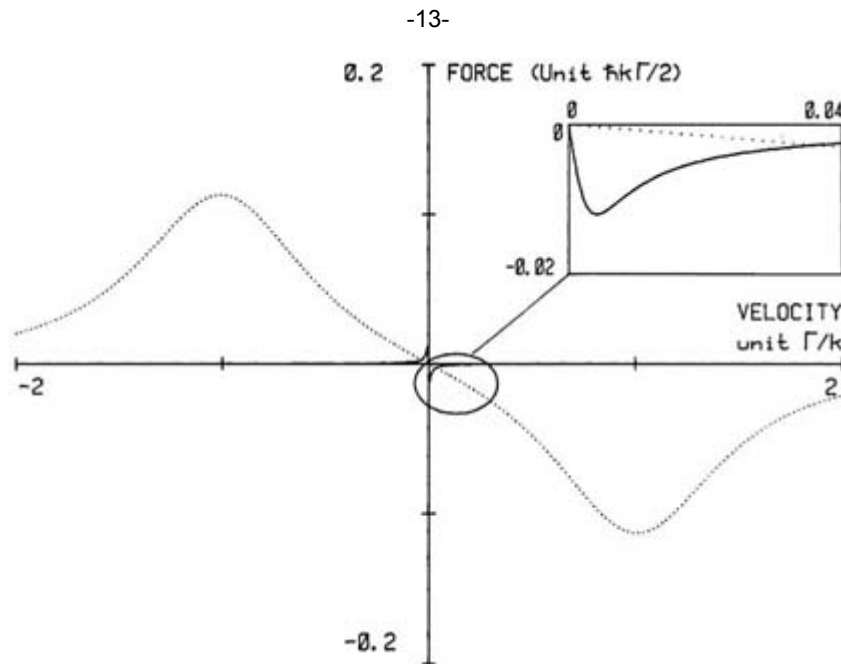


Figure C1.4.6. Comparison of capture velocity for Doppler cooling and ‘lin-perp-lin’ sub-Doppler cooling. Notice that the slope of the curves, proportional to the friction coefficient, is much steeper for the sub-Doppler mechanism. (After [17].)

The second mechanism operates with the two counterpropagating beams circularly polarized in opposite senses. When the two counterpropagating beams have the same amplitude, the resulting polarization is always linear and orthogonal to the propagation axis, but the tip of the polarization axis traces out a helix around the propagation axis with a pitch of λ . [Figure C1.4.7](#) illustrates this case. The physics of the sub-Doppler mechanism does not rely on hill-climbing and spontaneous emission, but on an imbalance in the photon scattering rate from the two counterpropagating light waves as the atom moves along the z axis. This imbalance leads to a velocity-dependent restoring force acting on the atom. The essential factor leading to the differential scattering rate is the creation of *population orientation* along the z axis among the sublevels of the atom *ground state*. Those sublevels with more population scatter more photons. Now it is evident from a consideration of the energy level diagram and the

Clebsch–Gordan coefficients coupling ground and excited levels that $J_g = \frac{1}{2} \leftrightarrow J_e = \frac{3}{2}$ transitions coupled by linearly polarized light cannot produce a population orientation in the ground state. In fact the simplest system to exhibit this effect is $J_g = 1 \leftrightarrow J_e = 2$, and a measure of the orientation is the magnitude of the $\langle J_z \rangle$ matrix element between the $J_{gz} = \pm$ sublevels. If the atom remained stationary at $z = 0$, interacting with the light polarized along y , the light shifts Δ_0, Δ_1 of the three ground-state sublevels would be

$$\Delta_{+1} = \Delta_{-1} = \frac{3}{4} \Delta_0 \quad (\text{C1.4.36})$$

and the steady-state populations 4/17, 4/17 and 9/17 respectively. Evidently, linearly polarized light will not produce a net steady-state orientation, $\langle J_z \rangle$. As the atom begins to move along z with velocity v , however, it sees a linear polarization precessing around its axis of propagation with an angle $\varphi = -kz = -kvt$. This precession gives rise to a new

-14-

term in the Hamiltonian, $V = kvJ_z$. Furthermore, if we transform to a rotating coordinate frame, the eigenfunctions belonging to the Hamiltonian of the moving atom in this new ‘inertial’ frame become linear combinations of the basis functions with the atom at rest. Evaluation of the steady-state orientation operator, J_z , in the inertial frame is now nonzero,

$$\langle J_z \rangle = \frac{40 \hbar kv}{17 \Delta_0} = \hbar[\Pi_{+1} - \Pi_{-1}]. \quad (\text{C1.4.37})$$

Notice that the orientation measure is only nonzero when the atom is moving. In equation (C1.4.37) we denote the populations of the $|\pm\rangle$ sublevels as Π_{\pm} , and we interpret the nonzero matrix element as a direct measure of the population difference between the $|\pm\rangle$ levels of the ground state. Note that since Δ_0 is a negative quantity (red detuning), equation (C1.4.37) tells us that the Π_- population is greater than the Π_+ population. Now, if the atom travelling in the $+z$ direction is subject to two light waves, one with polarization σ_- (σ_+) propagating in the $-z$ ($+z$) direction, the preponderance of population in the $|- \rangle$ level will result in a higher scattering rate from the wave travelling in the $-z$ direction. Therefore the atom will be subject to a net force opposing its motion and proportional to its velocity. The differential scattering rate is $\frac{40 kv}{17 \Delta_0} \Gamma'$ and, with an $\hbar k$ momentum quantum transferred per scattering event, the net force is

$$F = -\frac{40 \hbar k^2 v \Gamma'}{17 \Delta_0}. \quad (\text{C1.4.38})$$

The friction coefficient α_{cp} is evidently

$$\alpha_{cp} = -\frac{40}{17} \hbar k^2 \frac{\Gamma'}{\Delta_0} \quad (\text{C1.4.39})$$

which is a positive quantity since Δ_0 is negative from red detuning. Contrasting α_{cp} with α_{lpl} we see that α_{cp} must be much smaller since the assumption has been all along that the light shifts Δ were much greater than the line widths Γ' . It turns out, however [17], that the heating rate from recoil fluctuations is also much smaller so that the ultimate temperatures reached from the two mechanisms are comparable.

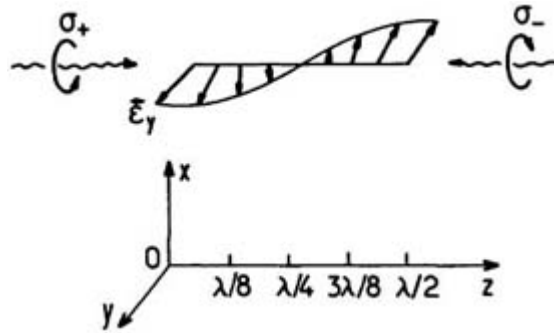


Figure C1.4.7. Spatial variation of the polarization from the field resulting from two counterpropagating, circularly polarized fields with equal amplitude but polarized in opposite senses. Note that the polarization remains linear but that the axis rotates in the x - y plane with a helical pitch along the z axis of length λ .

Although the Doppler cooling mechanism also depends on a scattering imbalance from oppositely travelling light waves, the imbalance in the scattering rate originates from a difference in the scattering probability per photon due to the Doppler shift induced by the moving atom. In the sub-Doppler mechanism the scattering probabilities from the two light waves are equal but the ground-state populations are not. The state with the greater population experiences the greater rate.

(C) THE MAGNETO-OPTICAL TRAP (MOT)

Basic notions. [18] originally suggested that the spontaneous light force could be used to trap neutral atoms. The basic concept exploited the internal degrees of freedom of the atom as a way of circumventing the optical Earnshaw theorem (OET) proved by [19]. This theorem states that if a force is proportional to the light intensity, its divergence must be null because the divergence of the Poynting vector, which expresses the directional flow of intensity, must be null through a volume without sources or sinks of radiation. This null divergence rules out the possibility of an inward restoring force everywhere on a closed surface. However, when the internal degrees of freedom of the atom are considered, they can change the proportionality between the force and the Poynting vector in a position-dependent way such that the OET does not apply. Spatial confinement is then possible with spontaneous light forces produced by counterpropagating optical beams. Using these ideas to circumvent the OET, Raab *et al* [20] demonstrated a trap configuration that is currently the most commonly employed. It uses a radial magnetic field gradient produced by a quadrupole field and three pairs of circularly polarized, counterpropagating optical beams, detuned to the red of the atomic transition and intercepting at right angles in the position where the magnetic field is zero. The magneto-optical trap exploits the position-dependent Zeeman shifts of the electronic levels when the atom moves in the radially increasing magnetic field. The use of circularly polarized light, red-detuned $\sim\Gamma$ results in a spatially dependent transition probability whose net effect is to produce a restoring force that pushes the atom toward the origin.

To make clear how this trapping scheme works, consider a two-level atom with a $J = 0 \rightarrow J = 1$ transition moving along the z direction. We apply a magnetic field $B(z)$ increasing linearly with distance from the origin. The Zeeman shifts of the electronic levels are position dependent, as shown in [figure C1.4.8\(a\)](#). We also apply counterpropagating optical fields along the $\pm z$ directions carrying oppositely circular polarization and detuned to the red of the atomic

transition. It is clear from [figure C1.4.8](#) that an atom moving along $+z$ will scatter σ^- photons at a faster rate than

σ^+ photons because the Zeeman effect will shift the $\Delta M_J = -1$ transition closer to the light frequency.

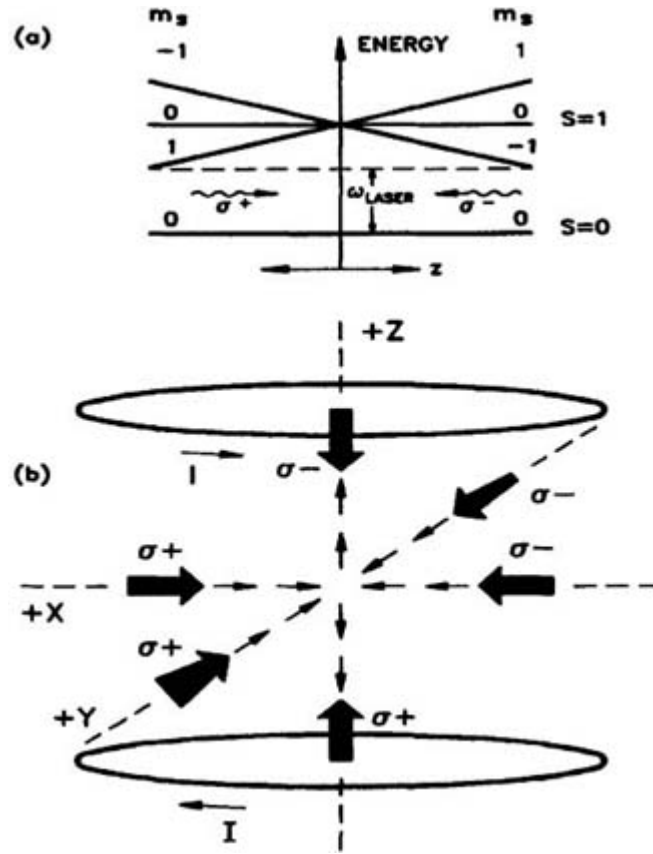


Figure C1.4.8. (a) An energy level diagram showing the shift of Zeeman levels as the atom moves away from the $z = 0$ axis. The atom encounters a restoring force in either direction from counterpropagating light beams. (b) A typical optical arrangement for implementation of a magneto-optical trap.

$$F_{1z} = -\frac{\hbar k}{2} \Gamma \frac{\Omega^2/2}{(\Delta + kv_z + \frac{\mu_B}{\hbar} \frac{dB}{dz} z)^2 + (\Gamma/2)^2 + \Omega^2/2}. \quad (\text{C1.4.40})$$

Similarly, if the atom moves along $-z$ it will scatter σ^+ photons at a faster rate from the $\Delta M_J = +1$ transition.

$$F_{2z} = +\frac{\hbar k}{2} \Gamma \frac{\Omega^2/2}{(\Delta - kv_z - \frac{\mu_B}{\hbar} \frac{dB}{dz} z)^2 + (\Gamma/2)^2 + \Omega^2/2}. \quad (\text{C1.4.41})$$

The atom will therefore experience a net restoring force pushing it back to the origin. If the light beams are red detuned $\sim \Gamma$, then the Doppler shift of the atomic motion will introduce a velocity-dependent term to the restoring force such that, for small displacements and velocities, the total restoring force can be expressed as the sum of a term linear in velocity and a term linear in displacement,

$$F_{\text{MOT}} = F_{1z} + F_{2z} = -\alpha\dot{z} - Kz.$$

Equation (C1.4.42) expresses the equation of motion of a damped harmonic oscillator with mass m ,

$$\ddot{z} + \frac{2\alpha}{m}\dot{z} + \frac{K}{m}z = 0. \quad (\text{C1.4.43})$$

The damping constant α and the spring constant K can be written compactly in terms of the atomic and field parameters as

$$\alpha = \hbar k \Gamma \frac{16|\Delta'|(\Omega')^2(k/\Gamma)}{[1 + 2(\Omega')^2]^2 [1 + \frac{4(\Delta')^2}{1+2(\Omega')^2}]^2} \quad (\text{C1.4.44})$$

and

$$K = \hbar k \Gamma \frac{16|\Delta'|(\Omega')^2(\frac{d\omega_0}{dz})}{[1 + 2(\Omega')^2]^2 [1 + \frac{4(\Delta')^2}{1+2(\Omega')^2}]^2} \quad (\text{C1.4.45})$$

where Ω', Δ' and $\frac{d\omega_0}{dz} = \frac{(\mu_B/\hbar)(\frac{dB}{dz})}{\Gamma}$ are Γ -normalized analogues of the quantities defined earlier. Typical MOT operating conditions fix $\Omega' = 1/2$, $\Delta' = 1$, so α and K reduce to

$$\alpha \simeq (0.132)\hbar k^2 \quad (\text{C1.4.46})$$

and

$$K \simeq (1.16 \times 10^{10})\hbar k \frac{dB}{dz}. \quad (\text{C1.4.47})$$

The extension of these results to three dimensions is straightforward if one takes into account that the quadrupole field gradient in the z direction is twice the gradient in the x, y directions, so that $K_z = 2K_x = 2K_y$. The velocity dependent damping term implies that kinetic energy E dissipates from the atom (or collection of atoms) as $E/E_0 = e^{-\frac{2\alpha}{m}t}$ where m is the atomic mass and E_0 the kinetic energy at the beginning of the cooling process. Therefore, the dissipative force term cools the collection of atoms as well as combining with the displacement term to confine them. The damping time constant $\tau = \frac{m}{2\alpha}$ is typically tens of microseconds. It is important to bear in mind that an MOT is anisotropic since the restoring force along the z axis of the quadrupole field is twice the restoring force in the x - y plane. Furthermore, an MOT provides a dissipative rather than a conservative trap and it is therefore more accurate to characterize the maximum capture velocity rather than the trap ‘depth’.

Early experiments with MOT-trapped atoms were carried out by initially slowing an atomic beam to load the trap [20, 21]. Later, a continuous uncooled source was used for that purpose, suggesting that the trap could be loaded with the slow atoms of a room-temperature vapour [22]. The next advance in the development of magneto-optical trapping was the introduction of the vapour-cell magneto-optical trap (VCMOT). This variation captures cold atoms directly from the low-velocity edge of the Maxwell–Boltzmann distribution always present in a cell

background vapour [23]. Without the need to load the MOT from an atomic beam, experimental apparatus became simpler and now many groups around the world use the VCMOT for applications ranging from precision spectroscopy to optical control of reactive collisions.

Densities in an MOT. The VCMOT typically captures about a million atoms in a volume less than a millimetre in diameter, resulting in densities $\sim 10^{10} \text{ cm}^{-3}$. Two processes limit the density attainable in an MOT: (1) collisional trap loss and (2) repulsive forces between atoms caused by reabsorption of scattered photons from the interior of the trap [21, 24]. Collisional loss in turn arises from two sources: hot background atoms that knock cold atoms out of the MOT by elastic impact and binary encounters between the cold atoms themselves. Trap loss due to cold collisions is the topic of section C1.4.3. The ‘photon-induced repulsion’ or photon trapping arises when an atom near the MOT centre spontaneously emits a photon which is reabsorbed by a another atom before the photon can exit the MOT volume. This absorption results in an increase of $2\hbar k$ in the relative momentum of the atomic pair and produces a repulsive force proportional to the product of the absorption cross section for the incident light beam and scattered fluorescence. When this outward repulsive force balances the confining force, further increase in the number of trapped atoms leads to larger atomic clouds, but not to higher densities.

(D) DARK SPOT

In order to overcome the ‘photon-induced repulsion’ effect, Ketterle *et al* [25] proposed a method that allows the atoms to be optically pumped to a ‘dark’ hyperfine level of the atom ground state that does not interact with the trapping light. In a conventional MOT one usually employs an auxiliary ‘repumper’ light beam, copropagating with the trapping beams but tuned to a neighbouring transition between hyperfine levels of ground and excited states. The repumper recovers population that leaks out of the cycling transition between the two levels used to produce the MOT. As an example, figure C1.4.9 shows the trapping and repumping transitions usually employed in an Na MOT. The scheme, known as a dark spontaneous-force optical trap (dark SPOT), passes the repumper through a glass plate with a small black dot shadowing the beam such that the atoms at the trap centre are not coupled back to the cycling transition but spend most of their time ($\sim 99\%$) in the ‘dark’ hyperfine level. Cooling and confinement continue to

function on the periphery of the MOT but the centre core experiences no outward light pressure. The dark SPOT increases density by almost two orders of magnitude.

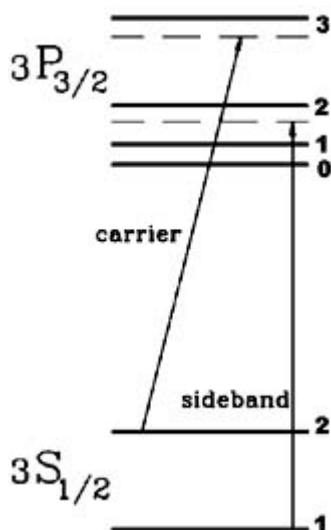


Figure C1.4.9. Usual cooling (carrier) and repumping (sideband) transitions when optically cooling Na atoms. The repumper frequency is normally derived from the cooling transition frequency with electro-optic modulation.

Dashed lines show that lasers are tuned about one natural linewidth to the red of the transition frequencies.

(E) THE FAR-OFF RESONANCE TRAP (FORT)

Although an MOT functions as a versatile and robust ‘reaction cell’ for studying cold collisions, light frequencies must tune close to atomic transitions and an appreciable steady-state fraction of the atoms remain excited. Excited-state trap-loss collisions and photon-induced repulsion limit achievable densities.

A far-off resonance trap (FORT), in contrast, uses the dipole force rather than the spontaneous force to confine atoms and can therefore operate far from resonance with negligible population of excited states. A hybrid MOT/dipole-force trap was used by a NIST–Maryland collaboration [26] to study cold collisions, and a FORT was demonstrated by Miller *et al* [27] for ^{85}Rb atoms. The FORT consists of a single, linearly polarized, tightly focused Gaussian-mode beam tuned far to the red of resonance. The obvious advantage of large detunings is the suppression of photon absorption. Note from [equation \(C1.4.12\)](#) that the spontaneous force, involving absorption and reemission, falls off as the square of the detuning, while [equation \(C1.4.11\)](#) shows that the potential derived from the dipole force falls off only as the detuning itself. At large detunings and high field gradients (tight focus) [equation \(C1.4.11\)](#) becomes

$$U \simeq \frac{\hbar\Omega^2}{4\Delta\omega_L} \tag{C1.4.48}$$

which shows that the potential becomes directly proportional to light intensity and inversely proportional to detuning. Therefore, at far detuning but high intensity the depth of the FORT can be maintained but most of the atoms will not absorb photons. The important advantages of FORTs compared to MOTs are (1) high density ($\sim 10^{12} \text{ cm}^{-3}$) and (2) a well defined polarization axis along which atoms can be aligned or oriented (spin polarized). The main disadvantage is the small number of trapped atoms due to small FORT volume. The best number achieved is about 10^4 atoms [28].

(F) MAGNETIC TRAPS

Pure magnetic traps have also been used to study cold collisions and they are critical for the study of dilute gas-phase Bose–Einstein condensates (BECs) in which collisions figure importantly. We anticipate, therefore, that magnetic traps will play an increasingly important role in future collision studies in and near BEC conditions.

The most important distinguishing feature of all magnetic traps is that they do not require light to provide atom containment. Light-free traps reduce the rate of atom heating by photon absorption to zero, an apparently necessary condition for the attainment of BEC. Magnetic traps rely on the interaction of atomic spin with variously shaped magnetic fields and gradients to contain atoms. The two governing equations are

$$U = -\mu_S B = -\frac{g_S \mu_B}{\hbar} \mathbf{S} B = -\frac{g_S \mu_B}{\hbar} M_S B \tag{C1.4.49}$$

and

$$\mathbf{F} = -\frac{g_S \mu_B}{\hbar} M_S \nabla B. \tag{C1.4.50}$$

If the atom has nonzero nuclear spin I then $\mathbf{F} = \mathbf{S} + \mathbf{I}$ substitutes for \mathbf{S} in equation (C1.4.49), the g-factor generalizes to

$$g_F \cong g_S \frac{F(F+1) + S(S+1) - I(I+1)}{2F(F+1)} \quad (\text{C1.4.51})$$

and

$$\mathbf{F} = -\frac{g_F \mu_B}{\hbar} M_F \nabla B. \quad (\text{C1.4.52})$$

Depending on the sign of U and F , atoms in states whose energy increases or decreases with magnetic field are called ‘weak-field seekers’ or ‘strong-field seekers’, respectively. One could, in principle, trap atoms in any of these states,

-21-

needing only to produce a minimum or a maximum in the magnetic field. Unfortunately only weak-field seekers can be trapped in a static magnetic field because such a field in free space can only have a minimum. Dynamic traps have been proposed to trap both weak- and strong-field seekers [29]. Even when weak-field seeking states are not in the lowest hyperfine levels they can still be used for trapping because the transition rate for spontaneous magnetic dipole emission is $\sim 10^{-10} \text{ s}^{-1}$. However, spin-changing collisions can limit the maximum attainable density. The first static magnetic field trap for neutral atoms was demonstrated by Migdall *et al* [30]. An anti-Helmholtz configuration, similar to an MOT, was used to produce an axially symmetric quadrupole magnetic field. Since this field design always has a central point of vanishing magnetic field, nonadiabatic Majorana transitions can take place as the atom passes through the zero point, transferring the population from a weak-field to a strong-field seeker and effectively ejecting the atom from the trap. This problem can be overcome by using a magnetic bottle with no point of zero field [31, 32, 33 and 34]. The magnetic bottle, also called the Ioffe–Pritchard trap, was recently used to achieve BEC in a sample of Na atoms pre-cooled in an MOT [35]. Other approaches to eliminating the zero-field point are the time-averaged orbiting potential (TOP) trap [36] and an optical ‘plug’ [37] that consists of a blue-detuned intense optical beam aligned along the magnetic trap symmetry axis and producing a repulsive potential to prevent atoms from entering the null-field region. Trap technology continues to develop and the recent achievement of BEC will stimulate more robust traps containing greater numbers of atoms. At present $\sim 10^7$ atoms can be trapped in a BEC loaded from an MOT containing $\sim 10^9$ atoms.

C1.4.3 INELASTIC EXOERGIC COLLISIONS IN MOTS

An exoergic collision converts internal atomic energy to kinetic energy of the colliding species. When there is only one species in the trap (the usual case) this kinetic energy is equally divided between the two partners. If the net gain in kinetic energy exceeds the trapping potential or the ability of the trap to recapture, the atoms escape; and the exoergic collision leads to trap loss.

Of the several trapping possibilities described in the last section, by far the most popular choice for collision studies has been the magneto-optical trap (MOT). An MOT uses spatially dependent resonant scattering to cool and confine atoms. If these atoms also absorb the trapping light at the initial stage of a binary collision and approach each other on an excited molecular potential, then during the time of approach the colliding partners can undergo a fine-structure-changing collision (FCC) or relax to the ground state by spontaneously emitting a photon. In either case, electronic energy of the quasimolecule converts to nuclear kinetic energy. If both atoms are in their electronic ground states from the beginning to the end of the collision, only elastic and hyperfine changing (HCC) collisions

can take place. Elastic collisions (identical scattering entrance and exit states) are not exoergic but figure importantly in the production of Bose–Einstein condensates (BECs). At the very lowest energies only s waves contribute to the elastic scattering and in this regime the collisional interaction is characterized by the scattering length. The sign of the scattering length determines the properties of a weakly interacting Bose gas and the magnitude controls the rate of evaporative cooling needed to achieve BEC. The HCC collisions arise from ground-state splitting of the alkali atoms into hyperfine levels due to various orientations of the nonzero nuclear spin. A transition from higher to lower molecular hyperfine level during the collisional encounter releases kinetic energy. In the absence of external light fields HCCs often dominate trap heating and loss.

-22-

If the collision starts on the excited level, the long-range dipole–dipole interaction produces an interatomic potential varying as $\pm C_3/R^3$. The sign of the potential, attractive or repulsive, depends on the relative phase of the interacting dipoles. For the trap-loss processes that concern us in this chapter we concentrate on the attractive long-range potential, $-C_3/R^3$. Due to the extremely low energy of the collision, this long-range potential acts on the atomic motion even when the pair are as far apart as $\lambda/2\pi$ (the inverse of the light-field wave vector k). Since the collision time is comparable to or greater than the excited-state lifetime, spontaneous emission can take place during the atomic encounter. If spontaneous emission occurs, the quasimolecule emits a photon red shifted from atomic resonance and relaxes to the ground electronic state with some continuum distribution of the nuclear kinetic energy. This conversion of internal electronic energy to external nuclear kinetic energy can result in a considerable increase in the nuclear motion. If the velocity is not too high, the dissipative environment of the MOT is enough to cool this radiative heating, allowing the atom to remain trapped. However, if the transferred kinetic energy is greater than the recapture ability of the MOT, the atoms escape the trap. This process constitutes an important trap-loss mechanism termed radiative escape (RE), and was first pointed out by Vigué [38]. For alkalis there is also another exoergic process involving excited–ground collisions. Due to the existence of fine structure in the excited state ($P_{3/2}$ and $P_{1/2}$), the atomic encounter can result in FCC, releasing Δ_{FS} of kinetic energy, shared equally between both atoms. For example in sodium $\frac{\Delta E_{FS}}{2} \simeq 12 \text{ K}$, which can easily cause the escape of both atoms from the MOT, typically 1 K deep.

These three effects, HCC, RE and FCC, are the main exoergic collisional process that take place in an MOT. They are the dominant loss mechanisms which usually limit the maximum attainable density and number in MOTs. They are not, however, the only type of collision in the trap.

C1.4.3.1 PHOTOASSOCIATION AT AMBIENT AND ULTRACOLD TEMPERATURES

The first measurement of a free–bound photoassociative absorption appeared long before the development of optical cooling and trapping, about two decades ago, when Scheingraber and Vidal [39] reported the observation of photoassociation in collisions between magnesium atoms. In this experiment fixed UV lines from an argon ion laser excited free–bound transitions from the thermal continuum population of the ground $X^1X^1\Sigma_g^+$ state to bound levels of the $A^1A^1\Sigma_u^+$ state of Mg_2 . Scheingraber and Vidal analysed the subsequent fluorescence to bound and continuum states from which they inferred the photoassociative process. The first unambiguous photoassociation *excitation spectrum*, however, was measured by Inoue *et al* [40] in collisions between Xe and Cl at 300 K. In both these early experiments the excitation was not very selective due to the broad thermal distribution of populated continuum ground states. Jones *et al* [41], with a technically much improved experiment, reported beautiful free–bound vibration progressions in KrF and XeI $X \rightarrow B$ transitions; and, from the intensity envelope modulation, were able to extract the functional dependence of the transition moment on the internuclear separation. Although individual vibrational levels of the B state were clearly resolved, the underlying rotational manifolds were not. Jones *et al* [41] simulated the photoassociation structure and line shapes by assuming a thermal distribution of rotational levels at 300 K. Photoassociation and dissociation processes prior to the cold and ultracold epoch have been reviewed by Tellinghuisen [42].

A decade after Schenigraber and Vidal reported the first observation of photoassociation, Thorsheim *et al* [43]

proposed that high-resolution free-bound molecular spectroscopy should be possible using optically cooled and confined atoms. Figure C1.4.10 shows a portion of their calculated $X \rightarrow A$ absorption spectrum at 10 mK for sodium atoms. This figure illustrates how cold temperatures compress the Maxwell-Boltzmann distribution to the point where

-23-

individual rotational transitions in the free-bound absorption are clearly resolvable. The marked differences in peak intensities indicate scattering resonances, and the asymmetry in the line shapes, tailing off to the red, reflect the thermal distribution of ground-state collision energies at 10 mK. Figure C1.4.11 plots the photon-flux-normalized absorption rate coefficient for singlet $X^1X^1\Sigma_g^+ \rightarrow A^1A^1\Sigma_u^+$ and triplet $a^3A^1\Sigma_u^+ \rightarrow 1^3X^1\Sigma_g^+$ molecular transitions over a broad range of photon excitation, red detuned from the Na ($^2S \rightarrow ^2P$) atomic resonance line. The strongly modulated intensity envelopes are called Condon modulations, and they reflect the overlap between the ground-state continuum wavefunctions and the bound excited vibrational wavefunctions. We shall see later that these Condon modulations reveal detailed information about the ground state scattering wave function and potential from which accurate s-wave scattering lengths can be determined. Thorsheim *et al* [44], therefore, predicted all the notable features of ultracold photoassociation spectroscopy later to be developed in many experiments: (1) precision measurement of vibration-rotation progressions from which accurate excited-state potential parameters can be determined, (2) line profile measurements and analysis to determine collision temperature and threshold behaviour and (3) spectral intensity modulation from which the ground-state potential, the scattering wave function and the s-wave scattering length can be characterized with great accuracy.

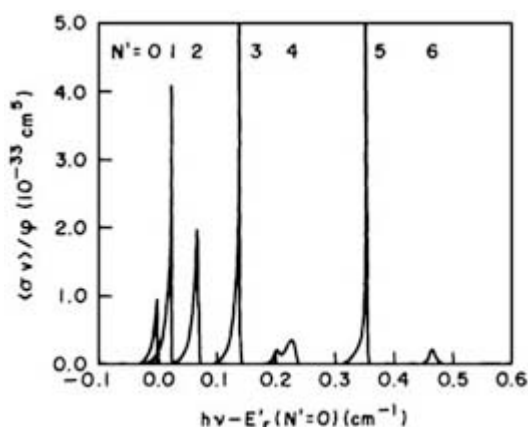


Figure C1.4.10. Calculated free-bound photoassociation spectrum at 10 mK.

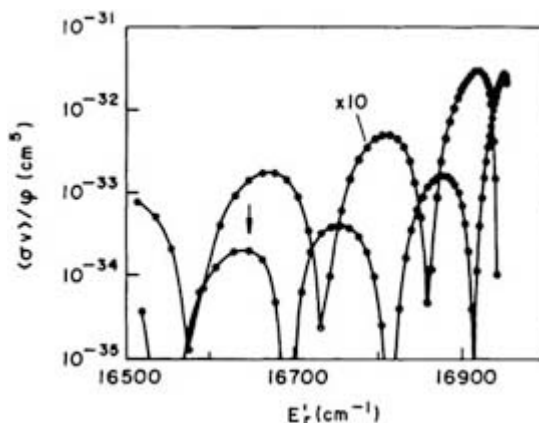


Figure C1.4.11. Calculated absorption spectrum of photoassociation in Na at 10 mK, showing Condon fluctuations.

An important difference distinguishes ambient-temperature photoassociation in rare-gas halide systems and sub-milliKelvin temperature photoassociation in cooled and confined alkali systems. At temperatures found in MOTs and FORTs (and within selected velocity groups in atomic beams) the collision dynamics are controlled by long-range electrostatic interactions, and Condon points R_C are typically at tens to hundreds of a_0 . In the case of the rare-gas halides the Condon points are in the short-range region of chemical binding and, therefore, free-bound transitions take place at much smaller internuclear distances, typically less than ten a_0 . For the colliding A,B quasimolecule the pair density n as a function of R is given by

$$n = n_A n_B 4\pi R^2 e^{-\frac{V(R)}{kT}} \quad (\text{C1.4.53})$$

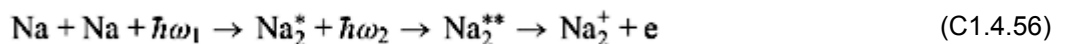
so the density of pairs varies as the square of the internuclear separation. Although the pair-density R dependence favours long-range photoassociation, the atomic reactant pressures are quite different with the $n_A n_B$ product of the order of 10^{35} cm^{-6} for rare-gas halide photoassociation and only about 10^{22} cm^{-6} for optically trapped atoms. Therefore the effective pair density available for rare-gas halide photoassociation greatly exceeds that for cold alkali photoassociation, permitting fluorescence detection and dispersion by high-resolution (but inefficient) monochromators.

C1.4.3.2 ASSOCIATIVE AND PHOTOASSOCIATIVE IONIZATION

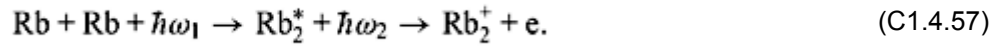
Conventional associative ionization (AI) occurring at ambient temperature proceeds in two steps: excitation of isolated atoms followed by molecular autoionization as the two atoms approach on excited molecular potentials. In sodium for example [44]



The collision event lasts a few picoseconds, fast compared to radiative relaxation of the excited atomic states (\sim tens of nanoseconds). Therefore the incoming atomic excited states can be treated as stationary states of the system Hamiltonian, and spontaneous radiative loss does not play a significant role. In contrast, cold and ultracold photoassociative ionization (PAI) must always start on ground states because the atoms move so slowly that radiative lifetimes become short compared to collision duration. The partners must be close enough at the Condon point, where the initial photon absorption takes place, so that a significant fraction of the excited scattering flux survives radiative relaxation and goes on to populate the final inelastic channel. Thus PAI is also a two-step process: (1) photoexcitation of the incoming scattering flux from the molecular ground-state continuum to specific vibration-rotation levels of a bound molecular state and (2) subsequent photon excitation either to a doubly excited molecular autoionizing state or directly to the molecular photoionization continuum. For example, in the case of sodium collisions the principal route is through doubly excited autoionization [45]



whereas for rubidium atoms the only available route is direct photoionization in the second step [46]



Collisional ionization can play an important role in plasmas, flames and atmospheric and interstellar physics and chemistry. Models of these phenomena depend critically on the accurate determination of absolute cross sections and rate coefficients. The rate coefficient is the quantity closest to what an experiment actually measures and can be regarded as the cross section averaged over the collision velocity distribution,

$$K = \int_0^\infty v\sigma(v)f(v)dv. \quad (\text{C1.4.58})$$

The velocity distribution $f(v)$ depends on the conditions of the experiment. In cell and trap experiments it is usually a Maxwell–Boltzmann distribution at some well defined temperature, but $f(v)$ in atomic beam experiments, arising from optical excitation velocity selection, deviates radically from the normal thermal distribution [47]. The actual signal count rate, $\frac{d(X_2^+)}{dt}$, relates to the rate coefficient through

$$\frac{1}{V\alpha} \frac{d(X_2^+)}{dt} = K[X]^2 \quad (\text{C1.4.59})$$

where V is the interaction volume, α the ion detection efficiency and $[X]$ the atom density. If rate constant or cross section measurements are carried out in crossed or single atomic beams [44, 47, 48] special care is necessary to determine the interaction volume and atomic density.

PAI was the first measured collisional process observed between cooled and trapped atoms [26]. The experiment was performed with atomic sodium confined in a hybrid laser trap, utilizing both the spontaneous radiation pressure and the dipole force. The trap had two counterpropagating, circularly polarized Gaussian laser beams brought to separate foci such that longitudinal confinement along the beam axis was achieved by the spontaneous force and transversal confinement by the dipole force. The trap was embedded in a large (~1 cm diameter) conventional optical molasses loaded from a slowed atomic beam. The two focused laser beams comprising the dipole trap were alternately chopped with a 3 μs ‘trap cycle’, to avoid standing-wave heating. This trap cycle for each beam was interspersed with a 3 μs ‘molasses cycle’ to keep the atoms cold. The trap beams were detuned about 700 MHz to the red of the $3s^2 S_{1/2}(F=2) \rightarrow 3p^2 P_{3/2}(F=3)$ transition while the molasses was detuned only about one natural line width (~10 MHz). The atoms captured from the molasses (~ 10^7 cm^{-3}) were compressed to a much higher excited atom density (~ $5 \times 10^9 \text{ cm}^{-3}$) in the trap. The temperature was measured to be about 750 μK . Ions formed in the trap were accelerated and

focused toward a charged-particle detector. To assure the identity of the counted ions, Gould *et al* [26] carried out a time-of-flight measurement; the results of which, shown in [figure C1.4.12](#) clearly establish the Na_2^+ ion product. The linearity of the ion rate with the square of the atomic density in the trap supported the view that the detected Na_2^+ ions were produced in a binary collision. After careful measurement of ion rate, trap volume and excited atom density, the value for the rate coefficient was determined to be $K = (1.1_{-0.5}^{+1.3}) \times 10^{-11} \text{ cm}^3 \text{ s}^{-1}$. Gould *et al* [26], following conventional wisdom, interpreted the ion production as originating from collisions between two *excited* atoms,

$$\frac{dN_1}{dt} = K \int n_e^2(\mathbf{r}) d^3\mathbf{r} = K \bar{n}_e N_e \quad (\text{C1.4.60})$$

where $\frac{dN_1}{dt}$ is the ion production rate, $n_e(\mathbf{r})$ the excited-state density, N_e the number of excited atoms in the trap ($= \int n_e(\mathbf{r}) d^3(\mathbf{r})$) and \bar{n}_e the ‘effective’ excited-state trap density. The value for K was then determined from these measured parameters. Assuming an average collision velocity of 130 cm s^{-1} , equivalent to a trap temperature of $750 \mu\text{K}$, the corresponding cross section was determined to be $\sigma = (8.6^{+10.0}_{-3.8}) \times 10^{-14} \text{ cm}^2$. In contrast the cross section at $\sim 575 \text{ K}$ had been previously determined to be $\sim 1.5 \times 10^{-16} \text{ cm}^2$ [49, 50 and 51]. Gould *et al* [26] rationalized the difference in cross section size by invoking the difference in de Broglie wavelengths, the number of participating partial waves and the temperature dependence of the ionization channel probability. The quantal expression for the cross section in terms of partial wave contributions l and inelastic scattering probability S_{12} is

$$\sigma_{12}(\epsilon) = \left(\frac{\pi}{k^2}\right) \sum_{l=0}^{\infty} (2l+1) |S_{12}(\epsilon, l)|^2 \cong \pi \left(\frac{\lambda_{\text{dB}}}{2\pi}\right)^2 (l_{\text{max}} + 1)^2 P_{12} \quad (\text{C1.4.61})$$

where λ_{dB} is the entrance channel de Broglie wavelength and P_{12} is the probability of the ionizing collision channel averaged over all contributing partial waves of which l_{max} is the greatest. The ratio of $(l_{\text{max}} + 1)^2$ between 575 K and $750 \mu\text{K}$ is about 400 and the de Broglie wavelength ratio factor varies inversely with temperature. Therefore, in order that the cross section ratio be consistent with low- and high-temperature experiments, $\frac{\sigma_{12}(575 \text{ K})}{\sigma_{12}(750 \mu\text{K})} \sim 1.7 \times 10^{-3}$, Gould *et al* [26] concluded that P_{12} must be about three times greater at 575 K than at $750 \mu\text{K}$.

-27-

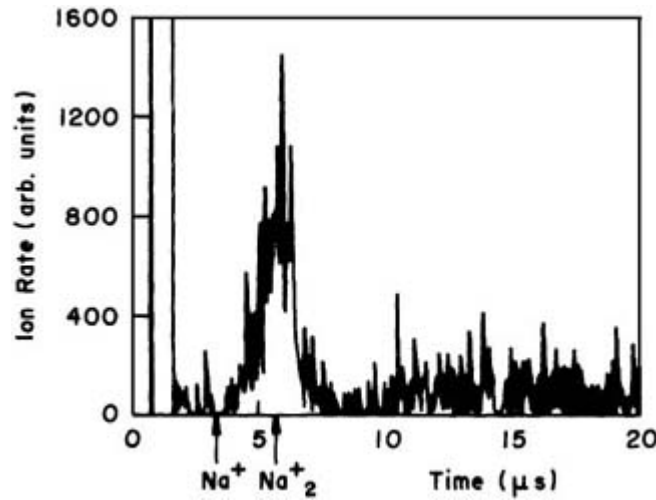


Figure C1.4.12. Time-of-flight spectrum clearly showing that the ions detected are Na_2^+ and not the atomic ion.

However, it soon became clear that the conventional picture of associative ionization, starting from the excited atomic states, could not be appropriate in the cold regime. Julienne [52] pointed out the essential problem with this picture. In the molasses cycle the optical field is only red detuned by one line width, and the atoms must therefore be excited at very long range, near $1800 a_0$. The collision travel time to the close internuclear separation where associative ionization takes place is long compared to the radiative lifetime, and most of the population decays to the ground state before reaching the autoionization zone. During the trap cycle, however, the excitation takes place at much closer internuclear distances due to a 70 line-width red detuning and high-intensity field dressing. Therefore, one might expect excitation survival to be better on the trap cycle than on the molasses cycle, and the NIST group set up an experiment to test the predicted cycle dependence of the ion rate.

Lett *et al* [53] performed a new experiment using the same hybrid trap. This time, however, the experiment measured ion rates and fluorescence separately as the hybrid trap oscillated between ‘trap’ and ‘molasses’ cycles. The results from this experiment are shown in figure C1.4.13. While keeping the total number and density of atoms (excited atoms plus ground-state atoms) essentially the same over the two cycles and while the excited state fraction changed only by about a factor of two, the ion rate increased in the trapping cycle by factors ranging from 20 to 200 with most observations falling between 40 and 100. This verified the predicted effect qualitatively even if the magnitude was smaller than the estimated 10^4 factor of Julienne [52]. This modulation ratio is orders of magnitude more than would be expected if excited atoms were the origin of the associative ionization signal. Furthermore, by detuning the trapping lasers over 4 GHz to the red, Lett *et al* continued to measure ion production at rates comparable to those measured near the atomic resonance. At such large detunings, reduction in atomic excited-state population would have led to reductions in ion rate by over four orders of magnitude, had the excited atoms been the origin of the collisional ionization. Not only did far off-resonance trap cycle detuning maintain the ion production rate, but Lett *et al* [53] observed evidence of peak structure in the ion signal as the dipole trap cycle detuned to the red.

-28-

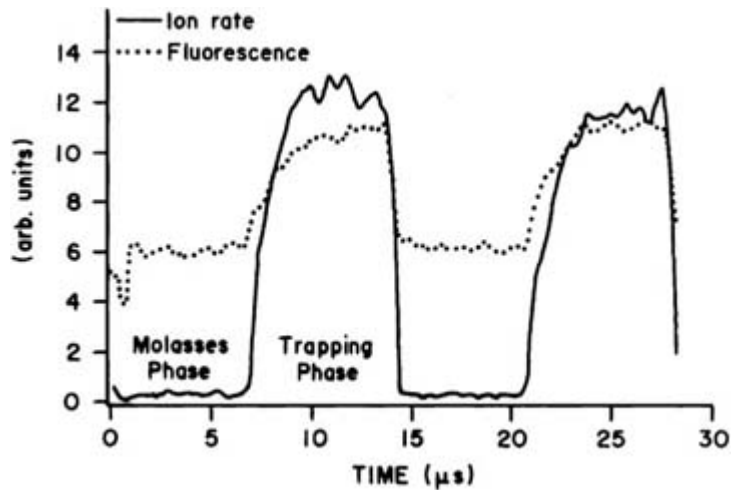


Figure C1.4.13. Trap modulation experiment showing much greater depth of ion intensity modulation (by more than one order of magnitude) than fluorescence or atom number modulation, demonstrating that excited atoms are not the origin of the associative ionizing collisions.

To interpret this experiment, Julienne and Heather [45] proposed a mechanism that has become the standard picture for cold and ultracold photoassociative ionization. Figure C1.4.14 details the model.

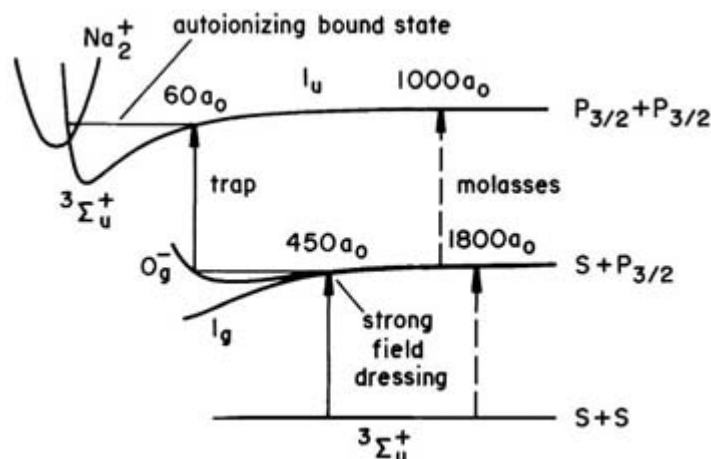


Figure C1.4.14. Photoassociative ionization (PAI) in Na collisions.

Two colliding atoms approach on the molecular ground-state potential. During the molasses cycle with the optical fields detuned only about one line width to the red of atomic resonance, the initial excitation occurs at very long range, around a Condon point at $1800 a_0$. A second Condon point at $1000 a_0$ takes the population to a 1_u doubly excited potential that, at shorter internuclear distance, joins adiabatically to a $3 \Sigma_u^+$ potential, thought to be the principal short-range entrance channel to associative ionization [54, 55]. More recent calculations suggest other entrance channels are important as well [56]. The long-range optical coupling to excited potentials in regions with little curvature implies that spontaneous radiative relaxation will depopulate these channels before the approaching partners reach the region of small internuclear separation where associative ionization takes place. The overall probability for collisional ionization during the molasses cycle remains therefore quite low. In contrast, during the trap cycle the optical fields are detuned 60 line widths to the red of resonance, the first Condon point occurs at $450 a_0$; and, if the trap cycle field couples to the 0_g^- long-range molecular state [57], the second Condon point occurs at $60 a_0$. Survival against radiative relaxation improves greatly because the optical coupling occurs at much shorter range where excited-state potential curvature accelerates the two atoms together. Julienne and Heather [45] calculate about a three-orders-of-magnitude enhancement in the rate constant for collisional ionization during the trap cycle. The dashed and solid arrows in figure C1.4.14 indicate the molasses-cycle and trap-cycle pathways, respectively. The strong collisional ionization rate constant enhancement in the trap cycle calculated by Julienne and Heather [45] is roughly consistent with the measurements of Lett *et al* [53], although the calculated modulation ratio is somewhat greater than what was actually observed. Furthermore, Julienne and Heather calculate structure in the trap detuning spectrum. As the optical fields in the dipole trap tune to the red, a rather congested series of ion peaks appear, which Julienne and Heather ascribed to free-bound association resonances corresponding to vibration-rotation bound levels in the 1_g or the 0_g^- molecular excited states. The density of peaks corresponded roughly to what Lett *et al* [53] had observed; these two tentative findings together were the first evidence of a new photoassociation spectroscopy. In a subsequent full paper expanding on their earlier report, Heather and Julienne [58] introduced the term ‘photoassociative ionization’ to distinguish the two-step optical excitation of the quasimolecule from the conventional associative ionization collision between excited atomic states. In a very recent paper, Pillet *et al* [59] have developed a perturbative quantum approach to the theory of photoassociation, which can be applied to the whole family of alkali homonuclear molecules. This study presents a useful table of photoassociation rates which reveals an important trend toward lower rates of molecule formation as the alkali mass increases and provides a helpful guide to experiments designed to detect ultracold molecule production.

REFERENCES

- [1] Phillips W D, Prodan J V and Metcalf H J 1985 Laser cooling and electromagnetic trapping of neutral atoms *J.Opt.Soc.Am.* B **2** 1751–67
- [2] Dalibard J and Cohen-Tannoudji C 1985 Dressed-atom approach to atomic motion in laser light: the dipole force revisited *J.Opt.Soc.Am.* B **2** 1707–20
- [3] Metcalf H and van der Straten P 1994 Cooling and trapping of neutral atoms *Phys. Rep.* **244** 203–86
- [4] Adams C S and Riis E 1997 Laser cooling and trapping of neutral atoms *Prog. Quant. Electr.* **21** 1–79
- [5] Adams C S, Carnal O and Mlynek J 1994 Atom interferometry *Adv. At. Mol. Opt. Phys.* **34** 1–33
- [6] Adams C S, Sigel M and Mlynek J 1994 Atom optics *Phys. Rep.* **240** 143–210
- [7] Morinaga M, Yasuda M, Kishimoto T and Shimizu F 1996 Holographic manipulation of a cold atomic beam *Phys.Rev.Lett.* **77** 802–5

- [8] Jessen P S and Deutsch I H 1996 Optical lattices *Adv. At. Mol. Opt. Phys.* **37** 95–138
- [9] Suominen K-A 1996 Theories for cold atomic collisions in light fields *J.Phys.B:At.Mol.Opt.Phys.* **29** 5981–6007
- [10] Frisch C R 1933 Experimenteller Nachweis des Einsteinschen Strahlungsruckstosses *Z.Phys.* **86** 42–8
- [11] Ashkin A 1970 Acceleration and trapping of particles by radiation pressure *Phys.Rev.Lett.* **24** 156–9
- [12] Stenholm S 1986 The semiclassical theory of laser cooling *Rev.Mod.Phys.* **58** 699–739
- [13] Cohen-Tannoudji C, Dupont-Roc J and Grynberg G 1992 *Atom–Photon Interactions: Basic Processes and Applications* (New York: Wiley)
- [14] Cook R J 1979 Atomic motion in resonant radiation: an application of Earnshaw’s theorem *Phys.Rev. A* **20** 224–8
- [15] Cook R J 1980 Theory of resonant-radiation pressure *Phys.Rev. A* **22** 1078–98
- [16] Lett P D, Watts R N, Westbrook C I, Phillips W D, Gould P L and Metcalf H J 1988 Observation of atoms, laser-cooled below the Doppler limit *Phys.Rev.Lett.* **61** 169–72
- [17] Dalibard J and Cohen-Tannoudji C 1989 Laser cooling below the Doppler limit by polarization gradients: simple theoretical models *J.Opt.Soc.Am. B* **6** 2023–45
- [18] Pritchard D E, Raab E L, Bagnato V, Wieman C E and Watts R N 1986 Light traps using spontaneous forces *Phys.Rev.Lett.* **57** 310–13
- [19] Ashkin A and Gordon J P 1983 Stability of radiation-pressure particle traps: an optical Earnshaw theorem *Opt. Lett.* **8** 511–13
- [20] Raab E, Prentiss M, Cable A, Chu S and Pritchard D E 1987 Trapping of neutral sodium atoms with radiation pressure *Phys.Rev.Lett.* **59** 2631–4
- [21] Walker T, Sesko D and Wieman C 1990 Collective behavior of optically trapped neutral atoms *Phys.Rev.Lett.* **64** 408–11

- [22] Cable A, Prentiss M and Bigelow N P 1990 Observation of sodium atoms in a magnetic molasses trap loaded by a continuous uncooled source *Opt. Lett.* **15** 507–9
- [23] Monroe C, Swann W, Robinson H and Wieman C 1990 Very cold trapped atoms in a vapor cell *Phys.Rev.Lett.* **65** 1571–4
- [24] Sesko D W, Walker T G and Wieman C 1991 Behavior of neutral atoms in a spontaneous force trap *J.Opt.Soc.Am. B* **8** 946–58
- [25] Ketterle W, Davis K B, Joffe M A, Martin A and Pritchard D 1993 High densities of cold atoms in a dark spontaneous-force optical trap *Phys.Rev.Lett.* **70** 2253–6
- [26] Gould P L, Lett P D, Julienne P S, Phillips W D, Thorsheim H R and Weiner J 1988 Observation of associative ionization of ultracold laser-trapped sodium atoms *Phys.Rev.Lett.* **60** 788–91
- [27] Miller J D, Cline R A and Heinzen D J 1993 Far-off-resonance optical trapping of atoms *Phys.Rev. A* **47** R4567–70
- [28] Miller J D, Cline R A and Heinzen D J 1993 Photoassociation spectrum of ultracold Rb atoms *Phys.Rev.Lett.* **71** 2204–7
- [29] Lovelace R V E, Mehanian C, Tommila T J and Lee D M 1985 Magnetic confinement of a neutral gas *Nature* **318** 30–6
- [30] Migdall A L, Prodan J V, Phillips W D, Bergman T H and Metcalf H J 1985 First observation of magnetically trapped neutral atoms *Phys.Rev.Lett.* **54** 2596–9
- [31] Gott Y V, Ioffe M S and Telkovsky V G 1962 *Nuclear Fusion Suppl.* part 3 (Vienna: International Atomic Energy Agency) p 1045
- [32] Pritchard D E 1983 Cooling neutral atoms in a magnetic trap for precision spectroscopy *Phys.Rev.Lett.* **51** 1336–9
- [33] Bagnato V S, Lafyatis G P, Martin A C, Raab E L, Ahmad-Bitar R and Pritchard D E 1987 Continuous stopping and trapping of neutral atoms *Phys.Rev.Lett.* **58** 2194–7

- [34] Hess H F, Kochanski G P, Doyle J M, Greytak T J, and Kleppner D 1986 Spin-polarized hydrogen maser *Phys.Rev. A* **34** 1602–4
- [35] Mewes M-O, Andrews M R, van Druten N J, Kurn D M, Durfee D S and Ketterle W 1996 Bose–Einstein condensation in a tightly confining DC magnetic trap *Phys.Rev.Lett.* **77** 416–19
- [36] Anderson M H, Ensher J R, Matthews M R, Wieman C E and Cornell E A 1995 Observation of Bose–Einstein condensation in a dilute atomic vapor *Science* **269** 198–201
- [37] Davis K B, Mewes M-O, Andrews M R, van Druten N J, Durfee D S, Kurn D M and Ketterle W 1995 Bose–Einstein condensation in a gas of sodium atoms *Phys.Rev.Lett.* **75** 3969–73
- [38] Vigué J 1986 Possibility of applying laser-cooling techniques to the observation of collective quantum effects *Phys.Rev. A* **34** 4476–9
- [39] Scheingraber H and Vidal C R 1977 Discrete and continuous Franck–Condon factors of the $\text{Mg}_2\text{A}^1\text{A}^1\Sigma_u^+ - \text{X}^1\text{X}^1\Sigma_g^+$ system and their J dependence *J.Chem.Phys.* **66** 3694–704
- [40] Inoue G, Ku J K and Setser D W 1982 Photoassociative laser induced fluorescence of XeCl *J.Chem.Phys.* **76** 733–4
- [41] Jones R B, Schloss J H and Eden J G 1993 Excitation spectra for the photoassociation of Kr–F and Xe–I collision pairs in the ultraviolet (209–258 nm) *J.Chem.Phys.* **98** 4317–34

-32-

- [42] Tellinghuisen J 1985 *Photodissociation and Photoionization (Advances in Chemical Physics LX)* ed K P Lawley (New York: Wiley) pp 299–369
- [43] Thorsheim H R, Weiner J and Julienne P S 1987 Laser-induced photoassociation of ultracold sodium atoms *Phys.Rev.Lett.* **58** 2420–3
- [44] Weiner J, Masnou-Seeuws F and Guisti-Suzor A, Associative ionization: experiments, potentials, and dynamics *Advances in Atomic, Molecular and Optical Physics* vol 26, ed D Bates and B Bederson (Boston: Academic) pp 209–96
- [45] Julienne P S and Heather R 1991 Laser modification of ultracold atomic collisions: theory *Phys.Rev.Lett.* **67** 2135–8
- [46] Leonhardt D and Weiner J 1995 Direct two-color photoassociative ionization in a rubidium magneto-optic trap *Phys.Rev. A* **52** R4332–R4335
- [47] Tsao C-C, Napolitano R, Wang Y and Weiner J 1995 Ultracold photoassociative ionization collisions in an atomic beam: optical field intensity and polarization dependence of the rate constant *Phys.Rev. A* **51** R18–21
- [48] Thorsheim H R, Wang Y and Weiner J 1990 Cold collisions in an atomic beam *Phys.Rev. A* **41** 2873–6
- [49] Bonanno R, Boulmer J and Weiner J 1983 Determination of the absolute rate constant for associative ionization in crossed-beam collision between $\text{Na } 3^2\text{P}_{3/2}$ atoms *Phys.Rev. A* **28** 604–8
- [50] Wang M-X, Keller J, Boulmer J and Weiner J 1986 Strong velocity dependence of the atomic alignment effect in $\text{Na}(3p) + \text{Na}(3p)$ associative ionization *Phys.Rev. A* **34** 4497–500
- [51] Wang M-X, Keller J, Boulmer J and Weiner J 1987 Spin-selected velocity dependence of the associative ionization cross section in $\text{Na}(3p) + \text{Na}(3p)$ collisions over the collision energy range from 2.4 to 290 meV *Phys.Rev. A* **35** 934–7
- [52] Julienne P S 1988 Laser modification of ultracold atomic collision in optical traps *Phys.Rev.Lett.* **61** 698–701
- [53] Lett P D, Jessen P S, Phillips W D, Rolston S L, Westbrook C I and Gould P L 1991 Laser modification of ultracold collisions: experiment *Phys.Rev.Lett.* **67** 2139–42
- [54] Dulieu O, Guisti-Suzor A and Masnou-Seeuws F 1991 Theoretical treatment of the associative ionization reaction between two laser-excited sodium atoms. Direct and indirect processes *J.Phys.B:At.Mol.Opt.Phys.* **24** 4391–408
- [55] Henriot A, Masnou-Seeuws F and Dulieu O 1991 Diabatic representation for the excited states of the Na_2 molecule: application to the associative ionization reaction between two excited sodium atoms *Z.Phys. D* **18** 287–98
- [56] Dulieu O, Magnier S and Masnou-Seeuws F 1994 Doubly-excited states for the Na_2 molecule: application to the dynamics of the associative ionization reaction *Z.Phys. D* **32** 229–40

- [57] Stwalley W C, Uang Y-H and Pichler G 1978 Pure long-range molecules *Phys.Rev.Lett.* **41** 1164–6
- [58] Heather R W and Julienne P S 1993 Theory of laser-induced associative ionization of ultracold Na *Phys.Rev. A* **47** 1887
- [59] Pillet P, Crubellier A, Bleton A, Dulieu O, Nosbaum P, Mourachko I and Masnou-Seeuws F 1997 Photoassociation in a gas of cold alkali atoms: I. Perturbative quantum approach *J.Phys.B:At.Mol.Opt.Phys.* **30** 2801–20
-

-33-

FURTHER READING

A good introduction to the physics of laser cooling and trapping can be found in two special issues of the *Journal of the Optical Society of America B*. These are:

1985 The mechanical effects of light *J.Opt.Soc.Am.* **B 2** 11

1989 Laser cooling and trapping of atoms *J.Opt.Soc.Am.* **B 6** 11

Two recent reviews recount subsequent research in the physics of neutral-atom cooling and trapping [3, 4].

For an introduction to current research in alkali-atom BEC see the special issue on BEC in the *Journal of Research of the National Institute of Standards and Technology*:

1996 *Journal of Research of the National Institute of Standards and Technology* **101** 4

Another useful review can be found by:

Meschede D *et al* 1998 *Naturwissenschaften* **85** 203–18

-1-

C1.5 Single molecule spectroscopy

Anne Myers Kelley

C1.5.1 INTRODUCTION

Until the late 1980s, virtually all molecular spectroscopic measurements involved observing a signal having contributions from a large number of different molecules, ‘large’ meaning much greater than one. While spectroscopists have long had the ability to detect and count individual photons or ions each originating from a single molecule, the statistical averaging required to make a meaningful measurement generally required observing many such events from many different molecules. Only recently have experimental techniques been developed that allow interrogation of fundamentally quantum mechanical entities on a one-by-one basis. These developments are driving revolutionary changes in the way molecular scientists make and interpret physical measurements. For example, simple organic molecules that are chemically identical, distinguished only by slightly different local environments within a solid, have been shown to have distinctly different electronic and vibrational spectra, linewidths, electric field and pressure-induced spectral shifts, and fluorescence lifetimes and quantum yields. The ability to correlate these various spectroscopic properties on a molecule-by-molecule basis is providing powerful insight into the details of intermolecular interactions. Single-molecule techniques have also shown that apparently pure, homogeneous enzyme preparations contain molecules having a wide range of catalytic activities and that an

individual enzyme's catalytic activity retains a 'memory' of its past history. This new information is stimulating a re-evaluation of established models for the chemical kinetics of biological systems. Single-molecule experiments involve sequential measurements of a given observable on the same molecule at different times and, if a time average is equivalent to an ensemble average (the ergodic hypothesis), no additional information is gained by probing individual members. The value of single-molecule measurements lies precisely in the fact that in many systems of interest, different members of the ensemble remain distinct on time scales much longer than that required to perform an experiment.

A wide variety of measurements can now be made on single molecules, including electrical (e.g. scanning tunnelling microscopy), magnetic (e.g. spin resonance), force (e.g. atomic force microscopy), optical (e.g. near-field and far-field fluorescence microscopies) and hybrid techniques. This contribution addresses only those techniques that are at least partially optical. Single-particle electrical and force measurements are discussed in the sections on scanning probe microscopies (B1.19) and surface forces apparatus (B1.20).

C1.5.2 HISTORY

The approach to, and finally the achievement of, detection and spectroscopy of single molecules proceeded almost independently from three separate directions.

C1.5.2.1 SPECTRAL SELECTION IN CRYOGENIC SOLIDS

The development of tunable, narrow-bandwidth dye laser sources in the early 1970s gave spectroscopists a new tool for selectively exciting small subsets of molecules within inhomogeneously broadened ensembles in the solid state. The technique of fluorescence line-narrowing [1, 2 and 3] takes advantage of the fact that relatively rigid chromophoric

-2-

molecules in dilute mixed solids at low temperatures often have very narrow electronic origin transitions even when the bulk absorption spectrum is severely broadened by an inhomogeneous distribution of spectrally distinct environments. By tuning a narrow-band excitation source into resonance with the low-energy side of the absorption band, only those molecules that absorb at that precise frequency can be excited. This subensemble, having a well defined electronic origin frequency, will then produce spectrally sharp emissions. As the sample is made increasingly dilute, and/or the excitation is tuned progressively farther toward the red edge of the absorption spectrum, the number of molecules on resonance with the laser decreases, eventually becoming either zero or one.

The first clearly demonstrated optical detection of single *chromophores* was published by Moerner's group at IBM Almaden 1989 on the mixed crystal system pentacene in *p*-terphenyl at 1.6 K [4, 5]. They utilized a double-modulation direct absorption technique employing frequency modulation of the laser source coupled with electric field or ultrasonic strain modulation of the absorption line. Direct absorption is not known for its high sensitivity, and these preliminary experiments achieved only modest signal-to-noise ratios. Shortly thereafter, Orrit's group at Bordeaux demonstrated single-molecule detection in the same system with the fundamentally much more sensitive technique of fluorescence excitation—fluorescence line-narrowing carried to the extreme of a single resonant chromophore at a time (figure C1.5.1) [6]. Fluorescence excitation has since been the technique of choice for nearly all single-molecule optical experiments, although some refinements in direct absorption detection have recently been demonstrated [7].

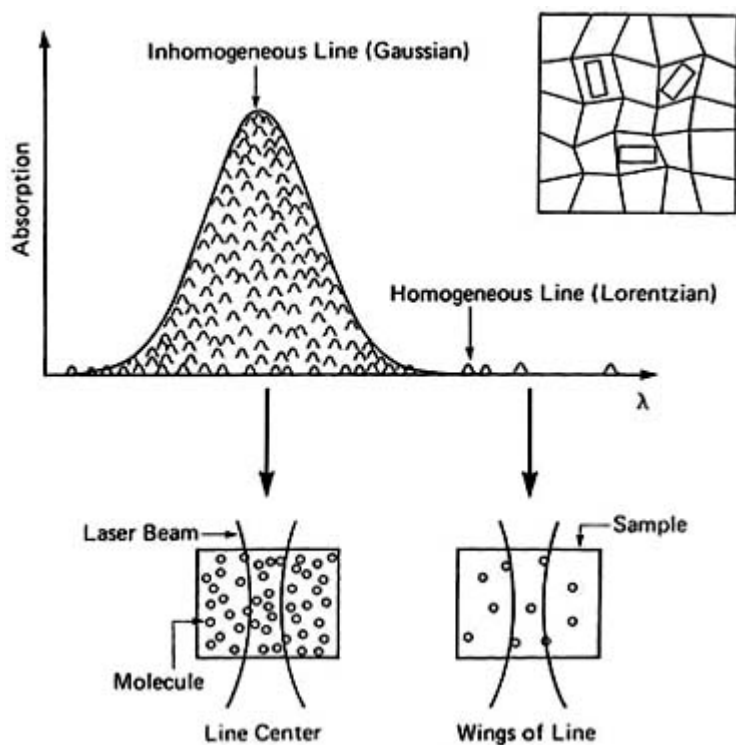


Figure C1.5.1. (A) Schematic diagram showing the principle of single-molecule spectral selection in solids at low temperatures. The inhomogeneously broadened electronic origin is composed of a superposition of the Lorentzian profiles of individual molecules, with a Gaussian distribution of centre frequencies caused by random strains and defects in the surrounding environment. (B) The number of dopant molecules in the probed volume on resonance with a narrow-band laser can be controlled by tuning the laser wavelength to the red side of the inhomogeneous band. Reprinted with permission from Moerner [178]. Copyright 1994 American Association for the Advancement of Science.

-3-

C1.5.2.2 FLUORESCENCE STATISTICS IN LIQUID SOLUTIONS

Fluorescence, because of its essentially background-free nature, has long been appreciated by both analytical and physical chemists for its high sensitivity. Photon counting techniques allow detection of single photons each emitted by a single molecule, although under normal conditions the total signal is composed of photons arising from many different molecules. Hirschfeld was apparently the first to demonstrate that individual molecules (in his case, large antibodies each tagged with 80–100 fluorophores) in a highly dilute solution could be detected by the burst of fluorescence emitted by each molecule as it diffused through the observation volume of an optical microscope [8]. The Keller [9] and Mathies [10] groups subsequently combined this idea with a more detailed analysis of the photon burst statistics to detect single molecules of the highly fluorescent multichromophoric protein phycoerythrin. The first demonstration of genuine single-*chromophore* detection in liquid solution was published by Shera and co-workers in 1990 on the laser dye rhodamine 6G in water [11], using pulsed excitation and time-gated detection to reduce background counts. Somewhat later, Nie *et al* [12] demonstrated that with some modifications to the detection system, a commercial laser confocal microscope could also be made sensitive enough to detect single chromophores diffusing in and out of the detection volume.

C1.5.2.3 SPATIAL SELECTION IN SOLIDS AND ON SURFACES

A third approach to single-molecule optical detection began with the development of near-field scanning optical microscopy (NSOM) in the 1980s (see section 1.19). NSOM allows optical measurements on surfaces to be made with a resolution approaching $\lambda/40$ in the best cases [13], which for visible light is comparable to the size of single large molecules such as proteins and polymers. The first observation of single molecules by NSOM was reported in

1993 by Betzig and Chichester [14]. Shortly thereafter, several groups demonstrated that the techniques of ordinary far-field fluorescence microscopy, if coupled with highly sensitive, low-noise detectors, can also detect single molecules as long as they are spaced far enough apart that the limiting resolution of about $\lambda/2$ is adequate. Far-field fluorescence microscopy is technically simpler than NSOM and has the advantage of not being restricted to surfaces, and has become the technique of choice for spatial selection of single molecules as long as the sample can be made sufficiently dilute that the additional resolution of the near-field technique is not needed.

C1.5.3 PRINCIPLES AND TECHNIQUES OF SINGLE-MOLECULE OPTICAL

C1.5.3.1 FLUORESCENCE

The vast majority of single-molecule optical experiments employ one-photon excited spontaneous fluorescence as the spectroscopic observable because of its relative simplicity and inherently high sensitivity. Many molecules fluoresce with quantum yields near unity, and spontaneous fluorescence lifetimes for chromophores with large oscillator strengths are a few nanoseconds, implying that with a sufficiently intense excitation source a single molecule should be able to absorb and emit of the order of 10^8 photons per second. Additionally, in most molecules much of the emitted light is sufficiently red-shifted from the excitation frequency to allow detection against a nominally near-zero background.

A number of experimental and physical realities cloud this rosy picture. Inevitably many emitted photons are lost due to the finite solid angle over which the fluorescence is collected, losses at the various filters, lenses, windows, and other

-4-

optical elements between the sample and the detector, and the quantum efficiency of the photodetector. Most experimental configurations actually detect emitted photons with an overall efficiency of only a few per cent. All molecules have nonzero, if small, quantum yields for forming long-lived metastable states (often triplet states), reducing the number of absorption–emission cycles that can be accomplished per second. The intensity of the excitation source has to be kept low enough that it does not induce undesired nonlinear optical effects (section B1.1 and section B1.5). Finally, all known molecules undergo photochemical degradation with some nonzero yield, and this photobleaching is generally the limiting factor in determining the total number of photons that can be collected from a single molecule. Nevertheless, it is routinely possible with strongly fluorescent chromophores to detect 10^3 – 10^4 photons per second. The challenging part of performing single-molecule fluorescence detection is not the absolute size of the signal, which is huge compared with that typical of many conventional ensemble-averaged spectroscopies such as Raman, but rather ascertaining that the signal arises from only one molecule, reducing the background count level, and obtaining a statistically significant amount of data from the molecule under observation before it photobleaches.

(A) SPECTRAL SELECTION

The spectral selection approach [15, 16 and 17] is based on the fact that purely lifetime-limited line widths for electronic transitions of molecules rarely exceed 10–20 MHz, whereas the apparent width of the electronic origin in condensed-phase molecular spectra is typically orders of magnitude greater even in single crystals and certainly in polymers and glasses. At temperatures below 4 K, where little thermal population of phonons is possible, this additional width is ascribed to slightly different local environments for different chromophores, each giving rise to a slightly different spectral shift. When a spectrally narrow laser (bandwidth on the order of the intrinsic lifetime-limited width) is tuned through the ensemble-broadened electronic origin, only those chromophores on resonance with that particular laser frequency can absorb, and as the laser is tuned into the wings of the spectral line, the number of molecules on resonance approaches either zero or one (figure C1.5.1) and figure C1.5.2. In practice this works best when tuning to the red side of the electronic origin; on the blue side, the zero-phonon lines (pure

electronic origin transitions) of blue-shifted molecules are degenerate with the much broader phonon sidebands (electronic excitation of the chromophore coupled with excitation of low-frequency matrix or intermolecular vibrations) of redder-shifted molecules. Typically the emission is collected with mirror or lens systems placed inside the cryostat to collect light over the largest possible solid angle, and observed with a high quantum efficiency detector, either a photomultiplier tube or an avalanche photodiode, through one or more long-pass or bandpass filters to block scattered or transmitted laser light. Alternatively, the emitted light may be dispersed with a spectrograph and detected with a high-efficiency array photodetector, generally a CCD.

-5-

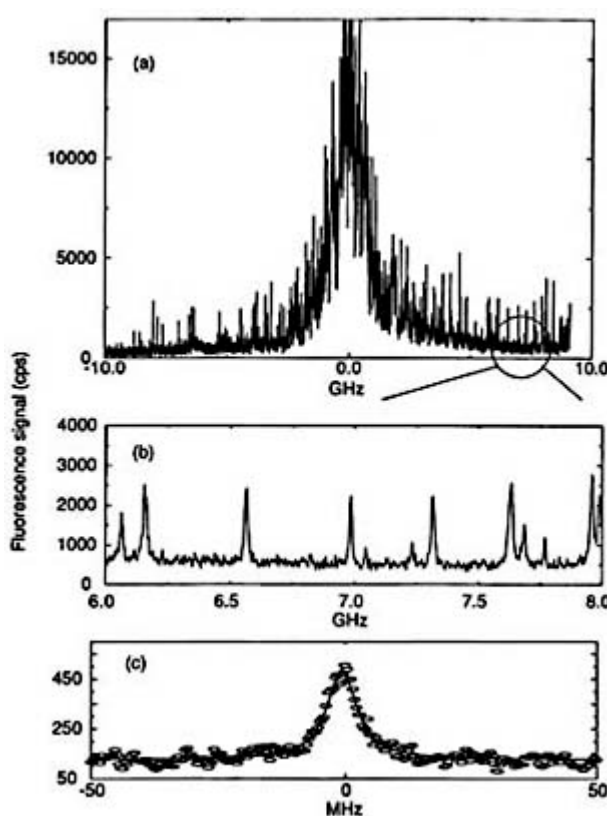


Figure C1.5.2. Fluorescence excitation spectra (cps = counts per second) of pentacene in *p*-terphenyl at 1.5 K. (A) Broad scan of the inhomogeneously broadened electronic origin. The spikes are repeatable features each due to a different single molecule. The laser detuning is relative to the line centre at 592.321 nm. (B) Expansion of a 2 GHz region of this scan showing several single molecules. (C) Low-power scan of a single molecule at 592.407 nm showing the lifetime-limited width of 7.8 MHz and a Lorentzian fit. Reprinted with permission from Moerner [198]. Copyright 1994 American Association for the Advancement of Science.

The criterion that single molecules are being observed is often taken to be the appearance in a fluorescence excitation frequency scan of well separated peaks that have about the same intensity and width, separated by regions of flat background. However, one cannot be certain that a given peak does not arise from two molecules with accidentally degenerate resonant frequencies. Stronger evidence is provided by observing that when spontaneous or photoinduced spectral jumps or photobleaching occur, the fluorescence excitation feature jumps to a new frequency or disappears entirely from the scanned region in a quantized, all-or-nothing manner. Probably the best evidence for single-molecule observation comes from the statistics of photon emission on short time scales. Since a single molecule must experience a nonzero time interval between successive photon emissions (after emitting a photon, it must absorb one prior to its next emission), the probability of emitting two photons with zero time delay goes to zero for a single molecule, the phenomenon known as photon antibunching.

The requirement of a very sharp and strong electronic origin absorption line limits the technique to strongly absorbing and fluorescing, relatively rigid chromophores and matrices having little Franck–Condon activity in low-frequency

vibrations. The requirement that the electronic origin be spectrally quite stable limits the technique to very low temperatures and to chromophore–matrix combinations in which spectral diffusion and photophysical hole-burning processes are slow. Thus, while the intrinsically high spectral resolution of this technique allows detailed spectroscopic and dynamical studies on individual molecules, the number of material systems to which it can be applied seems to be quite limited.

(B) SPATIAL SELECTION WITH NEAR-FIELD OPTICS

The development of near-field scanning optical microscopy (NSOM) as a viable experimental technique opened up the possibility of performing optical measurements with spatial resolution on molecular scales, just as scanning tunnelling microscopy (STM) allows imaging through electrical measurements down to the atomic level. By forcing the excitation light to pass through a metallized tip with an aperture much smaller than the wavelength of the light, and placing the sample in the near field of the tip (much less than a wavelength away), a variety of optical microscopies can be performed at a resolution much better than the classical far-field limit of $\approx\lambda/2$ [13, 14, 15, 16, 17 and 18]. Apertureless variants of NSOM have also been described [19 and 20]. The NSOM technique in general is described in more detail in [section B1.20](#).

NSOM in fluorescence mode can easily detect single molecules that are spatially separated by tens of nanometres or more, as long as they are sufficiently photostable to withstand the necessary number of excitation–emission cycles [14, 15, 16, 17, 18, 19, 20 and 21]. The usual criterion for single-molecule observation is the appearance of single isolated spots in the NSOM fluorescence image ([figure C1.5.3](#)), although two or more molecules that are accidentally in very close proximity are not always resolvable. The quantized nature of photobleaching events is another good criterion for single-molecule observation in spatially as well as spectrally selected techniques. NSOM is difficult to perform at low temperatures, but cryogenic near-field microscopes have been described [22] and demonstrated at the single-molecule level [23 and 24]. A significant limitation is that the chromophore of interest must be at or very near a surface to allow the tip to be brought into close proximity. The tips are notoriously difficult to fabricate in a reproducible manner, particularly when a very small aperture is desired, and can become quite hot during operation due to the laser power dissipated in the metal coating. The proximity of the metal tip to the chromophore can induce artifacts that have been discussed in some detail [25, 26 and 27]. In applications where sub-diffraction-limited optical resolution is needed, for example in studying biological or engineered supermolecular structures, NSOM is the only viable technique. If, however, the goal is merely to study single molecules and the samples can be prepared such that the molecules are spaced arbitrarily far apart, conventional far-field fluorescence microscopies are technically more straightforward and less subject to artifacts.

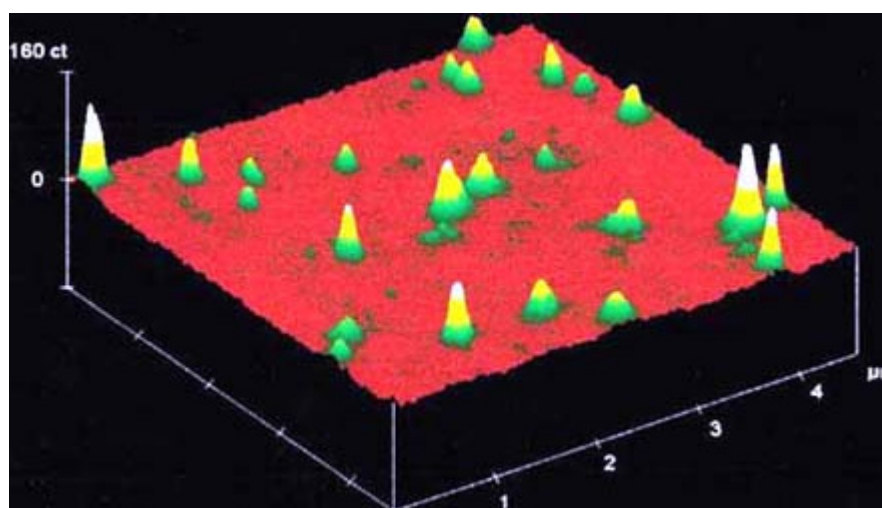


Figure C1.5.3. Near-field fluorescence image (4.5 μm square) of single oxazine 720 molecules dispersed on the surface of a PMMA film. Each peak (fwhm 100 nm) is due to a single molecule. The different intensities are due to different molecular orientations and spectra. Reprinted with permission from Xie [122]. Copyright 1996 American Chemical Society.

An intriguing alternative to NSOM is to engineer a tip bearing a single fluorescent molecule that can be excited at one wavelength, emitting light at a Stokes-shifted wavelength to be used as a highly spatially localized probe light source [28 and 29].

(C) SPATIAL SELECTION WITH FAR-FIELD OPTICS

Confocal scanning laser fluorescence microscopy is a well established optical technique (see section B1.19) that combines the transverse resolution common to any optical microscopy with a high degree of depth resolution that comes from requiring the fluorescence emission to follow the same optical path as the exciting laser light and pass through a pinhole conjugate to a pinhole through which the exciting laser was focused. A three-dimensional fluorescence image is mapped out by raster scanning either the exciting laser beam or the sample in the transverse plane and stepping the focusing lens to sample distance along the depth axis. If the fluorescence collection and detection systems are efficient enough, this technique can be sufficiently sensitive to detect emission from single molecules [12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29 and 30]. The limiting transverse resolution is determined by the numerical aperture of the laser focusing lens but cannot exceed about $\lambda/2$, while the resolution along the depth axis is typically of the same order. Thus, for single molecules to be resolved directly they must be separated on average by roughly 0.2–0.5 μm or more, although sub-diffraction resolution may be achieved through subsequent numerical processing of the images [31] or by making clever use of optical interference effects [32, 33]. Figure C1.5.4 compares near-field and far-field images of the same set of single molecules at a surface. Once a strong feature due to a single molecule has been identified, the centre of the focused spot can be moved to that molecule and more detailed spectroscopic or kinetic measurements made. The low resolution makes it difficult to be certain that a fluorescent spot truly arises from a single molecule, and it is useful to have other evidence such as photon emission statistics and/or quantized jumping or photobleaching to verify single-molecule observation.

-8-

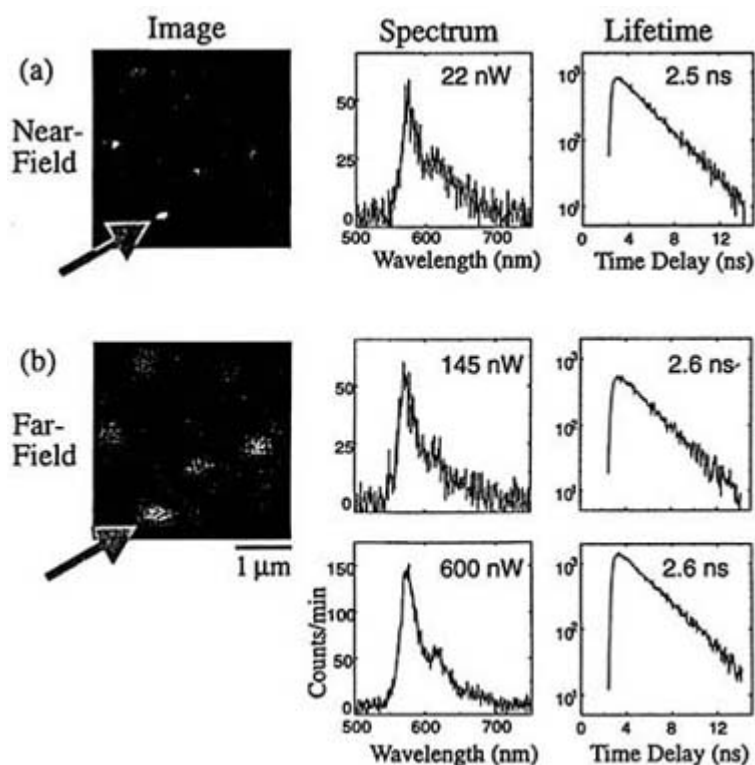


Figure C1.5.4. Comparison of near-field and far-field fluorescence images, spectra and lifetimes for the same set of isolated single molecules of a carbocyanine dye at a PMMA–air interface. Note the much higher resolution of the near-field image. The spectrum and lifetime of the molecule indicated with the arrow were recorded with near-field excitation and with far-field excitation at two different excitation powers. Reproduced with permission from Trautman and Macklin [125].

Forming an image by scanning the laser spot across the sample, or vice versa, minimizes the light dose received by each molecule and reduces photobleaching. The tradeoff is that it requires some time to gather an image. A fluorescent image can be obtained much more rapidly by irradiating a larger area in the transverse plane and imaging the emission from the entire area at once onto a two-dimensional photodetector. This approach is most useful for highly-photostable molecules at low temperatures [34, 35 and 36]. Photobleaching can be further reduced by employing an automatic positioning system with feedback to locate and centre the excitation on a single molecule as rapidly as possible [37] and also by excluding oxygen [38] and/or working at very low temperatures where most chromophores are more stable, although the latter adds considerable complexity to the experimental configuration [39, 40 and 41].

(D) STATISTICAL DETECTION METHODS

The statistics of the detected photon bursts from a dilute sample of chromophores can be used to count, and to some degree characterize, individual molecules passing through the illumination and detection volume. This can be achieved either by flowing the sample rapidly through a narrow fluid stream that intersects the focused excitation beam or by allowing individual chromophores to diffuse into and out of the beam. If the sample is sufficiently dilute that

chromophores pass through the beam effectively one at a time, repetitive excitation and emission of each molecule during its passage time generates a burst of emitted photons, superimposed on a random background count level due to stray room and laser light, Raman scattering from the solvent, etc [9, 10 and 11, 42, 43, 44, 45 and 46]. Figure C1.5.5 shows representative data from a configuration allowing free diffusion of chromophores into and out of the beam. A variety of statistical analyses of the photon bursts can be performed to improve the fidelity of detection and/or to discriminate between chromophores of different chemical species based on spectral and/or temporal features of the emission [44, 45, 47, 48, 49, 50, 51 and 52]. These statistical methods are employed mainly for counting molecules in analytical applications. The comparatively short observation time for each molecule limits the extent to which photophysical, spectroscopic or dynamic properties can be examined at the single-molecule level.

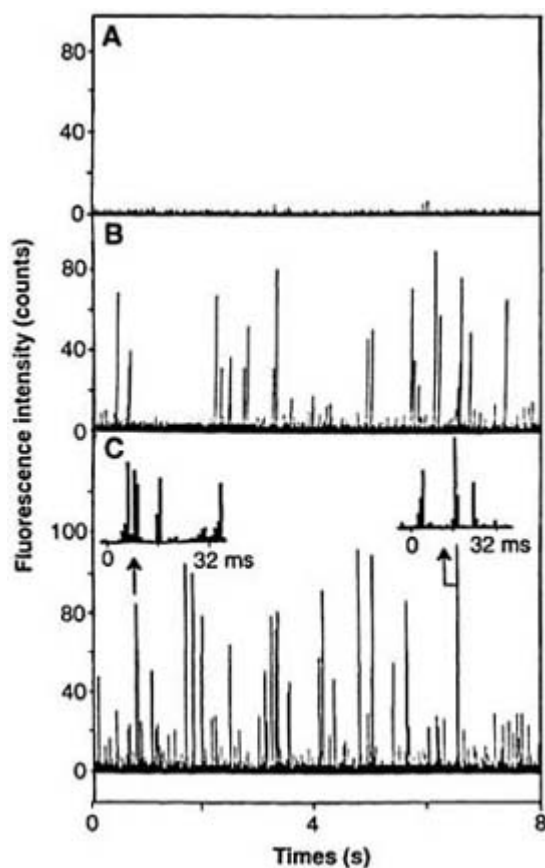


Figure C1.5.5. Time-dependent fluorescence signals observed from liquid solutions of rhodamine 6G by confocal fluorescence microscopy. Data were obtained with 514.5 nm excitation and detected through a 540–580 nm bandpass filter. The integration time is 1 ms per point. (A) Blank. (B) 2×10^{-11} M rhodamine 6G in water. Each peak represents detection of a single molecule diffusing into and then out of the detection volume. The inserts are close up views of the peaks indicated, showing the multiple detection of a single molecule. (C) 5×10^{-11} M rhodamine 6G in ethanol. Reprinted with permission from Nie *et al* [12]. Copyright 1994 American Association for the Advancement of Science.

C1.5.3.2 RAMAN SCATTERING

Raman scattering (section B1.3) is a notably weak process. In most experimental configurations no more than one in 10^{10} laser photons is scattered into a given Raman line. While resonance enhancement may improve this to one in 10^6 , Raman would still appear to be a highly unpromising technique for single-molecule detection. Nevertheless, at least four different groups have recently claimed single-molecule sensitivity using surface-enhanced Raman scattering (SERS), an enhancement mechanism whose physical basis is still subject to some controversy. Nie and Emory [53] examined rhodamine 6G dye molecules bound to particles of colloidal silver that were immobilized on a substrate such that they could be imaged via confocal microscopy. Using very low dye concentrations and exciting with 514.5 nm light, they observed Raman scattering from only a very few particles that they attributed to particularly stable and highly surface-enhanced ‘hot’ binding sites (figure C1.5.6) and figure C1.5.7. The apparent enhancement is of the order of 10^{14} over ordinary unenhanced Raman, whereas generally accepted values for SERS enhancements from ordinary bulk experiments are around 10^6 . More recent work by Brus *et al* essentially confirms and extends these observations 54. Kneipp *et al* examined crystal violet and a cyanine dye on colloidal silver with excitation in the near-infrared, away from both molecular electronic resonances and the principal particle plasmon resonance. They probed the particles in free solution and based their claim of single-chromophore detection on the excitation volume, dye-to-silver concentration ratio, and detection statistics [55, 56]. Finally, Käll *et al* examined the SERS spectra of haemoglobin attached to silver nanoparticles and concluded that single-molecule SERS is possible only for protein molecules situated between and bound to more than one silver particle

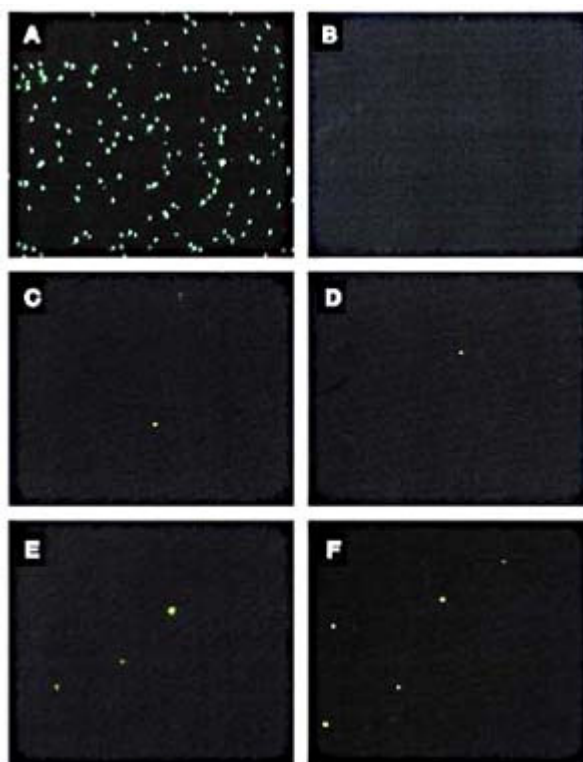


Figure C1.5.6. Single Ag nanoparticles imaged with evanescent-wave excitation. (A) Unfiltered photograph showing scattered laser light (514.5 nm) from Ag particles immobilized on a polylysine-coated surface. (B) Bandpass filtered (540–580 nm) photograph taken from a blank Ag colloid sample incubated with 1 mM NaCl and no dye. (C) and (D) Filtered photographs taken from an Ag colloid sample incubated with 2×10^{-11} M rhodamine 6G. Each image shows at least one Raman scattering particle. (E) and (F) Filtered photographs of Ag colloid incubated with higher concentrations of rhodamine 6G (2×10^{-10} M and 2×10^{-9} M, respectively). Each image shows several Raman scattering particles. Reprinted with permission from Nie and Emory [53]. Copyright 1997 American Association for the Advancement of Science.

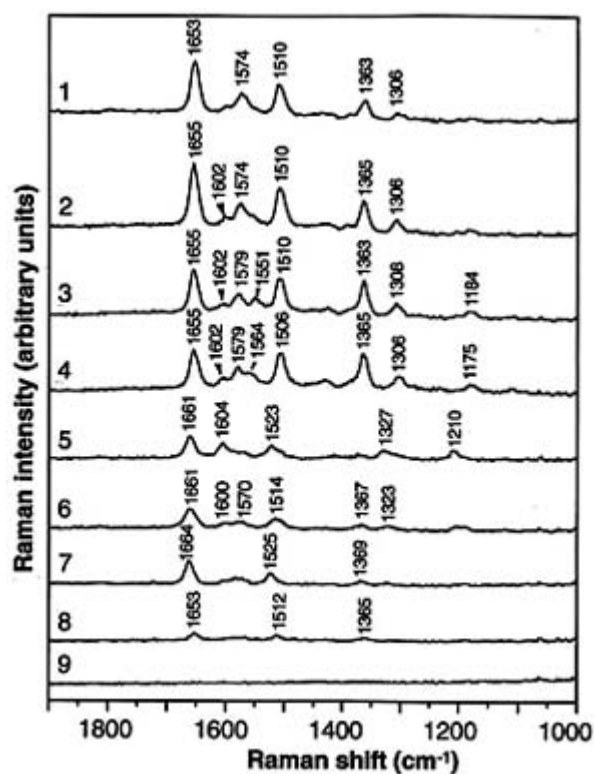


Figure C1.5.7. Surface-enhanced Raman spectra of a single rhodamine 6G particle on silver recorded at 1 s intervals. Over 300 spectra were recorded from this particle before the signals disappeared. The nine spectra displayed here were chosen to highlight several as yet unexplained sudden changes in both frequency and intensity. Reprinted with permission from Nie and Emory [53]. Copyright 1997 American Association for the Advancement of Science.

It now seems clear that, under certain conditions, massive enhancements of what is normally a very weak process can be achieved. The ability to obtain vibrational spectra would be a great advance in the characterization of single molecules if methods could be found to reproducibly observe all molecules in a sample, not only those that happen to bind to special sites on the colloid.

C1.5.3.3 MULTIPHOTON EXCITATION

Electronic excitation to a fluorescent state can be accomplished not only by direct one-photon absorption to that state, but also by the ‘simultaneous’ absorption of two or more photons whose energy sums to that of the excited state. Multiphoton absorption cross sections are sufficiently small that these processes are negligibly weak with cw light sources, but two- and three-photon absorption become viable processes for single-molecule detection at the high peak intensities provided by focusing femtosecond pulses to diffraction-limited spot sizes. In particular, the development of Ti:sapphire lasers reliably delivering femtosecond pulses in the far-red and near-infrared regions of the spectrum has enabled two- and three-photon excited fluorescence microscopy of blue and ultraviolet chromophores [58, 59]. Multiphoton absorption is sometimes referred to as ‘intrinsically confocal’ since the probability of absorbing two or three photons depends on the second or third power of the laser intensity, respectively, greatly enhancing the

contribution to the signal from molecules at the beam focus over those outside the focus, even without any additional aperturing. The transverse resolution can be somewhat enhanced for the same reason, although this advantage is partially offset by the larger focused spot size imposed by the longer wavelength of the laser. In addition, the negligible absorption of the unfocused red or infrared light by many samples, particularly biological samples, allows probing at comparatively great depths within samples that often absorb strongly in the UV.

Two-photon excited fluorescence detection at the single-molecule level has been demonstrated for chromophores in cryogenic solids [60], room-temperature surfaces [61], membranes [62] and liquids [63, 64 and 65]. Although multiphoton excited fluorescence has been embraced with great enthusiasm as a technique for both ordinary confocal microscopy and single-molecule detection, it is not a panacea; in particular, photochemical degradation in multiphoton excitation may be more severe than with ordinary linear excitation, probably due to absorption of more than the desired number of photons from the intense laser pulse (e.g. triplet excited state absorption) [61].

C1.5.4 SYSTEMS AND PHENOMENA

C1.5.4.1 SPECTROSCOPY AND PHOTOPHYSICS

(A) CRYOGENIC STUDIES

Studies of single molecules in cryogenic solids, while limited to a relatively small number of chromophore/ matrix combinations (and small variations thereupon), have covered a wide range of spectroscopic and dynamic processes [66, 67].

The spectral and temporal characteristics of the fluorescence excitation spectra have been examined for several chromophores in a variety of single crystalline, Sh'polskii, and amorphous matrices. Two distinct methods are employed in these studies: direct measurements in which the laser frequency is scanned repetitively across a single-molecule electronic origin while counting emitted photons, and correlation techniques in which the laser frequency is fixed and the autocorrelation statistics of emitted photons measured. Both techniques can provide information about the dynamics of fluctuations in the electronic origin frequency of a single molecule, although the correlation techniques are useful over a broader range of time scales. The direct technique also permits the electronic origin linewidths for different single molecules to be measured, with the limitation that the time required to accurately characterize the lineshape by scanning the laser frequency ranges from fractions of a second to minutes, and any fluctuations occurring on faster time scales are folded into the apparent linewidth. In all systems examined to date, at least some molecules exhibit time-dependent fluctuations in electronic origin frequency ranging from a few MHz to many GHz or more. These occur on time scales from microseconds to hours, and may be apparently spontaneous (usually denoted 'spectral diffusion') and/or light-induced (often called 'hole burning'). In general the narrowest lines, as narrow as the lifetime limit based on the fluorescence lifetime measured in bulk experiments, and most stable spectra are observed in single crystals: pentacene [6, 68 and 69] and terrylene [70, 71] in *p*-terphenyl, and terrylene [72], dibenzoterrylene [73] and dibenzanthanthrene [74] in naphthalene. These and closely related chromophores in *n*-alkane Sh'polskii matrices (terrylene in *n*-alkanes [75, 76, 77 and 78], perylene in *n*-nonane [79], dibenzanthanthrene [80] and terrylenediimide [81] in *n*-hexadecane) tend to show more spectral diffusion and a broader distribution of apparent linewidths, up to several times the lifetime limit. Even more spectral lability is observed in amorphous (polymer) matrices; the electronic origin linewidths measured over any finite period of time often exceed the lifetime limit by more than an order of

magnitude, and a rich variety of spectral diffusion and spectral jumping behaviours are observed (see figure C1.5.8 [78, 81, 82, 83, 84, 85, 86 and 87]). The spontaneous spectral diffusion that persists even down to 1.5 K is attributed to tunnelling between 'two-level systems', different configurations of the matrix separated by small barriers that are nearly isoenergetic for the chromophore in its ground electronic state but have significantly different energies in the excited electronic state, thus shifting the electronic transition frequency. Considerable theoretical work has been and continues to be performed to understand the physical origin of these two-level systems and to interpret the observed linewidth distributions and dynamics of the spectral diffusion [88, 89, 90, 91, 92, 93, 94 and 95].

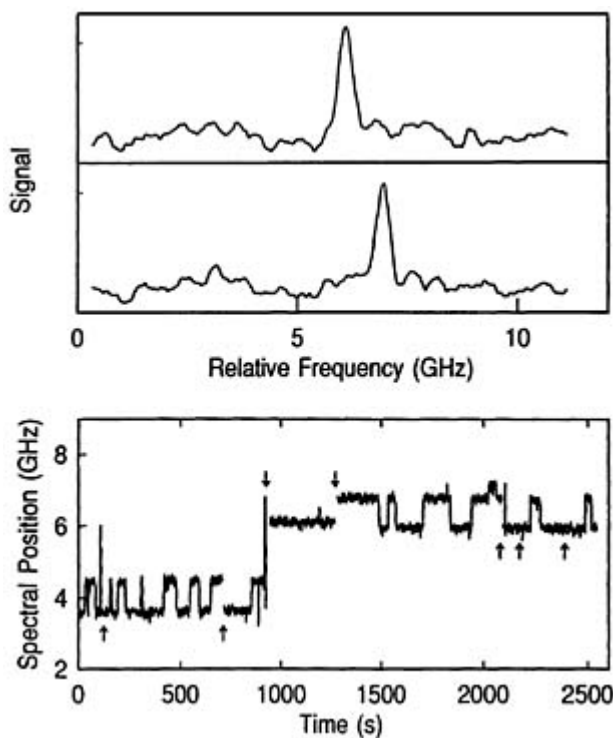


Figure C1.5.8. Spectral jumping of a single molecule of terrylene in polyethylene at 1.5 K. The upper trace displays fluorescence excitation spectra of the same single molecule taken over two different 20 s time intervals, showing the same molecule absorbing at two distinctly different frequencies. The lower panel plots the peak frequency in the fluorescence excitation spectrum as a function of time over a 40 min trajectory. The molecule undergoes discrete jumps among four (briefly five) different resonant frequencies during this time period. Arrows represent scans during which the molecule had jumped entirely outside the 10 GHz scan window. Adapted from [199].

The Stark effect (shifting of the absorption frequency in response to an external electric field) has been measured for nominally centrosymmetric single molecules both in mixed crystals (pentacene in *p*-terphenyl) [96] and in amorphous solids (terrylene in polyethylene) [97]. In the former system the energy shifts are dominated by the term quadratic in applied field as expected for a centrosymmetric system, but in the latter the Stark shifts are dominated by the linear term and vary widely in both magnitude and sign among different molecules. This is excellent evidence for large and variable dipole moments induced in nominally centrosymmetric chromophores by the disorder in the environment, and dramatically demonstrates the strength of the local electric fields produced by even the moderately polar bonds of a simple hydrocarbon environment.

-15-

Line shifts as a function of pressure have been studied for pentacene and terrylene in *p*-terphenyl [98, 99]. Both exhibited linear and reversible spectral red shifts with increasing pressure. Modest variations (factors of 1.3–1.6) in the pressure shifts among molecules were attributed to slightly different local environments.

Fluorescence lifetimes have been measured directly by time-correlated single-photon counting for pentacene in *p*-terphenyl [100]. This experiment requires careful selection of the laser pulse characteristics such that the pulse duration is short enough to resolve the 23 ns decay time, yet has a bandwidth narrow enough to allow spectral selection of individual molecules. Four different molecules had the same lifetime to within experimental uncertainty, indicating that the principal contributions to the S_1 state decay (radiation and internal conversion to S_0) are not strongly sensitive to the local environment in this relatively homogeneous crystalline matrix.

The polarization properties of single-molecule fluorescence excitation spectra have been explored and utilized to determine both the molecular transition dipole moment orientation and the depth of single pentacene molecules in a *p*-terphenyl crystal, taking into account the rotation of the polarization of the excitation light by the birefringent

crystal [101, 102].

Dispersed fluorescence spectra showing resolution of the ground-state vibrations have been reported for single molecules of pentacene in *p*-terphenyl [103, 104], terrylene in *p*-terphenyl [71] and terrylene in polyethylene [105, 106 and 107]. In the former system all molecules were found to have quite similar spectra, but the small variations between crystallographically distinct O₁ and O₂ sites were shown to be reproducible enough to distinguish between sites [104], and other small variations due to either natural abundance isotopic substitution or local defect induced redistributions of intensity were noted [104]. Terrylene in both crystalline and amorphous matrices exhibits significant spectral variations among molecules (see figure C1.5.9). The two distinct types of spectra observed in polyethylene were suggested to arise from terrylene molecules in two very different local environments, perhaps amorphous and crystalline, although the possibility of a chemical impurity could not be ruled out at the time and has since been suggested as the correct explanation [108]. These experiments highlight the promise of low-temperature single-molecule spectroscopy for making very detailed spectroscopic measurements probing correlations among various spectroscopic observables, e.g., electronic and vibrational frequencies. They also suggest the possibility that single-molecule spectroscopy could provide the vibrational frequencies of ¹³C substituted isotopomers, essential in classical vibrational analysis, without the need to synthesize specifically labelled material.

-16-

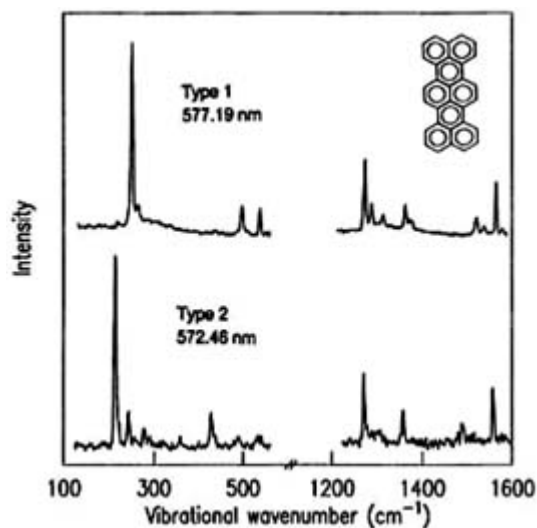


Figure C1.5.9. Vibrationally resolved dispersed fluorescence spectra of two different single molecules of terrylene in polyethylene. The excitation wavelength for each molecule is indicated and the spectra are plotted as the difference between excitation and emitted wavenumber. Each molecule's spectrum was recorded on a CCD detector at two different settings of the spectrograph grating to examine two different regions of the emission spectrum. 'Type 1' and 'type 2' spectra were tentatively attributed to terrylene molecules in very different local environments, although the possibility that type 2 spectra arise from a chemical impurity could not be ruled out. Further details are given in Tchénio [105–107].

A long-range goal of some single-molecule work is to engineer single-molecule optical devices. Toward this, two groups have demonstrated the ability to use light to drive single molecules of terrylene among two or more stable states in a predictable manner—a single-molecule optical 'switch' [77, 109]. The optical properties of terrylene in *n*-octane have been modified by exciting nearby triphenylene molecules to their triplet state [110], and this idea has been extended to probe the dynamics of triplet excitons in isotopically mixed naphthalene crystals using frequency shifts of single molecules of terrylene as a local environmental probe [111].

Finally, the ability to optically address single molecules is enabling some beautiful experiments in quantum optics. The non-Poissonian photon arrival time distributions expected theoretically for single molecules have been observed directly, both antibunching at short times [112] and bunching on longer time scales [6, 112 and 113]. The fluorescence excitation spectra of single molecules bound to spherical microcavities have been examined as a probe

of the optical resonances of small particles [114]. A variety of nonlinear optical effects have been observed at high laser intensities, including the AC Stark effect and Rabi oscillations, and exceptional agreement with theoretical predictions has been shown (see figure C1.5.10) [115, 116, 117 and 118]. In a particularly exciting application of quantum optics with single molecules, dibenzanthanthrene in *n*-hexadecane has been made to act as a triggered source of single photons by using the method of adiabatic following to prepare a single molecule in its fluorescent state [119].

-17-

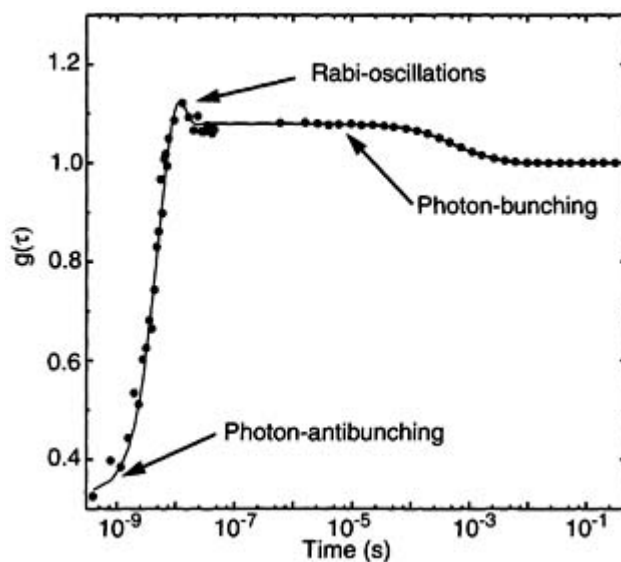


Figure C1.5.10. Normalized fluorescence intensity correlation function for a single terrylene molecule in *p*-terphenyl at 2 K. The solid line is the theoretical curve. Regions of deviation from the long-time value of unity due to photon antibunching (the finite lifetime of the excited singlet state), Rabi oscillations (absorption-stimulated emission cycles driven by the laser field) and photon bunching (dark periods caused by intersystem crossing to the triplet state) are indicated. Reproduced with permission from Plakhotnik *et al* [66], adapted from [118].

(B) ROOM TEMPERATURE STUDIES

Spatial selection techniques, both near-field and far-field, have been employed to examine a variety of spectroscopic and photophysical properties of single molecules at room temperature. Near-field and far-field images, fluorescence spectra, and lifetimes have been compared and found to be consistent under appropriate conditions (see figure C1.5.11), although the close proximity of the metallic tip used in near-field experiments can also alter fluorescence lifetimes as discussed above. While most of these studies involve dye molecules that have considerably more conformational flexibility than the rigid aromatics used in the low-temperature spectral selection experiments, single-molecule data over the full range from superfluid helium to room temperature have been obtained for a few chromophore/matrix combinations [81, 120]. The spectral linewidths (integrated over the required accumulation times of milliseconds or more) as well as the observed spectral jumps are orders of magnitude larger in the room temperature experiments, but similar phenomena of highly variable linewidths and resonant frequencies among different molecules, spontaneous and light-induced spectral jumping and ‘blinking’ of the fluorescence when exciting at a fixed frequency are observed [26, 121, 122, 123, 124, 125, 126, 127, 128 and 129]. Photon bunching in the emission correlation function due to triplet state formation has also been observed at room temperature [129, 130].

-18-

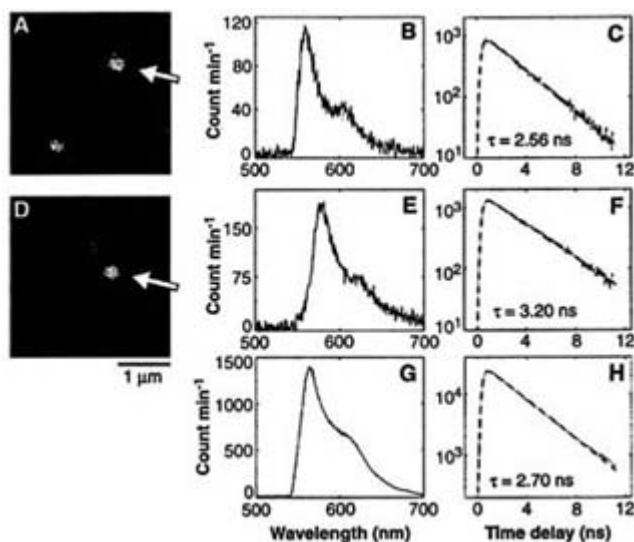


Figure C1.5.11. Far-field fluorescence images (A and D), corresponding fluorescence spectra (B and E), and fluorescence decays (C and F) for two different molecules of a carbocyanine dye at a PMMA–air interface. Lifetimes were fitted to a single exponential (dotted curves) with decay times of 2.56 ns ($\chi^2 = 1.05$) in (C) and 3.20 ns ($\chi^2 = 1.16$) in (F). For comparison, an ensemble measurement averaged over several hundred molecules is shown in (G) and (H). A single exponential fit to the lifetime yields a decay time of 2.70 ns ($\chi^2 = 6.7$). The larger χ^2 indicates a deviation from single-exponential behaviour, reflecting the ensemble average over a distribution of lifetimes. Reprinted with permission from Macklin *et al* [126]. Copyright 1996 American Association for the Advancement of Science.

The three-dimensional orientation of a single molecule's transition dipole has been determined using near-field optics, taking advantage of the longitudinal component of the electric field near the tip [14]. Similar determinations have been made using far-field optics with a 'donut mode' laser beam [123] and by analysing the intensity patterns from far-field fluorescence images under slightly aberrating conditions [131]. The far-field techniques in particular should prove invaluable for following orientational motions of single molecules in environments that permit such motion, as discussed briefly below.

C1.5.4.2 MAGNETIC RESONANCE OF CHROMOPHORES IN SOLIDS

Magnetic resonance techniques, while powerful spectroscopic probes of molecular structure (sections B1.12–B1.16), typically have quite low sensitivities, and direct detection of single nuclear or electron spins has yet to be demonstrated. However, electron spin resonance at the single-molecule level has been demonstrated through the indirect technique of optically detected magnetic resonance. The original experiments exploited the dependence of the time-averaged fluorescence intensity on the rate of intersystem crossing from the fluorescent singlet to the essentially nonemissive triplet state. The splittings among the magnetic components of the triplet state were detected by sweeping the RF field while measuring the total fluorescence intensity from a single molecule selected out of the inhomogeneous ensemble by its fluorescence excitation frequency [132, 133]. This idea has subsequently been extended to examine isotope effects on the rf resonant linewidths [134, 135] and to demonstrate single-spin coherence and spin echo phenomena [136, 137].

C1.5.4.3 CHEMICAL REACTIONS

Chemical reactions can be studied at the single-molecule level by measuring the fluorescence lifetime of an excited state that can undergo reaction in competition with fluorescence. Reactions involving electron transfer (section C3.2) are among the most accessible via such techniques, and are particularly attractive candidates for study as a means of testing relationships between charge-transfer optical spectra and electron-transfer rates. If the physical parameters that determine the reaction probability, such as overlap between the donor and acceptor orbitals,

thermodynamic driving force for the reaction and nuclear reorganization energies, are not constant across the ensemble, different molecules will exhibit different electron transfer ‘rates’, as defined through multiple measurements on the same molecule. A very broad distribution of lifetimes, and thereby electron transfer rates, was observed for excited cresyl violet molecules transferring an electron to an indium tin oxide surface (figure C1.5.12) [138]. In contrast, a study of the Os(VIII) ion-catalysed redox reaction between Ce(IV) and As(III), producing a fluorescent Ce(III) product, found uniform catalytic activities for different Os(VIII) ions [139]. Single-molecule rate measurements coupled with spectroscopy hold great promise for allowing the factors that dictate electron transfer and other chemical reactions to be teased apart.

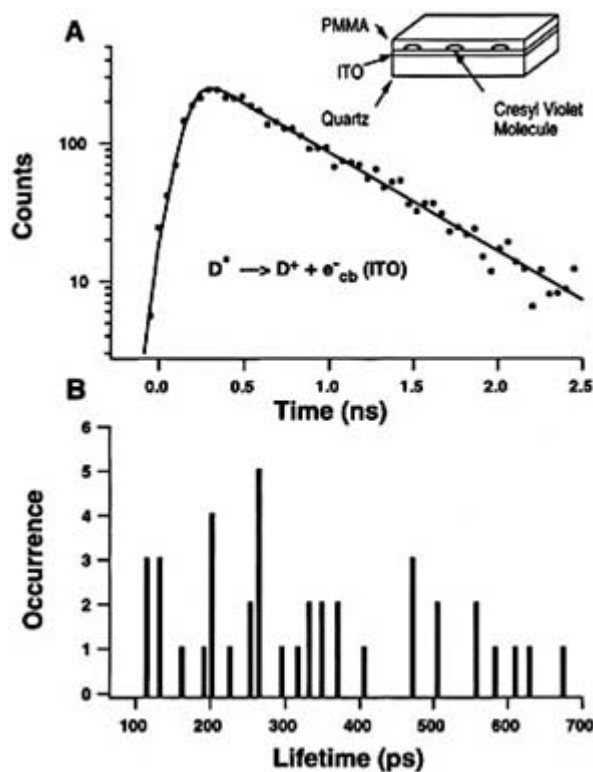


Figure C1.5.12.(A) Fluorescence decay of a single molecule of cresyl violet on an indium tin oxide (ITO) surface measured by time-correlated single photon counting. The solid line is the fitted decay, a single exponential of 480 ± 5 ps convolved with the instrument response function of 160 ps fwhm. The decay, which is considerably faster than the natural fluorescence lifetime of cresyl violet, is due to electron transfer from the excited cresyl violet (D^*) to the conduction band or energetically accessible surface electronic states of ITO. (B) Distribution of lifetimes for 40 different single molecules showing a broad distribution of electron transfer rates. Reprinted with permission from Lu and Xie [138]. Copyright 1997 American Chemical Society.

C1.5.4.4 TRANSLATIONAL AND ROTATIONAL MOTIONS

While translational and rotational diffusion are very well understood at the bulk level, there are advantages to being able to observe trajectories of individual molecules, particularly in nonhomogeneous materials. Fluorescence microscopy has been used to follow single-molecule translation over long time scales in hindered and/or anisotropic environments such as polymers [140, 141], gels [142], engineered submicrometre channels [143] and lipid membranes [144], as well as in aqueous solution [145] and to quantify the ‘optical tweezers’ effect whereby a polarizable molecule is attracted into the focus of an intense light source [146, 147]. The three-dimensional rotational motions of chromophores bound to polymers can be monitored by the intensity distribution patterns in confocal fluorescence microscopy (figure C1.5.13) figure C1.5.14 and figure C1.5.15 [148, 149]. Polarization modulation techniques have been employed to probe rotational motions for chromophores on surfaces and bound to polymers [127, 150 and 151] and near-field excitation with two polarization detection channels has been used to

examine rotation of dye molecules on glass surfaces and in polymers [141]. These techniques seem likely to find broad applicability as probes of local motion in nano- and mesostructured materials, as well as in monitoring the conformational dynamics of biopolymers such as DNA and proteins.

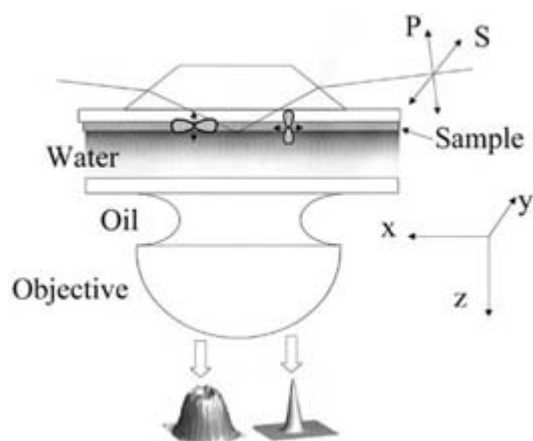


Figure C1.5.13. Schematic diagram of an experimental set-up for imaging 3D single-molecule orientations. The excitation laser with either s- or p-polarization is reflected from the polymer/water boundary. Molecular fluorescence is imaged through an aberrating thin water layer, collected with an inverted microscope and imaged onto a CCD array. Aberrated and unaberrated emission patterns are observed for z- and xy-orientated molecules, respectively. Reprinted with permission from Bartko and Dickson [148]. Copyright 1999 American Chemical Society.

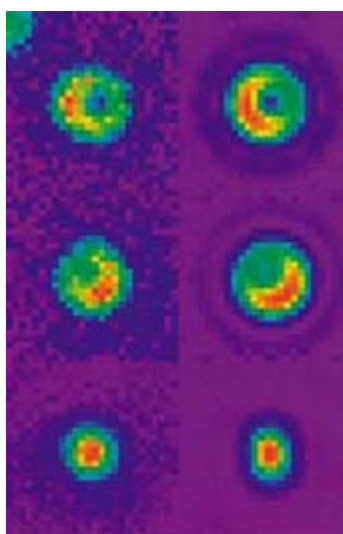


Figure C1.5.14. Fluorescence images of three different single molecules observed under the imaging conditions of figure C1.5.13. The observed dipole emission patterns (left column) are indicative of the 3D orientation of each molecule. The right-hand column shows the calculated fit to each observed intensity pattern. Molecules 1, 2 and 3 are found to have polar angles of $(\theta, \phi) = (4.5^\circ, -24.6^\circ)$, $(-5.3^\circ, 51.6^\circ)$ and $(85.4^\circ, -3.9^\circ)$, respectively. Reprinted with permission from Bartko and Dickson [148]. Copyright 1999 American Chemical Society.

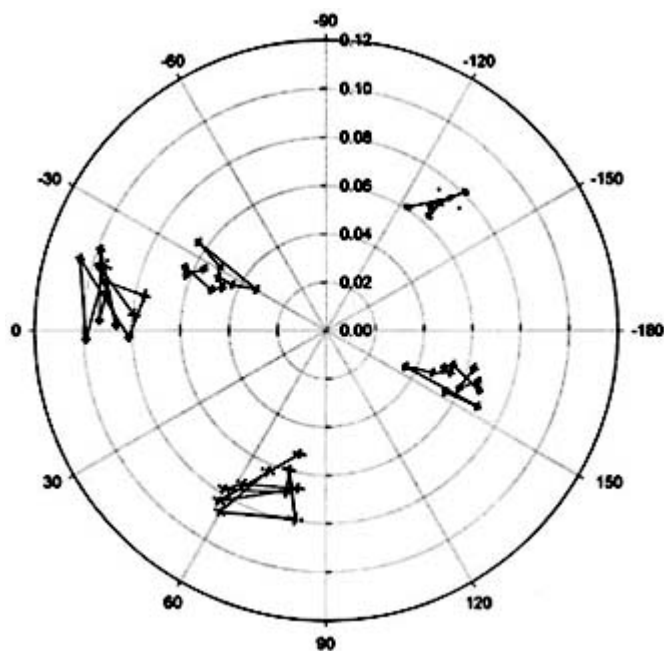


Figure C1.5.15. Molecular orientational trajectories of five single molecules. Each step in the trajectory is separated by 300 ms and is obtained from the fit to the dipole emission pattern such as is shown in figure C1.5.14. The radial component is displayed as $\sin \theta$ and the angular variable as ϕ . The lighter dots around the average orientation represent ± 1 standard deviation. Reprinted with permission from Bartko and Dickson [148]. Copyright 1999 American Chemical Society.

-22-

C1.5.4.5 CONJUGATED POLYMERS, CHROMOPHORE AGGREGATES, AND MICROHETEROGENEOUS MATERIALS

Materials that possess fundamental optical inhomogeneities are among the most appealing candidates for single-molecule techniques. These include conjugated polymers such as poly(phenylenevinylene) and its relatives, in which different polymer chains differ not only in their physical length but also in their effective electronic conjugation length, and noncovalent structures such as submicroscopic crystals and the J-aggregates formed by certain cyanine dyes, in which different ‘supermolecules’ differ in both numbers of chromophores and their spatial relationships. Large reversible fluorescence intensity fluctuations observed in far-field single-molecule studies on conjugated polymers have been interpreted in terms of excited-state quenching by a photochemically produced charge-separated state [129, 152 and 153]. Near-field techniques are particularly valuable in these systems, which have a functionally important structure on nanometre length scales. Near-field scanning optical microscopy, including analysis of the excitation and emission polarization properties, has been used to probe spatial relationships between excitation and emission in J-aggregates, providing information about the degree of order in the aggregate and the spatial extent of exciton migration [154, 155 and 156]. NSOM has also been applied to small molecular crystals to examine spatial inhomogeneities, energy transfer, and excitation trapping [156, 157 and 158].

Single molecules also have promise as probes for local structure when doped into materials that are themselves nonfluorescent. Rhodamine dyes in both silicate and polymer thin films exhibit a distribution of fluorescence maxima indicative of considerable heterogeneity in local environments, particularly for the silicate material [159]. A bimodal distribution of fluorescence intensities observed for single molecules of crystal violet in a PMMA film has been suggested to result from high and low viscosity local sites within the polymer that give rise to slow and fast internal conversion, respectively [160].

C1.5.4.6 METALLIC AND SEMICONDUCTOR NANOPARTICLES

Metallic and semiconductor ‘nanoparticles’ or ‘nanocrystals’—chunks of matter intermediate in size and physical properties between single atoms and the macroscopic bulk materials—are of great interest both for their

fundamental properties and for technological applications (section C2.17). Although methods have been developed to synthesize some of these materials with high monodispersity, even the best preparations have significant variations from particle to particle in the number of atoms and/or the geometry of the particle, making the ability to interrogate single particles particularly valuable. Generally these materials can be prepared such that the individual particles are spaced arbitrarily far apart, making far-field microscopy the optical technique of choice.

Size-dependent optical properties of single silver and gold nanoparticles have been studied [161, 162], as have the surface Raman enhancements of organic molecules bound to these nanoparticles, as discussed above [53]. Several beautiful spectroscopic studies have been carried out on single II–VI semiconductor nanocrystals, revealing very well resolved low-temperature fluorescence spectra having well defined phonon substructures [163, 164 and 165] (see figure C1.5.16), large Stark shifts in these spectra [166], pronounced spectral diffusion [164, 166, 167] and dramatic intensity-dependent effects on the spectra [165]. The spectral diffusion was interpreted as being due to randomly fluctuating local electric fields caused by charge carriers trapped at surface defects [166]. Recently, single-particle techniques have also been used to resolve luminescence from the individual ‘chromophores’ of nanostructured porous silicon [168].

-23-

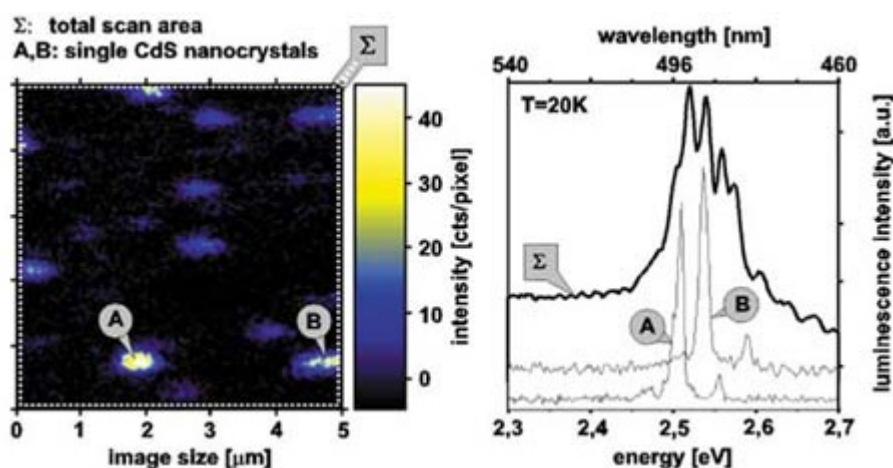


Figure C1.5.16. Spatial selection of single CdS nanocrystals on a quartz cover slip. Left: fluorescence image of isolated CdS nanocrystals recorded via scanning in a low-temperature confocal microscope ($T=20$ K). Total fluorescence was excited at 442 nm. Fifty-five per cent of the fluorescence was directed to a spectrometer during the scan, resulting in the Σ spectrum on the right. Spectra A and B were taken after moving the scanner to the bright spots marked in the fluorescence image, reducing the excitation intensity and increasing the acquisition time. Under these conditions both spectra show a second emission line shifted 5 meV (40 cm^{-1}) to the blue of the main peak. Reprinted with permission from Koberling *et al* [165]. Copyright 1999 by the American Physical Society.

C1.5.4.7 BIOLOGICAL SYSTEMS

The application of single molecule optical techniques to biological phenomena is an area currently seeing explosive growth [169].

The most obvious biological studies to undertake with optical techniques at the level of single functional units (generally not single *chromophores*) involve probing the photophysics of systems whose function, whether naturally evolved or engineered by man, involves responding to light. Photosynthetic light-harvesting complexes from green algae were imaged by near-field techniques at room temperature and fluorescence lifetimes measured with picosecond time resolution [170]. Subsequently, far-field microscopy has been used to obtain fluorescence excitation spectra of single bacterial light-harvesting complexes at 1.2 K [171, 172] and as a function of temperature [173] providing detailed information about individual chromophores’ site energies and energetic disorder, dipolar couplings between chromophores and spectral diffusion. Confocal fluorescence microscopy of single light-harvesting complexes at room temperature showed that photobleaching of just one bacteriochlorophyll

molecule of the 18-member assembly provides an energy trap that effectively quenches fluorescence from the entire assembly [174]. Single molecules of green fluorescent protein, widely used as a fluorescent tag in molecular biology, exhibit pronounced on-off 'blinking' effects whose origin has yet to be understood [175, 176].

Single-molecule optical methods have also been adapted to probe biological systems whose function is unrelated to light. Much of the research on single-molecule detection in liquids and gels is directed toward rapid DNA sequencing, utilizing either intrinsic fluorescence or, more likely, exogenous dyes bound to specific DNA bases [44, 45, 64, 177 and 178]. Single-molecule optical techniques are being used to study a variety of chemical and physical properties of

-24-

oligonucleotides, RNA and DNA including base pairing [179], electron transfer between dyes and DNA bases [180], ligand-induced conformational changes in RNA [181] and conformational dynamics in oligonucleotides [127, 150, 182, 183, 184 and 185], as well as to selectively cleave DNA [186]. Conformational dynamics in proteins that may have functional significance are being accessed through polarization studies, fluorescence quenching and energy transfer techniques carried out at the single-molecule level [187, 188]. Total internal reflection fluorescence microscopy has been used to visualize the motions of single molecules of fluorescently labelled kinesin, the motor protein that powers organelle transport along microtubules [189]. Dunn and co-workers are applying single-molecule techniques to study the morphology and dynamics of microenvironments in model biological membranes [190, 191].

Perhaps the most exciting application of single-molecule techniques in biology is the probing of enzymatic reactions at the level of individual turnovers [192, 193], utilizing systems in which either enzyme or substrate or both undergo large changes in optical properties (e.g. switching from fluorescent to nonfluorescent states) during the course of the reaction. These studies allow the possibility of heterogeneity in reaction rates among nominally identical enzyme molecules to be assessed, and make it possible to probe 'memory' effects, the extent to which an enzyme's binding constant or turnover rate depends upon its previous reaction history. Generally the enzyme is either bound to a surface [194] or confined within the pores of a gel [142] or nanoengineered structure [139, 143], and the substrate and product molecules allowed to diffuse toward and away from the immobilized enzyme. The catalytic activity among different molecules of nominally identical lactate dehydrogenase enzyme was found to vary by up to a factor of four, an observation tentatively attributed to the presence of multiple stable conformers of the enzyme [195]. A detailed study of single molecules of mammalian alkaline phosphatase revealed even more pronounced heterogeneities among different molecules (more than tenfold differences in turnover rate and more than two-fold differences in activation energy), which were attributed at least in part to post-translational modification producing chemically nonidentical enzyme molecules [196]. In contrast, single molecules of highly purified bacterial alkaline phosphatase have indistinguishable enzymatic activities [137]. Lengthy turnover trajectories of cholesterol oxidase, a flavoprotein that catalyses the oxidation of cholesterol by oxygen (figure C1.5.17), have been analysed to obtain the distribution of 'on' and 'off' times in the Michaelis-Menten catalytic mechanism, and also revealed evidence for memory effects probably due to slow conformational fluctuations in the protein [192, 193].

-25-

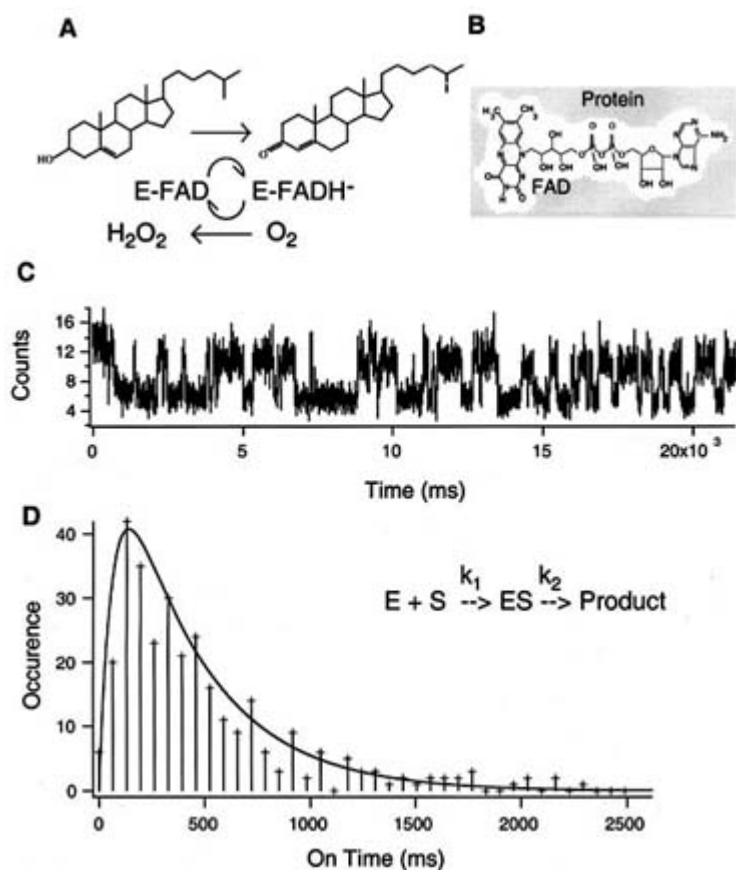


Figure C1.5.17.(A) Enzymatic cycle of cholesterol oxidase, which catalyses the oxidation of cholesterol by molecular oxygen. The enzyme's naturally fluorescent FAD active site is first reduced by a cholesterol substrate, generating a nonfluorescent FADH⁻, and then re-oxidized by molecular oxygen. (B) Structure of FAD. (C) Fluorescence intensity trajectory of an individual cholesterol oxidase enzyme, immobilized in an agarose gel, undergoing reaction with cholesterol and oxygen. Each on-off cycle of emission corresponds to one enzymatic turnover. (D) Distribution of emission on-times derived from the intensity trajectory. The nonexponential distribution reflects the fact that the forward reaction is not a single elementary step but involves an intermediate enzyme-substrate complex as shown in the inset. The solid line is a curve simulated by convolving two exponential distributions with $k_1[S]=2.5 \text{ s}^{-1}$ and $k_2[S]=15.3 \text{ s}^{-1}$. Reproduced with permission from Xie and Trautman [123]

C1.5.5 CONCLUSION

The ability to make optical measurements on individual molecules and submicroscopic aggregates, one at a time, is a valuable new tool in several areas of molecular science. By eliminating inhomogeneous broadening it allows pure spectroscopy to be performed with unprecedented precision in certain condensed phase systems. As an analytical method it permits the rapid detection of certain analytes with unmatched sensitivity. Finally, it is revolutionizing our

understanding of the relationships among nominally identical members of molecular ensembles. The deepest and most lasting contribution of this technique is likely to be in its application to complex systems such as biological macromolecules, polymers, and engineered nanostructures, where ensembles of members that are 'identical' to the best of nature's or man's efforts still exhibit differences on time scales that are of functional importance.

The techniques and applications of single molecule spectroscopy are currently in a state of rapid development, making this a difficult field to summarize at any given time. This contribution is, at best, a single frame of a movie

featuring plenty of action, improvisation, and unexpected plot twists.

ACKNOWLEDGMENTS

I am grateful to W E Moerner, Sunney Xie, Taras Plakhotnik, Felix Koberling, Shuming Nie, Dehong Hu, and Rob Dickson for providing the figures used in this contribution and/or for communicating preprints of their work prior to publication, and to Professor Dan Higgins for his comments on an early version of this chapter.

REFERENCES

- [1] Orrit M, Bernard J and Personov R I 1993 High-resolution spectroscopy of organic molecules in solids: from fluorescence line narrowing and hole burning to single molecule spectroscopy *J. Phys. Chem.* **97** 10 256–68
 - [2] Abram I I, Auerbach R A, Birge R R, Kohler B E and Stevenson J M 1975 Narrow-line fluorescence spectra of perylene as a function of excitation wavelength *J. Chem. Phys.* **63** 2473–8
 - [3] Lee H W H, Walsh C A and Fayer M D 1985 Inhomogeneous broadening of electronic transitions of chromophores in crystals and glasses: analysis of hole burning and fluorescence line narrowing experiments *J. Chem. Phys.* **82** 3948–58
 - [4] Moerner W E and Kador L 1989 Optical detection and spectroscopy of single molecules in solids *Phys. Rev. Lett.* **62** 2535–8
 - [5] Kador L, Horne D E and Moerner W E 1990 Optical detection and probing of single dopant molecules of pentacene in a *p*-terphenyl host by means of absorption spectroscopy *J. Phys. Chem.* **94** 1237–48
 - [6] Orrit M and Bernard J 1990 Single pentacene molecules detected by fluorescence excitation in a *p*-terphenyl crystal *Phys. Rev. Lett.* **65** 2716–19
 - [7] Kador L, Latychevskaia T, Renn A and Wild U P 1999 Absorption spectroscopy on single molecules in solids *J. Phys. Chem* **111** 8755–8
 - [8] Hirschfeld T 1976 Optical microscopic observation of single small molecules *Appl. Opt.* **15** 2965–6
 - [9] Nguyen D C, Keller R A, Jett J H and Martin J C 1987 Detection of single molecules of phycoerythrin in hydrodynamically focused flows by laser-induced fluorescence *Anal. Chem.* **59** 2158–61
 - [10] Peck K, Stryer L, Glazer A N and Mathies R A 1989 Single-molecule fluorescence detection: autocorrelation criterion and experimental realization with phycoerythrin *Proc. Natl Acad. Sci. USA* **86** 4087–91
 - [11] Shera E B, Seitzinger N K, Davis L M, Keller R A and Soper S A 1990 Detection of single fluorescent molecules *Chem. Phys. Lett.* **174** 553–7
 - [12] Nie S, Chiu D T and Zare R N 1994 Probing individual molecules with confocal fluorescence microscopy *Science* **266** 1018–21
-

- [13] Betzig E and Trautman J K 1992 Near-field optics: microscopy, spectroscopy, and surface modification beyond the diffraction limit *Science* **257** 189–95
- [14] Betzig E and Chichester R J 1993 Single molecules observed by near-field scanning optical microscopy *Science* **262** 1422–5
- [15] Moerner W E 1994 Fundamentals of single molecule spectroscopy in solids *J. Lumin.* **60–61** 997–1002
- [16] Moerner W E 1996 High-resolution optical spectroscopy of single molecules in solids *Acc. Chem. Res.* **29** 563–71
- [17] Orrit M, Bernard J, Brown R and Lounis B 1996 Optical spectroscopy of single molecules in solids *Progress in Optics* vol 35, ed E Wolf (Amsterdam: Elsevier) pp 61–144
- [18] Heinzelmann H and Pohl D W 1994 Scanning near-field optical microscopy *Appl. Phys. A* **59** 89–101
- [19] Zenhausern F, Martin Y and Wickramasinghe H K 1995 Scanning interferometric apertureless microscopy: optical imaging at 10 Ångstrom resolution *Science* **269** 1083–5

- [20] Hamann H F, Gallagher A and Nesbitt D J 1999 Enhanced sensitivity near-field scanning optical microscopy at high spatial resolution *Appl. Phys. Lett.* **75** 1469–71
- [21] Kopelman R and Tan W 1993 Near-field optics: imaging single molecules *Science* **262** 1382–4
- [22] Grober R D, Harris T D, Trautman J K and Betzig E 1994 Design and implementation of a low temperature near-field scanning optical microscope *Rev. Sci. Instrum.* **65** 626–31
- [23] Moerner W E, Plakhotnik T, Irrgartinger T, Wild U P, Pohl D W and Hecht B 1994 Near-field optical spectroscopy of individual molecules in solids *Phys. Rev. Lett.* **73** 2764–7
- [24] Durand Y, Woehl J C, Viellerobe B, Göhde W and Orrit M 1999 New design of a cryostat-mounted scanning near-field optical microscope for single molecule spectroscopy *Rev. Sci. Instrum.* **70** 1318–25
- [25] Ambrose W P, Goodwin P M, Martin J C and Keller R A 1994 Alterations of single-molecule fluorescence lifetimes in near-field optical microscopy *Science* **265** 364–7
- [26] Xie X S and Dunn R C 1994 Probing single molecule dynamics *Science* **265** 361–4
- [27] Bian R X, Dunn R C, Xie X S and Leung P T 1995 Single molecule emission characteristics in near-field microscopy *Phys. Rev. Lett.* **75** 4772–5
- [28] Sekatskii S K and Ketokhov V S 1996 Single fluorescence centres on the tips of crystal needles: first observation and prospects for application in scanning one-atom fluorescence microscopy *Appl. Phys. B* **63** 525–30
- [29] Kopelman R, Tan W and Birnbaum D 1994 Subwavelength spectroscopy, exciton supertips and mesoscopic light-matter interactions *J. Lumin.* **58** 380–7
- [30] Eigen M and Rigler R 1994 Sorting single molecules: application to diagnostics and evolutionary biotechnology *Proc. Natl Acad. Sci. USA* **91** 5740–7
- [31] van Oijen A M, Köhler J and Schmidt J 1999 Far-field fluorescence microscopy beyond the diffraction limit *J. Opt. Soc. Am. A* **16** 909–15
- [32] Hell S W and Nagorni M 1998 4Pi confocal microscopy with alternate interference *Opt. Lett.* **23** 1567–9
- [33] Klar T A and Hell S W 1999 Subdiffraction resolution in far-field fluorescence microscopy *Opt. Lett.* **24** 954–6
- [34] Güttler F, Irrgartinger T, Plakhotnik T, Renn A and Wild U P 1994 Fluorescence microscopy of single molecules *Chem. Phys. Lett.* **217** 393–7
- [35] Croci M, Irrgartinger T, Renn A and Wild U P 1994 Single molecule microscopy *Exp. Tech. Phys.* **41** 249–57
- [36] Jasny J, Sepiol J, Irrgartinger T, Traber M, Renn A and Wild U P 1996 Fluorescence microscopy in superfluid helium: single molecule imaging *Rev. Sci. Instrum.* **67** 1425–30
- [37] Ha T, Chemla D S, Enderle T and Weiss S 1997 Single molecule spectroscopy with automated positioning *Appl. Phys. Lett.* **70** 782–4

- [38] Weston K D, Carson P J, DeAro J A and Buratto S K 1999 Single-molecule detection fluorescence of surface-bound species in vacuum *Chem. Phys. Lett.* **308** 58–64
- [39] Göhde W, Tittel J, Basché T, Bräuchle C, Fischer U C and Fuchs H 1997 A low-temperature scanning confocal and near-field optical microscope *Rev. Sci. Instrum.* **68** 2466–74
- [40] Vácha M, Yokoyama N, Tokizaki T, Furuki M and Tani T 1999 Laser scanning microscope for low temperature single molecule and microscale spectroscopy based on gradient index optics *Rev. Sci. Instrum.* **70** 2041–5
- [41] Fleury L, Gruber A, Dräbenstedt A, Wrachtrup J and von Borczyskowski C 1997 Low-temperature confocal microscopy on individual molecules near a surface *J. Chem. Phys. B* **101** 7933–8
- [42] Zander C, Sauer M, Drexhage K H, Ko D-S, Schulz A, Wolfrun J, Brand L, Eggeling C and Seidel C A M 1996 Detection and characterization of single molecules in aqueous solution *Appl. Phys. B* **63** 517–23
- [43] Kung C, Barnes M D, Lermer N, Whitten W B and Ramsey J M 1999 Single-molecule analysis of ultradilute solutions with guided streams of 1 μm water droplets *Appl. Opt.* **38** 1481–7
- [44] Keller R A, Ambrose W P, Goodwin P M, Jett J H, Martin J C and Wu M 1996 Single-molecule fluorescence analysis in solution *Appl. Spectrosc.* **50** 12A–32A
- [45] Goodwin P M, Ambrose W P and Keller R A 1996 Single-molecule detection in liquids by laser-induced fluorescence *Acc. Chem. Res.* **29** 607–13
- [46] Soper S A and Legendre B L Jr 1998 Single-molecule detection in the near-IR using continuous-wave diode laser

excitation with an avalanche photon detector *Appl. Spectrosc.* **52** 1

- [47] Enderlein J, Robbins D L, Ambrose W P, Goodwin P M and Keller R A 1997 Statistics of single-molecule detection *J. Phys. Chem. B* **101** 3626–32
- [48] Keller R A 1998 Single-molecule identification in flowing sample streams by fluorescence burst size and intraburst fluorescence decay rate *Anal. Chem.* **70** 1444–51
- [49] Enderlein J, Goodwin P M, Van Orden A, Ambrose W P, Erdmann R and Keller R A 1997 A maximum likelihood estimator to distinguish single molecules by their fluorescence decays *Chem. Phys. Lett.* **270** 464–70
- [50] Müller R *et al* 1996 Time-resolved identification of single molecules in solution with a pulsed semiconductor diode laser *Chem. Phys. Lett.* **262** 716–22
- [51] Fries J R, Brand L, Eggeling C, Köllner M and Seidel C A M 1998 Quantitative identification of different single molecules by selective time-resolved confocal fluorescence spectroscopy *J. Phys. Chem. A* **102** 6602–13
- [52] Wilkerson C W Jr, Goodwin P M, Ambrose W P, Martin J C and Keller R A 1993 Detection and lifetime measurement of single molecules in flowing sample streams by laser-induced fluorescence *Appl. Phys. Lett.* **62** 2030–2
- [53] Nie S and Emory S R 1997 Probing single molecules and single nanoparticles by surface-enhanced Raman scattering *Science* **275** 1102–6
- [54] Michaels A M, Nirmal M and Brus L E 1999 Surface enhanced Raman spectroscopy of individual rhodamine 6G molecules on large Ag nanocrystals *J. Am. Chem. Soc.* **121** 9932–9
- [55] Kneipp K, Kneipp H, Deinum G, Itzkan I, Dasari R R and Feld M S 1998 Single-molecule detection of a cyanine dye in silver colloidal solution using near-infrared surface-enhanced Raman scattering *Appl. Spectrosc.* **52** 175–8
- [56] Kneipp K, Wang Y, Kneipp H, Perelman L T, Itzkan I, Dasari R R and Feld M S 1997 Single molecule detection using surface-enhanced Raman scattering (SERS) *Phys. Rev. Lett.* **78** 1667–70
- [57] Xu H, Bjerneld E J, Käll M and Börjesson L 1999 Spectroscopy of single hemoglobin molecules by surface enhanced Raman scattering *Phys. Rev. Lett.* **83** 4357–60
- [58] Strickler J H and Webb W W 1990 Two-photon excitation in laser scanning fluorescence microscopy *Proc. SPIE* **13948** 107–18

-29-

- [59] Brakenhoff G J, Squier J, Norris T, Bliton A C, Wade M H and Athey B 1996 Real-time two-photon confocal microscopy using a femtosecond, amplified Ti:sapphire system *J. Microscopy* **181** 253–9
- [60] Plakhotnik T, Walser D, Pirodda M, Renn A and Wild U P 1996 Nonlinear spectroscopy on a single quantum system: two-photon absorption of a single molecule *Science* **271** 1703–5
- [61] Sánchez E J, Novotny L, Holtom G R and Xie X S 1997 Room-temperature fluorescence imaging and spectroscopy of single molecules by two-photon excitation *J. Chem. Phys. A* **101** 7019–23
- [62] Sonnleitner M, Schütz G J and Schmidt T 1999 Imaging individual molecules by two-photon excitation *Chem. Phys. Lett.* **300** 221–6
- [63] Brand L, Eggeling C, Zander C, Drexhage K H and Seidel C A M 1997 Single-molecule identification of coumarin-120 by time-resolved fluorescence detection: comparison of one- and two-photon excitation in solution *J. Chem. Phys. A* **101** 4313–21
- [64] van Orden A, Cai H, Goodwin P M and Keller R A 1999 Efficient detection of single DNA fragments in flowing sample streams by two-photon fluorescence excitation *Anal. Chem.* **71** 2108–16
- [65] Mertz J, Xu C and Webb W W 1995 Single-molecule detection by two-photon-excited fluorescence *Opt. Lett.* **20** 2532–4
- [66] Plakhotnik T, Donley E A and Wild U P 1997 Single-molecule spectroscopy *Ann. Rev. Phys. Chem.* **48** 181–212
- [67] Moerner W E and Orrit M 1999 Illuminating single molecules in condensed matter *Science* **283** 1670–6
- [68] Ambrose W P and Moerner W E 1991 Fluorescence spectroscopy and spectral diffusion of single impurity molecules in a crystal *Nature* **349** 225–7
- [69] Ambrose W P, Basché T and Moerner W E 1991 Detection and spectroscopy of single pentacene molecules in a *p*-terphenyl crystal by means of fluorescence excitation *J. Phys. Chem.* **95** 7150–63
- [70] Kummer S, Basché T and Bräuchle C 1994 Terrylene in *p*-terphenyl: a novel single crystalline system for single molecule spectroscopy at low temperatures *Chem. Phys. Lett.* **229** 309–16
- [71] Kummer S, Kulzer F, Kettner R, Basché T, Tietz C, Glowatz C and Kryschi C 1997 Absorption, excitation, and

emission spectroscopy of terrylene in *p*-terphenyl: bulk measurements and single molecule studies *J. Phys. Chem* **107** 7673–84

- [72] Donley E A, Burzomato V, Wild U P and Plakhotnik T 1999 The distribution of linewidths of single probe molecules in a crystalline host at milliKelvin temperatures *J. Lumin.* **83–84** 255–9
- [73] Jelezko F, Tamarat P, Lounis B and Orrit M 1996 Dibenzoterrylene in naphthalene: a new crystalline system for single molecule spectroscopy in the near infrared *J. Chem. Phys.* **100** 13 892–4
- [74] Jelezko F, Lounis B and Orrit M 1997 Pump-probe spectroscopy and photophysical properties of single dibenzanthanthrene molecules in a naphthalene crystal *J. Phys. Chem* **107** 1692–702
- [75] Vacha M, Liu Y, Nakatsuka H and Tani T 1997 Inhomogeneous and single molecule line broadening of terrylene in a series of crystalline *n*-alkanes *J. Phys. Chem* **106** 8324–31
- [76] Plakhotnik T, Moerner W E, Irngartinger T and Wild U P 1994 Single molecule spectroscopy in Shpol'skii matrices *Chimia* **48** 31–2
- [77] Moerner W E, Plakhotnik T, Irngartinger T, Croci M, Palm V and Wild U P 1994 Optical probing of single molecules of terrylene in a Shpol'skii matrix: a two-state single-molecule switch *J. Chem. Phys.* **98** 7382–9
- [78] Kozankiewicz B, Bernard J and Orrit M 1994 Single molecule lines and spectral hole burning of terrylene in different matrices *J. Phys. Chem* **101** 9377–83
- [79] Pirotta M, Renn A, Werts M H V and Wild U P 1996 Single molecule spectroscopy. Perylene in the Shpol'skii matrix *n*-nonane *Chem. Phys. Lett.* **250** 576–82
- [80] Boiron A-M, Lounis B and Orrit M 1996 Single molecules of dibenzanthanthrene in *n*-hexadecane *J. Phys. Chem* **105** 3969–74
- [81] Mais S, Tittel J, Basché T and Bräuchle C 1997 Terrylenediimide: a novel fluorophore for single-molecule spectroscopy and microscopy from 1.4 K to room temperature *J. Chem. Phys. A* **101** 8435–40
- [82] Basché T and Moerner W E 1992 Optical modification of a single impurity molecule in a solid *Nature* **355** 335–7

-30-

- [83] Basché T, Ambrose W P and Moerner W E 1992 Optical spectra and kinetics of single impurity molecules in a polymer: spectral diffusion and persistent spectral hole burning *J. Opt. Soc. Am. B* **9** 829–36
- [84] Basché T, Kummer S and Bräuchle C 1995 Direct spectroscopic observation of quantum jumps of a single molecule *Nature* **373** 132–4
- [85] Tittel J, Kettner R, Basché T, Bräuchle C, Quante H and Müllen K 1995 Spectral diffusion in an amorphous polymer probed by single molecule spectroscopy *J. Lumin.* **64** 1–11
- [86] Zilker S J, Kador L, Friebel J, Vainer Y G, Kol'chenko M A and Personov R I 1998 Comparison of photon echo, hole burning, and single molecule spectroscopy data on low-temperature dynamics of organic amorphous solids *J. Phys. Chem* **109** 6780–90
- [87] Zumbusch A, Fleury L, Brown R, Bernard J and Orrit M 1993 Probing individual two-level systems in a polymer by correlation of single molecule fluorescence *Phys. Rev. Lett.* **70** 3584–7
- [88] Reilly P D and Skinner J L 1993 Spectral diffusion of single molecule fluorescence: a probe of low-frequency localized excitations in disordered crystals *Phys. Rev. Lett.* **71** 4257–60
- [89] Reilly P D and Skinner J L 1995 Spectral diffusion of individual pentacene molecules in *p*-terphenyl crystal: theoretical model and analysis of experimental data *J. Phys. Chem* **102** 1540–52
- [90] Skinner J L 1997 Theoretical models for the spectral dynamics of individual molecules in solids *Single Molecule Optical Detection, Imaging and Spectroscopy* ed T Basché, W E Moerner, M Orrit and U P Wild (Weinheim: VCH)
- [91] Geva E, Reilly P D and Skinner J L 1996 Spectral dynamics of individual molecules in glasses and crystals *Acc. Chem. Res.* **29** 579–84
- [92] Geva E and Skinner J L 1997 Theory of single-molecule optical line-shape distributions in low-temperature glasses *J. Chem. Phys. B* **101** 8920–32
- [93] Pfüegl W, Brown F L H and Silbey R J 1998 Variance and width of absorption lines of single molecules in low temperature glasses *J. Phys. Chem* **108** 6876–83
- [94] Geva E and Skinner J L 1998 Optical line shapes of single molecules in glasses: temperature and scan-time dependence *J. Phys. Chem* **109** 4920–6
- [95] Zumofen G and Klafter J 1994 Spectral random walk of a single molecule *Chem. Phys. Lett.* **219** 303–9

- [96] Wild U P, Güttler F, Pirota M and Renn A 1992 Single molecule spectroscopy: stark effect of pentacene in *p*-terphenyl *Chem. Phys. Lett.* **193** 451–5
- [97] Orrit M, Bernard J, Zumbusch A and Personov R I 1992 Stark effect on single molecules in a polymer matrix *Chem. Phys. Lett.* **196** 595–600
- [98] Croci M, Müschenborn H-J, Güttler F, Renn A and Wild U P 1993 Single molecule spectroscopy: pressure effect on pentacene in *p*-terphenyl *Chem. Phys. Lett.* **212** 71–7
- [99] Müller A, Richter W and Kador L 1995 Pressure effects on single molecules of terylene in *p*-terphenyl *Chem. Phys. Lett.* **241** 547–54
- [100] Pirota M, Güttler F, Gyax H, Renn A, Sepiol J and Wild U P 1993 Single molecule spectroscopy: fluorescence lifetime measurements of pentacene in *p*-terphenyl *Chem. Phys. Lett.* **208** 379–84
- [101] Güttler F, Sepiol J, Plakhotnik T, Mitterdorfer A, Renn A and Wild U P 1993 Single molecule spectroscopy: fluorescence excitation spectra with polarized light *J. Lumin.* **56** 29–38
- [102] Güttler F, Croci M, Renn A and Wild U P 1996 Single molecule polarization spectroscopy: pentacene in *p*-terphenyl *Chem. Phys.* **211** 421–30
- [103] Tchénio P, Myers A B and Moerner W E 1993 Dispersed fluorescence spectra of single molecules of pentacene in *p*-terphenyl *J. Chem. Phys.* **97** 2491–3

-31-

- [104] Fleury L, Tamarat P, Lounis B, Bernard J and Orrit M 1995 Fluorescence spectra of single pentacene molecules in *p*-terphenyl at 1.7 K *Chem. Phys. Lett.* **236** 87–95
- [105] Tchénio P, Myers A B and Moerner W E 1993 Vibrational analysis of the dispersed fluorescence from single molecules of terylene in polyethylene *Chem. Phys. Lett.* **213** 325–32
- [106] Myers A B, Tchénio P and Moerner W E 1994 Vibronic spectroscopy of single molecules: exploring electronic–vibrational frequency correlations within an inhomogeneous distribution *J. Lumin.* **58** 161–7
- [107] Myers A B, Tchénio P, Zgierski M Z and Moerner W E 1994 Vibronic spectroscopy of individual molecules in solids *J. Chem. Phys.* **98** 10 377–90
- [108] Fleury L, Tamarat P, Kozankiewicz B, Orrit M, Lapouyade R and Bernard J 1996 Single-molecule spectra of an impurity found in n-hexadecane and polyethylene *Mol. Cryst. Liq. Cryst.* **283** 81–7
- [109] Kulzer F, Kummer S, Matzke R, Bräuchle C and Basché T 1997 Single-molecule optical switching of terylene in *p*-terphenyl *Nature* **387** 688–91
- [110] Bach H, Renn A and Wild U P 1997 Excitation-induced frequency shifts of single molecules *Chem. Phys. Lett.* **266** 317–22
- [111] Bach H, Renn A, Zumofen G and Wild U P 1999 Exciton dynamics probed by single molecule spectroscopy *Phys. Rev. Lett.* **82** 2195–8
- [112] Basché T, Moerner W E, Orrit M and Talon H 1992 Photon antibunching in the fluorescence of a single dye molecule trapped in a solid *Phys. Rev. Lett.* **69** 1516–19
- [113] Bernard J, Fleury L, Talon H and Orrit M 1993 Photon bunching in the fluorescence from single molecules: a probe for intersystem crossing *J. Phys. Chem* **98** 850–9
- [114] Norris D J, Kuwata-Gonokami M and Moerner W E 1997 Excitation of a single molecule on the surface of a spherical microcavity *Appl. Phys. Lett.* **71** 297–9
- [115] Lounis B, Jelezko F and Orrit M 1997 Single molecules driven by strong resonant fields: hyper-Raman and subharmonic resonances *Phys. Rev. Lett.* **78** 3673–6
- [116] Brunel C, Lounis B, Tamarat P and Orrit M 1998 Rabi resonances of a single molecule driven by rf and laser fields *Phys. Rev. Lett.* **81** 2679–82
- [117] Tamarat P, Lounis B, Bernard J, Orrit M, Kummer S, Kettner R, Mais S and Basché T 1995 Pump-probe experiments with a single molecule: ac-Stark effect and nonlinear optical response *Phys. Rev. Lett.* **75** 1514–17
- [118] Kummer S, Mais S and Basché T 1995 Measurement of optical dephasing of a single terylene molecule with nanosecond time resolution *J. Chem. Phys.* **99** 17 078–81
- [119] Brunel C, Lounis B, Tamarat P and Orrit M 1999 Triggered source of single photons based on controlled single molecule fluorescence *Phys. Rev. Lett.* **83** 2722–5
- [120] Kulzer F, Koberling F, Christ T, Mews A and Basché T 1999 Terylene in *p*-terphenyl: single-molecule experiments at

room temperature *Chem. Phys.* submitted

- [121] Ambrose W P, Goodwin P M, Martin J C and Keller R A 1994 Single molecule detection and photochemistry on a surface using near-field optical excitation *Phys. Rev. Lett.* **72** 160–3
 - [122] Xie X S 1996 Single-molecule spectroscopy and dynamics at room temperature *Acc. Chem. Res.* **29** 598–606
 - [123] Xie X S and Trautman J K 1998 Optical studies of single molecules at room temperature *Ann. Rev. Phys. Chem.* **49** 441–80
 - [124] Trautman J K, Macklin J J, Brus L E and Betzig E 1994 Near-field spectroscopy of single molecules at room temperature *Nature* **369** 40–2
 - [125] Trautman J K and Macklin J J 1996 Time-resolved spectroscopy of single molecules using near-field and far-field optics *Chem. Phys.* **205** 221–9
-

-32-

- [126] Macklin J J, Trautman J K, Harris T D and Brus L E 1996 Imaging and time-resolved spectroscopy of single molecules at an interface *Science* **272** 255–8
- [127] Ha T, Enderle T, Chemla D S, Selvin P R and Weiss S 1996 Single molecule dynamics studied by polarization modulation *Phys. Rev. Lett.* **77** 3979–82
- [128] Lu H P and Xie X S 1997 Single-molecule spectral fluctuations at room temperature *Nature* **385** 143–6
- [129] Yip W-T, Hu D, Yu J, Vanden Bout D A and Barbara P F 1998 Classifying the photophysical dynamics of single- and multiple-chromophoric molecules by single molecule spectroscopy *J. Chem. Phys. A* **102** 7564–75
- [130] Ha T, Enderle T, Chemla D S, Selvin P R and Weiss S 1997 Quantum jumps of single molecules at room temperature *Chem. Phys. Lett.* **271** 1–5
- [131] Dickson R M, Norris D J and Moerner W E 1998 Simultaneous imaging of individual molecules aligned both parallel and perpendicular to the optic axis *Phys. Rev. Lett.* **81** 5322–5
- [132] Köhler J, Disselhorst J A J M, Donckers M C J M, Groenen E J J, Schmidt J and Moerner W E 1993 Magnetic resonance of a single molecular spin *Nature* **363** 242–4
- [133] Wrachtrup J, von Borczyskowski C, Bernard J, Orrit M and Brown R 1993 Optical detection of magnetic resonance in a single molecule *Nature* **363** 244–5
- [134] Köhler J, Brouwer A C J, Groenen E J J and Schmidt J 1994 Fluorescence detection of single molecule magnetic resonance for pentacene in *p*-terphenyl. The hyperfine interaction of a single triplet spin with a single ¹³C nuclear spin *Chem. Phys. Lett.* **228** 47–52
- [135] Köhler J, Brouwer A C J, Groenen E J J and Schmidt J 1995 Single molecule electron paramagnetic resonance spectroscopy: hyperfine splitting owing to a single nucleus *Science* **268** 1457–60
- [136] Wrachtrup J, von Borczyskowski C, Bernard J, Orrit M and Brown R 1993 Optically detected spin coherence of single molecules *Phys. Rev. Lett.* **71** 3565–8
- [137] Wrachtrup J, von Borczyskowski C, Bernard J, Brown R and Orrit M 1995 Hahn echo experiments on a single triplet electron spin *Chem. Phys. Lett.* **245** 262–7
- [138] Lu H P and Xie X S 1997 Single-molecule kinetics of interfacial electron transfer *J. Chem. Phys. B* **101** 2753–7
- [139] Tan W and Yeung E S 1997 Monitoring the reactions of single enzyme molecules and single metal ions *Anal. Chem.* **69** 4242–8
- [140] Bopp M A, Meixner A J, Tarrach G, Zschokke-Gränacher I and Novotny L 1996 Direct imaging single molecule diffusion in a solid polymer host *Chem. Phys. Lett.* **263** 721–6
- [141] Ruiter A G T, Veerman J A, Garcia-Parajo M F and van Hulst N F 1997 Single molecule rotational and translational diffusion observed by near-field scanning optical microscopy *J. Chem. Phys. A* **101** 7318–23
- [142] Dickson R M, Norris D J, Tzeng Y-L and Moerner W E 1996 Three-dimensional imaging of single molecules solvated in pores of poly(acrylamide) gels *Science* **274** 966–9
- [143] Nie S 1997 Confinement and detection of single molecules in submicrometer channels *Anal. Chem.* **69** 3400–5
- [144] Schmidt T, Schütz G J, Baumgartner W, Gruber H J and Schindler H 1995 Characterization of photophysics and mobility of single molecules in a fluid lipid membrane *J. Chem. Phys.* **99** 17 662–8
- [145] Xu X-H and Yeung E S 1997 Direct measurement of single-molecule diffusion and photodecomposition in free solution *Science* **275** 1106–9

- [146] Osborne M A, Balasubramanian S, Furey W S and Klenerman D 1998 Optically biased diffusion of single molecules studied by confocal fluorescence microscopy *J. Chem. Phys. B* **102** 3160–7
- [147] Chiu D T and Zare R N 1996 Biased diffusion, optical trapping and manipulation of single molecules in solution *J. Am. Chem. Soc.* **118** 6512–13
-

-33-

- [148] Bartko A P and Dickson R M 1999 Three-dimensional orientations of polymer-bound single molecules *J. Chem. Phys. B* **103** 3053–6
- [149] Bartko A P and Dickson R M 1999 Imaging three-dimensional single molecule orientations *J. Chem. Phys. B* **103** 11 237–41
- [150] Ha T, Glass J, Enderle T, Chemla D S and Weiss S 1998 Hindered rotational diffusion and rotational jumps of single molecules *Phys. Rev. Lett.* **80** 2093–7
- [151] Ha T, Laurence T A, Chemla D S and Weiss S 1999 Polarization spectroscopy of single fluorescent molecules *J. Chem. Phys. B* **103** 6839–50
- [152] Vandebout D A, Yip W T, Hu D H, Fu D K, Swager T M and Barbara P F 1997 *Science* **277** 1074–7
- [153] Hu D, Yu J and Barbara P F 1999 Single-molecule spectroscopy of the conjugated polymer MEH-PPV *J. Am. Chem. Soc.* **121** 6936–7
- [154] Higgins D A and Barbara P F 1995 Excitonic transitions in J-aggregates probed by near-field scanning optical microscopy *J. Chem. Phys.* **99** 3–7
- [155] Higgins D A, Reid P J and Barbara P F 1996 Structure and exciton dynamics in J-aggregates studied by polarization-dependent near-field scanning optical microscopy *J. Chem. Phys.* **100** 1174–80
- [156] Vanden Bout D A, Kerimo J, Higgins D A and Barbara P F 1997 Near-field optical studies of thin-film mesostructured organic materials *Acc. Chem. Res.* **30** 204–12
- [157] Vanden Bout D A, Kerimo J, Higgins D A and Barbara P F 1996 Spatially resolved spectral inhomogeneities in small molecular crystals studied by near-field scanning optical microscopy *J. Chem. Phys.* **100** 11 843–9
- [158] Higgins D A, Vanden Bout D A, Kerimo J and Barbara P F 1996 Polarization-modulation near-field scanning optical microscopy of mesostructured materials *J. Chem. Phys.* **100** 13 794–803
- [159] Wang H, Bardo A M, Collinson M M and Higgins D A 1998 Microheterogeneity in dye-doped silicate and polymer films *J. Chem. Phys. B* **102** 7231–7
- [160] Ishikawa M, Ye J Y, Maruyama Y and Nakatsuka H 1999 Triphenylmethane dyes revealing heterogeneity of their nanoenvironment: femtosecond, picosecond, and single-molecule studies *J. Chem. Phys. A* **103** 4319–31
- [161] Emory S R, Haskins W E and Nie S 1998 Direct observation of size-dependent optical enhancement in single metal nanoparticles *J. Am. Chem. Soc.* **120** 8009–10
- [162] Krug J T II, Wang G D, Emory S R and Nie S 1999 Efficient Raman enhancement and intermittent light emission observed in single gold nanocrystals *J. Am. Chem. Soc.* **121** 9208–14
- [163] Tittel J, Göhde W, Koberling F, Basché T, Kornowski A, Weller H and Eychmüller A 1997 Fluorescence spectroscopy on single CdS nanocrystals *J. Chem. Phys. B* **101** 3013–16
- [164] Empedocles S A, Norris D J and Bawendi M G 1996 Photoluminescence spectroscopy of single CdSe nanocrystallite quantum dots *Phys. Rev. Lett.* **77** 3873–6
- [165] Koberling F, Mews A and Basché T 1999 Single dot spectroscopy of CdS nanocrystals and CdS/HgS heterostructures *Phys. Rev. B* **60** 1921–7
- [166] Empedocles S A and Bawendi M G 1997 Quantum-confined Stark effect in single CdSe nanocrystalline quantum dots *Science* **278** 2114–17
- [167] Blanton S A, Hines M A and Guyot-Sionnest P 1996 Photoluminescence wandering in single CdSe nanocrystals *Appl. Phys. Lett.* **69** 3905–7
- [168] Mason M D, Credo G M, Weston K D and Buratto S K 1998 Luminescence of individual porous Si chromophores *Phys. Rev. Lett.* **80** 5405–8
- [169] Weiss S 1999 Fluorescence spectroscopy of single biomolecules *Science* **283** 1676–83
-

-34-

- [170] Dunn R C, Holtom G R, Mets L and Xie X S 1994 Near-field fluorescence imaging and fluorescence lifetime measurement of light harvesting complexes in intact photosynthetic membranes *J. Chem. Phys.* **98** 3094–8
- [171] van Oijen A M, Ketelaars M, Köhler J, Aartsma T J and Schmidt J 1998 Spectroscopy of single light-harvesting complexes from purple photosynthetic bacteria at 1.2 K *J. Chem. Phys. A* **102** 9363–6
- [172] van Oijen A M, Ketelaars M, Köhler J, Aartsma T J and Schmidt J 1999 Unraveling the electronic structure of individual photosynthetic pigment-protein complexes *Science* **285** 400–2
- [173] Tietz C, Chekhlov O, Dräbenstedt A, Schuster J and Wrachtrup J 1999 Spectroscopy on single light-harvesting complexes at low temperature *J. Chem. Phys. B* **103** 6328–33
- [174] Bopp M A, Jia Y, Li L, Cogdell R J and Hochstrasser R M 1997 Fluorescence and photobleaching dynamics of single light-harvesting complexes *Proc. Natl Acad. Sci. USA* **94** 10 630–5
- [175] Dickson R M, Cubitt A B, Tsien R Y and Moerner W E 1997 On/off blinking and switching behaviour of single molecules of green fluorescent protein *Nature* **388** 355–8
- [176] Pierce D W, Hom-Booher N and Vale R D 1997 Imaging individual green fluorescent proteins *Nature* **388** 338
- [177] Williams J G K 1997 Single-molecule detection of specific nucleic acid sequences in unamplified genomic DNA *Anal. Chem.* **69** 3915–20
- [178] Haab B B and Mathies R A 1999 Single-molecule detection of DNA separations in microfabricated capillary electrophoresis chips employing focused molecular streams *Anal. Chem.* **71** 5137–45
- [179] Schmidt T 1998 Detection of individual oligonucleotide pairing by single-molecule microscopy *Anal. Chem.* **71** 279–83
- [180] Sauer M, Drexhage K H, Lieberwirth U, Müller R, Nord S and Zander C 1998 Dynamics of the electron transfer reaction between an oxazine dye and DNA oligonucleotides motored on the single-molecule level *Chem. Phys. Lett.* **284** 153–63
- [181] Ha T, Zhuang X, Kim H D, Orr J W, Williamson J R and Chu S 1999 Ligand-induced conformational changes observed in single RNA molecules *Proc. Natl Acad. Sci. USA* **96** 9077–82
- [182] Edman L, Mets Ü and Rigler R 1996 Conformational transitions monitored for single molecules in solution *Proc. Natl Acad. Sci. USA* **93** 6710–15
- [183] Wennmalm S, Edman L and Rigler R 1997 Conformational fluctuations in single DNA molecules *Proc. Natl Acad. Sci. USA* **94** 10 641–6
- [184] Eggeling C, Fries J R, Brand L, Günther R and Seidel C A M 1998 Monitoring conformational dynamics of a single molecule by selective fluorescence spectroscopy *Proc. Natl Acad. Sci. USA* **95** 1556–61
- [185] Deniz A A, Dahan M, Grunwell J R, Ha T, Faulhaber A E, Chemla D S, Weiss S and Schultz P G 1999 Single-pair fluorescence resonance energy transfer on freely diffusing molecules: observation of Förster distance dependence and subpopulations *Proc. Natl Acad. Sci. USA* **96** 3670–5
- [186] Nie S 1998 A dual-beam optical microscope for observation and cleavage of single DNA molecules *Anal. Chem.* **70** 1743–8
- [187] Ha T, Ting A Y, Liang J, Caldwell W B, Deniz A A, Chemla D S, Schultz P G and Weiss S 1999 Single-molecule fluorescence spectroscopy of enzyme conformational dynamics and cleavage mechanism *Proc. Natl Acad. Sci. USA* **96** 893–8
- [188] Warshaw D M, Hayes E, Gaffney D, Lauzon A-M, Wu J, Kennedy G, Trybus K, Lowey S and Berger C 1998 Myosin conformational states determined by single fluorophore polarization *Proc. Natl Acad. Sci. USA* **95** 8034–9
- [189] Vale R D, Funatsu T, Pierce D W, Romberg L, Harada Y and Yanagida T 1996 Direct observation of single kinesin molecules moving along microtubules *Nature* **380** 451–3
- [190] Talley C E and Dunn R C 1999 Single molecules as probes of lipid membrane microenvironments *J. Chem. Phys. B* **103** 10 214–20
- [191] Hollars C W and Dunn R C 2000 Probing single molecule orientations in model lipid membranes with near-field scanning optical microscopy *J. Phys. Chem* **112** 7822–30

- [192] Lu H P, Xun L and Xie X S 1998 Single-molecule enzymatic dynamics *Science* **282** 1877–82
- [193] Xie X S and Lu H P 1999 Single-molecule enzymology *J. Biol. Chem.* **274** 15 967–70

- [194] Tokunaga M, Kitamura K, Saito K, Iwane A H and Yanagida T 1997 Single molecule imaging of fluorophores and enzymatic reactions achieved by objective-type total internal reflection fluorescence microscopy *Biochem. Biophys. Res. Commun.* **235** 47–53
- [195] Xue Q and Yeung E S 1995 Differences in the chemical reactivity of individual molecules of an enzyme *Nature* **373** 681–3
- [196] Craig D B, Arriaga E A, Wong J C Y, Lu H and Dovichi N J 1996 Studies on single alkaline phosphatase molecules: reaction rate and activation energy of a reaction catalyzed by a single molecule and the effect of thermal denaturation—the death of an enzyme *J. Am. Chem. Soc.* **118** 5245–53
- [197] Polakowski R, Craig D B, Skelley A and Dovichi N J 2000 Single molecules of highly purified bacterial alkaline phosphatase have identical activity *J. Am. Chem. Soc.* **122** 4853–5
- [198] Moerner W E 1994 Examining nanoenvironments in solids on the scale of a single, isolated impurity molecule *Science* **265** 46–53
- [199] Tchénio P, Myers A B and Moerner W E 1993 Optical studies of single terrylene molecules in polyethylene *J. Lumin.* **56** 1–14
-

C2.1 Polymers

Pierre Robyr

C2.1.1 INTRODUCTION

Polymers are substances consisting of large molecules also known as macromolecules. The molecules are built up of many subunits called monomers which are linked together, usually by covalent bonds. In a polymer, the number of subunits is generally larger than 100 [1]. Assemblies of less than 100 subunits are often referred to as oligomers. Macromolecules make up many of the materials in living organisms, as for example cellulose, lignin, proteins and nucleic acids. The latter two have highly specific roles in life. Proteins control many biochemical processes and nucleic acids store genetic information. Many polymers are man-made materials, and are therefore called synthetic polymers. These polymers have a great industrial importance because they offer an attractive compromise between ease of processability and final mechanical and thermal properties. This article focuses on the general properties of polymers, without dealing with the specific roles of natural polymers, such as proteins and nucleic acids.

In contrast to low-molecular-weight compounds and polymers with specific roles in biochemical processes, most polymers consist of similar molecules with different molecular weights. The mean value and the distribution of the molecular weight depend on the preparation conditions and decisively influence the material properties. Both quantities can be obtained from different techniques such as light scattering, viscosity measurements or gel permeation chromatography. However, each technique provides a different average molecular weight [2]. The most important ones are the *number* average molecular weight

$$M_n = \frac{\sum_i n_i M_i}{\sum_i n_i}$$

where n_i is the number of molecules with molecular weight M_i , and the *weight* average molecular weight:

$$M_w = \frac{\sum_i n_i M_i M_i}{\sum_i n_i M_i}$$

The ratio M_w / M_n is one for a strictly uniform molecular-weight distribution and larger than one for molecular-weight distributions with finite widths. The variance of the molecular weight distribution is $(M_w/M_n - 1)M_n^2$ [3], therefore $M_w / M_n - 1$ is often mentioned as a measure for molecular-weight dispersion.

In polymers made of dis-symmetric monomers, such as, for example, poly(propylene), the structure may be irregular and constitutional isomerism can occur as shown in [figure C2.1.1\(a\)](#). The succession of the relative configurations of the asymmetric centres can also vary between stretches of the chain. Configuration isomerism is characterized by the succession of dyads which are named either *meso*, if the two asymmetric centres have the same relative configurations, or *racemo* if the configurations differ ([figure C2.1.1\(b\)](#)). A polymer is called isotactic if it contains only one type of dyad and syndiotactic if the dyad sequence strictly alternates between the *meso* and *racemo* forms.

Polymers without configurational regularity are called atactic. Configurationally regular polymers can form crystalline structures, while atactic polymers are almost always amorphous. Many polymers consist of linear molecules, however, nonlinear chain architectures are also important ([figure C2.1.2](#)).

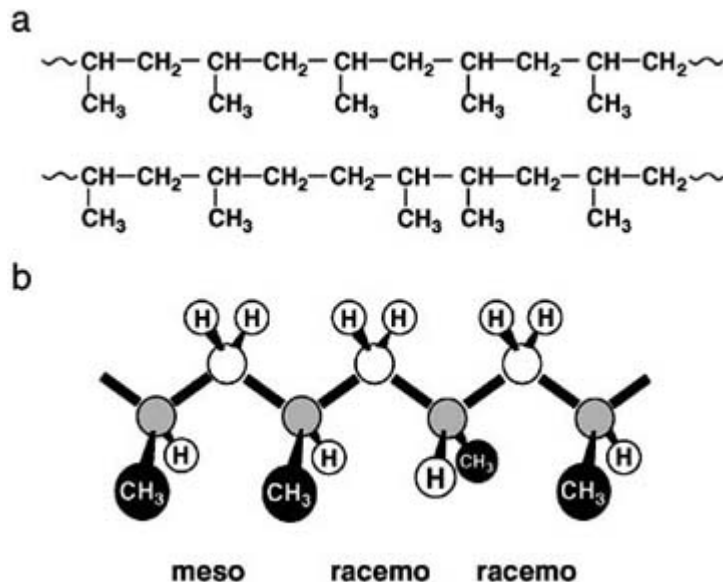


Figure C2.1.1. (a) Constitutional isomerism of poly (propylene). The upper chain has a regular constitution. The lower one contains a constitutional defect. (b) Configurational isomerism of poly(propylene). Depending on the relative configurations of the asymmetric carbons of two successive monomer units, the corresponding dyad is either *meso* or *racemo*.

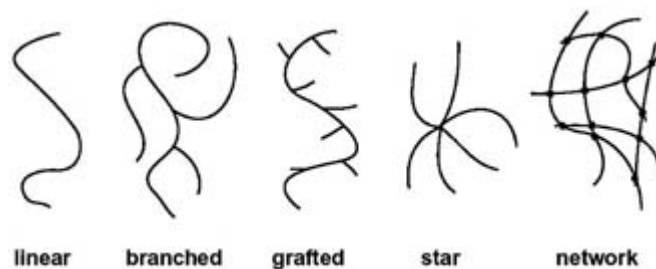
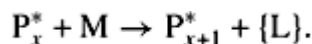


Figure C2.1.2. Polymers with linear and nonlinear chain architectures. The nonlinear polymers can have *branched* chains. Short chains of oligomers can be *grafted* to the main chain. The chains may form a *star*-like structure. The chains can be cross-linked and form a *network*.

Most properties of linear polymers are controlled by two different factors. The chemical constitution of the monomers determines the interaction strength between the chains, the interactions of the polymer with host molecules or with interfaces. The monomer structure also determines the possible local conformations of the polymer chain. This relationship between the molecular structure and any interaction with surrounding molecules is similar to that found for low-molecular-weight compounds. The second important parameter that controls polymer properties is the molecular weight. Contrary to the situation for low-molecular-weight compounds, it plays a fundamental role in polymer behaviour. It determines the slow-mode dynamics and the viscosity of polymers in solutions and in the melt. These properties are of utmost importance in polymer rheology and condition their processability. The mechanical properties, solubility and miscibility of different polymers also depend on their molecular weights.

The successful preparation of polymers is achieved only if the macromolecules are stable. Polymers are often prepared in solution where entropy destabilizes large molecular assemblies. Therefore, monomers have to be strongly bonded together. These links are best realized by covalent bonds. Moreover, reaction kinetics favourable to polymeric materials must be fast, so that high-molecular-weight materials can be produced in a reasonable time. The polymerization reaction must also be fast compared to side reactions that often hinder or preclude the formation of the desired product.

Polymerization reactions are generally divided into two main categories according to the mechanism of chain growth [4]. In the first category, called chain polymerization, chain growth proceeds exclusively by reaction between a monomer and the reactive site on the polymer chain with regeneration of the reactive site at the end of each growth step and possible production of a side-product L:



Chain polymerization involves at least two steps: initiation, i.e. formation of reactive sites, which are generally radicals or ions; and propagation, through which the chains grow. In most practical conditions termination reactions play an important role. Depropagation, i.e. the reduction of the degree of polymerization by one unit is also possible. In termination reactions, one or two reactive sites vanish and chain growth is stopped. The possibility of growth termination of a chain exists at every addition step and, therefore, the probability of termination increases with the degree of polymerization. This increase leads to the flattening of the average molecular weight as a function of the degree of monomer conversion, as shown in [figure C2.1.3](#) [5]. In the limiting situation of fast initiation, irreversible propagation and the absence of termination, the molecular weight increases linearly with the degree of conversion ([figure C2.1.3](#)). These conditions are approached in some ionic chain polymerizations performed under very clean conditions. Such reactions are called living polymerizations. Radical chain polymerization and also many chain polymerizations with simple ions as reactive sites produce polymers with irregular configuration. To prepare highly-regular polymers special catalysts have been developed. Two types widely used are Ziegler--Natta catalysts [6] and metallocene-based catalysts [7].

-4-

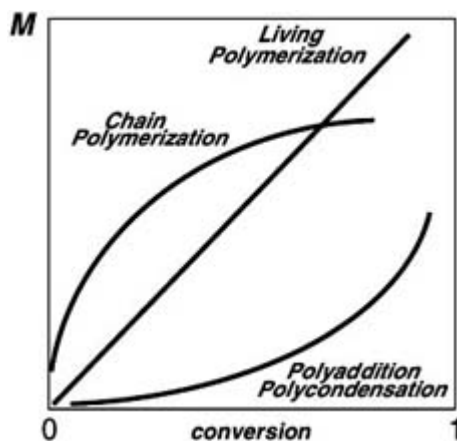
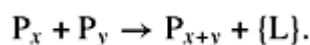


Figure C2.1.3. Schematic dependence of the molecular weight of a polymer as a function of the degree of monomer conversion for different polymerization reactions.

The second category of polymerization reactions does not involve a chain reaction and is divided into two groups: polyaddition and polycondensation [4]. In both reactions, the growth of a polymer chains proceeds by reactions between molecules of all degrees of polymerization. In polycondensations a low-molecular-weight product L is eliminated, while polyadditions occur without elimination:



In polycondensation reactions, the equilibrium can be easily shifted to the side of the higher molecular-weight compounds by allowing the eliminated compound to escape the reaction vessel or by actively removing it. Since molecules of all degrees of polymerization can react with each other, the average molecular weight grows faster with a higher degree of conversion (figure C2.1.3). Consequently, the preparation of polymers with high molecular weights requires a large degree of conversion [5].

Copolymerization involves the reaction of at least two different monomers A and B. In the case of chain copolymerization, the reactivity ratios r_A and r_B are important, $r_A = k_{AA}/k_{AB}$ and $r_B = k_{BB}/k_{BA}$, where k_{YX} is the rate constant of the reaction of a free monomer Y with a chain ending with a unit X [5]. Three different situations can be envisaged. (i) If $r_A r_B = 1$, $k_{AA}/k_{AB} = k_{BA}/k_{BB}$. The ratio of the probabilities of adding a monomer A or B to a chain end A is equal to the ratio of the probabilities of adding a monomer A or B to the chain end B. Consequently, the probabilities of adding a monomer A or B are independent of the chain end. This type of copolymerization is called ideal or Bernoullian copolymerization. (ii) If $r_A r_B < 1$, the sequence of the monomers tends to alternate. In the limiting case where $r_A = r_B = 0$, the monomer sequence strictly alternates. (iii) If $r_A r_B > 1$, the two types of monomers tend to sequence and build block structures along the chain.

We have tacitly assumed that the rate constants depend only on the last unit of the chain. In such a situation, the copolymerization is called a Markov copolymerization of first order. The special case (i), $r_A r_B = 1$, is a Markov copolymerization of order zero. If reactivity also depends on the penultimate unit of the chain, the polymerization is a Markov copolymerization of second order.

C2.1.3 CONFORMATION OF A SINGLE CHAIN

C2.1.3.1 INTRODUCTION

Polymer chains possess a huge number of degrees of freedom, which can be divided into two categories: on the one hand, bond angles and bond lengths and, on the other hand, torsional angles. Those of the first category undergo fast oscillations on the 10–100 fs time scale and vary little from one monomer to the other. Torsional angles are much softer degrees of freedom and set the conformation of the polymer chain. Two conformations are often encountered: random coils and helices. Random coils are found in polymer solutions, melts, and polymer glasses. Denatured proteins also adopt random coil structures. Helical structures often occur in crystalline polymer or as subunits in folded proteins.

The energy difference between the *trans* and *gauche* conformations in alkanes was investigated with Raman scattering [9]. It was found that the difference between the two energy minima in solution is 2–3 kJ mol⁻¹ and that the energy barrier is about 12 kJ mol⁻¹ (figure C2.1.4). Consequently, at room temperature the torsional angles of an alkane chain are most of the time either in the *gauche* or the *trans* state with roughly equal probabilities, and from time to time a chain segment collects sufficient thermal energy so that a conformational change can occur.

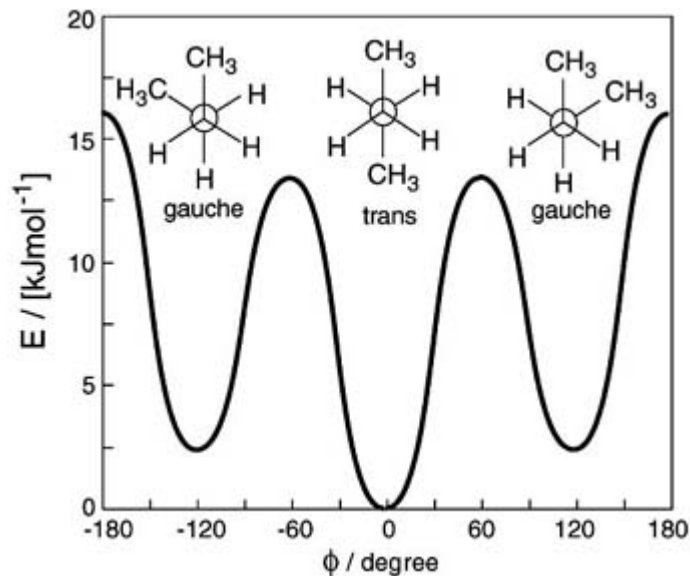


Figure C2.1.4. Potential energy as a function of the rotation about the central C–C bond in butane. The sketches show the projection of the molecule along the central C–C bond.

-6-

C2.1.3.2 GAUSSIAN CHAINS

At first sight, one might think that any treatment of the properties of a polymer chain has to emanate from its microscopic chemical structure since it determines the populations of the different rotational isomeric states of a given torsional angle. However, in polymers without well defined conformations, the correlation between the torsional angles along the chain decays rapidly. Beyond that length scale, typically a few nanometres and referred to as persistence length [10], one can think of the chain as a succession of jointed sticks, with unrestricted angles at the junctions. Such a chain with a large number of segments is called Gaussian. It can be easily shown [11] that for a chain consisting of N_s segments of length a_s the root mean square of the distance between the two chain ends averaged over all possible conformations is

$$\sqrt{\langle R^2 \rangle} = N_s^{1/2} a_s. \quad (\text{C2.1.1})$$

This relation also applies to any portion of the chain segments as long as the number of segments in the portion is sufficient. Therefore, if one proceeds n_s segmental steps, starting from a point in the interior of the chain, the resulting average displacement is of the order of $\sqrt{n_s} a_s$. Conversely, the number of monomers contained in a sphere of radius r scales as $n \propto n_s \propto r^2$. Thus, the Gaussian chain fills only partially a three-dimensional space and its fractal dimension is two. The mean monomer density c_m in a volume of size V as a function of the degree of polymerization N scales as

$$c_m(N) \propto \frac{N}{\langle R^2 \rangle^{3/2}} \propto N^{-1/2}. \quad (\text{C2.1.2})$$

The mean monomer density decreases with the increasing degree of polymerization.

It is not possible to apply (C2.1.1) down to the level of monomers and replace N_s by the degree of polymerization N and a_s^2 by the sum of the squares of the bond lengths in the monomer a_b^2 , because the chemical constitution imposes some stiffness to the chain on the length scale of a few monomer units. This effect is accounted for by introducing the characteristic ratio C_∞ defined as $C_\infty = \langle R^2 \rangle / (N a_b^2)$. The characteristic ratio can be determined from viscosity or scattering measurements.

Light scattering techniques play an important role in polymer characterization. In very dilute solution, where the polymer chains are isolated from one another, the inverse of the scattering function $S(q)$ can be expressed in the limit of vanishing scattering vector $q \rightarrow 0$ as [12]

$$\frac{1}{S(q)} = N^{-1} \left(1 + q^2 \frac{\langle R^2 \rangle}{18} + \dots \right). \quad (\text{C2.1.3})$$

Thus, by plotting S^{-1} as a function of q^2 in the limit of small q the mean square of the end-to-end distance can be obtained. For large value of q , $q^2 \langle R^2 \rangle \gg 1$,

-7-

$$S(q)q^2 = \frac{12N}{\langle R^2 \rangle} = \frac{12}{C_\infty a_b^2} \quad (\text{C2.1.4})$$

so that the characteristic ratio can be evaluated from the plateau value of $S(q)q^2 a_b^2$ at large q [12].

C2.1.3.3 EXCLUDED-VOLUME EFFECTS

The Gaussian chain model considers only the interactions between neighbouring monomers along the chains, which determine the characteristic ratio, but neglects the fact that two monomers distant along the chain by more than the persistence length cannot occupy the same volume. At first sight this shortcoming might seem intolerable. However, exactly this situation, where excluded-volume effects between distant monomers vanish, holds in poor solvents and in the melt. In a poor solvent, the interactions between the solvent molecules and the monomers are not favourable and the chain tends to contract onto itself. Solutions in which the poorness of the solvent exactly compensates for the excluded-volume effects are called theta solutions, or solutions at Θ conditions [13, 14].

Theta conditions in dilute polymer solutions are similar to the state of van der Waals gases near the Boyle temperature. At this temperature, excluded-volume effects and van der Waals attraction compensate each other, so that the second virial coefficient of the expansion of the pressure as a function of the concentration vanishes. On dealing with solutions, the quantity of interest becomes the osmotic pressure Π rather than the pressure. Its virial expansion may be written as

$$\Pi = RTc_w \left(\frac{1}{M_n} + \tilde{A}_2 c_w + \tilde{A}_3 c_w^2 + \dots \right) \quad (\text{C2.1.5})$$

where c_w is the weight concentration of the polymer. Since the interactions between the monomers and the solvent are temperature dependent, the conditions at which the second virial coefficient vanishes can be found by varying the temperature. The Θ conditions of a dilute solution of polystyrene in cyclohexane occur at $T = 35^\circ\text{C}$ and ambient pressure (figure C2.1.5).

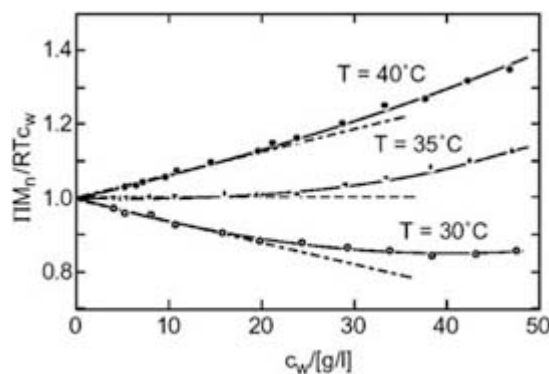


Figure C2.1.5. Reduced osmotic pressure $\Pi M_n / (RTc_w)$ as a function of the weight concentration c_w of polystyrene ($M_n = 130\,000\text{ g mol}^{-1}$) in cyclohexane at different temperatures. At $T = 35\text{ }^\circ\text{C}$ and ambient pressure, the solution is at the Θ conditions. (Figure from [74], reprinted by permission of EDP Sciences.)

-8-

C2.1.3.4 EXPANDED CHAINS

Polymer chains at low concentrations in good solvents adopt more expanded conformations than ideal Gaussian chains because of the excluded-volume effects. A suitable description of expanded chains in a good solvent is provided by the ‘self-avoiding random walk’ model. Flory [15] showed, using a mean field approximation, that the root mean square of the end-to-end distance of an expanded chain scales as

$$\sqrt{\langle R^2 \rangle} = N_s^{3/5} a_s = C_\infty^{1/2} N^{3/5} a_F \quad (\text{C2.1.6})$$

where $a_F = a_b$ in the Θ state. These results were later put on a firmer theoretical basis by de Gennes [16]. Using the same arguments as in the case of Gaussian chains, the fractal dimension of expanded chain is found to be five-thirds and the scaling behaviour of the scattering function for $q\sqrt{\langle R^2 \rangle} \gg 1$ is

$$S(q) \propto \frac{1}{q^{5/3}} \quad (\text{C2.1.7})$$

compared to $S(q) \propto q^{-2}$ in the case of Gaussian chains (equation (C2.1.4))

The neutron scattering data of figure C2.1.6 show that if the excluded volume effects are activated by increasing the temperature from the Θ point, and thus increasing the goodness of the solvent, the transition from the Gaussian chain behaviour into that of an expanded chain depends on the length scale. The transition from the regime $S^{-1} \propto q^2$ to $S^{-1} \propto q^{5/3}$ occurs first for low values of q , i.e. for long distances, while below a temperature-dependent length, referred to as the thermic correlation length, the chain is still Gaussian. The whole chain may be seen as a self-avoiding walk of Gaussian blobs (figure C2.1.7).

-9-

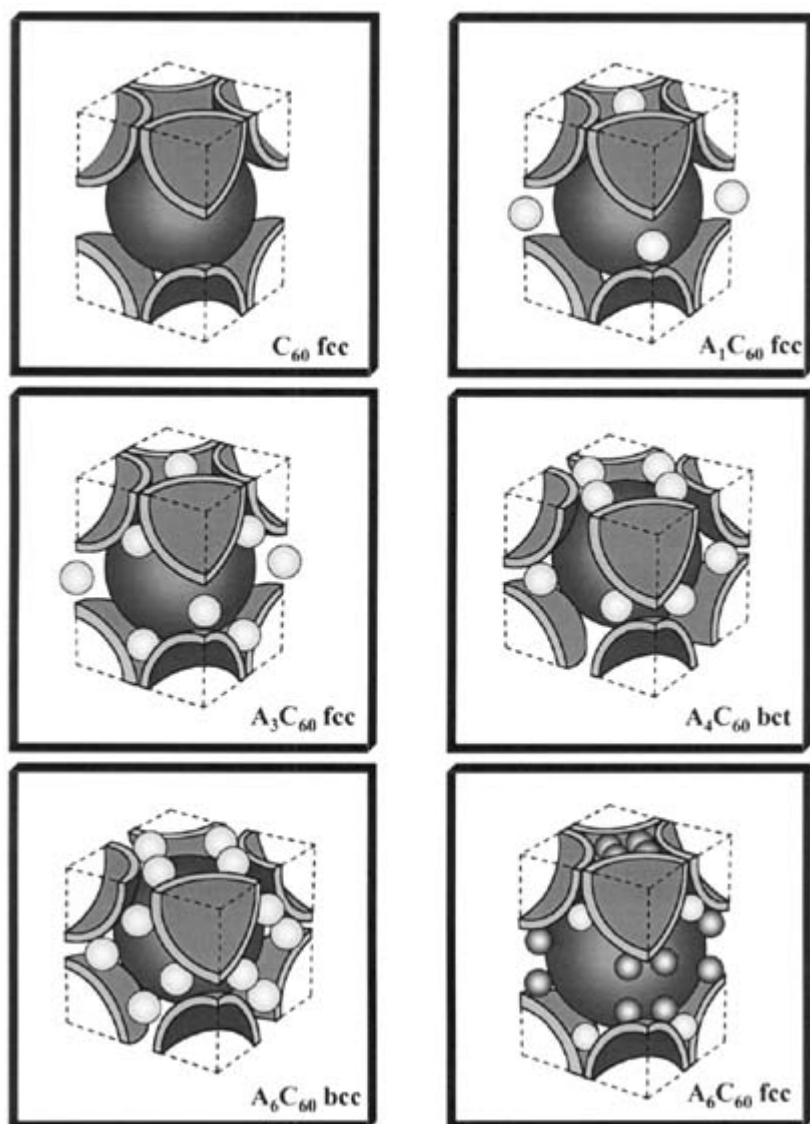


Figure C1.2.6. Summary of fcc [60] fullerene structure and alkali-intercalation composites of [60] fullerene.

-10-

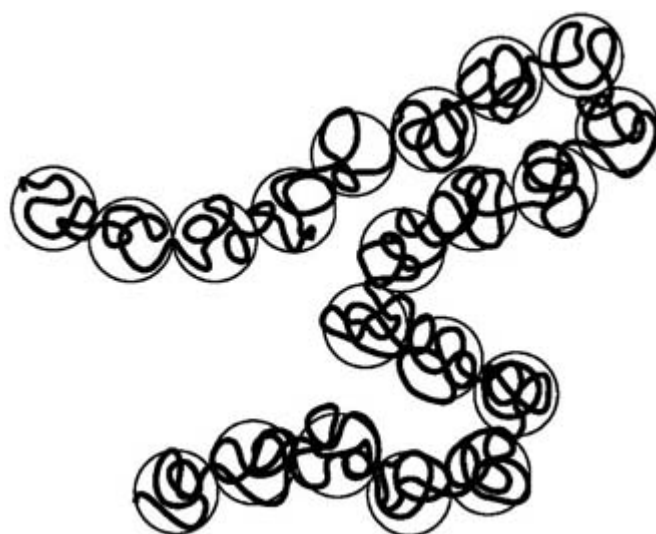


Figure C2.1.7. Schematic drawing of a polymer chain that behaves as a Gaussian chain on short length scales and

as an expanded chain on longer length scales. The cross-over distance corresponds to the size of the pearls.

C2.1.3.5 ROTATIONAL ISOMERIC STATE MODELS

The quantitative description of many properties of polymer chains, such as energy, entropy and detailed conformation, necessitates the consideration of the monomer structure and its influence on the possible states of the torsional angles. The rotational isomeric state (RIS) theory focuses on such detailed descriptions of a polymer chain [9, 17, 18]. The RIS theory takes into account only the interactions between nearest-neighbour monomers and assumes that each torsional angle can take only a few possible conformations. The nearest-neighbour interactions control the populations of the few allowed conformations. Using these approximations, the statistical weight of a given chain conformation can be calculated from the sum of the pair interactions; by summing over all the conformations, the partition function of a polymer chain can be obtained. From the partition function and the conformational Helmholtz energy, estimates of the conformational entropy and internal energy can be calculated. The RIS theory also provides a simple way to calculate quantities such as the characteristic ratios C_∞ , the mean square end-to-end distance $\langle R^2 \rangle$ and the scattering curves.

C2.1.3.6 POLYELECTROLYTES

The polymers considered hitherto were electrically neutral molecules in neutral solvents. If the monomers contain functional groups that can dissociate into ions in polar solvents, the behaviour of the polymer chain is significantly influenced by the long-range Coulombic interactions [19]. For example, if the ionic strength of the solvent is increased by adding some salt, the repulsive interaction between the equivalent charges along the polyelectrolyte chain are reduced and the polymer shrinks. Concomitantly, the viscosity of the solution which largely depends on the overall polymer dimension, diminishes drastically. Polyelectrolytes have also attracted much attention as ionic conductors in the development of light and efficient dry batteries [20].

C2.1.4 SOLUTION, MELT AND GLASS

C2.1.4.1 DILUTE AND SEMI-DILUTE SOLUTIONS

In dilute solutions, the polymer chains are isolated from one another and only interact during brief encounters. With increasing polymer concentration, a point is reached where the chains start to overlap, this point c_m^* is referred to as the critical concentration of monomers at the overlap limit and can be approximated as $c_m^* = N/R_F^3$ with $R_F = \sqrt{\langle R^2 \rangle}$ in a good solvent. Using equation (C2.1.6), the volume fraction of the polymer at the overlap limit is $\phi^* < N^{-4/5}$ [21]. Thus, for a polymer with $N = 10^4$, the chains are isolated in solution only if the polymer volume fraction is less than about 0.001.

In discussion of the solution behaviour, the osmotic pressure is a quantity of primary interest. According to equation (C2.1.5), the osmotic pressure at very low concentrations is inversely proportional to the molecular weight of the polymer. This behaviour is indeed observed in figure C2.1.8. With increasing polymer concentration, the dependence of the osmotic pressure on the molecular weight vanishes and only the weight concentration of the polymer is relevant. In this regime the chains strongly overlap and the osmotic pressure scales as $\Pi/RT \propto c_w^{9/4}$.

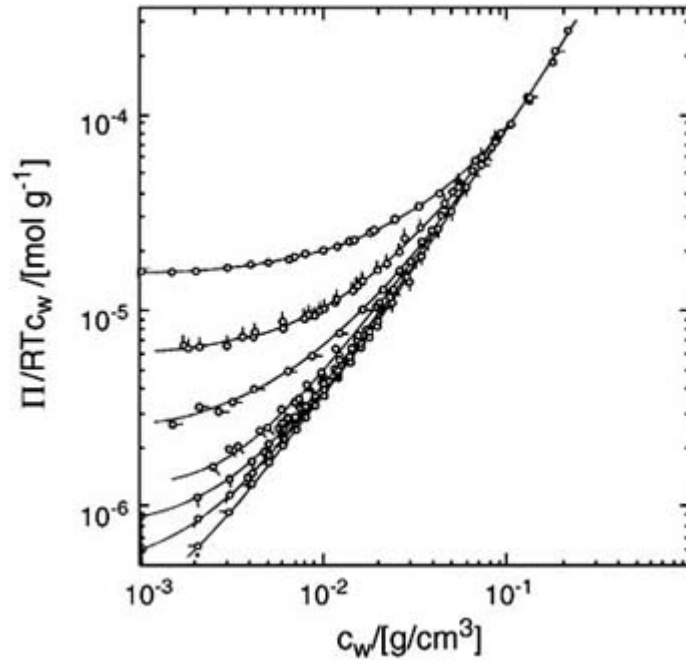


Figure C2.1.8. Reduced osmotic pressure $\Pi/(RTc_w)$ as a function of the polymer weight concentration c_w for solutions of poly(α -methylstyrene) in toluene at 25 °C. The molecular weight of poly(α -methylstyrene) varies between $M_n = 7 \times 10^4 \text{ g mol}^{-1}$ (uppermost curve) and $M_n = 7.47 \times 10^6 \text{ g mol}^{-1}$ (lowest curve). (Figure from [76], reprinted by permission of the American Chemical Society.)

-12-

The inverse scattering function of dilute polymer solutions for small scattering wavenumber ($qR_F \ll 1$) obeys the so-called Zimm relation [22]:

$$\frac{1}{S(q, c_w)} = \frac{1}{N} \left(1 + q^2 \frac{R_F^2}{18} + \dots \right) (1 + 2\tilde{A}_2 M_n c_w + \dots). \quad (\text{C2.1.8})$$

By extrapolating the measurement of $1/S(q, c_w)$ to $q \rightarrow 0$ and $c_w \rightarrow 0$, \tilde{A}_2 and R_F can be measured. The behaviour of the osmotic pressure and that of the scattering function can be satisfactorily explained in the dilute regime, the cross-over region and the semi-dilute range by using only three parameters, namely, the polymer concentration c_w , the Flory radius R_F and the thermic correlation length. Beyond the semi-dilute range, about $\phi > 0.1$, the description based on the three parameters only is no longer valid.

C2.1.4.2 MELTS AND SCREENING EFFECT

Consider the monomer-density distribution for a single chain averaged over all possible conformations. The distribution has a bell-like shape with its middle point at the gravity centre of the chain. With this picture in mind, the excluded-volume effect encountered in good solvents can be understood as an entropic force which acts on the monomers in the direction of lower concentration. Overall, these forces lead to an expansion of the polymer chain. In the melt, the monomer density is constant since the monomers of the other chains compensate for the density gradient of a particular chain. This compensation is referred to as the ‘screening effect’ and is responsible for the Gaussian behaviour of chains in the melt. These qualitative arguments [23] are expressed quantitatively in the fact that, in polymer melts, the second virial coefficient of the local osmotic pressure scales as $1/N$ and thus vanishes for long chains [24].

In summary, we see now how the change from the expanded chains in dilute solutions to the ideal chains in a melt is accomplished. With increasing polymer concentration, the chain overlap increases and the length scale over

which excluded-volume effects are screened decreases. As the screening length decreases to the thermic correlation length, all excluded-volume effects disappear. Simultaneously, the polymer shrinks to the size of a Gaussian chain.

C2.1.4.3 GLASSY STATE

Upon cooling the melt of a polymer that cannot crystallize, the system becomes glassy, i.e. hard and void of long-range order. In (figure C2.1.9(a)) the specific volume of poly (vinyl acetate) is plotted as a function of the temperature. The specific volume is measured upon heating the sample after having quenched it at $-20\text{ }^{\circ}\text{C}$. Clearly, near $T = 30\text{ }^{\circ}\text{C}$ the slope of the graphs changes. The change in slope from the glassy region to the melt is not accompanied by a discontinuity, as in the case of the fusion of crystalline materials. This type of transition, with continuous specific volume or heat capacity, is typical of glassy materials and the temperature at which the linear extrapolations of the measured quantity cross defines the glass temperature T_g . However, the exact value of the glass transition temperature depends on the rate of temperature change at which the measurement is performed. This dependence on the measurement kinetics supports the view that no real phase transition occurs, but rather that the system ‘freezes’ in a non-equilibrium state.

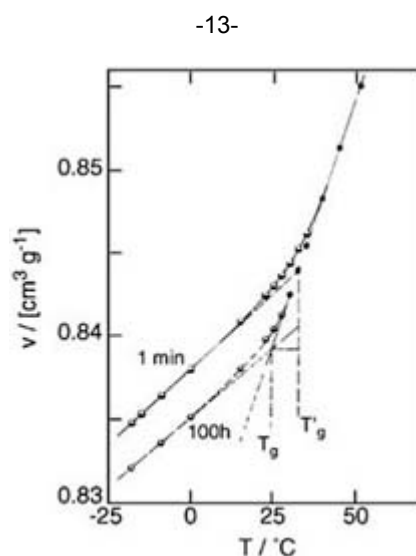


Figure C2.1.9. Specific volume of poly(vinyl acetate) as a function of the temperature measured during heating two samples which were preliminary quenched from the melt to $-20\text{ }^{\circ}\text{C}$. One sample was stored for 1 min and the other for 100 h at $-20\text{ }^{\circ}\text{C}$ before heating. (Figure from [77], reprinted by permission of John Wiley and Sons Inc).

The variation of the specific volume as a function of heating in figure C2.1.9 is plotted for two samples which were stored for different times after a quench to $-20\text{ }^{\circ}\text{C}$. The specific volume in the glassy region and the glass temperature depend on storage time. This dependence shows that, on a practical time scale, glassy polymers are not in equilibrium. A second point of interest about the structure of glassy polymers is the presence of local order. Some experimental results suggest the occurrence of more local order in the glass than that found in the melt, where the chains behave as Gaussian with hardly any specific order between each other. However, many investigations support the view that very little difference exists between the average conformation of polymer chain in the glass and that in the melt [25].

Several interpretations of the glass transition have been put forward. They can be grouped into three main categories [26]. First, some theories are based on the concept of the free volume in the form of voids as a requirement for the onset of cooperative motion. Upon cooling of the melt, the free volume is continuously squeezed out of the system. The transition into the glassy state occurs when no free-volume is left. Second, kinetic theories of the glass transition relate the onset of the segmental motion of the polymer to the transition from the glassy state to the melt. Third, thermodynamic theories are based on the extrapolation to the glass transition at infinitely long measurement times. In this hypothetical regime, T_g becomes independent of the experimental procedure and a second-order transition occurs between two phases at equilibrium. Each of the theories can predict some of the observed changes at the glass temperature. However, none can explain them all [26].

C2.1.5 THERMODYNAMICS AND PHASE TRANSITION OF POLYMER MIXTURES

Often it is difficult to achieve the desired mechanical properties with a system made of a single polymer. Polystyrene, for example, possesses a good stiffness but has a low impact strength, it is a brittle material. By mixing small amounts of polybutadiene with polystyrene, a tough material is obtained which retains satisfactory stiffness. Blending polystyrene with a small amount of polybutadiene results in a two-phase structure, in which polybutadiene droplets are embedded in the polystyrene matrix. Other binary systems with good mechanical properties, like blends of polystyrene and poly(dimethylphenylene oxide), form homogeneous structures. In order to correlate the mechanical properties with the structures of the blend, it is of primary interest to understand the formation of the different types of structures.

C2.1.5.1 FLORY–HUGGINS THEORY

The Flory–Huggins theory of polymer mixtures [15, 27] is based on two main assumptions: first, the screening of excluded-volume effects renders all polymer chains effectively Gaussian; second, the interaction between the monomers of different chains can be treated using a mean field approximation, i.e. all monomers feel, on average, the same environment. These assumptions allow the derivation of the Gibbs energy of mixing of a binary polymer system:

$$\Delta G_{\text{mix}} = RTn_c \left(\frac{\phi_A}{N_A} \ln \phi_A + \frac{\phi_B}{N_B} \ln \phi_B + \chi \phi_A \phi_B \right) \quad (\text{C2.1.9})$$

where n_c is the number of reference units, approximately equal to the number of monomers, ϕ_A and ϕ_B are the volume fractions of polymers A and B, N_A and N_B are the degrees of polymerization. The quantity χ is the Flory–Huggins parameter; it is dimensionless and determines in an empirical manner the change of the local Gibbs energy per reference unit upon mixing [15, 27]. The first two terms in the parentheses arise from the increase of the translational entropy of the centres of mass of the chains after mixing, the third from the interactions between the chain segments.

The gain of translational entropy in mixing macromolecules is small compared to that achieved in mixing low-molecular-weight compounds. This poor gain manifests itself in the occurrence of the degrees of polymerization in the denominators of the two entropy terms in equation (C2.1.9). Consequently, homogeneous mixing occurs only when χ is negative, or positive but small. In [figure C2.1.10\(a\)](#) the Gibbs energy of mixing of two polymers with the same degrees of polymerization N is plotted for different values of the product χN . If $\chi N \leq 2$, mixing occurs at any ϕ_A value. For $\chi N > 2$, the shape of ΔG_{mix} as a function of ϕ_A changes. Consider the curve for $\chi N = 2.4$ at $\phi_A = 0.4$; the system can lower its Gibbs energy by separating into two homogeneous phases with volume fractions ϕ'_A and ϕ''_A . A similar behaviour occurs for all values of χ with $\chi N > 2$. The ensemble of volume fractions resulting from the phase separation form the binodal ([figure C2.1.10\(b\)](#)) [28]. Within the binodal, the system has a two-phase structure. If the condition $N_A = N_B = N$ is released, the graphs in [figure C2.1.10](#) become asymmetric, but the essential features qualitatively remain.

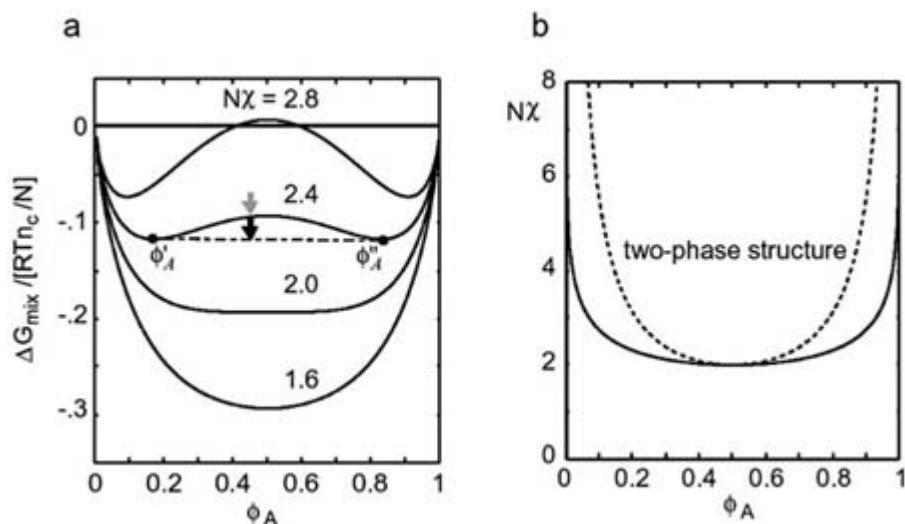


Figure C2.1.10. (a) Gibbs energy of mixing as a function of the volume fraction of polymer A for a symmetric binary polymer mixture $N_A = N_B = N$. The curves are obtained from [equation \(C2.1.9\)](#). (b) Phase diagram of a symmetric polymer mixture $N_A = N_B = N$. The full curve is the binodal and delimits the homogeneous region from that of the two-phase structure. The broken curve is the spinodal.

In many applications the phase structure as a function of the temperature is of interest. The discussion of this issue requires the knowledge of the temperature dependence of the Flory–Huggins parameter $\chi(T)$. If the interactions between the different monomers weakly depend on temperature, χ is proportional to T^{-1} because of the prefactor RT in [\(C2.1.9\)](#). In this situation, the axis in [figure C2.1.10\(b\)](#) indicates decreasing temperature and the region where the separation into a two-phase structure occurs is referred to as the ‘lower miscibility gap’. In some cases, χ increases with increasing temperature with a sign change from negative to positive. Such a temperature dependence of the Flory–Huggins parameter leads to the existence of an ‘upper miscibility gap’.

It is not always possible to describe the variations of $\Delta G_{\text{mix}}(\phi_A, T)$ with χ depending only on temperature. For some systems, the dependence of χ on the volume fraction must be considered too. However, the Flory–Huggins theory provides a very useful insight into structural behaviour of polymer mixtures.

C2.1.5.2 MECHANISMS OF PHASE SEPARATION

Phase separation of a polymer mixture is induced when the conditions change from the one-phase region into a miscibility gap. Usually, phase separation is provoked by a change in temperature. Two mechanisms of phase separation are known: ‘nucleation and growth’ and ‘spinodal decomposition’ [29]. Nucleation and growth processes start with the formation of the nuclei of a new composition. Subsequently, domains of the new composition grow from these nuclei. Systems that undergo spinodal decomposition behave differently. The sizes of the domains with the new compositions do not vary much until the latter stages of the phase separation process. In contrast, the compositions of the new phases change continuously.

The occurrence of one or the other phase-separation mechanisms can be predicted from the consideration of the change of the local Gibbs energy, δg , associated with a spontaneous fluctuation of the local composition about the composition of the homogeneous mixture ϕ_{A0} . If ϕg is positive, restoring forces will bring back the system to the homogeneous composition. In this regime, phase separation has to overcome an energy barrier, this is achieved in the nucleation step before the growth of the domains with a new composition. If ϕg is negative, no restoring force exists and the amplitude of the composition fluctuation grows so that the system separates into different phases spontaneously. Approximately, the sign of ϕg is given by that of the second derivative of $\Delta G_{\text{mix}}(\phi_A)$ at ϕ_{A0} . Therefore, the locus of the values ϕ_A with a vanishing second derivative of ΔG_{mix} delimits the region of the miscibility gap in which spinodal decomposition occurs. This locus is referred to as the spinodal ([figure C2.1.10 \(b\)](#)). The length scale of the concentration fluctuations at the beginning of the separation process is controlled by

the distance from the spinodal. If a system is brought into the region between the spinodal and the binodal, phase separation follows the nucleation and growth mechanism.

C2.1.5.3 BLOCK COPOLYMERS

In block copolymers [8, 30], long segments of different homopolymers are covalently bonded to each other. A large part of synthesized compounds are di-block copolymers, which consist only of two blocks, one of monomers A and one of monomers B. Tri- and multi-block assemblies of two types of homopolymer segments can be prepared. Systems with three types of blocks are also of interest, since in ternary systems the mechanical properties and the material functionality may be tuned separately.

Similarly to polymer mixtures, block copolymers can form an homogeneous phase but also separate into phases of different compositions. However, the presence of covalent bonds between the different blocks has important consequences on the structural arrangement after phase separation. Each of the different types of monomer segregate and almost pure domains are formed, but the domains have mesoscopic dimensions corresponding to the sizes of the blocks. Furthermore, since the block lengths usually have uniform sizes, the arrangement of the different domains are ordered. In the case of di-block copolymers the type of order depends on the ratio of the degree of polymerizations of block A and block B (figure C2.1.11). Tri- and multi-block binary systems exhibit qualitatively the same phase behaviour as di-block polymers. Changes occur for ternary systems. Their structures still exhibit periodic order, but the lattices are more complex [30].

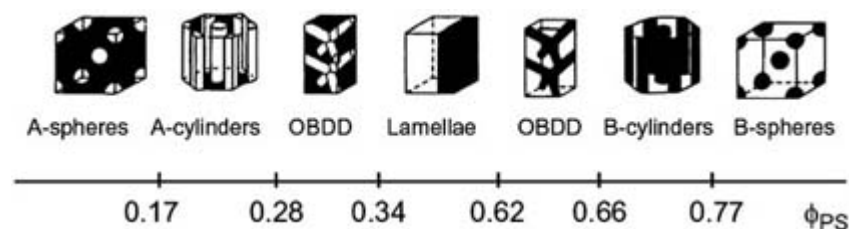


Figure C2.1.11. Morphologies of a microphase-separated di-block copolymer as function of the volume fraction of one component. The values here refer to a polystyrene–polyisoprene di-block copolymer and ϕ_{PS} is the volume fraction of the polystyrene blocks. OBDD denotes the ordered bicontinuous double diamond structure. (Figure from [78], reprinted by permission of Annual Reviews.)

C2.1.6 PARTIALLY CRYSTALLINE POLYMERS

C2.1.6.1 SEMI-CRYSTALLINE STRUCTURES

Polymers with a regular configuration can crystallize. However, because of the presence of chain entanglements, chain defects and chains with different molecular weights, only partial crystallization occurs and an amorphous part always subsists. Semi-crystalline polymers exhibit a hierarchical structure. On the molecular level, the chains are fully stretched or form regular helices which pack parallel to each other into lamellar crystallites. The chain axes are perpendicular to the lamellar plane and the lamellar thickness is of the order of 10 nm. In single-polymer crystals prepared from highly dilute solutions, the chain folds back in a regular manner and builds the adjacent helix. In most of the systems, however, the chain may also re-enter the same lamella at a distant position or leave it definitively and possibly participate in another lamella after crossing an amorphous layer. Many lamellae form larger spherical assemblies with amorphous interstices called spherulites (figure C2.1.12). These assemblies have sizes from a few micrometres up to centimetres. The directions of the polymer chains or helices are always perpendicular to the radius of the spherulite.

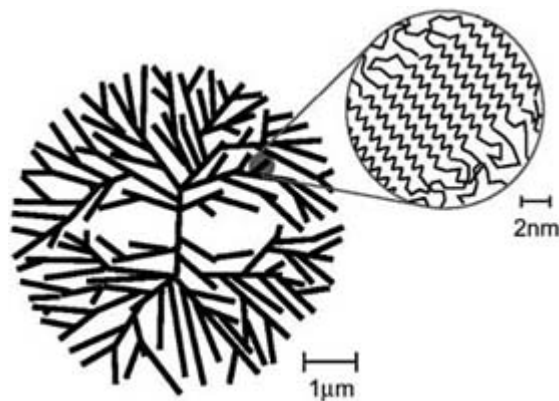


Figure C2.1.12. Schematic drawing of the cross section through a spherulite. The lines indicate the connectivity of the crystalline lamellae. The inner structure of a lamella is also shown and consists of parallel polymer chains with their axes perpendicular to the spherulite radius.

The content of crystalline material in a system can be characterized by several techniques. Density measurements provide the volume fraction occupied by the crystallites. Measuring the heat of fusion provides the weight fraction of the crystalline domains. More details on the structure of semi-crystalline polymers can be obtained from electron microscopy, scattering techniques, Raman spectroscopy and NMR [31]. It is important to realize that the formation of crystalline polymeric structures is hindered because of the length of the chains and the presence of entanglements. Therefore, structure formation is governed by kinetical criteria rather than equilibrium thermodynamics: the structure that develops at a given temperature is that with the maximal growth rate rather than that with the lowest Gibbs energy.

-18-

C2.1.6.2 PRIMARY CRYSTALLIZATION

The usual and therefore most important situation where polymers crystallize is in melts cooled below the point of the fusion of a crystallite of infinite dimensions. Then, crystallization occurs by the nucleation and growth of spherulites. Another crystallization process is sometimes encountered in oriented melts and glasses. In such systems, the crystallization seems to occur at once in the whole sample and not at the interface between the growing crystallites and the amorphous matrix. Despite numerous studies, the crystallization process is not fully understood. Scattering measurements suggest a preliminary spinodal decomposition of the undercooled isotropic melt in phases with and without chain ends and chain defects before the formation of the crystallites [32].

In supercooled isotropic melts, crystallite growth occurs after nucleation. Initially, the lateral size of the lamellae increases at a constant rate, but as soon as the spherulites start to touch each other, the growth rate decreases. Overall, the sigmoidal time dependence of volume fraction of the crystalline regions is well described by the Avrami equation: $\phi_c \propto 1 - \exp(-(zt)^\beta)$, in which z is related to the rate constant of the lateral growth of the lamellae and β is a phenomenological parameter called the 'Avrami exponent'. The rate constant of the lateral growth is controlled by the balance of two factors; namely, the thermodynamic driving force of crystallization and the mobility of the chain segments. The former increases with undercooling, the latter decreases, so that a maximal rate of crystallization exists at a given undercooling. The minimal thickness of the lamellae also depends on the undercooling: it decreases with increasing undercooling.

The study of the crystallization of oriented melts and glasses shows that density variations on the length scale of tens of nanometres, as measured with small-angle x-ray scattering, occur before Bragg peaks appear in wide-angle x-ray scattering measurements [33]. These observations exclude a lateral growth of crystalline lamellae and support a continuous phase separation of the whole sample into crystalline and amorphous domains. This process is similar to that found in spinodal decomposition of polymer mixtures. The reason for this alternative crystallization mechanism might be related to an increased mobility of defects along the preoriented chains.

C2.1.6.3 SECONDARY CRYSTALLIZATION

After completion of the primary crystallization at a given temperature, crystallization does not come to an end, but resumes upon cooling. Two modes of secondary crystallization have been identified [34]. The more common one consists of the ‘insertion’ of new crystallites between those formed during primary crystallization. The inserted crystallites have a smaller thickness because they are formed at a lower temperature. The second mode is called ‘surface crystallization’. The increase of the thickness with decreasing temperature is possible if the mobility is still high enough so that the defects concentrated at the crystal surface after primary crystallization can move further into the amorphous domains. Then, because of the higher thermodynamic driving force to crystallization due to the lower temperature, crystal thickness augments.

C2.1.7 POLYMER DYNAMICS AND MECHANICAL BEHAVIOUR

Polymers have found widespread applications because of their mechanical behaviour. They combine the mechanical properties of elastic solids and viscous fluids. Therefore, they are regarded as viscoelastic materials. Viscoelastic

-19-

behaviour does not mean a simple superposition of the two properties, but includes a new phenomenon called anelasticity in which elastic response and viscous flow are coupled. When a load is applied, part of the deformation, although reversible, requires a certain time to occur.

C2.1.7.1 MICROSCOPIC DYNAMICAL MODELS

(A) ROUSE MODEL

A polymer chain can be approximated by a set of balls connected by springs. The springs account for the elastic behaviour of the chain and the beads are subject to viscous forces. In the Rouse model [35], the elastic force due to a spring connecting two beads is $f = b\Delta r$, where Δr is the extension of the spring and the spring constant is $b = 3kT/a_R^2$; a_R is the root-mean-square distance of two successive beads. The viscous force that acts on a bead is the product of the bead velocity u and of the friction coefficient ξ_R of a bead. With these assumptions, one finds for the slowest relaxation mode of the Rouse chain [35], which corresponds to the motion of the end-to-end vector,

$$\tau_R = \frac{1}{3\pi^2} \frac{(\xi_R/a_R^2)}{kT} (R^2)^2. \quad (C2.1.10)$$

Since $\langle R^2 \rangle \propto N$, we have $\tau_R \propto N^2$. It is important to note that τ_R should be independent of the length of the Rouse unit given by a_R . Since a_R^2 is proportional to the number of monomer units between two beads, ξ_R should scale equally. This is the case in a melt, but not for an isolated polymer chain in a solvent where hydrodynamic interactions strongly affect the motion.

Using the fluctuation-dissipation theorem [36], which relates microscopic fluctuations at equilibrium to macroscopic behaviour in the limit of linear responses, the time-dependent shear modulus can be evaluated [37]:

$$G(t) = c_p kT \sum_{m=1}^{N_R-1} \exp\left(-\frac{2t}{\tau_m}\right). \quad (C2.1.11)$$

The summation extends over the $N_R - 1$ Rouse modes with relaxation time τ_m , $m = 1, \dots, N_R - 1$, and c_p is the number of polymer chains per unit of volume. Integrating equation (C.2.1.11) leads to $G(t) \propto t^{-1/2}$. From $G(t)$, the viscosity at the zero shear rate can be calculated [37], $\eta_0 = \int_0^\infty G(t) dt$, and results in $\eta_0 \propto N$. This is indeed found in many melts of polymer chains shorter than the entanglement molecular weight.

(B) ENTANGLEMENT EFFECTS AND REPTATION MODEL

With increasing molecular weight, polymer chains interpenetrate and become entangled. A critical molecular weight at the entanglement limit, M_c , is defined, above which effects of entanglements become apparent. Two of these effects are the occurrence of the rubber–elastic plateau in the mechanical-response functions (see section C.2.1.7.2) and a change in the dependence of the viscosity on the molecular weight.

-20-

Microscopically, entanglements mainly hinder the lateral motion of a polymer chain. On the basis of this idea, De Gennes [38] and Doi and Edwards [39] proposed that the chain motion occurs in a tube formed by the obstacles set by the adjacent chains (figure C2.1.13). The reptation model assumes that the average over the rapid wriggling along the cross section of the tube defines the primitive path of length l_{pr} along which the chain has to diffuse. The time τ_D required for the chain to leave the tube along the curvilinear path is $\tau_D \cong l_{pr}^2/D$, since $D = kT / (N_R \zeta_R)$, $\tau_D \propto N^3$. Experimentally, it was found that $\tau_D \propto N^\nu$, with $3.2 < \nu < 3.6$. The viscosity scales also as $\eta_0 \propto N^\nu$, in contrast to the situation below M_c , where $\eta_0 \propto N$. It is also interesting to compare the dependence of the diffusion coefficient on the molecular weight: in the absence of entanglements, $D \approx \langle R^2 \rangle / \tau_D \propto N^{-1}$, while in the presence of entanglements, $D \approx \langle R^2 \rangle / \tau_D \propto N_R / N_R^3$, hence $D \propto M^{-2}$.

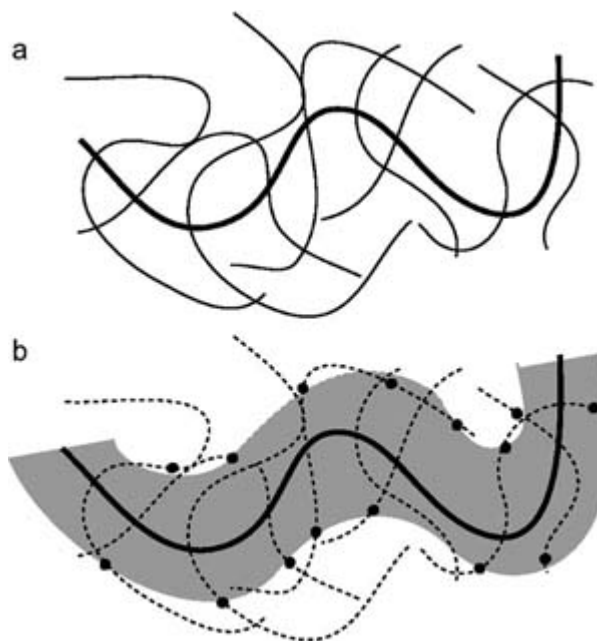


Figure C2.1.13. (a) Schematic representation of an entangled polymer melt. (b) Restriction of the lateral motion of a particular chain by the other chains. The entanglement points that restrict the motion of a chain define a temporary tube along which the chain reptates.

(C) HYDRODYNAMIC INTERACTIONS IN SOLUTIONS

In dilute solutions, the dependence of the diffusion coefficient on the molecular weight is different from that found in melts, either entangled or not. This difference is due to the presence of hydrodynamic interactions among the solvent molecules. Such interactions arise from the necessity to transfer solvent molecules from the front to the back of a moving particle. The motion of the solvent gives rise to a flow field which couples all molecules over a

distance larger than the size of the moving particle. A well known result, derived by Stokes, relates the friction coefficient of a sphere to its radius R_h : $\xi = 6\pi R_h \eta_c$, where η_c is the solvent viscosity.

-21-

In dilute polymer solutions, hydrodynamic interactions lead to a concerted motion of the whole polymer chain and the surrounding solvent. The folded chains can essentially be considered as impermeable objects whose hydrodynamic radius is $\frac{2}{3}R_g$; R_g is the gyration radius defined as

$$R_g^2 = \frac{1}{N} \sum_{i=1}^N \langle |r_i - r_c|^2 \rangle \quad (\text{C2.1.12})$$

where r_i is the position of monomer i and r_c is the position of the centre of gravity. The radius of the gyration is related to the end-to-end distance: for a Gaussian chain, $R_g^2 = \langle R^2 \rangle / 6 = C_\infty N a_0^2 / 6$. Using Einstein's and Stoke's relations we have $D = kT / \xi = kT / (4\pi \eta_s R_g)$. Together with the fact that the radius of gyration is proportional to the root mean square of the end-to-end distance we get $D \propto M^{-\nu}$, with $\nu = 0.5$ for Gaussian chains and $\nu = 0.6$ for expanded chains. The diffusion coefficient can be measured using light scattering and provides the radius of gyration. From the radius of gyration the degree of polymerization can be obtained.

Another simple way to obtain the molecular weight consists of measuring the viscosity of a dilute polymer solution. The intrinsic viscosity $[\eta]$ is defined as the excess viscosity of the solution compared to that of the pure solvent at the vanishing weight concentration of the polymer [40]:

$$[\eta] = \lim_{c_w \rightarrow 0} \frac{\eta - \eta_s}{\eta_s} \frac{1}{c_w} \quad (\text{C2.1.13})$$

The Mark–Houwink–Sakurada equation relates the intrinsic viscosity to the polymer weight:

$$[\eta] = K M^\mu \quad \mu = 3\nu - 1. \quad (\text{C2.1.14})$$

Two limiting cases exist: $\mu = 0.5$, corresponding to $\nu = 0.5$ for Gaussian chains, and $\mu = 0.8$, or $\nu = 0.6$ for expanded chains.

C2.1.7.2 MECHANICAL RESPONSES

(A) RESPONSE FUNCTIONS

Several functions are used to characterize the response of a material to an applied strain or stress [41]. The tensile relaxation modulus $E(t)$ describes the response to the application of a constant tensile strain e_{zz}^0 : $\mathbf{E}(t) = \sigma_{zz}(t) / e_{zz}^0$. Here $\sigma_{zz}(t)$ is the tensile stress and $e_{zz}^0 = \Delta L_z / L_z$, where L_z is the initial length of the sample and ΔL_z is the sample elongation. In shear experiments, the shear relaxation modulus $G(t)$ is defined as $\mathbf{G}(t) = \sigma_{xz}(t) / e_{xz}^0$, where e_{xz}^0 is the constant shear strain applied and $\sigma_{xz}(t)$ is the shear stress. The dynamical shear modulus $G^*(\omega)$ measures the response $\sigma_{xz}^0 \exp(i(\omega t + \delta))$ to a small oscillatory shear strain $e_{xz}^0 \exp(i\omega t)$, $G^*(\omega) = \sigma_{xz}^0 \exp(i\delta) / e_{xz}^0$. Another quantity,

-22-

which is often encountered is the dynamical shear compliance $J^*(\omega) = 1 / G^*(\omega)$, which characterizes the response

to a small oscillatory shear stress. The dynamical shear compliance and the dynamical shear modulus have a real and an imaginary part. The real part corresponds to the elastic response to the oscillatory field applied, while the imaginary part is characteristic of the viscous response and quantifies the work expended on the driven system.

(B) MECHANICAL RELAXATION PROCESSES

Before discussing the complex mechanical behaviour of polymers, consider a simple system whose mechanical response is characterized by a single relaxation time τ , due to the transition between two states. For such a system, the dynamical shear compliance is [42]

$$J^*(\omega) = J' - iJ'' = \frac{\Delta J}{1 + \omega^2\tau^2} - i\frac{\Delta J\omega\tau}{1 + \omega^2\tau^2} \tag{C2.1.15}$$

where ΔJ is related to the equilibrium value of the strain under the shear stress σ_{xz}^0 in a creep experiment: $\Delta J\sigma_{xz}^0 = e_{xz}(t \rightarrow \infty)$. The real and imaginary parts of the dynamical compliance are shown in figure C2.1.14. If the angular frequency of the applied shear stress is much faster than the transition rate constant τ^{-1} , the system reacts to the average stress only, which is zero; if the angular frequency is much slower than τ^{-1} , thermal equilibrium is maintained throughout the deformation process. In the intermediate range, $\omega\tau \approx 1$, the system absorbs energy from the applied stress field.

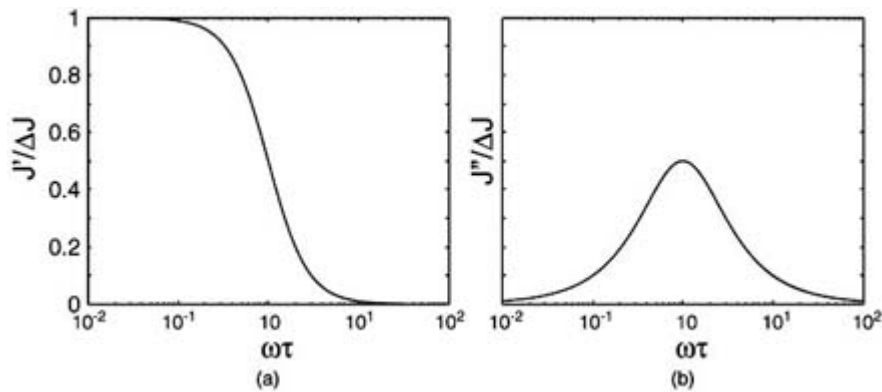


Figure C2.1.14. (a) Real part and (b) imaginary part of the dynamic shear compliance of a system whose mechanical response results from the transition between two different states characterized by a single relaxation time τ .

In the limit of a small deformation, a polymer system can be considered as a superposition of a two-state system with different relaxation times. Phenomenologically, the different relaxation processes are designated by Greek letters, α , β , and γ . The α processes are those with slow relaxation rate constants τ^{-1} , in which several monomers move cooperatively. They are usually associated with the glass transition. On the other hand, the symbol γ is used for fast processes, which are generally localized within a monomer unit. The different processes span a time scale of more than ten orders of magnitude.

The mechanical behaviour of a polymer as a function of temperature is summarized in figure C2.1.15. The compliance is about 10^{-9} N m^{-2} in the glassy state and increases to about 10^{-5} N m^{-2} after the glass–rubber transition. The width of the rubber plateau depends on the density of entanglement. For chains below the critical molecular weight at the entanglement limit, M_c , the plateau disappears and the polymer directly enters the terminal flow region. It is important to note that even in this region polymers still behave as viscoelastic liquids. This is in contrast to low-molecular-weight compounds above their melting point.

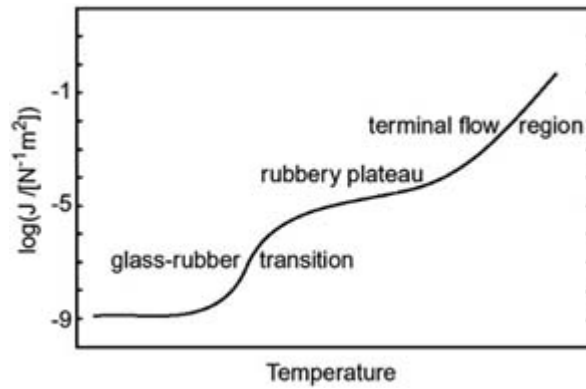


Figure C2.1.15. Schematic representation of the typical compliance of a polymer as a function of temperature.

(C) VOGEL–FULCHER AND WILLIAMS–LANDEL–FERRY EQUATIONS

Most response functions of polymers obey a time–temperature or frequency–temperature superposition [43, 44]. A change in temperature is equivalent to a shift of the logarithmic frequency axis:

$$G^*(T, \log \omega) = G^*(T_0, \log \omega + \log a_T). \quad (\text{C2.1.16})$$

In amorphous polymers, this relation is valid for processes that extend over very different length scales. Modes which involved a few monomer units as well as terminal relaxation processes, in which the chains move as a whole, obey the superposition relaxation. On the basis of this finding an empirical expression for the temperature dependence of viscosity at a zero shear rate and that of the mean relaxation time of α modes were derived:

$$\eta_0(T) = B \exp \frac{T_A}{T - T_V} \quad \tau_\alpha(T) = \tau_0 \exp \frac{T_A}{T - T_V}. \quad (\text{C2.1.17})$$

These are the Vogel–Fulcher equations [44]. In addition to the prefactors, two common parameters appear, namely the activation temperature T_A , typically $T_A = 1000 - 2000$ K, and the Vogel–Fulcher temperature T_V , which is generally 30–70 K below the glass temperature. Using the Vogel–Fulcher equations, Williams, Landel and Ferry derived an expression for the shift parameter $\log a_T$. This expression is known in the literature under the name ‘WLF equation’ [45, 46]:

-24-

$$\log a_T = -C_1 \frac{T - T_0}{T - T_0 + C_2}. \quad (\text{C2.1.18})$$

The parameters C_1 and C_2 are defined as

$$C_1 = \log e \cdot \frac{T_A}{T_0 - T_V} \quad \text{and} \quad C_2 = T_0 - T_V.$$

The WLF equation relates the dependence of the mechanical responses on frequency to that on temperature.

C2.1.8 NONLINEAR MECHANICAL BEHAVIOUR

In the last section we considered the mechanical behaviour of polymers in the linear regime where the response is proportional to the applied stress or strain. This section deals with the nonlinear behaviour of polymers under large deformation. Microscopically, the transition into the nonlinear regime is associated with a change of the polymer structure under mechanical loading.

C2.1.8.1 ELASTICITY OF IDEAL RUBBER

Rubbers are crosslinked polymers above the glass transition. These materials can be stretched by a large factor, sometimes exceeding 10. After removing the load, the system generally recovers its initial shape. In the elongated state, the applied stress is balanced by restoring forces of a mainly entropic nature. Using the extension ratio λ defined as

$$\lambda = \frac{L_z + \Delta L_z}{L_z}$$

where L_z is the original length of the sample and $L_z + \Delta L_z$ is its length under the applied stress, the restoring force f can be written as

$$f(\lambda) = \frac{1}{L_z} \left(\frac{\partial E}{\partial \lambda} \right)_{V,T} - \frac{T}{L_z} \left(\frac{\partial S}{\partial \lambda} \right)_{V,T} \quad (\text{C2.1.19})$$

where E is the internal energy, S is the entropy, and V is the volume of the system. Experimentally, it has been found that the energetic contribution to the force is usually small [47]. If the restoring force is purely entropic, the system is referred to as ideal rubber. The reduction of entropy upon elongation which gives rise to the restoring force is easy to understand: part of the chain conformations accessible in the undeformed state cannot be accessed after elongation. If the chain junctions are assumed to be fixed and deformed in affine manner, an expression for the engineering tensile stress (force divided by the initial cross sectional area) as a function of the elongation can be derived [47]:

-25-

$$\sigma_{zz} = c_p k T \left(\lambda - \frac{1}{\lambda^2} \right) \quad (\text{C2.1.20})$$

c_p is the number of elasticity active chains per volume unit. The comparison between experimental data and the prediction by (C2.1.20) shows a reasonable agreement up to large deformation (figure C2.1.16). For large values of λ , strain hardening arises because of the limited extensibility of the chains or because of shear-induced crystallization.

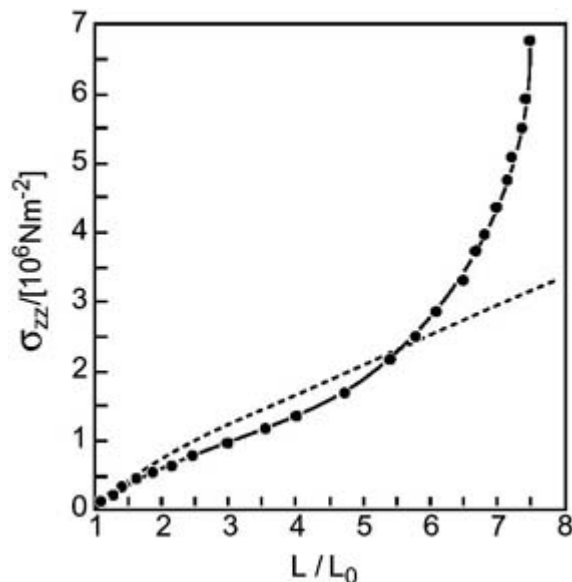


Figure C2.1.16. Tensile stress as a function of the extension ratio registered for a sample of natural rubber (circles). The broken curve is calculated from equation (C2.1.20). (Data from [79].)

C2.1.8.2 SHEAR THINNING AND NORMAL STRESS IN POLYMER MELTS

Polymers owe much of their attractiveness to their ease of processing. In many important techniques, such as injection moulding, fibre spinning and film formation, polymers are processed in the melt, so that their flow behaviour is of paramount importance. Because of the viscoelastic properties of polymers, their flow behaviour is much more complex than that of Newtonian liquids for which the viscosity is the only essential parameter. In polymer melts, the recoverable shear compliance, which relates to the elastic forces, is used in addition to the viscosity in the description of flow [48].

Flow behaviour of polymer melts is still difficult to predict in detail. Here, we only mention two aspects. The viscosity of a polymer melt decreases with increasing shear rate. This phenomenon is called shear thinning [48]. Another particularity of the flow of non-Newtonian liquids is the appearance of stress normal to the shear direction [48]. This type of stress is responsible for the expansion of a polymer melt at the exit of a tube that it was forced through. Shear thinning and normal stress are both due to the change of the chain conformation under large shear. On the one hand, the compressed coil cross section leads to a smaller viscosity. On the other hand, when the stress is released, as for example at the exit of a tube, the coils fold back to their isotropic conformation and, thus, give rise to the lateral expansion of the melt.

C2.1.8.2 YIELD PROCESS AND FRACTURE

Glassy and semicrystalline polymers exhibit complex stress–strain diagrams. The stress–strain relation for the plane-strain compression of bisphenol-A polycarbonate is shown in figure C2.1.17. For small strains the material response is elastic. This behaviour persists up to the yield point, after which further elongation is not accompanied by an increase of stress. Finally, strain hardening sets in, whose signature is a steep exponential-like increase of stress. Ultimately, the sample fractures. If deformation is stopped before this ultimate step, the sample only slightly retracts. However, most of the deformation is reversed when the polymer is heated above the glass transition [49].

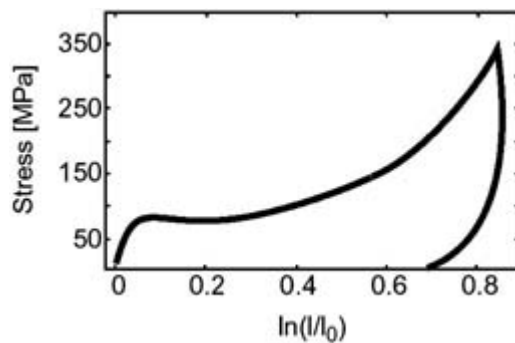


Figure C2.1.17. Stress–strain curve measured from plane-strain compression of bisphenol-A polycarbonate at 25 ° C. The sample was loaded to a maximum strain and then rapidly unloaded. After unloading, most of the deformation remains.

Under compression or shear most polymers show qualitatively similar behaviour. However, under the application of tensile stress, two different deformation processes after the yield point are known. Ductile polymers elongate in an irreversible process similar to flow, while brittle systems whiten due the formation of microvoids. These voids rapidly grow and lead to sample failure [50, 51]. The reason for these conspicuously different deformation mechanisms are thought to be related to the local dynamics of the polymer chains and to the entanglement network density.

Deformation recovery upon heating above the glass transition suggests that the structural changes due to deformation might be similar to those occurring in rubber, where the chains are stretched between two cross-linking points. In glassy polymers, the chain entanglements could act as temporary cross-linking points. Another observation supports the view that entanglements are of primary importance in the deformation of glassy and semicrystalline polymers. Fibres spun from gels with a low density of entanglements have a much higher drawability than fibres spun from entangled melts [52].

C2.1.9 DIFFUSION IN POLYMERS

Small molecules can penetrate and permeate through polymers. Because of this property, polymers have found widespread use in separation technology, protection coating, and controlled delivery [53]. The key issue in these applications is the selective permeability of the polymer, which is determined by the diffusivity and the solubility of a given set of low-molecular-weight compounds. The diffusion of a small penetrant occurs as a series of jumps

from one hole in the polymer matrix to another. Obviously, the size of the holes must be sufficient to accommodate the moving molecule. Because of this restriction, glassy polymers often possess a lower diffusivity but a higher selectivity than rubbers. However, diffusion of small molecules in polymers is a manyfold process. On the one hand, it can be relatively simple in the case where the diffusivity is independent of the local concentration of the penetrant and the polymer matrix is left unchanged by the motion of the small molecules. On the other hand, the penetrant may interact with the matrix and render the diffusivity strongly concentration dependent.

The diffusion of small molecules in polymers can be described using Fick's first and second laws. In a one-dimensional situation, the flux $J(c, x)$ as a function of the concentration c and the position x is given by

$$J = -D(c) \frac{\partial c}{\partial x} \quad (\text{C2.1.21})$$

where $D(c)$ is the diffusion coefficient. The concentration variation is obtained by evaluating the net change in flux within an elementary volume:

$$\frac{\partial c}{\partial t} = -\frac{\partial J}{\partial x} = \frac{\partial}{\partial x} D(c) \frac{\partial c}{\partial x} = D(c) \frac{\partial^2 c}{\partial x^2} + \frac{\partial D}{\partial c} \left(\frac{\partial c}{\partial x} \right)^2. \quad (\text{C2.1.22})$$

If the diffusion coefficient is independent of the concentration, equation (C2.1.22) reduces to the usual form of Fick's second law. Analytical solutions to diffusion equations for several types of boundary conditions have been derived [54]. In the particular situation of a steady state, the flux is constant. Using Henry's law ($c = kp$) to relate the concentration on both sides of the membrane to the partial pressure, the constant flux can be written as

$$J = -Dk \frac{\Delta p}{l} \quad (\text{C2.1.23})$$

where k is Henry's constant, Δp is the partial-pressure difference on the two sides of the membrane, and l is the membrane thickness. The product $P = Dk$ is called the permeability. Often Henry's law may not be applied or the diffusion coefficient may be concentration dependent, so that the permeability is only a phenomenological parameter with practical relevance, but little fundamental significance [55].

Several ideas have been put forward to calculate the diffusion coefficient of small molecules in polymers. Glasstone *et al* [56] proposed an expression based on transition-state theory

$$D = \lambda^2 \frac{kT}{h} \frac{Z^\ddagger}{Z} \exp\left(-\frac{\Delta E^\ddagger}{RT}\right) \quad (\text{C2.1.24})$$

where λ is the root mean square of the jump distance, Z^\ddagger and Z are the partition functions of the system in the activated and normal states, respectively and ΔE^\ddagger is the activation energy of the jump process at 0 K.

-28-

Other expressions for the diffusion coefficient are based on the concept of free volume [57], i.e. the amount of volume in the sample that is not occupied by the polymer molecules. Computer simulations have also been used to quantify the mobility of small molecules in polymers [58]. In a first approach, the partition functions of the ground and activated states and the energy of activation ΔE^\ddagger of a jump process are estimated from computer simulations. The simulations also provides an estimate of the mean jump distance λ , so that D can be calculated from (C2.1.24). Another straightforward approach is to use molecular dynamics simulations and evaluate the diffusion coefficient directly from $\langle |r(t) - r(0)|^2 \rangle = 6Dt$, where $r(t)$ is the position of a penetrant molecule at time t . However, this method is restricted to penetrants with high mobility because the simulation time is limited by the computational power available.

In glassy polymers the interactions of the penetrant molecules with the polymer matrix differ from one sorption site to another. A limiting description of the interaction distribution is known under the name of the dual-sorption model [59, 60]. In this model, the concentration of the penetrant molecules consists of two parts. One obeys Henry's law and the other a Langmuir isotherm:

$$c = c_H + c_L = kp + \frac{C_H bp}{1 + bp} \quad (\text{C2.1.25})$$

where b is the hole affinity constant and C_H is the hole saturation constant. The simplest transport model assumes that the molecules sorbed in the Langmuir mode are immobile. Other models assume different mobilities for the two modes. In all models equilibrium is maintained between the two modes.

In sorption experiments, the weight of sorbed molecules scales as the square root of the time, $M(t) \propto t^{1/2}$, if diffusion obeys Fick's second law. Such behaviour is called case I diffusion. For some polymer/penetrant systems, $M(t)$ is proportional to t . This situation is named case II diffusion [61, 62]. In these systems, sorption strongly changes the mechanical properties of the polymers and a sharp front of penetrant advances in the polymer at a constant speed (figure C2.1.18). Intermediate behaviours between case I and case II have also been found. The occurrence of one mode, or the other, is related to the time the polymer matrix needs to accommodate the structural changes induced by the progression of the penetrant.

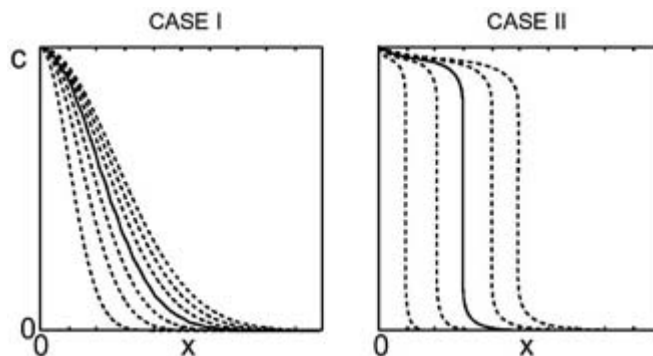


Figure C2.1.18. Schematic representation of the time dependence of the concentration profile of a low-molecular-weight compound sorbed into a polymer for case I and case II diffusion. In both diagrams, the concentration profiles are calculated using a constant time increment starting from zero. The solvent concentration at the surface of the polymer, $x = 0$, is constant.

C2.1.10 COMPUTER SIMULATIONS

The complexity of polymeric systems make the development of an analytical model to predict their structural and dynamical properties difficult. Therefore, numerical computer simulations of polymers are widely used to bridge the gap between the theoretical concepts and the experimental results. Computer simulations can also help the prediction of material properties and provide detailed insights into the behaviour of polymer systems. A simulation is based on two elements: a more or less detailed model of the polymer and a related force field which allows the calculation of the energy and the motion of the system using molecular mechanisms, molecular dynamics, or Monte Carlo techniques [63].

The objective of molecular mechanics is to generate static minimum-energy configurations at the prescribed density, corresponding to the local minima of the total potential energy. Such energy minimizations are used in the generation of realistic conformations suitable as starting structures for molecular dynamics and Monte Carlo simulations. They also allow the estimation of phase stability from the calculation of chemical potential differences through a procedure called thermodynamic integration [64].

Molecular dynamics tracks the temporal evolution of a microscopic model system through numerical integration of the equations of motion for the degrees of freedom considered. The main asset of molecular dynamics is that it provides directly a wealth of detailed information on dynamical processes.

Monte Carlo simulations generate a large number of conformations of the microscopic model under study that conform to the probability distribution dictated by macroscopic constraints imposed on the systems. For example, a Monte Carlo simulation of a melt at a given temperature T produces an ensemble of conformations in which conformation i with energy E_i occurs with a probability proportional to $\exp(-E_i/kT)$. An advantage of the Monte Carlo method is that, by judicious choice of the elementary moves, one can circumvent the limitations of molecular dynamics techniques and effect rapid equilibration of multiple chain systems [65]. However, Monte Carlo

simulations do not provide truly dynamical information.

The complexity of polymer systems prevents their simulation in full structural and dynamical detail [66]. First, the relevant length scales of polymer systems range from about 1 Å, the length of a bond, to hundreds of Ångstroms, the size of the chains. Second, the time scale important to polymer systems covers more than ten orders of magnitude. Consequently, a trade-off between the level of structural detail, the size of the system, and the time scale of the processes under study has to be made. On one extreme, atomistically detailed models [67, 68] provide the specific behaviour of a particular polymer, but the dynamics can be followed up to a few nanoseconds only [67]. On the other extreme, coarse-grained models permit the study of dynamics in the melt and of phase separation processes, but they reveal only universal features, the particular behaviour of different polymers is lost [66]. When none of the extreme models suit, it may be possible to identify the elementary move of Monte Carlo simulations with a relative time step, this method is known under the name dynamical Monte Carlo simulation [66]. Alternatively, the observation window of a molecular dynamics simulation can be shifted to longer times by freezing the fastest degrees of freedom and increasing the duration of the integration time step. For example, in so-called constrained molecular dynamics simulations the bond lengths and the bond angles are kept fixed, so that the integration time step can be chosen one or two orders of magnitude longer than 1 fs, the time step typically used in unconstrained atomistic molecular dynamics simulations [67].

-30-

Atomistically detailed models account for all atoms. The force field contains additive contributions specified in terms of bond lengths, bond angles, torsional angles and possible crossterms. It also includes non-bonded contributions as the sum of van der Waals interactions, often described by Lennard-Jones potentials, and Coulomb interactions. Atomistic simulations are successfully used to predict the transport properties of small molecules in glassy polymers, to calculate elastic moduli and to study plastic deformation and local motion in quasi-static simulations [67, 68]. The atomistic models are also useful to interpret scattering data [69] and NMR measurements [70] in terms of local order.

Coarse-grained models represent the polymer chain as a sequence of beads connected by springs similar to the Rouse model (see section C2.1.7.1(a)). Frictional forces acting on the beads, elastic forces between two connected beads and van der Waals forces between non-connected beads are taken into account in the force field. Several coarse-grained models exist [66]. They are grouped into two main categories depending on whether the bead positions are restricted to a lattice or not. Lattice models permit one to consider excluded-volume effects simply and to sample the possible conformations efficiently. Coarse-grained models are used in the study of melt dynamics, glass transition and entanglement effects [71–73]. They have also contributed to a better understanding of the phase behaviour of polymer blends and copolymers [71, 72].

REFERENCES

- [1] Flory P J 1953 *Principles of Polymer Chemistry* (Ithaca, NY: Cornell University Press) p 3
- [2] Flory P J 1953 *Principles of Polymer Chemistry* (Ithaca, NY: Cornell University Press) p 266
- [3] Strobl G 1997 *The Physics of Polymers* (Berlin: Springer) p 2
- [4] Jenkins A D, Kratochví I P, Stepto R F T and Suter U W 1996 Glossary of basic terms in polymer science *Pure. Appl. Chem.* **68** 2287– 311
- [5] McGrath J E 1992 Polymers synthesis *Encycl. Phys. Sci.* **13** 279– 300
- [6] Odian G 1991 *Principles of Polymerization* (New York: Wiley) p 630
- [7] Mashima K, Nakayama Y and Nakamura A 1997 Recent trends in polymerization of α -olefins catalyzed by organometallic complexes of early transition metals *Adv. Polym. Sci.* **133** 1– 54

- [8] Flory P J 1953 *Principles of Polymer Chemistry* (Ithaca, NY: Cornell University Press) p 178
- [9] Flory P J 1988 *Statistical Mechanics of Chain Molecules* (Munich: Hanser) p 56
- [10] Strobl G 1997 *The Physics of Polymers* (Berlin: Springer) p 21
- [11] Strobl G 1997 *The Physics of Polymers* (Berlin: Springer) p 23
- [12] Strobl G 1997 *The Physics of Polymers* (Berlin: Springer) p 34
- [13] Flory P J 1953 *Principles of Polymer Chemistry* (Ithaca, NY: Cornell University Press) p 425
- [14] De Gennes P G 1979 *Scaling Concepts in Polymer Physics* (Ithaca, NY: Cornell University Press) p 113
-

-31-

- [15] Flory P J 1953 *Principles of Polymer Chemistry* (Ithaca, NY: Cornell University Press) p 495
- [16] De Gennes P G 1979 *Scaling Concepts in Polymer Physics* (Ithaca, NY: Cornell University Press) p 38
- [17] Mattice L M and Suter U W 1994 *Conformational Theory of Large Molecules* (New York: Wiley)
- [18] Rehahn M, Mattice W L and Suter U W 1997 Rotational isomeric state models in macromolecular systems *Adv. Polym. Sci.* **131/132**
- [19] Mandel M 1988 *Polyelectrolytes* (Dordrecht: Reidel)
- [20] Chandrasekhar V 1998 Polymer solid electrolytes synthesis and structure *Adv. Polym. Sci.* **135** 139– 205
- [21] Strobl G 1997 *The Physics of Polymers* (Berlin: Springer) p 65
- [22] Strobl G 1997 *The Physics of Polymers* (Berlin: Springer) p 73
- [23] Flory P J 1949 The configuration of real polymer chains *J. Chem. Phys.* **17** 303– 10
- [24] De Gennes P G 1979 *Scaling Concepts in Polymer Physics* (Ithaca, NY: Cornell University Press) p 56
- [25] Sperling L H 1992 *Introduction to Physical Polymer Science* (New York: Wiley) p 175
- [26] Sperling L H 1992 *Introduction to Physical Polymer Science* (New York: Wiley) p 270
- [27] Doi M 1996 *Introduction to Polymer Physics* (Oxford: Clarendon) p 38
- [28] Strobl G 1997 *The Physics of Polymers* (Berlin: Springer) p 93
- [29] Strobl G 1997 *The Physics of Polymers* (Berlin: Springer) p 100
- [30] Bates F S and Fredrickson G H 1999 Block copolymers—designer soft materials *Phys. Today* **52** 32– 8
- [31] Strobl G 1997 *The Physics of Polymers* (Berlin: Springer) p 145
- [32] Terill N J, Fairclough P A, Towns-Andrews E, Komanshek B U, Young R J and Ryan A J 1998 Density fluctuations: the nucleation event in isotactic polypropylene crystallization *Polymer* **39** 2381– 5
- [33] Cakmak M, Teitge A, Zachman H G and White J L 1993 On-line small-angle and wide-angle x-ray scattering studies on melt-spinning poly(vinylidene fluoride) tape using synchrotron radiation *J. Polym. Sci.* **31** 371– 81
- [34] Strobl G 1997 *The Physics of Polymers* (Berlin: Springer) p 176
- [35] Doi M and Edwards S F 1986 *The Theory of Polymer Dynamics* (Oxford: Clarendon) p 91
- [36] Landau L D and Lifshitz E M 1986 *Statistical Physics Part 1 (Course of Theoretical Physics 5)* (Oxford: Pergamon)

p 384

- [37] Strobl G 1997 *The Physics of Polymers* (Berlin: Springer) p 272
 - [38] De Gennes P G 1979 *Scaling Concepts in Polymer Physics* (Ithaca, NY: Cornell University Press) p 219
 - [39] Doi M and Edwards S F 1986 *The Theory of Polymer Dynamics* (Oxford: Clarendon) p 188
 - [40] Strobl G 1997 *The Physics of Polymers* (Berlin: Springer) p 295
-

-32-

- [41] Ferry J D 1980 *Viscoelastic Properties of Polymers* (New York: Wiley)
- [42] Strobl G 1997 *The Physics of Polymers* (Berlin: Springer) p 207
- [43] Ferry J D 1980 *Viscoelastic Properties of Polymers* (New York: Wiley) p 271
- [44] Strobl G 1997 *The Physics of Polymers* (Berlin: Springer) p 227
- [45] Ferry J D 1980 *Viscoelastic Properties of Polymers* (New York: Wiley) p 274
- [47] Strobl G 1997 *The Physics of Polymers* (Berlin: Springer) p 229
- [47] Treloar L R G 1975 *The Physics of Rubber Elasticity* (Oxford: Clarendon) p 24
- [48] Graessley W W 1993 Viscoelasticity and flow in polymer melts and concentrated solutions *Physical Properties of Polymers* ed J E Mark *et al* (Washington, DC: ACS) pp 97– 143
- [49] Ward I M 1971 *Mechanical Properties of Solid Polymers* (New York: Wiley) p 329
- [50] Kausch H H (ed) 1990 Crazing in polymers *Adv. Polym. Sci.* **91/92**
- [51] Argon A S 1993 Inelastic deformation and fracture of glassy solids *Materials Science and Technology* vol 6 (Weinheim: VCH) pp 462– 508
- [52] Lemstra P J, Kirschbaum R, Ohta T and Yasuda H 1987 *Developments in Oriented Polymers—2* ed I M Ward (Amsterdam: Elsevier) p 39
- [53] Vieth W R 1991 *Diffusion In and Through Polymers* (Munich: Hanser) p 1
- [54] Crank J 1975 *The Mathematics of Diffusion* (Oxford: Clarendon)
- [55] Vieth W R 1991 *Diffusion In and Through Polymers* (Munich: Hanser) p 20
- [56] Glasstone S, Laidler K J and Eyring H 1941 *The Theory of Rate Processes* (New York: McGraw-Hill) p 522
- [57] Frisch H L and Stern S A 1983 Diffusion of small molecules in polymers *Crit. Rev. Solid State Mater. Sci.* **11** 123
- [58] Gusev A A and Suter U W 1995 Relationship between helium transport and molecular motions in a glassy polycarbonate *Macromolecules* **28** 2582– 4
- [59] Gusev A A and Suter U W 1991 Theory for solubility in static systems *Phys. Rev. A* **43** 6488– 94
- [60] Vieth W R 1991 *Diffusion In and Through Polymers* (Munich: Hanser) p 29
- [61] Vieth W R 1991 *Diffusion In and Through Polymers* (Munich: Hanser) p 41
- [62] Crank J 1975 *The Mathematics of Diffusion* (Oxford: Clarendon) p 254
- [63] Allen M P and Tildesley D J 1987 *Computer Simulations of Liquids* (Oxford: Clarendon)

- [64] King P M 1993 *Computer Simulations of Biomolecular Systems* vol 2, ed W F van Gunsteren *et al* (Leiden: ESCOM) pp 315– 48
- [65] Leontidis E, Forrest B M, Widmann A H and Suter U W 1995 Monte Carlo algorithms for the atomistic simulation of condensed polymer phases *J. Chem. Soc. Farad. Trans.* **91** 2355– 68
-

-33-

- [66] Binder K 1995 General aspects of computer simulation techniques and their application to polymer physics *Monte Carlo and Molecular Dynamics Simulations in Polymer Science* ed K Binder (Oxford: Oxford University Press) pp 3–46
- [67] Gentile F T and Suter U W 1993 Amorphous polymer microstructure *Materials Science and Technology, Structure and Properties of Polymers* vol 12, ed E L Thomas (Weinheim: VCH) pp 33– 77
- [68] Kotelyanskii M 1997 Simulation methods for modelling amorphous polymers *Trends Polym. Sci.* **5** 192–8
- [69] Kotelyanskii M, Wagner N J and Paulaitis M E 1996 Building large amorphous polymer structures: atomistic simulation of glassy polymers *Macromolecules* **29** 8497– 506
- [70] Robyr P, Gan Z and Suter U W 1998 Conformation of racemo and meso dyads in glassy polystyrenes from ¹³C polarization-transfer NMR *Macromolecules* **31** 8918– 23
- [71] Binder K (ed) 1995 *Monte Carlo and Molecular Dynamics Simulations in Polymer Science* (Oxford: Oxford University Press)
- [72] Binder K (ed) 1987 *Application of the Monte Carlo Method in Statistical Physics (Topics in Current Physics, vol 36)* (Berlin: Springer)
- [73] Roe R J (ed) 1991 *Computer Simulation of Physics* (Englewood Cliffs, NJ: Prentice-Hall)
- [74] Jannink G 1988 *Les Polymères en Solution: leur Modélisation et leur Structure* (Les Ulis, France: EDP) p 610
- [75] Farnoux *et al* 1978 *J. Physique* **39** 77
- [76] Noda *et al* 1981 *Macromolecules* **14** 668
- [77] Kovacs 1958 *J. Polym. Sci.* **30** 131
- [78] Bates F S and Fredrickson G H 1990 *Ann. Rev. Phys. Chem.* **41** 525
- [79] Treloar L R G 1975 *The Physics of Rubber Elasticity* (Oxford: Clarendon) p 87
-

FURTHER READING

Flory P J 1953 *Principles of Polymer Chemistry* (Ithaca, NY: Cornell University Press)

A fundamental work by a pioneer of polymer science. May be not appropriate as an introductory textbook, but very valuable reading for anyone who is more familiar with polymer science.

Strobl G 1997 *The Physics of Polymers* (Berlin: Springer)

A very good introduction to the physics of polymers with an excellent list of further reading.

Odian G 1991 *Principles of Polymerization* (New York: Wiley)

A classic reference on polymerization.

A very good overview on the behaviour of polymers and their characterization.

De Gennes P G 1979 *Scaling Concepts in Polymer Physics* (Ithaca, NY: Cornell University Press)

An elegant theoretical treatise of polymer physics which conveys an intuitive understanding of the behaviour of macromolecules.

Charrier J-M 1990 *Polymeric Materials and Processing; Plastics, Elastomers and Composites* (Munich: Hanser)

A book that covers many engineering aspects of polymers.

C2.2 Liquid crystals

I W Hamley

INTRODUCTION

We are all familiar with the three states of matter: gases, liquids and solids. In the 19th century the liquid crystal state was discovered [1 and 2]; this can be considered as the fourth state of matter [3]. The essential features and properties of liquid crystal phases and their relation to molecular structure are discussed here. Liquid crystals are encountered in liquid crystal displays (LCDs) in digital watches and other electronic equipment. Such applications are also considered later in this section. Surfactants and lipids form various types of liquid crystal phase but this is discussed in [section C2.3](#). This section focuses on low-molecular-weight liquid crystals, polymer liquid crystals being discussed in the previous section.

The label 'liquid crystal' seems to be a contradiction in terms since a crystal cannot be liquid. However, the term refers to a phase formed between a crystal and a liquid, with a degree of order intermediate between the molecular disorder of a liquid and the regular structure of a crystal. What we mean by order here needs to be defined carefully. The most important property of liquid crystal phases is that the molecules have long-range orientational order. For this to be possible the molecules must be anisotropic, whether this results from a rodlike or dislike shape.

Molecules that are capable of forming liquid crystal phases are called mesogens and have properties that are mesogenic. From the same root, the term mesophase can be used instead of liquid crystal phase. A substance in a liquid crystal phase is termed a liquid crystal. These conventions follow those in the *Handbook of Liquid Crystals*, [4, 5 and 6] the nomenclature of which [7] for various liquid crystal phases is adopted elsewhere in this section.

C2.2.1 TYPES OF LIQUID CRYSTAL

C2.2.1.1 CLASSIFICATION

Liquid crystal phases can be divided into two classes. *Thermotropic* liquid crystal phases are formed by pure

mesogens in a certain temperature range, hence the prefix *thermo* referring to phase transitions in which heat is generated or consumed. About 1% of all organic molecules melt from the solid crystal phase to form a thermotropic liquid crystalline phase before eventually transforming into an isotropic liquid at still higher temperature. In contrast, *lyotropic* liquid crystal phases form in solution and, thus, concentration controls the liquid crystallinity (hence *lyo*, referring to concentration) in addition to temperature. Thermotropic liquid crystals do not need a solvent in order to form. Lyotropic liquid crystal phases are formed by amphiphiles in solution.

-2-

C2.2.1.2 THERMOTROPIC LIQUID CRYSTALS

Thermotropic liquid crystal phases are formed by anisotropic molecules with long-range orientational order and in many types of structure with some degree of translational order. The main types of mesogen are those that are rodlike or calamitic and those that are disclike or discotic.

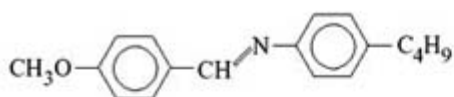
An understanding of the correlation between molecular structure and physical properties of thermotropic mesogens is important in order to optimize parameters such as the operating temperature range. A detailed discussion of such structure–property relationships is beyond the scope of this chapter. Further details can be found elsewhere [8 and 9]. The key feature of calamitic mesogens is a rigid aromatic core to which one or more alkyl chains are attached. Often the core is formed from linked 1,4-phenyl groups. Within a homologous series it is often found that the nematic phase is stable when the alkyl chain is short, whereas smectic phases are found with longer chains. The groups that link aromatic moieties in the core should maintain its linearity, whilst additionally increasing the length and polarizability of the core if liquid crystal phase formation is to be enhanced. Terminal units such as cyano groups also favour the formation of liquid crystal phases, due to polar attractive interactions between pairs of molecules. Lateral substituents are also used to control molecular packing. These are groups attached to the side of a molecule, usually in the aromatic core. Suitable lateral substituent such as fluoro groups can enhance molecular polarizability. On the other hand, they can disrupt molecular packing and thus reduce the nematic–isotropic phase transition temperature. Perhaps the most important use of lateral substitution is to generate the tilted smectic C phase by creating a lateral dipole. This is especially important in the chiral smectic C phase that is the basis of ferroelectric displays (section C2.2.4.3).

(A) NEMATIC PHASE

This is the simplest liquid crystal phase. It is formed by calamitic or discotic mesogens, typical examples of the former being shown in [figure C2.2.1](#). The molecules have no long-range translational order, just as in a normal isotropic liquid. However, they do possess long-range orientational order, in contrast to a liquid. The nematic phase can thus be considered to be an anisotropic liquid. It is denoted N, and an illustration of its structure is included in [figure C2.2.2](#). The most successful theories for orientational order in liquid crystals deal with the nematic phase.

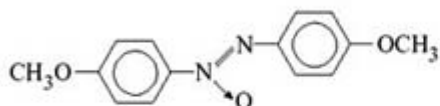
-3-

N-(4-methoxybenzylidene)-4'-butylaniline (MBBA)



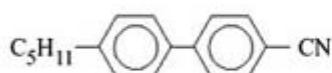
Cr 27 N 47 I

4,4-dimethoxyazoxybenzene (*p*-azoxyanisole, PAA)



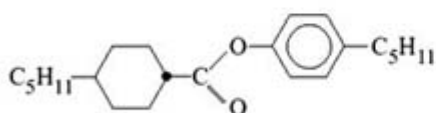
Cr 118 N 136 I

4-pentyl-4'-cyanobiphenyl (5CB)



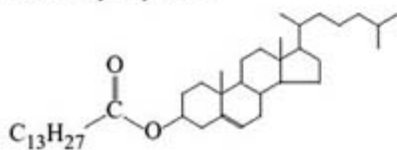
Cr 23 N 35 I

4-pentylphenyl-*trans*-4'-pentylcyclohexylcarboxylate



Cr 37 N 47 I

cholesteryl myristate



Cr 71 SmA 81 N* 86.5 I

Figure C2.2.1. Examples of rodlike nematogens.

The nematic phase formed by chiral molecules is itself chiral. This used to be called the cholesteric phase, because the mesogen for which it was first observed contained a cholesterol derivative. However, it is now called the chiral nematic phase, denoted N*, because it has been observed for other types of mesogen. The chiral nematic phase is illustrated in [figure C2.2.2](#). The director (average direction of molecules) twists round in a helix. It is important to note that this helical twist refers to the *average orientation* of molecules and not the packing of molecules themselves, because they do not have long-range translational order. The helical structure has a characteristic pitch, or repeat distance along the helix, which can range from about 100 nm to near infinity. When the pitch length is comparable to the wavelength of light, the chiral nematic phase scatters or reflects visible light, producing colours. Furthermore, the pitch and, thus, colour are sensitive to temperature, which is the basis of thermochromic devices, i.e. those that produce colour changes in response to temperature ([section C2.2.4.6](#)).

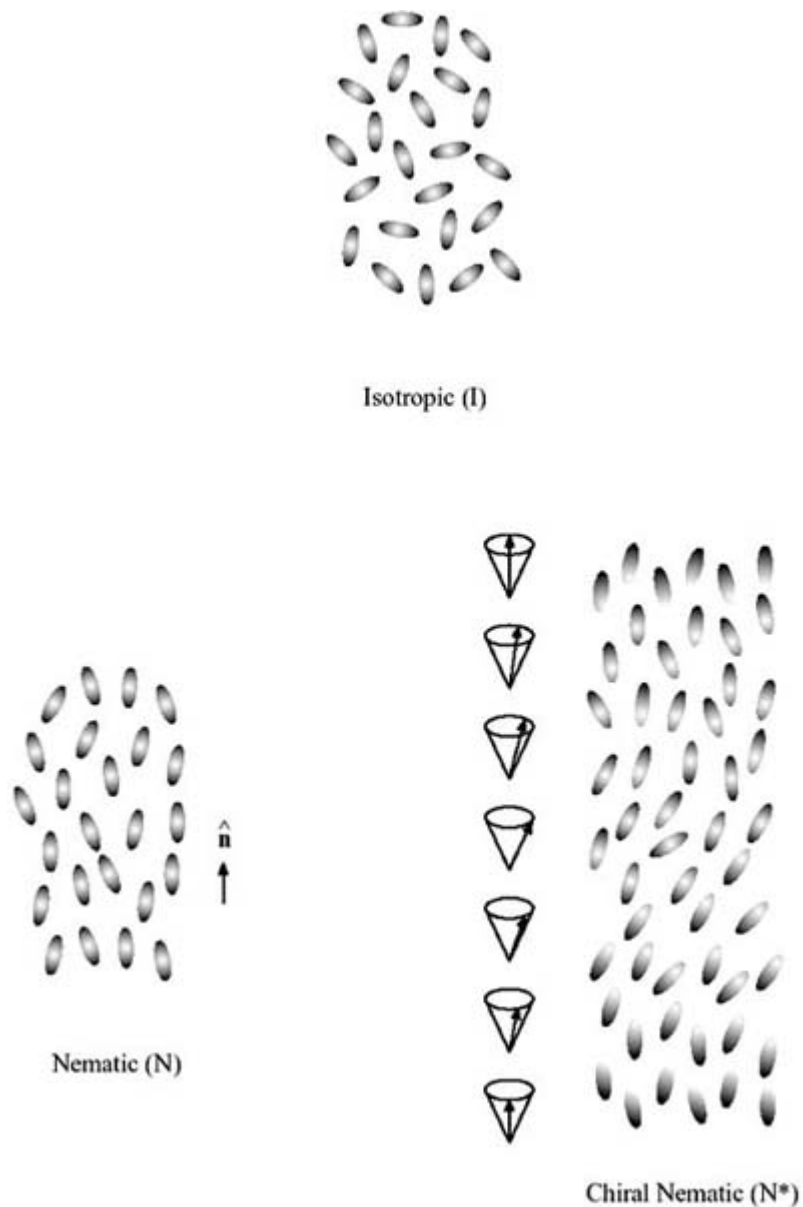


Figure C2.2.2. Isotropic, nematic and chiral nematic phases. Here \hat{n} denotes the director. In the chiral nematic phase, the director undergoes a helical rotation, as schematically indicated by its reorientation around a cone.

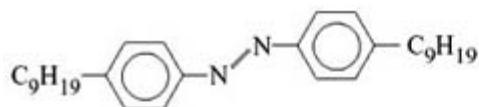
Although in figure C2.2.2 they are sketched with rodlike molecules, both nematic and chiral nematic phases can also be formed by discotic molecules.

(B) SMECTIC PHASES

The notation follows the discovery of different smectic phases, largely on the basis of miscibility experiments which did not provide information on the molecular arrangement. Some phases originally thought to be smectic (e.g. smectic D) turned out not to be so [10]; thus the modern nomenclature system is not very systematic. Typical mesogens forming smectic phases are shown in figure C2.2.3. Smectic phases are characterized by weak layering of molecules. This layering is usually so weak that the density modulation is essentially sinusoidal normal to the 'layers'. In a smectic A (SmA) phase the molecules are, on average, normal to the layers (figure C2.2.4). In contrast, in smectic C (SmC) phases the director is tilted with respect to the layers (figure C2.2.4). Different alignments of this structure are possible in which the molecules are aligned with an external field and the layers are tilted, or if grown from an SmA phase in a weak aligning field, the layer orientation can stay the same and the

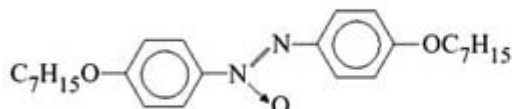
molecules can tilt.

4,4'-dinonylazobenzene



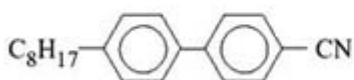
Cr 37 SmB 40 SmA 53 I

4,4'-diheptylazoxybenzene (HOAB)



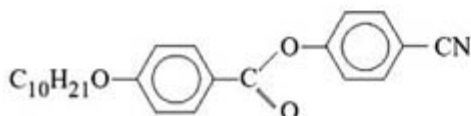
Cr 74.5 SmC 95.5 N 124 I

4-octyl-4'-cyanobiphenyl (8CB)



Cr 21 SmA 32.5 N 40 I

4-cyanophenyl-*trans*-4'-decyloxyphenylcarboxylate



Cr 79 SmA 79 N 86.5 I

Figure C2.2.3. Examples of smectogens.

-6-

In the smectic A₁ (SmA₁) phase, the molecules point up or down at random. Thus, the density modulation can be described as a Fourier series of cosines:

$$c_m(N) \propto \frac{N}{(R^2)^{3/2}} \propto N^{-1/2}. \quad (\text{C2.2.1})$$

Here the ρ_n are the amplitudes of the harmonics of the density, q_s is the wavenumber and the Φ_n are arbitrary phase angles, which are necessary for a complete theoretical description of this structure (see [section C2.2.3.4](#)). The z direction is, by convention, normal to the layers.

The smectic A phase is a liquid in two dimensions, i.e. in the layer planes, but behaves elastically as a solid in the remaining direction. However, true long-range order in this one-dimensional solid is suppressed by logarithmic growth of thermal layer fluctuations, an effect known as the Landau–Peierls instability [[11](#), [12](#) and [13](#)]

Detailed x-ray diffraction studies on polar liquid crystals have demonstrated the existence of multiple smectic A and smectic C phases [[14](#), [15](#) and [16](#)]. The first evidence for a smectic A–smectic A phase transition was provided by the optical microscopy observations of Sigaud *et al* [[17](#)] on binary mixtures of two smectogens. Different structures exist due to the competing effects of dipolar interactions (which can lead to alternating head–tail or interdigitated structures) and steric effects (which lead to a layer period equal to the molecular length). These

phases are thus sometimes referred to as frustrated smectics to reflect the simultaneous presence of two, sometimes incommensurate, length scales [18, 19 and 20]. Observed smectic A and smectic C structures are shown in [figure C2.2.5](#). Here the arrows denote longitudinal molecular dipoles. In the SmA₁ phase, the layer periodicity, d , is equal to the molecular length l . The molecules are interdigitated in the SmA_d phase, due to overlap between aromatic cores in antiparallel dimers of polar molecules (e.g. with NO₂ or CN terminal groups), leading to typical values of $d = (1.4-1.8)l$. In the SmA₂ phase, the polar molecules are arranged in an antiparallel arrangement with $d = 2l$ [21,22]. There are also two modulated smectic phases [21,22]. In the Sm \bar{A} phase, there is an alternation of antiferroelectric ordering producing a ‘ribbon’ structure, in which the ribbons are arranged on a centred lattice. In the SmA_{cre} ‘crenellated’ phase, on the other hand, the ribbons lie on a primitive lattice, i.e. there is an alternation in the lateral size of ‘up’ and ‘down’ domains ([figure C2.2.5](#)). An Sm \bar{C} phase has also been observed, with an alternation of bilayers in which the molecules are tilted with respect to the layers ([figure C2.2.5](#)). Finally, so-called ‘incommensurate’ SmA phases have been identified, in which SmA_d and either SmA₁ or SmA₂ periodic density waves coexist along the layer normal producing SmA_{1,inc} and SmA_{2,inc} phases, respectively. Such phases are quite difficult to represent in real space, so are not shown in [figure C2.2.5](#). In the case of a weakly coupled phase, the two independent and incommensurate waves coexist almost independently of each other, whereas in a strongly coupled incommensurate (soliton) SmA_{inc} phase regions of ‘locked’ SmA ordering are separated by smaller regions where the coexisting density waves are out of phase [21,23, 24 and 25]. X-ray diffraction is an invaluable technique to elucidate the structure of frustrated smectics, because Bragg peaks are obtained that are reciprocally related to the periodicities in the structure. In an oriented sample, the orientation of these peaks furthermore indicates the direction of these periodicities. Excellent reviews have been provided of the experimental evidence for frustrated smectic phases [21, 22] and of their theoretical description [18, 19, 21, 23, 25].

Both SmA and SmC phases are characterized by liquid-like ordering within the layer planes. Other types of smectic phase with in-plane order have been identified [9, 20, 26, 27]. These phases exhibit bond orientational order but short-range positional order within the smectic layers. The layers themselves are stacked with quasi-long-range order. True long-range ordering is suppressed due to the Landau–Peierls instability, as in all smectic phases. Bond-orientational order refers to the orientation of the vectors defining the in-plane lattice; it was proposed to exist in a smectic phase [28, 29] before being observed [30, 31]. As illustrated in [figure C2.2.4](#) in smectic B, smectic I and smectic F phases there is sixfold bond-orientational order, i.e. the lattice orientation is retained in the layers but the translational order is lost within a few intermolecular distances. The smectic B (SmB, sometimes known as hexatic B) phase resembles the SmA phase, but with long-range ‘hexatic’ bond-orientational order. The SmI and SmF phases are tilted versions of SmB. In the SmI phase, the molecules are tilted towards a vertex (nearest neighbour) whereas in the SmF phase they are tilted towards an edge of the hexagonal net.

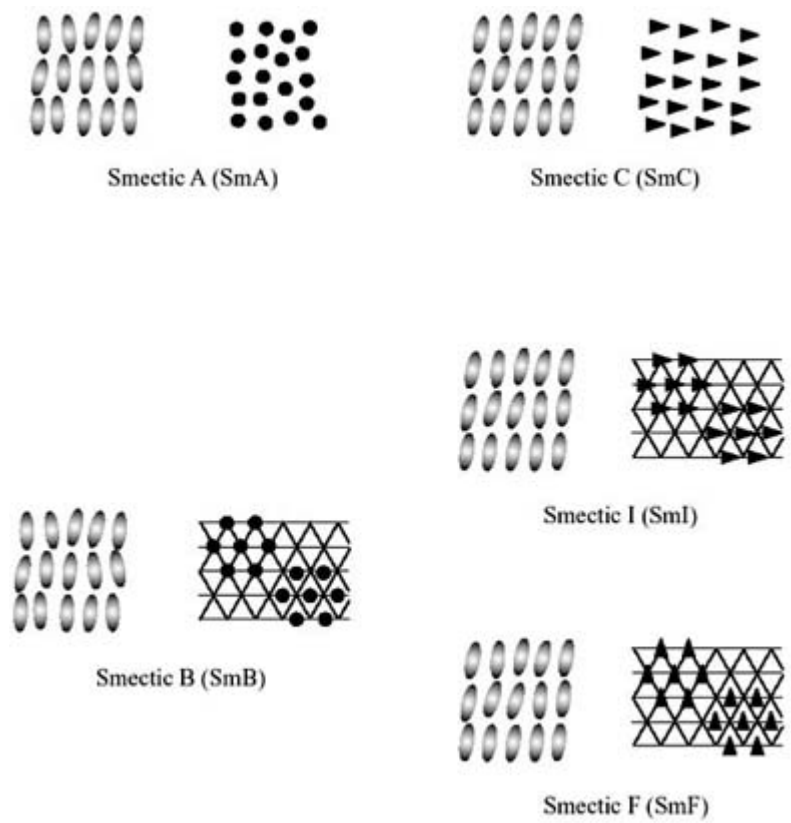


Figure C2.2.4. Types of smectic phase. Here the layer stacking (left) and in-plane ordering (right) are shown for each phase. Bond orientational order is indicated for the hexB, SmI and SmF phases, i.e. long-range order of lattice vectors. However, there is no long-range translational order in these phases.

-8-

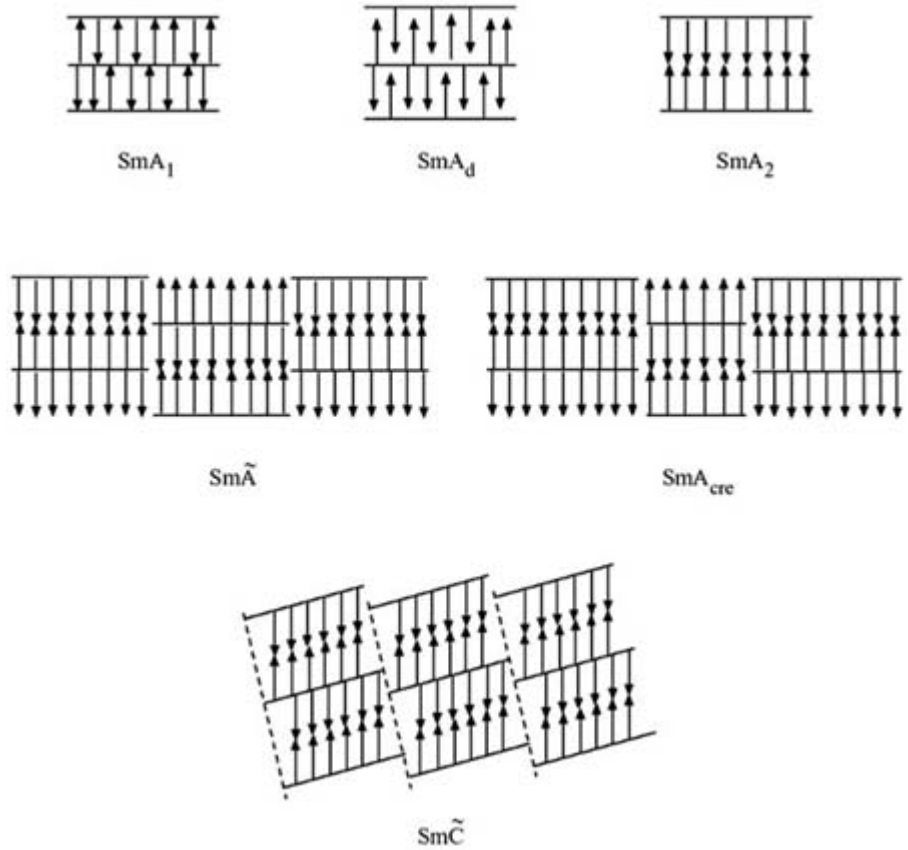


Figure C2.2.5. Frustrated smectic phases. Here the arrows denote longitudinal molecular dipoles.

-9-

Versions of the SmB, SmI and SmF phases with a higher degree of order were originally classified as smectic phases; however, they are now known to be ‘soft-crystal’ phases, with true long-range positional order in three dimensions. The layers are, however, very weakly attached to each other and this was the source of the original misidentification. The crystal version of SmB is now termed crystal B (abbreviated simply as B [7]) and crystal J and crystal G are three-dimensionally ordered versions of SmI and SmF, respectively. A further soft-crystal phase, confused in the early literature with a smectic, is the crystal E phase in which the molecules have a ‘herringbone’ or ‘chevron’ packing, which results from the quenching of sixfold rotational disorder in the B phase to produce long-range ordering of the short molecular axes. Tilted versions of this phase, called crystal H and crystal K, are derived from G and J phases respectively [20, 22, 26 and 27].

As with the nematic phase, a chiral version of the smectic C phase has been observed and is denoted SmC*. In this phase, the director rotates around the cone generated by the tilt angle [9,32]. This phase is helielectric, i.e. the spontaneous polarization induced by dipolar ordering (transverse to the molecular long axis) rotates around a helix. However, if the helix is unwound by external forces such as surface interactions, or electric fields or by compensating the pitch in a mixture, so that it becomes infinite, the phase becomes ferroelectric. This is the basis of ferroelectric liquid crystal displays (section C2.2.4.4). If there is an alternation in polarization direction between layers the phase can be ferrielectric or antiferroelectric. A smectic A phase formed by chiral molecules is sometimes denoted SmA*, although, due to the untilted symmetry of the phase, it is not itself chiral. This notation is strictly incorrect because the asterisk should be used to indicate the chirality of the phase and not that of the constituent molecules.

(C) COLUMNAR PHASES

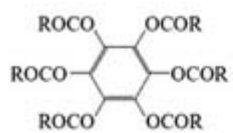
Columnar phases are formed by discotic mesogens [33], examples of which are shown in figure C2.2.6. An excellent review of molecules that form discotic phases has recently appeared [34]. Discotic molecules can form a nematic phase (termed N_D) just like calamitic mesogens. In addition, several types of columnar phase have been observed (figure C2.2.7) [35]. The recommended abbreviation for these phases is col [7], although D is often encountered, especially in the early literature. In the col_{hd} phase there is a disordered stacking of discotic molecules in the columns which are packed hexagonally. Hexagonal columnar phases where there is an ordered stacking sequence (col_{ho}) or where the mesogens are tilted within the columns (col_t) are also known [9, 20, 34, 35 and 36]. It is, however, important to note that individual columns are one-dimensional stacks of molecules and long-range positional order is not possible in a one-dimensional system, due to thermal fluctuations and, therefore, a sharp distinction between col_{hd} and col_{ho} is not possible [20]. Phases where the columns have a rectangular (col_{rd}) or oblique packing ($col_{ob,d}$) of columns with a disordered stacking of mesogens have also been observed [9, 20, 25, 34, 35 and 36].

C2.2.1.3 LYOTROPIC LIQUID CRYSTALS

Lyotropic liquid crystals are discussed in section C2.3 and section C2.6 of this encyclopedia and will not be considered further here.

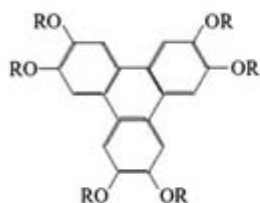
-10-

Hexaester of benzene



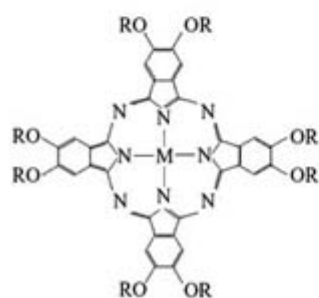
R=C₆H₁₃: Cr 81 Col_h 86 I

Triphenylene derivative



R=C₁₂H₂₅O: Cr 83 Col_r 99 Col_h 118 I

Peripherally substituted octa-alkoxyphthalocyanine



R=C₁₂H₂₅: Cr 105 Col_{ho} 310 I

M=Cu

Figure C2.2.6. Examples of dislike mesogens.

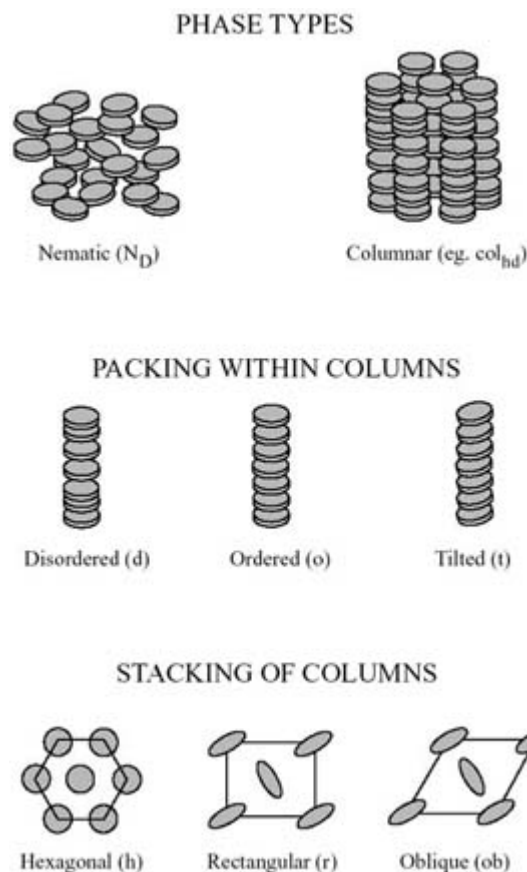


Figure C2.2.7. Schematic illustrating the classification and nomenclature of discotic liquid crystal phases. For the columnar phases, the subscripts are usually used in combination with each other. For example, D_{rd} denotes a rectangular lattice of columns in which the molecules are stacked in a disordered manner (after [33])

C2.2.2 CHARACTERISTICS OF LIQUID CRYSTAL PHASES

C2.2.2.1 IDENTIFICATION OF LIQUID CRYSTAL PHASES

(A) TEXTURES

Liquid crystal phases possess characteristic textures when viewed in polarized light under a microscope. These textures, which can often be used to identify phases, result from defects in the structure. Compendia of micrographs showing typical textures exist to facilitate phase identifications [37, 38]. These monographs also discuss the origins of defect structures in some detail.

As in crystals, defects in liquid crystals can be classified as point, line or wall defects. Dislocations are a feature of liquid crystal phases where there is translational order, since these are line defects in this ‘lattice’ order. Unlike crystals, there is a type of line defect unique to liquid crystals termed disclination [39]. A disclination is a discontinuity of orientation of the director field.

Disclinations in the nematic phase produce the characteristic ‘Schlieren’ texture, observed under the microscope using crossed polars for samples between glass plates when the director takes nonuniform orientations parallel to the plates. In thicker films of nematics, textures of dark flexible filaments are observed, whether in polarized light or not. This texture, in fact, gave rise to the term nematic (from the Greek for ‘thread’) [40]. The director fields

around disclinations of different ‘strength’, s , are shown in figure C2.2.8 (the lines run normal to the page). The variation in director orientation can be mapped out by rotating the sample between crossed polars. If the director $\hat{\mathbf{n}}$ in the xy -plane is denoted by a vector $\hat{\mathbf{n}} = [\cos \theta(\mathbf{r}), \sin \theta(\mathbf{r})]$, then it can be shown [37, 41, 42] that θ varies with $\mathbf{r} = (x, y)$ as

$$\theta = s \tan^{-1} \left(\frac{y}{x} \right) + \theta_0 \quad (\text{C2.2.2})$$

where θ_0 is a fixed angle. The chiral nematic phase has textures distinct from those of the non-chiral phase, which depend on the director orientation with respect to the confining glass slides. These are discussed in [37] and [43]

Wall defects are also very important in nematic phases, especially in electric or magnetic fields. This will be considered further in [section C2.2.4.1](#), which discusses Fréedericksz transitions in a nematic in an electric or magnetic field.

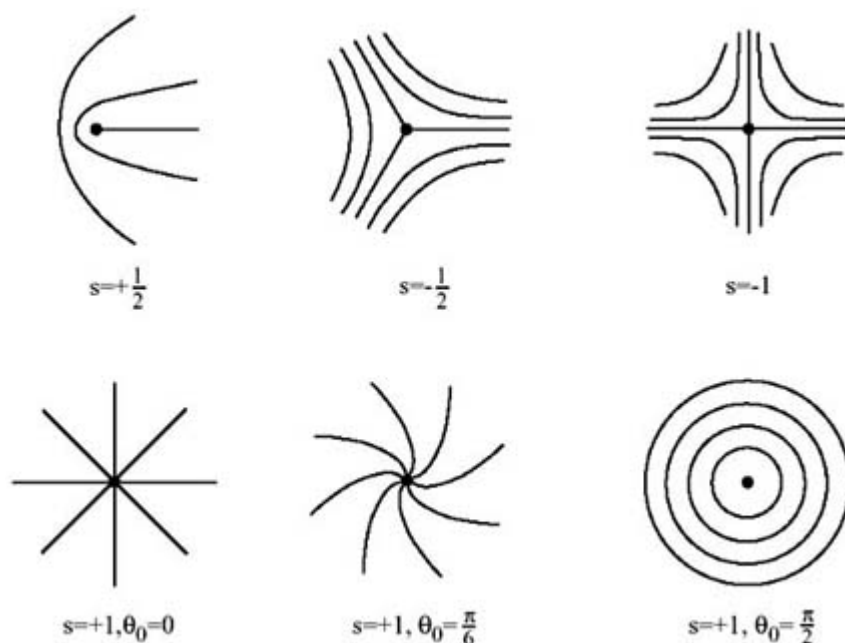


Figure C2.2.8. The director field in a nematic around disclinations of various strengths, s . The director fields are given by equation (C2.2.2).

Smectic A phases, in which the layers are not uniformly parallel to the glass slides confining the sample (non-planar orientation [37, 43]) are characterized by ‘fanlike’ textures, made up of ‘focal conics’ [20, 37, 43, 44]. A focal conic is an intersection in the plane of a geometric object called a Dupin cyclide, that results from lamellae forming a concentric roll (like a swiss roll or jelly roll) being bent into an elliptical torus of non-uniform cross-section. The fan texture results from disclinations in layers perpendicular to the plane of the confining glass plates, usually from focal conics packed into polygonal domains, producing ellipses lying in the plane (figure C2.2.9) [20, 37, 44]. For smectic A phases with layers oriented parallel to the confining slides, defects such as steps at the edge of lamellae are common [37] although other types of defect have been observed [20]. If a smectic C phase is prepared by cooling a smectic A phase, regions of Schlieren texture develop in areas of the sample that previously appeared dark under crossed polars, that can coexist with a fan structure. The hexatic B phase is characterized by ‘mosaic’ or ‘broken’ fan textures [9, 37, 38] in non-planar orientations, whilst SmI and SmF phases show a Schlieren texture similar to that of SmC, and can often be difficult to distinguish [9]. The crystal phases are characterized by mosaic, platelet or batonnet structures. Further details can be found elsewhere [37, 38].

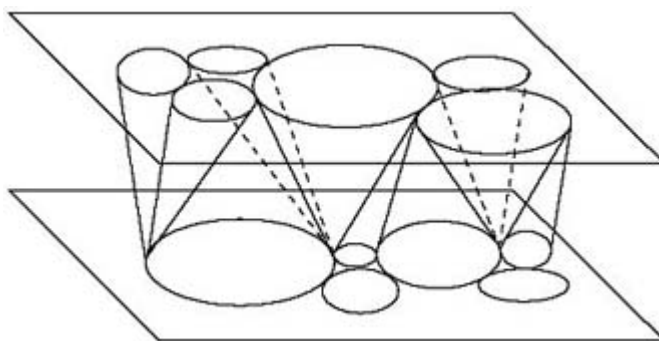


Figure C2.2.9. Polygonal domains of focal conics in a smectic A phase confined between parallel plates.

(B) LIGHT SCATTERING

The milky appearance of nematics is due to variations in refractive index on the length scale of the wavelength of visible light, which result from thermal fluctuations of director orientation. It is possible to analyse the angular dependence of scattered light intensity in static light scattering experiments [45, 46] to obtain ratios of Frank elastic constants (defined in section C2.2.3.3). However, dynamic light scattering (DLS) proves far more powerful since it yields information on hydrodynamic modes as well as the static elasticities. It can thus be used to obtain the Leslie coefficients related to the viscosity of nematic phases (see section C2.2.3.3). In samples oriented in specific geometries, it is also possible to measure the Frank elastic constants K_1 , K_2 and K_3 individually using DLS rather than just their ratios [45, 46].

The aforementioned light scattering techniques probe long-range fluctuations of the director. Other spectroscopic techniques can provide information on molecular ordering. For example, polarized Raman spectroscopy has been used to measure orientational order parameters in nematic and smectic phases based on measurements of the depolarization of light in oriented samples [47]. These measurements depend on anisotropic molecular polarizabilities, associated with specific Raman active bond vibrations, in liquid crystal phases [47]. Brillouin scattering is a type of inelastic scattering characterized by smaller frequency shifts than Raman bands [48]. It results from light scattered from alternate layers

-14-

of compression and rarefaction produced by phonons in a material and has been applied to nematic and smectic liquid crystals [49, 50]. It provides a measure of the velocity and absorption of second sound, and has been used to obtain elastic constants in the smectic phase [50].

(C) X-RAY AND NEUTRON DIFFRACTION

X-ray diffraction is one of the primary methods to determine the structure of a liquid crystal phase [22, 51]. Smectic phases are characterized by Bragg spots, which result from the layer periodicity. If the sample is oriented, usually in an electric or magnetic field, this yields further information for example on the tilt in SmC phases or on the structure of modulated phases. An oriented nematic phase is characterized by diffuse arcs that result from local anisotropic intermolecular interactions. Orientated smectic and nematic phases are characterized by wide-angle scattering arcs which result from the side-to-side packing of molecules. The anisotropy of these arcs is related to the extent of orientational ordering; indeed, the azimuthal angular dependence of the scattered intensity can be used to obtain an orientational order parameter [52]. However, this analysis does not provide information on the single molecule orientational distribution function (section C2.2.3.1). A powerful method to obtain this exploits the contrast variation possible in small-angle neutron scattering using deuterium labelling. A mixture of normal and deuterated mesogens produces purely single-molecule scattering at low angles, and this can be directly analysed to provide order parameters and to reconstruct the orientational distribution function [53, 54]. In fact, diffraction is capable of providing, in principle, the full distribution function, unlike spectroscopic methods [55]. Neutron scattering has also been used to probe molecular diffusive motions (rotational and translational), via incoherent quasi-elastic neutron scattering [54].

(D) SPECTROSCOPIC TECHNIQUES

Of spectroscopic techniques, nuclear magnetic resonance (NMR) has been most widely used to measure orientational ordering in liquid crystals [56, 57 and 58]. Most commonly, changes of line splittings in the spectra of deuterium-labelled molecules are used, specifically ^2H quadrupolar splittings or intermolecular dipole–dipole couplings between pairs of protons. If molecules are partially deuterated then information on the ordering of the labelled segments can be obtained. In addition, the dipolar spectra are easier to analyse than those of fully protonated molecules, when deuterium decoupling techniques are exploited. Further details are provided in [57]. Another method to obtain spectra that can easily be interpreted is to use rigid solutes dissolved in the liquid crystal phase [57]. If the structure of the solute molecules is not too complicated, the dipolar couplings can be analysed to provide orientational order parameters. Of course, the method only provides information on orientational ordering in liquid crystal–solute mixtures. However, since the form of the anisotropic interactions should be the same as in the pure liquid crystal phase, studies on solutes can provide indirect information on the ordering of the mesogens. The analysis of chemical shift anisotropies has recently been exploited to probe orientational ordering [56].

NMR is not the best method to identify thermotropic phases, because the spectrum is not directly related to the symmetry of the mesophase, and transitions between different smectic phases or between a smectic phase and the nematic phase do not usually lead to significant changes in the NMR spectrum [56]. However, the nematic–isotropic transition is usually obvious from the discontinuous decrease in orientational order. NMR can, however, be used to identify lyotropic liquid crystal phases, using ^2H NMR on solutions in D_2O . Cubic, hexagonal and lamellar phases can be distinguished due to different averages of the quadrupolar interaction which result from differences in the curvature and symmetry of the amphiphile–water interface [56, 57 and 58].

-15-

(E) DIFFERENTIAL SCANNING CALORIMETRY

This method is used to locate phase transitions via measurements of the endothermic enthalpy of phase transition. Details of the technique are provided elsewhere [25, 58]. Typically, the enthalpy change associated with transitions between liquid crystal phases or from a liquid crystal phase to the isotropic phase is much smaller than the melting enthalpy. Nevertheless, it is possible to locate such transitions with a commercial DSC, since typical enthalpies are $1\text{--}5\text{kJ mol}^{-1}$ [9]. These relatively small values indicate that transitions between liquid crystal phases and between these and the isotropic phase involve much more delicate structural changes than those that accompany the crystal–liquid crystal melting transition. Most liquid crystal phase transitions are first order (discontinuous in enthalpy), although some, such as the SmC to SmA transition can be second order (continuous). The latter can be difficult to locate, because the heat capacity is small and it is then necessary to turn to a higher-resolution technique such as adiabatic scanning calorimetry [59].

C2.2.3 THEORY

Thermotropic liquid crystal phases are formed by rodlike or disclike molecules. However, in the following we consider orientational ordering of rodlike molecules for definiteness, although the same parameters can be used for discotics. In a liquid crystal phase, the anisotropic molecules tend to point along the same direction. This is known as the director, which is a unit vector denoted \hat{n} .

C2.2.3.1 DEFINITION OF AN ORIENTATIONAL ORDER PARAMETER

Long-range orientational order of the constituent molecules is the defining characteristic of liquid crystals. It is therefore important to be able to quantify the degree of orientational order. To do this, an orientational order parameter is introduced, which describes the average orientation of the molecules. In general, the orientational distribution for a rigid molecule is a function of the three Euler angles $\Omega = (\alpha, \beta, \gamma)$ with respect to \hat{n} . However, for a uniaxial phase of cylindrically symmetric molecules, only the polar angle β is relevant. The orientational

distribution function then describes the probability for molecules to be oriented at an angle β with respect to the average, i.e. with respect to the director. It is usually denoted by $f(\beta)$ and in terms of the anisotropic potential of mean torque, $U(\beta)$, is defined as

$$f(\beta) = Z^{-1} \exp[-U(\beta)/kT] \quad (\text{C2.2.3})$$

where Z is the orientational partition function

$$Z = \int \exp[-U(\beta)/kT] d(\cos \beta). \quad (\text{C2.2.4})$$

Typical shapes of the orientation distribution function are shown in [figure C2.2.10](#). In a liquid crystal phase, the more highly oriented the phase, the more $f\beta$ tends to be sharply peaked near $\beta=0$. However, in the isotropic phase, a molecule has an equal probability of taking on any orientation and then $f\beta$ is constant.

An orientational order parameter can be defined in terms of an ensemble average of a suitable orthogonal polynomial. In liquid crystal phases with a mirror plane of symmetry normal to the director, orientational ordering is specified,

-16-

to lowest order, by the order parameter

$$\bar{P}_2 = \overline{\frac{3}{2} \cos^2 \beta - \frac{1}{2}}. \quad (\text{C2.2.5})$$

Here the bar indicates an average over the orientational distribution function. Here $P_2(\cos \beta) = (\frac{3}{2} \cos^2 \beta - \frac{1}{2})$ is the second rank Legendre polynomial. This average takes the value $\bar{P}_2 = 0$ for an isotropic phase. For a completely oriented phase $\bar{P}_2 = 1$. The order parameter is sometimes denoted by the symbol S [20]. The average in equation (C2.2.5) can be written explicitly in terms of the orientational distribution function:

$$\bar{P}_2 = \int P_2(\cos \beta) f(\beta) d(\cos \beta). \quad (\text{C2.2.6})$$

To completely specify the orientational ordering, the complete set of orientational order parameters, $\bar{P}_L, L = 0, 2, 4, \dots$, is required. Only the even rank order parameters are non-zero for phases with a symmetry plane perpendicular to the director (e.g. N and SmA phases).

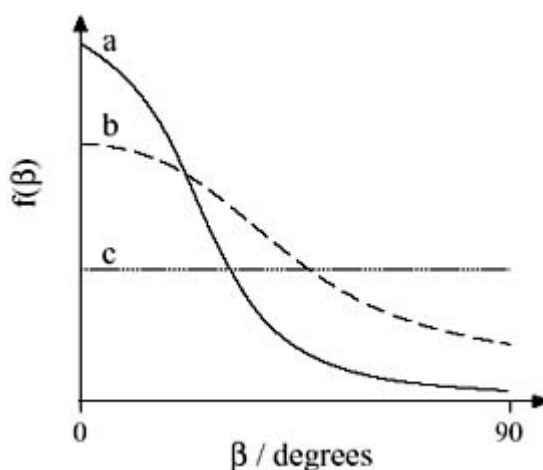


Figure C2.2.10. Orientational distribution functions for (a) a highly oriented liquid crystal phase, (b) a less well

oriented liquid crystal phase and (c) an isotropic phase.

C2.2.3.2 THEORIES FOR ORIENTATIONAL ORDER

There are basically two types of theory for orientational ordering in liquid crystals. The first considers long-range attractive dispersion interactions. The Maier–Saupe theory for orientational ordering in nematic phases belongs to this category. The second type of theory assumes that orientational order results from short-range steric interactions. The first example of this type of theory was the Onsager model, in which the excluded volume for rodlike particles is calculated as a function of their volume fraction. At sufficiently large volume fractions, the theory is able to predict a nematic phase.

We consider first the Maier–Saupe theory and its variants. In its original formulation, this theory assumed that orientational order in nematic liquid crystals arises from long-range dispersion forces which are weakly anisotropic [60, 61 and 62]. However, it has been pointed out [63] that the form of the Maier–Saupe potential is equivalent to one in

-17-

which there are both long-range attractive and short-range attractive contributions to the intermolecular potential. The general form of this potential is

$$U(\cos \beta) = \bar{u}_2 \bar{P}_2 P_2(\cos \beta). \quad (\text{C2.2.7})$$

This can be inserted in equation (C2.2.3) to give the orientational distribution function, and thus into equation (C2.2.6) to determine the orientational order parameters. These are determined self-consistently by variation of the interaction ‘strength’ \bar{u}_2 in equation (c2.2.7). As pointed out by de Gennes and Prost [20] it is possible to obtain the Maier–Saupe potential from a simple variational, maximum entropy method based on the lowest-order anisotropic distribution function consistent with a nematic phase.

A generalization of the Maier–Saupe theory to account for terms higher than second rank in the potential was presented by Humphries *et al* [64]. The model has also been extended to account for the orientational ordering of non-cylindrically symmetric (biaxial) nematogens [65]. Exploiting the rotational isomeric state model [66] to generate conformers, a molecular field theory for the orientational order of flexible nematogens has been developed, where the orientational ordering of each segment of a molecule is described by a second-rank tensor [67, 68]. The ability of such models to describe the orientational order of nematogens containing terminal alkyl chains has been assessed by making comparisons with order parameters extracted from NMR experiments [69, 70]. An odd–even variation of segmental orientational order parameters with the number of carbon atoms in the chain is one of the observed features that these theories can reproduce. The nematic–isotropic phase transition temperature and entropy also show an odd–even variation [67, 69].

The Maier–Saupe theory was developed to account for ordering in the smectic A phase by McMillan [71]. He allowed for the coupling of orientational order to the translational order, by introducing a translational order parameter which depends on an ensemble average of the first harmonic of the density modulation normal to the layers as well as \bar{P}_2 . This model can account for both first- and second-order nematic–smectic A phase transitions, as observed experimentally.

Turning now to theories for the nematic phase based on short-range repulsive intermolecular interactions, we consider first the Onsager model [72]. This theory has been used to describe nematic ordering in solutions of rodlike macromolecules such as tobacco mosaic virus or poly(γ -benzyl-L-glutamate). Here, the orientational distribution is calculated from the volume excluded to one hard cylinder by another. The theory assumes that the rods cannot interpenetrate. Denoting the length of rods by L and the diameter by D , it is assumed that the volume fraction $\Phi = c \frac{1}{4} \pi L D^2$ (c =concentration) is much less than unity and that the rods are very long $L \gg D$. It is found that the nematic phase exists above a volume fraction $\Phi_c = 4.5D/L$ [20]. The Onsager theory predicts jumps in density and order parameter \bar{P} at the isotropic–nematic phase transition on cooling, that are much larger than those

observed for thermotropic liquid crystals [20]. It is an athermal model so that quantities like the transition density are independent of temperature. For these reasons, it has not proved very successful for thermotropic liquid crystals, for which the (thermal) Maier–Saupe theory and its extensions are more suitable.

It has not proved possible to develop general analytical hard-core models for liquid crystals, just as for normal liquids. Instead, computer simulations have played an important role in extending our understanding of the phase behaviour of hard particles. Frenkel and Mulder found that a system of hard ellipsoids can form a nematic phase for ratios $L/D > 2.5$ (rods) or $L/D < 0.4$ (discs) [73]; however, such a system cannot form a smectic phase, as can be shown by a scaling

-18-

argument within the statistical mechanical theory [74]. However, simulations show that a smectic phase can be formed by a system of hard spherocylinders [75, 76]. The critical volume fractions for stability of a smectic A phase depend on whether the model is that of parallel spherocylinders [75], or, more realistically, freely rotating spherocylinders [74].

C2.2.3.3 CONTINUUM THEORY FOR ELASTIC PROPERTIES

An aligned monodomain of a nematic liquid crystal is characterized by a single director $\hat{\mathbf{n}}$. However, in imperfectly aligned or unaligned samples the director varies through space. The appropriate tensor order parameter to describe the director field is then

$$Q_{\alpha\beta}(\mathbf{r}) = Q(T)[\hat{n}_\alpha(\mathbf{r})\hat{n}_\beta(\mathbf{r}) - \frac{1}{3}\delta_{\alpha\beta}] \quad (\text{C2.2.8})$$

where $\alpha, \beta = 1, 2, 3$ and $\delta_{\alpha\beta}$ is the Kronecker delta function. In the continuum theory, it is assumed that $\hat{\mathbf{n}}$ varies slowly and smoothly with spatial position \mathbf{r} , so that details on a molecular scale can be neglected. This model is an extension of the elastic theory for solid bodies and was first applied to liquid crystals by Oseen [77]. The modern version of the theory is due to Frank [39] and its relationship to hydrodynamic theory has been considered [78]. Details can be found elsewhere [20, 39 and 79]. The result is that the elastic energy per unit volume has the form

$$F_{el} = \frac{1}{2}K_1[\nabla \cdot \hat{\mathbf{n}}]^2 + \frac{1}{2}K_2[\hat{\mathbf{n}} \cdot (\nabla \times \hat{\mathbf{n}})]^2 + \frac{1}{2}K_3[\hat{\mathbf{n}} \times (\nabla \times \hat{\mathbf{n}})]^2. \quad (\text{C2.2.9})$$

Here K_1, K_2 and K_3 are elastic constants. The first, K_1 is associated with a splay deformation, K_2 is associated with a twist deformation and K_3 with bend (figure C2.2.11). These three elastic constants are termed the Frank elastic constants of a nematic phase. Since they control the variation of the director orientation, they influence the scattering of light by a nematic and so can be determined from light-scattering experiments. Other techniques exploit electric or magnetic field-induced transitions in well-defined geometries (Fréedericksz transitions, see section (C2.2.4.1) [20, 80].

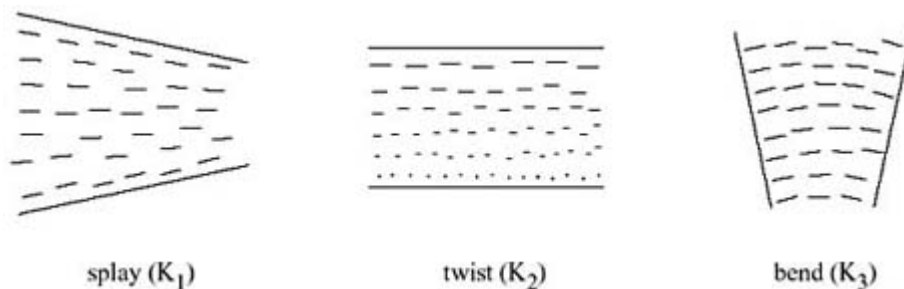


Figure C2.2.11. (a) Splay, (b) twist and (c) bend deformations in a nematic liquid crystal. The director is indicated by a dot, when normal to the page. The corresponding Frank elastic constants are indicated (equation(C2.2.9)).

Continuum theory has also been applied to analyse the dynamics of flow of nematics [77, 80, 81 and 82]. The equations provide the time-dependent velocity, director and pressure fields. These can be determined from equations for the fluid acceleration (in terms of the total stress tensor split into reversible and viscous parts), the rate of change of director in terms of the velocity gradients and the molecular field and the incompressibility condition [20].

-19-

Further details can be found elsewhere [20, 78, 82 and 84]. An approach to the dynamics of nematics based on analysis of microscopic correlation functions has also been presented [85]. Various combinations of elements of the viscosity tensor of a nematic define the so-called Leslie coefficients [20, 84].

As for crystals, the elasticity of smectic and columnar phases is analysed in terms of displacements of the lattice with respect to the undistorted state, described by the field $\mathbf{u}(\mathbf{r})$. This represents the distortion of the layers in a smectic phase and, thus, $u(\mathbf{r})$ is a one-dimensional vector (conventionally defined along z), whereas the columnar phase is two dimensional, so that $\mathbf{u}(\mathbf{r})$ is also. The symmetry of a smectic A phase leads to an elastic free energy density of the form [86]

$$F = F_0 + \frac{B}{2} \left(\frac{\partial u}{\partial z} \right)^2 + \frac{K_1}{2} \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right)^2. \quad (\text{C2.2.10})$$

Here F_0 is the free energy of the isotropic phase. As usual, the z direction is normal to the layers. Thus, two elastic constants, B (compression) and K_1 (splay), are necessary to describe the elasticity of a smectic phase [20, 79, 86]. A simple derivation of this equation based on the lowest-order derivative (curvature) of the layer displacement field $u(\mathbf{r})$ has been provided [87]. A similar expression can be obtained for a uniaxial columnar phase [20] (with the columns lying in the z direction):

$$F = F_0 + \frac{B}{2} \left(\frac{\partial u_x}{\partial x} + \frac{\partial u_y}{\partial y} \right)^2 + \frac{C}{2} \left[\left(\frac{\partial u_x}{\partial x} - \frac{\partial u_y}{\partial y} \right)^2 + \left(\frac{\partial u_x}{\partial y} + \frac{\partial u_y}{\partial x} \right)^2 \right] + \frac{K_3}{2} \left[\left(\frac{\partial^2 u_x}{\partial z^2} \right)^2 - \left(\frac{\partial^2 u_y}{\partial z^2} \right)^2 \right]. \quad (\text{C2.2.11})$$

Here B is again a compressional elastic constant, K_3 is a bend elastic constant and the elastic constant C results from an elliptical deformation of the rods (this term is absent if the column is liquid).

C2.2.3.4 THEORIES FOR PHASE TRANSITIONS

(A) NEMATIC–SMECTIC A TRANSITION

The nematic to smectic A phase transition has attracted a great deal of theoretical and experimental interest because it is the simplest example of a phase transition characterized by the development of translational order [88]. Experiments indicate that the transition can be first order or, more usually, continuous, depending on the range of stability of the nematic phase. In addition, the critical behaviour that results from a continuous transition is fascinating and allows a test of predictions of the renormalization group theory in an accessible experimental system. In fact, this transition is analogous to the transition from a normal conductor to a superconductor [89], but is more readily studied in the liquid crystal system.

When a nematic phase is cooled towards a smectic A phase, fluctuations of smectic order build up. These fluctuations were called ‘cybotactic clusters’ in the early literature. Regardless of the physical picture of such fluctuations,

it has been observed that the cluster size grows as the transition is approached [20, 90]. Furthermore, these clusters are anisotropic, being elongated along the director. They also grow faster along this direction as the transition is approached from above [20, 90].

Undoubtedly the most successful model of the nematic–smectic A phase transition is the Landau–de Gennes model [20]. It is applied in the case of a second-order phase transition by combining a Landau expansion for the free energy in terms of an order parameter for smectic layering with the elastic energy of the nematic phase [20]. It is first convenient to introduce an order parameter for the smectic structure, which allows both for the layer periodicity (at the first harmonic level, cf. equation (C2.2.1)) and the fluctuations of layer position $u\mathbf{r}$ [20]:

$$\psi(\mathbf{r}) = \rho_1(\mathbf{r}) e^{i\Phi(\mathbf{r})} \quad (\text{C2.2.12})$$

where $\Phi(\mathbf{r}) = -q_s u(\mathbf{r})$ is a phase factor.

Using this order parameter, the free energy in the nematic phase close to a transition to the smectic phase can be shown to be given by [20, 88, 89, 91]

$$F = F_0 + \frac{1}{2}A|\psi|^2 + \frac{1}{4}C|\psi|^4 + \frac{1}{6}E|\psi|^6 + C_{\parallel}|\nabla_{\parallel}|^2 + C_{\perp}|(\nabla_{\perp} - iq_s\delta\hat{\mathbf{n}})|^2 + F_N. \quad (\text{C2.2.13})$$

Here A , C and E are phenomenological coefficients in the Landau expansion in terms of the smectic ordering; C_{\parallel} and C_{\perp} account for gradients of the smectic order parameter; the fifth term also allows for director fluctuations, $\delta\hat{\mathbf{n}}$. The term F_N is the elastic free-energy density of the nematic phase, given by equation (C2.2.9). In the smectic A phase itself, the amplitude of the density modulation is constant and twist and splay distortions are forbidden, thus the expression for the free energy density simplifies to equation (C2.2.10).

High-resolution heat capacity measurements showed that the exponent for the temperature dependence of the heat capacity followed the predictions of the 3D XY model [92, 93] in systems where the nematic phase was large [94]. High-resolution x-ray scattering experiments in the nematic phase close to the continuous transition to a smectic A phase provided definitive evidence that the transition belongs to the 3D XY universality class [90]. The critical exponents obtained for the growth of correlation lengths were in excellent agreement with the renormalization group theory predictions, and provide strong support for the 3D XY model [92, 93].

(B) SMECTIC A–SMECTIC C TRANSITION

This transition is usually second order [18, 19 and 20]. The SmC phase differs from the SmA phase by a tilt of the director with respect to the layers. Thus, an appropriate order parameter contains the polar (θ) and azimuthal (ϕ) angles of the director:

$$\psi(\mathbf{r}) = \theta(\mathbf{r})e^{i\phi(\mathbf{r})}. \quad (\text{C2.2.14})$$

Obviously $\theta = 0$ corresponds to the SmA phase. This transition is analogous to the normal–superfluid transition in liquid helium and the critical behaviour is described by the XY model. Further details can be found elsewhere [18, 19 and 20].

(C) NAC POINT

A point at which nematic, SmA and SmC phases meet was demonstrated experimentally in the 1970s [95, 96]. The NAC point is an interesting example of a multicritical point because lines of continuous transition between N and

SmA phases, and SmA and SmC phases, meet the line of discontinuous transitions between the N and SmC phase. The latter transition is first order due to fluctuations of SmC order, which are continuously degenerate, being concentrated on two rings in reciprocal space rather than two points in the case of the N–SmA transition [18, 19 and 20]. Because the NAC point corresponds to the meeting of lines of continuous and discontinuous transitions it is an example of a Lifshitz point (a precise definition of this critical point is provided in [18, 19 and 20]). The NAC point and associated transitions between the three phases are described by the Chen–Lubensky model [97], which is able to account for the topology of the experimental phase diagram. In the vicinity of the NAC point, universal behaviour is predicted and observed experimentally [20].

(D) FRUSTRATED SMECTICS

Prost [21, 25, 98] showed that the properties and structure of frustrated smectic phases, as sketched in [figure C2.2.5](#) can be described by two order parameters. These are the mass density order parameter $\rho(\mathbf{r})$ ([equation \(C2.2.1\)](#)) and the polarization order parameter $P(\mathbf{r})$, which describes long-range correlations of dipoles. He then constructed a phenomenological Landau mean-field theory in which the free energy contains terms up to the quartic in these order parameters, their gradients and coupling terms. The number of terms in the free energy reflects the symmetry of the particular frustrated SmA and SmC phase under consideration. The theory has been comprehensively reviewed [18, 19, 20 and 21, 25].

(E) PHASE TRANSITIONS INVOLVING SMECTIC B PHASES

The transition from smectic A to smectic B phase is characterized by the development of a sixfold modulation of density within the smectic layers ('hexatic' ordering), which can be seen from x-ray diffraction experiments where a sixfold symmetry of diffuse scattering appears. This sixfold symmetry reflects the bond orientational order. An appropriate order parameter to describe the SmA–SmB phase transition is then [18, 19 and 20]

$$\psi_6 = \rho_6 e^{6i\phi} \quad (\text{C2.2.15})$$

where ϕ is the angle about the C_6 axis, and ρ_6 denotes a constant density. That such hexatic order could be created by dislocations was predicted by Halperin and Nelson and by Young [28, 99]. Again, high-resolution heat capacity measurements have also been useful in elucidating critical behaviour close to the transition in bulk samples [100, 101]. Calorimetry experiments on thin, freely suspended liquid crystal films have provided a great deal of information on the crossover between two- and three-dimensional behaviour at the SmA–HexB transition and they have confirmed that the transition is continuous [101, 102].

(F) PHASE TRANSITIONS IN DISCOTICS

McMillan's model [71] for transitions to and from the SmA phase ([section C2.2.3.2](#)) has been extended to columnar liquid crystal phases formed by discotic molecules [36, 103]. An order parameter that couples translational order to orientational order is again added into a modified Maier–Saupe theory, that provides the orientational order parameter. The coupling order parameter allows for the two-dimensional symmetry of the columnar phase. This theory is able to account for stable isotropic, discotic nematic and hexagonal columnar phases.

Monte Carlo computer simulations of spheres sectioned into a 'disc' [104, 105] show that steric interactions alone can produce a nematic phase of discotic molecules. Columnar phases are also observed [104, 105].

C2.2.4 APPLICATIONS OF LIQUID CRYSTALS

C2.2.4.1 NEMATIC LIQUID CRYSTAL DISPLAYS (LCDS)

The anisotropy of liquid crystal molecules leads to a susceptibility to electric and magnetic fields. Such fields can be used to change the average orientation of molecules and this is the basis for liquid crystal displays. In fact, the basic physics underlying nematic LCDs was worked out by Fréedericksz (a.k.a. Frederiks) in the 1930s [106]. It relies on the strong interactions of liquid crystal molecules with surfaces as well as their susceptibility to electromagnetic fields. Consider a nematic liquid crystal sandwiched in a thin film (about 10 μm thick) between two pieces of glass that have been treated to produce preferential orientation at the surface. In nematic liquid crystal displays, molecules are oriented parallel to the glass using a thin layer of rubbed polyimide polymer. Rubbing produces microscopic grooves in the polymer and, hence, alignment of the mesogens. To remain in an undeformed state, the bulk of the sample also adopts this orientation, which is termed homogeneous (figure C2.2.12) left). Now an electric or magnetic field is applied normal to the surface. In the bulk of the sample where the molecules are not pinned by the surface, the director tends to reorient in the direction of the field, as shown in figure C2.2.12 (right). If we compare with figure C2.2.11, we can see that this deformation involves bend and splay of the director field. This field-induced transition in director orientation is called a Fréedericksz transition [9, 106, 107]. We can also define Fréedericksz transitions when the director and field are both parallel to the surface, but mutually orthogonal or when the director is normal to the surface and the field is parallel to it. It turns out there is a threshold voltage for attaining orientation in the middle of the liquid crystal cell, i.e. a deviation of the angle of the director [9, 107]. For all three possible geometries, the threshold voltage takes the form [9, 107]

$$V_{th} = \pi \sqrt{\frac{K}{\epsilon_0 \Delta\epsilon}}. \quad (C2.2.16)$$

Here d is the thickness of the cell, K is either K_1 , K_2 or K_3 depending on the geometry (e.g. K_1 in the case of figure C2.2.12) and $\Delta\epsilon$ is the anisotropy in permittivity in the nematic liquid crystal. Note that in equation (C2.2.16) the threshold voltage, that is the relevant quantity for display operation, is independent of cell thickness.

-23-

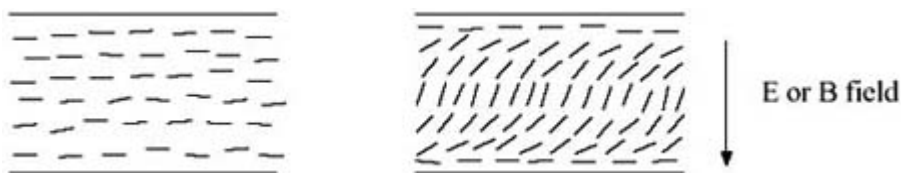


Figure C2.2.12. A Fréedericksz transition involving splay and bend. This is sometimes called a splay deformation, but only becomes purely splay in the limit of infinitesimal displacements of the director from its initial position [106]. The other two Fréedericksz geometries (‘bend’ and ‘twist’) are described in the text.

This equation is derived by accounting for the energies of the electric field and of the distorted director fields. Further details are provided in [9] and [107]. It is also possible to obtain expressions for the switching time [107], using the appropriate expressions for the hydrodynamics of nematic liquid crystals in an external field [108, 109]. Clearly, both threshold voltage and switching times are critical factors in the application of Fréedericksz transitions in LCDs. Both can be tuned by varying the cell thickness and elastic constants. The latter can be varied by appropriate choice of molecule or, in commercial devices, mixtures of molecules. It should also be noted that equation (C2.2.16) provides a means of measuring the Frank elastic constant in each of the three Fréedericksz geometries.

C2.2.4.2 TWISTED NEMATIC (TN) AND SUPERTWISTED NEMATIC (STN) LCDS

The first stable commercial liquid crystal display (LCD) device was the twisted nematic (TN) [110], still widely

used in watches and calculators. A TN display is sketched in [figure C2.2.13](#) [9, 110, 111 and 112]. It relies on the Fréedericksz transition described in the preceding section. The cell consists of two glass plates coated with rubbed polyimide to induce orientation of the director parallel to the surface. In addition, there is a thin layer of the transparent conducting material, indium tin oxide. This is used to apply an electric field across the liquid crystal sandwich, which is about 10 μm thick (controlled by spacers). The display also needs polarizers on the top and bottom plates (actually these are the most expensive part of TN displays) with their polarization axes parallel to the rubbing direction. The bottom plate is twisted with respect to the top one, so that the surface-aligned director is rotated through 90° and the polarizers are crossed. This induces a twist to the nematic phase, hence the name for the device. In this state, as light passes through the cell, its polarization axis is guided through 90° so it is transmitted through the bottom plate. In a normal device, the light is then reflected from the back plate and passes through the device again. This produces a silver or grey state. However, when an electric field is applied across the cell, the director switches to orient parallel to the field in the middle of the cell. Then light passing through the cell does not have its polarization axis rotated and so cannot be transmitted through the bottom polarizer. The cell then looks dark in the ‘on’ state. This can be used to create dark characters against a light background, as in most LCDs.

The reverse contrast, i.e. bright characters on dark, can be achieved by orientating one polarizer parallel to the rubbing direction and the other one perpendicular to it, so that the device is dark in the ‘off’ state. Backlit versions of this device are used in car dashboard displays [111].

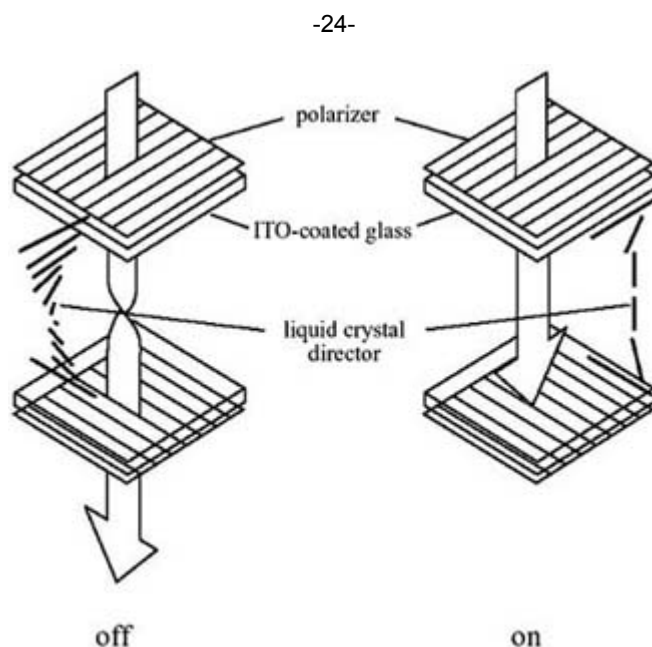


Figure C2.2.13. Principle of operation of a twisted nematic display. In the ‘off state’ the liquid crystal is in a homogeneous orientation throughout the cell. The director is oriented by rubbing the glass plates, which are placed such that the rubbing directions are parallel to the polarization direction in the adjacent polarizer. These polarizers are crossed. Light entering in the cell is rotated by the director and can pass through the cell. With a backlight or reflector, the cell appears ‘white’ (actually usually grey). However, upon application of an electric field via the indium tin oxide (ITO)-coated glass plates, the cell is switched ‘on’. The director undergoes a Fréedericksz transition to reorient parallel to the field and, thus, light entering the cell does not undergo rotation. With crossed polars the cell appears dark. This geometry describes displays with black pixels on a light background.

Important considerations for construction of a liquid crystal display include the following [111]:

- [1] *Operating conditions.* The nematic phase must be stable over the temperature range for which the device is to function, usually -20 to 80°C . At the same time, the mesogens must be rugged, i.e. chemically stable and capable of being switched many times. The development of LCDs in the 1970s was driven by the discovery of the alkylcyanobiphenyl ([figure C2.2.1](#)) class of mesogens that satisfy these requirements. No single compound, however, is able to provide the full desired operating temperature range, and in devices eutectic

mixtures are usually used.

- [2] *Threshold voltage.* Batteries can only supply low voltages, so for portable appliances the switching voltage or threshold voltage must be sufficiently small.
 - [3] *Current drawn.* LCDs usually draw small currents, which is one of their main advantages.
 - [4] *Sharpness.* This describes the steepness of the electro-optical switching as a function of voltage. This is defined in terms of the ratio of voltages required to achieve 90% compared to 10% transmission of light. This ratio should be as close to unity as possible.
-

-25-

- [5] *Contrast.* This can be quantified as the ratio of transmitted light intensity in the bright state compared to the dark state.
- [6] *Viewing angle.* The optical activity of the liquid crystal cell and polarizers depends strongly on viewing angle, leading to degradation in image quality if not viewed straight on (as confirmed by viewing a calculator display at different angles).
- [7] *Switching speed.* Typical switching times for TN devices are 20 to 50ms, which is quite slow and has limited their use in television displays.

Parameters (ii)–(vii) depend on the dielectric, mechanical and optical properties of the mesogens. To optimize a display, a compromise between different molecular characteristics is often required and mixtures of liquid crystals are usually used in commercial displays.

The desire to improve sharpness and viewing angle range led to the development of supertwisted nematic displays. As the name suggests, STN displays have higher twist angles than the TN display, typically 220–270°. They are widely used in laptop computers.

Often a seven-segment array is sufficient for TN devices where numbers are displayed [113]. The limited number of segments means that each can be addressed directly. However, dot matrix or VGA displays with large numbers of pixels are required to create alphabetical characters (not just in the Roman alphabet, but also more complex symbols such as those in C

C2.2.4.3 THIN-FILM TRANSISTOR (TFT) LCDS

Compared to STN displays, active matrix addressing in TFTs allows enhanced sharpness and greater multiplexing. In each liquid crystal pixel is addressed by a transistor, which thus primarily governs the response of the device. TFTs allow a greater number of pixels (higher resolution) and number of colour levels than STN devices. They are widely used in laptop computers, although they are more expensive than STN displays. Further details can be found elsewhere [115].

C2.2.4.4 FERROELECTRIC LCDS

Ferroelectric LCDs have potential as very fast displays and also do not require active matrix addressing technology. However, due to fabrication technology problems, they have yet to find extensive commercial application [112, 116, 117]. Thus, only the principles of operation is included here. The method relies on orientating the mesogens in a ferroelectric SmC* phase on the surfaces of the cell, but with no preferred in-plane orientation. This produces a so-called surface-stabilized ferroelectric liquid crystal (SSFLC) [112, 116, 117, 118, and 119] if the cell is sufficiently thin. In this so-called ‘bookshelf’ geometry the director vector lies perpendicular to the plane of the cell in either the ‘up’ or ‘down’ states and can be reoriented in response to an applied voltage. The reorientation of the polarization is coupled to molecular tilt and, hence, the optical axis, which can be used as an optical switch if the change in tilt angle is sufficient. Each of the two orientation states has the

same energy and so the device is bistable indefinitely. Further, bistable switching of spontaneous polarization occurs much faster than the polarization changes induced by reorientation of the director required in TN and STN display cells. However, there are a number of technical constraints that have to be overcome before SSFLC technology can be reliably used in displays: specifically, the cell has to be very thin (1 μm or less) and the director alignment and smectic layer orientation have to be controlled very carefully over large areas. Furthermore, the alignment in the optimal ‘bookshelf’ geometry can be destroyed by mechanical effects, e.g. simply by pressure of a finger, due to the high viscosity of the smectic phases, which are ‘soft solids’. In addition, there are problems with achieving a surface stabilized state, since other configurations are possible, particularly the so-called ‘chevron’ structure [120]. Further details are provided by Lagerwall [116]. Many of these problems can be solved and, indeed, prototype FLC displays have been demonstrated by several manufacturers [116].

C2.2.4.5 POLYMER DISPERSED LIQUID CRYSTAL (PDLC) DISPLAYS

A display which does not need polarizers can be made by dispersing a nematic liquid crystal in a polymer—a so-called polymer-dispersed liquid crystal (PDLC). The liquid crystals form microdroplets which scatter light in the ‘off’ state, but allow it to pass in the ‘on’ state, switched by an electric field. An example of the application of PDLCs is thus as switchable windows for privacy since the display can be switched between opaque and transparent states. For the device to function in this way, it is necessary for the director in the liquid crystal droplet to be oriented in a tangential orientation, which leads to two poles (bipolar droplet [121]). In addition, the refractive index of the polymer must be equal to the index of refraction for light polarized perpendicular to the director. In the off state, the two poles are oriented at random, but in an electric field they orient along the direction of the field (figure C2.2.14). In the off state, there is on average a difference in the refractive index of the liquid crystal droplets compared to the polymer, which leads to light scattering, whereas in the on state, the refractive indices are matched for light incident along the field direction and the display appears clear [9, 114, 121]. Thus, PDLCs are useful as switchable windows, e.g. for privacy or sunlight shading.

There are two basic methods of preparing PDLCs [121]. In the first, the liquid crystal is dispersed as an emulsion in an aqueous solution of a film-forming polymer (often polyvinyl alcohol) [121]. This emulsion is then coated onto a conductive substrate and the emulsion is dried to form the dispersed liquid crystal-in-polymer film. In the second method, the phase separation of a polymer is exploited to disperse the liquid crystal [123]. In the method of polymerization-induced phase separation [121], polymerization is induced through the application of heat, light or radiation e.g. through cross-linking of a network. A commonly used example exploits the cross-linking of epoxy adhesives to form a solid structure containing phase-separated liquid crystal droplets. In thermally induced phase separation, the liquid crystal is mixed with a thermoplastic polymer at high temperatures. When the system is cooled, the liquid crystal phase separates from the solidifying polymer. In solvent-induced phase separation, a polymer and a liquid crystal are mixed to form a single-phase mixture in an organic solvent. Evaporation of the solvent then drives the phase separation of polymer and liquid crystal [121].

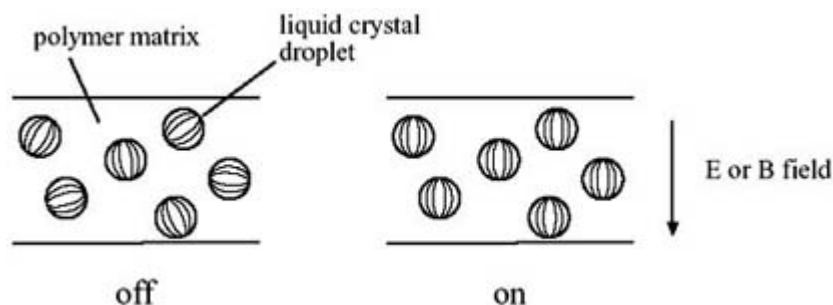


Figure C2.2.14. Principle of operation of a polymer-dispersed liquid crystal display. The contours of the liquid

crystal droplets in the polymer matrix correspond to the director orientation, which here is dipolar. In the off state, the cell scatters light and appears opaque, due to refractive index variations between the liquid crystal and the polymer. However, when an electric field is applied and the liquid crystal directors reorient, the refractive index along the field is matched to that of the matrix, and the cell becomes clear.

Further details of PDLCs can be found in the excellent monograph by Drzaic [121]. A review of the non-linear optical properties of PDLCs has also been presented [124].

C2.2.4.6 OTHER APPLICATIONS

The main non-display applications of liquid crystals can be subdivided into two classes. The first exploits their anisotropic optical properties in spatial light modulators [125] or their non-linear optical properties (optical wave mixing etc) [48, 124, 125]. Spatial light modulators are usually based on the ferroelectric SmC* phase aligned in a thin film. Liquid crystal spatial light modulators find advanced applications such as the storage of computer-generated holograms [125]. The second class of non-display applications exploits temperature-dependent colour (thermochromic) changes in the chiral nematic phase [9, 126]. Chiral nematic phases can appear coloured due to scattering of light by the helical structure, which can have a pitch as small as 100 nm. The pitch 'unwinds' as temperature is decreased, usually as a smectic A phase is approached, leading to observable colour changes. These have been exploited in medical thermography, where heat variations across the body surface are mapped [126, 127]. This is especially important in oncology. The technique has also found applications in engineering and aerodynamic research. Models of aircraft, for example, in wind tunnels can be coated with a chiral nematic liquid crystal. The flow of air over the model leads to heat variations (turbulent flow leading to 'cold spots') which can be visualized directly [126]. Gimmick applications of thermochromic liquid crystals include colour-changing clothes or beer mats.

ACKNOWLEDGMENTS

This review is dedicated to Professor G R Luckhurst (University of Southampton) on the occasion of his 60th birthday. He is also thanked for comments on a draft version of this chapter.

REFERENCES

- [1] Reinitzer F 1888 Beiträge zur Kenntnis des cholesterins *Monatsh. Chem.* **9** 421–41
Translated into English in 1998 *Liq. Cryst.* **5** 7–18 (part of a series of volumes associated with the 1988 *International Liquid Crystal Conference* commemorating the centennial of their discovery)
- [2] Lehmann O 1889 Über fließende krystalle *Z. Phys. Chem.* **4** 462–7
- [3] Templer R and Attard G 1991 The fourth state of matter *New Scientist* **130** 25–9
- [4] Demus D, Goodby J, Gray G W, Spiess H-W and Vill V (eds) 1998 *Handbook of Liquid Crystals: Vol 1. Fundamentals* (New York: Wiley-VCH)
- [5] Demus D, Goodby J, Gray G W, Spiess H-W and Vill V (eds) 1998 *Handbook of Liquid Crystals: Vol 2A. Low Molecular Weight Liquid Crystals I* (New York: Wiley-VCH)
- [6] Demus D, Goodby J, Gray G W, Spiess H-W and Vill V (eds) 1998 *Handbook of Liquid Crystals: Vol 2B. Low Molecular Weight Liquid Crystals II* (New York: Wiley-VCH)
- [7] Goodby J W and Gray G W 1998 Guide to the nomenclature and classification of liquid crystals *Handbook of Liquid*

Crystals: Vol 1. Fundamentals ed D Demus, J Goodby, G W Gray, H-W Spiess and V Vill (New York: Wiley-VCH)

- [8] Gray G W 1990 Low-molar-mass thermotropic liquid crystals *Phil. Trans. R. Soc. A* **330** 73
- [9] Collings P J and Hird M 1997 *Introduction to Liquid Crystals* (London: Taylor and Francis)
- [10] Diele S, Brand P and Sackmann H 1972 X-ray diffraction and polymorphism of smectic liquid crystals. II. D and E modifications *Mol. Cryst. Liq. Cryst.* **17** 163–9
- [11] Peierls R 1935 Quelques propriétés typiques des corps solides *Ann. Inst. Henri Poincaré* **5** 177–222
- [12] Landau L D 1937 Zur Theorie der Phasenumwandlungen II *Phys. Z. Sowjet Union* **11** 545–65
English translation in D ter Haar (ed) 1965 *The Collected Papers of LD Landau* (Oxford: Pergamon) pp 210–211
- [13] Als-Nielsen J, Litster J D, Birgeneau R J, Kaplan M and Safinya C R 1980 Lower marginal dimensionality. X-ray scattering from the smectic-A phase of liquid crystals *Ordering in Strongly Fluctuating Condensed Matter Systems* ed T Riste (New York: Plenum)
- [14] Levelut A M, Tarento R J, Hardouin F, Achard M F and Sigaud G 1981 Number of S_A phases *Phys.Rev. A* **24** 2180–6
- [15] Hardouin F, Levelut A M, Achard M F and Sigaud G 1983 Polymorphism in polar mesogens. I—Physico-chemistry and structural aspects *J Chem.Phys.* **80** 53
- [16] Shashidhar R and Ratna B R 1989 Phase-transitions and critical phenomena in polar smectic-A liquid-crystals—plenary lecture *Liq. Cryst.* **5** 421–42
- [17] Sigaud G, Hardouin F and Achard M F 1979 A possible polar smectic A–non-polar smectic A transition line in a binary system *Phys.Lett. A* **72** 24
- [18] Barois P 1992 Phase transitions in liquid crystals: introduction to phase transition theories *Phase Transitions in Liquid Crystals* ed S Martellucci and A N Chester (New York: Plenum)

-29-

- [19] Barois P 1998 Phase Transitions *Handbook of Liquid Crystals: Vol 1. Fundamentals* ed D Demus, J Goodby, G W Gray, H-W Spiess and V Vill (New York: Wiley-VCH)
- [20] de Gennes P G and Prost J 1993 *The Physics of Liquid Crystals* (Oxford: Oxford University Press)
- [21] Prost J 1984 The Smectic State *Adv. Phys.* **33** 1–46
- [22] Seddon J M 1998 Structural studies of liquid crystals by x-ray diffraction *Handbook of Liquid Crystals: Vol 1. Fundamentals* ed D Demus, J Goodby, G W Gray, H-W Spiess and V Vill (New York: Wiley-VCH)
- [23] Prost J and Barois P 1983 Polymorphism in polar mesogens. II—Theoretical Aspects *J.Chem.Phys* **80** 65–81
- [24] Hardouin F and Levelut A M 1980 X-ray study of reentrant polymorphism $N-S_A-N-S_A$ in a pure liquid crystal compound *J.Physique* **41** 41–56
- [25] de Gennes P G and Prost J 1993 *The Physics of Liquid Crystals* (Oxford: Oxford University Press) ch 10
- [26] Pershan P S 1988 *Structure of Liquid Crystal Phases* (Singapore: World Scientific)
- [27] Goodby J W 1998 Phase structures of calamitic liquid crystals *Handbook of Liquid Crystals: Vol 2A. Low Molecular Weight Liquid Crystals I* ed D Demus, J Goodby, G W Gray, H-W Spiess and V Vill 1998 (New York: Wiley-VCH)

- [28] Halperin B I and Nelson D R 1978 Theory of two-dimensional melting *Phys.Rev.Lett* **41** 121–4 Erratum: **41** 2514
- [29] Birgeneau R J and Litster J D 1978 Bond-orientational order model for smectic B liquid crystals *J. Physique Lett.* **39** 399–402
- [30] Leadbetter A J, Frost J C and Mazid M A 1979 Interlayer correlations in smectic B phases *J. Physique Lett.* **40** 325–9
- [31] Pindak R, Moncton D E, Davey S C and Goodby J W X-ray observation of a stacked hexatic liquid-crystal B phase *Phys.Rev.Lett* **46** 1135–8
- [32] Goodby J W 1998 Symmetry and chirality in liquid crystals *Handbook of Liquid Crystals: Vol 1. Fundamentals* ed D Demus, J Goodby, G W Gray, H-W Spiess and V Vill (New York: Wiley–VCH)
- [33] Chandrasekhar S, Sadashiva B K and Suresh K A 1977 Liquid crystals of disc-like molecules *Pramana* **9** 471–80
- [34] Cammidge A N and Bushby R J 1998 Synthesis and structural features *Handbook of Liquid Crystals: Vol 2B. Low Molecular Weight Liquid Crystals I* ed D Demus, J Goodby, G W Gray, H-W Spiess and V Vill (New York: Wiley–VCH)
- [35] Levelut A M 1983 Structure des phases mésomorphes formées de molécules discoïdes *J. Chim. Phys.* **80** 149–61
- [36] Chandrasekhar S 1998 Columnar, discotic, nematic and lamellar liquid crystals: Their structures and physical properties *Handbook of Liquid Crystals: Vol 2B. Low Molecular Weight Liquid Crystals I* ed D Demus, J Goodby, G W Gray, H-W Spiess and V Vill (New York: Wiley–VCH)
- [37] Demus D and Richter L 1978 *Textures of Liquid Crystals*(Weinheim: Chemie)
- [38] Gray G W and Goodby J G W 1984 *Smectic Liquid Crystals:Textures and Structures* (Glasgow: Hill)
- [39] Frank F C 1958 On the theory of liquid crystals *Discuss.Faraday Soc.* **25** 19–28
- [40] Friedel G 1922 Les états mésomorphes de la matière *Ann. Phys., Paris* **19** 273–474
- [41] Dunmur D and Toriyama K 1998 Elastic properties *Handbook of Liquid Crystals: Vol 1. Fundamentals* ed D Demus, J Goodby, G W Gray, H-W Spiess and V Vill (New York: Wiley–VCH)

- [42] Chandrasekhar S and Ranganath G S 1986 The structure and energetics of defects in liquid crystals *Adv. Phys.* **35** 507–96
- [43] Bouligand Y 1998 Defects and textures *Handbook of Liquid Crystals: Vol 1. Fundamentals* ed D Demus, J Goodby, G W Gray, H-W Spiess and V Vill (New York: Wiley–VCH)
- [44] Friedel G and Grandjean F 1910 *Bull. Soc. Franc. Miner.* **33** 192
Friedel G and Grandjean F 1910 *Bull. Soc. Franc. Miner.* **33** 402
- [45] Gleeson H F 1998 Light scattering from liquid crystals *Handbook of Liquid Crystals: Vol 1. Fundamentals* ed D Demus, J Goodby, G W Gray, H-W Spiess and V Vill (New York: Wiley–VCH)
- [46] Orsay Liquid Crystal Group 1970 Theory of light scattering by nematics *Liquid Crystals and Ordered Fluids* vol1, ed J F Johnson and R S Porter (New York: Plenum)

- [47] Jen S, Clark N A, Pershan P S and Priestley E B 1977 Polarized Raman scattering of orientational order in uniaxial liquid crystalline phases *J. Chem. Phys.* **66** 4635–61
- [48] Khoo I-C 1995 *Liquid Crystals. Physical Properties and Nonlinear Optical Phenomena* (New York: Wiley) ch 5
- [49] Schaetzing R and Litster J D 1979 Light scattering studies of liquid crystals *Advances in Liquid Crystals* vol4 (London: Academic)
- [50] Gleeson H F 1998 Brillouin scattering from liquid crystals *Handbook of Liquid Crystals: Vol 1. Fundamentals* ed D Demus, J Goodby, G W Gray, H-W Spiess and V Vill (New York: Wiley-VCH)
- [51] Leadbetter A J 1979 Structural studies of nematic, smectic A and smectic C phases *The Molecular Physics of Liquid Crystals* ed G R Luckhurst and G W Gray (London: Academic)
- [52] Leadbetter A J and Norris E K 1979 Distribution functions in three liquid crystals from x-ray diffraction measurements *Molec. Phys.* **38** 669–86
- [53] Hamley I W, Garnett S, Luckhurst G R, Roskilly S J, Pedersen J S, Richardson R M and Seddon J M 1996 Orientational ordering in the nematic phase of a thermotropic liquid crystal: A small angle neutron scattering study *J. Chem. Phys.* **104** 10 046–54
- [54] Richardson R M 1998 Neutron scattering *Handbook of Liquid Crystals: Vol 1. Fundamentals* ed D Demus, J Goodby, G W Gray, H-W Spiess and V Vill (New York: Wiley-VCH)
- [55] de Gennes P G 1972 Remarques sur la diffusion des rayons X par les fluides nematiques *C.R. Acad. Sci. Paris* **274** 142–4
- [56] Schmidt C and Spiess H W 1998 Magnetic resonance *Handbook of Liquid Crystals: Vol 1. Fundamentals* ed D Demus, J Goodby, G W Gray, H-W Spiess and V Vill (New York: Wiley-VCH)
- [57] Emsley J W (ed) 1985 *Nuclear Magnetic Resonance of Liquid Crystals* (Dordrecht: Reidel)
- [58] Dong R Y 1997 *NMR of Liquid Crystals* (New York: Springer)
- [59] Thoen J 1998 Thermal methods *Handbook of Liquid Crystals: Vol 1. Fundamentals* ed D Demus, J Goodby, G W Gray, H-W Spiess and V Vill (New York: Wiley-VCH)
- [60] Maier W and Saupe A 1958 Eine einfache molekulare theorie des nematischen kristallinflüssigen zustandes *Z. Naturf. A* **13** 564–66
- [61] Maier W and Saupe A 1959 Eine einfache molekular-statische theorie der nematische kristallinflüssigen phase. Teil I *Z. Naturf. A* **14** 882–9

- [62] Maier W and Saupe A 1960 Eine einfache molekular-statische theorie der nematische kristallinflüssigen phase. Teil II *Z. Naturf. A* **15** 287–92
- [63] Luckhurst G R and Zannoni C 1977 Why is the Maier–Saupe theory of nematic liquid crystals so successful? *Nature* **267** 412–14
- [64] Humphries R L, James P G and Luckhurst G R 1972 Molecular field treatment of nematic liquid crystals *J. Chem. Soc. Faraday Trans. II* **68** 1031–44
- [65] Luckhurst G R, Zannoni C, Nordio P L and Segre U 1975 A molecular field theory for uniaxial nematic

- liquid crystals formed by non-cylindrically symmetric molecules *Molec. Phys.* **30** 1345–58
- [66] Flory P J 1969 *Statistical Mechanics of Chain Molecules* (New York: Interscience)
- [67] Marelja S 1974 Chain ordering in liquid crystals. I. Even–odd effect *J. Chem. Phys.* **60** 3599–604
- [68] Emsley J W, Luckhurst G R and Stockley C P 1982 A theory of orientational ordering in uniaxial liquid-crystals composed of molecules with alkyl chains *Proc. R. Soc. A* **381** 117–28
- [69] Luckhurst G R 1985 Molecular field theories of nematics: systems composed of uniaxial, biaxial or flexible molecules *Nuclear Magnetic Resonance of Liquid Crystals* ed J W Emsley (Dordrecht: Reidel)
- [70] Zannoni C 1985 An internal order parameter formalism for non-rigid molecules *Nuclear Magnetic Resonance of Liquid Crystals* ed J W Emsley (Dordrecht: Reidel)
- [71] McMillan W L 1971 Simple molecular model for the smectic A phase of liquid crystals *Phys. Rev. A* **4** 1238–46
- [72] Onsager L 1949 The effects of shape on the interaction of colloidal particles *Ann. N. Y. Acad. Sci.* **51** 627–59
- [73] Frenkel D and Mulder B 1985 The hard ellipsoid-of-revolution fluid. 1. Monte-Carlo simulations *Mol. Phys.* **55** 1171–92
- [74] Frenkel D 1992 Computer simulations of phase transitions in liquid crystals *Phase Transitions in Liquid Crystals* ed S Martellucci and A N Chester (New York: Plenum)
- [75] Stroobants A, Lekkerkerker H N W and Frenkel D 1986 Evidence for smectic order in a fluid of hard parallel spherocylinders *Phys. Rev. Lett.* **57** 1452–5; *Erratum* **57** 2331
- [76] Frenkel D 1988 Thermodynamic stability of a smectic phase in a system of hard rods *Nature* **332** 822–3
- [77] Oseen C W 1929 The theory of liquid crystals *Trans. Faraday Soc.* **29** 883–99
- [78] Ericksen J L 1966 Some magnetohydrodynamic effects in liquid crystals *Arch. Ration. Mech. Analysis* **23** 266–75
- [79] Landau L D and Lifshitz E M 1986 *Theory of Elasticity* (Oxford: Pergamon) chVI
- [80] Stannarius R 1998 Elastic properties of nematic liquid crystals 1998 *Handbook of Liquid Crystals: Vol 2A. Low Molecular Weight Liquid Crystals I* ed D Demus, J Goodby, G W Gray, H-W Speiss and V Vill (New York: Wiley–VCH)
- [81] Ericksen J L 1976 Equilibrium theory of liquid crystals *Adv. Liq. Cryst.* **2** 233
- [82] Leslie F M 1968 Some constitutive equations for liquid crystals *Arch. Ration. Mech. Analysis* **28** 265–83
- [83] Parodi O 1970 Stress tensor for a nematic liquid crystal *J. Physique* **31** 581–4

- [84] Leslie F M 1998 Continuum theory for liquid crystals *Handbook of Liquid Crystals: Vol 1. Fundamentals* ed D Demus, J Goodby, G W Gray, H-W Speiss and V Vill (New York: Wiley–VCH)
- [85] Martin P C, Parodi O and Pershan P S 1972 Unified hydrodynamic theory for crystals, liquid crystals and normal fluids *Phys. Rev. A* **6** 2401–20

- [86] de Gennes P G 1969 Conjectures sur l'état smectique" *J.Physique Coll. C* **4** 65–71
- [87] de Gennes P G and Prost J 1993 *The Physics of Liquid Crystals* (Oxford: Oxford University Press) ch1
- [88] Litster J D and Birgeneau R J 1982 Phases and phase transitions *Phys. Today* **35** 26–33
- [89] de Gennes P G 1972 An analogy between superconductors and smectics A *Solid State Commum* **10** 753–6
- [90] Bouwman W G and de Jeu W H 1992 3D XY behavior of a nematic–smectic A phase transition: confirmation of the de Gennes model *Phys.Rev.Lett* **68** 800–3
- [91] de Gennes P G 1973 Some remarks on the polymorphism of smectics *Molec. Cryst. Liq. Cryst.* **21** 49–76
- [92] Lubensky T C 1983 The nematic to smectic A transition: a theoretical overview *J. Chim. Phys.* **80** 31–43
- [93] Le Guillou J C and Zinn-Justin J 1985 Accurate critical exponents from the epsilon-expansion *J. Physique. Lett.* **46** 137
- [94] Garland C W, Nounesis G and Stine J J 1989 XY behavior for the heat-capacity at nematic–smectic–A₁ liquid-crystal transitions *Phys.Rev. A* **39** 4919–22
- [95] Sigaud G, Hardouin F and Achard M F 1977 An experimental system for a nematic–smectic A–smectic C Lifshitz's point *Solid Saste Commum* **23** 35–6
- [96] Johnson D, Allender D, Dehoff D, Maze C, Oppenheim E and Reynolds R 1977 Nematic–smectic A–smectic C polycritical point: Experimental evidence and a Landau theory *Phys.Revs B* **16** 470–5
- [97] Chen J-H and Lubensky T C 1976 Landau–Ginzburg mean-field theory for the nematic to smectic C and nematic to smectic A phase transitions *Phys.Rev. A* **14** 1202–7
- [98] Prost J 1979 Smectic A to smectic A phase transition *J.Physique* **40** 581–7
- [99] Young A P 1979 Melting and the vector Coulomb gas in two dimensions *Phys.Rev. B* **19** 1855–66
- [100] Huang C C, Viner J M, Pindak R and Goodby J W 1981 Heat-capacity study of the transition from a stacked-hexatic-B phase to a smectic-A phase 1981 *Phys.Rev.Lett* **46** 1289–92
- [101] Huang C C 1998 Physical properties of non-chiral smectic liquid crystals *Handbook of Liquid Crystals: Vol 2A. Low Molecular Weight Liquid Crystals I* ed D Demus, J Goodby, G W Gray, H-W Speiss and V Vill (New York: Wiley–VCH)
- [102] Geer R, Huang C C, Pindak R and Goodby J W 1989 Heat capacity anomaly from 4-layer liquid-crystal films *Phys.Rev.Lett* **63** 540–3
- [103] Chandrasekhar S 1983 Liquid crystals of disc-like molecules *Phil. Trans. R. Soc. A* **309** 93–103
- [104] Frenkel D 1989 Columnar ordering as an excluded-volume effect *Liq. Cryst.* **5** 929–40
- [105] Veerman J A C and Frenkel D 1992 Phase-behavior of disk-like hard-core mesogens *Phys.Rev. A* **45** 5632–48

- [106] Fréedericksz V and Zolina V 1933 Forces causing the orientation of an anisotropic fluid *Trans. Faraday*

- [107] Blinov L M 1983 *Electro-optical and Magneto-optical Properties of Liquid Crystals* (Chichester:Wiley)
- [108] Brochard F, Pieranski P and Guyon E 1972 Dynamics of the orientation of a nematic-liquid-crystal film in a variable magnetic field *Phys.Rev.Lett* **28** 1681–3
- [109] Pieranski P, Brochard F and Guyon E 1973 Static and dynamic behavior of a nematic liquid crystal in a magnetic field. Part II: Dynamics *J.Physique* **34** 35–48
- [110] Schadt M and Helfrich W 1971 Voltage-dependent optical activity of a twisted nematic liquid crystal *Appl. Phys. Lett.* **18** 127–8
- [111] Hirschmann H and Reiffenrath V 1998 TN, STN displays *Handbook of Liquid Crystals: Vol 2A. Low Molecular Weight Liquid Crystals I* ed D Demus, J Goodby, G W Gray, H-W Speiss and V Vill (New York: Wiley–VCH)
- [112] Sage I 1998 Displays *Handbook of Liquid Crystals: Vol 1. Fundamentals* ed D Demus, J Goodby, G W Gray, H-W Speiss and V Vill (New York: Wiley–VCH)
- [113] Scheffer T J, Nehring J, Kaufmann M, Amstutz H, Heimgartner D and Eglin P 1985 24 × 80 character LCD panel using the supertwisted birefringence effect *Dig. Tech. Papers Int. Symp. Soc. Information Display* **16** 120–3
- [114] Collings P J 1990 *Liquid Crystals. Nature's Delicate Phase of Matter* (Princeton: Princeton University Press)
- [115] Kaneko E Active matrix addressed displays *Handbook of Liquid Crystals: Vol 2A. Low Molecular Weight Liquid Crystals I* ed D Demus, J Goodby, G W Gray, H-W Speiss and V Vill (New York: Wiley–VCH)
- [116] Lagerwall S T 1998 Ferroelectric liquid crystals *Handbook of Liquid Crystals: Vol 2B. Low Molecular Weight Liquid Crystals II* ed D Demus, J Goodby, G W Gray, H-W Speiss and V Vill (New York: Wiley–VCH)
- [117] Goodby J W, Blinc R, Clark N A, Lagerwall S T, Osipov M A, Pikin S A, Sakurai T, Yoshino K and Zeks B 1991 *Ferroelectric Liquid Crystals Principles, Properties and Applications* (Philadelphia: Gordon and Breach)
- [118] Clark N A and Lagerwall S T 1980 Submicrosecond bistable electro-optic switching in liquid crystals *Appl. Phys. Lett.* **36** 899–901
- [119] Clark N A, Handschy M A and Lagerwall S T 1983 Ferroelectric liquid crystal electro-optics using the surface stabilized structure *Molec. Cryst. Liq. Cryst.* **94** 213–34
- [120] Rieker T P, Clark N A, Smith G S, Parmar D S, Sirota E B and Safinya C R 1987 'Chevron' local layer structure in surface-stabilized ferroelectric smectic-C cells *Phys.Rev.Lett* **59** 2658–61
- [121] Drzaic P S 1995 *Liquid Crystal Dispersions* (Singapore: World Scientific)
- [122] Ferguson J L 1984 Encapsulated liquid crystal and method *US Patent* 4 435 047
- [123] Doane J W 1986 Field controlled light scattering from nematic micro droplets *Appl. Phys. Lett.* **48** 269–71
- [124] Simoni F 1997 *Nonlinear Optical Properties of Liquid Crystals and Polymer-Dispersed Liquid Crystals* (Singapore: World Scientific)
- [125] Crossland W A and Wilkinson T D 1998 Nondisplay applications of liquid crystals *Handbook of Liquid Crystals: Vol 1. Fundamentals* ed D Demus, J Goodby, G W Gray, H-W Speiss and V Vill (New York: Wiley–VCH)
- [126] Gleeson H F 1998 Thermography using liquid crystals *Handbook of Liquid Crystals: Vol 1. Fundamentals* ed D Demus, J Goodby, G W Gray, H-W Speiss and V Vill (New York: Wiley–VCH)

[127] Gautherie M 1969 Application des cristaux liquides cholestériques a la thermographie cutanée *J. Physique. Coll. C* 4 122–6

FURTHER READING

Chandrasekhar S 1992 *Liquid Crystals* (Cambridge: Cambridge University Press)

Collings P J 1990 *Liquid Crystals. Nature's Delicate Phase of Matter* (Princeton: Princeton University Press)

Collings P J and Hird M 1997 *Introduction to Liquid Crystals* (London: Taylor and Francis)

Demus D, Goodby J, Gray G W, Spiess H-W and Vill V (eds) 1998 *Handbook of Liquid Crystals* 4 vols (Weinheim: Wiley–VCH)

Luckhurst G R and Gray G W (eds) 1979 *The Molecular Physics of Liquid Crystals* (London: Academic)

de Gennes P G and Prost J 1993 *The Physics of Liquid Crystals* 2nd edn (Oxford: Oxford University Press)

Vertogen G and de Jeu W H 1988 *Thermotropic Liquid Crystals, Fundamentals* (Heidelberg: Springer).

-1-

C2.3 Micelles

John Texter

C2.3.1 INTRODUCTION

Surfactants are the primary molecular constituents of micelles. They are also called *amphiphiles*, and certain classes of surfactants are detergents. Surfactants are amphiphilic molecules having separate lyophilic or solvophilic (solvent-loving) groups and lyophobic or solvophobic (solvent-hating) groups (see [section C2.3.3](#)). Having both types of group makes a molecule *amphiphilic* and *amphipathic*. Micelles are the most prevalent aggregate structure in surfactant solutions and form over a narrow range in surfactant concentration centred around the critical micelle concentration, *cmc* (see [section C2.3.4](#)). This process of micellization is taken, in the so-called pseudophase approximation, as the formation of a two-phase system comprising the continuous solvent phase and a pseudophase consisting of the oily micellar cores. This approximation is convenient for many purposes, but it must not be taken literally, as micellar solutions are generally single-phase isotropic solutions and micelles are thermodynamically reversible aggregates, in dynamic equilibrium with surfactant ‘monomers’ in solution.

Idealized structures of some normal and reverse (inverse) micelles are illustrated in figure C2.3.1. Suctants self-assemble (aggregate) in order to lower the solution free energy. This process involves the creation of an interface separating the solvent (aqueous) phase from the solvophobic (hydrophobic) portions of the surfactant. When a solvophilic headgroup interacts attractively with a solvent, the solvophobic portions aggregate to form an oily nanodroplet of interpenetrating tail portions that is separated from the solvent by the headgroups. The headgroups serve to define a boundary between the solvent and solvophobic portion. Whether headgroups are charged or not,

their packing is influenced by repulsion forces between headgroups and solvation interactions with the solvent. Such aggregates define 'normal' micelles. When the surfactant headgroup is not solvated by the solvent, the headgroups tend to aggregate together, forming an internal structure separated from the solvent by an interface defined by the tailgroups. Such micelles are called inverse micelles or reverse micelles.

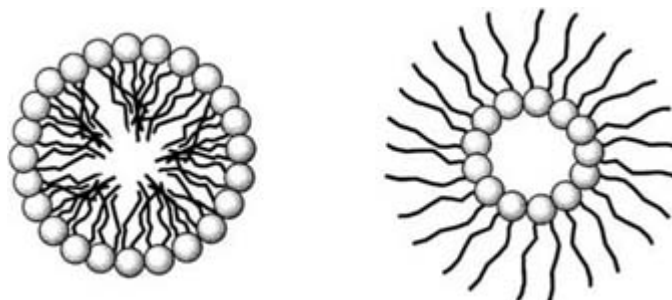


Figure C2.3.1. Idealized cross-sections of normal spherical micelles (left) and of reverse micelles (right).

-2-

C2.3.2 HISTORICAL OVERVIEW

Aqueous surfactant aggregation to form micelles appears to have first been suggested by McBain in 1913 [1]. A spherical model for micelle structure was put forward by Hartley [2, 3] where the hydrophilic headgroups were pictured to lie upon a roughly spherical surface, the internal volume was filled with the tailgroups and an electrical double layer developed radially out from the headgroups. The tailgroups form an oily nanodroplet that is covered with a polar shell. This early, and admittedly, idealized model was criticized on various grounds [4], although Hartley pointedly noted that the assumption of perfectly spherical symmetry was done for 'mathematical simplicity'. Menger proposed [4] taking more detailed account of the filling of the internal volume by the tailgroups and he explicitly introduced the use of *gauche* kinks in the hydrocarbon chains to help fill the requisite space.

This more careful attention to space filling was also addressed by Fromherz [5] and by Dill and Flory [6]. Fromherz showed that it was possible to produce space filling and nominally spherical micellar structures, with rough headgroup regions, while still allowing straight chain surfactants to aggregate with extended tailgroups, but allowing *gauche* defects to modify the orientation of headgroups. This model, along with that of Menger and with that of Dill and Flory, resulted in rough headgroup regions and significant contacts between the solvent and regions of the tailgroups. These approaches were put on a statistical dynamical basis by Dill and Flory using a lattice model illustrated in figure C2.3.2. In this lattice model the surfactant is imagined to be composed of a headgroup and tailgroup chain segments. Each of the headgroup and segments occupies a single lattice site, as illustrated in figure C2.3.2. The lattice comprises constant radial interlayer spacing and sites of equal volume. This approach set the stage for advances in structural modelling using the modern computational tools of statistical mechanics (see [section C2.3.7](#)).



Figure C2.3.2. Two-dimensional radial lattice representation of micelle structure using the approach of Dill and Flory [6]. Each lattice site is considered to be equal in volume to the others. Reproduced by permission from [6].

-3-

C2.3.3 SURFACTANTS

Surfactants are generally pictured as having a lyophilic (hydrophilic) headgroup of some type and a hydrocarbon tailgroup. The word surfactant derives as a contraction of the phrase *surface active agent*. They can also be exemplified by simple species such as short chain alcohols and amides that are usually surface active, but do not necessarily lower interfacial tensions significantly. It is difficult to draw clear boundaries with respect to molecular structure for defining what constitutes a surfactant, so we adopt an operational definition. A surfactant is a surface active amphiphile that aggregates in water or other solvent to form microstructures such as micelles and bilayers or segregates at interfaces to form monolayer assemblies. This definition is inclusive of many important polymeric dispersants, such as moderate molecular weight block copolymeric surfactants.

C2.3.3.1 CLASSIFICATION OF SURFACTANTS

Schemes for classifying surfactants are based upon physical properties or upon functionality. Charge is the most prevalent physical property used in classifying surfactants. Surfactants are charged or uncharged, ionic or nonionic. Charged surfactants are further classified as to whether the amphipathic portion is anionic, cationic or zwitterionic. Another physical classification scheme is based upon overall size and molecular weight. Copolymeric nonionic surfactants may reach sizes corresponding to 10 000–20 000 Daltons. Physical state is another important physical property, as surfactants may be obtained as crystalline solids, amorphous pastes or liquids under standard conditions. The number of tailgroups in a surfactant has recently become an important parameter. Many surfactants have either one or two hydrocarbon tailgroups, and recent advances in surfactant science include even more complex assemblies [7, 8 and 9].

Surfactants derive their general classification from their surface activity and tendency to preferentially segregate at liquid–gas, liquid–liquid and liquid–solid interfaces. They always contain at least two functional parts, a lyophilic or solvophilic part that preferentially solvates in the solvent and a lyophobic or solvophobic part that is poorly solvated in the same solvent. Solvophilicity (hydrophilicity in the case of water) is imparted by functional groups that have high solvent affinity. Carboxylates, sulphates and sulphonates are examples of charged functional groups that have relatively high affinity for water as a solvent. Uncharged hydrophilic groups may include almost any uncharged polar group. Hydroxyl and ethylene oxide groups are the most prevalent.

Solvophobicity (hydrophobicity with respect to water) is most often exemplified as a linear or branched hydrocarbon chain. Fluorocarbon chains and siloxane chains are also hydrophobic. Many commercially important

surfactants have more complex solvophobic groups such as substituted phenyl and naphthyl ring systems. The number of solvophobic tailgroups (single tail, double tail) in a surfactant is an important parameter because it dramatically affects solubility and surfactant packing in micellar aggregates (see [section C2.3.6](#)).

The majority of practical micellar systems of 'normal' micelles use water as the main solvent. Reverse micelles use water immiscible organic solvents, although the cores of reverse micelles are usually hydrated and may contain considerable quantities of water. Polar solvents such as glycerol, ethylene glycol, formamide and hydrazine are now being used instead of water to support 'regular' micelles [10]. Critical fluids such as critical carbon dioxide are

-4-

also being exploited for various micellar applications. Copolymeric surfactants of polystyrene and poly(1,1-dihydroperfluorooctyl acrylate) have been developed [11] to form polydisperse micelles in critical carbon dioxide and to solubilize polystyrene oligomers in such micellar solutions.

Charged surfactants impart electrostatic forces and ion-pairing interactions when they are aggregated as micelles. Charged surfactant solubility is greatly affected by counter-ions and the binding affinity of counter-ions. For example, lithium surfactant salts are much more hygroscopic than sodium salts and tend to have greater aqueous solubility. In such cases counter-ion *charge density* has a dramatic effect on binding affinity and, hence, on solubility. This ion-ion association also affects the formation and structure of micelles, since headgroup repulsion interactions are dramatically affected by counter-ion binding and the screening of repulsive electrostatic forces by counter-ions. An extensive practical listing of surfactants, as well as a key to their synthesis, may be found in an exhaustive review [12].

(A) ANIONIC

A selection of important anionic surfactants is displayed in [table C2.3.1](#). Carboxylic acid salts or the soaps are the best known anionic surfactants. These materials were originally derived from animal fats by saponification. The ionized carboxyl group provides the anionic charge. Examples with hydrocarbon chains of fewer than ten carbon atoms are too soluble and those with chains longer than 20 carbon atoms are too insoluble to be useful in aqueous applications. They may be prepared with cations other than sodium.

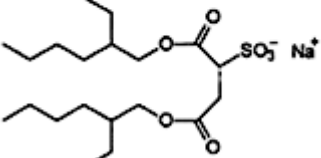
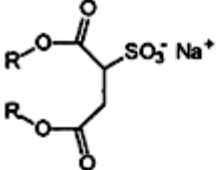
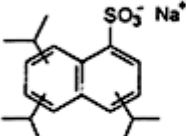
The blocking and elimination of the carboxyl group and replacement of anionic charge by sulphate and sulphonate substitution provided a revolution in detergency. Detergents such as sodium dodecylbenzene sulphonate (SDBS) have replaced soaps as laundry cleansing agents because of their efficacy and low cost. Such sulpho compounds may be readily derived from many natural products and synthetic precursors. Alkyl sulphates, such as sodium dodecylsulphate (SDS), alkyl ether sulphates, alkyl sulphonates, secondary alkyl sulphonates, aryl sulphonates such as alkylbenzene sulphonates, methylester sulphonates, α -olefinsulphonates and sulphonates of alkylsuccinates are important classes of anionic surfactants. Fatty acids and sulpho compounds illustrate three important anionic groups, carboxylate ($-\text{CO}_2^-$), sulphate ($-\text{OSO}_3^-$) and sulfonate ($-\text{SO}_3^-$). Phosphates such as mono- ($-\text{P}(\text{OH})\text{O}_2^-$) and dianions ($-\text{PO}_3^{2-}$) are also important. These dianions are basic and initially protonate in the neutral to slightly alkaline range, but they remain negatively charged to relatively low pH. Various data [13] have led to the following rank ordering of these groups with respect to their relative hydrophilicity:



The negative charge density is greatly affected by the group size. The carboxylate group is the smallest, attains the highest charge density and is the most hydrophilic in this series. However, it generally protonates in the pH 4–5 range, so its range of usefulness is very sensitive to pH. This follows also for the phosphates, but the sulphates and sulphonates generally remain charged, even to pH as low as 1.

-5-

Table C2.3.1 Structures of key anionic surfactants.

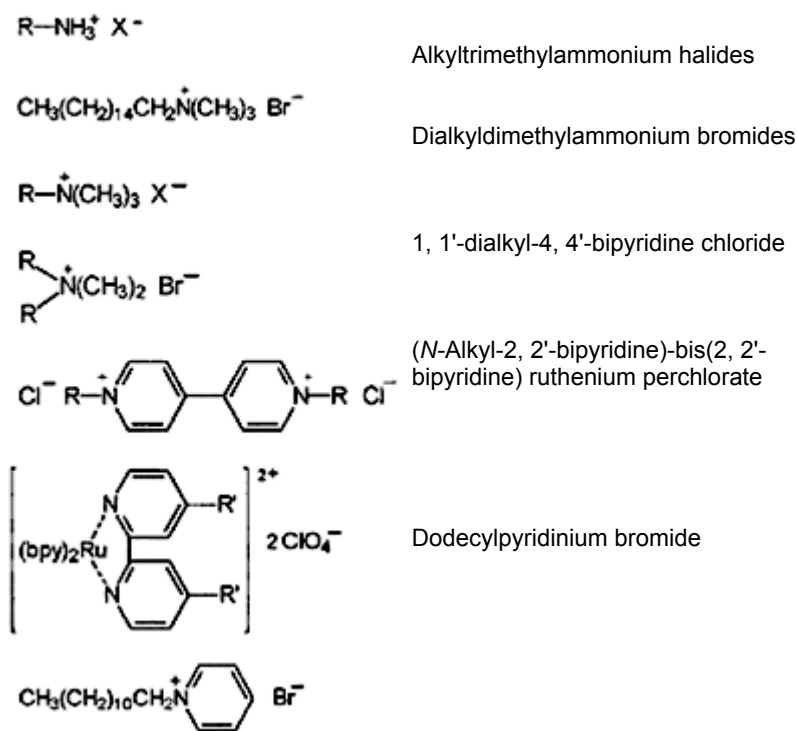
$\text{CH}_3(\text{CH}_2)_{10}\text{CH}_2\text{OSO}_3^- \text{Na}^+$	Sodium dodecylsulphate, SDS
$\text{R}-\text{CH}_2\text{OSO}_3^- \text{Na}^+$	Sodium alkylsulphates
$\text{CH}_3(\text{CH}_2)_{10}\text{CH}_2-\text{C}_6\text{H}_4-\text{SO}_3^- \text{Na}^+$	Sodium dodecylbenzene sulphonate, SDBS
$\text{R}-\text{C}_6\text{H}_4-\text{SO}_3^- \text{Na}^+$	Sodium alkylbenzenesulphonates
	Sodium bis(2-ethylhexyl) sulphosuccinate, AOT
	Sodium succinate esters
	Sodium di-, tri-isopropyl-naphthalene sulphonate, DTINS
$\text{R}-\text{CH}(\text{SO}_3^- \text{Na}^+)-\text{C}(=\text{O})-\text{OCH}_3$	α -sulpho fatty acid methyl esters
$\text{R}-\text{O}-\text{P}(=\text{O})(\text{O}^- \text{Na}^+)-\text{O}-\text{R}$	Dialkylphosphates

(B) CATIONIC

Most cationic surfactants derive from the quaternarization of nitrogen and most of the key cationic surfactant classes are illustrated in [table C2.3.2](#). Alkylammonium halides such as dodecylammonium bromide, for example, are good hydrogen bond donors and interact strongly with water. They also often can give up a proton and are then transformed into a nonionic surfactant. Chemical blocking of this hydrogen bond donating capability by full alkylation to yield the tetra-alkylammonium group results in cations that interact relatively weakly with halide counterions but strongly with organic anions.

Table C2.3.2 Structures of key anionic surfactants.

Alkylammonium halides
Hexadecyltrimethylammonium bromide, CTAB



A wide class of aryl-based quaternary surfactants derives from heterocycles such as pyridine and quinoline. The *N*-alkyl pyridinium halides are easily synthesized from alkyl halides, and the paraquat family, based upon the 4, 4'-bipyridine species, provides many interesting surface active species widely studied in electron donor-acceptor processes. Cationic surfactants are not particularly useful as cleansing agents, but they play a widespread role as charge control (antistatic) agents in detergency and in many coating and thin film related products.

(C) ZWITTERIONIC AND AMPHOTERIC

α -amino acids in the isoelectric pH range are true zwitterions and result from apparent intramolecular proton transfer. This class and other related surfactants are depicted in [table C2.3.3](#). The term zwitterionic surfactant is now taken to mean just about any combination of an anionic and a cationic group in a single amphiphilic molecule, whether or not either or both of these groups may be neutralized at some pH. In this sense zwitterionic is taken as synonymous with amphoteric. Often such species contain only one readily ionizable (or neutralizable) group. When such a group is the carboxyl group, the expected changes in charge and physical properties with pH must be borne in mind. Examples include tri-alkylammonioalkanoates, where the quaternary nitrogen and carboxylate are separated by the alkanolate carbon chain and wherein the nitrogen quaternerizes with the ω -carbon of the alkanolate. The relative hydrophilicities of carboxy, sulphonate and sulphate ammonio zwitterionics decrease in the following order:



This ordering is the same as discussed above for anionic surfactants and follows from the charge density of the particular acid group.

Table C2.3.3 Structures of key zwitterionic and amphipathic surfactants.

	Alkylamino acids
	Alkylbetaines
	Amidoalkylbetaines
	Alkylsulphobetaines
	Imidazolium betaines
	2,3-dimethyl-3-dodecyl-1,2,4-triazolium-5-thiolate
	Dialkanoyl lecithins

The various forms of betaines are very important for their charge control functions in diverse applications and include alkylbetaines, amidoalkylbetaines and heterocyclic betaines such as imidazolium betaines. Some surfactants can only be represented as resonance forms having formal charge separation, although the actual atoms bearing the formal charge are not functionally ionizable. Such species are *mesoionic* and an example of a triazolium thiolate is illustrated in table C2.3.3.

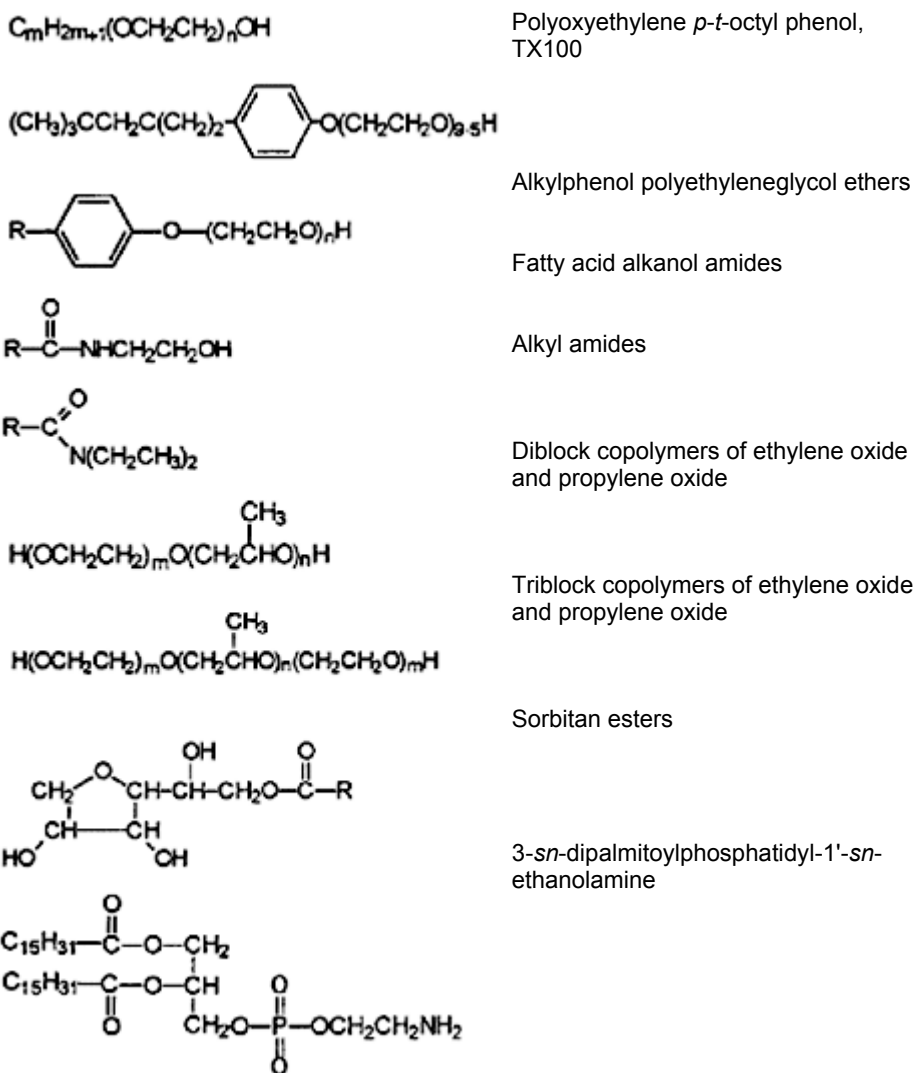
-8-

(D) NONIONIC

A selection of nonionic surfactant structures is listed in table C2.3.4. Many of these are analogous to some of the anionic and cationic surfactants, with the exception that the headgroup is uncharged. Steric and osmotic forces rather than electrostatic forces control interactions between nonionic surfactant headgroups. Oligomers of ethylene oxide have become the most important class of headgroup in nonionic surfactants and the alkanol polyethyleneglycol ethers, C_mE_n , have become the mostly widely studied class of nonionics. The ethylene oxide groups are generally believed to hydrate with about three water molecules. A related class is that of the alkylphenol ethoxylates. Triton X-100 (TX100) is the best known member of this class.

Table C2.3.4 Structures of key nonionic surfactants.

Alkanol polyethyleneglycol ethers,
 C_mE_n



-9-

The fastest growing class of block copolymeric nonionic surfactants derive from oligomeric ethyleneoxide (EO) and from oligomeric propyleneoxide (PO). They are obtained as diblocks, AB, as triblocks, ABA and as triblocks, BAB, where A denotes a hydrophilic block (such as EO) and B denotes a hydrophobic block (such as PO). Poly (butylene oxide) and polystyrene oligomers are also available for the more hydrophobic B blocks. These surfactants are now widely available commercially. An excellent review of their properties in aqueous solution is given by Alexandridis and Hatton [14]. Other nonionics include alkanol amides such as ethanolamides and diethanolamides, alkylamides and amine ethoxylates. Lecithin and other glycerol-based nonionics are physiologically important and are finding widespread pharmaceutical applications.

C2.3.3.2 PHYSICAL STATE OF SURFACTANTS

The physical state of surfactants affects how easily micellar solutions may be prepared. At room temperature surfactants are usually obtained as amorphous or crystalline solids, but increasing numbers of liquid surfactants are being derived, particularly as nonionics. The amorphous solid physical state is often obtained as a waxy paste. The surfactants in such pastes are usually in a liquid crystalline packing mode. Pure ionic surfactants of moderate formula weight (<500) are often obtained as crystalline powders. Published crystal structures suggest factors important to understanding how surfactants pack in micelles and in planar assemblies such as monolayers and bilayers. These factors include chain tilting and layering.

Molecular packing of surfactants in crystals has been reviewed at some length [15]. An almost universal factor

observed, at least for surfactants having long alkyl tailgroups, is that surfactants intrinsically pack in tail-to-tail or head-to-head bilayers in the absence of excess solvent. This factor makes it easy to understand the genesis of bilayer structures encountered in various surfactant mesophases, such as lamellar mesophases. Such bilayer packing typically allows for the compartmentalized separation of hydrophobic and hydrophilic domains in the crystal, wherein the hydrophobic tailgroups pack among themselves and the headgroups, often with solvent or water of crystallization, define a distinct hydrophilic or polar region. The tailgroups may pack end to end, as observed in a monohydrate of SDS [16], or they may pack in an interdigitated array such as in dodecylammonium bromide [17].

The extent of headgroup hydration provides relative control of the effective headgroup size and the tilt angles of hydrocarbon chains, relative to the normal, defined by the plane of the headgroups. As the effective headgroup size increases with increasing hydration this size must be accommodated by a concomitant amount of chain tilting. If the headgroup is small, little chain tilting is required in order to accommodate close packing of all of the molecules. As the headgroup size increases, the chains tilt to more densely fill space.

This headgroup size--chain tilting phenomenon can be illustrated for the case of SDS. The crystal structure of SDS monohydrate is illustrated in [figure C2.3.3](#) [18]. The packing is of a bilayer type and there is distinct chain tilting of about 40° relative to the ab plane of the headgroups. This tilting results in a bilayer thickness of about 28.9 \AA and a headgroup (projected) area of 29.5 \AA^2 . In a more anhydrous SDS polymorph [16] containing only four water molecules per 32 SDS, distinctly less tilting appears. A tilt angle of about 21° relative to the plane of the headgroups results from the much lower solvation, smaller headgroup projected area, 20.9 \AA^2 , and a much increased bilayer thickness, 38.9 \AA . Such dramatic variations in molecular packing with changing headgroup hydration and effective headgroup size underscores the importance of factors that control headgroup repulsion, such as counterion binding affinity and solvation, on packing in micelles.

-10-

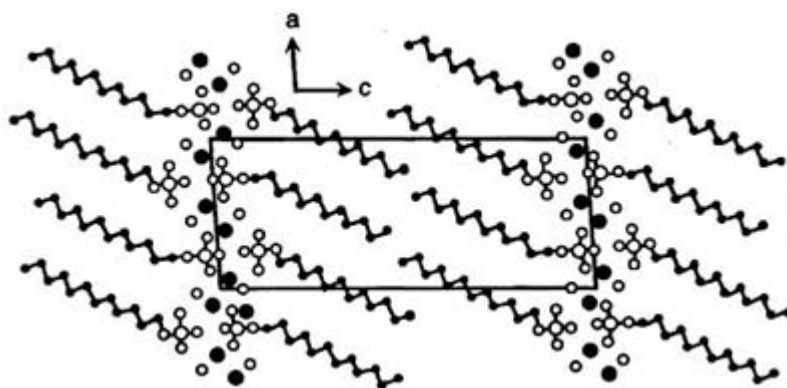


Figure C2.3.3. Molecular packing of SDS monohydrate viewed as projected on the ac plane. This polymorph crystallizes in a triclinic cell with unit cell constants a , b and c of 10.423 \AA , 5.662 \AA and 28.913 \AA , respectively, and with $\alpha = 86.70^\circ$, $\beta = 93.44^\circ$, $\gamma = 89.55^\circ$. There are four molecules per unit cell. Adapted from figure 2 of [18].

C2.3.4 EXPERIMENTAL METHODS FOR EXAMINING MICELLES AND MICELLIZATION

The concentration at which micellization commences is called the critical micelle concentration, cmc. Any experimental technique sensitive to a solution property modified by micellization or sensitive to some probe (molecule or ion) property modified by micellization is generally adequate to quantitatively estimate the onset of micellization. The determination of cmc is usually done by plotting the experimentally measured property or response as a function of the logarithm of the surfactant concentration. The intersection of asymptotes fitted to the experimental data or as a breakpoint in the experimental data denotes the cmc. A partial listing of experimental

techniques used to determine cmc is given in [table C2.3.5](#). Foremost among these methods is surface tension measurement, typically at the air–water interface. An example of such a measurement is illustrated in [figure C2.3.4](#) for the case of an equal weight mixture of sodium diisopropyl naphthalene sulphonate and sodium triisopropyl naphthalene sulphonate. The measurements illustrated were obtained [[19](#)] by the Wilhelmy plate technique. There are numerous other methods for measuring surface tensions, such as capillary rise, maximum bubble pressure, drop weight and pendant drop techniques [[20](#)]. Mention should be made of the largest compendium of cmc data, assembled by Mukerjee and Mysels [[21](#)], and of an excellent updating compendium of van Os *et al* [[22](#)].

-11-

Table C2.3.5 Survey of techniques and observables for determining cmc.

Density	Refractive index
Diffusion coefficient	Solubilization
Dye decomposition kinetics	Solubilization rate
Electromotive force	Specific heat
Conductance	Streaming current
ESR probe	Surface tensions
Flocculation rate	Taylor diffusion
Foaming power	Turbidometry
Freezing point	Turbidometric solubilization
Heat of dilution	Ultracentrifugation
Light scattering	Ultrafiltration
NMR	Vapour pressure lowering
Neutron scattering	Velocity of sound
Optical probe	Viscosity
Partial volume	Voltammetry of electroactive probe
Polarographic maximum	Wein effect
Potentiometry	X-ray scattering

Micellization is a second-order or continuous type phase transition. Therefore, one observes continuous changes over the course of micelle formation. Many experimental techniques are particularly well suited for examining properties of micelles and micellar solutions. Important micellar properties include micelle size and aggregation number, self-diffusion coefficient, molecular packing of surfactant in the micelle, extent of surfactant ionization and counterion binding affinity, micelle collision rates, and many others.

-12-

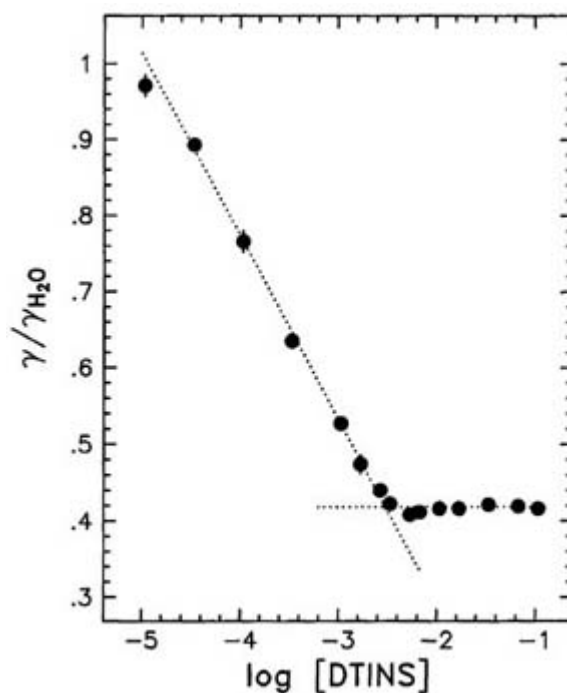


Figure C2.3.4. Relative surface tension of DTINS at 25 °C. The intersection of the dotted lines indicates a cmc of 3.0 mM. Reproduced by permission from figure 1 of [20].

C2.3.4.1 CRITICAL MICELLE CONCENTRATION

Measured cmcs range from large concentrations, such as approximately 0.6 M for sodium octanoate, to concentrations lower than 10^{-6} M for amphiphiles of very low solubility. The cmcs are significantly affected by a variety of physical properties associated with the hydrophilic headgroup, the hydrophobic portion (tailgroup) and the means by which these two groups are connected. Various aspects have already been discussed as affecting surfactant classification. The number of headgroups, the number of tailgroups and the connectivity of these groups can vary widely. Several linear free energy relationships between cmc and these molecular properties are discussed below and more rigorous thermodynamic modelling of micellization is presented in [section C2.3.5](#). These linear free energy relationships are based upon the thermodynamics of hydrophobic chain transfer from oil to water, solvent–hydrocarbon chain contacts, hydrocarbon chain packing and interactions between headgroups.

(A) CMC OF HOMOLOGOUS SERIES

Such linear free energy relationships are available for alkyl sulphates and for the C4 to C9 homologues of the dialkanoyl lecithins (see [table C2.3.3](#) for structure). Most of the naturally occurring phospholipids are too insoluble to form micelles, but the lower alkanoyl lecithins, also known as phosphatidylcholines, do form micelles. The cmcs for these homologues are listed in [table C2.3.6](#). The approximately linear free energy relationship between the alkyl chain length and log cmc is given by:

-13-

$$\log \text{ cmc} = a - bn. \quad (\text{C2.3.3})$$

The alkyl chain length is n , and a and b are fitting constants. The constant a usually varies with headgroup type and typically is 3.5 to 10. The b parameter is usually fairly uniform among different homologous alkyl chain surfactants. Values for b of about 0.5 are often obtained for single-chain nonionic surfactants and values of about 0.3 are obtained for single-chain ionic surfactants. For this series of alkyl sulphates, parameters $a = 4.39$ and $b = 0.29$ were obtained, and for this lecithin series, $a = 5.77$ and $b = 0.85$. Studies suggest that this b parameter is

proportional to the hydrophobic interaction energy of micelle formation.

Table C2.3.6 Cmc of homologous alkyl sulphates and dialkanoyl lecithins.

Homologue	cmc (mM)
Sodium hexylsulphate	420
Sodium heptylsulphate	220
Sodium octylsulphate	130
Sodium nonylsulphate	60
Sodium undecylsulphate	16
Sodium dodecylsulphate	8.2
Sodium tetradecylsulphate	2.05
Dibutanoyl lecithin	80
Dihexanoyl lecithin	14.6
Diheptanoyl lecithin	1.42
Diocetanoyl lecithin	0.265
Dinonanoyl lecithin	0.002 87

-14-

(B) SALT EFFECTS

The cmcs of ionic surfactants are usually depressed by the addition of inert salts. Electrostatic repulsion between headgroups is screened by the added electrolyte. This screening effectively makes the surfactants more hydrophobic and this increased hydrophobicity induces micellization at lower concentrations. A linear free energy relationship expressing such a salt effect is given by:

$$\log \text{cmc} = (\log \text{cmc})_0 - k_c c_i. \quad (\text{C2.3.4})$$

Here $(\log \text{cmc})_0$ is the log cmc in the absence of added electrolyte, k_c is related to the degree of counterion binding and electrostatic screening and c_i is the ionic strength (concentration) of inert electrolyte. Effects of added salt on cmc are illustrated in table C2.3.7.

Table C2.3.7 Salt effects on cmc.

Surfactant/salt	[salt] (M)	cmc (mM)
Sodium dodecylsulphate/NaCl	0	8.2
	0.1	1.4

	0.2	0.83
	0.4	0.52
Dodecylpyridinium bromide/KBr	0	12
	0.02	7.25
	0.05	4.70
	0.1	2.74
Dodecyltrimethylammonium bromide/NaBr	0	14.8
	0.0175	10.4
	0.05	7.0
	0.1	4.65

-15-

(C) SOLUBILITY

Headgroups and tailgroups have a big effect on solubility. Important surfactants, such as the phospholipids that make up significant fractions of cellular membranes, are nearly insoluble in water. Such insoluble and double-tail surfactants typically do not form micelles and generally prefer to pack as bilayers. A very important double-tail surfactant, AOT (see [table C2.3.1](#) for structure), is soluble in water, but is much more soluble in very many organic solvents. Such surfactants having high solubility in oils form the basis of reverse micellar (see [section C2.3.8](#)) and reverse microemulsion (see [section C2.3.11](#)) technology.

A quantitative treatment of surfactant solubility has been successfully made empirically using linear free energy relationships. An important relation is that for the linear free energy of transfer of alkanes to water [[23](#)]:

$$RT \ln S = aA. \quad (\text{C2.3.5})$$

In this relationship S is alkane solubility, A is the cavity surface area and a is the hydrophobic free energy per unit area. Extensive fitting of this equation [[24](#)] yields a value of $88 \text{ kJ mol}^{-1} \text{ \AA}^{-2}$ for the proportionality constant a . This value corresponds to an unfavourable free energy of about 3.6 kJ mol^{-1} for the transfer of a CH_2 group to aqueous solution.

(D) KRAFT POINT

The Kraft point (T_K) is the temperature at which the cmc of a surfactant equals the solubility. This is an important point in a temperature–solubility phase diagram. Below T_K the surfactant cannot form micelles. Above T_K the solubility increases with increasing temperature due to micelle formation. T_K has been shown to follow linear empirical relationships for ionic and nonionic surfactants. One found [[25](#)] to apply for various ionic surfactants is:

$$T_K = an + b \quad (\text{C2.3.6})$$

where n is hydrocarbon chain length and a and b are fitting parameters. The parameter b typically varies for

different types of headgroup. Various values (a , b) have been obtained such as (5.5, 11) for sodium alkylsulphates, (5.5, 29) for sodium alkylsulphonates (odd number of carbons), (5.5, 34) for sodium alkylsulphonates with an even number of carbon, and (5.5, 44) for sodium alkyloxyethylenesulphates.

(E) CLOUD POINT

The temperature at which a clear solution of surfactant just becomes turbid with heating is called the cloud point. This turbidity comes from light scattering by assemblies of micelles. These assemblies arise as a result of attractive interactions between micelles and may be thought of as clusters of micelles that in some cases become ordered mesophases of micelles. Such a phenomenon has been identified as a thermodynamic critical point and is most often exhibited by nonionic surfactants. As a practical matter it is known that cloud points can be raised by adding ionic groups to a surfactant (such as by adding a sulphate group to an oligomeric poly(propylene oxide)–poly(ethylene oxide) surfactant. At higher temperatures a phase separation into a water-rich phase and a surfactant-rich phase is generally obtained.

C2.3.5 THERMODYNAMICS OF MICELLIZATION

The free energy driving micellar aggregation primarily derives from intermolecular attractive and repulsive forces between the hydrophobic tailgroups and the decrease in free energy obtained when these tailgroups are no longer solvated by water or other polar solvent. Typically 10–100 monomers self-assemble to form a micelle. The spheroidal object formed by such aggregation allows one to put forward the pseudophase approximation, wherein the oily micellar interior constitutes a disperse oil phase (pseudophase) and the continuous aqueous phase represents the other pseudophase. Micellar solutions are, however, single-phase solutions, but this pseudophase approximation is very useful for keeping track of solute binding and partitioning into micelles, and phenomena related to such pseudophase concepts. A very important alternative approach to modelling micellization is the stepwise aggregation or chemical equilibrium approach. In this approach the detailed molecular or ionic growth of micelles is modelled in terms of sequential chemical equilibria, as surfactant merges with clusters to form slightly larger aggregates. Both of these approaches lead to thermodynamically equivalent results.

C2.3.5.1 CHEMICAL POTENTIAL

Assume that the chemical potential, μ , of surfactant in aggregates of size N in equilibrium with one another is uniform. One may therefore write

$$\mu = \mu_N \quad (\text{C2.3.7})$$

where μ is a constant and μ_N is the chemical potential of an aggregate of size N . We may also write

$$\mu_N = \mu_N^0 + \frac{kT}{N} \ln \left(\frac{X_N}{N} \right) \quad (\text{C2.3.8})$$

where (μ_N^0) is the standard part of the chemical potential and X_N is the mole fraction of surfactant aggregates of size N . After rearranging terms, X_N may be written:

$$X_N = N X_1^N e^{\frac{N(\mu_1^0 - \mu_N^0)}{kT}}. \quad (\text{C2.3.9})$$

The standard chemical potentials (μ_N^0) are approximately the same if the surfactant in each aggregate sees nearly the same interaction with the solvent. This simplifying assumption then gives

$$X_1 < 1 \quad (\text{C2.3.10})$$

$$X_N \ll X_1 \quad (\text{C2.3.11})$$

-17-

and yields the conclusion that the surfactant is in a monomeric state (aggregate of size 1). However, (μ_N^0) actually often decreases with increasing N . When this is the case, we see from [equation \(C2.3.9\)](#) that the mole fraction of surfactant in large aggregates may be relatively large. A necessary condition for the formation of large aggregates is this decrease of (μ_N^0) with increasing N .

C2.3.5.2 SHAPE EFFECTS

The variation of this standard chemical potential (μ_N^0) with aggregate size is important in determining whether micelles or aggregates will form. This reference potential also determines polydispersity and aggregate shape. For the sake of discussion we consider the formation of (one-dimensional aggregates) linear chains of surfactants. We approximate the pairwise binding energy (relative to separated species) as αkT . The standard reference potential is then written:

$$\mu_N^0 = - \left(1 - \frac{1}{N} \right) \alpha kT. \quad (\text{C2.3.12})$$

It appears that (μ_N^0) approaches (μ_∞^0) as $N \rightarrow \infty$. We therefore have in one dimension:

$$\mu_N^0 = \mu_\infty^0 + \frac{\alpha kT}{N}. \quad (\text{C2.3.13})$$

This expression can be generalized to two-dimensional aggregates (dislike micelles) and to spherical micelles, where

$$\mu_N^0 = \mu_\infty^0 + \frac{\alpha kT}{N^v} \quad (\text{C2.3.14})$$

and the exponent v takes values that depend on the aggregate shape. For dislike micelles, $v = 1/2$, and for spherical micelles, $v = 1/3$. The parameter α reflects the energy of intermolecular binding in units of the thermal energy.

C2.3.6 MORPHOLOGY AND STRUCTURE

The early Hartley model [2, 3] of a spherical micellar structure resulted, in later years, in some considerable debate. The self-consistency (inconsistency) of spherical symmetry with molecular packing constraints was subsequently noted [4, 5 and 6]. There is now no serious question of the tenet that unswollen micelles may readily deviate from spherical geometry, and ellipsoidal geometries are now commonly reported. Many micelles are essentially spherical, however, as deduced from many light and neutron scattering studies. Even ellipsoidal objects will appear

spherical when the time scale of the experimental probe is longer than the rotational period of the micelle. Cylindrical micelles presumably originate as spherical or ellipsoidal objects, and grow cylindrically as a consequence of competing energetics, wherein cylindrical extension is favoured over (hemispherical) end cap completion.

-18-

Deviations from spherical geometry are more prevalent in reverse micelles of low hydration, because spherical symmetry is more difficult to construct when concentrically arranging headgroups that may only interact repulsively. That these headgroups must fill space at the reverse micellar core necessarily introduces packing defects that mitigates against spherical symmetry. Normal and reverse micelles become more spherical when they are swollen with solvent, and thereby form microemulsion droplets ([section C2.3.11](#)). Regular spheroidal micelles, cylindrical micelles and branched cylindrical micelles have been imaged by cryo transmission electron microscopy, so that such approximate shapes have been fairly directly visualized [[26](#), [27](#)].

Resolution at the atomic level of surfactant packing in micelles is difficult to obtain experimentally. This difficulty is based on the fundamentally amorphous packing that is obtained as a result of the surfactants being driven into a spheroidal assembly in order to minimize surface or interfacial free energy. It is also based upon the dynamical nature of micelles and the fact that they have relatively short lifetimes, often of the order of microseconds to milliseconds, and that individual surfactant monomers are coming and going at relatively rapid rates.

In addition to these long contested arguments over the sphericity of micelles are arguments over the accessibility of solvent to the core of normal micelles. The morphological models such as that of Dill and Flory in [figure C2.3.2](#) portray micellar cores as composed of alkyl chains, and it has generally been believed that such an environment is a place where water may be expected to bind. Various experimental methods have, however, suggested the existence of water accessibility to these hydrocarbon chains and have caused some considerable disagreement in the literature. The simple packing models put forward by Fromherz showed, in a particularly clear way, that it is feasible to picture most parts of tailgroups as having a reasonable probability of accessing the surface of micelles. In other words, the micelle surface is not densely packed with headgroups, but also comprises intermediate and end of chain segments of the tailgroups. Such segments reasonably interact with water, consistent with dynamical measurements. Given that the lifetime of individual surfactants in micelles is of the order of microseconds and that of micelles is of the order of milliseconds, it is clear that the dynamical equilibria associated with micellar structures is one that brings most segments of surfactant into contact with water. The core of normal micelles probably remains fairly ‘dry’, however.

C2.3.6.1 PACKING PARAMETER

The surfactant number or surfactant parameter [[28](#), [29](#) and [30](#)], N_s , is defined as a dimensionless group:

$$N_s = v/la_h \quad (\text{c2.3.15})$$

where v is the volume of the surfactant tailgroup, l is the tailgroup length and a_h is the area of the head group. These volume and length parameters may be estimated from partial molecular structural studies of various homologous series of surfactants, from single-crystal x-ray diffraction studies and from molecular models. The key result is that this dimensionless group, based upon surfactant molecular properties, allows for the prediction of mesoscale packing morphology. A summary of the predictions obtained is given in [figure C2.3.5](#). See [table C2.3.1](#), [table C2.3.2](#), [table C2.3.3](#) and [table C2.3.4](#) for examples of the structures of some of the surfactants mentioned below.

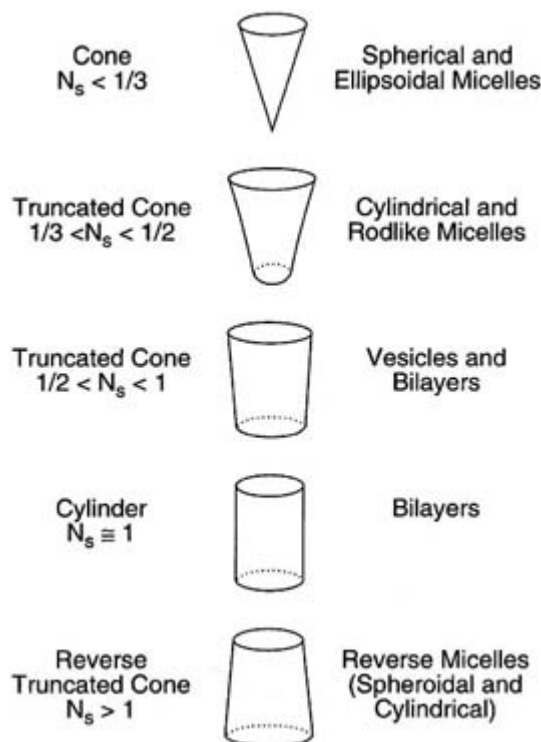


Figure C2.3.5. Mesoscale structures deriving from surfactants nominally exhibiting shapes corresponding to various packing parameter ranges.

(A) $N_s \leq 1/3$

This inequality indicates the amphiphile adopts a shape essentially equivalent to that of a cone with basal area a_h . Such cones self-assemble to form spheroidal micelles in solution or spheroidal hemimicelles on surfaces (see [section C2.3.15](#)). Single-chain surfactants with bulky headgroups, such as SDS, typify surfactants in this category.

(B) $1/3 < N_s < 1/2$

In this range the packing parameter yields a molecular shape similar to that of a truncated cone. Such cones may assemble to form rodlike structures and cylindrical micelles. Single-chain surfactants such as SDS and CTAB can fall within this range when the ionic strength is high enough to shield electrostatic repulsion between headgroups. This shielding is a *de facto* means for making the headgroups sterically smaller.

(C) $1/2 < N_s < 1$

This range yields more highly truncated cones. The main mesophase structure obtained from these units is a flexible bilayer such as that formed in vesicles and liposomes. These arrangements are often obtained from double-chain surfactants such as lecithin, double tailed cationic surfactants and AOT.

(D) $N_s \approx 1$

This parameter corresponds to cylindrical packing shapes. Surfactants and amphiphiles falling in this range often produce planar bilayers and lamellar mesophases. Such cylindrical building blocks also contribute to many

important liquid crystalline applications. Double-tailed surfactants with smaller headgroups, such as phosphatidyl ethanolamine, tend to form planar bilayers.

(E) $N_S > 1$

Surfactants having an inverted truncated cone shape yield inverted spheroidal micelles. Many double-chain surfactants such as AOT form such inverted micellar structures. These kinds of surfactant also form inverted anisotropic liquid crystalline phases.

C2.3.6.2 WORMLIKE MICELLES

Lengthy cylindrical micelles have become known as wormlike micelles, threadlike micelles and giant micelles. While the thickness of such micelles is typically of the order of two surfactant lengths (3–6 nm), the length of such micelles can approach 1000 nm or more [31]. These lengths are also known as persistence lengths, and usually are of the order of 30–200 nm. They are often studied in direct analogy to polymers, and much effort has been expended in applying the tenets of the statistical mechanics of polymer chains to wormlike micelles [32, 33]. Magid [34] gives an excellent review of the analogy of wormlike micelles to polymers. She presents a compelling picture of such giant micelles as living polymers.

C2.3.7 STATISTICAL MECHANICAL SIMULATIONS

In the absence of anisotropy introduced by specific surfactant–surfactant interactions, a spherical droplet model is reasonable because it tends to minimize the surface energy. Deviations from spherical symmetry occur because of the finite size and anisotropy of surfactant molecules and the anisotropy of interactions. Many early experimental data were interpreted on the assumption of spherical structures. In seminal Monte Carlo studies by Haan and Pratt [35], micelles simulating those of sodium octanoate were examined. They found that the chains adopted a spheroidal structure that was never close to perfectly spherical. An example packing configuration of the type observed is illustrated in [figure C2.3.6](#) for the case of an assembly involving 30 monomers. The shaded headgroups are mostly situated at the micellar surface, but it is obvious that much of the surface is also composed of methylene and methyl groups. This structure also obviously departs significantly from spherical symmetry. Spherical packing just is not energetically feasible when surfactant tailgroups must fill space. The situation changes dramatically when another solvent is permitted to fill the core region, as in microemulsions, and the surfactants can then pack in a more or less ‘planar’ manner at the oil–water interface. Similar conclusions have been upheld by much more time-consuming molecular dynamics simulations, such as those of Jönsoon *et al* [36]. A molecular dynamics snapshot of a sodium octanoate micelle is illustrated in [figure C2.3.7](#). This structure also shows that the micelle at a given instant is far from spherically symmetric. Of course, this structure is undergoing shape fluctuations as part of its dynamical equilibrium and it is constantly rotating in space. Such fluctuations and rotations tend to give an apparent spherical structure when averaged over time. This is why many structural studies based on neutron scattering,

for example, need to invoke spherical models. This molecular dynamics simulation also confirms the earlier Monte Carlo illustration of the accessibility of hydrocarbon chains to the micellar surface and the ensuing solvent contacts.

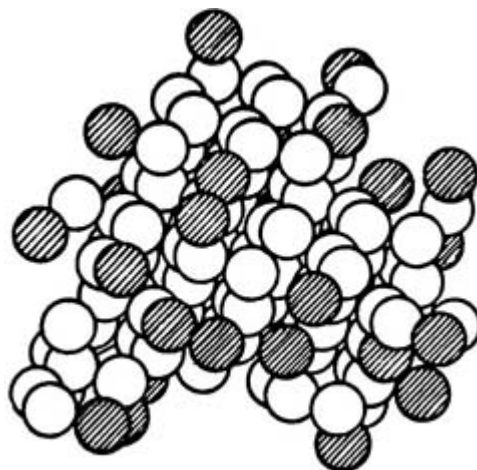


Figure C2.3.6. Illustration of micelle structure obtained by Monte Carlo simulations of model octanoate amphiphiles. There are 30 molecules simulated in this cluster. The shaded spheres represent headgroups. Reproduced by permission from figure 2 of [35].

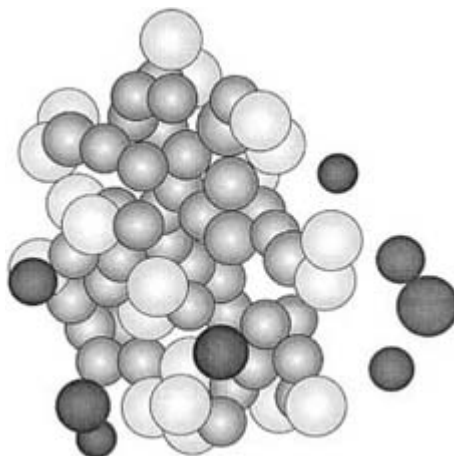


Figure C2.3.7. Snapshot of micelle of sodium octanoate obtained during molecular dynamics simulation. The darkest shading is for sodium counter-ions, the lightest shading is for oxygens and the medium shading is for carbon atoms. Reproduced by permission from figure 2 of [36].

Since these early simulations the art and science of simulating micellar structures has advanced significantly [37, 38 and 39]. Ionic micelles are being simulated with specific inclusion of solvent water. Micellization processes are being simulated dynamically using model united atom techniques [40, 41 and 42]; such methods are proving useful for understanding micelle and bilayer [43] structural evolution in solutions of small surfactants and for understanding micellization of polymeric surfactants (block copolymers) [44].

C2.3.8 REVERSE MICELLES

The idealized reverse micelle sketched in [figure C2.3.1](#) is an aggregate of a double-tail surfactant. In such systems the solvent is more compatible with the lyophobic part of the surfactant than with the headgroup. This preference

leads to the inverted structure illustrated. There is much less known in terms of conclusive physical data (e.g., cmc, aggregation number) on reverse micelles than on normal micelles. This, in part, is due to the hydrophilic nature of surfactant headgroups, and the fact that it is experimentally challenging to prepare and study reverse micelles in water immiscible solvents while keeping all water out of the system. The cartoon in [figure C2.3.1](#) is actually much more appropriate for a reverse *microemulsion* droplet, where the mole ratio of water to surfactant is of the order of ten or greater.

Such reverse droplets generally have tiny water pools in the core that exhibit many of the bulk properties of water. In such cases it is more sensible to imagine the surfactant headgroups aligned as a monolayer at a water–oil interface. However, in the absence of more than a few water molecules per surfactant, such idealized packing cannot be obtained without generating energetically unfavourable vacuum cores. Another point of controversy has been the question of whether core water is a necessary condition for the formation of reverse micelles [45]. There is no fundamental reason why dipolar headgroups of even ionizable surfactants cannot associate to form a reverse micellar core. The presence of some waters of hydration will tend to ameliorate such association, and provide hydrogen bonding as a means of forming such associations. However, data showing that water facilitates reverse micelle (microemulsion) formation are incontrovertible. NMR self-diffusion data [46] for reverse micelles of AOT in decane are illustrated in [figure C2.3.8](#) where the ratio of decane to decane plus water (0.6 % brine) is varied. In the limit of no added water, the self-diffusion of the surfactant is almost equal to that of water, and indicates that the reverse micelles formed have only a very few monomers in them. As water is added, both the self-diffusion of water and the self-diffusion of AOT decrease. This decrease indicates that the reverse micelles are controlling the diffusion rate of the water and AOT composing these micelles and are growing in size as more water is added. Reverse micelles of some surfactants will not form without added water, but many surfactants have been demonstrated to form them without added water [47].

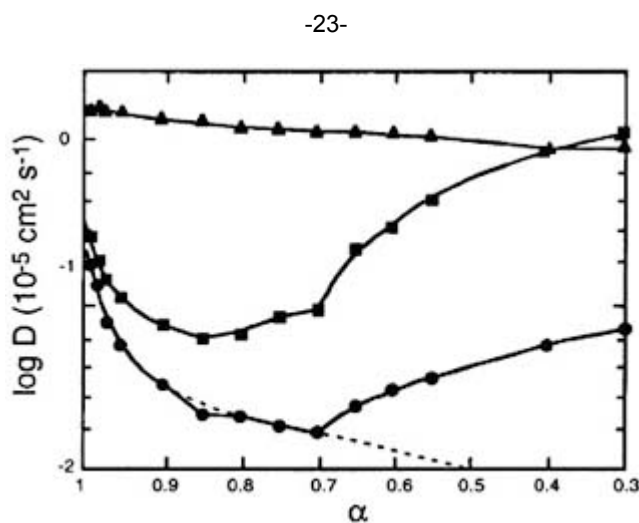


Figure C2.3.8. Self-diffusion coefficients at 45°C for AOT (●), water (■) and decane (▲) in ternary AOT, brine (0.6% aqueous NaCl) and decane microemulsion system as a function of composition, α . This compositional parameter, α , is the weight fraction of decane relative to decane and brine. Reproduced by permission from figure 3 of [46].

The issue of water in reverse micellar cores is important because water swollen reverse micelles (reverse microemulsions) provide means for carrying almost any water-soluble component into a predominantly oil-continuous solution (see discussions of microemulsions and micellar catalysis below). In the absence of water it appears that pre-micellar aggregates (pairs, trimers etc.) are commonly found in surfactant-in-oil solutions [47]. Critical micelle concentrations do exist (with some exceptions).

C2.3.9 SOLUBILIZATION AND PARTITIONING

Micelles are mainly important because they solubilize immiscible solvents in their cores. Normal micelles solubilize relatively large quantities of oil or hydrocarbon and reverse micelles solubilize large quantities of water. This is because the headgroups are water loving and the tailgroups are oil loving. These simple solubilization trends produce microemulsions (see [section C2.3.11](#)).

Other solubilization and partitioning phenomena are important, both within the context of microemulsions and in the absence of added immiscible solvent. In regular micellar solutions, micelles promote the solubility of many compounds otherwise insoluble in water. The amount of chemical component solubilized in a micellar solution will, typically, be much smaller than can be accommodated in microemulsion formation, such as when only a few molecules per micelle are solubilized. Such limited solubilization is nevertheless quite useful. The incorporation of minor quantities of pyrene and related optical probes into micelles are a key to the use of fluorescence depolarization in quantifying micellar aggregation numbers and micellar microviscosities [48]. Micellar solubilization makes it possible to measure acid–base or electrochemical properties of compounds otherwise insoluble in aqueous solution. Micellar solubilization facilitates micellar catalysis (see [section C2.3.10](#)) and emulsion polymerization (see [section C2.3.12](#)). On the other hand, there are untoward effects of micellar solubilization in practical applications of surfactants. When one has a multiphase

-24-

system as often encountered in cosmetic formulations or in photographic emulsion technology, one or more chemical components may be present as nanoparticulates (for example, an organic coupler that reacts to form image dye in a colour photographic element). If surfactant is present in excess, so as to form micelles, solubilization of such a component in the micelles often may lead to unwanted growth of such particulates via Ostwald ripening. Since micellar solubilization raises the effective solubility of the component, the ripening rates will increase and be exacerbated.

Micelles can solubilize gases. It has been demonstrated [49] that the Laplace model gives a good description of such solubilization for the case of ionic micelles:

$$\ln X_m = \ln X_b - \frac{2\sigma v}{\alpha n RT} \quad (\text{C2.3.16})$$

where X_m is the mole fraction of gas in the micelle and X_b is the mole fraction of gas in a bulk solvent equivalent to that of the surfactant tail, σ is the interfacial tension at the water–micelle interface, v is the partial molar volume of the gas in the micelle, n is the number of carbon atoms in the surfactant tail group, α is the segment length, T is temperature and R is the gas constant. This equation derives from the existence of a Laplace pressure differential across the micelle–water interface. Typical interfacial tensions applicable to micelles of hydrocarbon surfactants are in the neighbourhood of 30 dyn cm^{-1}

The solubilization of diverse solutes in micelles is most often examined in terms of partitioning equilibria, where an equilibrium constant K defines the ratio of the mole fraction of solute in the micelle (X_m) and the mole fraction of solute in the aqueous pseudophase. This ratio serves to define the free energy of solubilization ($-RT \ln K$). Recent monographs provide access to tabulations of such thermodynamic quantities [50, 51].

It is of particular interest to be able to correlate solubility and partitioning with the molecular structure of the surfactant and solute. ‘Likes dissolve like’ is a well-worn phrase that appears applicable, as we see in microemulsion formation where reverse micelles solubilize water and normal micelles solubilize hydrocarbons. Surfactant interactions, geometrical factors and solute loading produce limitations, however. There appear to be no universal models for solubilization that are readily available and that rest on molecular structure. Correlations of homologous solutes in various micellar solutions have been reviewed by Nagarajan [52]. Some examples of solubilization, such as for polycyclic aromatics in dodecyl sulphonate micelles, are driven by hydrophobic

interactions, while a variety of other types of micelle exhibit entropy driven solubilization. Other solutes and micelles involve solubilization in the headgroup region, and some of these cases are best modelled as specific binding phenomena (rather than partitioning). In cases where loading is fostered to the point of micellar swelling, where a core of neat solute is formed, we have the evolution of a microemulsion ‘droplet’.

C2.3.10 MICELLAR CATALYSIS

Reversibly formed micelles have long been of interest as models for enzymes, since they provide an amphipathic environment attractive to many substrates. Substrate binding (non-covalent), saturation kinetics and competitive inhibition are kinetic factors common to both enzyme reaction mechanism analysis and micellar binding kinetics.

-25-

Micellar catalysis has two important components. The first is the partitioning (localization, binding) of the reagent into or onto the micelle. Such localization provides increased interactions among the reagents, particularly with respect to the collision frequency in the case of bimolecular reactions. The second is a field or ‘solvent’ effect wherein the micellar environment provides catalysis by modifying the reaction transition state. The first of these components is the best understood and the most often encountered. Micellar catalysis differs from enzyme catalysis in some significant ways. Substrate specificity is usually much less significant in micellar catalysis. Also, the enhancement of reaction rates is much, much smaller in micellar catalysis than in enzyme catalysis. In addition, the substrate concentration in enzyme catalysis is usually much lower than the concentration of enzyme, but in micellar catalysis the concentration of micelles is usually of the same order as that of at least one of the reagents.

This localization phenomenon has also been shown to be important in a case of catalysis by *premicellar aggregates*. In such a case [53] premicellar aggregates of cetylpyridinium chloride (CPC) were shown to enhance the rate of the Fe(III) catalysed oxidation of sulphanic acid by potassium periodate in the presence of 1,10-phenanthroline as activator. This chemistry provides a lowering of the detection limit for Fe(III) by seven orders of magnitude. It must also be appreciated, however, that such premicellar aggregates of CPC actually constitute mixed micelles of CPC and 1,10-phenanthroline that are smaller than conventional CPC micelles.

The oily interiors of micelle cores often provide a driving force for substrate binding to micelles. Those interiors are not the only physical aspect that affect the chemistry. Studies have shown that both electrostatic and geometric factors may be important in micellar catalysis. In particular, the surfactant headgroup (its charge, its volume and its hydration) affects the catalytic power of the micelle. The role of surfactant headgroups in modifying transition states can be contrasted with the role of oily interiors in providing substrate binding. Micellar solubilization has also been shown to inhibit alkaline decomposition when the reactive site is buried *within* the micellar interior. However, many micellar catalysed reactions occur near the charged double layer in proximity to the ionic headgroups [54]. Nucleophilic aromatic substitutions, such as the attack by azide ion on 2,4-dinitrochloronaphthalene catalysed by CTAB micelles [55], are examples of micellar catalysis by field or solvent effects. Such substitutions typically involve charged transition states that can be significantly modified by micelle ionic structures.

C2.3.10.1 REACTION PATHWAYS

Alteration of reaction pathways by micellar catalysis often can yield modified product distributions. Such modification is most easily obtained when one pathway is catalysed and another is inhibited by the micellar environment. An excellent tabulation of product distribution variations obtained by micellar catalysis is given by Fendler [56]. An example is illustrated in [figure C2.3.9](#) for the CTAC (cetyltrimethylammonium chloride) catalysed photodecarbonylation of dissymmetrical ketones [57], A(CO)B. The CTAC micelles provide a cage effect that greatly enhances the joining of the A and B radicals produced by the photolysis. Although the localization effects and field effects provided in micellar catalysis can provide significant rate enhancements and

these environmental effects can provide dramatically altered product distributions, there has been little effective development of micellar catalysis in modifying stereoselectivities. The dynamical equilibria exhibited by micelles mitigates against easily developing stereoselective binding equilibria.

-26-

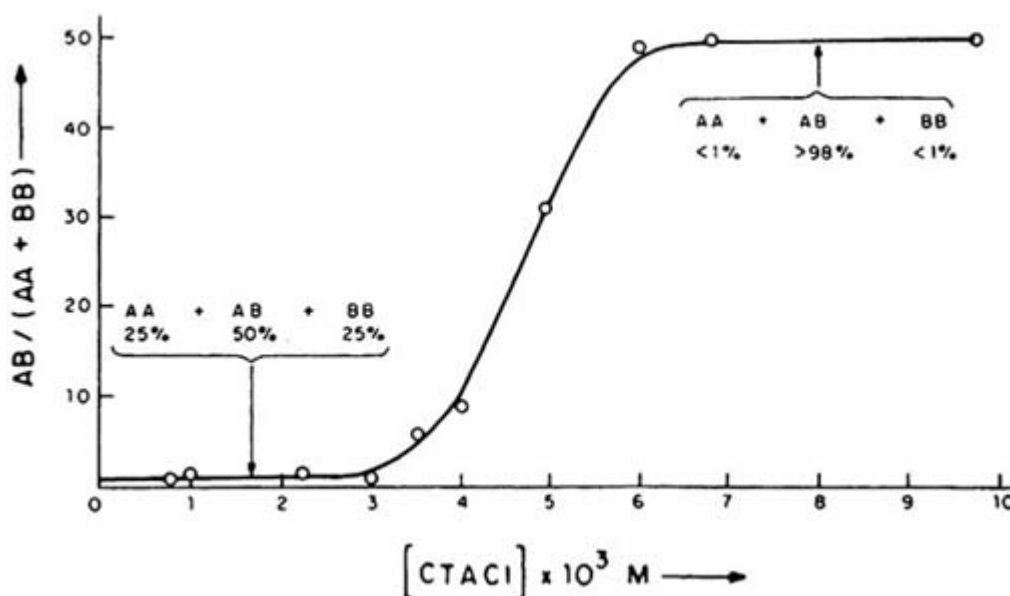


Figure C2.3.9. Product distribution of dissymmetrical ketone photolysis as influenced by cetyltrimethylammonium chloride (CTAC) micelles. The initial ketone, A(CO)B is photolysed to lose the carbonyl group and to produce three products, AA, AB and BB. These data are for benzyl (A) 4-methylbenzyl (B) ketone. Product AA is 1,2-diphenylethane, product BB is 1,2-ditolylethane and product AB is 1-phenyl-2-tolyl-ethane. At low CTAC concentration, in the absence of micelles, a random distribution of products is obtained. In the presence of micelles, however, the AB product is heavily favoured. Adapted with permission from [57].

C2.3.10.2 SELF-REPLICATION

A particularly interesting type of micellar catalysis is the autocatalytic self-replication of micelles [58]. Various examples have been described, but a particularly interesting case is the biphasic self-reproduction of aqueous caprylate micelles [59]. In this system ethyl caprylate undergoes hydroxyl catalysed hydrolysis to produce the free carboxylate anion, caprylate. Caprylate micelles then form. As these micelles form, they solubilize ethylcaprylate and catalyse further production of caprylate anion and caprylate micelles.

C2.3.10.3 REVERSE MICELLAR CATALYSIS

An even greater diversity of catalytic processes has been obtained in reverse micellar systems. Reverse micelles, as pictured in [figure C2.3.1](#) usually contain hydrating water molecules around the surfactant headgroups. Additional water solubilized in the core of reverse micelles produces reverse microemulsions ([section C2.3.11](#)), where the size of the water nanoreactor core can be simply adjusted by the amount of water added to the system. The same kinds of partitioning, field effects and headgroup charge effects encountered in normal micellar catalysis are also obtained in reverse micellar catalysis. However, the direct incorporation of varied catalysts, such as enzymes and cofactors, into the water pools provides a plethora of additional chemistries. Many of these chemistries are described in monographs [56, 60, 61], and include electron transfer reactions, donor-acceptor interactions, ester hydrolysis, carbohydrate hydrolysis, polymerization of olefinic monomers such as acrylates, methacrylates, acrylamides, acid dissociation,

Schiff base formation, photochemistry, protein partitioning, catalysis by chymotrypsin, lipase, peroxidase, phosphatase, catalase and alcohol dehydrogenase.

C2.3.11 MICROEMULSIONS

Microemulsions can be simply defined as solvent-swollen micellar solutions, wherein the swelling solvent is immiscible with the solvent of the pseudocontinuous phase. For example, in the case of normal micelles in aqueous solution, the swelling solvent typically is a water-immiscible organic solvent. This definition is not sufficient, however, because it only covers an important topological subset of microemulsions. This subset is one wherein the swelling solvent is present in objects readily identified as spheroidal or cylindrical ‘particles’ or ‘droplets’. Microemulsions also contain a distinct topological entity known as irregular, bicontinuous microemulsions, wherein interdigitated domains of immiscible solvents are separated by a monolayer of surfactant. These structures will be elaborated further in the discussion below.

All microemulsions have at least three chemical components, surfactant and two immiscible solvents; micellar solutions need only have two chemical components, surfactant and solvent. Microemulsions and micellar solutions are thermodynamically stable isotropic solutions. In this context the phrase ‘thermodynamically stable’ means that only moderate mixing is required to transform a mixture of the three main components into a transparent isotropic solution. There, of course, can be diverse kinetic barriers in such dissolution processing, and it often is convenient to dissolve the surfactant in one of the solvents before mixing in the other solvent, but, once formed, an isotropic microemulsion solution will not phase separate unless there is molecular decomposition or some field variable, temperature for example, is changed. Our use of the adjective isotropic in these contexts means that microemulsions are optically isotropic. That is to say the microstructures that exist in the aggregates in microemulsions are almost always significantly smaller than the wavelengths of visible light.

The composition of microemulsions is usefully considered in the context of ternary phase diagrams such as that illustrated in [figure C2.3.10](#). The region marked L_1 is the normal single-phase microemulsion domain wherein the solution microstructure is essentially that of micelles swollen with oil. Similarly, the L_2 domain essentially has water-swollen reverse micelles. At certain field variables these domains may be simply connected to one another. When such is the case, the microstructure in the ‘connecting’ region is predominantly that of an irregular, bicontinuous microemulsion. The physical, chemical and practical aspects of microemulsions fill many, many volumes, so no further attempt will be made to elucidate them here. Excellent monographs are available [[62](#), [63](#), [64](#), [65](#) and [66](#)].

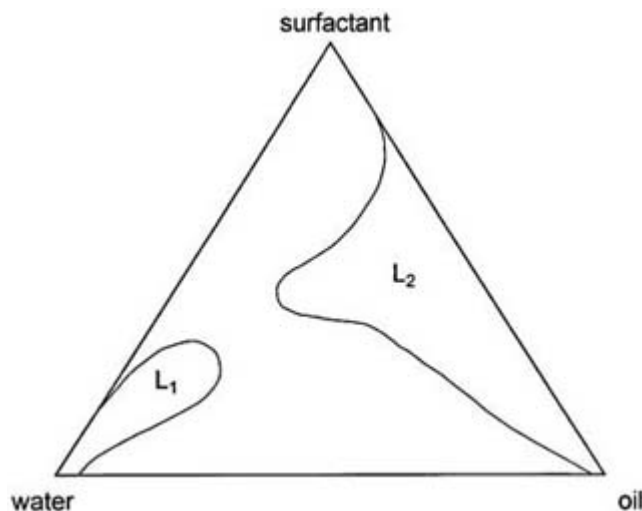


Figure C2.3.10. Ternary phase diagram of surfactant, oil and water illustrating the (regular) L₁ and (reverse) L₂ microemulsion domains.

C2.3.12 EMULSION POLYMERIZATION

The production of organic polymeric particles in the size range of 30–300 nm by emulsion polymerization has become an important technological application of surfactants and micelles. Emulsion polymerization is very well and extensively reviewed in many monographs and texts [67, 68], but we want to briefly illustrate the role of micelles in this important process.

Surfactants provide temporary emulsion droplet stabilization of monomer droplets in the two-phase reaction mixture obtained in emulsion polymerization. A cartoon of this process is given in [figure C2.3.11](#). There we see that a reservoir of polymerizable monomer exists in a relatively large droplet (of the order of the size of the wavelength of light or larger) kinetically stabilized by surfactant.

The role of micelles comes into play in nucleating the formation of the polymerized organic particles. The initial stages of polymerization, in the case of some monomers, may occur in the aqueous phase, but at some point of growth the aqueous solubility is no longer sufficient. The micelles provide a place, through solubilization, for small oligomers to continue growing thereafter. The micellar environment provides a region of intermediate solubility, more favourable for these oligomers than the aqueous phase or the reservoir phase. Transport between the monomer reservoirs (emulsion droplets) and the reacting, polymerizing polymer particle is also facilitated by micellar solubilization of monomer.

Three phases, initiation, growth and termination, are typically encountered in emulsion polymerization. The initiation stage involves the creation of monomeric radicals and ensuing oligomeric radicals. These oligomers become temporarily engulfed in a micelle. The initiation stage is followed by the growth stage, wherein monomer in the reservoir (emulsion) particles is depleted. Finally, the radical chemistry and polymerization is shut down in the termination stage, and the radical polymerization ceases.

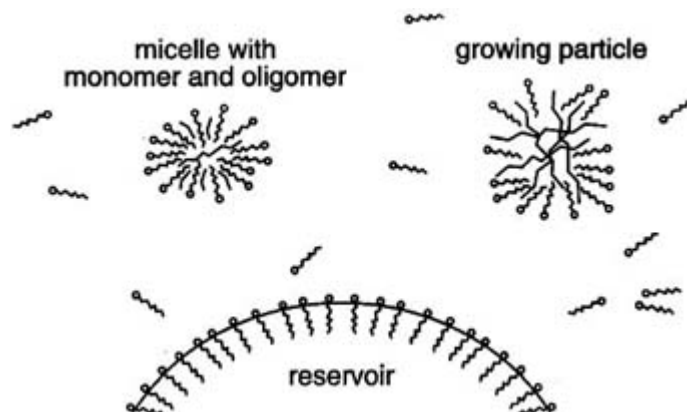


Figure C2.3.11 Key surfactant structures (not to scale) in emulsion polymerization: micelles containing monomer and oligomer, growing polymer particle stabilized by surfactant and an emulsion droplet of monomer (reservoir) also coated with surfactant. Adapted from figure 4-1 in [67].

C2.3.13 MICELLAR AND MICROEMULSION POLYMERIZATION

The polymerization of micelles composed of polymerizable surfactants while maintaining micellar morphology and size is a continuing challenge in colloid and polymer science. No significant case of polymerizing a micelle, while maintaining morphological integrity, has yet been convincingly reported. This situation will no doubt change in the near future, as the steric constraints of polymerization in related systems are being overcome [69, 70]. Since micelles and microemulsions are in dynamic equilibrium, it is difficult to polymerize surfactant monomers fast enough while maintaining the morphological integrity of the micelle. Also, steric constraints come into play once a monomer has been joined to another, so that it is difficult to carry on the polymerization while maintaining the nominal packing that existed before polymerization. It is noteworthy that Cussler and co-workers [71, 72] report what they believe to be polymerization in a bicontinuous microemulsion that preserves interfacial structure. However, it must be noted that the main evidence is that correlation lengths appear preserved, before and after polymerization, and that this preservation may be coincidental was not ruled out.

Irrespective of whether microemulsion and micellar interfaces can be polymerized using polymerizable surfactants, it is now very well established that monomers in microemulsions, whether part of or entirely composing the oil phase, whether solubilized in the aqueous phase or whether part of both pseudophases can be polymerized to form very small particles or interesting bulk materials. When oily monomers are polymerized inside conventional microemulsions or when aqueously soluble monomers (acrylamides, acrylates) are polymerized in reverse microemulsions, one usually obtains latexes similar to those obtained by emulsion polymerization (30–100 nm in diameter). The mechanism of microemulsion polymerization is essentially identical to that for emulsion polymerization, except that the initiation and termination intervals are often not connected by a very lengthy growth interval. This is because there are no large monomeric reservoirs and monomer transport is facile and essentially diffusion controlled. Excellent reviews on microemulsion polymerization are readily available [73, 74 and 75]. Very recent success has been obtained in obtaining very small particles in reverse microemulsion polymerization [76]. The key in the studies of Pileni and co-workers is

to use surfactants that have polymerizable counter-ions. Double-tail cationic surfactants with acrylate-type counter-ions yield ultrasmall particles (2–4 nm). Additional polymerizable monomer can be included in the water core, but having such species as part of the surfactant (headgroup) without covalently induced strain on the cationic surfactant packing appears to play a major factor in preserving the morphology through the polymerization interval.

C2.3.14 MICELLE-BASED MESOPHASES

Two-dimensional maps of single- and multiple-phase domains as a function of temperature and of surfactant/solvent mole fraction (or, alternatively, weight fraction) provide a useful means for characterizing ionic and nonionic surfactants. The variety of physical states discussed earlier, and more, are routinely exhibited in binary phase diagrams of surfactants. The formation of micellar and other aggregates is driven by hydrophobic interactions. As such aggregates become more concentrated and interact more strongly, supramolecular ordering of such aggregates occurs and the shape of such micelles and aggregates can change. These transitions yield a rich array of mesophases, such as lamellar phases where surfactant packs in infinite bilayers that can be swollen in the headgroup region by water or in the tailgroup region by organic solvent [77]. Some of these mesophases have building blocks identifiable as micelles.

Spherical and ellipsoidal micelles and rodlike micelles can form supramolecular assemblies having cubic symmetry. Ellipsoidal micelles at sufficiently high concentration may pack at cubic lattice sites to produce viscous cubic phases. For example, certain triblock copolymeric micelles form cubic mesophases [78], wherein the micelles aggregate in a cubic array. Such arrays are often thermoreversible gels that ‘melt’ on cooling, as isotropic and low viscosity micellar solutions form (as the number density of micelles decreases with concomitant increases in the cmc). Both fcc and bcc arrays have been reported for such cubic mesophases [79, 80, 81, 82, and 83]. Furthermore, such ordering is often induced by shear. Two-dimensional scattering from a shear-induced cubic mesophase of $\text{EO}_{96}\text{PO}_{39}\text{EO}_{96}$ (see table C2.3.4) is illustrated in figure C2.3.12 [84]. This is the type of scattering pattern expected for a bcc lattice.

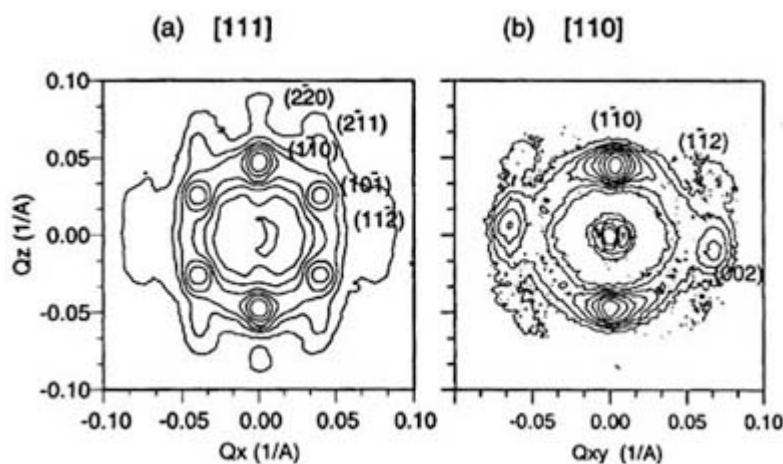


Figure C2.3.12. Two-dimensional neutron scattering by $\text{EO}_{96}\text{PO}_{39}\text{EO}_{96}$ (Pluronic F88) micellar solution under shear with (a) the sample shear axis parallel to the beam, and (b) the sample rotated 35° around the vertical axis. Reflections for several of the Miller indices expected for a bcc lattice are annotated. Reproduced by permission from figure 4 of [84].

Normal and reverse cylindrical micelles or rodlike micelles can pack hexagonally to form a variety of mesophases. When cylindrical rods (micelles) pack hexagonally in two dimensions to form the normal hexagonal (H_I) mesophase, one obtains what was known as the middle phase in soap technology. Similar arrays may form from reverse rodlike micelles, and such mesophases are called inverse hexagonal (H_{II}) phases. Such mesophases are liquid crystalline and highly viscous and are often formed by phospholipids. These hexagonal mesophases are illustrated in figure C2.3.13 [85]. A thorough theoretical analysis of nematic mesophase formation by rodlike micelles, in the framework of micellar growth coupling with micellar alignment, has been given by vander Schoot and Cates [86]. A summary of work done on trying to understand the formation of hexagonal mesophases has been

detailed by Odijk [87].

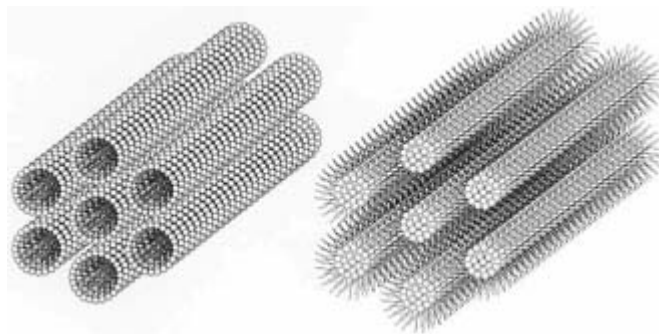


Figure C2.3.13. Normal (H_p , left) and inverse (H_{II} , right) hexagonal mesophases composed of rodlike micelles.

C2.3.15 ADSORBED MICELLES

The formation of surface aggregates of surfactants and adsorbed micelles is a challenging area of experimental research. A relatively recent summary has been edited by Sharma [51]. The details of how surfactants pack when aggregated on surfaces, with respect to the atomic level and with respect to mesoscale structure (geometry, shape etc.), are less well understood than for micelles free in solution. Various models have been considered for surface surfactant aggregates, but most of these models have been adopted without firm experimental support.

C2.3.15.1 ISOLATED SURFACTANT ADSORPTION

Individual and isolated surfactant adsorption onto a surface can be imagined to occur in various ways, such as those depicted in [figure C2.3.14](#). The headgroup end-on adsorption shown in [figure C2.3.14\(a\)](#) is the most commonly *assumed* mode of adsorption. It is generally invoked for charged surfactant adsorption, when it is assumed that the prevalent surface charge is opposite to that of the surfactant headgroup. This assumption has most often been applied to the adsorption of straight-chain cationic surfactants onto negatively charged surfaces (silica, metals etc.). The mode exhibited in [figure C2.3.14\(b\)](#) is one that requires interaction of the headgroup and a part of the surfactant chain. Modes depending on hydrophobic interactions with the surface are shown in parts (c) and (d). When both the headgroup and the tail interact with the surface, the structures in [figure C2.3.14\(e\)](#) and [figure C2.3.14\(f\)](#) would be expected. Recent Raman evidence for CTAB adsorption onto negatively charged silver according to modes such as those of (e) and (f) appears unequivocal [88]. Modes (c)–(f) appear to be the most often overlooked in the literature.

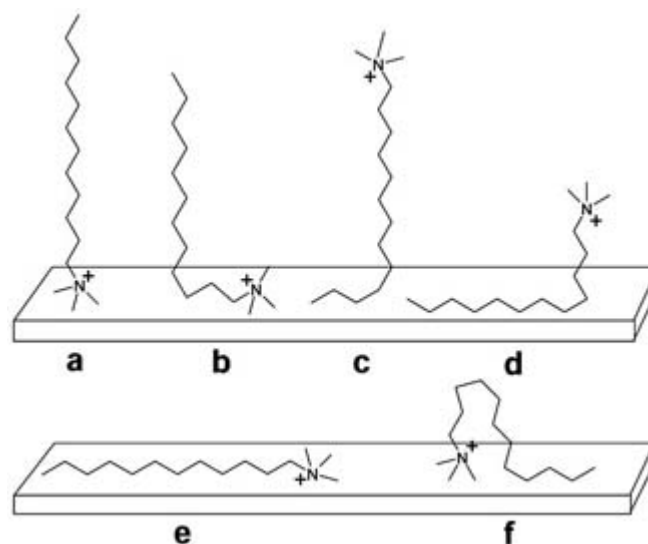


Figure C2.3.14. Isolated surfactant modes of adsorption at liquid–solid interfaces for a surfactant having a distinct headgroup and hydrophobic portion (dodecyltrimethylammonium cation): (a), (b) headgroup specific interaction; (c), (d) hydrophobic tail interaction, (e),(f) headgroup and tail interactions.

C2.3.15.2 HEMIMICELLE FORMATION

It is easy to see how extrapolation of isolated adsorption depicted in figure C2.3.14(a) can lead to a picture of hemimicelle structure as illustrated in figure C2.3.15(a). Combinations of the modes illustrated in figure C2.3.14 can be used to construct the alternative hemimicelle model of figure C2.3.15(b). This alternative model was much more popular in the 1960s and 1970s, but was supplanted by the evolution of dogma in favour of figure C2.3.15 (a) . It should be stressed that the structure for the hemispherical hemimicellar model takes simultaneous account of surfactant packing parameters and basic interfacial free energy, whilst the model of figure C2.3.15(a) ignores the highly unfavourable tail–solvent interactions that remain about the periphery after aggregation.

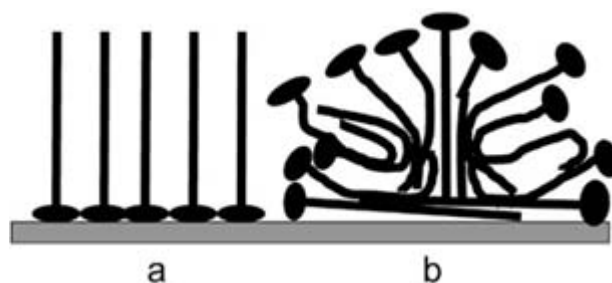


Figure C2.3.15. Hemimicelle structures: (a) monolayer type hemimicelle; (b) spheroidal, globular hemimicelle.

Adsorbed micelles, also called *admicelles*, are illustrated in cross-section in figure C2.3.16(a) and figure C2.3.16 (b) . The bilayer structure in (a), when capped at the ends, mitigates against the unfavourable solvent interactions maintained in the hemimicellar model of figure C2.3.15(a) . When suitable end caps are added, these two admicelle models in figure C2.3.16 become somewhat indistinguishable. It is possible for the bilayer admicelle, however, to propagate over large areas so that the detailed molecular packing remains fairly ordered and distinguishable from the more random short range packing exhibited in the model of figure C2.3.16(b) .

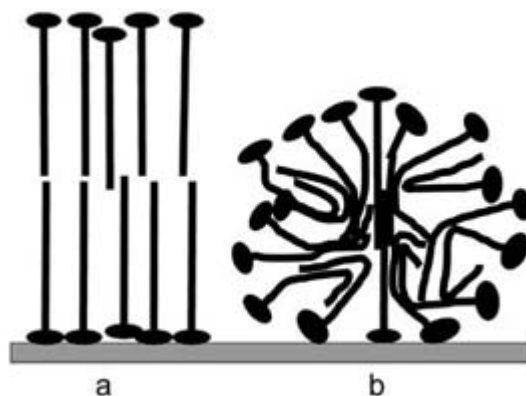


Figure C2.3.16. Adsorbed micelle structures: (a) bilayer admicelle; (b) spheroidal, globular adsorbed micelle.

Below the onset of surface aggregation through the adsorption of micelles, there are up to four regimes of adsorption observed. The first regime is that of isolated surfactant adsorption, such as depicted in [figure C2.3.14](#). The second regime is that in which hemimicelles form and/or admicelles adsorb. In log–log plots of adsorbed surfactant *versus* solution-phase surfactant concentration, the breakpoint between these first two regimes denotes the ‘surface cmc’ for hemimicelle formation. Additional surfactant uptake with increasing concentration in a subsequent regime has been interpreted as corresponding to completion of a surface bilayer. Asymptotic adsorption corresponds to a fourth regime, wherein solution micelles are believed to be in dynamical equilibrium with a surface bilayer-type admicelle, such as depicted in [figure C2.3.16\(a\)](#).

C2.3.15.3 DIRECT STRUCTURAL DATA

The qualitative resolution of the morphology and structure of surfactant aggregates on surfaces is experimentally formidable. Until only recently all such structural assignments were made on quite indirect bases. In the case of interpreting neutron scattering data, for example, generally only unbounded structures of the type illustrated in [figure C2.3.15\(a\)](#) and [figure C2.3.16\(a\)](#) were considered in the modelling and data fitting processes. Fortunately, recent AFM (atomic force microscopy) studies by Manne [89] and collaborators have provided direct morphological data for a variety of surfactants interacting with several different surfaces. The overall results are exciting because they illustrate a diversity of structure.

All of these results were obtained for solutions that were above the respective cmc in aqueous solution. Structural results [90] for tetradecyltrimethylammonium bromide are illustrated in [figure C2.3.17](#) where the surfactant is adsorbed onto a hydrophobic cleavage plane of MoS₂. The cartoon illustrated shows the structure deduced fairly directly from the atomic force micrographs. Other results are summarized in [table C2.3.8](#). The dodecyl- to hexadecyltrimethylammonium bromides, hexadecyltrimethylammonium chloride and SDS all yield similar parallel half-cylinders on a hydrophobic cleavage plane of graphite. These results contrast with earlier assignments of vertical monolayers (such as illustrated in [figure C2.3.15](#) for the cationic surfactants [93] and for SDS [94] and hemispheres for SDS [95]). A strikingly different result, parallel flexible cylinders [90], was obtained for the dodecyl- to hexadecyltrimethylammonium bromides and chlorides on an anionic cleavage plane of mica. These flexible cylinders remain parallel but undergo s-shaped shifts as the underlying lattice is covered. Earlier assignments [92, 96, 97 and 98] deduced uniform bilayer structures such as illustrated in [figure C2.3.16\(a\)](#). This AFM approach did yield an assignment of a uniform bilayer for the adsorption of the double-tail cationic bromide on mica in agreement with an earlier study [99]. The results obtained by AFM for adsorption of the cationic tetradecyltrimethylammonium bromide onto the anionic surface of SDS are particularly interesting, as spheres and spheroids, such as depicted in [figure C2.3.15\(b\)](#) and [figure C2.3.16\(b\)](#), were deduced in agreement with earlier assignments [100, 101] of spheres, but contrasted with other assignments inferring bilayer patches [102, 103, 104 and 105]. Much work remains in developing a firmer comprehension of hemimicellar structure at concentrations lower than those that produce micelles in bulk solution.

Table C2.3.8 Hemimicelle morphology by AFM.

Surface	Surfactant	Morphology
Graphite	$\text{CH}_3(\text{CH}_2)_{10}\text{CH}_2\text{N}^+(\text{CH}_3)_3 \text{Br}^-$	Parallel half-cylinders [90, 91]
	$\text{CH}_3(\text{CH}_2)_{12}\text{CH}_2\text{N}^+(\text{CH}_3)_3 \text{Br}^-$	
	$\text{CH}_3(\text{CH}_2)_{14}\text{CH}_2\text{N}^+(\text{CH}_3)_3 \text{Br}^-$ (CTAB)	Parallel half-cylinders [92]
Mica	$\text{CH}_3(\text{CH}_2)_{10}\text{CH}_2\text{OSO}_3^- \text{Na}^+$ (SDS)	Parallel flexible cylinders [91]
	$\text{CH}_3(\text{CH}_2)_{10}\text{CH}_2\text{N}^+(\text{CH}_3)_3 \text{Br}^-$	
	$\text{CH}_3(\text{CH}_2)_{12}\text{CH}_2\text{N}^+(\text{CH}_3)_3 \text{Br}^-$	Uniform bilayer [91]
	$\text{CH}_3(\text{CH}_2)_{14}\text{CH}_2\text{N}^+(\text{CH}_3)_3 \text{Br}^-$ (CTAB)	
MoS ₂	$\text{CH}_3(\text{CH}_2)_{10}\text{CH}_2 > \text{N}^+(\text{CH}_3)_2 \text{Br}^-$	Parallel half-cylinders [91]
	$\text{CH}_3(\text{CH}_2)_{10}\text{CH}_2$	Spheres and spheroids [91]
SiO ₂	$\text{CH}_3(\text{CH}_2)_{12}\text{CH}_2\text{N}^+(\text{CH}_3)_3 \text{Br}^-$ (CTAB)	
	$\text{CH}_3(\text{CH}_2)_{12}\text{CH}_2\text{N}^+(\text{CH}_3)_3 \text{Br}^-$ (CTAB)	

-35-

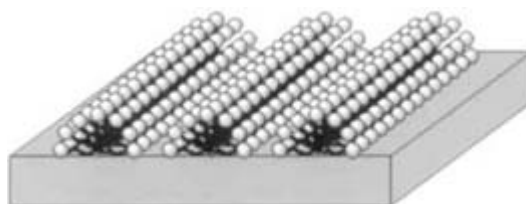


Figure C2.3.17. Model of half-cylindrical aggregates (hemimicelles) on a crystalline hydrophobic substrate, such as for tetradecyltrimethylammonium bromide on MoS₂ [91]. Adapted from figure 2 of [89].

C2.3.16 MICELLE–POLYMER INTERACTIONS

Many practical applications of surfactants and polymers utilize both types of component in the same single- or multi-phase formulation. For example, in the photographic industry, anionic surfactants are used as charge stabilizers for a myriad of organic and inorganic nanoparticulates and polymers (such as gelatin and other polyelectrolytes) are used as steric stabilizers for the same particulates. There are significant synergistic interactions between surfactants and polymers, and most of these interactions are based on how these polymers interact with micelles of these surfactants [106, 107, 108 and 109].

C2.3.16.1 EFFECTS ON CMC

An important mechanistic feature of such interactions is in the molecular detail of the interaction. Two general

types of interaction may be articulated: (1) nucleation of a micelle by some pendant group of the polymer, wherein a micelle grows about some part of the polymer; (2) polymeric adsorption onto the micellar surface, where one can picture a nonintegral type of binding between the micelle and polymer involving only the micellar headgroup region and active sites of the polymer [110]. These two limiting cases define boundaries between which mixtures of the two effects can be observed.

Additives, whether hydrophobic solutes, other surfactants or polymers, tend to nucleate micelles at concentrations lower than in the absence of additive. Due to this ‘nucleating’ effect of polymers on micellization there is often a measurable cmc, usually called a critical aggregation concentration or cac, below the regular cmc observed in the absence of added polymer. This cac is usually independent of polymer concentration. The size of these aggregates is usually smaller than that of free micelles, and this size tends to be small even in the presence of added salt (conditions where free micelles tend to grow in size).

These effects are illustrated in [figure C2.3.18](#) for the case of SDS micellization as influenced by poly(ethylene oxide), PEO, and salt [111]. The breakpoints in the figure denote the cmc or cac. From [table C2.3.6](#) and [table C2.3.7](#) we see that the cmc of SDS is 8.2 mM in the absence of salt and polymer and is 1.4 mM in 0.1 M NaCl. The open symbols in [figure C2.3.18](#) show that in the absence of salt, 35 000 Dalton PEO (at 0.1 % w/w) depresses the cac to about 3.5 mM and 60 000 Dalton PEO (also at 0.1 % w/w) depresses the cac even further to about 2.3 mM. In 0.1 M NaCl both molecular weight samples of PEO depress the cac, to about 1 mM (relative to 1.4 mM in the absence of polymer), but the relative depression is much less than in the absence of salt.

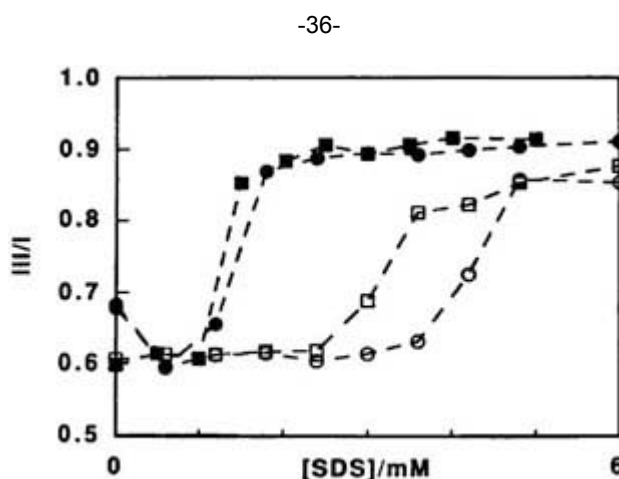


Figure C2.3.18. Vibronic peak fluorescence intensity ratio (III/I) as a function of SDS concentration for 0.1 % PEO solutions \circ , \bullet —35 000 Daltons; \square , \blacksquare —600 000 Daltons. Open symbols are for aqueous solution without added salt, and filled symbols are for 100 mM aqueous NaCl. Reproduced with permission from figure 2 of [111].

C2.3.16.2 EFFECTS ON RHEOLOGY

Solution viscosity in systems with strong polymer–micelle interactions generally increases with polymer–micelle binding. If one imagines the necklace model of Cabane and co-workers [112, 113 and 114], where micelles (beads) are bound to polymeric strands, two contributing features to viscosity may be identified. First, there exists an inertial effect due to the adsorbed micelles. These micelles serve to simply increase the effective molecular weight of the polymer and the effective friction coefficient of the polymer. In the neighbourhood of an adsorbed micelle, translation of the polymer strand in a reptation mode, for example, also requires translation of the attached micelle having considerably greater macromolecular cross-section. Second, occasional micelles may interact or bind to two different polymer strands. This kind of network formation also dramatically increases the effective molecular weight and friction coefficient of the polymer, as it introduces *de facto* cross-linking. This kind of cross-linking can also lead to the formation of gels at concentrations far below where the polymer would gel on its own.

A particularly interesting kind of sol–gel–sol sequence has been reported by Bloor *et al* [115] for SDS interactions

with ethyl-(hydroxyethyl)cellulose (EHEC). Below the cac the polymer and SDS are in solution and significant interactions among themselves or with each other are absent. As the cac is passed the SDS nucleates in micelles around the pendent ethyl groups. A fraction of these micelles attach to two or more pendent groups connecting different polymeric strands and thereby generating a gel network. This network grows with increasing SDS until a maximum number of cross-links are established. Further SDS additions displace some of these micelle–EHEC connections, and gradually dissolve the cross-links to produce another sol state comprising individual EHEC strands containing attached micelles.

REFERENCES

- [1] McBain J W 1913 Mobility of highly charged micelles *Trans. Faraday Soc.* **9** 99–101

-37-

- [2] Hartley G S 1935 The application of the Debye–Hückel theory to colloidal electrolyte *Trans. Faraday Soc.* **31** 31–50
- [3] Hartley G S 1948 State of solution of colloidal electrolytes *Q. Rev. Chem. Soc.* **2** 152–83
- [4] Menger F M 1979 On the structure of micelles *Accounts Chem. Res.* **12** 111–17
- [5] Fromherz P 1980 Micelle structure: a surfactant block model *Chem. Phys. Lett.* **77** 460
- [6] Dill K A and Flory P J 1981 Molecular organization in micelles and vesicles *Proc. Natl Acad. Sci. USA* **78** 676–80
- [7] Menger F M and Littau C A 1993 Gemini surfactants: Synthesis and properties *J. Am. Chem. Soc.* **113** 1451–2
- [8] Danino D, Talmon Y, Levy H, Beinert G and Zana R 1995 Branched threadlike micelles in an aqueous solution of a trimeric surfactant *Science* **269** 1420–1
- [9] Zana R 1996 Gemini (dimeric) surfactants *Curr. Opin. Colloid Interface Sci.* **1** 566–71
- [10] Friberg S E and Liang Y-C 1987 Nonaqueous microemulsions *Microemulsions: Structure and Dynamics* ed S E Friberg and Bothorel (Boca Raton, FL: Chemical Rubber Company) pp 79–91
- [11] McClain J B, Betts D E, Canelas D A, Samulski E T, DeSimone J, Londono J D, Cochran H D, Wignall G D, Chillura-Martino D and Triolo R 1996 Design of nonionic surfactants for supercritical carbon dioxide *Science* **274** 2049–52
- [12] Biermann M, Lange F, Piorr R, Ploog U, Rutzen H, Schinder J and Schmid R 1987 Synthesis of surfactants *Surfactants in Consumer Products* ed J Flabe (Berlin: Springer) pp 23–132
- [13] Laughlin R G 1978 Relative hydrophilicities among surfactant groups *Adv. Liq. Crystals* **3** 99–148
- [14] Alexandridis P and Hatton T A 1995 Poly(ethylene oxide)–poly(propylene oxide)–poly(ethylene oxide) block copolymer surfactants in aqueous solutions and at interfaces: thermodynamics, structure, dynamics, modeling *Colloids Surf. A* **96** 1–46
- [15] Texter J 1999 Characterization of surfactants *Surfactants—a Practical Handbook* ed K R Lange (Munich: Hanser) ch 1, pp 1–68
- [16] Sundell S 1977 The crystal structure of sodium dodecylsulfate *Acta Chem. Scand. A* **31** 799–807
- [17] Lundén B-M 1974 The crystal structure of *n*-dodecylammonium bromide *Acta Crystallogr. B* **42** 1756–60
- [18] Coiro V M, Manigrasso M, Mazza F and Pochetti G 1987 Structure of a triclinic phase of sodium dodecyl sulfate monohydrate. A comparison with other sodium dodecyl sulfate crystal phases *Acta Crystallogr. C* **43** 850–4
- [19] Adamson A W 1967 *Physical Chemistry of Surfaces* 2nd edn (New York: Interscience) pp 9–45
- [20] Texter J, Horch F R, Qutubuddin S and Dayalan E 1990 Voltammetric detection of micelle formation *J. Colloid Interface Sci.* **135** 263–71
- [21] Mukerjee P and Mysels K J 1970 *Critical Micelle Concentrations of Aqueous Surfactant Systems (National Standard Reference Data System, National Bureau of Standards Circular No 36)* (Springfield, VA: National Technical Information Service)
- [22] van Os N M, Haak J R and Rupert L A M 1993 *Physico-Chemical Properties of Selected Anionic, Cationic and Nonionic Surfactants* (Amsterdam: Elsevier)
- [23] Reynolds J A, Gilbert D B and Tanford C 1974 Empirical correlation between hydrophobic free energy and aqueous cavity surface area *Proc. Natl Acad. Sci. USA* **71** 2925–7

- [24] McAuliffe C 1966 Solubility in water of paraffin, cycloparaffin, olefin, acetylene, cycloolefin, and aromatic hydrocarbons *J. Phys. Chem.* **70** 1267–75
- [25] Gu T and Sjöblom J 1991 Empirical relationships between the Kraft points and the structural units in surfactants *Acta Chem. Scand.* **45** 762–5
- [26] Vinson P K 1990 Cryo-electron microscopy of microstructures in complex liquids *Thesis* University of Minnesota (Ann Arbor: University Microfilms)
- [27] Talmon Y 1996 Transmission electron microscopy of complex fluids: The state of the art *Ber. Bunsenges. Phys. Chem.* **100** 364–72

-38-

- [28] Israelachvili J N, Mitchell D J and Ninham B V 1976 Theory of self-assembly of hydrocarbon amphiphiles into micelles and bilayers *J. Chem. Soc. Faraday Trans. II* **72** 1525–68
- [29] Israelachvili J N 1992 *Intermolecular and Surface Forces* 2nd edn (London: Academic) pp 370–2
- [30] Evans D F and Wennerström H 1994 *The Colloidal Domain—Where Physics, Chemistry, Biology, and Technology Meet* (New York: VCH) pp 12–16
- [31] MacIntosh F C, Safran S A and Pincus P A 1990 Self-assembly of linear aggregates: the effect of electrostatics on growth *Europhys. Lett.* **12** 697–702
- [32] Lequeux F and Candau S J 1997 Structural properties of wormlike micelles *Theoretical Challenges in the Dynamics of Complex Fluids* ed McLeish (Dordrecht: Kluwer) pp 181–90
- [33] Lequeux F 1996 Structure and rheology of wormlike micelles *Curr. Opin. Colloid Interface Sci.* **1** 341–4
- [34] Magid L 1998 The surfactant–polyelectrolyte analogy *J. Phys. Chem. B* **102** 4064–74
- [35] Haan S W and Pratt L R 1981 Monte Carlo study of a simple model for micelle structure *Chem. Phys. Lett.* **79** 436–40
- [36] Jönsson R, Edholm O and Teleman O 1986 Molecular dynamics simulations of a sodium octanoate micelle in aqueous solution *J. Chem. Phys.* **85** 2259–71
- [37] Allen M P and Tildesley D J 1993 *Computer Simulation in Chemical Physics* (Dordrecht: Kluwer)
- [38] Allen M P and Tildesley D J 1987 *Computer Simulation of Liquids* (Oxford: Clarendon)
- [39] Goodfellow J M 1990 *Molecular Dynamics—Applications in Molecular Biology* (Boca Raton, FL: Chemical Rubber Company)
- [40] van Os N M and Karaborni S (eds) 1993 *Tenside* **30** 234–93
- [41] Karaborni S and O’Connell J P 1993 Molecular dynamics simulations of model chain molecules and aggregates including surfactants and micelles *Tenside* **30** 235–42
- [42] Smit B, Hilbers P A J and Esselink K 1993 Computer simulations of simple oil/water/surfactants systems *Tenside* **30** 287–93
- [43] Care C M, Dalby T and Desplat J-C 1997 Micelle formation in a lattice model of an amphiphile and solvent mixture *Prog. Colloid Polym. Sci.* **103** 130–37
- [44] Nguyen-Misra M, Misra S, Wang Y, Rodrigues K and Mattice W L 1997 Simulation of self-assembly in solution by triblock copolymers with sticky blocks at their ends *Prog. Colloid Polym. Sci.* **103** 138–45
- [45] Eicke H F and Christian H 1978 On the stability of micelles in apolar media. *J. Colloid Interface Sci.* **46** 417–27
- [46] Texter J, Antalek B and Williams A J 1997 Reverse micelle to sponge phase transition *J. Chem. Phys.* **106** 7869–72
- [47] Chevalier Y and Zemb T 1990 The structure of micelles and microemulsions *Rep. Prog. Phys.* **53** 279–371
- [48] Hasegawa M, Sugimura T, Shindo Y and Kitahara A 1996 Structure and properties of AOT reversed micelles as studied by the fluorescence probe technique *Colloids Surf. A* **109** 305–18
- [49] King A D 1995 Solubilization of gases *Solubilization in Surfactant Aggregates* ed S D Christian and J F Scamehorn (New York: Dekker) pp 35–58
- [50] Christian S D and Scamehorn J F (eds) 1995 *Solubilization in Surfactant Aggregates* (New York: Dekker)
- [51] Sharma R (ed) 1995 *Surfactant Adsorption and Surface Solubilization* (Washington, DC: American Chemical Society)
- [52] Nagarajan R 1996 Solubilization in aqueous solutions of amphiphiles *Curr. Opin. Colloid Interface Sci.* **1** 391–401
- [53] Alexiev A, Rubio S, Deyanova M, Stoyanova A, Sicilia D and Pérez-Bendito D 1994 Improved catalytic photometric determination of iron (III) in cetylpyridinium pre-micellar aggregates *Anal. Chim. Acta* **295** 211–19

- [54] Carreto M L, Rubio S and Pérez-Bendito D 1996 Organic microheterogeneous systems in kinetic analysis—Self-assembled systems *Analyst* **121** 33R–44R
- [55] Bunton C A, Moffatt J R and Rodenas E 1982 Abnormally high nucleophilicity of micelle-bound azide ion *J. Am. Chem. Soc.* **104** 2653–5
-

-39-

- [56] Fendler J H 1982 *Membrane Mimetic Chemistry* (New York: Wiley) chs 11 and 12, pp 293–491
- [57] Turro N J and Cherry W R 1978 Photoreaction in detergent solutions. Enhancement of regioselectivity resulting from the reduced dimensionality of substrates sequestered in a micelle *J. Am. Chem. Soc.* **100** 7431–2
- [58] Luisi P L 1996 Self-reproduction of micelles and vesicles: models for the mechanisms of life from the perspective of compartmented chemistry *Advances in Chemical Physics* vol XCII, ed I Prigogine and S A Rice (New York: Wiley) pp 425–38
- [59] Bachmann P A, Luisi P L and Lang J 1992 Autocatalytic self-replicating micelles as models for prebiotic structures *Nature* **357** 57–9
- [60] Pileni M-P (ed) 1989 *Structure and Reactivity in Reverse Micelles* (Amsterdam: Elsevier)
- [61] Luisi P L and Straub B E (eds) 1984 *Reverse Micelles* (New York: Plenum)
- [62] Friberg S E and Bothorel P (eds) 1987 *Microemulsions: Structure and Dynamics* (Boca Raton, FL: Chemical Rubber Company)
- [63] Rosano H L and Clause M (eds) 1987 *Microemulsion Systems* (New York: Dekker)
- [64] Chen S-H, Huang J S and Tartaglia P (eds) 1992 *Structure and Dynamics of Strongly Interacting Colloids and Supramolecular Aggregates in Solution* (Dordrecht: Kluwer) pp 229–429
- [65] Shah D O (ed) 1985 *Macro- and Microemulsions—Theory and Applications* (Washington, DC: American Chemical Society)
- [66] Zana R 1994 Microemulsions *Heterog. Chem. Rev.* **1** 145–57
- [67] Odian G 1981 *Principles of Polymerization* (New York: Wiley)
- [68] Piirma I 1976 *Emulsion Polymerization* (New York: Academic)
- [69] Lee Y-S, Yang J-Z, Sisson T M, Frankel D A, Gleeson J T, Aksay E, Keller S L, Gruner S M and O'Brien D F 1995 Polymerization of nonlamellar lipid assemblies *J. Am. Chem. Soc.* **117** 5573–8
- [70] Srisiri W, Sisson T M, O'Brien D F, McGrath K M, Han Y and Gruner S M 1997 Polymerization of the inverted hexagonal phase *J. Am. Chem. Soc.* **119** 4866–73
- [71] Burban J H, He M and Cussler E L 1995 Silica gels made by bicontinuous microemulsion polymerization *AIChE J.* **41** 159–65
- [72] Burban J H, He M and Cussler E L 1995 Organic microporous materials made by bicontinuous microemulsion polymerization *AIChE J.* **41** 907–14
- [73] Dunn A S 1989 Polymerization in micelles and microemulsions *Comprehensive Polymer Science—the Synthesis, Characterization, Reactions and Applications of Polymers* vol 4, ed G C Eastmond, A Ledwith, S Russo and P Sigwalt (New York: Pergamon) pp 219–24
- [74] Desai S D, Gordon R D, Gronda A M and Cussler E L 1996 Polymerized microemulsions *Curr. Opin. Colloid Interface Sci.* **1** 519–22
- [75] Barton J 1996 Free-radical polymerization in inverse microemulsions *Prog. Polym. Sci.* **21** 399–438
- [76] Hammouda A, Gulik T and Pileni M-P 1995 Synthesis of nanosize latexes by reverse micelle polymerization *Langmuir* **11** 3656–9
- [77] Alexandridis P, Olsson U and Lindman B 1997 Structural polymorphism of amphiphilic copolymers: Six lyotropic liquid crystalline and two solution phases in a poly(oxybutylene)–poly(oxyethylene)–water–xylene system *Langmuir* **13** 23–34
- [78] Wanka G, Hoffman H and Ulbricht W 1990 The aggregation behavior of poly-(oxyethylene)–poly(oxypropylene)–poly(oxyethylene)-block copolymers in aqueous solutions *Colloid Polym. Sci.* **268** 101–17
- [79] Luzzati V, Tardieu A, Gulik-Krzywicki T, Rivas E and Reiss-Husson F 1968 Structure of the cubic phases of lipid–water systems *Nature* **220** 485–8
- [80] Balmbra R R, Clunie J S and Godman J F 1969 Cubic mesomorphic phases *Nature* **222** 1159–60
- [81] Luzzati V, Delacroix H and Gulik A 1996 The micellar cubic phases of lipid-containing systems: Analogies with foams, relations with the infinite periodic minimal surfaces, sharpness of the polar/apolar partition *J. Physique. II* **6** 405–18

- [82] Mortensen K 1996 Structural studies of PEO–PPO–PEO triblock copolymers, their micellar aggregates and mesophases; a small-angle neutron scattering study *J. Phys.: Condens Matter* **8** A103–A104
- [83] Berrett J F, Molino F, Porte G, Diat O and Lindner P 1996 The shear-induced transition between oriented textures and layer-sliding-mediated flows in a micellar cubic crystal *J. Phys.: Condens Matter* **8** 9513–17
- [84] Mortensen K 1998 Structural properties of self-assembled polymeric micelles *Curr. Opin. Colloid Interface Sci.* **3** 12–19
- [85] Seddon K M, Hogan J L, Warrender N A and Pebay-Peyroula E 1990 Structural studies of phospholipid cubic phases *Prog. Colloid Polym. Sci.* **81** 189–97
- [86] van der Schoot P and Cates M E 1994 Growth, static light scattering and spontaneous ordering of rodlike micelles *Langmuir* **10** 670–9
- [87] Odijk T 1996 Ordered phases of elongated micelles *Curr. Opin. Colloid Interface Sci.* **1** 337–40
- [88] Tarazona A, Kreisig S, Koglin E and Schwuger M J 1997 Adsorption properties of two cationic surfactant classes on silver surfaces studied by means of SERS spectroscopy and *ab initio* calculations *Prog. Colloid Polym. Sci.* **103** 181–92
- [89] Manne S 1997 Visualizing self-assembly: Force microscopy of ionic surfactant aggregates at solid–liquid interfaces *Prog. Colloid Polym. Sci.* **103** 226–33
- [90] Manne S, Cleveland J P, Gaub H E, Stucky G D and Hansma P K 1994 Direct visualization of surfactant hemimicelles by force microscopy of the electrical double layer *Langmuir* **10** 4409–13
- [91] Manne S and Gaub H E 1995 Molecular organization of surfactants at solid–liquid interfaces *Science* **270** 1480–3
- [92] Wanless E J and Ducker W A 1996 Organization of sodium dodecyl sulfate at the graphite-solution interface *J. Phys. Chem.* **100** 3207–14
- [93] Saleeb F Z and Kitchener J A 1965 The effect of graphitization on the absorption of surfactants by carbon black *J. Chem. Soc.* 911–17
- [94] Zettlemoyer A C 1968 Hydrophobic surfaces *J. Colloid Interface Sci.* **28** 343–69
- [95] Pashley R M and Israelachvili J N 1981 A comparison of surface forces and interfacial properties of mica in purified surfactant solutions *Colloids Surf.* **2** 169–87
- [96] Kékicheff P, Christenson H K and Ninham B W 1989 Adsorption of cetyltrimethylammonium bromide to mica surface below the critical micellar concentration *Colloid Surf.* **40** 31–41
- [97] Helm C A, Israelachvili J N and McGuiggan P M 1989 Molecular mechanisms and forces involved in the adhesion and fusion of amphiphilic bilayers *Science* **246** 919–22
- [98] Richetti P and Kékicheff P 1992 Direct measurement of depletion and structural forces in a micellar system *Phys. Rev. Lett.* **68** 1951–4
- [99] Pashley R M, McGuiggan P M, Ninham B W, Brady J and Evans D F 1986 Direct measurements of surface forces between bilayers of double-chained quaternary ammonium acetate and bromide surfactants *J. Phys. Chem.* **90** 1637–42
- [100] Leimbach J, Sigg J and Rupprecht H 1995 The existence of small surface aggregates–surface micelles on polar charged surfaces *Colloids Surf. A* **94** 1–11
- [101] Gu Y and Huang Z 1989 Thermodynamics of hemimicellization of cetyltrimethylammonium bromide at the silica gel/water interface *Colloids Surf.* **40** 71–6
- [102] Rutland M W and Parker J L 1994 Surface forces between silica surfaces in cationic surfactant solutions: Adsorption and bilayer formation at normal and high pH *Langmuir* **10** 1110–21
- [103] Partyka S, Linsheimer M and Faucompre B 1993 Aggregate formation at the solid–liquid interface: the calorimetric evidence *Colloids Surf. A* **76** 267–81
- [104] Söderlind E and Stilbs P 1993 A ^2H NMR study of two cationic surfactants adsorbed on silica particles *Langmuir* **9** 2024–34

- [105] Yeskie M A and Harwell J H 1988 On the structure of aggregates of adsorbed surfactants: The surface charge

density at the hemimicelle/admicelle transition *J. Phys. Chem.* **92** 2346–52

- [106] Goddard G D 1986 Polymer–surfactant interaction. Part 1. Uncharged water-soluble polymers and charged surfactants *Colloid Surf.* **19** 255–300
- [107] Goddard G D 1986 Polymer–surfactant interaction. Part 2. Polymer and surfactant of opposite charge *Colloid Surf.* **19** 301–22
- [108] Lindman B and Thalberg K 1993 Polymer–surfactant interactions—recent developments *Interactions of Surfactants with Polymers and Proteins* ed E Goddard and K P Ananthapadmanabhan (Boca Raton, FL: Chemical Rubber Company) pp 203–76
- [109] Hansson P and Lindman B 1996 Surfactant polymer interactions *Curr. Opin. Colloid Interface Sci.* **1** 604–13
- [110] Cabane B and Duplessix R 1985 Neutron scattering study of water-soluble polymers adsorbed on surfactant micelles *Colloids Surf.* **13** 19–33
- [111] van Stam J, Brown W, Fundin J, Almgren M and Lindblad C 1993 Interaction between sodium dodecylsulfate and poly(ethylene oxide) in aqueous systems *Colloid–Polymer Interactions: Particulate, Amphiphilic, and Biological Surfaces* ed P L Dubin and P Tong (Washington, DC: American Chemical Society) pp 194–215
- [112] Cabane B 1977 Structure of some polymer–detergent aggregates in water *J. Phys. Chem.* **81** 1639–45
- [113] Cabane B and Duplessix R 1982 Organization of surfactant micelles adsorbed on a polymer molecule in water: A neutron scattering study *J. Physique* **43** 1529–42
- [114] Cabane B, Lindell K, Engstrom S and Lindman B 1996 Microphase separation in polymer + surfactant systems *Macromolecules* **29** 3188–97
- [115] Bloor D M, Wan-Yunis W M Z, Wan-Badhi W A, Li Y, Holzwarth J F and Wyn-Jones E 1995 Equilibrium and kinetic studies associated with the binding of sodium dodecyl sulfate to the polymers poly(propylene oxide) and ethyl-(hydroxyethyl)cellulose *Langmuir* **11** 3395–400
-

-1-

C2.4 Organics films (Langmuir-Blodgett films and self-assembled monolayers)

Georg Hähner

C2.4.0 INTRODUCTION

The goal of constructing stable organic molecular architectures with desired properties that modify surfaces independent from their bulk characteristics is of fundamental interest in many areas. Organic chemists have made significant progress over recent years in designing and constructing molecules that have certain desired physical and chemical properties. Assembling such molecules onto surfaces, as well as characterizing and studying the resulting layers, lies at the centre of interest in many laboratories, and a variety of techniques have been employed in order to reach this goal. Ultrathin and especially monomolecular organic films with a high degree of order are of special interest since they open up new fields of research and could establish an organic counterpart to inorganic crystals. Such films play an important role not only in fundamental science, where they often serve as model systems (e.g. for polymers) but also in applied sciences, where they are employed as corrosion inhibitors, lubricants, adhesion promoters and in biosensors, as well as in many other applications.

Inorganic surfaces can be covered with organic molecules by different methods. The Langmuir–Blodgett (LB) and the self-assembly (SA) techniques that are described here both offer the possibility of tailoring the properties of ultrathin organic films to a certain degree. The LB technique is suitable only for a limited number of molecules and substrates and requires special equipment for preparation. It was the first technique that gave chemists a practical method by which to prepare ordered molecular structures on surfaces.

Adsorption from solution, which is most often employed in connection with the SA technique, provides the easiest

route for studying the behaviour of organic molecules. Organic films can be prepared by immersing the inorganic substrates in dilute solutions of the surfactant. However, not all molecules and substrates are appropriate for establishing self-assembled monolayers (SAMs).

Apart from the techniques described in this chapter other methods of organic film formation are vacuum deposition or film formation by allowing a melt or a solution of the material to spread on the substrate and subsequently to solidify. Vacuum deposition is limited to molecules with a sufficiently high vapour pressure while a prerequisite for the latter is an even spreading of the solution or melt over the substrate, which depends on the nature of the intermolecular forces. This subject is of general relevance to the formation of organic films.

Excellent books and review articles covering LB and SA films have appeared recently. The following covers the basics and some selected topics are presented as examples. For a more comprehensive overview and more details on specific topics the reader is referred to the cited literature.

-2-

C2.4.1 LANGMUIR–BLODGETT FILMS

The systematic study of ultrathin and, in particular, monomolecular organic films that are ordered originates from the end of the nineteenth century, and it was at that time when the first truly quantitative studies of amphiphilic (i.e. one end of the molecule is polar and hydrophilic, the other end is hydrophobic) monolayers were made by Pockels [1]. She demonstrated that in the case of amphiphilic molecules such as stearic acid ($C_{17}H_{35}CO_2H$) there exists a unique form of layer at the air–water interface, having a definite ratio of mass of stearic acid to surface area of water on which it resides [2, 3 and 4]. Using these results and the known density of stearic acid, one can deduce a thickness of these layers of about 2.3 nm, which compares well with the modern value of 2.5 nm for the length of such a molecule. Lord Rayleigh suggested that these films were monolayers and thus gave a direct measure of molecular dimensions [5]. Subsequent work showed that only amphiphilic molecules form good monolayers whereas simple aliphatic ones do not [6].

In 1917, Langmuir published his systematic study of amphiphilic compounds at the air–water interface [7]. In 1920, he mentioned the transfer of films from this interface to a solid substrate [8]. In 1935 K Blodgett published an extensive report on the deposition of mono- and multilayers of fatty acids on a solid substrate from films existing at the air–water interface [9]. In the following 30 years a number of publications appeared dealing with the properties of such films. The term *Langmuir–Blodgett* films (LB films) is usually employed to denote mono- or multilayers transferred from a liquid–gas interface onto a solid substrate. The molecular film at the (liquid–gas) interface itself is denoted a *Langmuir* film.

C2.4.1.1 PREPARATION OF LB FILMS

(A) LB MOLECULES (AMPHIPHILES)

It is well known from everyday life that there are substances that dissolve in water and others that are insoluble. Many salts, for example, are soluble in water, while lipids are not. However, the latter are soluble in nonpolar solvents, such as CCl_4 . This is due to the different interactions between the solvent and the solute. Materials that ‘like’ water (i.e. polar ones) are called hydrophilic, while those which do not like water are called hydrophobic. An amphiphile is a molecule that is not soluble in water, but has a hydrophilic and a hydrophobic end group (figure C2.4.1). Therefore one end goes into water while the other points out, resulting in a spreading of the material on the water surface. A typical example is stearic acid ($C_{17}H_{35}CO_2H$), where the long hydrocarbon chain is hydrophobic while the carboxyl group ($-COOH$), which can dissociate in water and become negatively charged, is hydrophilic.

-3-

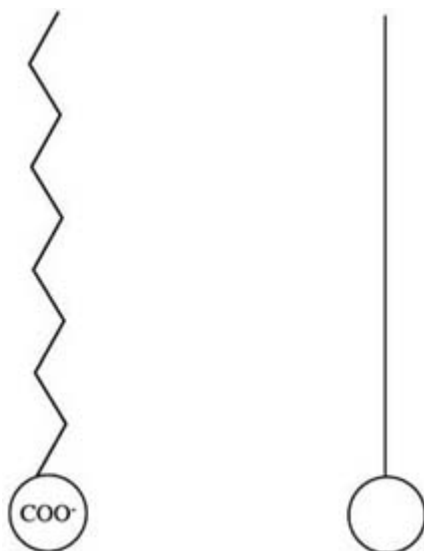


Figure C2.4.1. Schematic diagram of a fatty acid with a hydrophilic (COO^-) and a hydrophobic end group (CH_3) (left) and of an amphiphile in general (right).

(B) TROUGH

The basic equipment for the preparation of LB films is a trough containing the subphase on which the compound is spread and which is equipped with barriers in order to manipulate the film at the liquid–gas interface ([figure C2.4.2](#)). Both the trough and the barrier are frequently made out of Teflon. This material is very inert and can be cleaned with strong oxidizing materials without any damage. The movable barrier allows the pressure exerted on the film to be controlled. All movements are performed by a motor. The substrate can be lowered and raised by means of another motor with a gearbox in order to transfer the film from the interface onto the substrate (see *(e)* below). The balance used to measure the pressure in the film is most often a so-called Wilhelmy plate. Discussion of methods for surface pressure measurements can for example be found in [10]. The full automatization and computerization for the preparation of monomolecular and multilayer films started in the early 1970s [11].

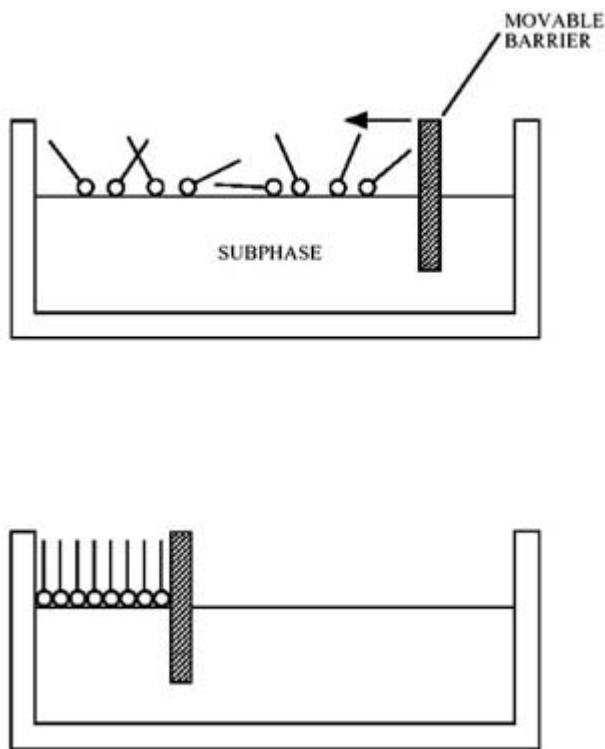


Figure C2.4.2. Schematic sideview of the trough. The movable barrier is used to push the molecules on the subphase together in the Langmuir film which is subsequently transferred to a solid substrate.

Demand for temperature controlled troughs came from the material scientists who worked with large molecules and polymers that establish viscous films. Such troughs allow a deeper understanding of the distinct phases and the transitions in LB films and give more complete pressure–area isotherms (see (d) below).

In general, extreme care has to be taken when LB films are prepared, since the quality of the resulting films depends crucially on the preparation conditions. The best place for an LB trough is a laboratory where the surroundings, i.e. temperature, humidity and atmosphere, are completely controlled. Often it is placed in a laminar flow box. Also, the trough should be installed in a shock-free environment.

(C) SUBPHASE

The most often used subphase is water. Mercury and other liquids [12], such as glycerol, have also occasionally been used [13, 14]. The water has to be of ultrapure quality. The pH value of the subphase has to be adjusted and must be controlled, as well as the ion concentration. Different amphiphiles are differently sensitive to these parameters. In general it takes some time until the whole system is in equilibrium and the final values of pressure and other variables are reached. Organic contaminants cannot always be removed completely. Such contaminants, as well as ions, can have a harmful influence on the film preparation. In general, all chemicals and materials used in the film preparation have to be extremely pure and clean.

(D) PRESSURE–AREA (Π –A) ISOTHERM

A drop of a dilute solution (1%) of an amphiphile in a solvent is typically placed on the water surface. The solvent evaporates, leaving behind a monolayer of molecules, which can be described as a two-dimensional gas, due to the large separation between the molecules (figure C2.4.3). The movable barrier pushes the molecules at the surface closer together, while pressure and area per molecule are recorded. The pressure–area isotherm yields information about the stability of monolayers at the water surface, a possible reorientation of the molecules in the two-dimensional system, phase transitions and changes in the conformation. While being pushed together, the layer at

the water surface goes through different states: gaseous, liquid and solid. If the pressure is increased further, the layer collapses due to mechanical instabilities. This collapse is detected as a sharp decrease in the pressure. This break-down pressure is a function of temperature, pH, the subphase and the velocity with which the barrier moves.

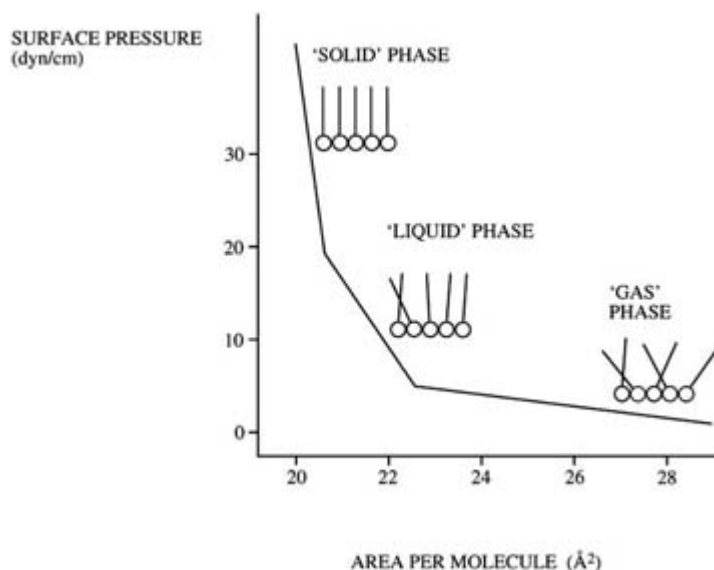


Figure C2.4.3. Pressure–area isotherm for a fatty acid. The molecules are in a gaseous, liquid or solid state, depending on the area per molecule available. If the pressure is further increased, a mechanical instability occurs and the film breaks down.

(E) FILM TRANSFER

Vertical deposition. The conventional method of transferring films from the air–water interphase onto a solid substrate is vertical deposition, which was demonstrated by Langmuir and Blodgett ([figure C2.4.4](#)). They showed that a monolayer of an amphiphile can be transferred to a substrate by moving a vertical plate through the film at the water–air interface. During transfer the Langmuir monolayer is held at constant surface pressure. The transfer process itself has not yet been fully understood, although it is known that there is a critical velocity above which it does not work. The effects of various parameters, such as viscosity, on the critical velocity have been investigated [[15](#)].

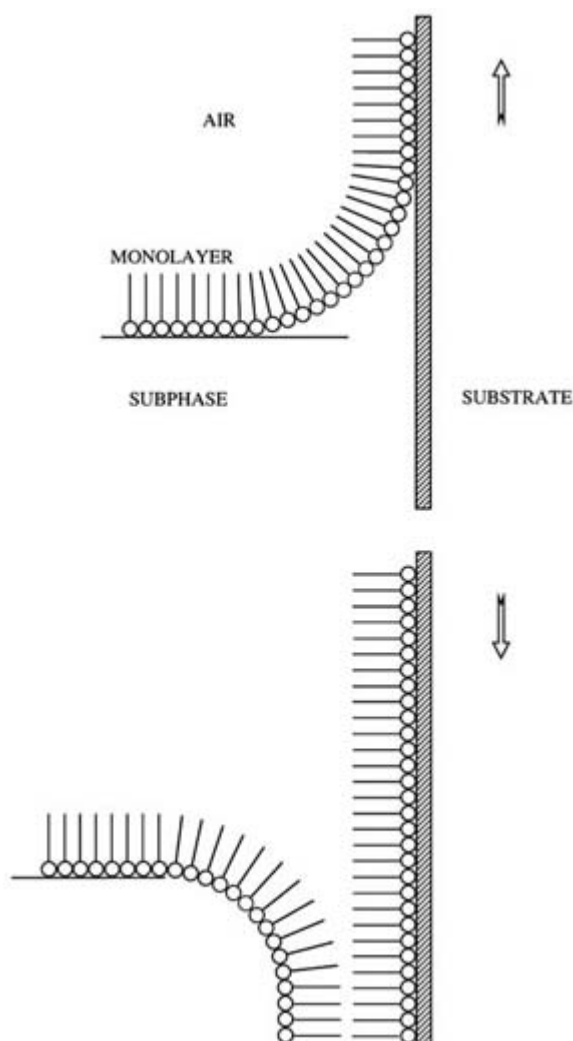


Figure C2.4.4. Schematic diagram of the transfer process of LB films onto a hydrophilic substrate. Vertical upward and downward strokes result in hydrophobic and hydrophilic surfaces, respectively.

If a substrate is moved through a monolayer at the water–air interface, a film can be deposited by both dipping the substrate into the water and retracting it. Usually the film is transferred while retracting, if the substrate is hydrophilic and the hydrophilic headgroups are interacting with the surface. On the other hand, if the substrate is hydrophobic, the film is transferred while dipping, as the hydrophobic alkyl chains interact with the surface. Thus, multilayers can be prepared by several subsequent transfer processes. If the transfer process starts with a hydrophilic substrate the surface will be hydrophobic after the first film transfer, hydrophilic after the second one and so on. This transfer mode is called Y-type deposition, resulting in multilayers with ‘head to head’ and ‘tail to tail’ configurations of the layers. Films can, however, also be transferred only in downstroke mode resulting in so-called X-type layers (‘head to tail’ configuration) or in upstroke mode only resulting in Z-type layers (‘tail to head’ configuration). The different transfer modes have specific advantages and disadvantages—in general, the Y-type (multi)layers are the most stable ones for very hydrophilic headgroups [16].

Horizontal transfer (Schaefer’s method). Another technique to prepare structures with LB (multi)layers is named after Schaefer [17]. This method is useful for depositing rigid films which can be described as two-dimensional solids. First, a compressed monolayer is established at the water–air interface. Subsequently a flat substrate is brought horizontally into contact with the film (figure C2.4.5). When the substrate is lifted and separated from the water surface a monolayer is transferred to the substrate while (theoretically) maintaining the molecular order.

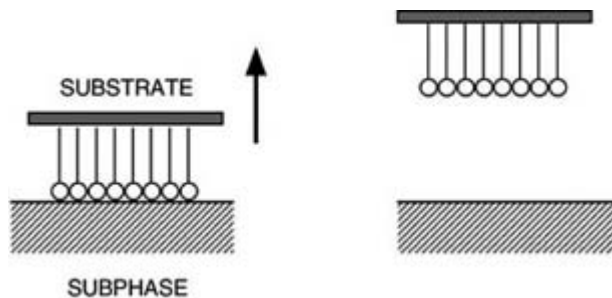


Figure C2.4.5. Horizontal transfer on a hydrophobic substrate. This method is useful for very rigid films that are in the 'solid state' in the π -A-diagram.

(F) SUBSTRATES

Monolayers can be transferred onto many different substrates. Most LB depositions have been performed onto hydrophilic substrates, where monolayers are transferred when pulling the substrate out from the subphase. Transparent hydrophilic substrates such as glass [18, 19] or quartz [20] allow spectra to be recorded in transmission mode. Examples of other hydrophilic substrates are aluminium [21, 22, 23 and 24], chromium [9, 25] or tin [26], all in their oxidized state. The substrate most often used today is silicon wafer. Gold does not establish an oxide layer and is therefore used chiefly for reflection studies. Also used are silver [27], gallium arsenide [27, 28] or cadmium telluride wafer [28] following special treatment.

C2.4.2 EXAMPLES OF LB FILMS

(A) FATTY ACIDS

LB films of fatty acids are still studied today, particularly by those researchers who are interested in the basic physics of the subject. The literature on this topic is very extensive. The study of such films by means of x-ray diffraction revealed that the order of LB films in the direction normal to the substrate and of the lattice planes is extremely good. This fact was initially responsible for the widespread enthusiasm for the study of LB films. Structural aspects of such films are discussed in detail in the review article by Schwartz [29]. A number of authors have found that long-chain fatty acid (and fatty acid salt) LB monolayers deposited on a variety of hydrophilic substrates and under a variety of conditions adopt a hexagonal packing with the chain normal to the surface (untilted) on average [29] and short-range translational order. However, there are exceptions, such as calcium arachidate (CaA_2), which is tilted by 20–30° from the normal on Si oxide [29, 30 and 31]. Also LB monolayers of phospholipids often display hexagonal packing of the chains [29]. However, in most cases multilayers display different packing from that of monolayers [29]. They are typically packed in a crystalline rectangular or triclinic lattice and in most cases the transition from monolayer to

bulk structure is abrupt [29]. Techniques that can give direct information on in-plane order in LB films are electron diffraction and polarized-light optical microscopy [28, 29].

In most cases, once a two-dimensional structure has been established, it will propagate through the film as further layers are deposited [32]. It has been suggested that disclination structures that can be found in the resulting films exist already at the water–air interface [33] and are transferred to the substrate in the initial layer. It has also been proposed that annealing of the film at the water–air interface can greatly reduce the density of disclinations so that subsequent multilayers will also contain a low disclination density [34]. It is supposed that the regions immediately associated with the disclinations are responsible for electron conduction through the film.

Electron tunnelling through monolayers of long-chain carboxylic acids is one aspect of interest since it was assumed that such films could be used as gate electrodes in field-effect transistors or even in devices depending on electron tunnelling [24, 26, 35, 36, 37 and 38]. It was found, however, that the whole subject depends critically on

the materials involved, especially on the metal used as substrate and electrode. It seems that conduction through monolayers can be best understood by conduction through defects [39, 40, 41 and 42]. In addition to the defects due to disclinations on a small scale, there will be fluctuations in the distances between molecules on a somewhat larger scale.

Apart from fatty acids, straight-chain molecules containing other hydrophilic end groups have been employed in numerous studies. In order to stabilize LB films chemical entities such as the alcohol group and the methyl ester group have been introduced, both of which are less hydrophilic than carboxylic acids and are largely unaffected by the pH of the subphase.

New factors for the establishment of multilayer structures are, for example, the replacement of the hydrocarbon chain by a perfluorinated chain and the use of a subphase containing multivalent ions [29]. The latter can become incorporated into an LB film during deposition. The amount depends on the pH of the subphase and the individual ion. The replacement of the hydrocarbon by a rodlike fluorocarbon chain is one way to increase van der Waals' interaction and therefore enhance order and stability in molecular assemblies [43].

Remarkably, such fluorocarbon monolayers show higher friction than their hydrocarbon counterparts [44], although fluorocarbons are known to have the lowest surface free energy of all organic materials.

Mechanical stability. The LB technique can be used to force the ordering of long-chain molecules. They are forced to order on a liquid subphase and are subsequently transferred to a solid substrate. Although order in these materials can be high, achieving this is a potentially non-equilibrium and difficult procedure. The monolayer-substrate bond is often weak (either van der Waals or hydrogen-bond interactions); as a result, the assemblies are not very mechanically or thermally stable [45, 46]. Stable and high-surface-free-energy monolayers are difficult to prepare with this technique. The mechanical stability of a variety of systems, mainly arachidic ones, has been investigated with the atomic force microscope. Details and references can be found in the review article by Schwartz [29].

Thermal stability. For applications of LB films, temperature stability is an important parameter. Different techniques have been employed to study this property for mono- and multilayers of arachidate LB films. In general, an increase in temperature is connected with a conformational disorder in the films and above 390 K the order present in the films seems to vanish completely [45, 46 and 47]. However, a comprehensive picture for order-disorder transitions in mono- and multilayer systems cannot be given. Nevertheless, some general properties are found in all systems [47]. *Gauche* conformations mostly reside at the ends of the chains at room temperature, but are also present inside the

films at higher temperatures. The energy connected with a *gauche* conformation in a densely packed film of calcium arachidate (CaA_2) was found to be more than an order of magnitude higher compared to the liquid state [48].

(B) RODLIKE STRUCTURES WITH CYCLIC GROUPS

In addition to the rodlike molecules described above, amphiphiles, which consist of a rodlike structure containing a cyclic group near its centre, have also been studied. Of great interest here are materials containing both the azobenzene and the stilbene structure (figure C2.4.6) [49]. Interestingly, these are extended conjugated structures involving two rings and are thus constrained to remain approximately in the same plane (*trans* configuration). There are, however, *cis* conformations for both molecules. The reversible change between the two conformations by an irradiation are connected with different absorption spectra. This so-called photo-chromic effect is interesting because of potential applications, for example, in recording media. Materials containing modifications of these groups have been employed in devices intended to generate optical second harmonics. Studies of amphiphilic materials containing these groups are for example described in [50, 51, 52, 53 and 54].

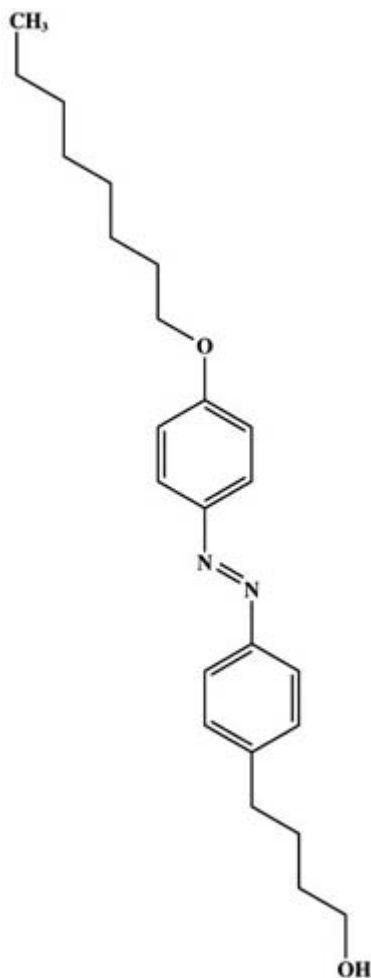


Figure C2.4.6. Azobenzene structure modified with hydrocarbon chains [53, 54].

-10-

Success of depositing compounds where an 18-carbon chain was attached to one end of an azobenzene group and various different hydrophilic groups attached to the other end has been reported in X and Z mode [52] and piezo- and pyroelectric effects were demonstrated.

Many workers have used the *cis-to-trans* structural change referred to above and brought about by UV irradiation to change some physical parameter of the LB films formed from azobenzene derivatives [55, 56, 57, 58, 59 and 60].

(C) PORPHYRINS AND PHTHALOCYANINES

The porphyrin is one of the most important among biomolecules. It is involved in the fundamental processes of life, such as oxygen transfer and storage, electron transfer, and synthesis of amino acids [61]. Phthalocyanine is a very stable, planar, synthetic aromatic macrocycle. The basic porphyrin and phthalocyanine structures are shown in [figure C2.4.7](#). In the case of porphyrins the positions bearing numbers are capable of having various groups attached to them, though many of these positions are usually occupied by hydrogen. In the case of phthalocyanine, groups can be attached to the periphery of the benzene rings. Both molecules have a planar structure and a number of relatively low-lying excited states. Due to the latter property, it seems likely that interesting devices could be made from films in which derivatives of these materials are incorporated, which has led to great interest in these compounds. In addition the structures are extremely stable. Porphyrins, for example, can survive the fractioning process applied to petroleum and the phthalocyanine group is stable to 400°C [62]. Both materials can be complexed with divalent metals, which reside at the centre of the ring.

LB films of porphyrin and phthalocyanine derivatives can be made in different ways.

Molecules that have been rendered amphiphilic are deposited in the Y mode so that the planes of the molecules are nearly vertical with respect to the film plane. In the mid-1980s it was shown that certain derivatives of porphyrin can be used to obtain good Y layers [63, 64 and 65]. On another derivative electron diffraction studies led to the conclusion that films consist of crystallites formed from tilted molecules [66]. It seems likely that all these materials rearrange after deposition to form many small crystallites, which do not have crystal planes corresponding to the original LB planar structure.

In some cases it has been possible to prepare LB films from porphyrin and phthalocyanine containing fourfold symmetry. However, it is not entirely clear how such materials can be deposited by the LB technique. A number of studies deals with LB films made of non-amphiphilic phthalocyanines. However, it is very difficult to form LB films from symmetric porphyrins [67]. True LB deposition of such compounds leads to an edge-on structure, whereas related techniques can lead to a structure in which the molecular planes lie parallel to the substrate [68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78 and 79].

Another possibility is that the ring structure may have long hydrocarbon chains attached at the corners so that they stand up at one side. These chains provide the hydrophobic component and the polarizable ring structure provides the hydrophilic moiety. There are studies with porphyrins bearing four long hydrocarbon chains and whose hydrophilic moieties are associated with the ring structure [80, 81]. However, these materials did not lead to the formation of ordered multilayers. The same general principle applied to phthalocyanines led to stable films at the water–air interface and could produce multilayers by the LB technique [82, 83].

-11-

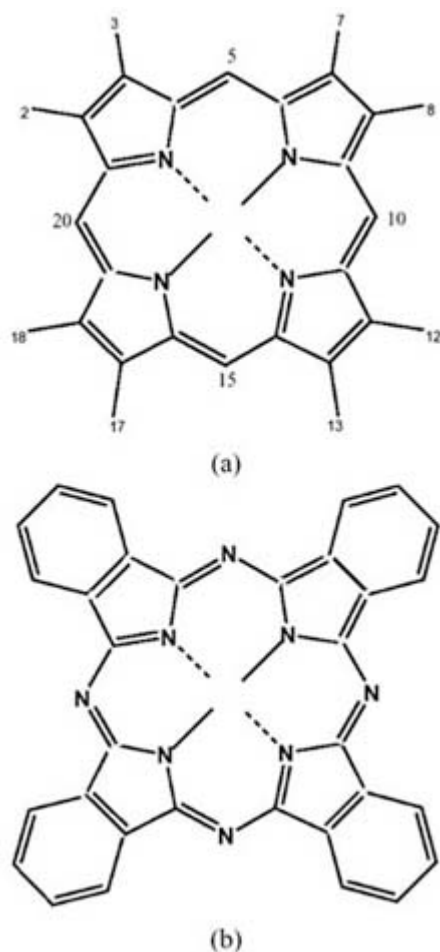


Figure C2.4.7. Porphyrin (a) and phthalocyanine (b) structures.

In summary, a vast number of materials has been used to form LB films. However, in the majority of cases an effort to characterize the film structure or even to show that a regular layer structure has been achieved is lacking. Work on the structure of films of disc-like molecules such as porphyrins and phthalocyanines is especially limited. Some references can be found in [29].

(D) MORE COMPLEX STRUCTURES (POLYMERS)

The generally low chemical, mechanical and thermal stability of LB films hinders their use in a wide range of applications. Two approaches have been studied to solve this problem. One is to spread a polymerizable monomer on the substrate and to polymerize it either before or following transfer to the substrate. The second is to employ preformed polymers containing hydrophilic and hydrophobic groups.

LB films made of more complex structures such as polymers can be divided into different classes [84].

-12-

(E) POST-FORMED POLYMERS FROM MONOMERS CONTAINING ONE OR MORE DOUBLE BONDS

These are systems in which multilayer structures are formed from molecules containing one or more double bonds and in which polymerization is subsequently initiated by appropriate means such as electron beam or UV light exposure.

A study of polymerization after deposition is of interest to polymer chemists since it is possible to arrange the monomers so that the polymerizable groups are adjacent to one another and to monitor changes in film structure arising from polymerization. The system under investigation is thus under control to a large degree. The first work published on this was upon multilayers of vinyl stearate and subsequent polymerization by γ -rays [85]. Under appropriate conditions a high level of polymerization was observed.

Polymerization of compounds performed with UV light was first reported in the 1970s [86] and was followed by further studies [87, 88 and 89]. Another study was concerned with the deposition and polymerization of multilayers of alcohols and acids incorporating the diene group, $-\text{CH}=\text{CH}-\text{CH}=\text{CH}-$, at the hydrophilic end of the molecule [90].

Other investigations dealt with straight-chain molecules (ω -tricosenoic acid) in which the penultimate and final carbon atoms at the hydrophobic end are connected by a double bond [91, 92]. The material does not polymerize as rapidly as those described before when irradiated by UV light, however, but it is readily polymerized when bombarded with an electron beam. It was thus thought to be an optimal material for the fabrication of electron beam resists.

The deposition and photo-polymerization of relatively complex amphiphilic compounds having two hydrophobic chains attached to a single hydrophilic headgroup have also been studied [93, 94]. This work is in the direction of materials for forming stable bio-compatible coatings for artificial organs, the most important outcome being the formation of stable multilayers having a hydrophilic outer surface.

(F) DIACETYLENES

Similar systems to those mentioned above exist where the constituent monomer contains the diacetylene group.

A summary of the studies performed on symmetrical compounds having a diacetylene group at the centre is given in [94]. Most of the materials studied in the context of LB films have been diyonic acids (figure C2.4.8).

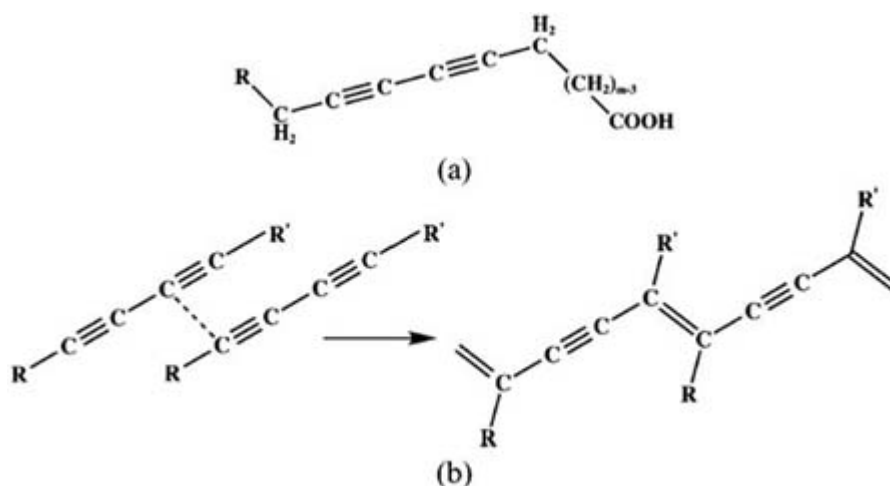


Figure C2.4.8 Diacetylene structure employed to prepare polymeric LB films (a) and principle in diacetylene polymerization (b).

If these materials are deposited as LB multilayers, polymerization can be induced either by thermal or optical means. This subject has been intensively studied [95, 96, 97, 98 and 99]. Since parameters such as m , subphase components, pH and polymerization before and after dipping, as well as temperature and wavelength employed for polymerization can be varied, the literature on diacetylenes is extensive and the reader is referred for example to the book of Tredgold [100].

(G) PREFORMED POLYMERS

Mono- and multilayers may be formed by the LB technique from polymers bearing both hydrophilic and hydrophobic side groups that are already spread as a polymer at the water–air interface.

Unwanted structures in the film plane—often found within LB films formed from simple rodlike molecules or from molecules polymerized after deposition—can be problematic, since many possible applications of such films require a uniform structure within the plane. On the other hand, however, the production of a system in which the structure within the plane is so disordered that there exist no structural features large enough to cause problems would also render applications possible. In three-dimensional materials, for example, both inorganic glasses and many polymers are capable of transmitting light without any appreciable scattering for substantial distances.

Studies of the waveguiding of light in multilayers of certain polymers showed that it is possible to propagate light with an attenuation that is still large compared to many other materials but small compared to other LB materials [101].

Another approach to the fabrication of LB films from preformed polymers is to form a hydrophobic main chain by reacting monomers terminated by a vinyl group [102, 103, 104, 105 and 106]. The side groups studied also included perfluorinated hydrocarbon chains, which tilt with respect to the normal to the plane of the film, whereas the analogous ordinary hydrocarbon chains do not [105].

Other polymers, such as polymethacrylates, have been studied, as well as esters of naturally occurring polysaccharides. References can be found in the literature cited in the list of further reading.

(H) RIGID-ROD POLYMERS

Finally, rigid-rod polymers can be deposited on a solid substrate by the LB technique. These materials have both

hydrophilic and hydrophobic characteristics, and are capable of residing with the rod axis horizontal at the water–air interface.

The polymers described so far have relatively flexible main chains which can result in complex conformations. In some cases, they can double back and cross over themselves. There are also investigations on polymers which are constrained to remain in a conformation corresponding, at least approximately, to a straight line, but which have amphiphilic properties that ensure that this line is parallel to the water surface. Chiral molecules are one example and many polypeptides fall into this class [107]. Another example is cofacial phthalocyanine polymers (figure C2.4.9).

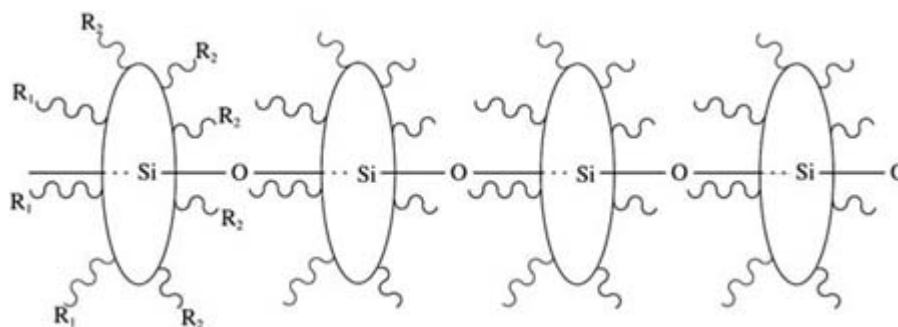


Figure C2.4.9. Part of a bridge-stacked polyphthalocyanine.

The species at the centre of the rings is usually Si or Ge and the bridging atom is oxygen. In one study the peripheral hydrogens on the phthalocyanine molecules were replaced by alkyl groups and the resulting polymers could be rendered soluble in ordinary organic solvents [108, 109 and 110]. Successful deposition of several of these materials has been achieved and different techniques were employed to study their structural properties [109, 111, 112, 113 and 114].

The variety of molecules used to prepare LB films is enormous, and only a small selection of examples can be presented here. Liquid crystals and biomolecules such as phospholipids, for example, can also be used to prepare LB films. The reader is referred to the literature for information about individual species.

C2.4.2 SELF-ASSEMBLED MONOLAYERS (SAMS)

Self-assembled monolayers (SAMs) are molecular layers that form spontaneously upon adsorption by immersing a substrate into a dilute solution of the surface-active material in an organic solvent [115]. This is probably the most comprehensive definition and includes compounds that adsorb spontaneously but are neither specifically bonded to the substrate nor have intermolecular interactions which force the molecules to organize themselves in the sense that a defined orientation is adopted. Some polymers, for example, belong to this class. They might be attached to the substrate via weak van der Waals' interactions only.

Most often the term SA is used in connection with compounds that attach strongly to the substrate and/or have significant intermolecular interactions. Order, orientation and stability in SA systems depend crucially on the compound involved. For establishing a lateral translational order the anchoring to the substrate and/or the intermolecular interactions are important. A highly ordered substrate, for example, may induce a high translational order if there is strong coupling between headgroups and substrate. This is, for example, the case for the alkyl thiol–gold system (see below).

The so-called self-assembly technique has its origin in 1946, when a paper was published by Bigelow *et al* [116] and thus is slightly younger than the LB technique. The authors noted that a hydrophilic surface exposed to an amphiphilic compound dissolved in a non-polar solvent induces the amphiphilic material to form a monolayer on it.

This idea was later extended to form multilayers by synthesizing a material with a hydrophilic group at one end and a further hydrophilic group, masked by a hydrophobic blocking group, at the other. After deposition on a hydrophilic surface by the technique introduced by Bigelow *et al*, the blocking groups are removed by a chemical reaction, revealing a further hydrophilic surface. Then the whole process can be repeated, which, in principle, should be capable of providing a simple way to produce ordered multilayers. In practice, however, there are many difficulties.

In the 1980s the study of SAMs was sparked by the use of octadecyltrichlorosilane (OTS) for film formation on silica surfaces [117] and by the investigation of dialkyldisulphide layers on gold [118], prepared by immersion of the substrates in diluted solutions of the surfactants. Higher-quality films on gold were obtained by the adsorption of structurally analogous alkyl thiols [119, 120]. These experiments initiated an avalanche of investigations with a plethora of SA systems, among which thiols on gold is the most intensively investigated combination to date.

C2.4.2.1 SA MOLECULES

Not all molecules are suited for establishing SAMs. The majority of cases studied have involved assembly of alkyl-chain-based entities. The molecules of self-organizing chemical compounds all have a similar structure. The spontaneous nature of film formation is due to the interaction energies of the monolayers. These can be considered in terms of three main components (figure C2.4.10) [121], which cooperatively establish stability, order and orientation in the monolayer.

-16-

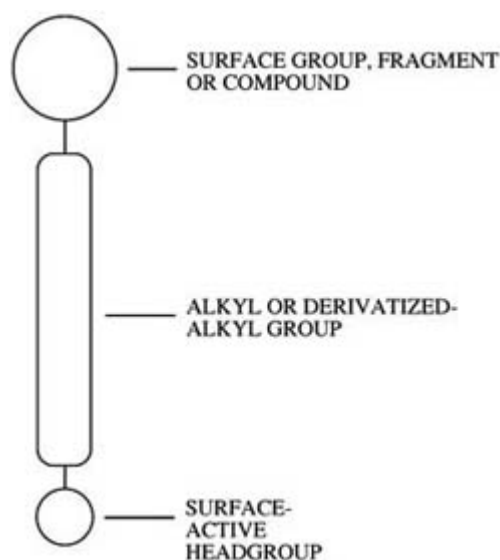


Figure C2.4.10. Schematic diagram of a self-assembling molecule.

The first part is the headgroup, which is responsible for the bonding to the substrate surface, which may be by chemisorption or physisorption.

In the case of chemisorption this is the most exothermic process and the strong molecule substrate interaction results in an anchoring of the headgroup at a certain surface site via a chemical bond. This bond can be covalent, covalent with a polar part or purely ionic. As a result of the exothermic interaction between the headgroup and the substrate, the molecules try to occupy each available surface site. Molecules that are already at the surface are pushed together during this process. Therefore, even for chemisorbed species, a certain surface mobility has to be anticipated before the molecules finally anchor. Otherwise the evolution of ordered structures could not be explained.

The spontaneous adsorption brings the molecules close enough together that intermolecular interactions become important; these—in the case of alkyl-chain-based molecules—consist of the short-range van der Waals’

interaction. The chains constitute the second part of alkane-based self-organizing molecules. If the molecules are densely packed in the final layer they are forced to stretch and are in a nearly all *trans*-configuration for sufficiently long chains at room temperature [122].

The self-organizing process of the amphiphilic alkane chains would not be possible solely due either to the interaction between the chains or the bonding of the molecules to the substrate. In fact it is a cooperative process of both factors and there are limiting cases where one or the other dominates. In the case of alkanethiols the first and, as mentioned earlier, the most important process, is chemisorption. The establishment of a well ordered and densely packed layer is only possible following the anchoring of molecules at the surface sites. Van der Waals' forces are the most important *intermolecular* interactions in the case of simple alkane chains. By substitution of the methylene groups in the chain by larger polar groups, long-range electrostatic interactions can also play a significant role and can become energetically more important than the short-range van der Waals' interactions.

-17-

If the coupling to the substrate is weak (physisorption), as is the case for alkylsiloxanes on a SiO_x surface in the presence of a water layer, for example, the packing may also be mainly driven by intermolecular forces. Stability in this system is provided by crosslinking between the molecules (see below).

The third part is the interaction between the terminal functionality, which in the case of simple alkane chains is a methyl group ($-\text{CH}_3$), and the ambient. These surface groups are disordered at room temperature as was experimentally shown by helium atom diffraction and infrared studies in the case of methyl-terminated monolayers [122]. The energy connected with this conformational disorder is of the order of some kT .

This third part can be substituted by a functional group, a small fragment or even a polymer, where alkanethiols are only used to attach the whole compound to the surface. This potential makes compounds modified with SA molecules attractive in a whole variety of areas and technologies.

C2.4.2.2 PREPARATION OF SAMs

In contrast to the preparation of LB films, that of SAMs is fairly simple and no special equipment is required. The inorganic substrate is simply immersed into a dilute solution of the surface active material in an organic solvent (typically in the mM range) and removed after an extended period (~24 h). Subsequently, the sample is rinsed extensively with the solvent to remove any excess material (wet chemical preparation).

Preparation of films for sufficiently volatile molecules can also be performed by evaporating the molecules in vacuum (gas-phase deposition) or by the use of a desiccator which contains the substrate and the dilute solution in a vessel separately and which is evacuated to 0.1 mbar and kept under vacuum for several hours (~24 h). This also results in a vapour-phase-like deposition of the molecules onto the substrates.

C2.4.2.3 EXAMPLES OF SAMs

A plethora of different SA systems have been reported in the literature. Examples include organosilanes on hydroxylated surfaces, alkanethiols on gold, silver, copper and platinum, dialkyl disulphides on gold, alcohols and amines on platinum and carboxyl acids on aluminium oxide and silver. Some examples and references can be found in [123]. More recently also phosphonic and phosphoric esters on aluminium oxides have been reported [124, 125]. Only a small selection out of this number of SA systems can be presented here and properties such as kinetics, thermal, chemical and mechanical stability are briefly presented for alkanethiols on gold as an example.

The molecules for SA monolayers are chosen or synthesized according to the substrate that should be coated. Thiol-terminated entities have been mostly used in connection with metal surfaces, but also on GaAs [126]. Chloro- and acid-terminated molecules are most often employed on oxide surfaces of metals or semiconductors. However, they have also occasionally been used with metal surfaces [127].

(A) MONOLAYERS FORMED FROM ACIDS

The first work published in this area was that of Bigelow mentioned above [116]. In 1957, monolayers of long-chain fatty acids were formed on thin films of silver, copper, iron and cadmium deposited on glass microscope slides [43].

-18-

Finally, in 1985, the results of an extensive investigation in which adsorption took place onto an aluminium oxide layer formed on a film of aluminium deposited *in vacuo* onto a silicon wafer was published by Allara and Nuzzo [127, 128]. Various carboxylic acids were dissolved in high-purity hexadecane and allowed to adsorb from this solution onto the prepared aluminium oxide surface. It was found that for chains with more than 12 carbon atoms, chains are nearly in a vertical orientation and are tightly packed. For shorter chains, however, no stable monolayers were found. The kinetic processes involved in layer formation can take up to several days.

More recently, alternative chemistries have been employed to coat oxide surfaces with SAMs. These have included carboxylic [129, 130], hydroxamic [131], phosphonic [124, 132] and phosphoric acids [133]. Potential applications of SAMs on oxide surfaces range from protective coatings and adhesive layers to biosensors.

(B) SILANES

Organosilanes, such as trichlorosilanes or trimethylsilanes, can establish SA monolayers on hydroxylated surfaces. Apart from their (covalent) binding to the surface these molecules can also establish a covalent intermolecular network, resulting in an enhanced mechanical stability of the films (figure C2.4.11). In 1980, work was published on the formation of SAMs of octadecyltrichlorosilane (OTS) [117]. Subsequently, the use of this material was extended to the formation of multilayers [134].

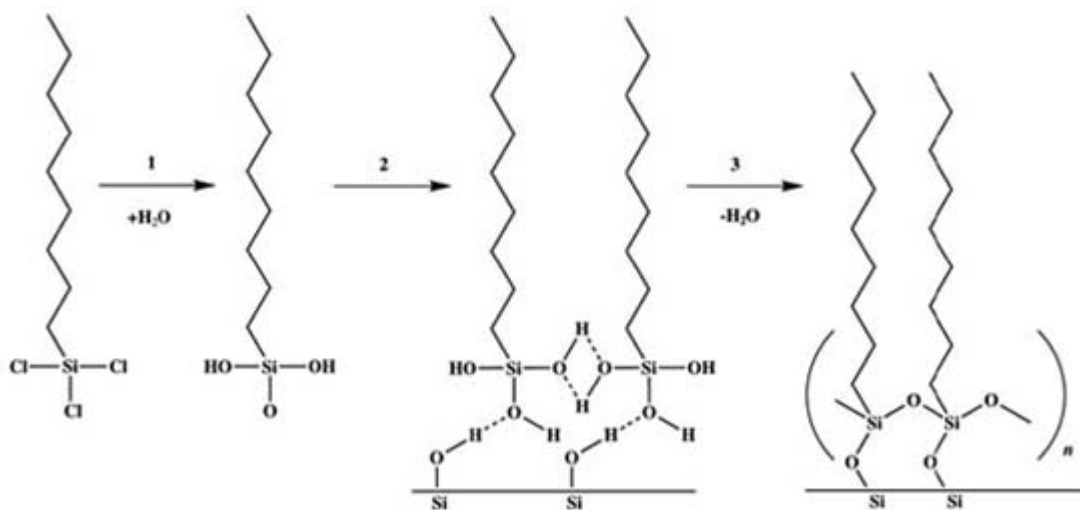


Figure C2.4.11. The formation of SAMs from OTS on a silicon oxide substrate.

Although it has been claimed in the literature that well defined and organized films have been achieved, there is still some debate about the quality of these systems. It has been suggested that the assembly mechanism of chlorosilanes on oxidized surfaces depends very crucially on the processing conditions involved. The assembling process shows an interesting dependence on pre-hydration and temperature [135], suggesting that water plays a central role. The stability in the films seems to be mostly due to the intermolecular network and not to the bonding to the substrate. This is supported by the observation that organized films are established on amorphous substrates.

-19-

Apart from these simple silanes, derivatives with aromatic groups at different places in the chain have also been investigated [136, 137]. It was found that the average tilt angle of these molecules depends on the specific functional entities contained in the chains. It is likely that apart from packing considerations—important for bulky groups, for example—other factors also influence the resulting tilt.

(C) THIOLS

In 1983, it was shown that certain sulphur compounds have a strong affinity to gold and bind strongly to a gold surface [118]. This led to considerable interest and activity in the study of SAMs of sulphur compounds on gold and today alkyl thiols on gold are the most extensively studied SA system. Although ethanol has been most frequently used as a solvent, however, other liquids have also been employed [122]. Polycrystalline gold substrates prepared by thermal evaporation predominantly display the (111) face [138, 139, 140, 141 and 142], which has the lowest surface energy [138]. Most of the work cited in the following has been performed on such surfaces.

Monolayers of alkanethiols adsorbed on gold, prepared by immersing the substrate into solution, have been characterized by a large number of different surface analytical techniques. The lateral order in such layers has been investigated using electron [143], helium [144, 145] and x-ray [146, 147] diffraction, as well as with scanning probe microscopies [122, 148]. Information about the orientation of the alkyl chains has been obtained by ellipsometry [149], infrared (IR) spectroscopy [150, 151] and NEXAFS [152].

The systematic study of alkanethiols ($\text{CH}_3(\text{CH}_2)_n\text{-SH}$) with different chain lengths revealed that for $n = 11$ or greater closely packed layers are obtained with a tilt angle of the alkyl chains between 30 and 35° from the surface normal and with an area per molecule of approximately 21.4 \AA^2 [122]. For $n < 11$, there is a gradual deterioration in order with decreasing length. It is generally accepted that adsorption takes place by elimination of the terminal hydrogen and that the thiol is bound to the gold by a true valence bond [122]. In contrast to chlorosilanes on oxidized substrates, alkanethiols establish a superlattice on gold and much of their stability is due to the anchoring of the molecules to the substrate.

There is a controversy about the bonding state of the sulphur. Most evidence suggests that it is bound in the form of a thiolate [122], while x-ray diffraction suggests that the sulphur atoms may dimerize [147]. However, not all of the observed overstructures can be explained with this latter assumption.

The growth and the structure of fully developed alkanethiol films on gold have been extensively investigated with STM [153, 154, 155, 156 and 157]. This work has confirmed that four phases exist in the final layer. There is a basic hexagonal lattice corresponding to the superlattice established by the sulphur atoms and three variants of a rectangular lattice, which was experimentally observed in the organization of the chains in long-chain ($>\text{C}_{12}$) thiolate monolayers on Au(111). The rectangular lattices are due to different twist angles of the chains around their axis [148].

The domain size in alkanethiol films on gold depends on the concentration used during preparation and is typically between 10 and 50 nm [158]. Thiol monolayers on gold have long-range angular order but quite short-range radial order. This can be explained in terms of tightly packed tilted molecules.

Order and dense packing are relative in the context of these systems and depend on the point of view. Usually the term order is used in connection with translational symmetry in molecular structures, i.e. in a two-dimensional monolayer with a crystal structure. Dense packing in organic layers is connected with the density of crystalline polyethylene.

Self-organizing monolayers are highly disordered compared to inorganic crystals, due to the defects that are always present. On the other hand, when compared to polymeric glasses or liquid paraffin, they are highly ordered systems. Hence, the terms order and densely packed in this context do not imply the absence of defects. The degree of order is comparable to that of Langmuir–Blodgett layers.

Apart from domain boundaries, some of the defects in alkanethiol monolayers (pitholes) are created by the thiol itself [159] by 'etching' processes. It was found that the solvent used for preparation also has some effect on the resulting defect density.

In contrast to the gold surface, on silver the chains adopt a lower tilt angle of 12° from the surface normal [160, 161]. This is attributed to the different nature of the bonding of sulphur to silver as compared to gold and the slightly different packing density. The coherence length determined with He atom diffraction was found to be 12 nm [162].

GaAs has been coated with thiols with a view to modifying devices [123]. For these films, S–As bonds are presumed to be present. An ordering of the chains for $n = 18$ has been reported. However, this system has generally been much less investigated than those involving metal substrates.

The lubricant properties of alkanethiols and fluorinated alkanes have been studied extensively by scanning probe techniques [163]. In agreement with experiments on LB monolayers it was found that the fluorocarbon monolayers show considerably higher friction than the corresponding hydrocarbon monolayers [164, 165 and 166] even though the fluorocarbons are known to have the lowest surface free energy of all organic materials.

Kinetics of film formation. The kinetics of film formation of SAMs is important, in order to establish a recipe that allows films of reproducible quality to be prepared, i.e. densely packed, well ordered monolayers. Ellipsometry and contact-angle measurements can give information about coverage and orientation present in the films. In general it has been reported that film formation is a two-step process, at least for dilute solutions (~ 1 mM). A fast first step where the contact angles and film thickness are already close to their limiting values (minutes) is followed by a second slower step, after which the final values are reached [120]. It was also shown that the order is established during this second step, where the last 5–10% of molecules are incorporated into the film and force the molecules on the surface to stretch [167]. For different chain lengths and solvents it was found that only Langmuir kinetics can explain the experimental data of thiolate films on polycrystalline gold, irrespective of the experimental conditions [168].

Lateral structuring of SAMs—microcontact printing. Of great interest is the application-specific chemical structuring of ultrathin organic films, for example for use in biomedical devices. Such structuring can be accomplished by lithographic means, including the so-called microcontact printing technique (μ CP) [169, 170 and 171]. This is a relatively simple technique and allows lateral patterning down to submicrometre scale. It is also known as soft-lithographic patterning. In short, a structured stamp made of poly(dimethylsiloxane) (PDMS) is inked with the diluted solution and brought into contact with the substrate (figure C2.4.12). Patterns with dimensions in the (sub) micrometre range can routinely be produced by this technique [172]. Although the contact time between stamp and surface is of the order of only 10–20 s, the resulting films of alkanethiols on gold are chemically not distinguishable from those prepared by immersion. Remarkably, even the order in areas that are prepared by μ CP is comparable to that in films prepared by immersion [173]. However, with lateral force microscopy, regions prepared by immersion can be distinguished from those prepared by μ CP [173, 174]. This is probably due to slightly different domain size distributions and thus different mechanical stabilities of the films on a nanometre scale [173].

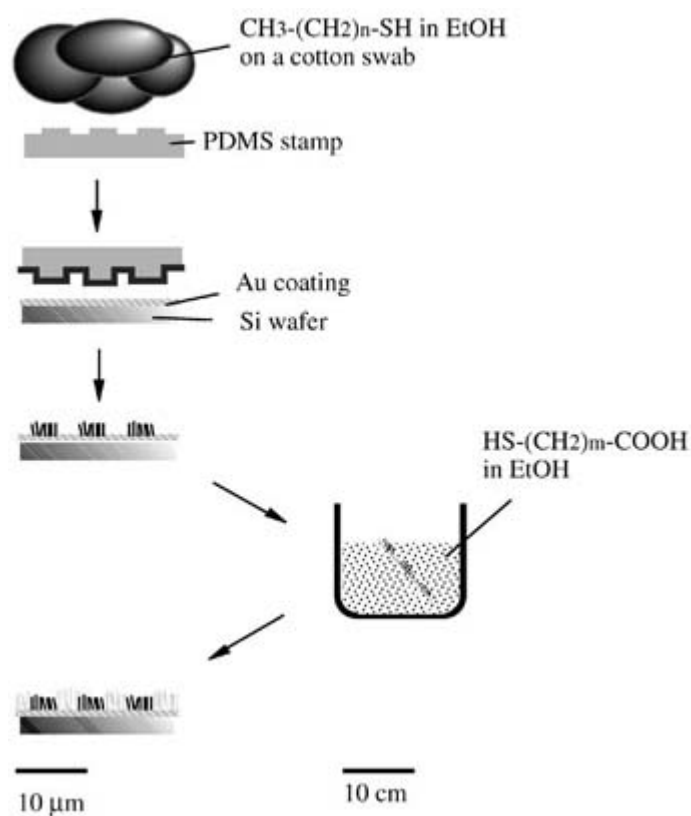


Figure C2.4.12. Principle of the microcontact printing process. Chemically patterned organic films with differently functionalized regions can be prepared by a combination of μCP and subsequent immersion.

Other lithographical means include micromachining [175], photopatterning [176] or electron beam patterning [177], which have been demonstrated on alkanethiolate/Au SAMs, alkanethiolate and organo-siloxane on Si and Ti and alkanethiolates on GaAs.

Chemical stability. The chemical stability of SA films is of interest in many areas. However, there is no general rule for it. The chemical stability of silane films is remarkable, due to their intermolecular crosslinking. Therefore, they are found to be more stable than LB films. Alkyltrichlorosilane monolayers provide structures that are stable to chemical conditions that most LB films could not stand. However, photopolymerized LB films also show considerable stability in organic solvents.

SAMs of thiolates on gold are generally resistant to strong acids or bases [175, 178 and 179], are not destroyed by solvents [180] and can withstand physiological environments [181, 182 and 183]. However, they also show some degradation if exposed to the ambient atmosphere for sufficiently extended periods [184].

Thermal stability. The thermal stability of SAMs is, similarly to LB films, an important parameter for potential applications. It was found that SA films containing alkyl chains show some stability before an increase in the number of *gauche* conformations occurs, resulting in melting and irreversible changes in the film. The disordering of the

chain structure as temperature is increased has been studied for both LB films and SAMs. In general, an all-*trans* structure dominates at room temperature for most of the systems. When the melting point is approached, however, the number of *gauche* defects increases. For thiols with more than 16 carbon atoms it was shown experimentally that *gauche* defects are mainly concentrated near the free ends of the chains [151, 185]. Unlike bulk hydrocarbons, these films do not show a sharp phase transition in the temperature range between 80 and 420 K but rather a gradual change to a progressively more ordered state as the temperature is lowered.

Similarly to LB films, the order of alkanethiols on gold depending on temperature has been studied with NEXAFS. It was observed that the barrier for a *gauche* conformation in a densely packed film is an order of magnitude higher than that of a free chain [48].

SAMs that are made out of structures capable of forming strong intermolecular hydrogen bonds have been studied especially in view of their expected high thermal and chemical stability [186, 187].

A good survey of the chemical and physical film characteristics of highly organized SAMs is given in [123].

Mechanical stability. Chemisorption to the surface, intermolecular interactions and crosslinking between adjacent compounds—if possible—all contribute to the resulting stability of the monolayer film. Lateral force microscopy investigations revealed that the mechanical stability towards lateral forces on the nanometre scale is likely to be determined by the defect density and the domain size on a nano- to micrometre scale [163, 173].

Experiments with chemically grafted SAMs displayed much larger wear resistance than films produced by the LB technique [188]. Also it was found that wear properties of SAMs can be further improved by chemically grafting C₆₀ molecules onto SAM surfaces [189].

(D) ALKANETHIOLS WITH FUNCTIONAL ENTITIES ATTACHED

The strong bond formed between the thiol endgroups and gold and silver surfaces allows the possibility of forming molecules that have a wide variety of different functional groups at the opposite end and thus of coating a noble metal surface with a variety of differently functionalized molecules and mixtures.

A large number of studies concerned with thiol-terminated molecules has been directed at the preparation of tailored organic surfaces, since their importance has been steadily increasing in various applications. Films of ω -functionalized alkanethiols have facilitated fundamental studies of interfacial phenomena, such as adhesion [190, 191], corrosion protection [192], electrochemistry [193], wetting [194], protein adsorption [195, 196] or molecular recognition [197, 198, 199, 200 and 201] to mention only a few.

Biological applications are attracting increasing attention and examples from this area are given below. An understanding of the mechanism of protein adsorption, the interaction of proteins with ‘artificial’ substrates and the way in which these interactions determine the biological activity of these substrates are of immense biomedical significance [202, 203, 204, 205 and 206]. SAMs play a particularly important role here, since they can serve as models of polymer surfaces, allowing surface chemical properties to be investigated independent from the effects of surface morphology. In this way, macroscopic concepts such as hydrophobicity, hydrophilicity, wettability and water

content, which are crucial to understanding cell adhesion and anchorage-dependent cell behaviour [207, 208, 209, 210 and 211], can be substituted by more fundamental, molecular-level concepts of surface organization, reactivity and structure. Efforts have been undertaken to engineer gradients of surface hydrophobicity/hydrophilicity on polymeric surfaces [212, 213, 214 and 215] and, more recently, on SAMs prepared from thiols [216].

Hydroxylated surfaces are of particular interest, due to the possibility of derivatizing the –OH groups with biologically active moieties. The spatial arrangement and density of –OH groups within a monolayer matrix are relevant, since they may regulate the accessibility of a specific functional group to biomolecules. One approach to controlling these properties uses mixed chain length CH₃- and OH-terminated alkanethiolate SAMs [216, 217, 218, 219 and 220]. Another approach is a thiol-terminated hexasaccharide with acetyloxy groups that can be replaced by hydroxyl groups, and which was also adsorbed on gold [221] (figure C2.4.13). It was found that deprotection and thus the degree of hydrophilicity can be tailored before as well as following adsorption on the surface. This allows the OH-group density to be adjusted to a high degree and might have advantages over the mixed monolayers, since it can be better controlled.

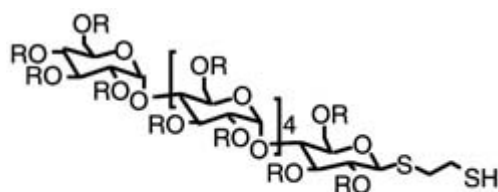


Figure C2.4.13. Thiol-terminated hexasaccharide, where acetoxy groups can be replaced by hydroxyl groups, both before and following adsorption [221].

In another study, coadsorption of simple *n*-alkanethiols, which acted as a scaffolding, and a synthetic receptor was studied on gold [222]. The design of the system mimics those of receptors bound to lipid membranes.

Poly(ethylene glycol) is often employed to render surfaces protein resistant. Oligo(ethylene glycol)- (OEG-) terminated alkanethiols were adsorbed on Ag and Au and were studied concerning their properties towards protein adsorption [223, 224]. Interestingly, the slight difference in the packing density and chain tilt of the alkane chains on gold and silver leads to a completely different behaviour of the OEG-terminated thiols concerning protein adsorption [223]. While the SAMs on gold are protein resistant, those on silver adsorb a certain amount of fibrinogen.

Regarding protein adsorption properties, differently terminated SAMs on gold have also been investigated [225]. It was found that the nature of the adsorbate chain structure was the most important parameter for the observed behaviour towards protein and cell adsorption.

Covalent immobilization of proteins on microstructured gold surfaces was studied in [226]. On these substrates, which were prepared by μ CP and etching, the immobilization sites of proteins could be spatially controlled using an amino-reactive SAM. The whole process, i.e. production of the micropatterned substrate including SAM exchange and protein immobilization, took a reasonably small amount of time (~24 h), providing some flexibility in the experimental work.

(E) THIOL-MODIFIED POLYMERS

In addition to simple alkanethiols or those functionalized with small groups or compounds, thiols can also be used to attach polymers to metal surfaces. However, there are specific problems, since the accessibility of the surface for the thiol groups of the modified polymers, for example, is often strongly restricted and a significant entropy change may be connected with the adsorption, depending on the polymer, the surface and the solvent.

The study of ultrathin polymer layers on metals is relevant in understanding the behaviour of polymers on surfaces, as well as in the areas of adhesion and corrosion. Gold and copper surfaces can be covered with monolayers of polymers by adsorption from solution [227, 228, 229, 230, 231, 232, 233, 234 and 235].

One example of a polymer layer on gold consists of adsorbed thiol-terminated poly(styrene)s of different molecular weights [234]. Poly(styrene) itself does not adsorb significantly on gold from tetrahydrofuran (THF) [232] or toluene [236]. As the average molecular weight (M_n) increases up to ~100 000, an increasing amount of the thiol-terminated polymer adsorbs on the surface [234]. The adsorbed mass remains constant up to M_n of ~200 000 or more, corresponding to a layer thickness of ~3 nm [234]. However, thiol-terminated poly(styrene) with $M_n = 500$ 000 does not adsorb at all [234]. It seems that the sulphur–gold interaction is no longer sufficient to overcome the entropy loss and loss of polymer–solvent interactions, which would accompany adsorption of very high-molecular-weight chains [234].

Structure, morphology and friction of thiol-terminated poly(styrene) have also been studied with atomic force microscopy [237, 238 and 239].

No adsorption of a block copolymer with a styrene/propylene sulphide molar ratio of 3:1 from THF was found when the propylene sulphide blocks were capped with ethyl groups [227]. Thiol-capped styrene-propylene sulphide block copolymers ($M_n = 60\ 000$), in contrast, adsorb on gold from THF [234]. The layer thickness of styrene-propylene sulphide block copolymers decreases with increasing propylene sulphide block size, since the propylene sulphide block interacts with and, as a result, adheres to the surface [234]. A strong interaction of the thiol endgroup seems, however, to be necessary for any further segmental adsorption of the polymer chain [227]. The styrene block most likely escapes from the gold surface, at least partially, and this could explain the observed changes in layer thickness. Analogously, poly(styrene) that is terminated with six or seven ethylene sulphide units with a thiol end group also adsorbs from toluene onto gold [235].

Poly(methyl methacrylate) (PMMA) with sulphide-modified side chains adsorbs onto gold from methyl ethyl ketone and dichloromethane (there is little influence of the solvents on the layer formation) [229]. Two polymers with one sulphide group per 10 and 100 monomer units were examined [229]. The thickness of the adsorbed layers was ~1.5–3 nm, depending on the concentration of the polymer in the solution and, to a degree, on the sulphur content [229]. Angle-dependent XPS measurements indicate that the sulphide groups are not concentrated at the interface but dispersed relatively uniformly throughout the film [229]. Unmodified poly(methyl methacrylate) also adsorbs under the same conditions, but the thickness of the corresponding layer is below 1 nm [229]. Also, a poly(benzylglutamate) with a disulphide endgroup yields thicker layers on gold than the corresponding non-modified poly(benzylglutamate) [228]. Moreover, poly(acrylate)s with disulphide moieties in the side chains have been studied. In this case it was found that the disulphide groups promote the adsorption [230, 232].

-25-

The examples described above are only a small selection out of a tremendous number of investigations of LB films and SAMs. This number is still increasing and it is expected that ultrathin organic films will play a central role in both fundamental and applied sciences in the future.

REFERENCES

- [1] Pockels A 1891 Surface tension *Nature* **43** 437–39
- [2] Pockels A 1892 On the relative contaminations of the water-surface by equal quantities of different substances *Nature* **46** 418–9
- [3] Pockels A 1893 Relationship between the surface-tension and relative contamination of water surfaces *Nature* **48** 152–4
- [4] Pockels A 1894 On the spreading of oil upon water *Nature* **50** 223–4
- [5] Lord Rayleigh 1899 Investigations in capillarity *Phil. Mag.* **48** 321–37
- [6] Hardy W B 1912 The tension of composite fluid surfaces and the mechanical stability of films of fluids *Proc. R. Soc. A* **86** 610–35
- [7] Langmuir I 1917 The constitutions and fundamental properties of solids and liquids. 2. Liquids *J. Am. Chem. Soc.* **39** 1848–906
- [8] Langmuir I 1920 The mechanism of the surface phenomena of flotation *Trans. Faraday Soc.* **15** 62–74
- [9] Blodgett K B 1935 Films built by depositing successive monomolecular layers on a solid surface *J. Am. Chem. Soc.* **57** 1007–22
- [10] Adamson A W 1990 *Physical Chemistry of Surfaces* (New York: Wiley) p 119

- [11] den Engelsen D, Hengst J H T and Honig E P 1976 An automated Langmuir trough for building monomolecular layers *Philips Tech. Rev.* **36** 44–6
- [12] Gaines G L 1966 *Insoluble Monolayers at Liquid–Gas Interfaces* (New York: Interscience)
- [13] Barraud A, Leloup J, Gouzerh A and Palacin S 1935 An automated trough to make alternate layers *Thin Solid Films* **133** 117–23
- [14] Richard J, Barraud A, Vandevyver M and Ruaudel-Teixier A 1988 A 2-step transfer of conducting Langmuir films from a glycerol subphase *Thin Solid Films* **159** 207–14
- [15] Buhaenko M R, Goodwin J W, Richardson R M and Daniel M F 1985 The influence of shear viscosity of spread monolayers on the Langmuir–Blodgett process *Thin Solid Films* **134** 217–26
- [16] Ulman A 1992 *Ultrathin Organic Films* (Boston: Academic) p 123
- [17] Langmuir I and Schaefer V I 1938 Activities of urease and pepsin monolayers *J. Am. Chem. Soc.* **60** 1351–60
- [18] Saint Pierre M and Dupeyrat M 1983 Measurement and meaning of the transfer process energy in the building up of Langmuir–Blodgett multilayers *Thin Solid Films* **99** 205–13
- [19] Roberts G G, McGinnity M, Barlow W A and Vincett P S 1979 Electroluminescence, photoluminescence and electroabsorption of a highly substituted anthracene Langmuir film *Solid State Commun.* **32** 683–6

-26-

- [20] Tieke B, Lieser G and Weiss K 1983 Parameters influencing the polymerization and structure of long-chain diynoic acids in multilayers *Thin Solid Films* **99** 95–102
- [21] Sugi M, Saito M, Fukui T and Iizima S 1983 Effect of dye concentration in Langmuir multilayer photoconductors *Thin Solid Films* **99** 17–20
- [22] Daniel M F, Lettington O C and Small M 1983 Investigation into the Langmuir–Blodgett film formation ability of amphiphiles with cyano head groups *Thin Solid Films* **99** 61–9
- [23] Schoeler U, Tews K H and Kuhn H 1974 Potential model of dye molecule from measurements of the photocurrent in monolayer assemblies *J. Chem. Phys.* **61** 5009–16
- [24] Tredgold R H, Jones S D, Evans S D and Williams P I 1986 Aluminium oxide as a substrate for the deposition of Langmuir–Blodgett films *J. Mol. Electron.* **2** 147–9
- [25] Girling I R and Milverton D R J 1984 A method for the preparation of an alternating multilayers film *Thin Solid Films* **115** 85–8
- [26] Hardy R M and Scala L C 1966 Electrical and structural properties of Langmuir films *J. Electrochem. Soc.* **113** 109–16
- [27] Peterson I R, Veale G and Montgomery C M 1986 The preparation of oleophilic surfaces for Langmuir–Blodgett deposition *J. Colloid Interface Sci.* **109** 527–30
- [28] Tredgold R H and El-Badawy Z I 1985 Increase of Schottky-barrier height at GaAs-surfaces by carboxylic-acid monolayers and multilayers *J. Phys. D: Appl. Phys.* **18** 103–9
- [29] Schwartz D K 1997 Langmuir–Blodgett film structure *Surf. Sci. Rep.* **27** 245–334
- [30] Outka D A, Stöhr J, Rabe J P, Swalen J D and Rothermund H H 1987 Orientation of arachidate chains in Langmuir–Blodgett monolayers on Si(111) *Phys. Rev. Lett.* **59** 1321–4
- [31] Kinzler M, Schertel A, Hähner G, Wöll C, Grunze M, Albrecht H, Holzhüter G and Gerber T 1994 Structure of mono- and multilayer Langmuir–Blodgett films from Cd arachidate and Ca arachidate *J. Chem. Phys.* **100** 7722–35
- [32] Peterson I R 1984 Optical observation of monomer Langmuir–Blodgett film structure *Thin Solid Films* **116** 357–66

- [33] Peterson I R 1987 Langmuir–Blodgett films: structure and application *J. Mol. Electron.* **3** 103–11
- [34] Bibo A M and Peterson I R 1989 Disclination recombination kinetics in water-surface monolayers of 22-tricosenoic acid *Thin Solid Films* **178** 81–92
- [35] Mann B and Kuhn H 1971 Tunneling through fatty acid salt monolayers *J. Appl. Phys.* **42** 4398–405
- [36] Gundlach K H and Kadlech J 1974 The influence of the oxide film on the current in Al–Al oxide–fatty acid monolayer–metal junctions *Chem. Phys. Lett.* **25** 293–5
- [37] Polymeropoulos E E 1977 Electron tunneling through fatty-acid monolayers *J. Appl. Phys.* **48** 2404–7
- [38] Polymeropoulos E E and Sagiv J 1978 Electrical conduction through adsorbed monolayers *J. Chem. Phys.* **69** 1836–47
- [39] Peterson I R 1980 Defect density in a metal–monolayer–metal cell *Aust. J. Chem.* **33** 1713–6
- [40] Procarione W L and Kauffman J W 1974 The electrical properties of phospholipid bilayer Langmuir films *Chem. Phys. Lipids* **12** 251–60

-27-

- [41] Tredgold R H and Winter C S 1981 Tunneling currents in Langmuir–Blodgett monolayers of stearic acid *J. Phys. D: Appl Phys.* **14** L185–8
- [42] Peterson I R 1986 A structural study of the conducting defects in fatty acid Langmuir–Blodgett monolayers *J. Mol. Electron.* **2** 95–9
- [43] Chapman J A and Tabor D 1957 An electron diffraction study of retracted monolayers *Proc. R. Soc. A* **242** 96–107
- [44] Briscoe B J and Evans D C B 1981 The shear properties of Langmuir–Blodgett layers *Proc. R. Soc. A* **380** 389–407
- [45] Cohen S R, Naaman R and Sagiv J 1986 Thermally induced disorder in organized organic monolayers on solid substrates *J. Phys. Chem.* **90** 3054–6
- [46] Rothberg L, Higashi G S, Allara D L and Garoff S 1987 Thermal disordering of Langmuir–Blodgett-films of cadmium stearate on sapphire *Chem. Phys. Lett.* **133** 67–72
- [47] Ulman A 1992 *Ultrathin Organic Films* (Boston: Academic) p 138
- [48] Schertel A, Hähner G, Grunze M and Wöll C 1996 Near edge x-ray adsorption fine structure investigation of the orientation and thermally induced order–disorder transition in thin organic films containing long chain hydrocarbons *J. Vac. Sci. Technol. A* **14** 1801–6
- [49] Tredgold R H 1994 *Order in Thin Organic Films* (Cambridge: Cambridge University Press) p 70
- [50] Heesemann J 1980 Studies on monolayers: 1. Surface tension and absorption spectroscopic measurements of monolayers of surface-active azo and stilbene dyes *J. Am. Chem. Soc.* **102** 2167–76
- [51] Nakahara H and Fukuda K 1983 Orientation of chromophores in monolayers and multilayers of azobenzene derivatives with long alkyl chains *J. Colloid Interface Sci.* **93** 530–9
- [52] Blinov L M, Dubinin N V, Mikhnev L V and Yudin S G Polar 1984 Langmuir–Blodgett films *Thin Solid Films* **120** 161–70
- [53] Jones R, Tredgold R H, Hoorfar A, Allen R A and Hodge P 1985 Crystal-formation and growth in Langmuir–Blodgett multilayers of azobenzene derivatives—optical and structural studies *Thin Solid Films* **134** 57–66
- [54] Tredgold R H, Allen R A and Hodge P 1987 X-ray-diffraction and optical studies of Langmuir–Blodgett films formed from azobenzene derivatives *Thin Solid Films* **155** 343–52

- [55] Nishiyama K and Fujihira M 1988 *Cis-trans* reversible photoisomerization of an amphiphilic azobenzene derivative in its pure LB film prepared as polyion complexes with polyallylamine *Chem. Lett.* 1257–60
- [56] Nakahara H, Fukuda K, Shimomura M and Kunitake T 1988 Molecular arrangements and photoisomerization of amphiphilic azobenzene derivatives in monolayers and multilayers *Nippon Kagaku Kaishi* 1001–10
- [57] Tachibana H, Nakamura T, Matsumoto M, Komizu H, Manda E, Niino H, Yabe A and Kawabata Y 1989 Photochemical switching in conductive Langmuir–Blodgett films *J. Am. Chem. Soc.* **111** 3080–1
- [58] Liu Z, Loo B H, Baba R and Fujishima A 1990 Excellent reversible photochromic behaviour of 4-octyl-4'-(5-carboxyl-pentamethyleneoxy)-azobenzene in organized monolayer assemblies *Chem. Lett.* 1023–6
- [59] Barnik M I, Kozenkov V M, Shtykov N N, Palto S P and Yudin S G 1989 Photoinduced optical anisotropy in Langmuir–Blodgett films *J. Mol. Electron.* **5** 53–6
- [60] Liu Z F, Loo B H, Hashimoto K and Fujishima A A 1991 Novel photoelectrochemical hybrid one-way process observed in the azobenzene system *J. Electroanal. Chem. Interfacial Electrochem.* **297** 133–44

-28-

- [61] Ulman A 1992 *Ultrathin Organic Films* (Boston: Academic) p 159
- [62] Tredgold R H 1994 *Order in Thin Organic Films* (Cambridge: Cambridge University Press) p 74
- [63] Jones R, Tredgold R H and Hodge P 1983 Langmuir–Blodgett films of simple esterified porphyrins *Thin Solid Films* **99** 25–32
- [64] Jones R, Tredgold R H, Hoorfar A and Hodge P 1984 Electrical-conductivity in Langmuir–Blodgett films of porphyrins—inplane and through-the-film studies *Thin Solid Films* **113** 115–28
- [65] Jones R, Tredgold R H and Hoorfar A 1985 Effects of thickness on surface-potential and surface conductivity in non-insulating Langmuir–Blodgett multilayers of porphyrins *Thin Solid Films* **123** 307–14
- [66] Luk S Y, Mayers F R and Williams J O 1988 Preparation and characterization of Langmuir–Blodgett films of mesoporphyrin-IX dimethylester indium chloride *Thin Solid Films* **157** 69–79
- [67] Bull R A and Bulkowski J E 1983 Tetraphenylporphyrin monolayers—formation at the air water interface and characterization on glass supports by absorption and fluorescence spectroscopy *J. Colloid Interface Sci.* **92** 1–12
- [68] Fujiki M, Tabei H and Kurihara T 1988 Self-assembling features of soluble nickel phthalocyanines *J. Phys. Chem.* **92** 1281–5
- [69] Nakahara H, Fukuda K, Katahara K and Nishi H 1989 Langmuir–Blodgett films of octa-alkyl phthalocyanines *Thin Solid Films* **178** 361–6
- [70] Ogawa K, Yonehara H, Shoji T, Kinoshita S, Maekawa E, Nakahara H and Fukuda K 1989 Inplane anisotropy in Langmuir–Blodgett films of a copper phthalocyanine derivative *Thin Solid Films* **178** 439–43
- [71] Nichogi K, Waragai K, Taomoto A, Saito Y and Asakawa S 1989 Lead phthalocyanine Langmuir–Blodgett films *Thin Solid Films* **178** 297–301
- [72] Fryer J R, McConnell C M, Hann R A, Eyres B L and Gupta S K 1990 The structure of some Langmuir–Blodgett films. 1. Substituted phthalocyanines *Phil. Mag. B* **61** 843–52
- [73] Brynda E, Kalvoda L, Koropecy I, Nespurek S and Rakusan J 1990 Copper tetra-4-tert-butylphthalocyanine Langmuir–Blodgett-films—photoelectrical and structural studies *Synth. Met.* **37** 327–33
- [74] Brynda E, Koropecy I, Kalvoda L and Nespurek S 1991 Electrical and photoelectrical properties of copper tetra[4-tert-butylphthalocyanine] Langmuir–Blodgett-films *Thin Solid Films* **199** 375–84
- [75] Fukui M, Katayama N, Ozaki Y, Araki T and Iriyama K 1991 Structural characterization of phthalocyanine Langmuir–Blodgett multilayer assemblies by FTIR spectroscopy *Chem. Phys. Lett.* **177** 247–51

- [76] Pace M D, Barger W R and Snow A W 1989 Molecular packing and iodine doping of oxovanadium-substituted and copper-substituted tetrakis(cumylphenoxy)phthalocyanine Langmuir–Blodgett films studied by ESR *Langmuir* **5** 973–8
- [77] Shutt J D, Batzel D A, Sudiwala R V, Rickert S E and Kenney M E 1988 Fabrication of thin-film dielectrics from an amphiphilic 2-ring phthalocyanine *Langmuir* **4** 1240–7
- [78] Shutt J D and Rickert S E 1989 Thermally stable high-field molecular multilayer dielectrics formed from an amphiphilic 2-ring phthalocyanine *J. Mol. Electron.* **5** 129–34
- [79] Liu Y, Shigehara K and Yamada A 1989 Purification of lutetium dipthalocyanine and electrochromism of its Langmuir–Blodgett films *Thin Solid Films* **179** 303–8
- [80] Ruaudel-Teixier A, Barraud A, Belbeoch B and Roulliay M 1983 Langmuir–Blodgett films of pure porphyrins *Thin Solid Films* **99** 33–40

-29-

- [81] Lesieur P, Vandevyver M, Ruaudel-Teixier A and Barraud A Orientational studies of Langmuir–Blodgett films of porphyrins with polarized resonant Raman spectroscopy *Thin Solid Films* **159** 315–22
- [82] Palacin S, Lesieur P, Stefanelli I and Barraud A 1988 Structural studies of intermolecular interactions in pure and diluted films of a redox-active phthalocyanine *Thin Solid Films* **159** 83–90
- [83] Palacin S and Barraud A 1989 Highly ordered Langmuir–Blodgett films based on semi-amphiphilic phthalocyanines *J. Chem. Soc. Chem. Commun.* 45–7
- [84] Tredgold R H 1994 *Order in Thin Organic Films* (Cambridge: Cambridge University Press) p 82
- [85] Cemel A, Fort T and Lando J B 1972 Polymerization of vinyl stearate multilayers *J. Polymer Sci. A-1* **10** 2061–83
- [86] Naegele D, Lando J B and Ringsdorf H 1977 Polymerization of cadmium octadecylfumarate in multilayers *Macromolecules* **10** 1339–44
- [87] Rabe J P, Rabolt J F, Brown C A and Swalen J D 1982 Order-disorder transitions in Langmuir–Blodgett films. 2. IR studies of the polymerization of Cd-octadecylfumarate and Cd-octadecylmaleate *J. Chem. Phys.* **84** 4096–102
- [88] Laschewsky A, Ringsdorf H and Schmidt G 1985 Polymerization of hydroxocarbon and fluorocarbon amphiphiles in Langmuir–Blodgett multilayers *Thin Solid Films* **134** 153–72
- [89] Laschewsky A, Ringsdorf H and Schmidt G 1988 Polymerization of eicosenoic acid and octadecyl fumarate in Langmuir–Blodgett multilayers *Polymer* **29** 448–56
- [90] Laschewsky A and Ringsdorf H 1988 Polymerization of amphiphilic dienes in Langmuir–Blodgett multilayers *Macromolecules* **21** 1936–41
- [91] Barraud A, Rosilio C and Ruaudel-Teixier A 1977 Solid-state electron-induced polymerization of ω -tricosenoic acid multilayers *J. Colloid Interface Sci.* **62** 509–23
- [92] Barraud A, Rosilio C and Ruaudel-Teixier A 1980 Polymerized monomolecular layers: a new class of ultrathin resins for microlithography *Thin Solid Films* **68** 91–8
- [93] Uchida M, Tanizaki T, Kunitake T and Kajiyama T 1989 Surface stability and functional property of polymerized Langmuir–Blodgett type films *Macromolecules* **22** 2381–7
- [94] Uchida M, Tanizaki T, Oda T and Kajiyama T 1991 Control of surface chemical-structure and functional property of Langmuir–Blodgett-film composed of new polymerizable amphiphile with a sodium-sulfonate *Macromolecules* **24** 3238–43
- [95] Tieke B, Graf H J, Wegner G, Naegele D, Ringsdorf H, Banerjee A, Day D and Lando J B 1977 Polymerization of mono- and multilayers forming diacetylenes *Colloid Polmer Sci.* **255** 512–31

- [96] Tieke B, Wegner G, Naegele D and Ringsdorf H 1976 Polymerization of tricoso-10,12-dienoic acid in multilayers *Angew. Chem. Int. Edn. Engl.* **15** 764–5
- [97] Tieke H, Lieser G and Wegner G 1979 Polymerization of diacetylenes in multilayers *J. Polymer Sci.: Polymer Chem. Edn.* **17** 1631–44
- [98] Tieke B, Enkelmann V, Kapp H, Lieser G and Wegner G 1981 Topochemical reactions in Langmuir–Blodgett multilayers *J. Macromol. Sci.—Chem. A* **15** 1045–58
- [99] Tieke B and Weiss K 1984 The morphology of Langmuir–Blodgett multilayers of amphiphilic diacetylenes—effects of the preparation conditions and the role of additives *J. Colloid Interface Sci.* **101** 129–48
-

-30-

- [100] Tredgold R H 1994 *Order in Thin Organic Films* (Cambridge: Cambridge University Press) p 86
- [101] Tredgold R H, Young M C J, Hodge P and Khoshdel E 1987 Lightguiding in Langmuir–Blodgett-films of preformed polymers *Thin Solid Films* **151** 441–9
- [102] Elbert R, Laschewsky A and Ringsdorf H 1985 Hydrophilic spacer groups in polymerizable lipids—formation of biomembrane models from bulk polymerized lipids *J. Am. Chem. Soc.* **107** 4134–41
- [103] Biddle M B, Lando J B, Ringsdorf H, Schmidt G and Schneider J 1988 Polymeric amphiphiles with hydrophilic main chain spacers—studies in monolayers and Langmuir–Blodgett multilayers *Colloid Polymer Sci.* **266** 806–13
- [104] Erdelen C, Laschewsky A, Ringsdorf H, Schneider J and Schuster A 1989 Thermal-behaviour of polymeric Langmuir–Blodgett multilayers *Thin Solid Films* **180** 153–66
- [105] Schneider J, Erdelen C, Ringsdorf H and Rabolt J F 1989 Structural studies of polymers with hydrophilic spacer groups. 2. Infrared-spectroscopy of Langmuir–Blodgett multilayers of polymers with fluorocarbon side-chains at ambient and elevated temperatures *Macromolecules* **22** 3475–80
- [106] Penner T L, Schildkraut J S, Ringsdorf H and Schuster A 1991 Oriented films from polymeric amphiphiles with mesogenic groups—Langmuir–Blodgett liquid-crystals *Macromolecules* **24** 1041–9
- [107] Tredgold R H 1994 *Order in Thin Organic Films* (Cambridge: Cambridge University Press) p 101
- [108] Orthmann E and Wegner G 1986 Catalysis of the polycondensation of dihydroxysiliconphthalocyanine *Macromol. Chem. Rapid Commun.* **7** 243–7
- [109] Orthmann E and Wegner G 1986 Preparation of ultrathin layers of molecularly controlled architecture from polymeric phthalocyanines by the Langmuir–Blodgett-technique 1986 *Angew. Chem. Int. Edn. Engl.* **25** 1105–7
- [110] Caseri W, Sauer T and Wegner G 1988 Soluble phthalocyaninato-polysiloxanes—rigid rod polymers of high molecular-weight *Macromol. Chem. Rapid Commun.* **9** 651–7
- [111] Rabe J P, Sano M, Batchelder D and Kalatchev A A 1988 Polymers on graphite and gold—molecular images and substrate defects *J. Microsc.* **152** 573–83
- [112] Sauer T, Arndt T, Batchelder D, Kalachev A A and Wegner G 1990 The structure of Langmuir–Blodgett-films from substituted phthalocyaninato-polysiloxanes *Thin Solid Films* **187** 357–74
- [113] Kalachev A A, Sauer T, Vogel V, Plate N A and Wegner G 1990 Influence of subphase conditions on the properties of Langmuir–Blodgett-films from substituted phthalocyaninato-polysiloxanes *Thin Solid Films* **188** 341–53
- [114] Crockett R G M, Campbell A J and Ahmed F R 1990 Structure and molecular-orientation of tetramethoxy-tetraoctoxy phthalocyaninato-polysiloxane Langmuir–Blodgett-films *Polymer* **31** 602–8

- [115] Ulman A 1992 *Ultrathin Organic Films* (Boston: Academic Press) p 237
- [116] Bigelow W C, Pickett D L and Zisman W A 1946 Oleophobic monolayers. 1. Films adsorbed from solution in non-polar liquids *J. Colloid Interface Sci.* **1** 513–38
- [117] Sagiv J 1980 Organized monolayers by adsorption. 1. Formation and structure of oleophobic mixed monolayers on solid surfaces *J. Am. Chem. Soc.* **102** 92–8
- [118] Nuzzo R G and Allara D L 1983 Adsorption of bifunctional organic disulphides on gold surfaces *J. Am. Chem. Soc.* **105** 4481–3
- [119] Porter M D, Bright T B, Allara D L and Chidsey C E D 1987 Spontaneously organized molecular assemblies. 4. Structural characterization of normal-alkyl thiol monolayers on gold by optical ellipsometry, infrared-spectroscopy, and electrochemistry *J. Am. Chem. Soc.* **109** 3559–68

-31-

- [120] Bain C D, Troughton, E B Tao Y T, Evall J and Whitesides G M 1989 Formation of monolayer films by the spontaneous assembly of organic thiols from solution onto gold *J. Am. Chem. Soc.* **111** 321–35
- [121] Ulman A 1992 *Ultrathin Organic Films* (Boston: Academic Press) p 237
- [122] Dubois L H and Nuzzo R G 1992 Synthesis, structure, and properties of model organic surfaces *Annu. Rev. Phys. Chem.* **43** 437–63
- [123] Allara D L 1995 *Biosensors and Bioelectronics* vol 10 (Amsterdam: Elsevier) pp 771–83
- [124] Gao W, Dickinson L, Grozinger C, Morin F G and Reven L 1996 Self-assembled monolayers of alkylphosphonic acids on metal oxides *Langmuir* **12** 6429–35
- [125] Bram C, Jung C and Stratmann M 1997 Self assembled molecular monolayers on oxidized inhomogeneous aluminum surfaces *Fres. J. Anal. Chem.* **358** 108–11
- [126] Sheen C W, Martensson J, Shi J, Parikh A N and Allara D L A 1992 New class of organized self-assembled monolayers—alkane thiols on GaAs(100) *J. Am. Chem. Soc.* **114** 1514–5
- [127] Allara D L and Nuzzo R G 1985 Spontaneously organized molecular assemblies. 1. Formation, dynamics, and physical-properties of normal-alkanoic acids adsorbed from solution on an oxidized aluminum surface *Langmuir* **1** 45–52
- [128] Allara D L and Nuzzo R G 1985 Spontaneously organized molecular assemblies. 2. Quantitative infrared spectroscopic determination of equilibrium structures of solution-adsorbed normal-alkanoic acids on an oxidized aluminum surface *Langmuir* **1** 52–66
- [129] Aronoff Y G, Chen B, Lu G, Seto C, Schwartz J and Bernasek S L 1997 Stabilization of self-assembled monolayers of carboxylic acids on native oxides of metals *J. Am. Chem. Soc.* **119** 259–62
- [130] Laibinis P E, Hickman J J, Wrighton M S and Whitesides G M 1989 Orthogonal self-assembled monolayers—alkanethiols on gold and alkane carboxylic-acids on alumina *Science* **245** 845–7
- [131] Folkers J P, Gorman C B, Laibinis P E, Buchholz S and Whitesides G M 1995 Self-assembled monolayers of long-chain hydroxamic acids on the native oxides of metals *Langmuir* **11** 813–24
- [132] Woodward J T, Ulman A and Schwartz D K Self-assembled monolayer growth of octadecylphosphonic acid on mica *Langmuir* **12** 3626–9
- [133] Brovelli D, Hähner G, Ruiz L, Hofer R, Kraus G, Waldner A, Schlösser J, Orszlan P, Ehrat M and Spencer N D 1999 Highly oriented alkanephosphate monolayers on tantalum(V)oxide surfaces *Langmuir* 4324–7
- [134] Netzer L and Sagiv J 1983 A new approach to construction of artificial monolayer assemblies *J. Am. Chem. Soc.* **105** 674–6

- [135] Parikh A N, Allara D L, Azouz I B and Rondelez F 1994 An intrinsic relationship between molecular-structure in self-assembled *n*-alkylsiloxane monolayers and deposition temperature *J. Phys. Chem.* **98** 7577–90
- [136] Tillman N, Ulman A, Schildkraut J S and Penner T L 1988 Incorporation of phenoxy groups in self-assembled monolayers of trichlorosilane derivatives—effects on film thickness, wettability, and molecular-orientation *J. Am. Chem. Soc.* **110** 6136–44
- [137] Tillman N, Ulman A and Elman J F 1990 A novel self-assembled monolayer film containing a sulfone-substituted aromatic group *Langmuir* **6** 1512–8
- [138] Folkers J P, Zerkowski J A, Laibinis P E, Seto C T and Whitesides G M 1992 Designing ordered molecular arrays in 2 and 3 dimensions *ACS Symp. Ser.* **499** 10–23

-32-

- [139] Laibinis P E, Fox M A, Folkers J P and Whitesides G M 1991 Comparisons of self-assembled monolayers on silver and gold—mixed monolayers derived from $\text{HS}(\text{CH}_2)_{21}\text{X}$ and $\text{HS}(\text{CH}_2)_{10}\text{Y}$ ($\text{X}, \text{Y} = \text{CH}_3, \text{CH}_2\text{OH}$) have similar properties *Langmuir* **7** 3167–73
- [140] Troughton E B, Bain C D, Whitesides G M, Nuzzo R G, Allara D L and Porter M D 1988 Monolayer films prepared by the spontaneous self-assembly of symmetrical and unsymmetrical dialkyl sulfides from solution onto gold substrates—structure, properties, and reactivity of constituent functional-groups *Langmuir* **4** 365–85
- [141] Alves C A, Smith E L and Porter M D 1982 Atomic scale imaging of alkanethiolate monolayers at gold surfaces with atomic force microscopy *J. Am. Chem. Soc.* **114** 1222–7
- [142] Widrig C A, Alves C A and Porter M D 1991 Scanning tunneling microscopy of ethanethiolate and normal-octadecanethiolate monolayers spontaneously adsorbed at gold surfaces *J. Am. Chem. Soc.* **113** 2805–10
- [143] Strong L and Whitesides G M 1988 Structures of self-assembled monolayer films of organosulphur compounds adsorbed on gold single-crystals—electron-diffraction studies *Langmuir* **4** 546–58
- [144] Chidsey C E D, Liu G, Rowntree P and Scoles G J 1989 Molecular order at the surface of an organic monolayer studied by low energy helium diffraction *J. Chem. Phys.* **91** 4421–3
- [145] Camillone N III, Chidsey C E D, Liu G and Scoles G J 1993 Superlattice structure at the surface of a monolayer of octadecanethiol self-assembled on Au(111) *J. Chem. Phys.* **98** 3503–11
- [146] Fenter P, Eisenberger P and Liang K S 1993 Chain-length dependence of the structures and phases of $\text{CH}_3(\text{CH}_2)_{n-1}\text{SH}$ self-assembled on Au(111) *Phys. Rev. Lett.* **70** 2447–50
- [147] Fenter P, Eberhardt A and Eisenberger P 1994 Self-assembly of *n*-alkyl thiols as disulphides on Au(111) *Science* **266** 1216–8
- [148] Delamarche E, Michel B, Biebuyck H A and Gerber C 1996 Golden interfaces: the surface of self-assembled monolayers *Adv. Mater.* **8** 719–29
- [149] Porter M D, Bright T B, Allara D L and Chidsey C E D 1987 Spontaneously organized molecular assemblies. 4. Structural characterization of normal-alkyl thiol monolayers on gold by optical ellipsometry, infrared-spectroscopy, and electrochemistry *J. Am. Chem. Soc.* **109** 3559–68
- [150] Nuzzo R G, Dubois L H and Allara D L 1990 Fundamental-studies of microscopic wetting on organic-surfaces. 1. formation and structural characterization of a self-consistent series of polyfunctional organic monolayers *J. Am. Chem. Soc.* **112** 558–69
- [151] Dubois L H, Zegarski B R and Nuzzo R G 1990 Temperature induced reconstruction of model organic-surfaces *J. Electron Spectrosc. Relat. Phenom.* **54/55** 1143–52
- [152] Hähner G, Kinzler M, Thümmel C, Wöll C and Grunze M 1992 Structure of self-organizing organic films: A near edge x-ray absorption fine structure investigation of thiol layers adsorbed on gold *J. Vac. Sci. Technol. A* **10** 2758–63

- [153] Poirier G E, Tarlov M J and Rushmeier H E 1994 Two-dimensional liquid phase and the $p \times \sqrt{3}$ phase of alkanethiol self-assembled monolayers on Au(111) *Langmuir* **10** 3383–6
- [154] Poirier G E and Tarlov M J 1994 The $c(4 \times 2)$ superlattice of *n*-alkanethiol monolayers self-assembled on Au(111) *Langmuir* **10** 2853–6
- [155] Poirier G E and Pylant E D 1996 The self-assembly mechanism of alkanethiols on Au(111) *Science* **272** 1145–8
- [156] Delamarche E, Michel B, Gerber C, Anselmetti D, Güntherodt H J, Wolf H and Ringsdorf H 1994 Real-space observation of nanoscale molecular domains in self-assembled monolayers *Langmuir* **10** 2869–71
-

-33-

- [157] Schönenberger C, Sondag-Hüthorst J A M, Jorritsima J and Fokink L J G 1994 What are the 'holes' in self-assembled monolayers of alkanethiols on gold? *Langmuir* **10** 611–4
- [158] Larsen N B, Biebuyck H, Delamarche E and Michel B 1997 Order in microcontact printed self-assembled monolayers *J. Am. Chem. Soc.* **119** 3017–26
- [159] Edinger K, Grunze M and Wöll C 1997 Corrosion of gold by alkanethiols *Ber. Bunsenges. Phys. Chem. Chem. Phys.* **101** 1811–5
- [160] Laibinis P E, Whitesides G M, Allara D L, Tao Y T, Parikh A N and Nuzzo R G 1991 Comparison of the structures and wetting properties of self-assembled monolayers of normal-alkanethiols on the coinage metal-surfaces, Cu, Ag, Au *J. Am. Chem. Soc.* **113** 7152–67
- [161] Walczak M M, Chung C, Stole S M, Widrig C A and Porter M D 1991 Structure and interfacial properties of spontaneously adsorbed normal-alkanethiolate monolayers on evaporated silver surfaces *J. Am. Chem. Soc.* **113** 2370–8
- [162] Fenter P, Eisenberger P, Li J, Camillone N, Bernasek S, Scoles G, Ramanarayanan T A and Liang K S 1991 Structure of $\text{CH}_3(\text{CH}_2)_{17}\text{SH}$ Self-assembled on the Ag(111) surface—an incommensurate monolayer *Langmuir* **7** 2013–6
- [163] Carpick R W and Salmeron M 1997 Scratching the surface: fundamental investigations of tribology with atomic force microscopy *Chem. Rev.* **97** 1163–94
- [164] DePalma V and Tillman N 1989 Friction and wear of self-assembled trichlorosilane monolayer films on silicon *Langmuir* **5** 868–72
- [165] Overney R M, Meyer E, Frommer J, Brodbeck D, Lüthi R, Howald L, Güntherodt H J, Fujihara M, Takano M and Gotoh Y 1992 Friction measurements on phase-separated thin-films with a modified atomic force microscope *Nature* **359** 133–5
- [166] Chaudhury M K and Owen M J 1993 Adhesion hysteresis and friction *Langmuir* **9** 29–31
- [167] Hähner G, Wöll C, Buck M and Grunze M 1993 Investigation of intermediate steps in the self-assembly of *n*-alkanethiols on gold surfaces by soft-x-ray spectroscopy *Langmuir* **9** 1955–8
- [168] Dannenberger O, Buck M and Grunze M 1999 Self-assembly of *n*-alkanethiols: A kinetic study by second harmonic generation *J. Phys. Chem. B* **103** 2202–13
- [169] Kumar A and Whitesides G M 1993 Features of gold having micrometer to centimeter dimensions can be formed through a combination of stamping with an elastomeric stamp and an alkanethiol 'ink' followed by chemical etching *Appl. Phys. Lett.* **63** 2002–4
- [170] Kumar A, Biebuyck H A and Whitesides G M 1994 Patterning self-assembled monolayers—applications in materials science *Langmuir* **10** 1498–511
- [171] Wilbur J L, Kumar A, Kim E and Whitesides G M 1994 Microfabrication by microcontact printing of self-assembled monolayers *Adv. Mater.* **6** 600–4

- [172] Xia Y and Whitesides G M 1994 Use of controlled reactive spreading of liquid alkanethiol on the surface of gold to modify the size of features produced by microcontact printing *J. Am. Chem. Soc.* **117** 3274–5
- [173] Fischer D, Marti A and Hähner G 1997 Orientation and order in microcontact-printed, self-assembled monolayers of alkanethiols on gold investigated with near edge x-ray absorption fine structure spectroscopy *J. Vac. Sci. Technol. A* **15** (4) 2173–80
- [174] Bar G, Rubin S, Parikh A N, Swanson B I, Zawodzinski T A and Wangbo M H 1997 Scanning force microscopy study of patterned monolayers of alkanethiols on gold. importance of tip-sample contact area in interpreting force modulation and friction force microscopy images *Langmuir* **13** 373–7
-

-34-

- [175] Kumar A, Biebuyck H A, Abbott N L and Whitesides G M 1992 The use of self-assembled monolayers and a selective etch to generate patterned gold features *J. Am. Chem. Soc.* **114** 9188–9
- [176] Huang J, Dahlgren D A and Hemminger J C 1994 Photopatterning of self-assembled alkanethiolate monolayers on gold: a simple monolayer photoresist utilizing aqueous chemistry *Langmuir* **10** 626–8
- [177] Lercel M J, Tiberio R C, Chapman P F, Craighead H G, Sheen C W, Parikh A N and Allara D L 1993 Self-assembled monolayer electron-beam resists on GaAs and SiO₂ *J. Vac. Sci. Technol. B* **11** 2823–8
- [178] Lee T R, Carey R L, Biebuyck H A and Whitesides G M 1994 The wetting of monolayer films exposing ionizable acids and bases *Langmuir* **10** 741–9
- [179] Creager S E and Clarke J 1994 Contact-angle titrations of mixed ω-mercaptoalkanoic acid/alkanethiol monolayers on gold. Reactive vs nonreactive spreading, and chain length effects on surface pK_a values *Langmuir* **10** 3675–83
- [180] Nuzzo R G, Fusco F A and Allara D L 1987 Spontaneously organized molecular assemblies. 3. Preparation and properties of solution adsorbed monolayers of organic disulphides on gold surfaces *J. Am. Chem. Soc.* **109** 2358–68
- [181] Lopez G P, Albers M W, Schreiber S L, Carrol R, Peralta E and Whitesides G M 1993 Convenient methods for patterning the adhesion of mammalian-cells to surfaces using self-assembled monolayers of alkanethiolates on gold *J. Am. Chem. Soc.* **115** 5877–8
- [182] Singhvi R, Kumar A, Lopez G P, Stephanopoulos G N, Wang D I C, Whitesides G M and Ingber D E 1994 Engineering cell-shape and function *Science* **264** 696–8
- [183] Delamarche E, Sundarababu G, Biebuyck H, Michel B, Gerber C, Sigrist H, Wolf H, Ringsdorf H, Xanthopoulos N and Mathieu H J 1996 Immobilization of antibodies on a photoactive self-assembled monolayer on gold *Langmuir* **12** 1997–2006
- [184] Horn A B, Russell D A, Shorthouse L J and Simpson T R E 1996 Ageing of alkanethiol self-assembled monolayers *J. Chem. Soc.-Farad. Trans.* **92** 4759–63
- [185] Nuzzo R G, Korenic E M and Dubois L H 1990 Studies of the temperature-dependent phase behaviour of long chain *n*-alkyl thiol monolayers on gold *J. Chem. Phys.* **93** 767–73
- [186] Lenk T J, Hallmark V M, Hoffmann C L, Rabolt J F, Castner D G, Erdelen C and Ringsdorf H 1994 Structural investigation of molecular organization in self-assembled monolayers of a semifluorinated amidethiol *Langmuir* **10** 4610–7
- [187] Tam-Chang S W, Biebuyck H A, Whitesides G M, Jeon N and Nuzzo R G 1995 Self-assembled monolayers on gold generated from alkanethiols with the structure RNHCOCH(2)SH *Langmuir* **11** 4371–82
- [188] Bhushan B, Kulkarni A V, Koinkar V N, Boehm M, Odoni L, Martelet C and Belin M 1995 Microtribological characterization of self-assembled and Langmuir–Blodgett monolayers by atomic and friction force microscopy *Langmuir* **11** 3189–98

- [189] Tsukruk V V, Everson M P, Lander L M and Brittain W J 1996 Nanotribological properties of composite molecular films: C₆₀ anchored to a self-assembled monolayer *Langmuir* **12** 3905–11
- [190] Stewart R, Whitesides G M, Godfried H P and Silvera I F 1986 Improved adhesion of thin conformal organic films to metal surfaces *Rev.Sci. Instrum.* **57** 1381–3
- [191] Young J T, Boerio F J, Zhang Z and Beck T L 1996 Molecular structure of monolayers from thiol-terminated polyimide model compounds on gold. 1. A spectroscopic investigation *Langmuir* **12** 1219–26
- [192] Volmer M, Stratmann M and Viefhaus H 1990 Electrochemical and electron spectroscopic investigations of iron surfaces modified with thiols *Surf. Interf. Anal.* **16** 278–82
-

-35-

- [193] Stern D A, Wellner E, Salaita G N, Laguren-Davidson L, Lu F, Batina N, Frank D G, Zapfen D C, Walton N and Hubbard A T 1988 Adsorbed thiophenol and related compounds studied at Pt(111) electrode by EELS, Auger-spectroscopy, and cyclic voltammetry *J. Am. Chem. Soc.* **110** 4885–93
- [194] Bain C D and Whitesides G M 1989 A study by contact-angle of the acid–base behaviour of monolayers containing omega-mercaptocarboxylic acids adsorbed on gold—an example of reactive spreading *Langmuir* **5** 1370–8
- [195] Prime K L and Whitesides G M 1991 Self-assembled organic monolayers—model systems for studying adsorption of proteins at surfaces *Science* **252** 1164–7
- [196] Pale-Grosdemange C, Simon E S, Prime K L and Whitesides G M 1991 Formation of self-assembled monolayers by chemisorption of derivatives of oligo(ethylene glycol) of structure HS(CH₂)₁₁(OCH₂CH₂)_nmeta-OH on gold *J. Am. Chem. Soc.* **113** 12–20
- [197] Häussling L, Michel B, Ringsdorf H and Rohrer H 1991 Direct observation of streptavidin specifically adsorbed on biotin-functionalized self-assembled monolayers with the scanning tunneling microscope *Angew. Chem. Int. Edn. Engl.* **30** 569–72
- [198] Häussling L, Ringsdorf H, Schmitt F J and Knoll W Biotin-functionalized self-assembled monolayers on gold—surface-plasmon optical studies of specific recognition reactions *Langmuir* **7** 1837–40
- [199] Spinke J, Liley M, Schmitt F J, Guder H J, Angermaier L and Knoll W 1993 Molecular recognition at self-assembled monolayers—optimization of surface functionalization *J. Chem. Phys.* **99** 7012–9
- [200] Spinke J, Liley M, Guder H J, Angermaier L and Knoll W 1993 Molecular recognition at self-assembled monolayers—the construction of multicomponent multilayers *Langmuir* **9** 1821–5
- [201] Schierbaum K D, Weiss T, Thoden van Velzen E U, Engbersen J F J, Reinhoudt D N and Göpel W 1994 Molecular recognition by self-assembled monolayers of cavitand receptors *Science* **265** 1413–5
- [202] Norde W and Lyklema J 1979 Thermodynamics of protein adsorption *J. Colloid Interface Sci.* **71** 350–66
- [203] Norde W 1986 Adsorption of proteins from solution at the solid–liquid interface *Adv. Colloid Interface Sci.* **25** 267–340
- [204] Andrade J D and Hlady V 1987 Plasma-protein adsorption—the big 12 *Ann. NY Acad. Sci.* 158-1-72
- [205] Brynda E, Hlady V and Andrade J D 1990 Protein packing in adsorbed layers studied by excitation-energy transfer *J. Colloid Interface Sci.* **139** 374–80
- [206] Golander C G, Lin Y S, Hlady V and Andrade J D 1990 Wetting and plasma-protein adsorption studies using surfaces with a hydrophobicity gradient *Colloids Surf.* **49** 289–302
- [207] Yamada K M and Kennedy D W 1978 Dualistic nature of adhesive protein function: fibronectin and its biologically active peptide fragments can autoinhibit fibronectin function *J. Cell Biol.* **99** 29–36

- [208] Lewandowska K, Balachander N, Sukenik C N and Culp L A 1989 Modulation of fibronectin adhesive functions for fibroblasts and neural cells by chemically derivatized substrates *J. Cell Physiol.* **141** 334–45
- [209] Grinnell F and Phan T V 1985 Platelet attachment and spreading on polystyrene surfaces—dependence on fibronectin and plasm-concentration *Thromb. Res.* **39** 165–71
- [210] Chinn J A, Horbett T A, Ratner B D, Schway M B, Haque Y and Hauschka S D 1989 Enhancement of serum fibronectin adsorption and the clonal plating of swiss mouse-3T3 fibroblast and MM14-mouse myoblast cells on polymer substrates modified by radiofrequency plasma deposition *J. Colloid Interface Sci.* **127** 67–87
- [211] Dekker A, Reitsma K, Beugeling T, Bantjes A, Feijen J and van Aken W G 1991 Adhesion of endothelial-cells and adsorption of serum-proteins on gas plasma-treated polytetrafluoroethylene *Biomaterials* **12** 130–8
-

-36-

- [212] Maroudas N G 1977 Sulphonated polystyrene as an optimal substratum for the adhesion and spreading of mesenchymal cells in monovalent and divalent saline solutions *J. Cell. Physiol.* **90** 511–20
- [213] van Wachem P B, Beugeling T, Feijen J, Bantjes A, Detmers J P and van Aken W G 1985 Interaction of cultured human endothelial cells with polymeric surfaces of different wettabilities *Biomaterials* **6** 403–8
- [214] Pratt K J, Williams S K and Jarrell B E 1989 Enhanced adherence of human adult endothelial cells to plasma discharge modified polyethylene *J. Biomed. Mater. Res.* **23** 1131–47
- [215] Dekker A, Beugeling T, Wind H, Poot A, Bantjes A, Feijen J and van Aken W G 1991 Deposition of cellular fibronectin and desorption of human serum-albumin during adhesion and spreading of human endothelial-cells on polymers *J. Mater. Sci.* **2** 227–33
- [216] Liedberg B and Tengvall P 1995 Molecular gradients of ω -substituted alkanethiols on gold: preparation and characterization *Langmuir* **11** 3821–7
- [217] Bain C D and Whitesides G M 1989 Formation of monolayers by the coadsorption of thiols on gold—variation in the length of the alkyl chain *J. Am. Chem. Soc.* **111** 7164–75
- [218] Bain C D, Evall J and Whitesides G M 1989 Formation of monolayers by the coadsorption of thiols on gold—variation in the head group, tail group, and solvent *J. Am. Chem. Soc.* **111** 7155–64
- [219] Bertilsson L and Liedberg B 1993 Infrared study of thiol monolayers assemblies on gold—preparation, characterization, and functionalization of mixed monolayers *Langmuir* **9** 141–9
- [220] Atre S V, Liedberg B and Allara D L 1995 Chain length dependence of the structure and wetting properties in binary composition monolayers of OH⁻ and CH₃-terminated alkanethiolates on gold *Langmuir* **11** 3882–93
- [221] Fritz M C, Hähner G, Spencer N D, Bürli R and Vasella A 1996 Self-assembled hexasaccharides: surface characterization of thiol-terminated sugars adsorbed on a gold surface *Langmuir* **12** 6074–82
- [222] Motesharei K and Myles D C 1994 Molecular recognition in membrane mimics—a fluorescence probe *J. Am. Chem. Soc.* **116** 7413–14
- [223] Harder P, Grunze M, Dahint R, Whitesides G M and Laibinis P E 1998 Molecular conformation in oligo(ethylene glycol)-terminated self-assembled monolayers on gold and silver surfaces determines their ability to resist protein adsorption *J. Phys. Chem. B* **102** 426–36
- [224] Feldman K, Hähner G, Spencer N D, Harder P, and Grunze M 1999 Probing resistance to protein adsorption of oligo(ethylene glycol)-terminated self-assembled monolayers by scanning force microscopy *J. Am. Chem. Soc.* at press
- [225] DiMilla P A, Folkers J P, Biebuyck H A, Härter R, Lopez G P and Whitesides G M 1994 Wetting and protein adsorption of the self-assembled monolayers of alkanethiolates which are supported on transparent films of gold *J. Am. Chem. Soc.* **116** 2225–6
- [226] Zaugg F G, Spencer N D, Wagner P, Kernen P, Vinckier A, Groscurth P and Semenza G 1999 Microstructured

bioreactive surfaces: covalent immobilization of proteins on Au(111)/silicon via aminoreactive alkanethiolate self-assembled monolayers *J. Mater. Sci.: Mater. Med.* **10** 255–63

- [227] Waldman D A, Kolb B U, McCarthy T J and Hsu S L 1988 Infrared study of adsorbed monolayers of poly(styrene-propylene sulphide) (PS-PPS) block copolymers *Polym. Mater. Sci. Eng.* **59** 326–33
- [228] Enriquez E P, Gray K H, Guarisco V F, Linton R W, Mar K D and Samulski E T 1992 Behaviour of rigid macromolecules in self-assembly at an interface *J. Vac. Sci. Technol. A* **10** 2775–82
-

-37-

- [229] Lenk T J, Hallmark V M, Rabolt J F, Häussling L and Ringsdorf H 1993 Formation and characterization of self-assembled films of sulphur-derivatized poly(methyl methacrylates) on gold *Macromolecules* **26** 1230–7
- [230] Sun F, Castner D G and Grainger D W Ultrathin 1993 Self-assembled polymeric films on solid-surfaces. 2. Formation of 11-(n-pentylidithio)undecanoate-bearing polyacrylate monolayers on gold *Langmuir* **9** 3200–7
- [231] Steiner U B, Caseri W R, Suter U W, Rehahn M and Schmitz L 1993 Ultrathin layers of low and high-molecular-weight imides on gold and copper *Langmuir* **9** 3245–54
- [232] Erdelen C, Häussling L, Naumann R, Ringsdorf H, Wolf H, Yang J, Liliey M, Spinke J and Knoll W 1994 Self-assembled disulphide-functionalized amphiphilic copolymers on gold *Langmuir* **10** 1246–50
- [233] Steiner U B, Caseri W R, Suter U W, Rehahn M and Rau I U 1994 Self-assembled layers of substituted poly(p-phenylene)s on gold and copper *Langmuir* **10** 1164–70
- [234] Stouffer J M and McCarthy T J 1988 Polymer monolayers prepared by the spontaneous adsorption of sulphur-functionalized polystyrene on gold surfaces *Macromolecules* **21** 1204–8
- [235] Stouffer J M and McCarthy T J 1986 Self-assembly of sulphur functionalization on the adsorption of polystyrene on gold *Polym. Prepr.* **27**(2) 242–5
- [236] Batchelder D N, Evans S D, Freeman T L, Häussling L, Ringsdorf H and Wolf H 1994 Self-assembled monolayers containing polydiacetylenes *J. Am. Chem. Soc.* **116** 1050–3
- [237] Koutsos V, van der Vegte E W, Pelletier E, Stamouli A and Hadziioannou G 1997 Structure of chemically end-grafted polymer chains studied by scanning force microscopy in bad-solvent conditions *Macromolecules* **30** 4719–26
- [238] Koutsos V, van der Vegte E W, Grim P C M and Hadziioannou G Isolated polymer chains via mixed self-assembled monolayers: morphology and friction studied by scanning force microscopy *Macromolecules* **31** 116–23
- [239] Koutsos V, van der Vegte E W and Hadziioannou G 1999 Direct view of structural regimes of end-grafted polymer monolayers: a scanning force microscopy study *Macromolecules* **32** 1233–6
-

-1-

C2.5 Introducing protein folding using simple models

D Thirumalai and D K Klimov

C2.5.1 INTRODUCTION

Most reactions in cells are carried out by enzymes [1]. In many instances the rates of enzyme-catalysed reactions are enhanced by a factor of a million. A significantly large fraction of all known enzymes are proteins which are made from twenty naturally occurring amino acids. The amino acids are linked by peptide bonds to form polypeptide chains. The primary sequence of a protein specifies the linear order in which the amino acids are linked. To carry out the catalytic activity the linear sequence has to fold to a well defined three-dimensional (3D) structure. In cells only a relatively small fraction of proteins require assistance from chaperones (helper proteins) [2]. Even in the complicated cellular environment most proteins fold spontaneously upon synthesis. The determination of the 3D folded structure from the one-dimensional primary sequence is the most popular protein folding problem.

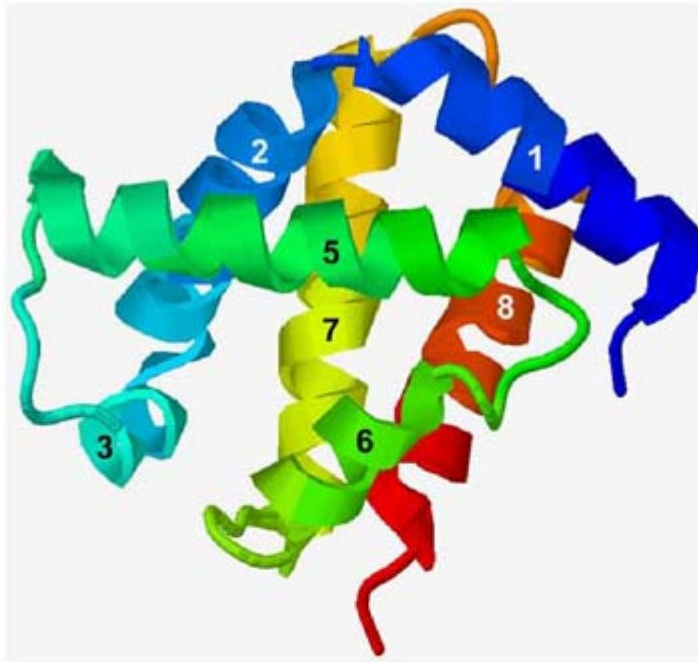
For a number of years the protein folding problem remained only of academic interest. The synthesis of proteins in cells was described by Crick in [1]. Schematically this process can be represented as DNA → RNA → Proteins. This proposal and Anfinsen's [3] demonstration that a denatured protein would fold to the native conformation under suitable conditions was sufficient to understand the role of protein folding in cells. However, in the last couple of decades many diseases have been directly linked to protein folding (especially misfolding) [4]. Thus, there is an urgency to understand the mechanisms in the formation of folded structures. The biotechnology industry is also interested in the problem because of the hope that by understanding the way polypeptide chains fold one can design molecules (using natural or synthetic constituents) of use in medicine. Finally, the full potential of the human genome project involves understanding what the genes encode. For all these profoundly important reasons the protein folding problem has taken centre stage in molecular biology.

Because this problem is complex several avenues of attack have been devised in the last fifteen years. A combination of experimental developments (protein engineering, advances in x-ray and nuclear magnetic resonance (NMR), various time-resolved spectroscopies, single molecule manipulation methods) and theoretical approaches (use of statistical mechanics, different computational strategies, use of simple models) [5, 6 and 7] has led to a greater understanding of how polypeptide chains reach the native conformation.

From our perspective there are four major problems that comprise the protein folding enterprise. They are:

- (a) Prediction of the 3D fold of a protein given only the amino acid sequence. This is referred to as the structure prediction problem. In [figure C2.5.1](#) we show the 3D structure of haemoglobin. The structure prediction problem involves determining this fold using the primary sequence which is given in the lower part of [figure C2.5.1](#).
-

Given that a sequence folds to a known native structure, what are the mechanisms in the transition from the unfolded conformation to the folded state? This is a kinetics problem, the solution of which requires elucidation of the pathways and transition states in the folding process.



1val - 2leu - 3ser - 4pro - 5ala - 6asp - 7lys - 8thr - 9asn - 10val - 11lys
 12ala - 13ala - 14trp - 15gly - 16lys - 17val - 18gly - 19ala - 20his - 21ala - 22gly
 23glu - 24tyr - 25gly - 26ala - 27glu - 28ala - 29leu - 30glu - 31arg - 32met - 33phe
 34leu - 35ser - 36phe - 37pro - 38thr - 39thr - 40lys - 41thr - 42tyr - 43phe - 44pro
 45his - 46phe - 47asp - 48leu - 49ser - 50his - 51gly - 52ser - 53ala - 54gln - 55val
 56lys - 57gly - 58his - 59gly - 60lys - 61lys - 62val - 63ala - 64asp - 65ala - 66leu
 67thr - 68asn - 69ala - 70val - 71ala - 72his - 73val - 74asp - 75asp - 76met - 77pro
 78asn - 79ala - 80leu - 81ser - 82ala - 83leu - 84ser - 85asp - 86leu - 87his - 88ala
 89his - 90lys - 91leu - 92arg - 93val - 94asp - 95pro - 96val - 97asn - 98phe - 99lys
 100leu - 101leu - 102ser - 103his - 104cys - 105leu - 106leu - 107val - 108thr - 109leu - 110ala
 111ala - 112his - 113leu - 114pro - 115ala - 116glu - 117phe - 118thr - 119pro - 120ala - 121val
 122his - 123ala - 124ser - 125leu - 126asp - 127lys - 128phe - 129leu - 130ala - 131ser - 132val
 133ser - 134thr - 135val - 136leu - 137thr - 138ser - 139lys - 140tyr - 141arg

Figure C2.5.1. The 3D native structure of haemoglobin visualized using RasMol 2.6 [8]. The linear sequence of amino acids of haemoglobin is given below the figure.

- (b)
- (c)
- (d)

How to design sequences that adopt a specified fold [9]? This is the inverse protein folding problem that is vital to the biotechnology industry. There are some proteins that do not spontaneously reach the native conformation. In the cells these proteins fold with the assistance of helper molecules referred to as chaperonins. The chaperonin-mediated folding problem involves an understanding of the interactions between proteins.

It is not the aim of this section to introduce the reader to all the areas listed above. Our goal is modest. We describe some of the theoretical developments which arose from studies of caricatures of proteins. Such models were designed in order to understand certain general features about protein structures and how these are kinetically reached. To keep the bibliography compact we mostly cite review articles. The interested reader can find the original papers in these cited works. We hope that this short introduction will entice the reader to delve into the ever surprising world of biological macromolecules.

C2.5.2 RANDOM HETEROPOLYMER AS A CARICATURE OF PROTEINS

In homopolymers all the constituents (monomers) are identical, and hence the interactions between the monomers and between the monomers and the solvent have the same functional form. To describe the shapes of a homopolymer (in the limit of large molecular weight) it is sufficient to model the chain as a sequence of connected beads. Such a model can be used to describe the shapes that a chain can adopt in various solvent conditions. A measure of shape is the dimension of the chain as a function of the degree of polymerization, N . If N is large then the precise chemical details do not affect the way the size scales with N [10]. In such a description a homopolymer is characterized in terms of a single parameter that essentially characterizes the effective interaction between the beads, which is obtained by integrating over the solvent coordinates.

Proteins are clearly not homopolymers because many energy scales are required to characterize the polypeptide chain. Besides the excluded volume interactions and hydrogen bonds the potential between the side chain depends on the nature of the residues [1]. Therefore, as a caricature of proteins the heteropolymer model is a better approximation. A convenient limit is the random heteropolymer for which approximate analytic treatments are possible [11]. In a random heteropolymer the interactions between the beads are assumed to be randomly distributed. Some of the interactions are attractive (which are responsible for conferring globularity to the chain) while others are repulsive and these residues are better accommodated in an extended conformation. In proteins water is a good solvent for polar residues while it is a poor solvent for hydrophobic residues. (In a good solvent contacts between the monomer and solvent are favoured whereas in a poor solvent the monomers are attracted to each other.) Because only 55% of the residues in proteins are hydrophobic it is clear that in a typical protein energetic frustration plays a role. In addition because of chain connectivity there is also topological frustration. This arises because residues that are proximal tend to form structures on short-length scales. The assembly of such short-length scale structures would typically be incompatible with the global fold giving rise to topological frustration. Even if energetic frustrations are eliminated a polypeptide chain (in fact any biomolecule) is topologically frustrated [7].

In the field of spin glasses and structural glasses such frustration effects are well known [12]. Thus, it was natural to suggest that random heteropolymers could serve as a simple representation of polypeptide chains. In appendix C2.5.A we sketch computational details for one model of random heteropolymers. Bryngelson and Wolynes [13] proposed, using phenomenological arguments, that the random energy model (REM) would be an appropriate description of some aspects of proteins. The rationale for this is the following: consider the exponentially large number of conformations. Because of the presence of several conflicting energies in a polypeptide chain it is natural to assume that these energies are randomly distributed. If there are no correlations between these energies and if the distribution is Gaussian one gets the REM. Of course, in the REM model the chain connectivity is ignored and there is no manifest for a spatial dependence of the chain coordinates. We show in appendix C2.5.B that in the compact phase the random heteropolymer is equivalent to REM.

The random heteropolymer models of proteins are interesting from a statistical mechanics perspective. However, they do not explain the key characteristics of proteins, such as reversible and cooperative folding to a unique native conformation. Moreover, the theories for heteropolymers suggest that, typically, the energy landscapes for these systems are extremely rugged consisting of many minima that are separated by barriers of varying heights [11]. This would mean that kinetically it would be impossible for chain with a typical realization of interactions to reach

the ground state in finite time scales. Thus, the dynamics of such random heteropolymer models typically exhibits glassy behaviour. Natural proteins do not exhibit any hallmarks of glassy dynamics at most temperatures of interest. It follows that a certain refinement of the random heteropolymers is required to capture protein-like properties. One of the important theoretical advances is the observation that very simple minimal models [14] can be constructed that capture many (not all) of the salient features observed in proteins [14]. The simplest manifestation of such models are the lattice representation of polypeptide chains. In the next section we introduce the models and describe a few results that have been obtained by numerically exploring their behaviour.

C2.5.3 LATTICE MODELS OF PROTEINS

The computational protocol for describing protein folding mechanisms is straightforward in principle. The dynamics is well described by the classical equations of motion. Simulations of a monomeric protein involves equilibrating the polypeptide chain in a box of water molecules at the desired temperature and density. If an appropriately long trajectory is generated then the dynamics of the protein can be directly monitored. There are two crucial limitations that prevent a straightforward application of this approach to a study of the folding of proteins. First, the interaction potentials or the force fields for such a complex system are not precisely known. Molecular dynamics simulations in the standard packages use potentials that rely on the transferability hypothesis, i.e. that interactions designed in one context can be used in aqueous medium and for larger systems. The need to compute potentials that can be used reliably in simulations of protein dynamics remains acute.

The second problem is related to the limitations in generating really long duration trajectories that can sample all the relevant conformational spaces of proteins. To observe reversible folding of even a moderate sized protein requires simulations that span the millisecond time scale. More importantly, making comparisons with experiments involves generating many (greater than perhaps 100) folding trajectories so that a reliable ensemble average is obtained. Thus, we need to make progress on both fronts (force fields and enhanced sampling techniques in long duration simulations) before straightforward all-atom simulations become routine.

In the light of the previously mentioned difficulties various simplified models of proteins have been suggested [14]. The main rationale for using such drastic simplifications is that a detailed study of such models can enable us to decipher certain general principles that govern the folding of proteins [5, 6 and 7]. For this class of model detailed computations without sacrificing accuracy is possible. Such an approach has yielded considerable insights into the mechanisms, time scales and pathways in the folding of polypeptide chains. In this section we will outline some of the results that have been obtained (largely from our group) with the aid of simple lattice models of proteins.

In the simple version of the lattice representation of proteins the polypeptide chain is modelled as a sequence of connected beads. The beads are confined to the sites of a suitable lattice. Most of the studies have used the cubic lattice. To satisfy the excluded volume condition only one bead is allowed to occupy a lattice site. If all the beads are identical we have a homopolymer model the characteristics of which on lattices have been extensively studied.

-5-

To introduce protein-like character the interactions between beads (those separated by at least three bonds) that are nearest neighbours on a lattice are assumed to depend on the nature of the beads. The energy of a conformation, specified by $\{r_i, i = 1, 2, \dots, N\}$, is

$$E(\{r_i\}) = \sum_{i < j} \Delta(|r_i - r_j| - a) B_{ij} \quad (\text{C2.5.1})$$

where N is the number of beads in the chain, a is the lattice spacing, and B_{ij} is the value of the contact interaction between beads i and j . We will consider different forms of B_{ij} . Since this model can be viewed as a coarse-grained representation of the α -carbons of the polypeptide chain the value of a is typically taken to be about 3.8 Å.

Lattice models have been used for a long time in polymer physics [15]. They were instrumental in computing many properties (scaling of the size of the polymer with N , distribution of end-to-end distance, etc.) of real homopolymer chains. In the context of proteins lattice models were first introduced by Taketomi *et al* [16]. The currently popular Go model [16] only considers interactions between residues (beads on the lattice) that occur in the native (ground) state. Thus, in this ‘strong specificity limit’ only native contacts are taken into account. It follows that in this version of the Go model the chain is forced to adopt the lowest energy conformation at low temperatures. Go also considered a variant of this model in which certain nonnative contacts are allowed. Although these models were insightful, Go and co-workers did not use them to obtain plausible general principles of protein folding. This was partly due to the fact that in their studies they typically used long chains, and hence exact enumeration was not possible.

Simple lattice models, with the express purpose of obtaining minimal representations of polypeptide chains, were first suggested by Chan and Dill [17]. To account for the major interactions in proteins these authors argued that the twenty naturally occurring amino acids can be roughly divided into two categories, namely, hydrophobic (H) and polar (P). Chan and Dill have suggested that this simple HP model can capture many salient features of proteins. They also suggested that many of the conceptual puzzles (the Levinthal paradox in particular) could be addressed by systematically studying short chains. This simple exactly enumerable HP model and their variants have been used to understand cooperativity, folding kinetics, and the designability of protein structures [14]. Thus, it is instructive to describe the calculations that have been done using lattice models. A study of such models indeed provides a good introduction to the computational aspects of protein folding.

C2.5.3.1 EMERGENCE OF STRUCTURES

The sequence space of proteins is extremely dense. The number of possible protein sequences is 20^N . It is clear that even by the fastest combinatorial procedure only a very small fraction of such sequences could have been synthesized. Of course, not all of these sequences will encode protein structures which for functional purposes are constrained to have certain characteristics. A natural question that arises is how do viable protein structures emerge from the vast sea of sequence space? The two physical features of folded structures are: (1) in general native proteins are compact but not maximally so. (2) The dense interior of proteins is largely made up of hydrophobic residues and the hydrophilic residues are better accommodated on the surface. These characteristics give the folded structures a lower free energy in comparison to all other conformations.

Lattice models are particularly suited for answering the questions posed. We will show that the two physical restrictions are sufficient to rationalize the emergence of very limited (believed to be only of the order of a thousand

-6-

or so) protein -like structures. To provide a plausible answer to this question using lattice models we need to specify the form of the interaction matrix elements B_{ij} . For purposes of illustration we consider the random bond (RB) model in which the elements B_{ij} are distributed as

$$P(B_{ij}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(B_{ij} - B_0)^2}{2\sigma^2}\right). \quad (\text{C2.5.2})$$

Here σ ($=1$) is the variance in B_{ij} , and the hydrophobicity parameter B_0 is the mean value. We chose $B_0 = 0$ (half the beads are hydrophobic) and $B_0 = -0.1$. The latter is motivated by the observation that in natural proteins roughly 55% of the residues are hydrophobic [18].

Protein-like structures are not only compact but also have low energy. With this in mind we have calculated the number of compact structures (CS) as CS with low energy for a given N . The number of CS in its most general form may be written as

$$C_N(\text{CS}) \simeq \bar{Z}^N Z_1^{N(d-1)/d} N^{\gamma_c-1} \quad (\text{C2.5.3})$$

where $\ln \bar{Z}$ is the conformational free energy (in units of $k_B T$), Z_1 is the surface fugacity, d is the spatial dimension, and γ_c represents the possible logarithmic corrections to the free energy. It is clear that natural proteins are relatively unique and hence their number on an average has to grow at rates that are much smaller than that given in (C2.5.3). To explore this we have calculated by exact enumeration the number of compact structures, $C_N(\text{CS})$ and the number of minimum energy structures $C_N(\text{MES})$ as a function of N (MES are compact, but not necessarily maximally compact).

We performed exhaustive enumeration of all self-avoiding conformations to explore the conformational space of the polypeptide chain of a given length. In order to reduce the sixfold symmetry on the cubic lattice we fixed the direction of the first monomeric bond in all conformations. The remaining conformations are related by eightfold symmetry on the cubic lattice (excluding the cases when conformations are completely confined to a plane or straight line). To decrease further the number of conformations to be analysed the Martin algorithm [19] was modified to reject all conformations related by symmetry.

We define MES as those conformations, the energies of which lie within the energy interval Δ above the lowest energy E_0 . Several values for Δ were used to ensure that no qualitative changes in the results are observed. We set Δ to be constant and equal to 1.2 (or 0.6) (definition (i)). We have also tested another definition for Δ , according to which $\Delta = 1.3|E_0 - t B_0|/N$, where t is the number of nearest-neighbour contacts in the ground state (definition (ii)). It is worth noting that in the latter case Δ increases with N . Both definitions yield equivalent results. Using these definitions for Δ , we computed $C(\text{MES})$ as a function of the number of residues N .

The computational technique involves exhaustive enumeration of all self-avoiding conformations for $N \leq 15$ on a cubic lattice. In doing so we calculated the energies of all conformations according to (C2.5.1) and then determined the number of MES. Each quantity, such as the number of MES, $C(\text{MES})$, the lowest energy E_0 , the number of nearest-neighbour contacts t in the lowest energy structures, is averaged over 30 sequences. Therefore, when referring to these quantities, we will imply their average values. To test the reliability of the computational results an additional

-7-

sample of 30 random sequences was generated. Note that in the case of $C(\text{MES})$ we computed the quenched averages, i.e., $C(\text{MES}) = \exp[\ln\langle c(\text{MES}) \rangle]$, where c is the number of MES for a given sequence.

The number of MES $C(\text{MES})$ is plotted as a function of the number of residues N in figure C2.5.2 for $B_0 = -0.1$ and $\Delta = 0.6$. A pair of squares at given N represents $C(\text{MES})$ computed for two independent runs of 30 sequences each. For comparison, the number of self-avoiding walks $C(\text{SAW})$ and the number of CS $C(\text{CS})$ are also plotted in this figure (diamonds and triangles, respectively). The most striking and important result of this graph is the following. As expected on general theoretical grounds, $C(\text{SAW})$ and $C(\text{CS})$ grow exponentially with N , whereas $C(\text{MES})$ exhibits drastically different scaling behaviour. There is no variation in $C(\text{MES})$ and its value remains practically constant within the entire interval of N starting with $N = 7$. We find (see figure C2.5.2) that $C(\text{MES}) \approx 10^1$. These results suggest that $C(\text{MES})$ grows (in all likelihood) only as $\ln N$ with N . Thus the restriction of compactness and low energy of the native states may impose an upper bound on the number of distinct protein folds.

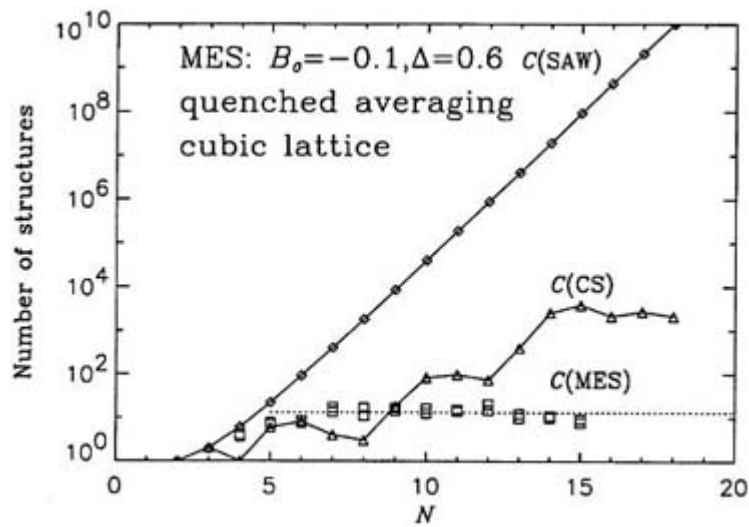


Figure C2.5.2. Scaling of the number of MES $C(\text{MES})$ (squares) is shown for the hydrophobic parameter $B_0 = -0.1$ and $\Delta = 0.6$. Data were obtained for the cubic lattice. The pairs of squares for each N represent the quenched averages for different samples of 30 sequences. The number of compact structures $C(\text{CS})$ and self-avoiding conformations $C(\text{SAW})$ are also displayed to underscore the dramatic difference of scaling behaviour of $C(\text{MES})$ and $C(\text{CS})$ (or $C(\text{SAW})$). It is clear that $C(\text{MES})$ remains practically flat, i.e. it grows no faster than $\ln N$.

C2.5.3.2 3D HP MODEL

The calculations described above suggest that upon imposing minimal restrictions on the structures (compactness and low energies) the structure space becomes sparse. As suggested before this must imply that each basin of attraction (corresponding to a given MES) in the structure space must contain numerous sequences. The way these sequences are distributed among the very slowly growing number (with respect to N) of MES, i.e. the density of sequences in structure space, is an important question. This was beautifully addressed in the paper by Li *et al* [20]. They considered a 3D ($N = 27$) cubic lattice. By using the HP model and restricting themselves to only maximally compact structures as putative native basins of attractions (NBAs) they showed that certain basins have a much larger number of sequences. In particular, they discovered that one of the NBAs serves as a ground state for 3794 (total number is 2^{27}) sequences

and hence was considered most designable (figure C2.5.3). The precise density of sequences among the NBAs is clearly a function of the interaction scheme. These calculations and the arguments presented in the previous subsection using the random bond model point out that since the number of NBA for the entire sequence space is small it is likely that proteins could have evolved randomly. Naturally occurring folds must correspond to one of the basins of attraction in the structure space so that many sequences have these folds as the native conformations, i.e. these are highly designable structures in the language of Li *et al* [20]. These ideas were further substantiated by Lindgard and Bohr [21], who showed that among maximally compact structures there are only very few folds that have protein-like characteristics. These authors also estimated, using geometrical characteristics and stability arguments, that the number of distinct folds is of the order of a thousand. All of these studies confirm that the density of the structure space is sparse. Thus, each fold can be designed by many sequences. From the purely structural point of view nature does have several options in the sense that many sequences can be ‘candidate proteins’. However, there is also evolutionary pressure to fold rapidly (i.e. a kinetic component to folding). This requirement further restricts the possible sequences that can be considered as proteins, because they must satisfy the dual criterion of reaching a definite fold on a biologically relevant time scale. These observations are schematically sketched in [figure C2.5.4](#).

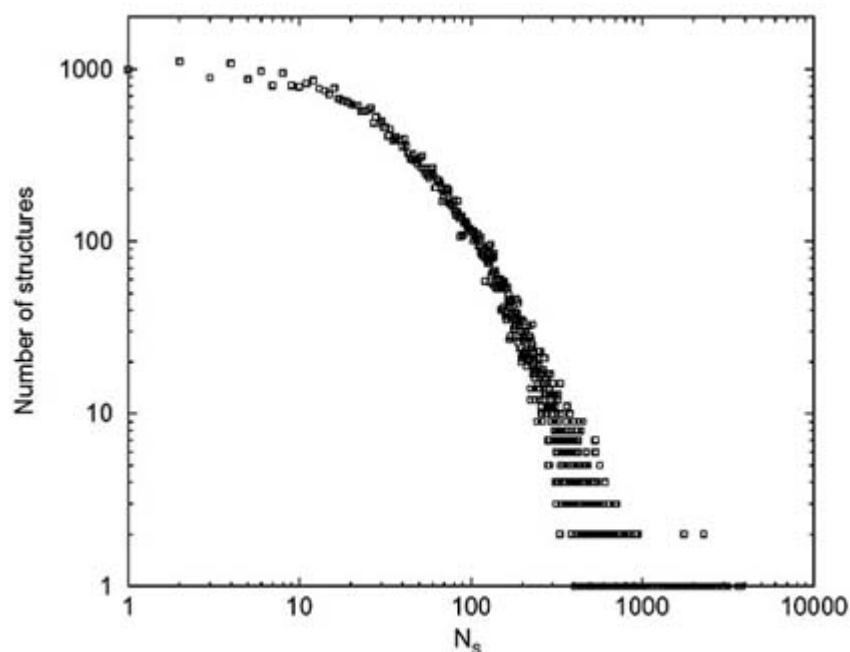


Figure C2.5.3. Histogram of the number of structures with a given number of associated sequences N_s for the $3D\ 3 \times 3$ case, in a log–log plot.

C2.5.3.3 SYMMETRY AND DESIGNABILITY

In the study by Li *et al* [20] it was noted that highly designable structures appear to be symmetric. Independently, in a thought provoking article Wolynes [22] has made a series of compelling arguments as to why nature might use symmetry (at least in an inexact manner) to generate symmetrical tertiary folds of proteins. Many enzymes are oligomers. Wolynes makes a number of observations about the symmetry aspects of protein structures: (a) the tetrameric haemoglobin (α -helical protein) molecule has an approximate two-fold symmetry.

-9-

(b) A striking example of approximate symmetry in β -proteins is found in the structure of a monomeric γ crystallin in which the shapes adopted by residues 1–88 and 89–174 are nearly the same. However, the two individual sequences do not bear much similarity. This, of course, is consistent with the notion that the structure space is so sparse that many sequences are forced to adopt similar shapes. The interesting conclusion from examining the γ crystallin structure is that the underlying symmetries in the shape are only inexact. (c) The obvious example of nearly symmetrical structures are helical proteins, with the four helix bundle being one the most prominent examples (see [figure C2.5.5](#)). (d) Various proteins with mixed topology (like triose phosphate isomerase (TIM) barrels and jelly rolls) appear to have the kind of inexact but apparently symmetrical arrangement discussed by Wolynes [23]. (e) He also conjectured that it is likely that the underlying approximate symmetry is reflected in the free energy landscape being funnel like. This would facilitate rapid folding which for many proteins may be a result of evolutionary pressure. The precise connections between the symmetries and the folding mechanisms and functional competence of biological molecules have not been worked out. Nevertheless, it appears that employing such ideas might be useful in the *de novo* design of proteins.

We note that equally striking are the kind of symmetrical arrangements found in RNA molecules [24]. The crystal structure of the P4–P6 domain of *Tetrahymena* self-splicing RNA clearly is highly symmetric with helices packed in a nearly regular arrangement. Since in an evolutionary sense the RNA world preceded the protein world it is interesting to speculate that the emergence of inexact symmetries may have been a biological necessity. The observation of inexact symmetries in a protein structure might be a consequence of the fact that they are present in the ‘parent’ molecules. In fact this evolutionary conservation may have been imprinted when evolution from the RNA world to the current scheme for protein synthesis took place. The most compelling reason for observing near regular patterns in biomolecular structures is because synthesis of symmetrical folds might be energetically

economical.

C2.5.3.4 EXPLORING THE PROTEIN FOLDING MECHANISM USING THE LATTICE MODEL

It is well known that proteins reach the biologically active native states in a relatively short time, which is of the order of a second for most single domain proteins [1]. Based on folding and refolding experiments on ribonuclease A, Anfinsen concluded that under appropriate conditions natural sequences of proteins spontaneously fold to their native conformation [3]. This implies that protein folding is a self-assembly process, i.e. the information needed for specifying the topology of the native state is contained in the primary sequence itself. This thermodynamic hypothesis does not, however, address the question of how the native state is accessed in a short time scale. This issue was raised by Levinthal who wondered how a polypeptide chain of reasonable length can navigate the astronomically large conformational space so rapidly. Levinthal posited that certain preferred pathways must guide the chain to the native state. The Levinthal paradox, simplistic as it is, has served as an intellectual impetus to gain an understanding of the ease with which a polypeptide chain reaches the native conformation [5, 6]. We use lattice models to describe the foldability of biological sequences of proteins. A sequence is foldable if it reaches the native state in a reasonable time and remains stable over some range of external conditions (pH, temperature).

C2.5.3.5 CHARACTERISTIC TEMPERATURES

The basic features of folding can be understood in terms of two fundamental equilibrium temperatures that determine the ‘phases’ of the system [7]. At sufficiently high temperatures (kT greater than all the attractive interactions) the shape of the polypeptide chain can be described as a random coil and hence its behaviour is the same as a self-avoiding walk. As the temperature is lowered one expects a transition at $T = T_\theta$ to a compact phase. This transition is very much in the spirit of the collapse transition familiar in the theory of homopolymers [10]. The number of compact

-10-

conformations at T_θ is still exponentially large. Because the polypeptide chains have additional energy scales that discriminate between the various compact conformations we expect a transition to the ground (native) state at a lower temperature T_F . Generally the transition at T_θ is second order, while the transition at T_F is similar to first order. Since we are considering finite systems, the notion of ‘phases’ should be used with care. These expectations, based on fairly general arguments, have been confirmed in various lattice simulations of protein-like heteropolymers. For the lattice models the collapse temperature T_θ is determined from the peak of the specific heat and the folding transition temperature is obtained from the fluctuations in the overlap function given by

$$\Delta\chi = \langle \chi^2 \rangle - \langle \chi \rangle^2 \quad (C2.5.4)$$

where

$$\chi = 1 - \frac{1}{N^2 - 3N + 2} \sum_{i < j+2} \delta(r_{ij} - r_{ij}^N) \quad (C2.5.5)$$

with r_{ij}^N referring to the native state. In [figure C2.5.6 a](#)) (for the structure displayed in [figure C2.5.7](#)) we plot the temperature dependence of C_v , which has a peak at $T_\theta = 0.83$. This figure also shows the variation of $d\langle R_g \rangle / dT$ with temperature. The peak of this curve (at 0.86) almost coincides with that of the specific heat indicating that this transition is associated with compaction of the chain. Hence, the maximum in C_v legitimately indicates the collapse temperature. X-ray scattering experiments have been used to obtain T_θ for a few proteins. In [figure C2.5.6 \(a\)](#) we also show the temperature dependence of $\Delta\chi$ from which the folding temperature T_F is determined to be 0.79.

(A) FOLDING RATES

The key question we want to answer is what are the intrinsic sequence dependent factors that not only determine the folding rates but also the stability of the native state? It turns out that many of the global aspects of the folding kinetics of proteins can be understood in terms of the equilibrium transition temperatures. In particular, we will show that the key factor that governs the foldability of sequences is the single parameter

$$\sigma_T = \frac{T_\theta - T_F}{T_\theta} \quad (\text{C2.5.6})$$

which indicates how far T_F is from T_θ . To establish a direct correlation between the folding time τ_F and σ_T we generated a number of sequences for $N = 27$. The folding time was taken to be equal to the mean first passage time. The first passage time for a given initial trajectory was calculated by determining the total number of Monte Carlo steps (MCS) needed to reach the native conformation for the first time. By averaging over an ensemble of initial trajectories (typically this number varies between 400–800 in our examples) the mean first passage time is obtained. The precise moves that are utilized in the simulations are described elsewhere [18]. The dependence of τ_F on σ_T (for the random bond model and for the other interaction schemes) is given in [figure C2.5.8](#). This figure shows a remarkable correlation between the folding time and σ_T . A small change in σ_T results in a dramatic effect (a few orders of magnitude) on the folding times. It is clear that both T_θ and T_F are dependent on the sequence. As a result mutations that preserve the native state can alter the folding rates due to the change in the σ_T values.

Using lattice models we have also established that folding rates correlate well with $Z = (E_N - E_{MS}) / \delta$, where E_N is the native state energy, E_{MS} is the average energy of the ensemble of misfolded structures, and δ is the dispersion in the contact energies. The relationship between σ and Z also suggests that, in general, the correlation between τ_F and σ should be superior. More importantly, experimental measurements of Z are difficult. On the other hand, both T_θ and T_F can be measured in scattering, CD, or fluorescence experiments. Other measures, such as energy gap (however it is defined), do not correlate with τ_F .

In the previous section we showed that because the structure space is very sparse there have to be many sequences that map onto the countable number of basins in the structure space. The kinetics here shows that not all the sequences, even for highly designable structures, are kinetically competent. Consequently, the biological requirements of stability and speed of folding severely restrict the number of evolved sequences for a given fold. This very important result is schematically shown in [figure C2.5.4](#).

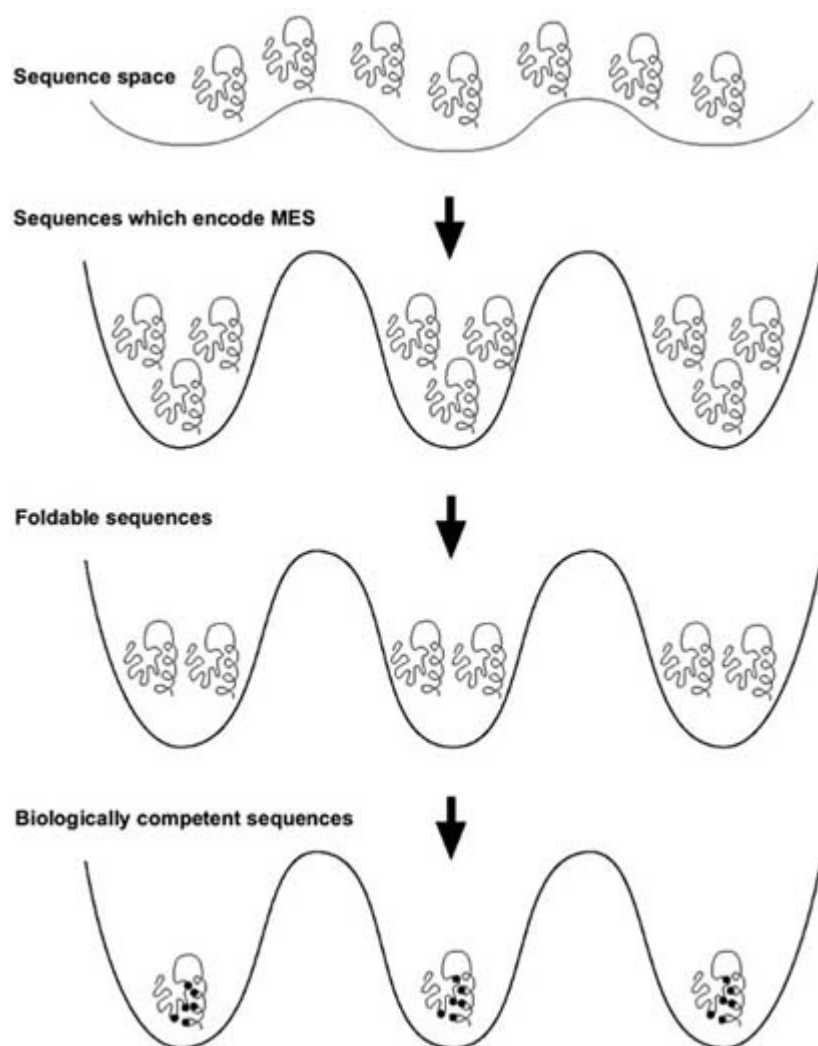


Figure C2.5.4. Schematic illustration of the stages in the drastic reduction of sequence space in the process of evolution to functionally competent protein structures.

It is important to point out that the simulations reported in [figure C2.5.8](#) were done at sequence-dependent temperatures using the condition $\langle \chi(T_s) \rangle = 0.21$. At these temperatures, all of which are below their respective folding transition temperatures, the native conformation has the highest occupation probability. In lattice models the native state is a single conformation (a microstate) which is, of course, physically unrealistic. In real systems there is a volume associated with the native basin of attraction and there are many conformations that map onto the NBA. The probability of being in the NBA at the various simulation temperatures is in excess of 0.5 so that under the conditions of our simulations the *stability criterion is automatically satisfied*. The results in [figure C2.5.8](#) therefore, shows that the *dual requirement of stability and the kinetic accessibility* of NBA is most easily satisfied by those sequences that have small values of σ_T . Thus rapid folding occurs when $\sigma_T \approx 0$, i.e. near a multicritical-like point. In this case there are no detectable intermediate ‘phases’. The sequence, whose native state is shown in [figure C2.5.7](#) has $\sigma_T = 0.05$. We found that this sequence folds rapidly.



Figure C2.5.5. Native structure of acyl-coenzyme A binding protein (first NMR structure out of 29 deposited to PDB). The figure was created using RasMol 2.6 [8].

-13-

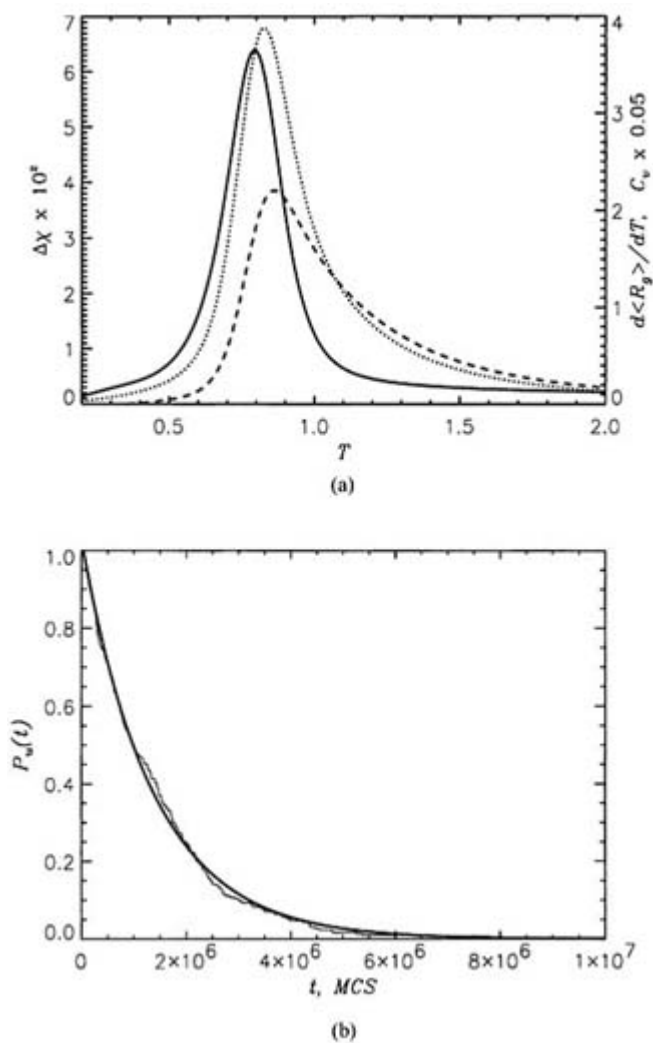


Figure C2.5.6. Thermodynamic functions computed for the sequence whose native state is shown in [figure C2.5.7](#). (a) Specific heat C_v (dotted curve) and derivative of the radius of gyration with respect to temperature $d\langle R_g \rangle/dT$ (broken curve) as a function of temperature. The collapse temperature T_θ is determined from the peak of C_v and found to be 0.83. T_θ is very close to the temperature at which $d\langle R_g \rangle/dT$ becomes maximum (0.86). This illustrates

that T_θ is indeed associated with the compaction of the chain. The temperature dependence of fluctuations of overlap function $\Delta\chi$ is given by the full curve. The folding transition temperature T_F is obtained from the peak of $\Delta\chi$ and for this sequence $T_F = 0.79$. The curves are scaled to fit one plot. (b) Time dependence of the fraction of unfolded molecules $P_u(t)$ for the sequence shown in [figure C2.5.7](#) calculated at folding conditions $T_s \lesssim T_F$. The function $P_u(t)$ is computed from a distribution of first passage times τ_{li} . The first passage time for a given initial condition is the first time the trajectory reaches the native conformation. Typically an adequately converged distribution is obtained by averaging over several hundred initial conditions. For the conditions used in this simulation folding is two state, therefore, $P_u(t)$ is adequately fitted with the single exponential (thick full curve). The folding time τ_F obtained from the fit is 1.4×10^6 MCS.

-14-

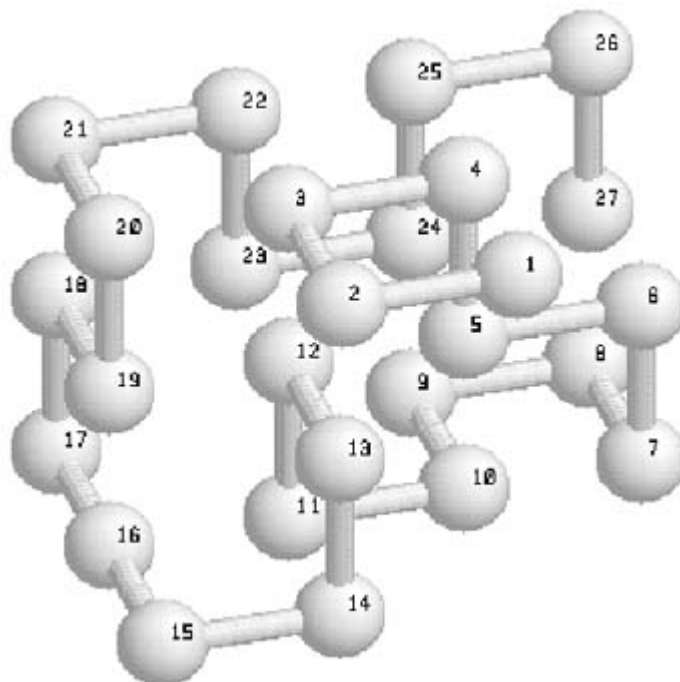


Figure C2.5.7. The native conformation of fast folding sequence ($N = 27$) with random bond potentials is shown. This structure has $c = 22$ non-bonded contacts, therefore it is not a maximally compact conformation for which $c = 28$. The figure was created using RasMol 2.6 [8].

-15-

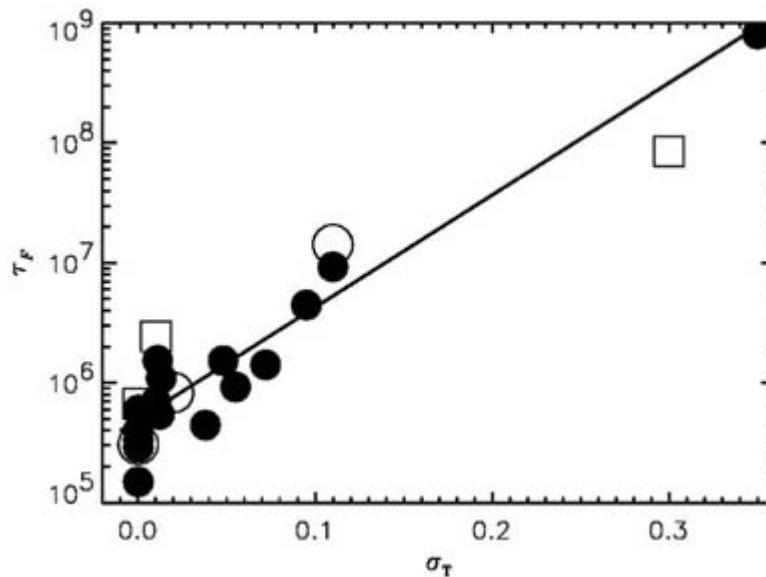


Figure C2.5.8. Plot of the folding times τ_F as a function of σ_T for the 22 sequences. This figure shows that under the external conditions when the NBA is the most populated there is a remarkable correlation between τ_F and σ_T . The correlation coefficient is 0.94. It is clear that over a four orders of magnitude of folding times $\tau_F \approx \exp(-\sigma_T/\sigma_0)$ where σ_0 is a constant. The filled and open circles correspond to different contact interactions used in [C2.5.1](#). The open squares are for $N = 36$.

(B) TOPOLOGICAL FRUSTRATION AND KINETIC PARTITIONING MECHANISM

Lattice models can also be used to obtain the outlines of the mechanisms for the folding of proteins. The qualitative aspects of the folding kinetics of biomolecules can be understood in terms of the concept of topological frustration. The primary sequence of proteins has about 55% hydrophobic residues. The linear density of hydrophobic residues along the polypeptide chain is roughly constant, implying that the hydrophobic residues are spread throughout the chain. As a result on any length scale l there is a propensity for the hydrophobic residues to form tertiary contacts under folding conditions. The resulting structures which are formed by contacts between residues that are in proximity would be in conflict with the global fold corresponding to the native state. The incompatibility of structures on local scales with the near unique native state on the global scale leads to topological frustration. It is important to realize that topological frustration is inherent to all foldable sequences, and is a direct consequence of the polymeric nature of proteins as well as competing interactions (hydrophobic residues which prefer the formation of compact structures and hydrophilic residues which are better accommodated by extended conformations). A consequence of topological frustration is that the underlying energy landscape is rugged, consisting of many minima that are separated by barriers of varying heights.

-16-

It is important to understand the nature of the low-lying minima in the rugged energy landscape. On the length scale l there are many ways of forming structures that are in conflict with the global fold. It is expected that most of these structures have high free energies and are unstable to thermal fluctuations. We expect a certain number of these structures to have low free energies and represent relatively deep minima. The similarity between these structures and the native fold could be considerable and hence these structures could be viewed as being native like. These competing basins of attraction (CBA) in which the polypeptide chain adopts native-like structures can act as kinetic traps that will slow down the folding process.

The basic consequences of topological frustration for mechanisms of folding can be understood in terms of the kinetic partitioning mechanism (KPM) [7]. Imagine an ensemble of denatured molecules in search of the native conformation. This is the experimental situation that arises when the concentration of denaturant molecules is decreased. It is clear that a fraction of molecules Φ would reach the NBA rapidly without being trapped in the low lying energy minima. The remaining fraction would be trapped in the minima and only on longer time scales do

fluctuations enable the chain to reach the NBA. The value of the partition factor Φ depends on the sequence and is explicitly determined by the σ_T value. Thus because of topological frustration the initial pool of denatured molecules partitions into fast folders and slow folders that reach the native state by indirect off-pathway processes.

From the description of the kinetic partitioning mechanism (KPM) given above it follows that generically the time dependence of the fraction of molecules that have not folded at time t , $P_u(t)$, is given by

$$P_u(t) = \Phi \exp\left(-\frac{t}{\tau_{\text{NC}}}\right) + \sum_k a_k \exp\left(-\frac{t}{\tau_k}\right) \quad (\text{C2.5.7})$$

where τ_{NC} is the time constant for reaching the native state by the fast process, τ_k is the time for escape from the CBA labelled k , and a_k is the ‘volume’ associated with the k th CBA. From this consideration we expect that for a given sequence, trajectories can be grouped into those that reach the native conformation rapidly (Φ being their fraction), and those that remain in one of the CBA for a discernible length of time. In [figure C2.5.9 a](#)) we show an example of a trajectory that reaches the native state directly from the random coil conformation. In contrast in [figure C2.5.9 \(b\)](#) we show an example of a trajectory for the same sequence at the same simulation temperature. This figure shows that on a very short time scale the chain gets trapped in conformations other than the NBA and only on a long time scale does it reach the native state. This figure illustrates the basic principle of KPM. If we perform an average over an ensemble of such trajectories the kinetic result given in (C2.5.7) ensues.

-17-

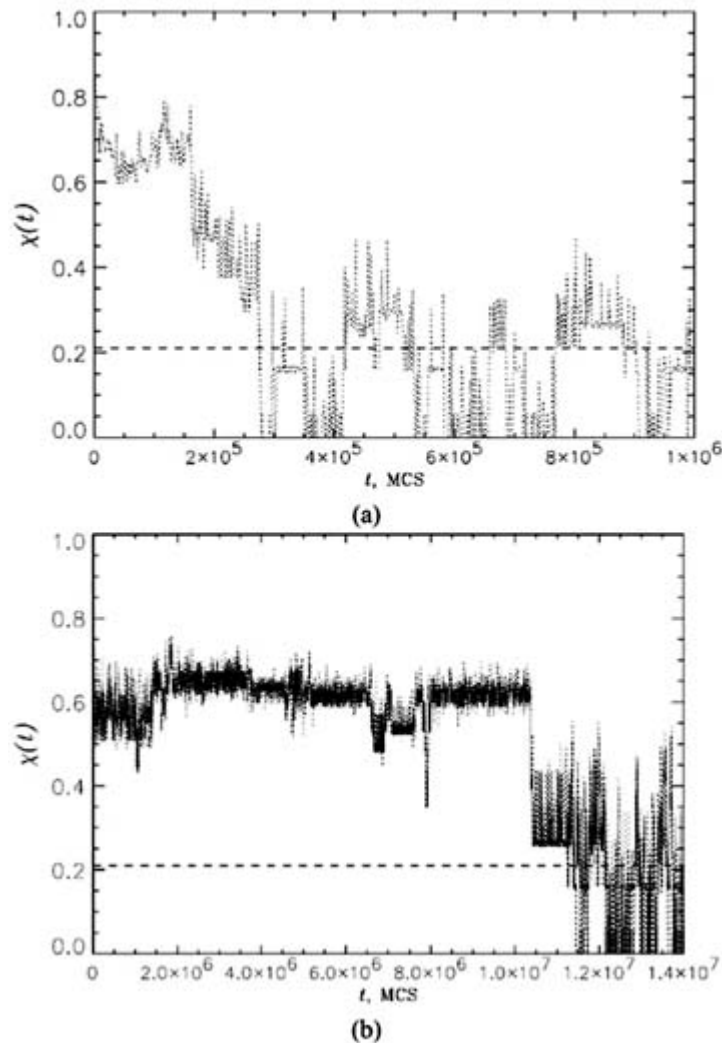


Figure C2.5.9. Examples of folding trajectories at $T = T_s$ derived from the condition $\langle \chi(T_s) \rangle = 0.21$. (a) Fast folding trajectory as monitored by $\chi(t)$. It can be seen that sequence reaches the native state very rapidly in a two-state manner without being trapped in intermediates. The first passage time for this trajectory is 277 912 MCS. (b) Slow folding trajectory for the same sequence. The sequence becomes trapped in several intermediate states with large χ *en route* to the native state. The first passage time is 11 442 793 MCS. Notice that the time scales in both panels are dramatically different.

(C) CLASSIFYING FOLDING MECHANISMS IN TERMS OF Σ_T

The various folding mechanisms expected in foldable sequences may be classified in terms of the σ_T . We have already shown that sequences that fold extremely rapidly have very small values of σ_T . Based on our study of several model proteins as well as analysis of real proteins we classify the folding kinetics of proteins in the following [7].

-18-

(D) FAST FOLDERS

For these sequences the value of σ_T is less than a certain small value σ_f . For such sequences the folding occurs directly from the ensemble of unfolded states to the NBA. The free energy surface is dominated by the NBA (or a funnel) and the volume associated with NBA is very large. The partition factor Φ is near unity so that these sequences reach the native state by two-state kinetics. The amplitudes a_k in (C2.5.7) are nearly zero. There are no intermediates in the pathways from the denatured state to the native state. Fast folders reach the native state by a nucleation–collapse mechanism which means that once a certain number of contacts (folding nuclei) are formed then the native state is reached very rapidly [25, 26]. The time scale for reaching the native state for fast folders (which are normally associated with those sequences for which topological frustration is minimal) is found to be

$$\tau_{\text{NCNC}} = \frac{\eta a}{\gamma} f(\sigma_T) N^\omega \quad (\text{C2.5.8})$$

where η is the solvent viscosity, a is the typical size of a residue, γ is the average surface tension between the residue and water, $f(\sigma_T)$ is typically an exponential function of σ_T , and the exponent ω is between 3.8 and 4.2. In general, only small proteins (N less than about 100) are fast folders.

(E) MODERATE FOLDERS

Sequences for which $\sigma_f \leq \sigma_T \leq \sigma_h$ (where σ_h is the upper boundary for moderate filters) can be classified as moderate folders. Unlike fast folding sequences the Φ values are fractional which means that a substantial fraction of molecules is essentially trapped in one of the CBAs for some length of time. For these sequences there are detectable intermediates and, for all but very small proteins, the rate determining step is the activated transition from one of the CBAs to the native state. The average time scale for transition from these misfolded structures to the native conformation is given by

$$\tau_F \approx \tau_0 \exp(\sqrt{N}) \quad (\text{C2.5.9})$$

at $T \approx T_F$. This shows that typical barriers for moderate folders are quite small. As a result the folding times even for long proteins ($N \approx 200$) are only of the order of a second. It is these small barriers that enable typical proteins to fold in a biologically relevant time scale without encountering the Levinthal paradox.

(F) SLOW FOLDERS AND CHAPERONES

For sequences with $\sigma_T \geq \sigma_h$, folding is extremely slow and these sequences may not reach the native state in a

biologically relevant time scale. The volume corresponding to NBA is very small in this case and as a result Φ is nearly zero. The free energy surface is dominated by CBAs. Under these circumstances spontaneous folding does not become viable. In cells such proteins are rescued by chaperones. Typically this happens when N is so large that $\tau_0 \exp(\sqrt{N})$ exceeds reasonable folding time scales. Thus in cells we expect that only those proteins which are large or whose biological functioning state has to be oligomers require chaperones.

-19-

C2.5.3.6 MINIMUM NUMBER OF RESIDUES FOR OBTAINING FOLDABLE PROTEIN STRUCTURES

Natural proteins are made up of twenty amino acid residues. An important question, from the perspective of protein design, is how many distinct types of residues are required for protein-like behaviour? Such a selection cannot be made arbitrarily because in natural proteins one should have polar, hydrophobic, and charged residues. In addition, for optimal packing of the core, hydrophobic residues with different van der Waals radii may be required. To explore the potential simplification of the number of residues Wang and Wang [27] have carried out a highly significant study using lattice models and standard statistical potentials for the contact interaction elements B_{ij} (C2.5.1). They discovered that a grouping of amino acid residues into five categories mimics the folding behaviour found using the standard twenty residues. To demonstrate this they used a cubic lattice with $N = 27$ and mostly focused on the maximally compact structures as ground states. Thus, structures such as ones given in [figure C2.5.6](#) are not explicitly considered. Nevertheless, the demonstration that a suitable set of five amino acid residue types is sufficient is an important result which should have implications for the protein design problem—the generation of primary sequences that can fold to a chosen target folded structure.

In their original article they mostly focused on various thermodynamic properties (nature and degeneracy of the ground states). They also carried out kinetic simulations to assess if the kinetic properties are altered by using a reduced number of residues. To test this idea Wang and Wang used the foldability index σ (which correlates well with folding rates) as a discriminator of sequence properties. The precise question addressed by Wang and Wang is the following: what is the minimum number of residues that are required to obtain foldable (characterized by having relatively small values of σ) sequences? We found that fast folding sequences have σ less than about a quarter. They carried out two sets of computations. In one set they initially optimized the stability gap [5] of various sequences using the twenty residues. They substituted the residues in these optimized sequences by the representative residue for each group. Four subgroups were considered with each containing five and a variant, three and two amino acids. The foldability index for the standard sample and their substitutes is shown in [figure C2.5.10](#) as full circles. In another set of computations they examined the foldability index (open diamonds in [figure C2.5.10](#)) for sequences that were optimized using the reduced sets of amino acids. Both these curves show that as long as the number of amino acid types exceeds five one can generate sequences with relatively small values of σ . [figure C2.5.10](#) also shows that smaller values of σ can be obtained if optimization is carried out with reduced sets of amino acids. Such sequences are foldable, i.e. the dual requirements of stability over a wide temperature range and the kinetic accessibility of their native states are simultaneously satisfied.

-20-

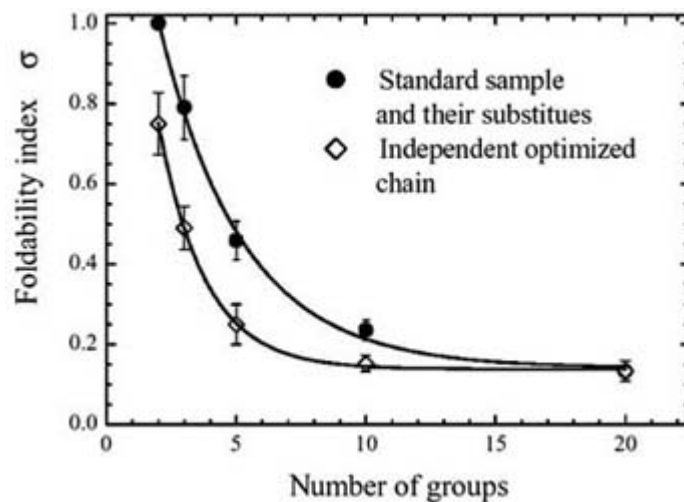


Figure C2.5.10. The figure gives the foldability index σ of 27-mer lattice chains with sets containing different number of amino acids. The sets are generated according to scheme described in [27]. The set of 20 amino acids is taken as a standard sample. Each sequence with 20 amino acids is optimized to fulfil the stability gap [5]. The residues in the standard samples are substituted with four different sets containing a smaller number of amino acids [27]. The foldability of these substitutions is indicated by the full circles. The open diamonds correspond to the sequences with same composition. However, the amino acids are chosen from the reduced representation and the resultant sequence is optimized using the stability gap [5].

C2.5.4 CONCLUSIONS

The examples of modelling discussed in [section C2.5.2](#) and [section C2.5.3](#) are meant to illustrate the ideas behind the theoretical and computational approaches to protein folding. It should be borne in mind that we have discussed only a very limited aspect of the rich field of protein folding. The computations described in [section C2.5.3](#) can be carried out easily on a desktop computer. Such an exercise is, perhaps, the best of way of appreciating the simple approach to get at the principles that govern the folding of proteins.

In this section we have not discussed experimental advances that are offering extraordinary insights into the way the denatured molecules reach the native state. Two remarkable experimental approaches hold the promise that in short order we will be able to watch the folding process from submicrosecond time scale until the native state is reached. A brief summary of these follow.

- (1) Eaton *et al* [28] have shown that optical triggers of folding can offer a window into the folding process from the microsecond time scale. Since this many laboratories have probed the plausible structure formations that occur on a short time scale. Fast folding experimental techniques have been used to obtain the detailed kinetics for the building blocks of proteins, namely, β -hairpin, α -helices, and loops. Very recent experiments have given compelling evidence that there are populated native-like intermediates even in proteins that were thought to follow two-state kinetics.

- (2) Perhaps the most exciting development in the last few years is the ability to nanomanipulate single biomolecules using atomic force microscopy and optical tweezer techniques [29]. So far such experiments have been used to provide a microscopic basis of elasticity in muscle proteins. If these stretching experiments can be combined with fluorescent resonance energy transfer experiments then it is possible to follow the folding of individual molecules as it passes through the transition state to the native conformation. It has been suggested on theoretical grounds that such two-dimensional single-molecule experiments can measure directly the distribution of folding rates (and the barrier distribution) in much the same way that mean first passage times are computed in minimal protein models (see [section C2.5.3](#)).

The challenges posed by these high precision experiments demand more refined models and further developments in computational techniques. For the theoretically inclined it will no longer be sufficient to describe kinetics only in terms of energy landscapes. The wealth of data that are being generated by experiments such as those mentioned above, requires a quantitative understanding of the various factors that govern the pathways, mechanisms, and the transition states in the folding process. These challenging issues will make the area of biomolecular folding an engaging one for many years to come.

ACKNOWLEDGMENTS

We are grateful to John D Weeks for useful comments and to Chao Tang for supplying [figure C2.5.3](#). We are indebted to Dr J Wang and Professor W Wang for kindly providing us with [figure C2.5.10](#) prior to publication.

APPENDIX C2.5.A

There are several versions of the random heteropolymer models. To keep the discussions technically simple we will consider one case—the so-called random hydrophilic–hydrophobic chain whose phases were studied by Garel *et al* (GLO) [11]. The GLO model consists of a polymer chain with N monomers. The GLO model can be viewed as a generalization of the popular Edwards model which was introduced in order to understand the swelling of real homopolymer chains in good solvents [10]. In the GLO model the chain is made up of hydrophobic (hydrophilic) residues that tend to collapse (swell) the chain when dispersed in a solvent. The solvent mediated interactions at each site are assumed to be random. The random interactions depend only on a given site i and the strength depends on the degree of hydrophilicity λ_i . Besides the term accounting for chain connectivity there are two- and higher-body interactions that determine the shape of the chain. In the GLO model the two-body interaction is given by

$$v_{ij} = v_0 + \beta(\lambda_i + \lambda_j)\delta[r_i - r_j] \quad (\text{C2.5.A1})$$

where v_0 is repulsive short-range interaction, λ_i is a quenched random variable which is distributed as

$$P(\lambda_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\lambda_i - \lambda_0)^2}{2\sigma^2}\right). \quad (\text{C2.5.A2})$$

-22-

If the mean λ_0 is positive then the majority of the residues are hydrophilic. A description of the collapsed phase of the chain requires introducing three- and four-body interaction terms. Thus, the total Hamiltonian is

$$\begin{aligned} \beta H = & \frac{1}{2} \sum_{i \neq j} v_{ij} + \frac{1}{6} \sum_{i \neq j \neq k} \omega_3 \delta(r_i - r_j) \delta(r_i - r_k) \\ & + \frac{1}{24} \sum_{i \neq j \neq k \neq l} \omega_4 \delta(r_i - r_j) \delta(r_j - r_k) \delta(r_k - r_l). \end{aligned} \quad (\text{C2.5.A3})$$

Since the charge variables λ_i are quenched the thermodynamics of the system requires averaging the free energy using the distribution $P(\lambda_i)$, i.e.

$$F = -k_B T \int \prod P(\lambda_i) \ln Z(\lambda_i) d\{\lambda_i\}. \quad (\text{C2.5.A4})$$

The average of $\ln Z(\lambda_i)$ is most conveniently done using the replicas through the relation

$$\ln Z = \lim_{n \rightarrow 0} \frac{Z^n - 1}{n}. \quad (\text{C2.5.A5})$$

Using (C2.5.A2)–(C2.5.A4) the required average can be carried out. This leads to a complicated expression for $\overline{Z^n}$ where the bar indicates the average over the quenched random variables λ_i . In terms of the order parameters

$$q_{ab}(r, r') = \int ds \delta(r_a(s) - r) \delta(r_b(s) - r') \quad (\text{C2.5.A6})$$

and

$$\rho_a(r) = \int ds \delta(r_a(s) - r) \quad (\text{C2.5.A7})$$

(a and b are replica indices) the expression for $\overline{Z^n}$ becomes

$$\overline{Z^n} = \int Dq_{ab}(r, r') D\hat{q}_{ab}(r, r') D\rho_a(r) D\phi_a(r) \exp[H_{\text{eff}}] \quad (\text{C2.5.A8})$$

where

$$H_{\text{eff}} = G(q_{ab}, \hat{q}_{ab}, \rho_a, \phi_a) + \ln \zeta(\hat{q}_{ab}, \phi_a) \quad (\text{C2.5.A9})$$

-23-

with

$$G = \int dr \sum \left(i\rho_a \phi_a - (v_0 + 2\beta\lambda_0) \frac{\rho_a^2}{2} - \frac{\omega'_3}{6} \rho_a^3 - \frac{\omega_4}{24} \rho_a^4 \right) + \int dr \int dr' \sum_{a < b} (i q_{ab}(r, r') \hat{q}_{ab}(r, r') + \beta^2 \lambda^2 q_{ab}(r, r') \rho_a(r) \rho_b(r'))$$

and

$$\zeta(\hat{q}_{ab}, \phi_a) = \int \prod_a D r_a(s) \exp(-H_T\{r_a(s)\}). \quad (\text{C2.5.A10})$$

In (C2.5.A10) $H_T\{r_a(s)\}$ is

$$H_T\{r_a(s)\} = \exp \left(-\frac{d}{2a^2} \int_0^N ds \sum_a \left(\frac{dr}{ds} \right)^2 - i \int_0^N ds \sum_a \phi_a(r_a(s)) - i \int_0^N ds \sum_{a < b} \hat{q}_{ab}(r_a, r_b) \right) \quad (\text{C2.5.A11})$$

and

$$\omega'_3 = \omega_3 - 3\beta^2 \lambda^2. \quad (\text{C2.5.A12})$$

The path integrals in (C2.5.A10) may be evaluated using the spectrum of the effective n -body Hamiltonian

$$H_n = -\frac{d}{2a^2} \sum_a \nabla_a^2 + \sum_a i\phi(r_a) + \sum_{a<b} i\hat{q}_{ab}(r_a, r_b) \quad (\text{C2.5.A13})$$

in the limit of $n \rightarrow 0$. If N is very large then we can use ground state dominance to evaluate the spectrum of H_n . This gives

and

$$\xi(\hat{q}_{ab}, \phi_a) \simeq \exp[-N \min_{\{\Psi(r)\}} \{(\Psi | H_n | \Psi) - E_0(\Psi | \Psi) - 1\}] \quad (\text{C2.5.A14})$$

where E_0 is the ground state energy of H_n . GLO evaluated the integral over q_{ab} (C2.5.A6) by a saddle point approximation which leads to

$$i\hat{q}_{ab}(r, r') = -\beta^2 \lambda^2 \rho_a(r) \rho_b(r'). \quad (\text{C2.5.A15})$$

-24-

From the above equation it follows that in the mean-field limit replica symmetry is not broken. This makes the GLO model conceptually simpler to interpret than the random bond heteropolymer model discussed in the appendix.

The total wavefunction $\Psi\{r_1, r_2, \dots, r_n\}$ is written as a product of single-particle functions (Hartree approximation). The various integrals are evaluated in the saddle point approximation. A simple Gaussian form for the trial one-particle wavefunction

$$\phi(r) = \left(\frac{1}{2\pi R^2}\right)^{d/4} \exp\left(-\frac{r^2}{2R^2}\right) \quad (\text{C2.5.A16})$$

is chosen with R being the single variational parameter. Upon performing the Gaussian integrals the free energy per monomer f becomes

$$\beta f = \frac{a^2}{8R^2} + \frac{1}{(2\sqrt{\pi})^d} \frac{(v_0 + 2\beta\lambda_0) N}{2} \frac{N}{R^d} + \Omega \quad (\text{C2.5.A17})$$

where

$$\Omega = \left(\frac{1}{(2\pi\sqrt{3})^d} \frac{\omega_3}{6} - \left(\frac{1}{2\pi}\right)^d (3^{-d/2} - 2^{-d})\right) \left(\frac{N}{R^d}\right)^2 + \left(\frac{1}{(32\pi^3)^d}\right)^{d/2} \frac{\omega_4}{24} \left(\frac{N}{R^d}\right)^3. \quad (\text{C2.5.A18})$$

At low temperatures the shape of the chain is determined by the sign of first term in (C2.5.A18). If the sign is negative then the positive four-body term is required for a stable theory.

The phase of the random hydrophobic–hydrophilic model is complicated and depends on the value of λ_n [11]. We

only describe the hydrophilic case when λ_0 is positive. In this case there is a first-order transition to a collapsed state ($R \sim N^{-1/d}$) induced by the negative three-body term. GLO have pointed out that this transition is neither the usual θ -point nor is it a freezing temperature because there is no replica symmetry breaking. In fact, this collapse transition resembles that seen in proteins where it is suspected that it is a first-order transition. The microscopic origin of the first-order transition upon collapse of polypeptide chains is not fully understood. Recent arguments suggest that it could arise because the burial of hydrophobic residues and the accommodation of the hydrophilic ones at the surface of proteins in water requires some work and perhaps this assembly happens in a discontinuous manner.

APPENDIX C2.5.B

In [section C2.5.2](#) we considered a variational-type theory to treat the thermodynamics of the random hydrophobic-hydrophilic heteropolymer. Here we describe a limiting behaviour of the random bond model [\[30\]](#).

-25-

In this appendix we show that the random bond model in the compact phase is identical to the random energy model (REM). Historically, REM was proposed as a caricature for proteins on phenomenological grounds [\[13\]](#). The heteropolymer with random bond interactions was treated using a variational theory which suggested that when the disorder increases beyond a limiting value the chain undergoes a thermodynamic glass transition. The nature of this transition is closely related to Potts glasses.

The random-bond heteropolymer is described by a Hamiltonian similar to (C2.5.A3) except that the short-range two-body term v_{ij} is taken to be random with a Gaussian distribution. In this case a three-body term with a positive value of ω_3 is needed to describe the collapsed phase. The Hamiltonian is

$$H = \sum_{i < j} (v_0 + v_{ij}) \delta(r_i - r_j) + \sum_{i \neq j \neq k} \delta(r_i - r_j) \delta(r_j - r_k) \quad (\text{C2.5.B1})$$

The distribution of the random couplings is given by

$$P(v_{ij}) = \frac{1}{\sqrt{2\pi v^2}} \exp\left(-\frac{v_{ij}^2}{2v^2}\right). \quad (\text{C2.5.B2})$$

In the collapse phase the monomer density $\rho = N/R^3$ is constant (for large N). Thus, the only conformation dependent term in (C2.5.A1) comes from the random two-body term. Because this term is a linear combination of Gaussian variables we expect that its distribution is also Gaussian and, hence, can be specified by the two moments. Let us calculate the correlation $\overline{E_1 E_2}$ between the energies E_1 and E_2 of two conformations $\{r_i^{(1)}\}$ and $\{r_i^{(2)}\}$ of the chain in the collapsed state. The mean square of E_1 is

$$\overline{E_1^2} = \frac{v^2}{2} \sum_{i,j} \delta(r_i^{(1)} - r_j^{(1)}) = \frac{v^2}{2} N\rho \quad (\text{C2.5.B3})$$

which is independent of the collapsed conformation. Similarly, we have

$$\begin{aligned}\overline{E_1 E_2} &= \frac{v^2}{2} \sum_{i,j} \delta(r_i^{(1)} - r_j^{(1)}) \delta(r_i^{(2)} - r_j^{(2)}) \\ &= \frac{v^2}{2} \sum_{r,r'} q_{12}^2(r, r')\end{aligned}$$

-26-

where $q_{12}(r, r')$ is the overlap between the two conformations. Because

$$\sum_{r,r'} q_{12}(r, r') = N \quad (\text{C2.5.B4})$$

and since the monomer density is constant we have $q_{12}(r, r') = \rho^2/N$. This implies

$$\overline{E_1 E_2} = \frac{v^2}{2} \rho^2. \quad (\text{C2.5.B5})$$

Thus the joint probability is

$$\lim_{N \rightarrow \infty} \sim \exp\left(-\frac{E_1^2 + E_2^2}{N\rho v^2}\right) \quad (\text{C2.5.B6})$$

which is equivalent to the behaviour in uncorrelated REM. Thus, it is not a surprise that in large dimensions (which are captured by variational type treatments) the random bond heteropolymer model yields exactly the same result as the REM.

REFERENCES

- [1] Stryer L 1988 *Biochemistry* (Freeman)
- [2] Lorimer G H 1996 A quantitative assessment of the role of the chaperonin proteins in protein folding *in vivo* *FASEB J.* **10** 5–9
- [3] Anfinsen C B 1973 Principles that govern the folding of protein chains *Science* **181** 223–30
- [4] Lansbury P T 1999 Evolution of amyloids: What normal protein folding can tell us about fibrillogenesis and disease *Proc. Natl Acad. Sci. (USA)* **96** 3342–4
- [5] Onuchic J N, Luthey-Schulten Z A and Wolynes P G 1997 Theory of protein folding: An energy landscape perspective *Ann. Rev. Phys. Chem.* **48** 545–600
- [6] Dill K A and Chan H S 1997 From Levinthal to pathways to funnels *Natur. Struct. Biol.* **4** 10–19
- [7] Thirumalai D and Klimov D K 1999 Deciphering the timescales and mechanisms of protein folding using minimal off-lattice models *Curr. Opin. Struct. Biol.* **9** 197–207
- [8] Sayle R and Milner-White E J 1995 Rasmol: Biomolecular graphics for all *Trends Biochem. Sci.* **20** 374–6

- [10] deGennes P G 1985 *Scaling Concepts in Polymer Physics* (Cornell: Cornell University Press)
- [11] Garel T, Orland H and Thirumalai D 1996 Analytical theories of protein folding *New Developments in Theoretical Studies of Protein Folding* ed R Elber (Singapore: World Scientific) pp 197–268
- [12] Virasoro M, Mezard M and Parisi G 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)
- [13] Bryngelson J D and Wolynes P G 1987 Spin glasses and the statistical mechanics of protein folding *Proc. Natl Acad. Sci. (USA)* **84** 7524–8
- [14] Dill K A, Bromberg S, Yue K, Fiebig K M, Yee D P, Thomas P D and Chan H S 1995 Principles of protein folding—a perspective from simple exact models *Protein Sci.* 561–602
- [15] Orr W J C 1947 Statistical treatment of polymer solutions at infinite dilution *Trans. Faraday Soc.* **43** 12–27
- [16] Taketomi H, Ueda Y and Go N 1975 Studies on protein folding, unfolding, and fluctuations by computer simulation *Int. J. Pept. Protein Res.* **7** 445–59
- [17] Chan H S and Dill K A 1989 Intrachain loops in polymers: effects of excluded volume *J. Chem. Phys.* **90** 493–509
- [18] Klimov D K and Thirumalai D 1996 Factors governing the foldability of proteins *Proteins: Struct. Funct. Genet.* **26** 411–41
- [19] Thirumalai D and Klimov D K 2000 Emergence of stable and fast folding protein structures *Stochastic Dynamics and Pattern Formation in Biological and Complex Systems* ed S Kim, K J Lee and W Sung (Melville, NY: American Institute of Physics) pp 95–111
- [20] Li H, Winfree N and Tang C 1996 Emergence of preferred structures in a simple model of protein folding *Science* **273** 666–9
- [21] Lindgard P-A and Bohr H 1996 Magic numbers in protein structures *Phys. Rev. Lett.* **77** 779–82
- [22] Wolynes P G 1996 Symmetry and the energy landscape of biomolecules *Proc. Natl Acad. Sci. (USA)* **93** 14 249–55
- [23] Wolynes P G 1997 Folding nucleus and energy landscapes of larger proteins within the capillarity approximation *Proc. Natl Acad. Sci. (USA)* **94** 6170–5
- [24] Cate J H, Gooding A R, Podell E, Zhou K, Golden B L, Kundrot C E, Cech T R and Doudna J A 1996 Crystal structure of a group I ribozyme domain: principles of RNA packing *Science* **273** 1678–85
- [25] Guo Z and Thirumalai D 1995 Kinetics of protein folding: nucleation mechanism, time scales and pathways *Biopolymers* **36** 83–103
- [26] Shakhnovich E I, Abkevich V and Ptitsyn O 1996 Conserved residues and the mechanism of protein folding *Nature* **379** 96–8
- [27] Wang J and Wang W 1999 A computational approach to simplifying the protein folding alphabet *Natur. Struct. Biol.* **6** 1033–8
- [28] Eaton W A, Munoz V, Thompson P A, Henry E R and Hofrichter J 1998 Kinetics and dynamics of loops, α -helices, β -hairpins, and fast-folding proteins *Acc. Chem. Res.* **31** 745–53
- [29] Fisher T E, Oberhauser A F, Carrion-Vazquez M, Marszalek P E and Fernandez J M 1999 The study of protein mechanics with the atomic force microscope *Trends Biochem. Sci.* **24** 379–84
- [30] Shakhnovich E and Gutin A 1989 Formation of unique structure in polypeptide chains. Theoretical

C2.6 Colloids

Jeroen S van Duijneveldt

C2.6.1 INTRODUCTION

C2.6.1.1 CLASSIFICATION OF COLLOIDS

The term colloid refers to systems where one phase is finely divided in another phase—with at least one of the dimensions in the range of about 1 nm to 1 μm . This encompasses a wide variety of systems, some of which will be mentioned below. In a narrower sense, the word colloid is often used to denote systems consisting of solid particles (or liquid droplets) suspended in a liquid. This contribution will mainly focus on such systems. On the one hand, these particles are (significantly) larger than the solvent molecules. On the other hand, they are sufficiently small to remain suspended and undergo vivid Brownian motion (after the British botanist Robert Brown, who published his observations on aqueous pollen suspensions in 1827). The term colloid (after the Greek word for ‘glue’) was coined by Thomas Graham in the 1860s, to denote substances such as gelatin, albumin and gums. In a solution, these would not pass a dialysis membrane.

First of all, a general classification can be made depending on the nature of the continuous and suspended phases: gas, liquid, or solid. The names of the corresponding colloidal systems are summarized in table C2.6.1. Traditionally, following Kruyt [1], colloids are further classified as either reversible or irreversible, depending on whether they redisperse spontaneously when they are added to a solvent. Polymer and micellar solutions would be reversible, for instance, whereas suspensions and emulsions would usually be irreversible. These terms are more or less equivalent to the terms lyophilic (solvent-loving) and lyophobic (solvent-hating), respectively, which are also used widely. Many systems encountered in technology or in nature are colloids. Some examples are given in [table C2.6.2](#).

Table C2.6.1 Classification of colloidal systems.

Disperse phase	Dispersion medium		
	Gas	Liquid	Solid
Gas	—	Foam	Solid foam
Liquid	Aerosol	Emulsion	Solid emulsion
Solid	Aerosol	Suspension	Solid dispersion

Table C2.6.2 Some practical examples of colloidal systems.

Aerosols	Inks
Agrochemicals	Milk
Blood	Paints
Carbon black	Pastes
Cosmetics	Polymer solutions
Drilling muds	Protein solutions
Fog	Soils
Ice-cream	Viruses

C2.6.1.2 SCOPE

In practice, e.g., in nature or in formulated products, colloidal suspensions (also denoted sols or dispersions) tend to be complex systems, consisting of many components that are often not very well defined, in terms of particle size for instance. Much progress has been made in the understanding of colloidal suspensions by studying well defined model systems, which allow for a quantitative modelling of their behaviour. Such systems will be discussed here.

Although the remainder of this contribution will discuss suspensions only, much of the theory and experimental approaches are applicable to emulsions as well (see [2] for a review). Some other colloidal systems are treated elsewhere in this volume. Polymer solutions are an important class—see [section C2.1](#). For surfactant micelles, see [section C2.3](#). The special properties of certain particles at the lower end of the colloidal size range are discussed in [section C2.17](#).

C2.6.1.3 COLLOIDS AS ATOMS

In addition to their practical importance, colloidal suspensions have received much attention from chemists and physicists alike. This is an interesting research area in its own right, and it is an important aspect of what is referred to as soft condensed matter physics. This contribution is written from such a perspective, and although a balanced account is aimed for, it is inevitably biased by the author's research interests. References to the original literature are included, but within the scope of this contribution only a fraction of the vast amount of literature on colloidal suspensions can be mentioned.

Colloidal particles can be seen as large, model 'atoms'. In what follows we assume that particles with a typical radius $a = 100$ nm are studied, about 10^3 times as large as atoms. Usually, the solvent is considered to be a homogeneous medium, characterized by bulk properties such as the density ρ and dielectric constant ϵ . A full statistical mechanical description of the system would involve all colloid and solvent degrees of freedom, which tend to be intractable. Instead, the potential of mean force, V , is used, in which the interactions between colloidal particles are averaged over

all solvent degrees of freedom [3, 4]. Usually, V is written as a sum of pair potentials. Its equivalent for an atomic system is the potential energy. Analogously, the osmotic pressure Π replaces the pressure p . As a consequence of

this colloid–atom analogy, for instance, colloidal suspensions at low concentrations obey van 't Hoff's law, $\Pi = nkT$, the equivalent of the ideal gas law. At higher concentrations, colloids can display a similar phase behaviour as simple liquids, including colloidal gas, liquid, and crystal phases, that differ in the arrangement of the particles within the solvent.

Model colloids have a number of properties that make them experimentally convenient and interesting systems to study. For instance, the timescale for 'structural relaxation' of a colloidal fluid can be estimated as the time for a particle to diffuse a distance equal to its radius,

$$t_R \approx \frac{a^2}{D}$$

where the Stokes diffusion coefficient of a sphere in a liquid of viscosity η is given by

$$D = \frac{kT}{6\pi\eta a}. \quad (\text{C2.6.1})$$

This typically yields t_R of order 0.01 s.

Due to the particle size, a colloidal crystal is much weaker than a normal solid material—the elastic moduli are proportional to the number density n , and therefore a colloidal solid would be about 10^9 times weaker than its atomic equivalent. The weakness of a colloidal solid means that a crystal can be broken up easily, by shaking the sample, for instance. The slow structural relaxation means that non-equilibrium behaviour is experimentally accessible, such as crystallization kinetics and glass or gel formation. Indeed, many colloidal systems probably never reach thermodynamic equilibrium on the timescale of experiments. Furthermore, as shown below, the interaction potential between colloidal particles can be tuned by varying the surface chemistry of the particles and the solvent conditions.

In the theory of the liquid state, the hard-sphere model plays an important role. For hard spheres, the pair interaction potential $V(r) = \infty$ for $r < d$, where d is the particle diameter, whereas $V(r) = 0$ for $r > d$. The structure of a simple fluid, such as argon, is very similar to that of a hard-sphere fluid. Hard-sphere atoms do, of course, not exist. Certain model colloids, however, come very close to hard-sphere behaviour. These systems have been studied in much detail and some results will be quoted below.

C2.6.1.4 OUTLINE

The remainder of this contribution is organized as follows. In [section C2.6.2](#), some well studied colloidal model systems are introduced. Methods for characterizing colloidal suspensions are presented in [section C2.6.3](#). An essential starting point for understanding the behaviour of colloids is a description of the interactions between particles. Various factors contributing to these are discussed in [section C2.6.4](#). Following on from this, theories of colloid stability and of the kinetics of aggregation are presented in [section C2.6.5](#). Finally, [section C2.6.6](#) is devoted to the phase behaviour of concentrated suspensions.

Encyclopedia of Chemical Physics and Physical Chemistry

C2.6.2 MODEL COLLOIDS

A huge variety of model colloids have been studied. In this section we will highlight a few of these, of particular interest to the discussion of concentrated suspensions in [section C2.6.6](#).

C2.6.2.1 POLYDISPERSITY

Even when carefully prepared, model colloids are almost never perfectly monodisperse. The spread in particle sizes, or polydispersity, is usually expressed as the relative width of the size distribution,

$$\sigma = \frac{s_a}{\bar{a}}$$

where s_a denotes the standard deviation of a . So-called monodisperse model systems tend to have polydispersities ranging from about $\sigma = 0.01$ to about 0.2. Suspensions encountered in practice tend to be much more polydisperse than that. When performing accurate quantitative experiments, the polydispersity needs to be taken into account. In some cases, such as in the formation of colloidal crystals (see [section C2.6.6](#)), the qualitative behaviour may also depend sensitively on the polydispersity.

C2.6.2.2 INORGANIC PARTICLES

Traditionally, most model studies were carried out using inorganic colloids, for instance gold and sulphur sols. A variety of particle types and shapes are provided by hydrous metal oxides [5] and silica [6]. Inorganic model suspensions are usually made using a nucleation and growth process. By controlling the nucleation step, monodisperse suspensions can be obtained. For instance, monodisperse silica spheres can be obtained by hydrolyzing alkoxysilanes in alcoholic solution [7]. Alternatively, seed particles can be prepared using microemulsions and then grown to the required size, resulting in very monodisperse suspensions (see [figure C2.6.1](#)). Near hard-sphere silica suspensions can then be obtained by coating the particles with a chemically grafted polymer layer and suspending them in organic solvents [8].

Encyclopedia of Chemical Physics and Physical Chemistry

-5-

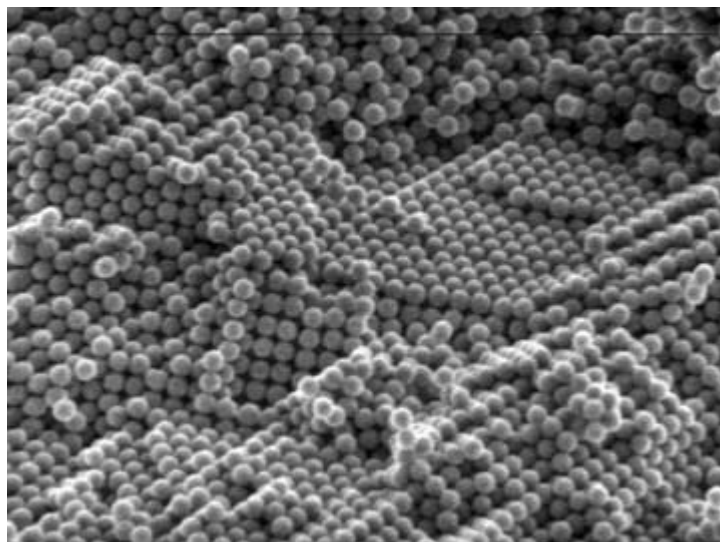


Figure C2.6.1. SEM image of silica spheres of radius $a = 75$ nm and polydispersity $\sigma < 0.01$ (courtesy of Professor A van Blaaderen)

C2.6.2.3 POLYMER LATTICES

An important step in the progress of colloid science was the development of monodisperse polymer latex suspensions in the 1950s. These are prepared by emulsion polymerization, which is nowadays also carried out industrially on a large scale for many different polymers. Perhaps the best-studied colloidal model system is that of polystyrene (PS) latex [9]. This is prepared with a hydrophilic group (such as sulphate) at the end of each molecule. In water this produces well defined spheres with a number of end groups at the surface, which (partly) ionize to

produce charged particles. In aqueous suspensions, near hard-sphere behaviour can be obtained by adding sufficient salt to screen the electrostatic repulsions (see [section C2.6.4](#)).

Another model system consists of polymethylmethacrylate (PMMA) latex, stabilized in organic solvents by a 'comb' polymer, consisting of a PMMA backbone with poly-12-hydroxystearic acid (PHSA) chains attached to it [10]. The PHSA chains form a steric stabilization layer at the surface (see [section C2.6.4](#)). Such particles can approach the hard-sphere model very well [11].

C2.6.2.4 PARTICLES WITH UNUSUAL PROPERTIES

In addition to the 'standard' model systems described above, more exotic particles have been prepared with certain unusual properties, of which we will mention a few. For instance, using seeded growth techniques, particles have been developed with a silica shell which surrounds a core of a different composition, such as particles with magnetic [12], fluorescent [13] or gold cores [14]. Another example is that of spheres of polytetrafluoroethylene (PTFE), which are optically anisotropic because the core is crystalline [15].

A different class, in between polymer lattices and polymer solutions, is that of microgels, consisting of weakly crosslinked polymer networks. Just as for polymer solutions, small changes in the solvency conditions may have large

Encyclopedia of Chemical Physics and Physical Chemistry

-6-

effects on their behaviour. They tend to undergo a swelling–deswelling transition, which for sufficiently weakly crosslinked particles may result in a particle size change by a factor of 5 [16].

C2.6.2.5 NON-SPHERICAL COLLOIDS

Although the majority of studies on model colloids involve (quasi-) spherical particles, there is a growing interest in the properties of non-spherical colloids. These tend to be either rod-like or plate-like.

One model for rod-like colloids is the tobacco mosaic virus (TMV), which consists of rods of diameter D about 18 nm and length L of 300 nm [17, 18]. These colloids have the advantage of being quite monodisperse, but are hard to obtain in large amounts. The *fd* virus gives longer, semi-flexible rods ($L = 880$ nm, $D = 9$ nm) [18, 19]. Inorganic boehmite rods have also been prepared successfully [20].

The major class of plate-like colloids is that of clay suspensions [21]. Many of these swell in water to give a stack of parallel, thin sheets, stabilized by electrical charges. Natural clays tend to be quite polydisperse. The synthetic clay laponite is comparatively well defined, consisting of discs of about 1 nm in thickness and 25 nm in diameter. It has been used in a number of studies (e.g. [22]).

C2.6.2.6 PURIFICATION

After preparation, colloidal suspensions usually need to undergo purification procedures before detailed studies can be carried out. A common technique for charged particles (typically in aqueous suspension) is dialysis, to deal with ionic impurities and small solutes. More extensive deionization can be achieved using ion exchange resins.

Another standard method is to use a (high-speed) centrifuge to sediment the colloids, replace the supernatant and redisperse the particles. Provided the particles are well stabilized in the solvent, this allows for a rigorous purification. Larger objects, such as particle aggregates, can be fractionated off because they settle first. A third method is (ultra)filtration, whereby larger impurities can be retained, particularly using membrane filters with accurately defined pore sizes.

C2.6.3 PROPERTIES AND CHARACTERIZATION METHODS

Even when well defined model systems are used, colloids are rather complex, when compared with pure molecular compounds, for instance. As a result, one often has to resort to a wide range of characterization techniques to obtain a sufficiently comprehensive description of a sample being studied. This section lists some of the most common techniques used for studying colloidal suspensions. Some of these techniques are discussed in detail elsewhere in this volume and will only be mentioned in passing. A few techniques that are relevant more specifically for colloids are introduced very briefly here, and a few advanced techniques are highlighted.

Although the behaviour of colloidal suspensions does in general depend on temperature, a more important control parameter in practice tends to be the particle concentration, often expressed as the volume fraction ϕ . In fact, for hard-sphere suspensions the phase behaviour is determined by ϕ only. For spherical particles $\phi = \frac{4}{3}\pi a^3 n$.

Encyclopedia of Chemical Physics and Physical Chemistry

-7-

In practice, there are various ways by which ϕ can be determined for a given sample, and the results may be (slightly) different. In particular, for sterically stabilized particles, the effective hard-sphere volume fraction will be different from the value obtained from the total solid content.

C2.6.3.1 OBSERVATION

Straightforward, direct observation is generally very useful to assess suspension stability, phase separations, etc. Light microscopy (see [section B1.19](#)) can, under some conditions, image particles directly. Often, however, this is prevented by sample turbidity or insufficient resolution. An enhanced resolution can be obtained by preparing core-shell particles with a fluorescent core. Even when the particles are touching, the cores can still be resolved using confocal scanning laser microscopy (CSLM), allowing for the determination of three-dimensional structures in dense suspensions [23] (see [figure C2.6.2](#)).

Electron microscopy (see [section B1.18](#)) is very valuable in characterizing particles (see, for instance, [figure C2.6.1](#)). The suspension structure is, of course, not represented well because of the vacuum conditions in the microscope. This can be overcome using environmental SEM [24].

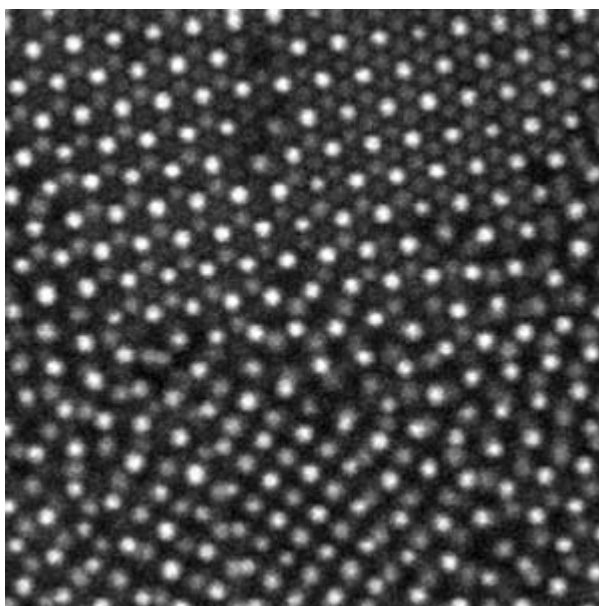


Figure C2.6.2. CSLM image of near hard-sphere silica particles of diameter $d = 1050$ nm with a fluorescent core of diameter 400 nm, showing fcc stacking (top), hcp stacking (bottom middle) and amorphous areas (image size $16.3 \mu\text{m} \times 16.3 \mu\text{m}$, courtesy of Professor A van Blaaderen)

C2.6.3.2 GENERAL PROPERTIES

Because model colloids tend to have a rather well defined chemical composition, elemental analysis can be used to obtain detailed information, such as the grafted amount of polymer in the case of sterically stabilized particles. More details about the chemical structure can be obtained using NMR techniques ([section B1.13](#)). In addition, NMR

Encyclopedia of Chemical Physics and Physical Chemistry

-8-

relaxation techniques can be used to quantify properties such as polymer adsorption to the particle surface (see [section B1.13](#)). Often further details, such as particle density and refractive index, need to be determined as well.

C2.6.3.3 SCATTERING TECHNIQUES

Because colloidal particles typically have a size similar to that of the wavelength of light, light scattering and diffraction are very useful in characterizing their suspensions. In particular, for small particles, x-ray and neutron scattering are employed as well (see [section B1.9](#)). Photon correlation spectroscopy (PCS) is a standard technique for particle size determination. Static light scattering is also used for this, and to characterize the structure of suspensions. Although, in an indirect way, both techniques can yield information on particle interactions as well.

C2.6.3.4 PARTICLE INTERACTIONS

The interactions between colloidal particles (see [section C2.6.4](#)) are central to the understanding of suspension behaviour. Although most work has had to rely on rather indirect ways to characterize these interactions, novel techniques are emerging that access these interactions more directly.

Particles can be manipulated in suspension using strongly focused laser beams ('optical tweezers') [[25](#)] or magnetic fields [[26](#)] and by collecting statistics on the particle movements using video microscopy, information on the particle interactions can be obtained.

Surfaces can be characterized using scanning probe microscopies (see [section B1.19](#)). In addition, by attaching a colloidal particle to the tip of an atomic force microscope, colloidal interactions can be probed as well [[27](#)]. Interactions between surfaces can be studied using the surface force apparatus (see [section B1.20](#)). This also helps one to understand the interactions between colloidal particles.

C2.6.3.5 RHEOLOGY

The study of the rheology, or flow behaviour, of suspensions is an important method for characterizing particle interactions. Controlling the rheological properties of suspensions is also crucial in practice, where during processing and in the final application the flow behaviour usually has to be within fairly narrow specifications. This field is only touched upon here; for more details consult [[28](#), [29](#), [30](#) and [31](#)], or the general colloid science texts [[32](#), [33](#) and [34](#)]. For Newtonian fluids, the shear stress (force/area) τ is proportional to the shear rate (velocity gradient), $\dot{\gamma}$,

$$\tau = \eta \dot{\gamma}. \quad (\text{C2.6.2})$$

For dilute dispersions of hard spheres, Einstein's viscosity equation predicts

$$\frac{\eta}{\eta_0} = 1 + 2.5\phi + \dots$$

where η_0 denotes the solvent viscosity. For concentrated hard-sphere suspensions, experimental data can be

using the Krieger–Dougherty equation,

$$\frac{\eta}{\eta_0} = [1 - (\phi/\phi_{\max})]^{-[\eta]\phi_{\max}}$$

where $[\eta]$ is called the intrinsic viscosity and the viscosity diverges at a concentration ϕ_{\max} . At low shear rate, $\phi_{\max} = 0.63$ (similar to the random close packing density; see [section C2.6.6.2](#)) and $[\eta] = 3.13$. The Krieger–Dougherty equation is also widely used to correlate data for other types of suspensions.

Colloidal dispersions often display non-Newtonian behaviour, where the proportionality in [equation \(C2.6.2\)](#) does not hold. This is particularly important for concentrated dispersions, which tend to be used in practice. [Equation \(C2.6.2\)](#) can be used to define an apparent viscosity, η_{app} , at a given shear rate. If η_{app} decreases with increasing shear rate, the dispersion is called shear thinning (pseudoplastic); if it increases, this is known as shear thickening (dilatant). The latter behaviour is typical of concentrated suspensions. If a finite shear stress has to be applied before the suspension begins to flow, this is known as the yield stress. The apparent viscosity may also change as a function of time, upon application of a fixed shear rate, related to the formation or breakup of particle networks. Thixotropic dispersions show a decrease in η_{app} with time, whereas an increase with time is called rheopexy.

C2.6.3.6 SEDIMENTATION AND DIFFUSION

In most colloidal suspensions the particles have a tendency to sediment. At infinite dilution, spherical particles with a density difference $\Delta\rho$ with the solvent will move at the Stokes velocity

$$U_0 = \frac{2a^2\Delta\rho g}{9\eta}.$$

At finite concentration, the settling rate is influenced by hydrodynamic interactions between the particles. For purely repulsive particle interactions, settling is hindered. Attractive interactions encourage particles to settle as a group, which increases the settling rate. For hard spheres, the first-order correction to the Stokes settling rate is given by [\[33\]](#)

$$\frac{U}{U_0} = 1 - 6.55\phi.$$

The tendency for particles to settle is opposed by their Brownian diffusion. The number density distribution of particles as a function of height z will tend to an equilibrium distribution. At low concentration, where van 't Hoff's law applies, the barometric height distribution is given by

$$n(z) \propto \exp\left(-\frac{mgz}{kT}\right)$$

where m is the buoyant mass of the particles and g the gravitational acceleration. Perrin has already used this to determine Avogadro's number [\[35\]](#).

Given the a^2 size dependence of the settling rate, sedimentation can be used for particle size analysis. Indeed, a quick

impression of particle size (or degree of aggregation of primary particles) is often obtained from the settling behaviour of dilute suspensions. A quantitative analysis of particle sizes can be carried out using the analytic ultracentrifuge (see, for instance, [34]).

In practice, sedimentation is an important property of colloidal suspensions. In formulated products, sedimentation tends to be a problem and some products are shipped in the form of weak gels, to prevent settling. On the other hand, in applications such as water clarification, a rapid sedimentation of impurities is desirable.

C2.6.3.7 ELECTROKINETIC PHENOMENA

In particular, in polar solvents, the surface of a colloidal particle tends to be charged. As will be discussed in [section C2.6.4.2](#), this has a large influence on particle interactions. A few key concepts are introduced here. For more details, see [32] (ch 13), [33] (ch 7), [36] (ch 4) and [34] (ch 12). The presence of these surface charges gives rise to a number of electrokinetic phenomena, in particular electrophoresis.

In electrophoresis, the motion of charged colloidal particles under the influence of an electric field is studied. For spherical particles, we can write

$$v = \mu_E E$$

where v is the particle velocity, E the electric field and μ_E is called the electrophoretic mobility. For low surface electrostatic potentials, it is given by Henry's equation,

$$\mu_E = 2 \frac{\epsilon \epsilon_0}{3\eta} \zeta f(\kappa a)$$

where κ is the inverse screening length (see [section C2.6.2](#)). $f(\kappa a)$ increases from 1 at $\kappa a = 0$ to 1.5 at $\kappa a \rightarrow \infty$. The zeta-potential ζ represents the electrostatic potential near the point where the diffuse double layer ([section C2.6.4.2](#)) starts.

Related phenomena are electro-osmosis, where a liquid flows past a surface under the influence of an electric field and the reverse effect, the streaming potential due to the flow of a liquid past a charged surface.

C2.6.4 PARTICLE INTERACTIONS

Many properties of colloidal suspensions, such as their stability, rheology, and phase behaviour, are closely related to the interactions between the suspended particles. The background of the most important contributing factors to these interactions is discussed in this section.

C2.6.4.1 VAN DER WAALS INTERACTIONS

Between any two atoms or molecules, van der Waals (or dispersion) forces act because of interactions between the fluctuating electromagnetic fields resulting from their polarizabilities (see [section A1.5](#), and, for instance,

Encyclopedia of Chemical Physics and Physical Chemistry

[37]). Similarly, van der Waals forces operate between any two colloidal particles in suspension. In the 1930s, predictions for these interactions were obtained from the pairwise addition of molecular interactions between two particles [38]. The interaction between two identical spheres is given by

$$V_{\text{vdW}}(r) = -\frac{A}{6} \left[\frac{2a^2}{r^2 - 4a^2} + \frac{2a^2}{r^2} + \ln \left(1 - \frac{4a^2}{r^2} \right) \right] \quad (\text{C2.6.3})$$

where A is the Hamaker constant, which typically is of order 10^{-20} J. Similar equations are obtained for other geometries [37, 39].

At large separation r , equation (C2.6.3) decays as $V_{\text{vdW}}(r) \propto r^{-6}$, just as the van der Waals interactions between molecules do. However, at large separation, say $r > 100$ nm, relativistic effects have to be taken into account and the so-called retarded van der Waals interactions decay as r^{-7} .

At short separations, equation (C2.6.3) tends to

$$V_{\text{vdW}}(r) \approx \frac{A a}{12 H} \quad (\text{C2.6.4})$$

where $H = r - 2a$ is the surface separation. At contact, equation (C2.6.4) would predict an infinitely strong attraction. In reality, this is prevented by steep Born repulsions at short distances. Nevertheless, the van der Waals interactions tend to create a deep interaction minimum near $r = 2a$, strong enough to result in aggregation of suspended particles, unless a stabilizing mechanism such as electrostatic interactions or steric stabilization is provided (see section C2.6.2 and section C2.6.3).

The Hamaker constant can be evaluated accurately using the continuum theory, developed by Lifshitz and coworkers [40]. A key property in this theory is the frequency dependence of the dielectric permittivity, $\epsilon(\omega)$. If this spectrum were the same for particles and solvent, then $A = 0$. Since the refractive index n is also related to $\epsilon(\omega)$, the van der Waals forces tend to be very weak when the particles and solvent have similar refractive indices. A few examples of values for A for interactions across vacuum and across water, obtained using the continuum theory, are given in table C2.6.3.

Encyclopedia of Chemical Physics and Physical Chemistry

-12-

Table C2.6.3 Hamaker constants A (10^{-20} J) (from [120]).

Material	Medium	
	Vacuum	Water
Water	3.7	—
Pentane	3.8	0.34
PS	6.6	0.95
PMMA	7.1	1.05
Fused silica	6.6	0.85

More generally, approximate relations can be used to estimate the Hamaker constant for particles 1 and 2, suspended in a medium 3, such as

$$A_{132} \approx (\sqrt{A_{11}} - \sqrt{A_{33}})(\sqrt{A_{22}} - \sqrt{A_{33}}) \quad (\text{C2.6.5})$$

where A_{ii} is the Hamaker constant for interaction of material i across a vacuum. Although the validity of such equations is limited, one interesting aspect of [equation C2.6.5](#) is that van der Waals interactions between two suspended particles can be repulsive, when the suspending medium has a Hamaker constant intermediate between that of the two particles.

C2.6.4.2 ELECTROSTATIC INTERACTIONS

Particularly in polar solvents, electrostatic charges usually have an important contribution to the particle interactions. We will first discuss the ion distribution near a single surface, and then the effect on interactions between two colloidal particles.

THE ELECTRICAL DOUBLE LAYER

Here a few core equations are presented from the simplest theory for the electric double layer: the Gouy–Chapman theory [41]. We consider a solution of ions of valency z_+ and z_- in a medium with dielectric constant ϵ . The ions are represented by point charges (they have no size) and it is assumed that the ions undergo rapid Brownian motion, and their average spatial distribution may be obtained through Boltzmann's distribution from the electrostatic potential. For simplicity, we restrict ourselves to symmetric electrolytes, $z = z_+ = z_-$. We write the electrostatic potential ϕ in dimensionless form as

$$\Phi = \frac{ze\phi}{kT}$$

Encyclopedia of Chemical Physics and Physical Chemistry

-13-

where e is the elementary charge. A combination of Poisson's law and the Boltzmann distribution gives the Poisson–Boltzmann equation, which here takes the form

$$\nabla^2 \Phi = \kappa^2 \sinh \Phi \quad (\text{C2.6.6})$$

and κ is given as a function of the bulk electrolyte concentration c_0 by

$$\kappa = \sqrt{\frac{2e^2 N_A c_0 z^2}{\epsilon \epsilon_0 kT}}. \quad (\text{C2.6.7})$$

This is an inverse length; κ^{-1} is known as the Debye screening length (or double layer thickness). As demonstrated below, it gives the length scale on which the ion distribution near a surface decays to the bulk value. [Table C2.6.4](#) gives a few numerical examples.

Table C2.6.4 Debye screening length κ^{-1} for aqueous solutions of a 1-1 electrolyte at 298 K (equation (C2.6.7)).

c_0 (mol dm⁻³) κ^{-1} (nm)

10⁻⁵ 97

10^{-3}	9.7
0.1	0.97

Surfaces in polar solvents and particularly in water tend to be charged, through dissociation of surface groups or by adsorption of ions, resulting in a charge density σ . Near a flat surface, ϕ only depends on the distance x from the surface. The solution of equation (C2.6.6) then is

$$\Phi = 2 \ln \left(\frac{1 + \gamma e^{-\kappa x}}{1 - \gamma e^{-\kappa x}} \right) \quad (\text{C2.6.8})$$

where $\gamma = \tanh(\frac{1}{4}\Phi_s)$ and the subscript s denotes values at the surface. The corresponding surface charge density is given by

$$\sigma = 2\sqrt{2\epsilon\epsilon_0 kT N_A c_0} \sinh \frac{1}{2} \Phi_s. \quad (\text{C2.6.9})$$

From Φ , the charge distribution can then be calculated using Boltzmann's distribution. An example of this is shown in

Encyclopedia of Chemical Physics and Physical Chemistry

-14-

figure C2.6.3, which plots the distribution of counterions (of opposite sign to the charged surface) and co-ions (of the same sign as the surface). More detailed descriptions of the ionic distribution take into account the non-uniform packing of ions and molecules close to the surface. A significant potential drop may occur across this so-called Stern layer adjacent to the surface. The potential ϕ_d outside the Stern layer then enters the description of the diffuse double layer. In practice, ϕ_d is usually equated to the ζ -potential, discussed in [section C2.6.3.7](#).

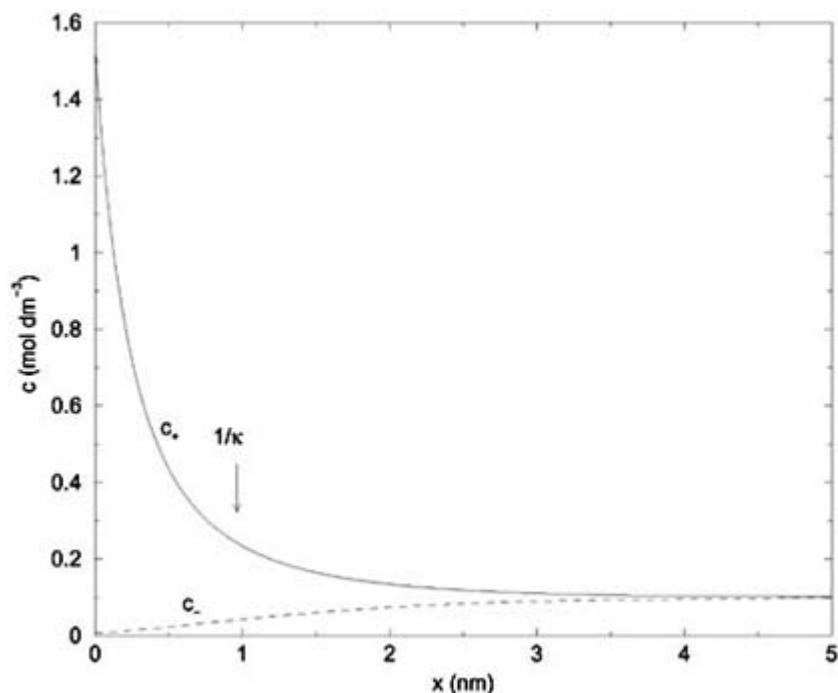


Figure C2.6.3. Distribution of positive and negative ions (c_+ , c_-) near a flat surface in water at 298 K ([equation \(C2.6.8\)](#)). Parameters: $z = 1$, $\phi_s = 70$ mV, $c_0 = 0.1$ mol dm⁻³, corresponding to $\sigma = -0.068$ C m⁻² ([equation \(C2.6.9\)](#)). The double layer thickness $\kappa^{-1} = 0.96$ nm is indicated.

DOUBLE LAYER INTERACTIONS

When describing the interactions between two charged flat plates in an electrolyte solution, [equation \(C2.6.6\)](#) cannot be solved analytically, so in the general case a numerical solution will have to be used. Several equations are available, however, to describe the behaviour in a number of limiting cases (see [41] for a detailed discussion). Here we present two limiting cases for the interactions between two charged spheres, surrounded by their counterions and added electrolyte, which will be referred to in further sections. This pair interaction V_R is always repulsive in the theory discussed here.

The first case is relevant in the discussion of colloid stability of [section C2.6.5](#). It uses the potential around a single sphere in the case of a double layer that is thin compared to the particle, $\kappa a \gg 1$. Furthermore, it is assumed that the surface separation is fairly large, such that $\exp(-\kappa H) \ll 1$, so the potential between two spheres can be calculated from the sum of single-sphere potentials. Under these conditions, V_R is approximated by [42]:

Encyclopedia of Chemical Physics and Physical Chemistry

-15-

$$V_R = \frac{32\pi\epsilon\epsilon_0ak^2T^2\gamma^2}{z^2e^2}e^{-\kappa H}. \quad (\text{C2.6.10})$$

Again, κ^{-1} is the length scale on which the interaction decays.

In the second case, a thick double layer, $\kappa a \ll 1$ (low ionic strength), is assumed. When the surface potential is low, $\Phi_s \ll 1$, a reasonable approximation is given by

$$V_R = 4\pi\epsilon\epsilon_0a^2\phi_s^2e^{2\kappa a}\frac{e^{-\kappa r}}{r}. \quad (\text{C2.6.11})$$

This r dependence is also known as a Yukawa potential. This type of potential has been used to describe the behaviour of latex suspensions at low ionic strength.

More sophisticated approaches to describe double layer interactions have been developed more recently. Using cell models, the full Poisson-Boltzmann equation can be solved for ordered structures. The approach by Alexander *et al* shows how the effective colloidal particle charge saturates when the 'bare' particle charge is increased [43]. Using integral equation methods, the behaviour of the 'primitive model' has been studied, in which all the interactions between the colloidal macro-ions and the small ions are addressed (see, for instance, [44, 45]).

C2.6.4.3 INTERACTIONS DUE TO SOLUBLE POLYMERS

In many colloidal systems, both in practice and in model studies, soluble polymers are used to control the particle interactions and the suspension stability. Here we distinguish three scenarios: interactions between particles bearing a grafted polymer layer, forces due to the presence of non-adsorbing polymers in solution, and finally the interactions due to adsorbing polymer chains. Although these cases are discussed separately here, in practice more than one mechanism may be in operation for a given sample.

STERIC STABILIZATION

The first case concerns particles with polymer chains attached to their surfaces. This can be done using chemically (end-)grafted chains, as is often done in the study of model colloids. Alternatively, a block copolymer can be used, of which one of the blocks (the anchor group) adsorbs strongly to the particles. The polymer chains may vary from short alkane chains to high molecular weight polymers (see also [section C2.6.2](#)). The interactions between such

'hairy' colloidal particles depend on many parameters, such as the nature of the polymer and the solvent, the molecular weight and grafting density. For theoretical approaches to describe the resulting behaviour, see [33, 46, 47]. Here a few general observations are made.

For so-called steric stabilization to be effective, the polymer needs to be attached to the particles at a sufficiently high surface coverage and a good solvent for the polymer needs to be used. Under such conditions, a fairly dense polymer brush with thickness L will be present around the particles. When two particles approach, such that $r < d + 2L$, the polymer layers may be compressed from their equilibrium configuration, thus causing a repulsive interaction.

Encyclopedia of Chemical Physics and Physical Chemistry

-16-

Alternatively, the polymer layers may overlap, which increases the local polymer segment density, also resulting in a repulsive interaction. Particularly on close approach, $r < d + L$, a steep repulsion is predicted to occur. When a relatively low molecular weight polymer is used, the repulsive interactions are rather short-ranged (compared to the particle size) and the particles display near hard-sphere behaviour (e.g., [11]).

When the solvent quality is reduced, by changing the solvent composition or temperature, for instance, steric stabilization may not occur. On close approach, a repulsive interaction will still result, but for partial overlap of the polymer layers an attractive interaction may arise. For a number of systems, steric stabilization was found to fail, resulting in particle aggregation, at the θ temperature of the polymer [46] (at the θ temperature, the second osmotic virial coefficient of the polymer solution is zero; see also section C2.1). The attractions tend to be of short range compared to the particles themselves. The behaviour of such systems has been modelled using a narrow square well potential [48]. In the limit of a very narrow attraction range, Baxter's adhesive sphere ('sticky' sphere) potential [49] is obtained. Many authors have interpreted their observations by modelling the particle interactions using this potential.

NON-ADSORBING POLYMER

The second case involves non-adsorbing polymer chains in solution. It was realized by Asakura and Oosawa (AO) [50] and separately by Vrij [51] that these chains will give rise to an effective attraction between colloidal particles. This is known as depletion attraction (see figure C2.6.4). We will summarize the AO theory to explain this.

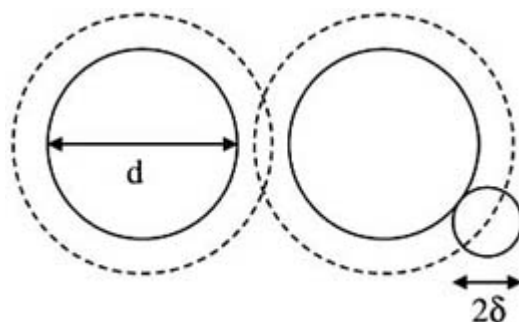


Figure C2.6.4. Graphical representation of the AO model. A depletion shell of thickness δ surrounds each particle.

The colloidal particles are represented by hard spheres with diameter d , and the polymer coils by spheres with radius δ . As a guide, δ is often taken to be equal to R_g , the radius of gyration of the polymer. Polymer molecules are considered not to have any interaction. A polymer 'sphere', however, cannot overlap with a colloidal particle (the colloidal particles and polymer molecules behave as hard spheres towards each other). This means that a polymer coil cannot enter a sphere with radius $a + \delta$, centred on a colloidal particle. In other words, there is a depletion shell of thickness δ around each particle. If two particles approach each other, their depletion zones will overlap when $r < d + 2\delta$. This gives rise to an osmotic pressure imbalance, which results in an effective attraction between the particles, given by

$$V_{\text{dep}} = -\Pi_p V_{\text{overlap}} \quad (d < r < d + 2\delta) \quad (\text{C2.6.12})$$

where Π_p is the polymer osmotic pressure and V_{overlap} is the overlap volume of the excluded spheres of the two particles. It is given by

Encyclopedia of Chemical Physics and Physical Chemistry

-17-

$$V_{\text{overlap}} = \left(1 - \frac{3r}{2d(1+\xi)} + \frac{1}{2} \left[\frac{r}{d(1+\xi)} \right]^3\right) \frac{\pi}{6} d^3 (1+\xi)^3$$

where

$$\xi = \delta/a$$

is the polymer/colloid size ratio. Figure C2.6.5 shows examples of the shape of this potential. A few points are worth noting about this potential. First, although the net effect is an attraction between the colloids, this is the result of purely repulsive interactions. Second, this interaction can easily be tuned experimentally: the range (2δ) is set by the polymer size (molecular weight), whereas the strength can be adjusted by the polymer concentration (at low concentration, van 't Hoff's law can again be applied: $\Pi_p = n_{\text{pol}}kT$). Some results obtained in this way will be discussed in section C2.6.6.4. For a more advanced discussion of depletion interactions, see [33, 46, 47 and 52]. The depletion picture also applies to other systems, such as mixtures of colloidal particles. However, whereas neglecting the interactions between polymer molecules may be reasonable, this cannot be done in the general case.

ADSORBING POLYMER

Finally, we briefly mention interactions due to adsorbing polymers. Block copolymers, with one block strongly adsorbing to the particles, have already been mentioned above. Here, we focus on homopolymers that adsorb moderately strongly to the particles. If this can be done such that a high surface coverage is achieved, the adsorbed polymer layer may again produce a steric stabilization between the particles.

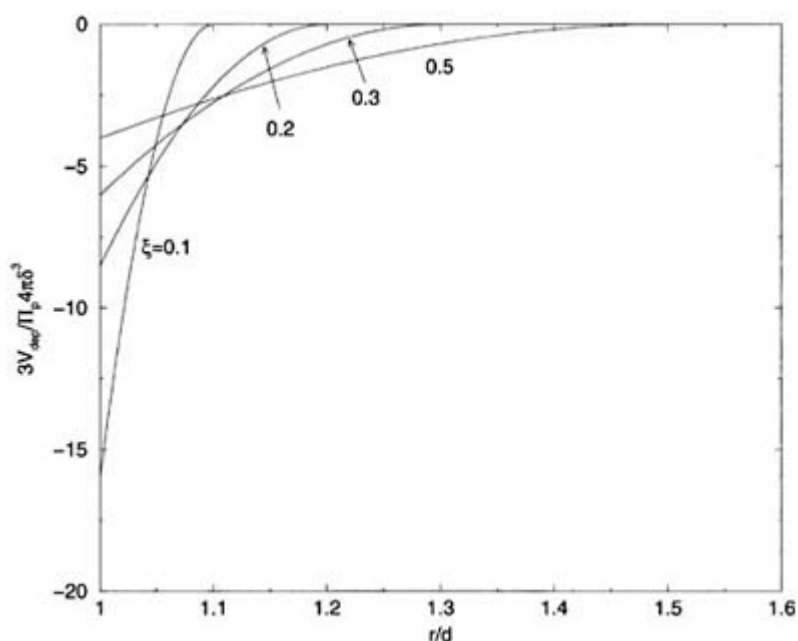


Figure C2.6.5. Examples of the AO potential, equation (C2.6.12). The values of ξ are indicated next to the curves. The hard-sphere repulsion at $r = d$ has not been drawn.

At lower surface coverage, however, the possibility exists that one polymer chain may attach itself to two particles. If the adsorption is strong enough, this results in an aggregation of the particles, known as bridging flocculation [33, 46, and 47].

C2.6.5 COLLOID STABILITY AND AGGREGATION

In this section we focus on the theory of stability of charged colloids. In section C2.6.5.1 it is shown how particles can be made to aggregate by adding sufficient electrolyte. The associated aggregation kinetics are discussed in [section C2.6.5.2](#), and the structure of the aggregates in [section C2.6.5.3](#). For more details, see the recent reviews [53, 54 and 55], or the colloid science textbooks [33, 39].

C2.6.5.1 CHARGED PARTICLES

In suspensions containing no soluble polymer, the van der Waals forces and electrostatic interactions are the main factors in controlling the particle interactions. The van der Waals interactions are often strong enough to cause an irreversible aggregation of the particles, unless a stabilizing mechanism is present. In polar solvents, particularly in water, the particle surfaces tend to be charged. At low salt concentration, the resulting double layer repulsions can be strong enough to prevent aggregation.

Here we consider the total interaction between two charged particles in suspension, surrounded by their counterions and added electrolyte. This is the celebrated DLVO theory, derived independently by Derjaguin and Landau and by Verwey and Overbeek [41]. By combining the van der Waals interaction ([equation \(C2.6.4\)](#)) with the repulsion due to the electric double layers ([equation \(C2.6.10\)](#)), we obtain

$$V_{DLVO} = V_{vdW} + V_R. \quad (C2.6.13)$$

For the repulsions, the approximate result of [equation \(C2.6.10\)](#) is appropriate for the typical conditions relevant for investigating the particle stability ($\kappa a \gg 1$, $\kappa H \approx 1$) [33, 41]. Examples of the shape of this potential are shown in [figure C2.6.6](#). At short range, a deep attractive minimum (called the primary minimum) is found, due to the van der Waals attractions. At slightly larger separation, a repulsive maximum is present at low salt concentration. Once two particles have reached the primary minimum this tends to be irreversible. They may, however, be kinetically stabilized against reaching this minimum when the repulsive maximum is sufficiently high, $V_{\max} \gg kT$.

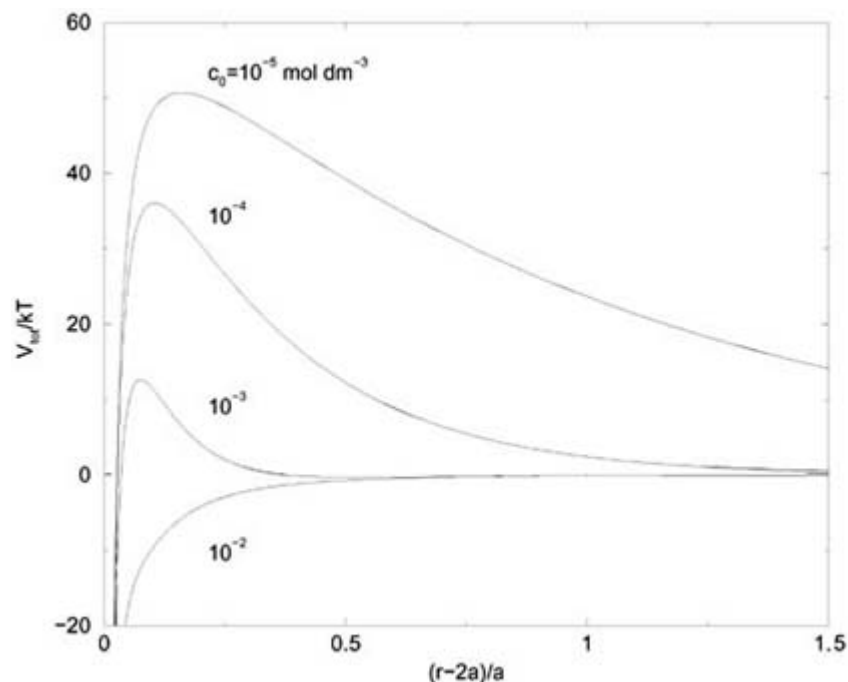


Figure C2.6.6. DLVO potential for gold spheres with $A/kT=25$, $a=100$ nm and $\Phi_s = 1$ in water, at a range of concentrations of 1-1 electrolyte (equations (C2.6.3), (C2.6.10) and (C2.6.13)).

As can be seen in figure C2.6.6 the repulsive maximum is reduced at high ionic strength. A rapid and irreversible aggregation into the primary minimum (also referred to as coagulation) is expected to occur when the maximum has become sufficiently small. Over a narrow range of electrolyte concentrations, a transition occurs from kinetic stabilization to rapid aggregation, when the maximum is about zero. By solving equation (C2.6.13) for $V_{DLVO} = 0$ and $dV_{DLVO}/dH = 0$, we obtain $\kappa H = 1$, and the corresponding electrolyte concentration, known as the critical coagulation concentration (c.c.c.), is given by

$$c_{ccc} = \frac{49.6\gamma^4}{N_A z^6 l_b^3} \left(\frac{kT}{A} \right)^2 \quad (\text{C2.6.14})$$

where $l_b = e^2/4\pi\epsilon\epsilon_0 kT$ is the Bjerrum length. For the conditions used in figure C2.6.6 c_{ccc} is calculated to be 1.3 mmol dm⁻³.

In equation (C2.6.14) it can be seen that the required salt concentration depends strongly on the valency of the ions z . At high surface potential $\gamma \rightarrow 1$, and $c_{ccc} \propto z^{-6}$. This had been observed experimentally and is known as the Schulze–Hardy rule. This result was one of the early successes of DLVO theory. At low surface potential, however, the valency dependence is less pronounced, $c_{ccc} \propto z^{-6}$. In reality, the behaviour of higher valency ions can be rather complicated, for instance, they may adsorb to the particle surface and even change the sign of the surface charge.

Encyclopedia of Chemical Physics and Physical Chemistry

At larger particle separation, a second minimum may occur in the potential energy. In many cases, this minimum is too shallow to be of much significance. For larger particles, however, the minimum may become of order kT . Aggregation in this minimum is referred to as secondary minimum flocculation.

For a more complete understanding of colloid stability, we need to address the kinetics of aggregation. The theory discussed here was developed to describe coagulation of charged colloids, but it does apply to other cases as well. First, we consider the case of so-called rapid coagulation, which means that two particles will aggregate as soon as they meet (at high salt concentration, for instance). This was considered by von Smoluchowski [56]; here we follow [39, 57].

It is assumed that irreversible aggregation occurs on contact. The rate of coagulation is expressed as the aggregation flux J of particles towards a central particle. Using a steady-state approximation, the diffusive flux is derived to be

$$J = 16\pi D a n$$

where D is again the diffusion coefficient of a single particle (equation (C2.6.1)) and n is the initial monomer concentration. As aggregation proceeds, a distribution of aggregate sizes (monomers, dimers, trimers, etc) is established, which evolves in time. This is described by

$$n_i = \frac{n(t/t_p)^{i-1}}{(1+t/t_p)^{i+1}} \quad (\text{C2.6.15})$$

where n_i denotes the number density of i -mers at time t . The half-life time of aggregation t_p , after which the total number of aggregates has halved, is given as a function of the volume fraction ϕ by

$$t_p = \frac{\pi \eta a^3}{kT \phi} \quad (\text{C2.6.16})$$

In table C2.6.5, a few numerical examples for t_p are shown. Smaller colloids are found to aggregate much faster and stabilizing them is therefore more difficult. The validity of equation (C2.6.15) has been confirmed experimentally (e.g. [58]).

Table C2.6.5 Rapid coagulation half-life time for particles in water at $T=300$ K (equation (C2.6.16)).

a	$\phi = 10^{-5}$	$\phi = 0.1$
100 nm	76 s	8 ms
1 μm	21 h	8 s

Encyclopedia of Chemical Physics and Physical Chemistry

The second case concerns situations where not all particle encounters result in aggregation. This is known as slow coagulation. This was addressed first by Fuchs [59]; again we follow [39, 57].

In slow coagulation, particles have to diffuse over an energy barrier (see the previous section) in order to aggregate. As a result, not all Brownian particle encounters result in aggregation. This is expressed using the stability ratio W , defined as

$$W = \frac{J_0}{J} \quad (\text{C2.6.17})$$

where J_0 denotes the rapid coagulation rate. By solving the diffusion equations under steady-state conditions, it was found that

$$W = 2a \int_{2a}^{\infty} \frac{\exp(V_{\text{DLVO}}/kT)}{r^2} dr. \quad (\text{C2.6.18})$$

Because of the exponential term, W is mainly determined by the potential energy maximum V_{max} , and it can be approximated as [57]

$$W \approx \frac{1}{2\kappa a} \exp(V_{\text{max}}/kT). \quad (\text{C2.6.19})$$

A combination of equation (C2.6.13), equation (C2.6.14), equation (C2.6.15), equation (C2.6.16), equation (C2.6.17), equation (C2.6.18) and equation (C2.6.19) then allows us to estimate how low the electrolyte concentration needs to be to provide kinetic stability for a desired length of time. This theory successfully accounts for a number of observations on slowly aggregating systems, but two discrepancies are found (see, for instance, [33]). First, the observed dependence of stability ratio on salt concentration tends to be much weaker than predicted. Second, the variation of the stability ratio with particle size is not reproduced experimentally. Recently, however, it was reported that for model particles with a low surface charge, where the DLVO theory is expected to hold, the aggregation kinetics do agree with the theoretical predictions (see [60], and references therein).

C2.6.5.3 AGGREGATE STRUCTURE

Although the theories of colloid stability and aggregation kinetics were developed several decades ago, the actual structure of aggregates has only been studied more recently. To describe the structure, we start with the relationship between the size of an aggregate (linear dimension), expressed as its radius of gyration R_g and its mass m :

$$m \propto R_g^{d_f}.$$

For compact, homogeneous objects in three dimensions, $d_f = 3$. Colloidal aggregates, however, tend to be rather open, fractal structures, with $d_f < 3$. For a general introduction to fractals, see [section C3.6](#) and [61].

Encyclopedia of Chemical Physics and Physical Chemistry

-22-

First, we consider the case of rapid, irreversible aggregation. In the literature on fractal aggregates, this is known as diffusion limited cluster aggregation (DLCA), where particles (monomers) and also the aggregates diffuse and aggregate when any two of them meet. Computer simulations predicted rather open structures with $d_f \approx 1.8$ under these conditions [62, 63 and 64], and experiments have confirmed this (figure C2.6.7) [65]. Various methods exist for measuring d_f . For instance, scattering (light, x-ray) experiments yield d_f from the variation of the scattered intensity I with wavevector Q , as

$$I \propto Q^{-d_f}.$$

This relationship holds for wavevectors that probe the appropriate size range, $1/R_g < Q < 1/a$ (see [section B1.10](#)).

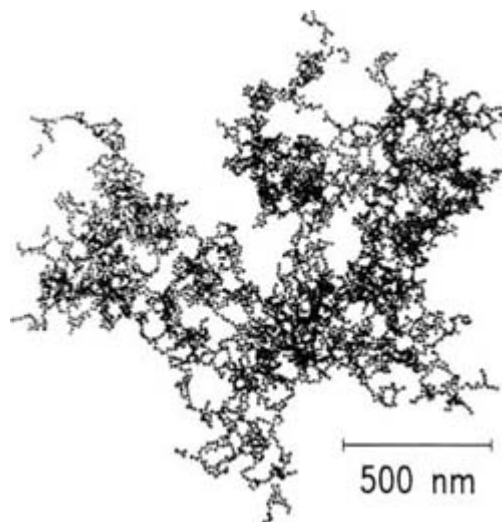


Figure C2.6.7. Fractal aggregate of gold particles with $a = 7.2 \pm 0.7$ nm, obtained under DLCA conditions, with $d_f = 1.74$ (reproduced with permission from [65]. Copyright 1984 Elsevier Science Publishers B.V).

More compact structures are obtained in the slow coagulation regime. Here aggregation is still irreversible but not every collision results in aggregation, and the clusters have more time to explore the available space. This is known as reaction limited cluster aggregation (RLCA). Here computer simulations predicted $d_f \approx 2.1$ [66], which was again confirmed experimentally [67]. The DLCA and RLCA regimes have been observed for a range of colloidal systems [64]. Only when the particle bonds are sufficiently weak that restructuring of a cluster can occur, are more dense clusters with $d_f \approx 3$ obtained (see, for instance, [54]).

Although this section has focused on the behaviour of charged particles, similar phenomena may be observed using sterically stabilized particles. As discussed in [section C2.6.4](#), these can also be given strong, short-ranged attractions, by changing the solvent quality or by adding non-adsorbing polymers. A similar aggregation behaviour to the charged spheres may then be observed [68].

Encyclopedia of Chemical Physics and Physical Chemistry

-23-

C2.6.6 BEHAVIOUR OF CONCENTRATED SUSPENSIONS

In the previous section, non-equilibrium behaviour was discussed, which is observed for particles with a deep minimum in the particle interactions at contact. In this final section, some examples of equilibrium phase behaviour in concentrated colloidal suspensions will be presented. Here we are concerned with purely repulsive particles (hard or soft spheres), or with particles with attractions of moderate strength and range (colloid–polymer and colloid–colloid mixtures). Although we shall focus mainly on equilibrium aspects, a few comments will be made about the associated kinetics as well [69, 70].

C2.6.6.1 COLLOIDAL CRYSTALS

One of the intriguing and beautiful properties of suspensions of well defined colloidal particles is their ability to order into a regular crystal lattice, called a colloidal crystal. The lattice spacing in colloidal crystals is set by the particle size and tends to be similar to the wavelength of light. Therefore, Bragg scattering (iridescence) can be observed using light (see [section B1.9](#)). Examples of this were found first in nature. For instance, tipula iridescent virus (TIV) particles were observed to assume face centred cubic (fcc) stackings [71], and opals are fossilized colloidal crystals consisting of silica [72]. For further background, see [69, 73, 74]. Colloidal crystals are used as model systems to study the freezing transition. Because of their optical properties, they are also being investigated for potential applications such as optical rejection filters (for instance, [75]).

C2.6.6.2 HARD SPHERES

Hard spheres are perhaps the simplest system to undergo a freezing transition. Freezing of hard spheres was observed using computer simulations [76]. The freezing and melting densities were found to be $\phi_F = 0.49$ and $\phi_M = 0.55$ [77]. The stable crystal structure is fcc (see [78] and references therein). So, although this may seem counter intuitive, no particle attractions are needed for a freezing transition to occur—at sufficiently high density (pressure), this will also occur for particles with purely repulsive interactions. The phase behaviour of hard spheres is summarized in figure C2.6.8.

Encyclopedia of Chemical Physics and Physical Chemistry

-24-

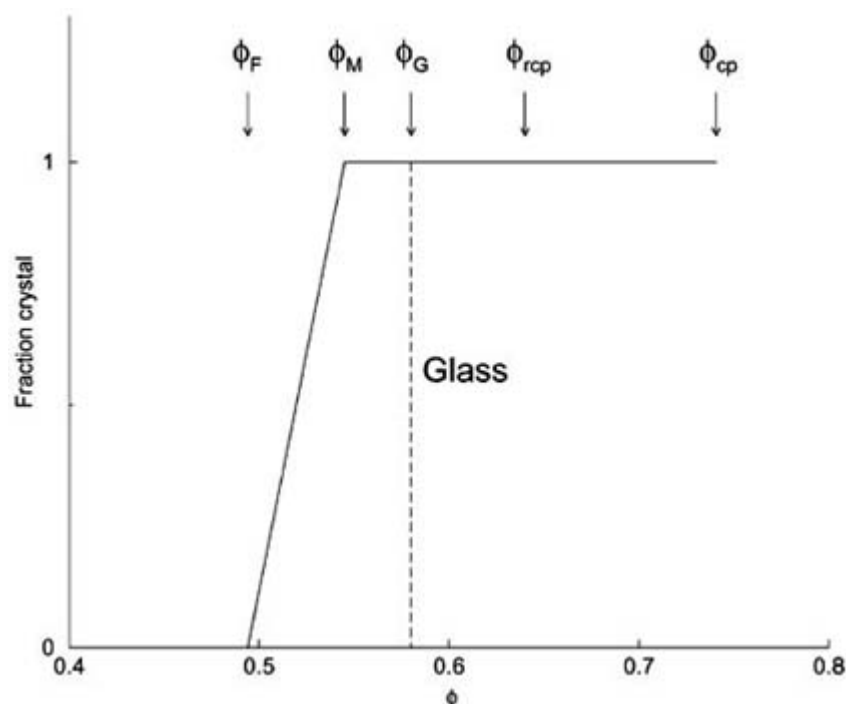


Figure C2.6.8. Phase diagram of hard spheres (see text for details)

Experimentally, the hard-sphere phase transition was observed using non-aqueous polymer lattices [79, 80]. Samples are prepared, brought into the fluid state by tumbling and then left to stand. Depending on particle size and concentration, colloidal crystals then form on a time scale from minutes to days. Experimentally, there is always some uncertainty in the actual volume fraction. Often the concentrations are therefore rescaled so freezing occurs at $\phi_F = 0.49$. The width of the coexistence region agrees well with simulations [11, 80].

On further increasing the concentration, the glass transition at $\phi_G \approx 0.58$ is reached [81]. At this concentration, the overall structure is arrested and particles can only undergo local diffusive motion. In other words, the sample is not ergodic anymore, as can be shown using dynamic light scattering, where the intermediate scattering function does not relax to zero (see section B1.10). The hard-sphere glass serves as a model to understand the glass transition in simple liquids (see also section C2.15 *Disordered Systems*).

Samples can be concentrated beyond the glass transition. If this is done quickly enough to prevent crystallization, this ultimately leads to a random close-packed structure, with a volume fraction $\phi_{rcp} \approx 0.64$. Close-packed structures, such as fcc, have a maximum packing density of $\phi_{cp} = 0.74$. The crystallization kinetics are strongly concentration dependent. The nucleation rate is fastest near the melting concentration. On increasing concentration, the nucleation process is arrested. This has been found to occur at the glass transition [82].

The formation of colloidal crystals requires particles that are fairly monodisperse—experimentally, hard sphere crystals are only observed to form in samples with a polydispersity below about 0.08 [69]. Using computer

simulations, a maximum polydispersity for the solid phase of 0.06 was predicted [83].

C2.6.6.3 SOFT SPHERES

Charged particles in polar solvents have soft-repulsive interactions (see section C2.6.4). Just as hard spheres, such particles also undergo an ordering transition. Important differences, however, are that the transition takes place at (much) lower particle volume fractions, and at low ionic strength (low κ) the solid phase may be body centred cubic (bcc), rather than the more compact fcc structure (see [69, 73, 84]). For the interactions, a Yukawa potential (equation (C2.6.11)) is often used. The phase diagram for the Yukawa potential was calculated using computer simulations by Robbins *et al* [85].

We will focus on one experimental study here. Monovoukas and Gast studied polystyrene particles with $a = 67$ nm in potassium chloride solutions [86]. They obtained a very good agreement between their observations and the predicted Yukawa phase diagram (see figure C2.6.9). In order to make the comparison they rescaled the particle charges according to Alexander *et al* [43] (see also [87]). At high electrolyte concentrations, the particle interactions tend to hard-sphere behaviour (see section C2.6.4) and the phase transition shifts to volume fractions around 0.5 [88].

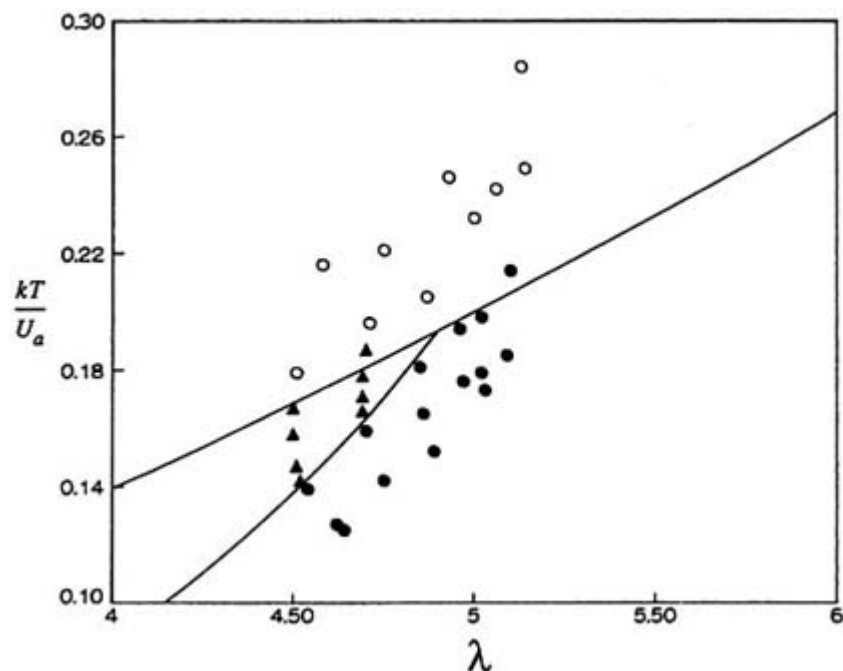


Figure C2.6.9. Phase diagram of charged colloidal particles. The solid lines are predictions by Robbins *et al* [85]. Fluid phase (open circles), fcc crystal (solid circles) and bcc crystal (triangles). U_a is the interaction energy at the mean particle separation $x = n^{-1/3}$, and $\lambda = \kappa x$ (reproduced with permission from [86]. Copyright 1989 Academic Press).

In extensively deionized suspensions, there are experimental indications for effective attractions between particles, such as long-lived void structures [89] and attractions between particles confined between charged walls [90]. Nevertheless, under these conditions the DLVO theory does seem to describe interactions of isolated particles at the pair level correctly [90]. It may be possible to explain the experimental observations by taking into account explicitly the degrees of freedom of both the colloidal particles and the small ions [91, 92].

C2.6.6.4 COLLOID–POLYMER MIXTURES

In section C2.6.4.3 it was shown how the addition of non-adsorbing polymer chains induces a depletion attraction between colloidal particles. If sufficient polymer is added, these attractions can be strong enough to induce a phase separation of the colloidal particles. An early application of this was the creaming of rubber latex [93].

Much later, experiments on model colloids revealed that the addition of polymer may either induce a gas–liquid type phase separation or a fluid–solid transition [94, 95, 96 and 97]. Using perturbation theories, these observations could be accounted for quite well [97, 98].

At equilibrium, in order to achieve equality of chemical potentials, not only the colloid but also the polymer concentrations in the different phases are different. We focus here on a theory that allows for this polymer partitioning [99]. Predictions for two polymer/colloid size ratios are shown in figure C2.6.10. A liquid phase is predicted to occur only when the range of attractions is not too small compared to the particle size, $\delta/a > 0.3$. Under these conditions a phase behaviour is obtained that is similar to that of simple liquids, such as argon. Because of the polymer partitioning, however, there is a three-phase triangle (rather than a triple point). For smaller polymer (narrower attractions), the gas–liquid transition becomes metastable with respect to the fluid–crystal transition. These predictions were confirmed experimentally [100]. The phase boundaries were predicted semi-quantitatively.

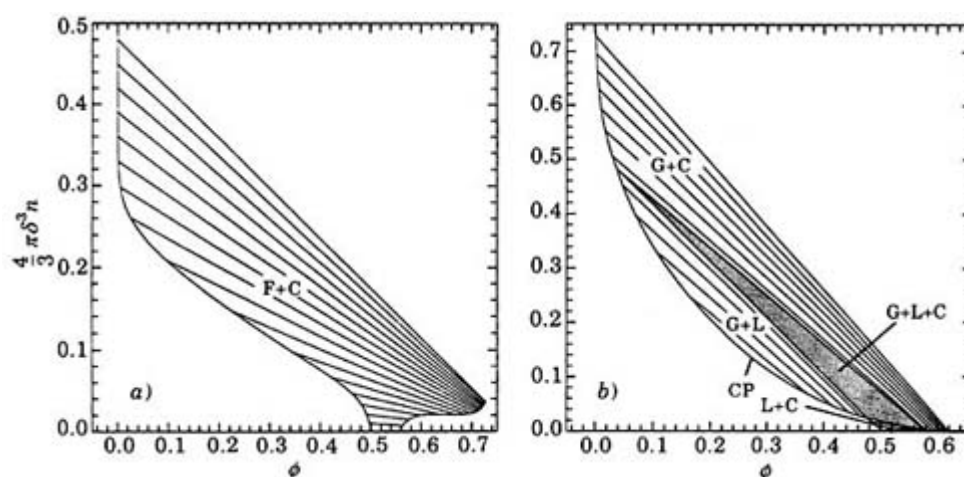


Figure C2.6.10. Phase diagram of colloid–polymer mixtures: polymer coil volume fraction $\frac{4}{3}\pi\delta^3 n$ vs particle volume fraction ϕ . (a) Narrow attractions, $\delta/a = 0.1$. Only a fluid–crystal transition is present. Tie lines indicate coexisting phases. (b) Longer range attractions, $\delta/a = 0.4$. Gas, liquid and crystal phases (G, L and C) are present, as well as a critical point (CP). The three-phase triangle is shaded (reproduced with permission from [99]. Copyright 1992 EDP Sciences).

In practice, colloidal systems do not always reach the predicted equilibrium state, which is observed here for the case of narrow attractions. On increasing the polymer concentration, a fluid–crystal phase separation may be induced, but at higher concentration crystallization is arrested and amorphous gels have been found to form instead [101, 102]. Close to the phase boundary, transient gels were observed, in which phase separation proceeded after a lag time.

The behaviour of these systems is similar to that of suspensions in which short-range attractions are induced by changing solvent quality for sterically stabilized particles (e.g. [103]). Another case in which narrow attractions arise is that of solutions of globular proteins. These crystallize only in a narrow range of concentrations [104].

C2.6.6.5 MIXTURES OF HARD SPHERES

As shown in [section C2.6.6.2](#), hard-sphere suspensions already show a rich phase behaviour. This is even more the case when binary mixtures of hard spheres are considered. First, we will mention the case of moderate size ratios, around 0.6. At low concentrations these form a mixed fluid phase. On increasing the overall concentration of mixtures, however, binary crystals of type AB_2 and AB_{13} were observed (where A represents the larger spheres), in addition to pure A or B crystals [[105](#), [106](#)]. An example of an AB_2 structure is shown in figure C2.6.11. Computer simulations confirmed the thermodynamic stability of the structures that were observed [[107](#), [108](#)].

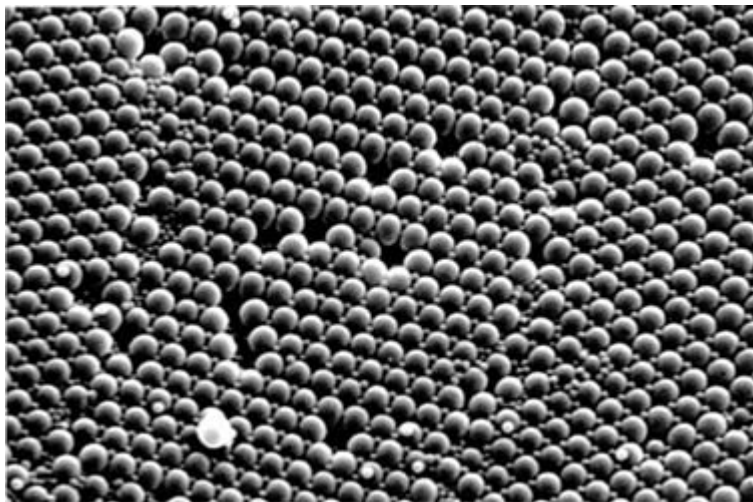


Figure C2.6.11. SEM of AB_2 structure, formed in aqueous mixtures of PS latex spheres with $a = 68$ nm and $a = 264$ nm (courtesy of Prof R H Ottewill).

A second case to be considered is that of mixtures with a small size ratio, <0.2 . For a long time it was believed that such mixtures would not show any instability in the fluid phase, but such an instability was predicted by Biben and Hansen [[109](#)]. This can be understood to be as a result of depletion interactions, exerted on the large spheres by the small spheres (see [section C2.6.4.3](#)). Experimentally, such mixtures were indeed found to display an instability [[110](#)]. The gas–liquid transition does, however, seem to be metastable with respect to the fluid–crystal transition [[111](#), [112](#)]. This was confirmed by computer simulations [[113](#)].

C2.6.6.6 NON-SPHERICAL COLLOIDS

Other possibilities for observing phase transitions are offered by suspensions of non-spherical particles. Such systems can display liquid crystalline phases, in addition to the isotropic liquid and crystalline phases (see also [section C2.2](#)). First, we consider rod-like particles (see [[114](#), [115](#)], and references therein). As shown by Onsager [[116](#), [117](#)], sufficiently elongated particles will display a nematic phase, in which the particles have a tendency to align parallel to

each other. Several experimental systems display this Onsager transition [[114](#)].

Hard spherocylinders (cylinders with hemispherical end caps) were studied using computer simulations [[118](#)]. In addition to a nematic phase, such particles also display a smectic-A phase, in which the particles are arranged in liquid-like layers. To observe this transition, rather monodisperse particles are needed. The smectic-A phase was indeed observed in suspensions of TMV particles [[17](#)].

Disc-like particles can also undergo an Onsager transition—here the particles form a discotic nematic, where the short particle axes tend to be oriented parallel to each other. In practice, clay suspensions tend to display sol–gel transitions, without a clear tendency towards nematic ordering (for instance, [[22](#)]). Using sterically stabilized platelets, an isotropic–nematic transition could be observed [[119](#)].

ACKNOWLEDGMENTS

The author is very grateful to Professors A van Blaaderen, R H Ottewill and D A Weitz for providing original photographs.

REFERENCES

- [1] Kruyt H R (ed) 1952 *Colloid Science* vol I (Amsterdam: Elsevier)
- [2] Bibette J, Leal Calderon F and Poulin P 1999 Emulsions: basic principles *Rep. Prog. Phys.* **62** 969–1033
- [3] Onsager L 1933 Theories of concentrated electrolytes *Chem. Rev.* **13** 73–89
- [4] McMillan W G and Mayer J E 1945 The statistical thermodynamics of multicomponent systems *J. Chem. Phys.* **13** 276–305
- [5] Matijević E 1976 Preparation and characterization of monodispersed metal hydrous oxide sols *Prog. Colloid Polym. Sci.* **61** 24–35
- [6] Iler R K 1979 *The Chemistry of Silica* (New York: Wiley)
- [7] Stöber W, Fink A and Bohn E 1968 Controlled growth of monodisperse silica spheres in the micron size range *J. Colloid Interface Sci.* **26** 62–9
- [8] van Helden A K, Jansen J W and Vrij A 1981 Preparation and characterization of spherical monodisperse silica dispersions in nonaqueous solvents *J. Colloid Interface Sci.* **81** 354–68
- [9] Vanderhoff J W, van den Hul H J, Tausk R J and Overbeek J Th G 1970 The preparation of monodisperse latexes with well-characterized surfaces *Clean Surfaces* ed G Goldfinger (New York: Dekker) pp 15–44
- [10] Antl L, Goodwin J W, Hill R D, Ottewill R H, Owens S M, Papworth S and Waters J A 1986 The preparation of poly (methyl methacrylate) lattices in non-aqueous media *Colloid Surf.* **17** 67–78

Encyclopedia of Chemical Physics and Physical Chemistry

-29-

- [11] Underwood S M, Taylor J R and van Megen W 1994 Sterically stabilised colloidal particles as model hard spheres *Langmuir* **10** 3550–4
- [12] Pathmamanoharan C and Philipse A P 1998 Preparation and properties of monodisperse magnetic cobalt colloids grafted with polyisobutene *J. Colloid Interface Sci.* **205** 304–53
- [13] van Blaaderen A and Vrij A 1992 Synthesis and characterisation of colloidal dispersions of fluorescent, monodisperse silica spheres *Langmuir* **8** 2921–31
- [14] Liz-Marzán L M, Giersig M and Mulvaney P 1996 Synthesis of nanosized gold-silica core-shell particles *Langmuir* **12** 4329–35
- [15] Piazza R, Bellini T and Degiorgio V 1993 Equilibrium sedimentation profiles of screened charged colloids: a test of the hard-sphere equation of state *Phys. Rev. Lett.* **71** 4267–70
- [16] Saunders B R and Vincent B 1999 Microgel particles as model colloids: theory, properties and applications *Adv. Colloid Interface Sci.* **80** 1–25
- [17] Wen X, Meyer R B and Caspar D L D 1989 Observation of smectic-A ordering in a solution of rigid-rod-like particles *Phys. Rev. Lett.* **63** 2760–3

- [18] Fraden S 1995 Phase transitions in colloidal suspensions of virus particles *Observation, Prediction and Simulation of Phase Transitions in Complex Fluids* ed M Baus, L F Rull and J P Hansen (Dordrecht: Kluwer) pp 113–64
- [19] Martin C, Weyerich B, Biegel J, Deike R, Johner C, Klein R and Weber R 1995 Electric-field light-scattering by rod-like polyelectrolytes in aqueous suspensions *J. Physique II* **5** 697–719
- [20] Buining P A, Veldhuizen Y S J, Pathmamanoharan C and Lekkerkerker H N W 1992 Preparation of a non-aqueous dispersion of sterically stabilized boehmite rods *Colloid. Surf.* **64** 47–55
- [21] van Olphen H 1977 *Introduction to Clay Colloid Chemistry* 2nd edn (New York: Wiley)
- [22] Mourchid A, Delville A, Lambard J, L&233;colier E and Levitz P 1995 Phase diagram of colloidal dispersions of anisotropic charged particles: equilibrium properties, structure, and rheology of Laponite suspensions *Langmuir* **11** 1942–50
- [23] van Blaaderen A and Wiltzius P 1995 Real-space structure of colloidal hard-sphere glasses *Science* **270** 1177–9
- [24] Donald A M 1998 Environmental scanning electron microscopy for the study of 'wet' systems *Curr. Op. Colloid Interface Sci.* **3** 143–7
- [25] Grier D G 1997 Optical tweezers in colloid and interface science *Curr. Op. Colloid Interface Sci.* **2** 264–70
- [26] Poulin P, Cabuil V and Weitz D A 1997 Direct measurement of colloidal forces in an anisotropic solvent *Phys. Rev. Lett.* **79** 4862–5
- [27] Ducker W A, Senden T J and Pashley R M 1991 Direct measurement of colloidal forces using an atomic force microscope *Nature* **353** 239–41
- [28] Barnes H A, Hutton J F and Walters K (ed) 1989 *An Introduction to Rheology* (Amsterdam: Elsevier)
- [29] Macosko C H 1994 *Rheology* (New York: Wiley)
- [30] Larson R G 1999 *The Structure and Rheology of Complex Fluids* (New York: Oxford University Press)
- [31] Liu S J and Masliyah J H 1996 Rheology of suspensions *Adv. Chem. Series* **251** 107–76

- [32] Hunter R J 1989 *Foundations of Colloid Science* vol II (Oxford: Oxford University Press)
- [33] Russel W B, Saville D A and Schowalter W R 1989 *Colloidal Dispersions* 2nd edn (Cambridge: Cambridge University Press)
- [34] Hiemenz P C and Rajagopalan R 1997 *Principles of Colloid and Surface Chemistry* 3rd edn (New York: Dekker)
- [35] Perrin J 1910 Mouvement brownien et moléculs *J. Physique, Paris* **9** 5–39
- [36] Lyklema J 1995 *Fundamentals of Interface and Colloid Science* vol 2 (London: Academic Press)
- [37] Israelachvili J N 1992 *Intermolecular and Surface Forces* 2nd edn (London: Academic)
- [38] Hamaker H C 1937 London-van der Waals attraction between spherical particles *Physica* **4** 1058–72
- [39] Hunter R J 1987 *Foundations of Colloid Science* vol I (Oxford: Oxford University Press)
- [40] Dzaloshinskii I E, Lifshitz E M and Pitaevskii L P 1961 The general theory of van der Waals forces *Adv. Phys.* **10** 165–208
- [41] Verwey E J W and Overbeek J Th G 1948 *Theory of the Stability of Lyophobic Colloids* (New York: Elsevier)

- [42] Reerink H and Overbeek J Th G 1954 The rate of coagulation as a measure of the stability of silver iodide sols *Discuss. Faraday Soc.* **18** 74–84
- [43] Alexander S, Chaikin P M, Grant P, Morales G J, Pincus P and Hone D 1984 Charge renormalisation, osmotic pressure, and bulk modulus of colloidal crystals: theory *J. Chem. Phys.* **80** 5776–81
- [44] Salgi P and Rajagopalan R 1993 Polydispersity in colloids—implications to static structure and scattering *Adv. Colloid Interface Sci.* **43** 169–288
- [45] Klein R and D’Aguanno B 1996 Scattering properties of colloidal suspensions *Light Scattering, Principles and Development* ed W Brown (Oxford: Clarendon) pp 30–102
- [46] Napper D H 1983 *Polymeric Stabilization of Colloidal Dispersions* (London: Academic)
- [47] Fleer G J, Cohen Stuart M A, Scheutjens J M H M, Cosgrove T and Vincent B 1993 *Polymers at Interfaces* (London: Chapman & Hall)
- [48] Jansen J W, de Kruif C G and Vrij A 1986 Attractions in sterically stabilised silica dispersions. I. Theory of phase separation *J. Colloid Interface Sci.* **114** 471–80
- [49] Baxter R J 1968 Percus–Yevick equation for hard spheres with surface adhesion *J. Chem. Phys.* **49** 2770–4
- [50] Asakura S and Oosawa F 1954 On interaction between two bodies immersed in a solution of macromolecules *J. Chem. Phys.* **22** 1255–6
- [51] Vrij A 1976 Polymers at interfaces and the interactions in colloidal dispersions *Pure Appl. Chem.* **48** 471–83
- [52] Jenkins P and Snowden M 1996 Depletion flocculation in colloidal dispersions *Adv. Colloid Interface Sci.* **68** 57–96
- [53] Hidalgo-Alvarez R, Martin A, Fernandez A, Bastos D, Martinez F and de las Nieves F J 1996 Electrokinetic properties, colloidal stability and aggregation kinetics of polymer colloids *Adv. Colloid Interface Sci.* **67** 1–118
- [54] Poon W C K and Haw M D 1997 Mesoscopic structure formation in colloidal aggregation and gelation *Adv. Colloid Interface Sci.* **73** 71–126

- [55] Asnaghi D, Carpineti M, Giglio M and Vailati A 1997 Small angle light scattering studies concerning aggregation processes *Curr. Op. Colloid Interface Sci.* **2** 246–50
- [56] von Smoluchowski M 1917 Versuch einer Mathematischen Theorie der Koagulationkinetik Kolloider Lösungen *Z. Phys. Chem.* **92** 129–68
- [57] Overbeek J Th G 1952 Kinetics of flocculation *Colloid Science* vol I, ed H R Kruyt (Amsterdam: Elsevier) pp 278–301
- [58] Higashitani K and Matsuno Y 1979 Rapid Brownian coagulation of colloidal dispersions *J. Chem. Eng. Japan* **12** 460–5
- [59] Fuchs N 1934 Über die Stabilität und Aufladung der Aerosole *Z. Phys.* **89** 736–43
- [60] Behrens S H, Borkovec M and Schurtenberger P 1998 Aggregation in charge-stabilized colloidal suspensions revisited *Langmuir* **14** 1951–4
- [61] Harrison A 1995 *Fractals in Chemistry* (Oxford: Oxford University Press)
- [62] Meakin P 1983 Formation of fractal clusters and networks by irreversible diffusion-limited aggregation *Phys. Rev. Lett.* **51** 1119–22

- [63] Kolb M, Botet R and Jullien R 1983 Scaling of kinetically growing clusters *Phys. Rev. Lett.* **51** 1123–6
- [64] Lin M Y, Lindsay H M, Weitz D A, Ball R C, Klein R and Meakin P 1989 Universality of fractal aggregates as probed by light scattering *Proc. R. Soc. A* **423** 71–87
- [65] Weitz D A and Huang J S 1984 Self-similar structures and the kinetics of aggregation of gold colloids *Kinetics of Aggregation and Gelation* ed F Family and D P Landau (Amsterdam: North-Holland) pp 19–28
- [66] Brown W D and Ball R C 1985 Computer simulation of chemically limited aggregation *J. Phys. A: Math. Gen.* **18** L517–21
- [67] Schaefer D W, Martin J E, Wiltzius P and Cannell D S 1984 Fractal geometry of colloidal aggregates *Phys. Rev. Lett.* **52** 2371–4
- [68] Rouw P W and de Kruijff C G 1989 Adhesive hard-sphere colloidal dispersions: fractal structures and fractal growth in silica dispersions *Phys. Rev. A* **39** 5399–408
- [69] Pusey P N 1991 Colloidal suspensions *Liquids, Freezing and Glass Transition* ed J P Hansen, D Levesque and J Zinn-Justin (Amsterdam: Elsevier) pp 763–942
- [70] Dhont J K G 1996 *An Introduction to Dynamics of Colloids* (Amsterdam: Elsevier)
- [71] Williams R C and Smith K M 1957 A crystallizable insect virus *Nature* **179** 119–20
- [72] Sanders J V 1964 Colour of precious opal *Nature* **204** 1151–3
- [73] Pieranski P 1983 Colloidal crystals *Chem. Phys.* **24** 25–73
- [74] Sood A K 1991 Structural ordering in colloidal suspensions *Solid State Phys.* **45** 1–73
- [75] Asher S A, Holtz J, Liu L and Wu Z 1994 Self-assembly motif for creating submicron periodic materials. Polymerized crystalline colloidal arrays *J. Am. Chem. Soc.* **116** 4997–8
- [76] Alder B J and Wainwright T E 1957 Phase transition for a hard sphere system *J. Chem. Phys.* **27** 1208–9
- [77] Hoover W G and Ree F H 1968 Melting transition and communal entropy for hard spheres *J. Chem. Phys.* **49** 3609–17

- [78] Bruce A D, Wilding N B and Ackland G J 1997 Free energies of crystalline solids: a lattice-switch Monte-Carlo method *Phys. Rev. Lett.* **79** 3002–5
- [79] Kose A and Hachisu S 1974 Kirkwood–Alder transition in monodisperse latexes. I. Nonaqueous systems *J. Colloid Interface Sci.* **46** 460–9
- [80] Pusey P N and van Megen W 1986 Phase behaviour of concentrated suspensions of nearly hard colloidal spheres *Nature* **320** 340–2
- [81] van Megen W and Underwood S M 1993 Dynamic light scattering study of glasses of hard colloidal spheres *Phys. Rev. E* **47** 248–61
- [82] van Megen W and Underwood S M 1993 Change in crystallization mechanism at the glass transition of colloidal spheres *Nature* **362** 616–18
- [83] Bolhuis P G and Kofke D A 1996 Monte Carlo study of freezing of polydisperse hard spheres *Phys. Rev. E* **54** 634–43
- [84] Arora A K and Tata B V R (ed) 1996 *Ordering and Phase Transitions in Charged Colloids* (New York: VCH)
- [85] Robbins M O, Kremer K and Grest G S 1988 Phase diagram and dynamics of Yukawa systems *J. Chem. Phys.* **88** 3286–312

- [86] Monovoukas Y and Gast A P 1989 The experimental phase diagram of charged colloidal suspensions *J. Colloid Interface Sci.* **128** 533–48
- [87] Palberg T, Mönch W, Bitzer F, Piazza R and Bellini T 1995 Freezing transition for colloids with adjustable charge: a test of charge renormalization *Phys. Rev. Lett.* **74** 4555–8
- [88] Hachisu S and Kobayashi Y 1974 Kirkwood–Alder transition in monodisperse latexes. II. Aqueous latexes of high electrolyte concentration *J. Colloid Interface Sci.* **46** 470–6
- [89] Ito K, Yoshida H and Ise N 1994 Void structure in colloidal dispersions *Science* **263** 66–8
- [90] Crocker J C and Grier D G 1996 When like charges attract: the effects of geometrical confinement on long-range colloidal interactions *Phys. Rev. Lett.* **77** 1897–900
- [91] Löwen H, Hansen J P and Madden P A 1993 Nonlinear counterion screening in colloidal suspensions *J. Chem. Phys.* **98** 3275–89
- [92] van Roij R and Hansen J P 1997 Van der Waals-like instability in suspensions of mutually repelling charged colloids *Phys. Rev. Lett.* **79** 3082–5
- [93] Vester C F 1938 Die Rahmung des Hevea-Latex mittels Kolloiden *Kolloid Z.* **84** 63–74
- [94] Kose A and Hachisu S 1976 Ordered structure in weakly flocculated monodisperse latex *J. Colloid Interface Sci.* **55** 487–98
- [95] de Hek H and Vrij A 1981 Interactions in mixtures of colloidal silica spheres and polystyrene molecules in cyclohexane *J. Colloid Interface Sci.* **84** 409–22
- [96] Sperry P R 1984 Morphology and mechanism in latex flocculated by volume restriction *J. Colloid Interface Sci.* **99** 97–108
- [97] Vincent B, Edwards J, Emmett S and Croot R 1988 Phase separation in dispersions of weakly-interacting particles in solutions of non-adsorbing polymers *Colloid Surf.* **31** 267–98
- [98] Gast A P, Hall C K and Russel W B 1983 Polymer-induced phase separations in nonaqueous colloidal suspensions *J. Colloid Interface Sci.* **96** 251–67

- [99] Lekkerkerker H N W, Poon W C K, Pusey P N, Stroobants A and Warren P B 1992 Phase behaviour of colloid + polymer mixtures *Europhys. Lett.* **20** 559–64
- [100] Ilett S M, Orrock A, Poon W C K and Pusey P N 1995 Phase behaviour of a model colloid–polymer mixture *Phys. Rev. E* **51** 1344–52
- [101] Poon W C K, Pirie A D and Pusey P N 1995 Gelation in colloid–polymer mixtures *Faraday Discuss.* **101** 65–76
- [102] Verhaegh N A M, Asnaghi D, Lekkerkerker H N W, Giglio M and Cipelletti L 1997 Transient gelation by spinodal decomposition in colloid–polymer mixtures *Physica A* **242** 104–18
- [103] Chen M and Russel W B 1991 Characteristics of flocculated silica dispersions *J. Colloid Interface Sci.* **141** 564–77
- [104] Rosenbaum D, Zamora P C and Zukoski C F 1996 Phase behaviour of small attractive colloidal particles *Phys. Rev. Lett.* **76** 150–3
- [105] Bartlett P, Ottewill R H and Pusey P N 1990 Freezing of binary mixtures of colloidal hard spheres *J. Chem. Phys.* **93** 1299–312
- [106] Bartlett P, Ottewill R H and Pusey P N 1992 Superlattice formation in binary mixtures of hard-sphere colloids *Phys. Rev. Lett.* **68** 3801–4

- [107] Eldridge M D, Madden P A and Frenkel D 1993 Entropy-driven formation of a superlattice in a hard-sphere binary mixture *Mol. Phys.* **79** 105–20
- [108] Eldridge M D, Madden P A and Frenkel D 1993 Entropy driven formation of a superlattice in a hard-sphere binary mixture *Nature* **365** 35–7
- [109] Biben T and Hansen J P 1991 Phase separation of asymmetric binary hard-sphere fluids *Phys. Rev. Lett.* **66** 2215–18
- [110] van Duijneveldt J S, Heinen A W and Lekkerkerker H N W 1993 Phase separation in bimodal dispersions of sterically stabilized silica particles *Europhys. Lett.* **21** 369–74
- [111] Dinsmore A D, Yodh A G and Pine D J 1995 Phase diagrams of nearly hard-sphere binary colloids *Phys. Rev. E* **52** 4045–57
- [112] Imhof A and Dhont J K G 1995 Experimental phase diagram of a binary colloidal hard-sphere mixture with a large size ratio *Phys. Rev. Lett.* **75** 1662–5
- [113] Dijkstra M, van Roij R and Evans R 1999 Direct simulation of the phase behaviour of binary hard-sphere mixtures: test of the depletion potential description *Phys. Rev. Lett.* **82** 117–20
- [114] Lekkerkerker H N W, Buining P, Buitenhuis J, Vroege G J and Stroobants A 1995 Liquid crystal phase transitions in dispersions of rodlike colloidal particles *Observation, Prediction and Simulation of Phase Transitions in Complex Fluids* ed M Baus, L F Rull and J P Hansen (Dordrecht: Kluwer) pp 53–112
- [115] Vroege G J and Lekkerkerker H N W 1992 Phase transitions in lyotropic colloidal and polymer liquid crystals *Rep. Prog. Phys.* **55** 1241–309
- [116] Onsager L 1942 Anisotropic solutions of colloids *Phys. Rev.* **62** 558
- [117] Onsager L 1949 The effects of shape on the interaction of colloidal particles *Ann. NY Acad. Sci.* **51** 627–59
- [118] Frenkel D, Lekkerkerker H N W and Stroobants A 1988 Thermodynamic stability of a smectic phase in a system of hard rods *Nature* **332** 822–3

Encyclopedia of Chemical Physics and Physical Chemistry

-34-

- [119] van der Kooij F M and Lekkerkerker H N W 1998 Formation of nematic liquid crystals in suspensions of hard colloidal platelets *J. Phys. Chem. B* **102** 7829–32
- [120] Hough D B and White L R 1980 The calculation of Hamaker constants from Lifshitz theory with applications to wetting phenomena *Adv. Colloid Interface Sci.* **14** 3–41

FURTHER READING

Fennel Evans D and Wennerström H 1994 *The Colloidal Domain* (New York: VCH)

Introductory text with an emphasis on self-assembly systems and emulsions

Hiemenz P C and Rajagopalan R 1997 *Principles of Colloid and Surface Chemistry* 3rd edn (New York: Marcel Dekker)

General textbook on colloid and surface science, including details about characterization methods

Hunter R J 1987 and 1989 *Foundations of Colloid Science* vols I and II (Oxford: Clarendon Press)

Extensive, general introduction to colloid science

Pusey P N 1991 Colloidal suspensions *Liquids, Freezing and Glass Transition* ed J P Hansen, D Levesque and J Zinn-Justin

Les Houches, session LI (Amsterdam: North-Holland) pp 763–942

Advanced review paper, with a detailed section on colloidal dynamics

Russel W B, Saville D A and Schowalter W R 1989 *Colloidal Dispersions* (Cambridge: Cambridge University Press)

General textbook, emphasizing the physical equilibrium and non-equilibrium properties of colloids

Shaw D J 1996 *Introduction to Colloid and Surface Chemistry* (Oxford: Butterworth-Heinemann)

Short introductory textbook

Encyclopedia of Chemical Physics and Physical Chemistry

-1-

C2.7 Catalysis

Bruce C Gates

C2.7.1 INTRODUCTION

The idea of a *catalyst* is one of the most fascinating and significant in science and the word is one of the few that have carried over broadly from scientific into nonscientific language. A *catalyst* speeds up a chemical reaction without being consumed substantially—the occurrence of a reaction accelerated by a catalyst is called *catalysis*. At first, one might think that catalysis seems too good to be true, but the principles are well understood; a catalyst works by forming chemical bonds with reactants, generating intermediates that react more readily to give products than the reactants would alone—and giving back the catalyst. A catalyst affects the rate of approach to equilibrium of a reaction but not the position of the equilibrium. Catalysts provide subtle control of chemical conversions; a good catalyst increases the rate of a desired reaction but not the rates of undesired side reactions. Catalysis is ubiquitous in biology and technology and is the key to the efficiency of most chemical conversions. Only temperature provides a comparable means for increasing reaction rates, but high temperatures are often unacceptable—for example, because they harm biological organisms; high temperatures in chemical technology often mean high costs, e.g., because reaction in a liquid at high temperature requires a high pressure to maintain the liquid state, and high-pressure equipment is expensive.

The *activity* of a catalyst is a measure (e.g., a rate or a rate constant) of how fast a catalytic reaction occurs under some standard conditions. Activities vary from catalyst to catalyst and depend on the variables that influence reaction rates (temperature, reactant concentrations etc). The *selectivity* of a catalyst is a measure of how fast it causes one reaction to proceed relative to others; selectivity might be defined, e.g., as the rate of formation of a desired product divided by the rate of formation of all products under some standard conditions. The *stability* of a catalyst accounts for how fast it loses activity during operation; the ideal catalyst is infinitely stable, but real catalysts undergo changes causing loss of activity and selectivity, and they must be replaced or regenerated (at intervals ranging from seconds to years). The *regenerability* of a technological catalyst is a measure of how well it responds to treatments to bring back its activity and selectivity after deactivation; many solid catalysts are regenerated by burning off carbonaceous deposits formed during operation with organic reactants.

Biology and chemical technology as we know them are hardly imaginable without catalysis. Almost all biological reactions are catalytic and the number of biological catalysts is huge. Most large-scale technological reactions are also catalytic, generating products valued in trillions of dollars annually—roughly two orders of magnitude more than the annual cost of the catalyst purchase and replacement. Products of catalytic technology (with examples) include fuels (gasoline), polymeric materials (polypropylene), clothing (nylon), pharmaceuticals (pain relievers), foods (hydrogenated fats), solvents (methanol) and chemicals (sulphuric acid). Catalysis is also essential to the

minimization of environmental pollutants, e.g., by conversion of automobile and power-plant emissions (e.g., CO and NO_x). By converting these harmful emissions into benign gases (e.g., CO₂ and N₂) and by reducing the production of harmful byproducts of chemical processes that had earlier been dumped, catalysis has dramatically improved the quality of the earth's air and water.

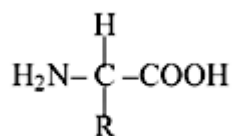
-2-

Catalysis spans chemistry, chemical engineering, materials science and biology. The goal here is to enliven the subject with diverse examples showing the microscopic details of catalysis.

C2.7.2 CLASSIFICATION OF CATALYSTS AND CATALYSIS

Catalysis in a single fluid phase (liquid, gas or supercritical fluid) is called *homogeneous catalysis* because the phase in which it occurs is relatively uniform or homogeneous. The catalyst may be molecular or ionic. Catalysis at an interface (usually a solid surface) is called *heterogeneous catalysis*; an implication of this term is that more than one phase is present in the reactor, and the reactants are usually concentrated in a fluid phase in contact with the catalyst, e.g., a gas in contact with a solid. Most catalysts used in the largest technological processes are solids. The term *catalytic site* (or *active site*) describes the groups on the surface to which reactants bond for catalysis to occur; the identities of the catalytic sites are often unknown because most solid surfaces are nonuniform in structure and composition and difficult to characterize well, and the active sites often constitute a small minority of the surface sites.

Most biological catalysts are enzymes, i.e., proteins, which are macromolecules (polypeptides) formed by biopolymerization of amino acids (with elimination of water); some enzymes are huge, with hundreds of monomer units. The 20 amino acid monomers occurring in nature,



have R groups including, e.g. H (in glycine), CH₂OH (in serine), CH₂COOH (in aspartic acid), CH₂CH₂COOH (in glutamic acid) and CH₂(CH₂)₃NH₂ (in lysine). Some of the R groups are ligands for metal ions (e.g., Zn, Fe, Cu, Mo and Co) that play catalytic roles. Acidic, basic and metal groups typically constitute the active sites of enzymes. The structures of a number of enzymes are known from x-ray crystallography. The polymer chains are usually folded to give a precise juxtaposition of the active sites on the interior surface of a clamlike structure. It is not a straightforward matter to classify enzyme catalysis as simple homogeneous or heterogeneous catalysis, because, in biological cells, enzymes exist both in solution (e.g., in cytoplasm) and within membranes—some enzymes are even arranged in assembly-line fashion so that the product molecules from one pass directly to the next.

The subject of catalysis has evolved with little integration of homogeneous, heterogeneous, and biological catalysis, as is reflected in the general references cited in the further reading section.

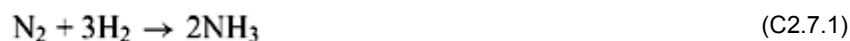
-3-

C2.7.3 A BIT OF HISTORY—THE AMMONIA SYNTHESIS REACTION

A definition of catalysis similar to that given above was stated first in about 1895 by Wilhelm Ostwald, whose work on catalysis was recognized with a Nobel prize. Sixty years before, Jakob Berzelius had coined the term

‘catalysis’, recognizing that a single concept could account for changes in compositions of numerous substances resulting from their mere contact with liquids, solids or ‘ferments’. Berzelius’s insight bears on phenomena that had earlier vexed the alchemists, who, aware of the then mysterious actions of these ‘ferments’ and other substances (‘contacts’), sought vainly for a philosopher’s stone to change base metals into gold.

Ostwald’s definition of catalysis rests on reaction kinetics and, indeed, at about the time he stated it, the beginnings of physical chemistry were emerging in the quantitative representation of the thermodynamics and kinetics of chemical reactions. The first concepts of reaction thermodynamics and kinetics came into focus in the early 1900s in work by Nernst and by Fritz Haber and coworkers; Haber carried out research motivated by the goal of synthesizing ammonia on a large scale from nitrogen and hydrogen, to produce fertilizers—and also explosives. The ammonia synthesis reaction,



which is still of great technological importance, takes place at almost negligibly low rates, except in the presence of a catalyst, and the reaction is strongly equilibrium limited. As the reaction is exothermic, the equilibrium conversion decreases with increasing temperature, so that the advantage of increasing the rate by increasing temperature is offset by a decrease in the attainable (equilibrium) conversion. Thus, an advantage of a highly active catalyst such as iron or ruthenium is that by increasing the reaction rate, the catalyst lowers the temperature at which the reaction can practically be carried out, thereby increasing the attainable conversion. Haber’s understanding of the interplay of thermodynamics and kinetics was pivotal to the development of the early concepts of physical chemistry; Haber’s Nobel prize is one of several that recognize research in catalysis.

Figure C2.7.1 is a potential energy diagram for the ammonia synthesis reaction taking place on the surface of an iron catalyst [1]. The reaction proceeds via chemisorbed intermediates (i.e., those chemically bonded to the iron surface) N, H, NH and NH₂, and the energy barriers shown in figure C2.7.1 are much lower than those that would pertain if the intermediates were N, H, NH and NH₂ in the gas phase; thus, figure C2.7.1 shows why iron is a good ammonia synthesis catalyst (and a good ammonia decomposition catalyst). Notice that by causing the reaction to proceed via steps whereby chemisorbed H combines with N, NH and then NH₂, the catalyst provides a sequence with each elementary step characterized by only a moderately high activation energy barrier; notice also that iron’s ability to cause dissociation of H₂ and N₂, leading to the formation of the chemisorbed atomic species, allows the formation of intermediates that are not too stable (not too low in energy)—if they were much more stable, then the barriers would be higher, the overall reaction not so fast and the catalyst not so active. Good catalysts generally provide efficient pathways for both the formation and conversion of highly reactive intermediates.

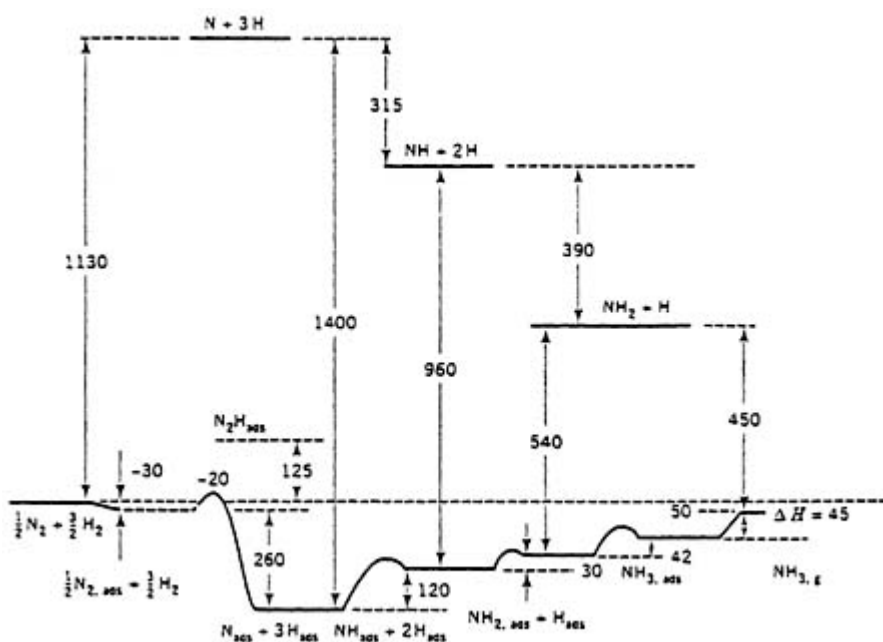
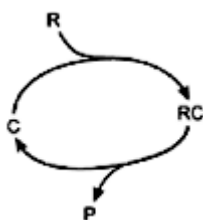


Figure C2.7.1. Schematic potential energy diagram for the catalytic synthesis and decomposition of ammonia on iron. The energies are in kJ mol^{-1} ; the subscript 'ads' refers to species adsorbed on iron [1].

The ammonia synthesis reaction provides another important lesson; it illustrates how fundamental science developed from research that was motivated by a technological need. Catalysis is one of the most essential and enduring enabling technologies, and it has been pulling researchers into unexplored territory from the beginning, with no end in sight. For example, much of ultrahigh-vacuum surface science (section A1.7) has emerged from work directed toward understanding of catalysis by solids and much of organometallic chemistry has emerged from work directed toward understanding of catalysis by transition metal compounds in solution.

C2.7.4 CATALYTIC CYCLES

Because a good catalyst is not consumed to a significant degree as it functions, catalysis is a cyclic process, and compact representations of catalysis are cycles that show the various intermediate species, illustrated by the following simple example, where C is the catalyst, R the reactant, P the product and RC the intermediate:



(using figure C2.7.1 can you write a cycle for the ammonia synthesis reaction?).

A well-understood catalytic cycle is that of the Wilkinson alkene hydrogenation (figure C2.7.2) [2]. Like most catalytic cycles, that shown in figure C2.7.2 is complex, involving intermediate species in the cycle (inside the dashed line) and other species outside the cycle and in dead-end paths. Knowledge of all but a small number of catalytic cycles is only fragmentary because of the complexity and because, if the catalyst is active, the cycle turns over rapidly and the concentrations of the intermediates are minute; thus, these intermediates are often not even

identified. Typically, as understanding develops, a more complex model emerges—i.e., a cycle with more intermediates. Determination of the important intermediates and quantitative representation of the kinetics and thermodynamics of the separate elementary reactions (e.g., figure C2.7.2) are often daunting tasks requiring the concerted application of various physical/chemical techniques. The challenges have helped to motivate the development of more and more sensitive spectroscopic methods, transient kinetics methods and so forth, some of which are mentioned below and described elsewhere in this encyclopedia.

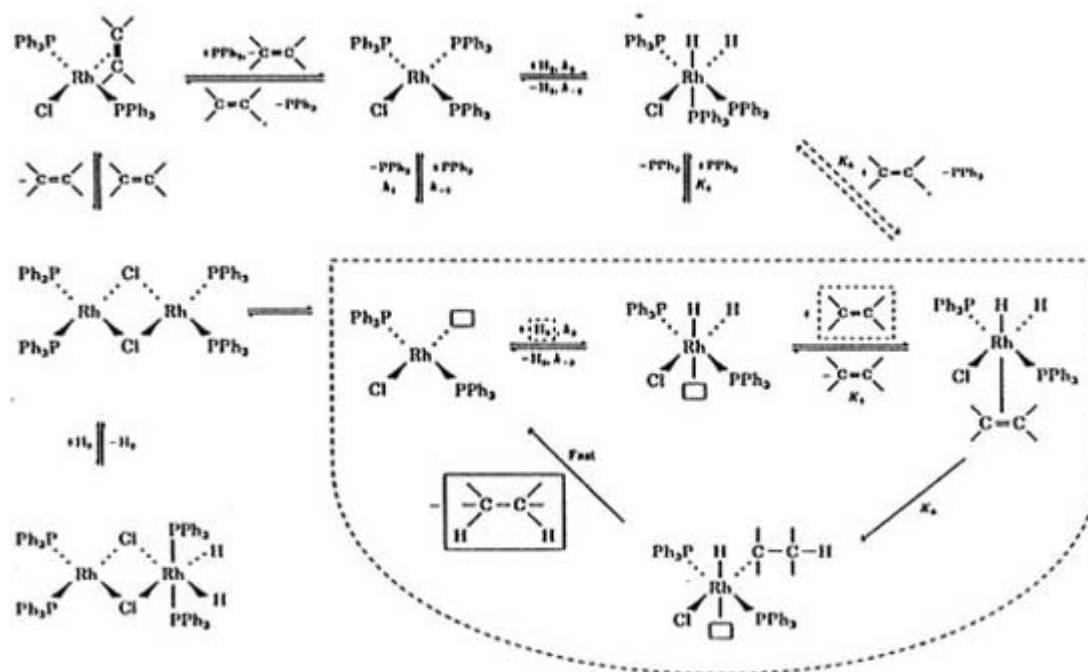


Figure C2.7.2. Catalytic cycle (within dashed lines) for the Wilkinson hydrogenation of alkene [2]. Values of rate and equilibrium constants are given in [2]

An important point about kinetics of cyclic reactions is that if an overall reaction proceeds via a sequence of elementary steps in a cycle (e.g., figure C2.7.2), some of these steps may be equilibrium limited so that they can proceed at most to only minute conversions. Nevertheless, if a step subsequent to one that is so limited is characterized by a large enough rate constant, then the equilibrium-limited step may still be fast enough for the overall cycle to proceed rapidly. Thus, the step following an equilibrium-limited step in the cycle pulls the cycle along—it drains the intermediate that can form in only a low concentration because of an equilibrium limitation and allows the overall reaction (the cycle) to proceed rapidly. A good catalyst accelerates the steps that most need a boost.

C2.7.5 MACROSCOPIC PHYSICAL PROPERTIES OF CATALYSTS

C2.7.5.1 SOLUBLE CATALYSTS

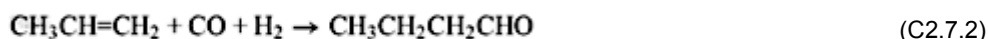
If a catalyst is to work well in solution, it (and the reactants) must be sufficiently soluble and stable. Most polar catalysts (e.g., acids and bases) are used in water and most organometallic catalysts (compounds of metals with organic ligands bonded to them) are used in organic solvents. Some enzymes function in aqueous biological solutions, with their solubilities determined by the polar functional groups (R groups) on their outer surfaces.

A solution containing both reactants and a catalyst may be mixed mechanically to bring the constituents into efficient contact—otherwise, the rate of the catalytic reaction would be affected by mass transport (e.g., diffusion)

and thus the reaction would proceed more slowly than in the absence of significant concentration gradients.

In technology, an economic separation of the products of a reaction from the solution containing the catalyst is necessary. Distillation is a commonly used method and, for it to work successfully, the products and catalyst must be stable at the temperatures of the distillation, which are often relatively high; some organometallic compounds, for example, may not meet this criterion.

As the separation is often an expensive part of a process, simplifications may be valuable. For example, in a process for hydroformylation of propene,



the reaction is carried out in a mixed reactor with a gas and two separate liquid phases, one organic, containing most of the product, and the other aqueous, containing almost all the catalyst, which is an organometallic compound of rhodium with phosphine ligands that are sulphonated to make them water soluble. The separation of the product from the catalyst is simple and economical [3] (figure C2.7.3): product liquid flows continuously to a tank (comparable to a separatory funnel) where it settles into two layers: one organic, containing the product; the other aqueous, containing the catalyst. The aqueous stream is recycled to the reactor so that the catalyst is continuously reused; thus, the process as well as the reaction is cyclic.

-7-

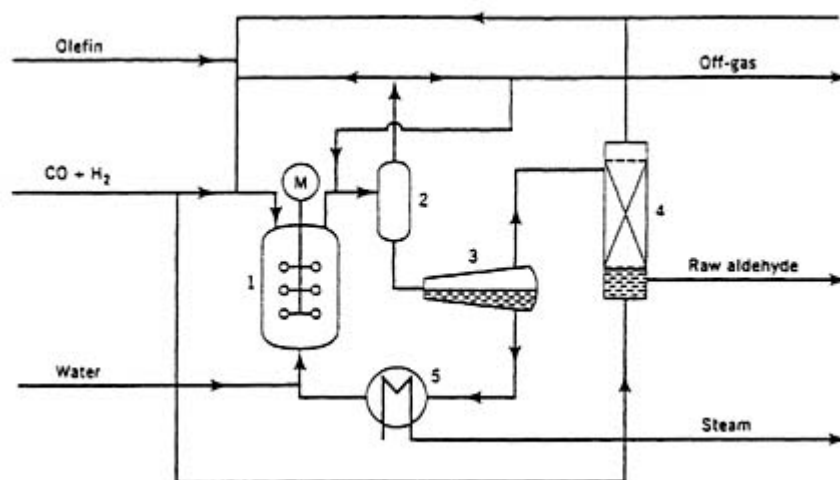


Figure C2.7.3. Process flow diagram for hydroformylation of propene; 1, reactor; 2, separator; 3, phase separator; 4, stripping column; 5, heat exchanger.

C2.7.5.2 SOLID CATALYSTS

Macroscopic properties often influence the performance of solid catalysts, which are used in reactors that may simply be tubes packed with catalyst in the form of particles—chosen because gases or liquids flow through a bed of them (usually continuously) with little resistance (little pressure drop). Catalysts in the form of honeycombs (monoliths) are used in automobile exhaust systems so that a stream of reactant gases flows with little resistance through the channels and heat from the exothermic reactions (e.g., CO oxidation to CO₂) is rapidly removed.

Efficient use of a catalyst requires high rates of reaction per unit volume and, since reaction takes place on the surface of a solid, catalysts have high surface areas per unit volume. Therefore, the typical catalyst is porous, with an internal surface area often exceeding 100 m² g⁻¹. Porous materials often consist of aggregates of nonporous (crystalline) microparticles, with the void spaces between them constituting a labyrinth of internal pores with

diameters roughly equal to those of the microparticles; reaction takes place on the microparticle surfaces. The pores may have average diameters <2.0 nm (micropores) because these imply high surface areas per unit volume, but larger pores are also common.

Because catalysts lose activity and need to be regenerated and replaced periodically, they must be robust enough to withstand these processes. Catalyst particles used in large reactors must be strong enough to resist crushing under the weight of the particles above them. Some catalyst particles are used in entrained and fluidized bed reactors, where they are in constant motion (for rapid heat transfer), and they must be resistant to abrasion. Many catalysts must withstand use at high temperatures (~ 800 K).

Physical properties affecting catalyst performance include the surface area, pore volume and pore size distribution ([section B1.26](#)). These properties regulate the tradeoff between the rate of the catalytic reaction on the internal surface and the rate of transport (e.g., by diffusion) of the reactant molecules into the pores and the product molecules out of the pores: the higher the internal area of the catalytic material per unit volume, the higher the rate of the reaction

-8-

per unit volume—up to a limit; beyond the limit, an increase in internal surface area requires such small microparticles (and hence such small pores) that the restrictions of the pores limit the rate of transport through them and lead to the existence of significant concentration gradients within the catalyst particles. Thus, the reactant concentration becomes less at the particle interior than at the edge, and the overall reaction rate is reduced and no longer proportional to the internal surface area. If internal area is gained at the expense of increases in pore volume, a limit is reached at which the material no longer meets the crush strength requirement.

Catalyst particles are usually cylindrical in shape because it is convenient and economical to form them by extrusion—like spaghetti. Other shapes may be dictated by the need to minimize the resistance to transport of reactants and products in the pores; thus, the goal may be to have a high ratio of external (peripheral) surface area to particle volume and to minimize the average distance from the outside surface to the particle centre, without having particles that are so small that the pressure drop of reactants flowing through the reactor will be excessive.

C2.7.5.3 CONSTITUENTS OF SOLID CATALYSTS

A solid catalyst is usually a composite, consisting of a material called a support or carrier (which often lacks catalytic activity) and other components, including those with catalytic activity, and perhaps still others, called promoters. The support is usually the principal component, sometimes being 99% or more of the catalyst mass. Thus, the physical properties are largely determined by the support. Supports are usually ceramic materials ([section C2.12](#)), the most common being transition aluminas such as γ - Al_2O_3 . Others include silica (SiO_2), carbon and zeolites. Transition aluminas offer the advantages of being inexpensive, robust, stable and formable into particles with wide ranges of shapes, internal surface areas and pore size distributions. The typical catalyst incorporates small amounts (e.g., 1 wt%) of catalytically active components (e.g., metals, metal oxides or metal sulphides) on the internal surface of the support. Since these components are often expensive, they are dispersed as small particles on the support surface; for example, particles of Pt on Al_2O_3 may be less than 1 nm in diameter, so small that almost all the Pt atoms are exposed at a surface where reactant molecules can bond to them and catalysis can occur. Supported metal catalysts have been applied for decades and are among the first *nanomaterials* ([section C1.2](#)) to find industrial applications.

The components in catalysts called promoters lack significant catalytic activity themselves, but they improve a catalyst by making it more active, selective, or stable. A *chemical promoter* is used in minute amounts (e.g., parts per million) and affects the chemistry of the catalysis by influencing or being part of the catalytic sites. A *textural (structural) promoter*, on the other hand, is used in massive amounts and usually plays a role such as stabilization of the catalyst, for instance, by reducing the tendency of the porous material to collapse or sinter and lose internal surface area, which is a mechanism of deactivation.

C2.7.6 EXAMPLES OF CATALYSIS

C2.7.6.1 WILKINSON HYDROGENATION OF ALKENES CATALYSED BY A RHODIUM COMPLEX

The hydrogenation of alkenes catalysed by organometallic compounds is exemplified by the Wilkinson catalyst of [figure C2.7.2](#) invented by the Nobel-prize-winning chemist Geoffrey Wilkinson. The cycle of [figure C2.7.2](#) elucidated by the group of Halpern [[4](#), [5](#) and [6](#)], has a number of characteristics that are so common that it can be regarded as

-9-

a prototype. The rhodium compounds (complexes) that react with the H_2 and alkene are coordinatively unsaturated and their reactions with these reactants lead to dissociation of the reactants (which become ligands bonded to the metal). The ligands that are formed, H and alkyl, combine with each other while bonded to the Rh, leading to the formation of the alkane product and regeneration of the catalyst (closure of the cycle). Thus, the role of the Rh centres in the Wilkinson catalyst is similar to that of the iron sites on the surface of the ammonia synthesis catalyst in that they both cause dissociation of the reactants and position the resultant ligands so that they combine with each other to give the product. This generalization is often valid for metals in catalysts, whether they are in compounds in solution, on metal surfaces or on surfaces of metal oxides or sulphides, for example.

There is more to the Wilkinson hydrogenation mechanism than the cycle itself; a number of species in the cycle are drained away by reaction to form species outside the cycle. Thus, for example, PPh_3 (Ph is phenyl) drains rhodium from the cycle and thus it inhibits the catalytic reaction (slows it down). However, PPh_3 plays another, essential role—it is part of the catalytically active species and, as an electron-donor ligand, it affects the reactivities of the intermediates in the cycle in such a way that they react rapidly and lead to catalysis. Thus, there is a tradeoff that implies an optimum ratio of PPh_3 to Rh.

When a strong electron-donor ligand such as pyridine is added to the reaction mixture, it can bond so strongly to the Rh that it essentially drains off all the Rh and shuts down the cycle; it is called a catalyst *poison*. A poison for many catalysts is CO; it works as a physiological poison in essentially the same way as it works as a catalyst poison: it bonds to the iron sites of haemoglobin in competition with O_2 .

The reactivities of the species within the Wilkinson cycle are so great that they are not observed directly during the catalytic reaction; rather, they are present in a delicate dynamic balance during the catalysis in concentrations too low to observe easily, and only the more stable species outside the cycle (outside the dashed line in [figure C2.7.2](#) are the ones observed. Obviously it was no simple matter to elucidate this cycle; the research required piecing it together from observations of kinetics and equilibria under conditions chosen so that sometimes the cycle proceeded slowly or not at all.

Techniques such as NMR spectroscopy ([section B1.12](#)) and IR spectroscopy ([section B1.2](#)) are useful in such experiments. Furthermore, theory ([section B3.1](#)) has proceeded to the point of being successful in predicting some simple catalytic cycles.

C2.7.6.2 CHIRAL HYDROGENATION OF ALKENES CATALYSED BY A RHODIUM COMPLEX

A hydrogenation reaction closely related to that referred to in the preceding paragraphs is that represented in [figure C2.7.4](#) [[7](#), [8](#)]. The catalyst incorporates a bidentate (two-toothed) phosphine ligand that affects the stereochemistry (chirality) of the reaction. There are two pathways for reaction ([figure C2.7.5](#)) giving products of different stereochemistries: the pathway shown on the left shows the preferred (more stable) mode of binding of the reactant with the catalyst; the pathway on the right involves a minor isomer of the reactant–catalyst complex. The chirality of the product is determined predominantly by the latter pathway, which gives the major product because the reactant–catalyst complex on the right is much more reactive than that on the left, even though it is formed in lower concentrations than that on the left.

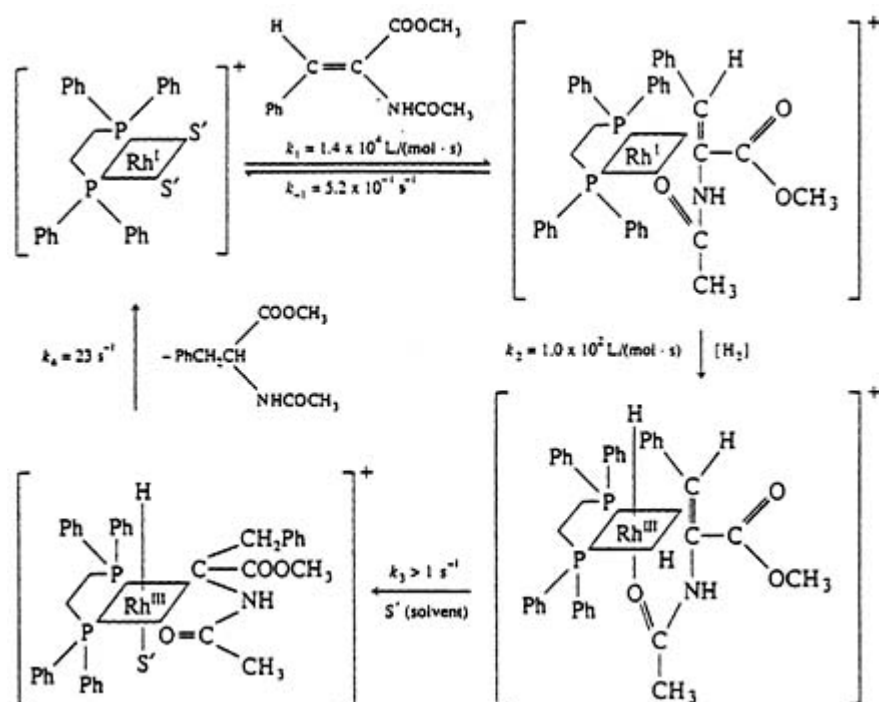


Figure C2.7.4. Catalytic cycle for hydrogenation of methyl-(*Z*)- α -acetamidocinnamate; the rate constants were measured at 298 K; S' is solvent [8].

This example illustrates a subtle control of a chemical reaction by a delicate manipulation of the stereochemical environment around a metal centre dictated by the selection of the ligands. This example hints at the subtlety of nature's catalysts, the enzymes, which are also typically stereochemically selective. Chiral catalysis is important in biology and in the manufacture of chemicals to regulate biological functions, i.e., pharmaceuticals.

C2.7.6.3 NO DECOMPOSITION CATALYSED BY A RUTHENIUM SURFACE

Scanning tunnelling microscopy (STM, [section B1.19](#)) has made it possible to observe microscopic details of catalyst surfaces, as shown by an investigation of one of the simplest solid catalysts, a single crystal of a pure metal (see [section A1.7](#)), Ru, which is active and selective for the decomposition of NO, which proceeds via dissociative chemisorption to give N and O atoms [9]. A clean Ru(0001) sample was prepared under ultrahigh vacuum and exposed to NO at 300 K, which dissociated completely on the surface. The distribution of N and O atoms on the surface was investigated by STM at 300 K, allowing the researchers to distinguish isolated N atoms (dark spots), islands of O atoms, and individual O atoms ([figure C2.7.6](#)); adsorbed NO molecules were not observed, as they move very rapidly on the surface [9]. Data were obtained at various times after the exposure of the crystal to NO.

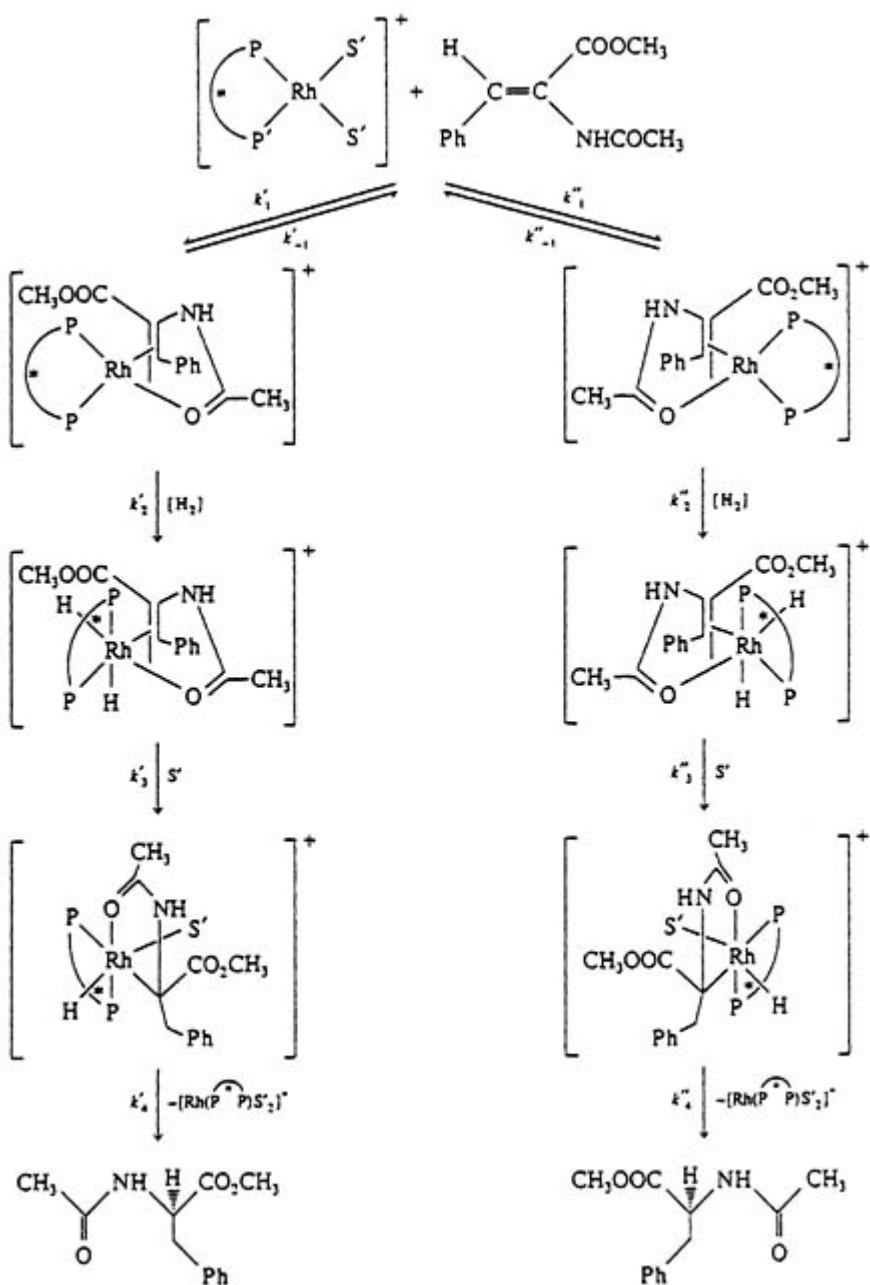


Figure C2.7.5. Pathways for the hydrogenation of methyl-(Z)- α -acetamidocinnamate catalysed by a rhodium complex with a chiral diphosphine ligand: S' is solvent [8].

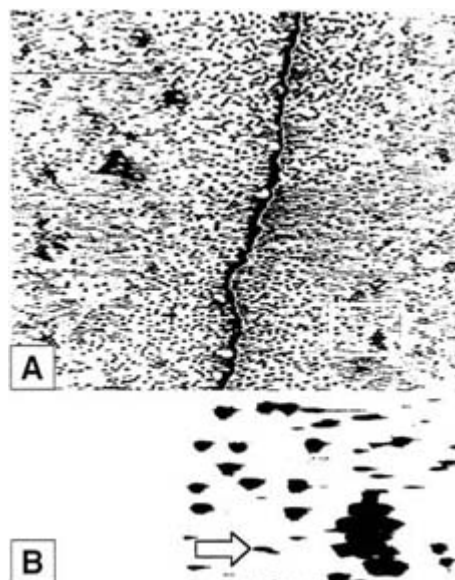


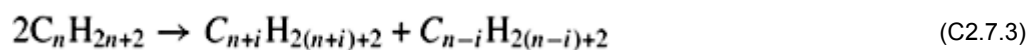
Figure C2.7.6. STM images of an Ru(0001) surface after dissociative adsorption of NO at 315 K. (A) Image (38 nm×33 nm) showing two terraces separated by a monatomic step (black stripe). (B) Close-up (6 nm×4 nm) showing an O island and individual N atoms. Individual O atoms are imaged as dashes (arrow) [9]

The Ru surface is one of the simplest known, but, like virtually all surfaces, it includes defects, evident as a step in figure C2.7.6. The observations show that the sites where the NO dissociates (active sites) are such steps. The evidence for this conclusion is the locations of the N and O atoms; there are gradients in the surface concentrations of these elements, indicating that the transport (diffusion) of the O atoms is more rapid than that of the N atoms; thus, the slow-moving N atoms are markers for the sites where the dissociation reaction must have occurred, where their surface concentrations are highest.

It has long been inferred that surface sites of low coordination (comparable to the metals in coordinatively unsaturated rhodium complexes in figure C2.7.2 are catalytically more active than those with larger numbers of neighbours and lower degrees of coordinative unsaturation, e.g., those on terraces on the Ru surface. Even more highly unsaturated are those sites at corners. The sites for one catalytic reaction may be different from those for another and, furthermore, the reactants may change the surface and create active sites. Active sites on surfaces more complex than those of single metal crystals are difficult to determine.

C2.7.6.4 PROPANE METATHESIS CATALYSED BY TANTALUM COMPLEXES ANCHORED TO SILICA

An alternative to elucidating the active sites on a surface is to synthesize them. For example, a new catalyst for metathesis of alkanes,



was synthesized on the surface of silica by the reaction of surface oxygen atoms with a reactive organometallic precursor, Ta(-CH₂CMe₃)₃(=CHCMe₃) (Me is methyl); two surface species were formed and identified by techniques including IR, NMR and EXAFS spectroscopies (section B1.6) [10]: (SiO-)Ta(-CH₂CMe₃)₂(=CHCMe₃) (~65%) and (SiO-)₂Ta(-CH₂CMe₃)₂(=CHCMe₃) (~35%). The authors suggested a catalytic cycle (although they lack the strongest evidence of the intermediates) (figure C2.7.7) in which the surface-bound Ta group is denoted with a subscript s. Thus, the catalysis is molecular and the silica just an enormous, rigid ligand. This anchoring helps to stabilize the tantalum complex in states of coordinative unsaturation that probably do not exist in solution—analogue complexes in solution could react with each other (leading to self-inhibition of catalysis).

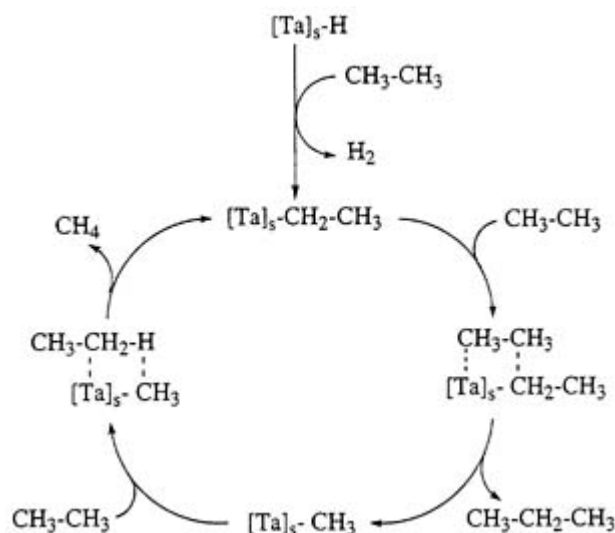


Figure C2.7.7. Catalytic cycle for the metathesis of propane [10]

C2.7.6.5 CATALYTIC ACTION OF THE ENZYME HALOALKANE DEHALOGENASE

An organism investigated for conversion of environmentally harmful halogenated compounds (e.g., dry-cleaning solvents) degrades them by dehalogenation catalysed by the enzyme haloalkane dehalogenase; 1-haloalkane and water react to give a primary alcohol and halide ion. The reaction mechanism was investigated by x-ray crystallography; enzyme crystals were soaked in dichloroethane to give a reactant–catalyst complex at 277 K and pH 5.0 (values less than those corresponding to the maximum reaction rate) and the complex was stable enough to allow an accurate structure determination [11]. Similar measurements were made with the sample warmed to room temperature and after standing long enough for the catalytic reaction to take place and the enzyme to become complexed with Cl^- product (but not 2-chloroethanol product).

The data led to the cycle shown in [figure C2.7.8](#). Here, only the active site on the interior enzyme surface ([section C2.6](#)) is depicted, consisting of R groups including aspartic acid, glutamic acid and others, represented with the shorthand Asp_{260} , Glu_{56} etc; the subscripts represent the positions on the polypeptide chain.

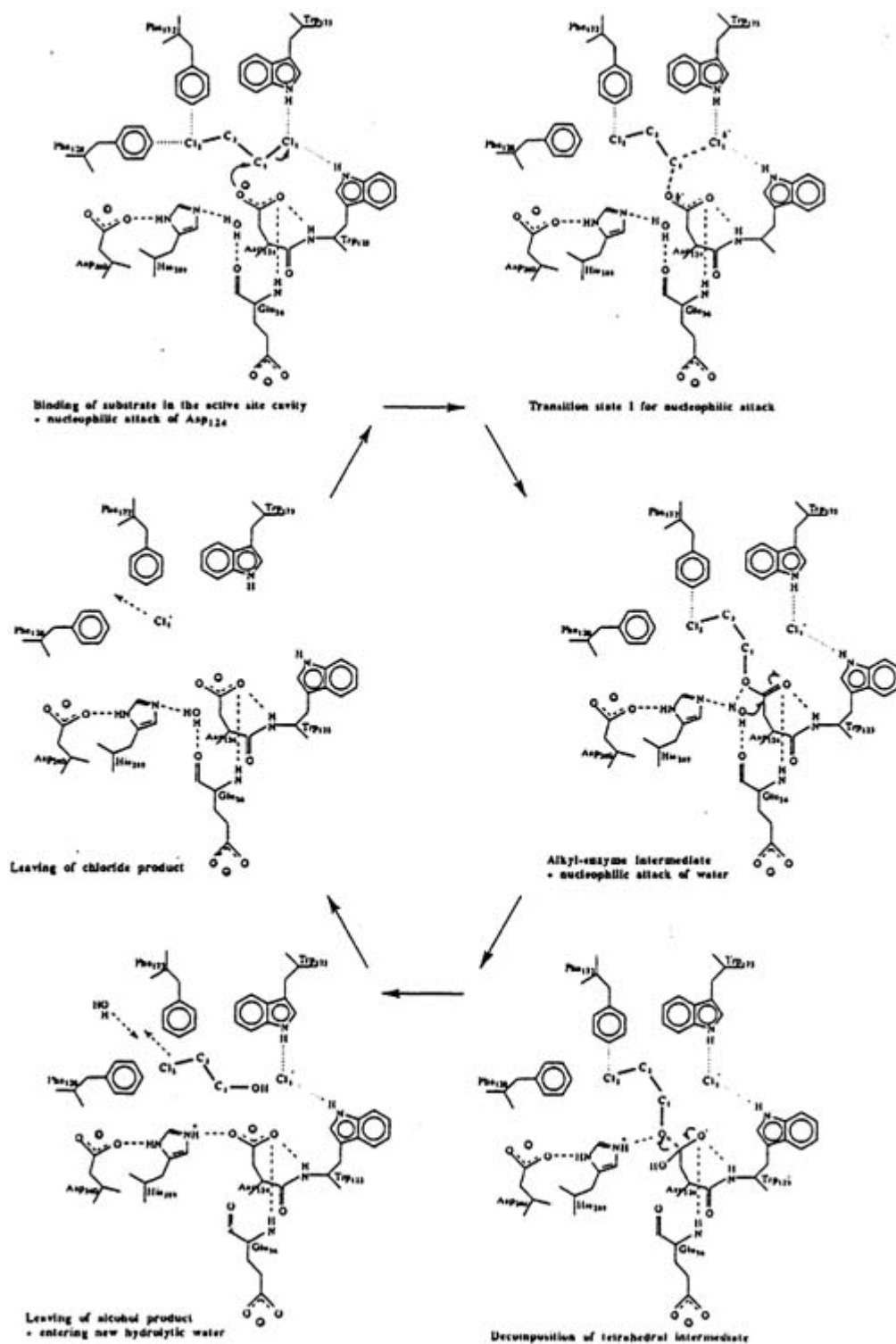


Figure C2.7.8. Catalytic cycle indicating the working of the enzyme haloalkane dehalogenase [11].

C2.7.6.6 CO OXIDATION CATALYSIS ON A PLATINUM SURFACE

In the preceding example, the structure of the catalyst combined with reactants and products was determined and the data were used to infer a cycle. Structures of the highly reactive intermediates in catalysis are generally elusive and information about them based only on inference. In prospect, the most incisive information about the workings of a catalyst can be obtained by observations of the catalyst in action. The following example illustrates this

strategy.

The CO oxidation occurring in automobile exhaust converters is one of the best understood catalytic reactions, taking place on Pt surfaces by dissociative chemisorption of O₂ to give O atoms and chemisorption of CO, which reacts with chemisorbed O to give CO₂, which is immediately released into the gas phase. Details are evident from STM observations focused on the reaction between adsorbed O and adsorbed CO [12].

Experiments were carried out with a Pt(111) single-crystal surface cleaned in an ultrahigh-vacuum system. To monitor the progress of the reaction as a function of time, the researchers brought O₂ into contact with the Pt to give a surface partially covered with O atoms, which aggregate into islands with a distinct periodic structure; the O atoms show up as dark dots in islands in the image of [figure C2.7.9](#) at time 0. At time 0, CO was introduced into the reactor and bonded with the surface, thereupon reacting with the chemisorbed O and reducing the sizes of the islands, as shown by the STM images at increasing times after the introduction of CO ([figure C2.7.9](#)). The CO molecules that must have been present on the surface at short times were not visible because of their high mobilities. But after 290 s, and more clearly after 600 s, the O islands had markedly shrunk because of reaction, and the CO had become evident as an additional ordered, streaky structure ([figure C2.7.9](#)); at these times the CO had formed into closely packed immobilized structures. After longer times, the O islands became smaller, indicating the progress of the catalytic reaction; by 2020 s, the O had been completely converted ([figure C2.7.9](#)). Note the contrast with the NO decomposition reaction; the CO oxidation takes place on the flat metal surface, not just on minority sites such as steps.

A striking feature of the images is the nonuniformity of the distribution of the adsorbed species. The reaction between O and CO takes place at the boundaries between the surface domains and it was possible to determine reaction rates by measuring the change in length L of the boundaries of the O islands. The kinetics is represented by the rate equation

$$r = -dn_{\text{O}}/dt = kL \quad (\text{C2.7.4})$$

where r is rate, n_{O} the number of surface O atoms, t time and k the rate constant. Data determined from the images are shown in [figure C2.7.10](#). Similarly, the rate was determined in the conventional way by macroscopic measurements of the surface coverages, assuming a random distribution of chemisorbed O and CO:

$$r = k'\theta_{\text{O}}\theta_{\text{CO}} \quad (\text{C2.7.5})$$

where k' is a rate constant. Data determined in this way are also shown in [figure C2.7.10](#). The agreement between the microscopic and macroscopic kinetics data provides a verification of the simplified representation of the conventional macroscopic kinetics, which works rather well, although the distribution of surface species is far from random.

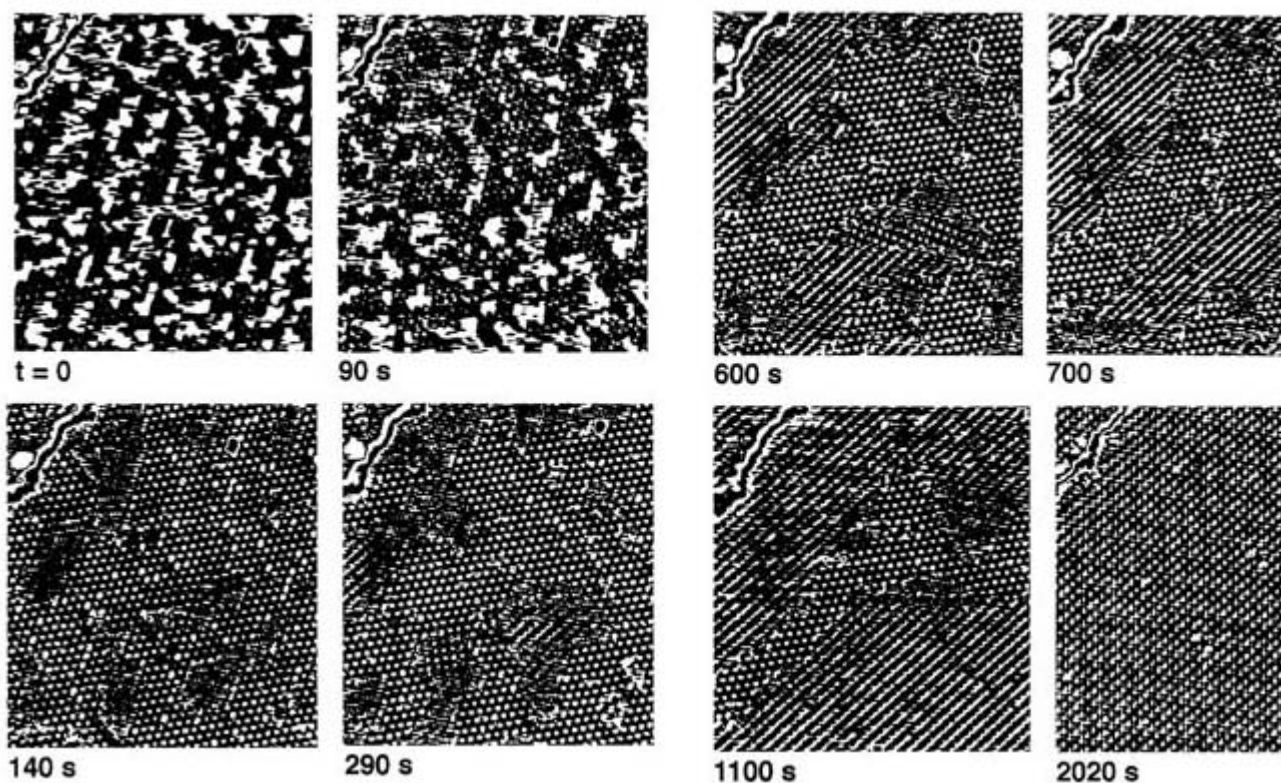


Figure C2.7.9. STM images recorded during reaction of adsorbed O atoms with adsorbed CO molecules on a Pt(111) crystal at 274 K; image size, 18 nm×17 nm. Times are those after addition of CO to the surface; see text for details [12].

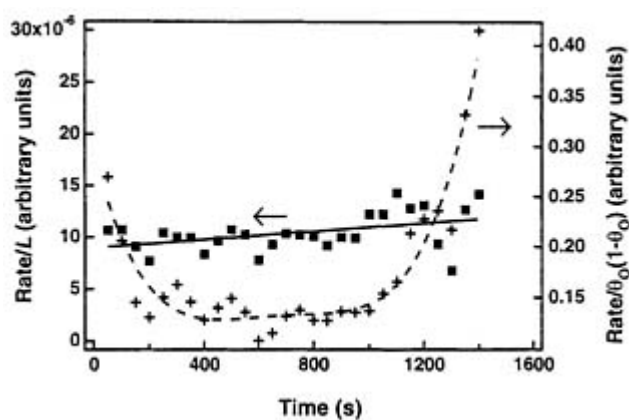


Figure C2.7.10. Rates of reaction of CO with O on Pt(111), determined from microscopic (■) and macroscopic (+) data; see text for details [12].

C2.7.6.7 SHAPE-SELECTIVE HYDROCARBON REACTIONS CATALYSED BY ZEOLITES

Zeolites (section C2.13) are unique because they have regular pores as part of their crystalline structures. The pores are so small (about 1 nm in diameter) that zeolites are *molecular sieves*, allowing small molecules to enter the pores, whereas larger ones are sieved out. The structures are built up of linked SiO₄ and AlO₄ tetrahedra that share O ions. The faujasites (zeolite X and zeolite Y) and ZSM-5 are important industrial catalysts. The structure of faujasite is represented in figure C2.7.11 and that of ZSM-5 in figure C2.7.12. The points of intersection of the lines represent Si or Al ions; oxygen is present at the centre of each line. This depiction emphasizes the zeolite framework structure and shows the presence of the intracrystalline pore structure. In the centre of the faujasite structure is an open space (supercage) with a diameter of about 1.2 nm. The pore structure is three dimensional,

with the supercages connected by apertures with diameters of about 0.74 nm. Molecules large enough to fit through the apertures can undergo catalytic reaction in the cages. ZSM-5 also has a three-dimensional structure, with straight parallel pores intersected by zig-zag pores.

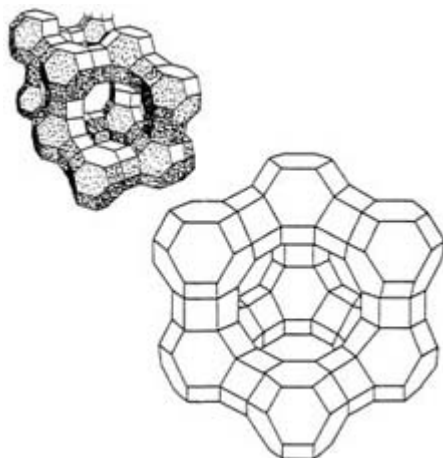


Figure C2.7.11. Framework structure of zeolites X and Y.

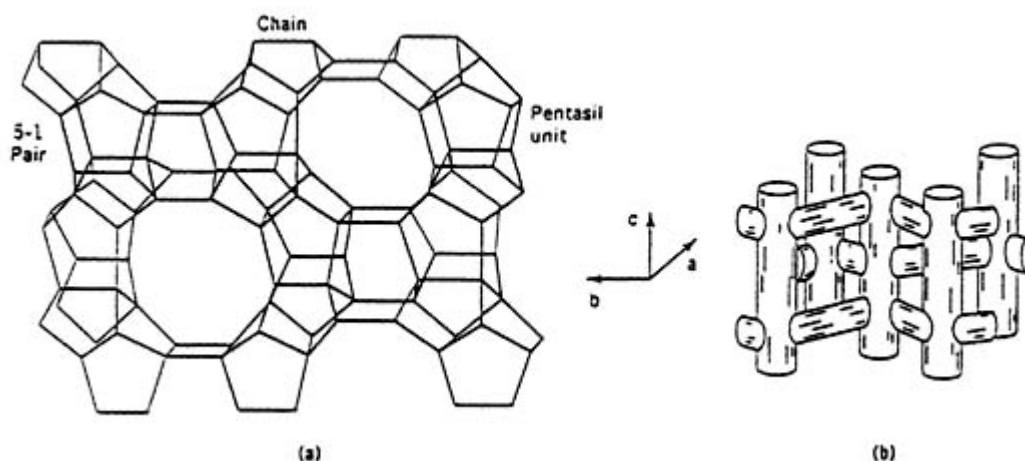


Figure C2.7.12. Structure of zeolite ZSM-5; (a) framework, (b) schematic representation of pores.

The zeolite frameworks are built up of SiO_4 tetrahedra, which are neutral, and AlO_4 tetrahedra, which have a charge of -1 . The charge of the AlO_4 tetrahedra is balanced by the charges of cations located at various crystallographically defined positions in the zeolite, many of them exposed at the internal surface. The cations are typically catalytically active sites. When the cations are H^+ (in OH groups), the zeolites are acidic. Acidic zeolite Y (HY) and HZSM-5 are applied as components of petroleum cracking catalysts to make gasoline. The OH groups located near AlO_4 tetrahedra are strong Brønsted acids and the catalytic sites for many reactions, including those mentioned below.

Zeolites are unique as shape-selective catalysts. *Mass transport shape selectivity* is a consequence of transport restrictions allowing some species to diffuse more rapidly than others in zeolite pores. Small molecules enter the pores and are catalytically converted, but larger molecules may pass through a flow reactor unconverted because they do not fit into the pores, where almost all the catalytic sites are located. Similarly, product molecules formed inside a zeolite may be so large that their diffusion out of the pores may be so slow that they are largely converted into other products before escaping into the product stream. Mass transport selectivity is illustrated by toluene disproportionation catalysed by HZSM-5 [13](#) (figure C2.7.13). The desired product is industrially valuable *p*-xylene.

The *ortho*- and *meta*-isomers are bulkier than the *para*-isomer and diffuse less readily in the zeolite pores. The transport restriction favours their conversion into the *para*-isomer, which is formed in excess of the equilibrium concentration. Because the selectivity is transport influenced, it is dependent on the path length for transport, which is the length of the zeolite crystallites.

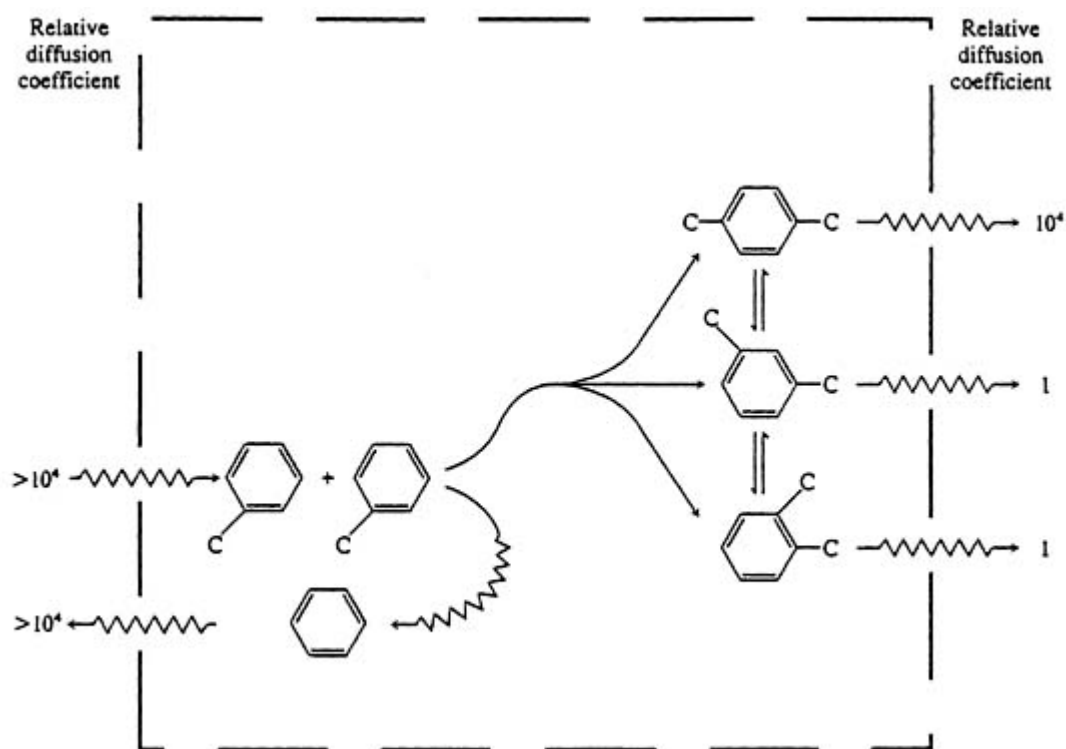


Figure C2.7.13. Schematic representation of diffusion and reaction in pores of HZSM-5 zeolite-catalysed toluene disproportionation; the numbers are approximate relative diffusion coefficients in the pores [13].

-19-

A different kind of shape selectivity is *restricted transition state shape selectivity*. It is related not to transport restrictions but instead to size restrictions of the catalyst pores, which hinder the formation of transition states that are too large to fit; thus reactions proceeding through smaller transition states are favoured. The catalytic activities for the cracking of hexanes to give smaller hydrocarbons, measured as first-order rate constants at 811 K and atmospheric pressure, were found to be the following for the reactions catalysed by crystallites of HZSM-5 14: *n*-hexane, 29; 3-methylpentane, 19; 2,2-dimethylbutane, 12 s^{-1} . The reaction rates were independent of the zeolite crystallite size, which rules out a transport effect. Instead, the selectivity is determined by the geometry of the transition states, which become bulkier with increasing branching of the molecules and are believed to be C_{12} carbenium ions resulting from the carbon-carbon bond formation reaction of the alkane with a carbenium ion formed from the alkane by abstraction of a hydride ion by another carbenium ion. These C_{12} carbenium ions easily fit in the zeolite pores when the reactant has no branches and barely fit when it has two branches; the crowding in the pores hinders the reaction.

REFERENCES

- [1] Ertl G 1980 *Surface Science and Catalysis: Proc. 7th Int. Congr. on Catalysis* part A, pp 21–35
- [2] Collman J P, Hegedus L S, Norton J R and Finke R G 1987 *Principles and Applications of Organotransition Metal Chemistry* (Mill Valley, CA: University Science)

- [3] Bach H, Gick W, Konkol W and Wiebus E 1988 The Ruhrchemie/Rhone-Poulenc (RHV/RP) process-latest variant of the fifty-year-old hydroformylation reaction *Proc. 9th Int. Congr. on Catalysis* vol 1, pp 254–9
- [4] Halpern J, Okamoto T and Zakhariiev A 1976 Mechanism of the chlorotris(triphenylphosphine)rhodium(I)-catalyzed hydrogenation of alkenes *J. Mol. Catal.* **2** 65–9
- [5] Halpern J and Wong C S 1973 Hydrogenation of tris(triphenylphosphine)chlororhodium(I) *Chem. Commun.* 629–30
- [6] Halpern J 1978 Mechanistic aspects of homogeneous catalysis *Trans. Am. Crystallogr. Assoc.* **14** 59–70
- [7] Halpern J, Riley D P, Chan A S C and Pluth J J 1977 Novel coordination chemistry and catalytic properties of cationic 1,2-bis(diphenylphosphino)ethanorhodium(I) complexes *J. Am. Chem. Soc.* **99** 8055–7
- [8] Halpern J 1982 Mechanism and stereoselectivity of asymmetric hydrogenation *Science* **217** 401–7
- [9] Zambelli T, Wintterlin J and Ertl G 1996 Identification of the 'active sites' of a surface-catalyzed reaction *Science* **273** 1688–90
- [10] Vidal V, Theolier A, Thivolle-Cazat and Basset J M 1997 Metathesis of alkanes catalyzed by silica-supported transition metal hydrides *Science* **276** 99–102
- [11] Verschueren K H G, Seljee F, Rozeboom J, Kalk K H and Dijkstra B W 1993 Crystallographic analysis of the catalytic mechanism of haloalkane dehalogenase *Nature* **363** 693–8
- [12] Wintterlin J, Völkening S, Janssens T V W, Zambelli T and Ertl G 1997 Atomic and macroscopic reaction rates of a surface-catalyzed reaction *Science* **278** 1931–4
- [13] Weisz P B 1981 Molecular shape selective catalysis *Proc. 7th Int. Congr. on Catalysis (Tokyo)* **1** 1
- [14] Haag W O, Lago R M and Weisz P B 1982 Transport and reactivity of hydrocarbon molecules in a shape-selective zeolite *Faraday Discuss. Chem. Soc.* **72** 317–30
-

FURTHER READING

Gates B C 1992 *Catalytic Chemistry* (New York: Wiley)

A textbook, intended for advanced undergraduates, attempting to state principles and demonstrate unity of catalysis as a field.

Fersht A R 1985 *Enzyme Structure and Mechanism* 2nd edn (New York: Freeman)

An introduction, with little physical chemistry.

Creighton T E 1984 *Proteins: Structure and Molecular Principles* (New York: Freeman)

Physical chemistry of enzymes.

Parshall G D and Iltel S D 1992 *Homogeneous Catalysis* 2nd edn (New York: Wiley)

A concise summary of chemistry of technologically important reactions catalysed by organometallic complexes in solution.

Cornils B and Herrmann W A (eds) 1996 *Applied Homogeneous Catalysis with Organometallic Compounds* (Weinheim: VCH)

A two-volume, multiauthored account with emphasis on industrial applications.

Ertl G, Knözinger H and Weitkamp J (eds) 1997 *Handbook of Heterogeneous Catalysis* (Weinheim: VCH)

A five-volume, multiauthored handbook giving principles, methods and applications.

Satterfield C N 1991 *Heterogeneous Catalysis in Industrial Practice* (New York: McGraw-Hill)

An introduction with an industrial flavour.

Thomas J M and Thomas W J 1996 *Principles and Practice of Heterogeneous Catalysis* (Weinheim: VCH)

Emphasis on characterization of solid catalysts with physical chemical methods.

Boudart M 1968 *Kinetics of Chemical Processes* (New York: Prentice-Hall)

Boudart M and Djega-Mariadassou G 1984 *Kinetics of Heterogeneous Catalytic Reactions* (Princeton, NJ: Princeton University Press)

Complementary books stating principles of kinetics of catalytic reactions.

Gates B C, Katzer J R and Schuit G C A 1979 *Chemistry of Catalytic Processes* (New York: McGraw-Hill)

Integrating chemistry and chemical engineering of industrial processes, homogeneous and heterogeneous catalysis.

Somorjai G A 1994 *Introduction to Surface Chemistry and Catalysis* (New York: Wiley)

Emphasis on ultrahigh-vacuum surface science as a foundation for understanding surface catalysis.

Advances in Catalysis continuing series (New York: Academic)

Monograph, multiauthored, approximately yearly.

C2.8 Corrosion

P Schmuki and M J Graham

C2.8.1 INTRODUCTION

Corrosion is the ‘gnawing away’ of materials due to exposure to different environments. Basically, a material is trying to return to its natural state, e.g., metallic iron oxidizes to form the ore from whence it came.

The most common form of corrosion (in terms of tons of materials lost) is electrochemical corrosion, which can occur for example in aqueous solutions, in the atmosphere and in the ground. Here, the actual corrosion reaction is invariably the anodic or oxidation reaction, whereby a metal dissolves while releasing electrons and ions. Thus one might say that ‘corrosion’ is a negative way of looking at an electrochemical dissolution or oxidation reaction. The reason for separating this topic from other dissolution or oxidation reactions which are of economic benefit, e.g. oxidation of silicon to form semiconductor devices, is based on the historic roots of corrosion science and the tremendous economic significance of material destruction. In industrialized countries, the cost of corrosion is estimated to be about 3.5% of the GNP. Areas such as construction materials, electronics and transportation are affected and thus an extensive number of reference books is available [[1](#), [2](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#), [18](#), [19](#), [20](#), [21](#), [22](#), [23](#), [24](#), [25](#), [26](#), [27](#), [28](#), [29](#), [30](#), [31](#), [32](#) and [33](#)]. Frequently, modes of corrosion are described according to the type of attack (e.g. uniform corrosion, localized corrosion) or the topic is categorized according to the specific material involved.

In the following, however, classification of corrosion processes is made according to the reaction mechanism rather than the phenomenology. Emphasis is placed on electrochemical reactions, including high temperature oxidation. Other types of corrosion such as purely physical processes (e.g. erosion, fretting) or mixed type (e.g. stress corrosion cracking) are only briefly mentioned, with reference to further reading, in [section C2.8.5](#).

C2.8.2 ELECTROCHEMICAL FUNDAMENTALS [34, 35, 36, 37 AND 38]

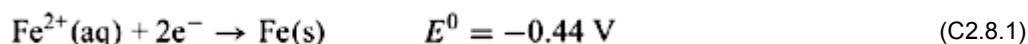
One can distinguish between a thermodynamic and kinetic stability to corrosion.

C2.8.2.1 THERMODYNAMIC CONSIDERATIONS

Thermodynamic stability is generally provided for noble metals in most media as their oxidation potential is more anodic than the reduction potential of species commonly occurring in the surrounding phase. However, for many materials of technological and industrial importance this is not the case.

For example, for iron in aqueous electrolytes, the thermodynamic warning of the likelihood of corrosion is given by comparing the standard electrode potential of the metal oxidation, with the potential of possible reduction reactions. The metal anodic oxidation reaction, $\text{Fe} \rightarrow \text{Fe}^{2+} + 2\text{e}^-$, can be written in the standard (reduction) notation as:

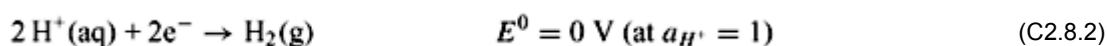
-2-



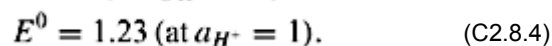
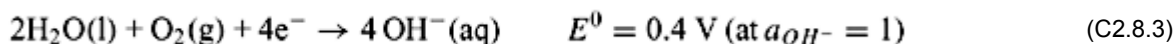
where E^0 denotes the standard redox potential of the reaction versus the normal hydrogen electrode (NHE).

Depending on the electrolyte pH, one of the following reduction (cathodic) reactions will dominate:

(a) in acidic solutions:



(b) in neutral and in alkaline solutions:



Both cathodic reactions can drive the metal oxidation. Of course, the potentials given above are only *standard* Gibbs values (E^0), and the effective electrode potential follows the Nernst equation (see [section C2.11](#)). For the oxidation (anodic) reaction, the potential (E_a) of the Nernst equation can be written as:

$$E_a = E_{\text{Fe}^{2+}/\text{Fe}}^0 + \frac{RT}{zF} \ln[\text{Fe}^{2+}]. \quad (\text{C2.8.5})$$

For the reduction reaction potential (E_c), the Nernst equations are:

(a)

$$E_c = E_{H^+/H}^0 + \frac{RT}{zF} \ln[H^+] \approx -0.059 \text{ pH} \quad (\text{C2.8.6})$$

(b)

$$E_c = E_{H^+/H}^0 + \frac{RT}{zF} \ln[H^+] \approx 1.23 \text{ V} - 0.059 \text{ pH}. \quad (\text{C2.8.7})$$

The right-hand term in both equations indicates the direct dependence of E_c on the pH of the medium (assuming otherwise standard conditions).

-3-

For the coupled redox cell, the e.m.f. (E) results as:

$$E = E_c - E_a \quad (\text{C2.8.8})$$

and the Gibbs free energy (ΔG) of the reaction as:

$$\Delta G = -zFE. \quad (\text{C2.8.9})$$

Thus, it can basically be predicted under what conditions (pH, concentration of redox species) the metal dissolution reaction ($\text{Fe} \rightarrow \text{Fe}^{2+}$) proceeds thermodynamically. From a practical point of view, the rate of the reaction and therefore the fate of the oxidized species (Fe^{2+}) is extremely important: they can either be solvated, i.e., to form $\text{Fe}(\text{H}_2\text{O})_6^{2+}$ complexes, and therefore be efficiently dissolved in the solution, or they can react with oxygen species of the solution to form a surface oxide layer (FeO , Fe_3O_4 , Fe_2O_3). Such oxide layers can represent effective kinetic barriers against corrosion (see [section C2.8.3](#) on passive films).

Thus it is important to take into account the thermodynamics of oxide formation and any additional electrochemical reactions such as the oxidation of the Fe^{2+} to Fe^{3+} . The results of the calculations are frequently represented as pH-potential diagrams (so-called Pourbaix diagrams [39]). The Pourbaix diagram for iron in an aqueous environment is shown in figure C2.8.1.

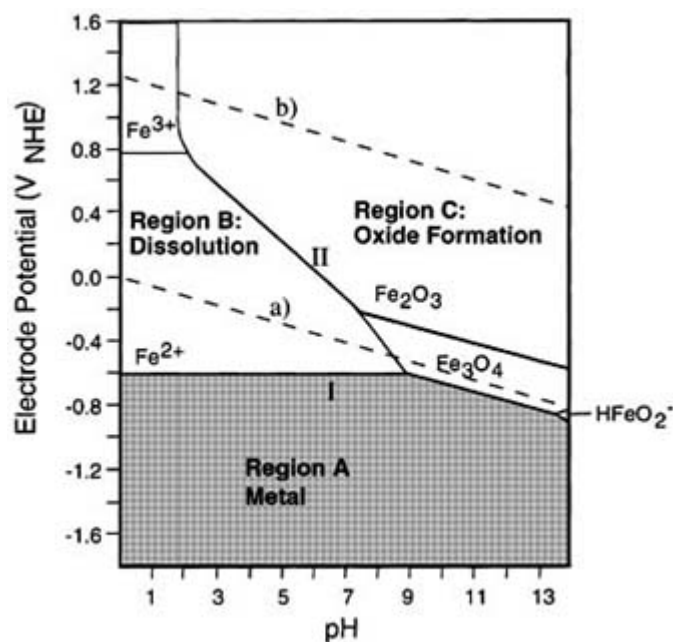


Figure C2.8.1. Simplified E/pH diagram (Pourbaix diagram) for the iron–water system at 25°C. The diagram is drawn for a concentration of dissolved Fe species of $10^{-6} \text{ mol l}^{-1}$. The potentials are given versus the normal hydrogen electrode (NHE) scale.

-4-

The diagram gives regions of existence, i.e. for a particular combination of pH and redox potential it can be predicted whether it is thermodynamically favourable for iron to be inert (stable) (region A), to actively dissolve (region B) or to form an oxide layer (region C).

The dotted lines represent the cases when the above cathodic reactions, (a) or (b), drive the reaction. The solid lines indicate the stability ranges for Fe and its corrosion products (Fe^{2+} , Fe^{3+} , Fe_3O_4 , Fe_2O_3 , HFeO_2^-).

Consider, for example, an acidic solution at pH 1: iron dissolves (formation of $\text{Fe}(\text{H}_2\text{O})_6^{2+}$); as E_c of reaction (a) is at -0.06 V NHE we are in region B (existence of Fe^{2+}). Additionally, it can be seen that the Fe^{2+} species can be further oxidized to Fe^{3+} if O_2 is present in the electrolyte (line (b) lies in the region of existence of Fe^{3+}).

Considering the case of $\text{pH} > 9$, the formation of an oxide film is favoured compared with Fe dissolution.

In the case of a neutral solution (e.g. $\text{pH} = 7$), depending on the corrosion potential all these three ranges (stability, dissolution or oxide formation) may be involved.

pH –potential diagrams are available for many elements in aqueous environments and are often a valuable tool in the preliminary assessment of the (thermodynamic) stability of a system. However, it should be pointed out that these calculations are based purely on thermodynamic considerations and, hence, this approach gives no information on the rate (kinetics) of the possible corrosion reactions.

C2.8.2.2 KINETIC CONSIDERATIONS

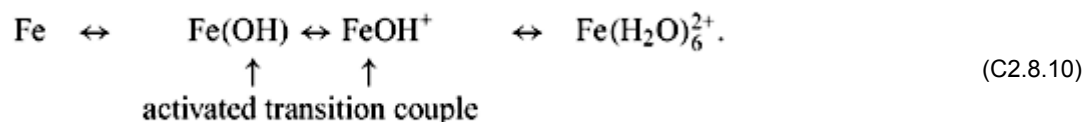
For many practically relevant material/environment combinations, thermodynamic stability is not provided, since $E_a > E_c$. Hence, a key consideration is how fast the corrosion reaction proceeds. As for other electrochemical reactions, a variety of factors can influence the rate determining step. In the most straightforward case the reaction is activation energy controlled; i.e. the ion transfer through the surface Helmholtz double layer involving migration and the adjustment of the hydration sphere to electron uptake or donation is rate determining. The transition state is

called an ‘activated surface complex’.

Alternatively, the mass transport properties in the solution can become rate determining—the reaction is then said to be diffusion controlled.

(A) ACTIVATION CONTROL (SEE ALSO SECTION C2.11 OF THIS BOOK)

Let us consider the oxidation of Fe(s) to Fe²⁺ (solvated), which can be described by the following reaction sequence [36, 40]:



-5-

The intermediate species Fe(OH) and FeOH⁺ can be regarded as constituting the activated surface complex.

The rate constant for formation and decay of this complex, *k*, can be written as

$$k = B e^{-\Delta G^\ddagger / RT} \quad (\text{C2.8.11})$$

As the reaction leading to the complex involves electron transfer it is clear that the activation energy ΔG^\ddagger for complex formation can be lowered or raised by an applied potential ($\Delta\Phi$). Of course, both the forward (oxidation) and well as the reverse (reduction) reaction are influenced by $\Delta\Phi$. If one expresses the reaction rate as a current flow (*j*), the above equation C2.8.11 can be expressed in terms of the Butler–Volmer equation (for a more detailed treatment see [section C2.11](#)). For the anodic reaction (Fe → Fe²⁺), the resulting anodic current density, *j_a*, upon applying externally an anodic voltage $\Delta\Phi$ has the form

$$j_a = j_0 e^{b\Delta\Phi} \quad (\text{C2.8.12})$$

For the cathodic reaction (Fe²⁺ → Fe), the cathodic current density *j_c* can analogously be written as

$$j_c = -j_0 e^{-b'\Delta\Phi} \quad (\text{C2.8.13})$$

where *j₀*, *b* and *b'* are constants. (*j₀* is the so-called exchange current density, i.e., the reaction current density in the absence of an external applied potential.)

Since any current resulting from the anodic reaction must be consumed by the cathodic reaction, the cathodic current, *j_c*, must be equal to the anodic current *j_a*. As a consequence, the equilibrium potential Φ_0 of a metal (e.g. Fe) that is immersed into an aqueous electrolyte will be adjusted by the condition that $j_a = |j_c| (=j_0)$. This is illustrated in [figure C2.8.2\(a\)](#). Under ideal conditions, $\Delta\Phi_0$ is the Nernst potential (E_{Fe}) of the Fe²⁺/Fe couple. For iron immersed in an aqueous electrolyte, the dominant reduction reaction is, however, one of the reactions involving the electrolyte (equations (C2.8.2), (C2.8.3) and (C2.8.4)). For this reaction equilibrium the same principle applies as outlined above for the Fe²⁺/Fe case (e.g. for H ↔ H₂ the concept of $j_a = j_c$ applies at E_H (an analogous diagram as for Fe in [figure C2.8.2 \(a\)](#) could be drawn for the H⁺/H couple).

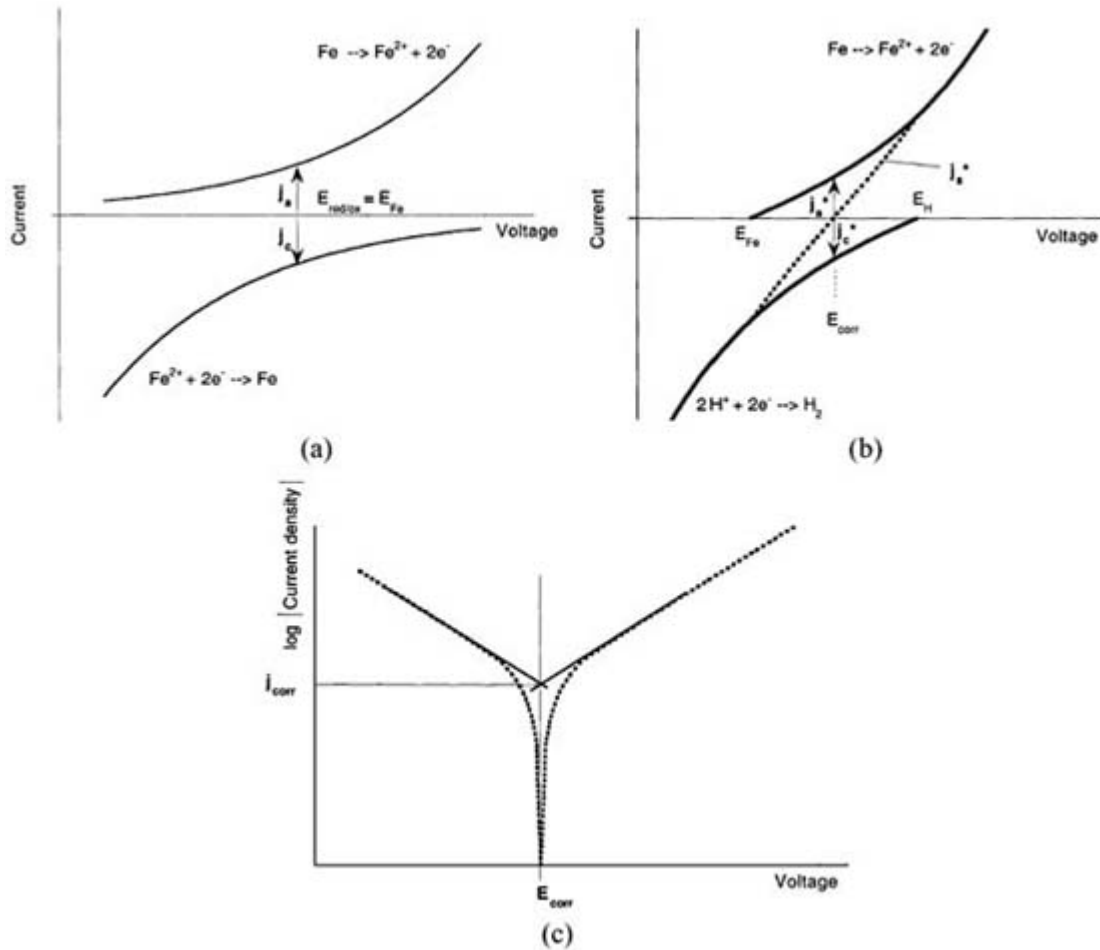


Figure C2.8.2. (a) Schematic polarization curves for the anodic and cathodic reaction of an Fe/Fe²⁺ electrode. The anodic and cathodic branches of the curve correspond to equations (C2.8.12) and (C2.8.13) respectively. The equilibrium potential of this electrode will adjust so that $|j_a| = |j_c|$; the corresponding potential is E_{redox} of Fe. (b) Schematic polarization curves for a mixed electrode (Fe in an aqueous solution in the presence of hydrogen ions), i.e. the redox system of Fe (figure 2(a)) is coupled with a second redox system: H⁺/H₂. Also in this mixed case the absolute values of the two partial current densities are equal ($|j_a^*| = |j_c^*|$) at equilibrium. The corresponding potential is the corrosion potential (E_{corr}); the corrosion current density is equal to the anodic current density at E_{corr} ($j_{corr}^* = j_a^*$). Experimentally, only the sum of the anodic and cathodic branch is accessible—the dotted line represents this sum of current density (j_s^*). (c) Determination of the corrosion current from a so-called Tafel plot ($\log(j_s^*)$ against U). The corrosion current density (j_{corr}^*) is obtained from the extrapolation of linear parts of the cathodic and anodic branches of j_s^* to the corrosion potential.

The coupled situation of both redox equilibria is described by the so-called ‘mixed potential theory’. The mixed oxidation (Fe → Fe²⁺) and reduction (H⁺ → H₂) systems will equilibrate to zero net current; the resulting potential, which lies between $E_{Fe^{2+}/Fe}$ and E_{H^+/H_2} , is called the corrosion potential (E_{corr}). This is illustrated in figure C2.8.2 (b) (which is obtained by ‘adding’ figure C2.8.2(a) and the analogous curves for the H⁺/H couple). The rate of corrosion is given by the current of metal ions leaving the metal surface in the anodic region. Thus, the corrosion current density, j_{corr} , can be identified with the anodic current of the coupled system, j_a^* .

As both processes, reduction and oxidation, take place on the same electrode surface (a short-circuited system), it is not possible to directly measure the corrosion current. Experimentally, only the sum of the anodic

current densities (j_s^*) is accessible (the dotted line in figures C2.8.2(b) and (c)). To obtain the corrosion rate, j_s^* can be measured as a function of an externally applied voltage. To acquire the polarization curves as in figure C2.8.2 (c), a traditional three-electrode set-up (figure C2.8.3) is mainly used where the system is compared with a reference electrode. The potential (voltage) between the metal and the reference electrode is varied using a counter-electrode and the current is registered (figure C2.8.3). From figure C2.8.2 (b) it is clear that, for potentials which are sufficiently far from the equilibrium value, $j_a^* \approx j_c^*$. Thus, in a plot of $\log(j_s)$ versus ΔU a linear portion is obtained which can be extrapolated back to E_{corr} as shown in figure C2.8.2(c). The corresponding current density value is j_{corr} . The semilog representation of figure C2.8.2 (c) is often referred to as a Tafel plot; the slope of the linear portions, which depends on the exact mechanism of the charge transfer reaction, is accordingly called the Tafel slope.

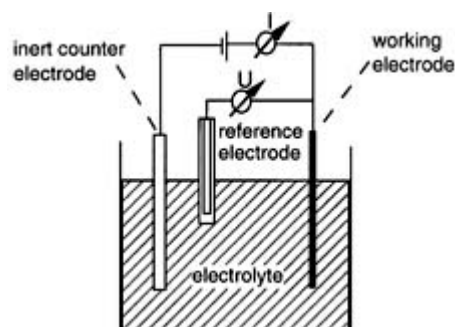


Figure C2.8.3. A three-electrode electrochemical set-up used for the measurement of polarization curves. A potentiostat is used to control the potential between the working electrode and a standard reference electrode. The current is measured and adjusted between an inert counter-electrode (typically Pt) and the working electrode.

The corrosion current can be converted into material loss (m_{corr}) using Faraday's law according to equation C2.8.14):

$$m_{corr} = (M/zF) j_{corr} t \quad (C2.8.14)$$

where M is the molar mass of the metal, z is the charge number of the ion, F is the Faraday constant and t is the time.

This, of course, assumes a 100% current efficiency regarding metal dissolution, i.e. no other competitive electrochemical reactions occur.

It should be pointed out that external polarization differs from the unbiased (open circuit) case in that after application of, say, an anodic voltage only the oxidation reaction takes place on the metal, whereas the cathodic reaction ($H \rightarrow H_2$) occurs at the external counter-electrode.

Other techniques to determine the corrosion rate use instead of DC biasing, an AC approach (electrochemical impedance spectroscopy). From the impedance spectra, the polarization resistance (R_p) of the system can be determined. The polarization resistance is indirectly proportional to j_{corr} . An advantage of an AC method is given by the fact that a small AC amplitude applied to a sample at the corrosion potential essentially does not remove the system from equilibrium.

(B) DIFFUSION CONTROL (SEE ALSO SECTION C2.11 OF THIS BOOK)

Electrochemical processes can become diffusion controlled if the formation of the activated complex is fast compared with the diffusion of the reacting anion to the surface or dissolving cations from the surface. In aqueous

solutions diffusion control of uniform corrosion is frequently encountered when the cathodic reaction depends on the supply of $O_2(g)$ —which is only sparingly soluble in water and therefore is present only in small concentrations.

Under diffusion controlled conditions the reaction rate depends, then, only on the supply of $O_2(g)$ to the surface which is determined by Fick's law:

$$\text{flux} = D(\partial N/\partial x) \tag{C2.8.15}$$

where D is the diffusion coefficient and $\partial N/\partial x$ is the particle (e.g. O_2) concentration gradient within the Nerstian diffusion layer.

Therefore, in the limiting case—the surface concentration of the reacting species is zero as all the arriving ions immediately react—the current density becomes voltage independent and depends only on diffusion, specifically, on the width of the Nerstian diffusion layer δ , and of course the diffusion coefficient and the bulk concentration of anions (c). The limiting current density (j_L) is then given by

$$j_L = (zFD/\delta)c. \tag{C2.8.16}$$

The diffusion layer width is very much dependent on the degree of agitation of the electrolyte. Thus, via the parameter δ , the hydrodynamics of the solution can be considered. Experimentally, defined hydrodynamic conditions are achieved by a rotating cylinder, disc or ring-disc electrodes, for which analytical solutions for the diffusion equation are available [37, 41, 42 and 43].

In the polarization curve of [figure C2.8.4](#) (solid line), the two regimes, activation control and diffusion control, are schematically shown. The anodic and cathodic plateau regions at high anodic and cathodic voltages, respectively, indicate diffusion control; the current is independent of the applied voltage and j_L is reached.

-9-

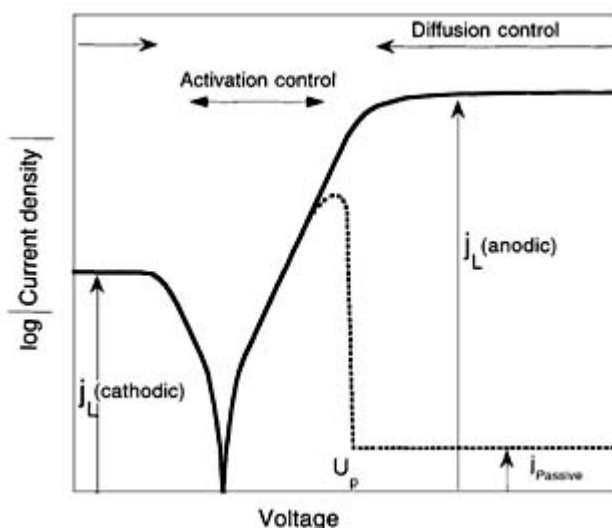


Figure C2.8.4. The solid line shows a typical semilogarithmic polarization curve ($\log j$ against U) for an active electrode. Different stages of reaction control are shown in the anodic and cathodic regimes: the linear slope according to an exponential law indicates activation control; at high anodic and cathodic potentials the current becomes independent of applied voltage, indicating diffusion control.

It is worth noting that under activation control the reaction rate depends on crystal orientation as the strength of the

bonds to the rest of the lattice, as well as the number of available bonds, directly influences the activation energy needed to create the activated complex. For instance, the kinetics of the dissolution of a Si(111) plane is much slower than for Si(100) due to stronger backbonding of an Si(111) surface atom. (An Si(111) surface is ‘attached’ by three backbonds to the lattice and has only one available (dangling) bond sticking out into the electrolyte, whereas an Si(100) surface has two backbonds and two of the bonds are dangling).

Under diffusion-controlled dissolution conditions (in the anodic direction) the crystal orientation has no influence on the reaction rate as only the mass transport conditions in the solution determine the process. In other words, the material is removed uniformly and electropolishing of the surface takes place.

C2.8.3 OXIDE FORMATION AND PASSIVITY [44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54 AND 55]

C2.8.3.1 PASSIVITY

(A) PASSIVE FILM FORMATION

In terms of an electrochemical treatment, passivation of a surface represents a significant deviation from ideal electrode behaviour. As mentioned above, for a metal immersed in an electrolyte, the conditions can be such as predicted by the Pourbaix diagram that formation of a second-phase film—usually an insoluble surface oxide film—is favoured compared with dissolution (solvation) of the oxidized anion. Depending on the quality of the oxide film, the formation of a surface layer can retard further dissolution and virtually stop it after some time. Such surface layers are called passive films. This type of film provides the comparably high chemical stability of many important construction materials such as aluminium or stainless steels.

Highly protective layers can also form in gaseous environments at ambient temperatures by a redox reaction similar to that in an aqueous electrolyte, i.e. by oxygen reduction combined with metal oxidation. The thickness of spontaneously formed oxide films is typically in the range of 1–3 nm, i.e., of similar thickness to electrochemical passive films. Substantially thicker anodic films can be formed on so-called valve metals (Ti, Ta, Zr, ...), which allow the application of anodizing potentials (high electric fields) without dielectric breakdown.

Passivation is manifested in a polarization curve ([figure C2.8.4](#) dashed line) by a dramatic decrease in current at a particular onset potential (the passivation potential, U_p). The corrosion reaction rate kinetics, i.e. the anodic current density, is lowered by several orders of magnitude.

The value and existence of a passivation potential U_p is based on the thermodynamics of oxide formation. Accordingly, passivation potentials and conditions for oxide formation can be predicted from a Pourbaix diagram ([figure C2.8.1](#)). A polarization curve as in [figure C2.8.4](#) can be perceived as reflecting a cross section through the Pourbaix diagram at a fixed pH. For example, at pH 7 one crosses, moving from cathodic to anodic potentials, first the active metal dissolution line (I), then the passivation line (II) at U_p .

It is generally believed that in the first stage of a passivation reaction—just below U_p —a precursor film is formed (e.g. a thin hydroxide layer), which then facilitates subsequent oxide formation.

The actual chemical mechanism of oxide formation has to address several factors, as schematically shown in [figure C2.8.5](#). Although in essence very similar, two slightly different mechanisms are distinguished: [figure C2.8.5 \(a\)](#) represents the growth of an oxide under open-circuit conditions, i.e. a piece of iron immersed in a passivating solution or exposed to an oxygen-containing environment. [Figure C2.8.5 \(b\)](#) shows the situation under an externally applied voltage in an electrolyte.

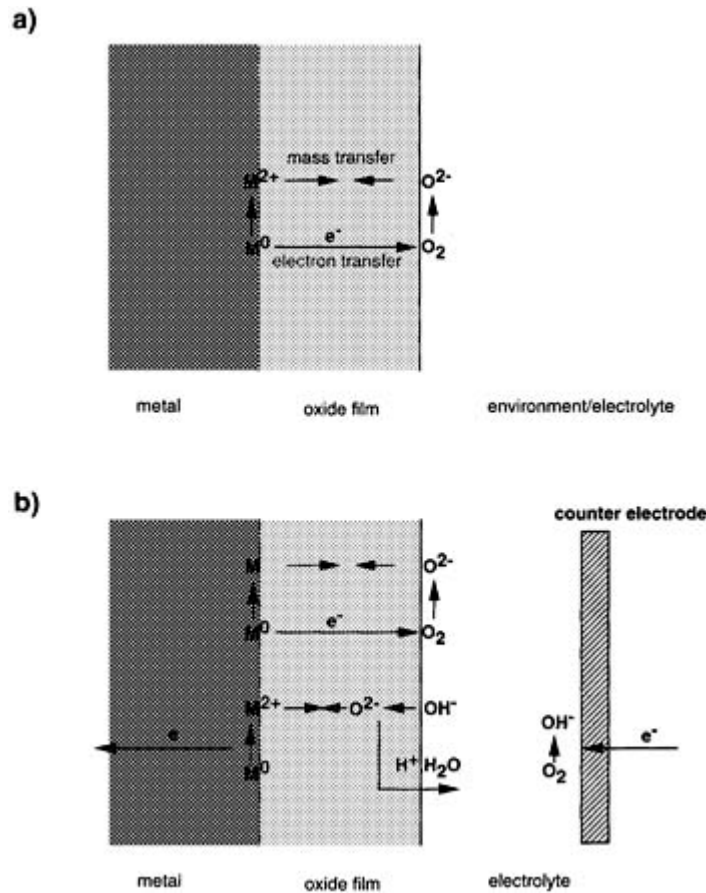


Figure C2.8.5. Growth of an oxide film on a metal surface. (a) In the absence of an externally applied potential: metal oxidation ($M \rightarrow M^{2+}$) occurring at the inner interface is coupled with oxygen reduction ($O_2 \rightarrow O^{2-}$) at the outer interface. For film growth one of the ionic species migrates predominantly—mass transfer is coupled with electron transfer through the layer. (This situation also corresponds to high-temperature corrosion.) (b) In the presence of an externally applied anodic potential (potentiostat): in addition to the mechanism of (a) film growth can also take place without electron transfer through the film as oxygen reduction happens at the counter-electrode. Mass transfer is not coupled with electron transfer.

In both cases, the anodic reaction occurs by oxidation at the metal/oxide interface:



For the situation in figure C2.8.5 a), the cathodic reaction is the reduction of O₂ at the oxide/gas or oxide/electrolyte interface:



At least one of the ionic species has to diffuse or migrate through the oxide and accordingly the layer grows usually either at the inner or outer interface. (Alternatively, the transport of ions through the film can be formulated as cation and anion vacancies moving through the lattice of the oxide film [44,56, 57].) As the cathodic and anodic

reactions are spatially separated by the oxide, electrons have also to be transferred through the layer and, thus, the conductance of the layer is essential to the process. Most oxides are semiconductors due to a non-equilibrated stoichiometry and, thus, either a negatively or positively charged species has the freedom to migrate through the lattice.

The driving force for migration is established by the different electrochemical potentials (ΔU) that exist at the two interfaces of the oxide. In other words, the electrochemical potential at the outer interface is controlled by the dominant redox species present in the electrolyte (e.g. O_2).

The situation in [figure C2.8.5\(b\)](#) is different in that, in addition to the mechanism in [figure C2.8.5\(a\)](#), reduction of the redox species can occur at the counter-electrode. Thus, electron transfer through the layer may not be needed, as film growth can occur with OH^- species present in the electrolyte involving a (field-aided) deprotonation of the film. The driving force is provided by the applied voltage, ΔU .

Quantitative approaches to describe the kinetics of film formation i.e. the mechanistic extraction of growth laws for film formation date back to the work of Cabrera and Mott [[58](#)] and Vetter [[59](#)]. It is essential that at low to moderate temperatures pure diffusion of ionic species is very small. Thus, film growth is controlled by the electric field across the layer and the lowering of the activation energy for ion or vacancy hopping (so-called high-field mechanism). This results in the so-called inverse logarithmic growth law.

$$1/x = A - B \log(t) \quad (C2.8.19)$$

where x is the film thickness and t is the time.

The growth according to this equation is self-limiting as the field strength F is lowered (at constant voltage) with an increasing film thickness x .

$$F(t) = \Delta U/x(t). \quad (C2.8.20)$$

In most practical cases (and at moderate voltages) the high-field growth law can control film growth, say up to only a maximum of 10 nm, as at this thickness the field strength effects become even less important than film growth due to diffusion of vacancies or ions.

The above rate law has been observed for many metals and alloys either anodically oxidized or exposed to oxidizing atmospheres at low to moderate temperatures—see e.g. [[60](#)]. It should be noted that a variety of different mechanisms of growth have been proposed (see e.g. [[61](#), [62](#)]) but they have in common that they result in either the inverse logarithmic or the direct logarithmic growth law. For many systems, the experimental data obtained up to now fit both growth laws equally well, and, hence, it is difficult to distinguish between them.

It should be mentioned that as well as for metals the passivation of semiconductors (particularly on Si, GaAs, InP) is also a subject of intense investigation. However, the goal is mostly not the suppression of corrosion but either the formation of a dielectric layer that can be exploited for devices (MIS structures) or the minimization of interface states (dangling bonds) on the semiconductor surface [[63](#), [64](#)].

(B) PROPERTIES OF PASSIVE FILMS

The protective quality of the passive film is determined by the ion transfer through the film as well as the stability of the film with respect to dissolution. The dissolution of passive oxide films can occur either chemically or electrochemically. The latter case takes place if an oxidized or reduced component of the passive film is more soluble in the electrolyte than the original component. An example of this is the oxidative dissolution of Cr_2O_3

films as CrO_4^{2-} [39, 65, 66].

From polarization curves the ‘protectiveness’ of a passive film in a certain environment can be estimated from the passive current density in [figure C2.8.4](#) which reflects the layer’s resistance to ion transport through the film, and chemical dissolution of the film. It is clear that a variety of factors can influence ion transport through the film, such as the film’s chemical composition, structure, number of grain boundaries and the extent of flaws and pores. The protectiveness and stability of passive films has, for instance, been based on percolation arguments [67, 68], structural arguments [69], ion/defect mobility [56, 57] and charge distribution [70, 71].

To illustrate some of the different approaches, let us consider passive films grown on Fe–Cr alloys. It has been established since 1911 [72] that an increase of Cr in the alloy increases the stability of the oxide film against dissolution.

The percolation argument is based on the idea that with an increasing Cr content an insoluble interlinked chromium oxide network can form which is also protective by embedding the otherwise soluble iron oxide species. As the threshold composition for a high stability of the oxide film is strongly influenced by solution chemistry and is different for different dissolution reactions [73], a comprehensive model, however, cannot be based solely on geometrical considerations but has in addition to consider the dissolution chemistry in a concrete way.

Other authors have attributed the improved corrosion resistance with increasing Cr content with the increasing tendency of the oxide to become more disordered [69]. This would then suggest that an amorphous oxide film is more protective than a crystalline one, due to a bond and structural flexibility in amorphous films.

This example illustrates that exact information on the chemistry and structure of the passive film is necessary to clarify the mechanisms relevant to stability and protectiveness of passive films.

The nature of passive oxide films on many technologically important metals and alloys has been the subject of investigation for many years. *Ex situ* surface analytical techniques such as x-ray photoelectron spectroscopy (XPS), Auger electron spectroscopy (AES) and secondary ion mass spectrometry (SIMS) provide useful information on the chemical composition and thickness of the films. Good agreement exists regarding a qualitative description of the chemistry of passive films on many metals. However, due to either different experimental approaches or data analysis, slightly different views can be found on the more detailed nature of the different films. Generally, it is important to note that the passive film, once formed should not be considered as a rigid layer, but instead as a system in dynamic equilibrium between film dissolution and growth. In other words, the passive film can adjust its composition and thickness to changing environmental factors. Principally, the chemical composition and the thickness of electrochemically formed passive films depend (apart from the base metal) on the passivation potential, time, electrolyte composition and temperature, i.e., on all passivation parameters and, hence, a detailed treatment is beyond the scope of this chapter. For further relevant literature the reader is referred to e.g. [74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88 and 89] and references therein.

The question of the structure of the passive film has been tackled by many research groups. Methods used to investigate the structure include x-ray scattering, diffraction and Mössbauer spectroscopy. For thick anodic oxide films or thick oxide films grown at elevated temperatures, the structure can be assessed by x-ray diffraction techniques. However, for thin passive films formed at low to moderate temperatures, the thickness of the films is usually less than 10 nm and, hence, it is experimentally difficult to investigate the structure by traditional x-ray diffraction. Another question often asked is whether the structure of a thin, mostly hydrated passive film formed under electrochemical conditions may change as it is removed from the conditions under which it was formed. Therefore, lately, new *in situ* techniques (STM, x-ray scattering using synchrotron radiation, EXAFS) to study the structure of thin oxide films have attracted considerable interest. In the case of the passive film on Fe, for instance, it could be shown with *in situ* STM [90] as well as with *in situ* x-ray scattering [91] that the passive film has a crystalline structure. Up to now, however, these investigations have been extended to only a few metals and, hence, the question of the structure of passive films remains to be investigated further.

As outlined above, electron transfer through the passive film can also be crucial for passivation and thus for the corrosion behaviour of a metal. Therefore, interest has grown in studies of the electronic properties of passive films. Many passive films are of a semiconductive nature [92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102 and 103] and therefore can be investigated with techniques borrowed from semiconductor electrochemistry—most typically photoelectrochemistry and capacitance measurements of the Mott–Schottky type [104]. Generally it is found that many passive films cannot be described as ideal but rather as amorphous or highly defective semiconductors which often exhibit doping levels close to degeneracy [105].

(C) PASSIVITY BREAKDOWN AND LOCALIZED CORROSION

The passive state of a metal can, under certain circumstances, be prone to localized instabilities. Most investigated is the case of localized dissolution events on oxide-passivated surfaces [51, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117 and 118]. The essence of localized corrosion is that distinct anodic sites on the surface can be identified where the metal oxidation reaction (e.g. $\text{Fe} \rightarrow \text{Fe}^{2+} + 2\text{e}^-$) dominates, surrounded by a cathodic zone where the reduction reaction takes place (e.g. $2\text{H}^+ + 2\text{e}^- \rightarrow \text{H}_2$). The result is the formation of an active pit in the metal, an example of which is illustrated in figure C2.8.6(a) and (b).

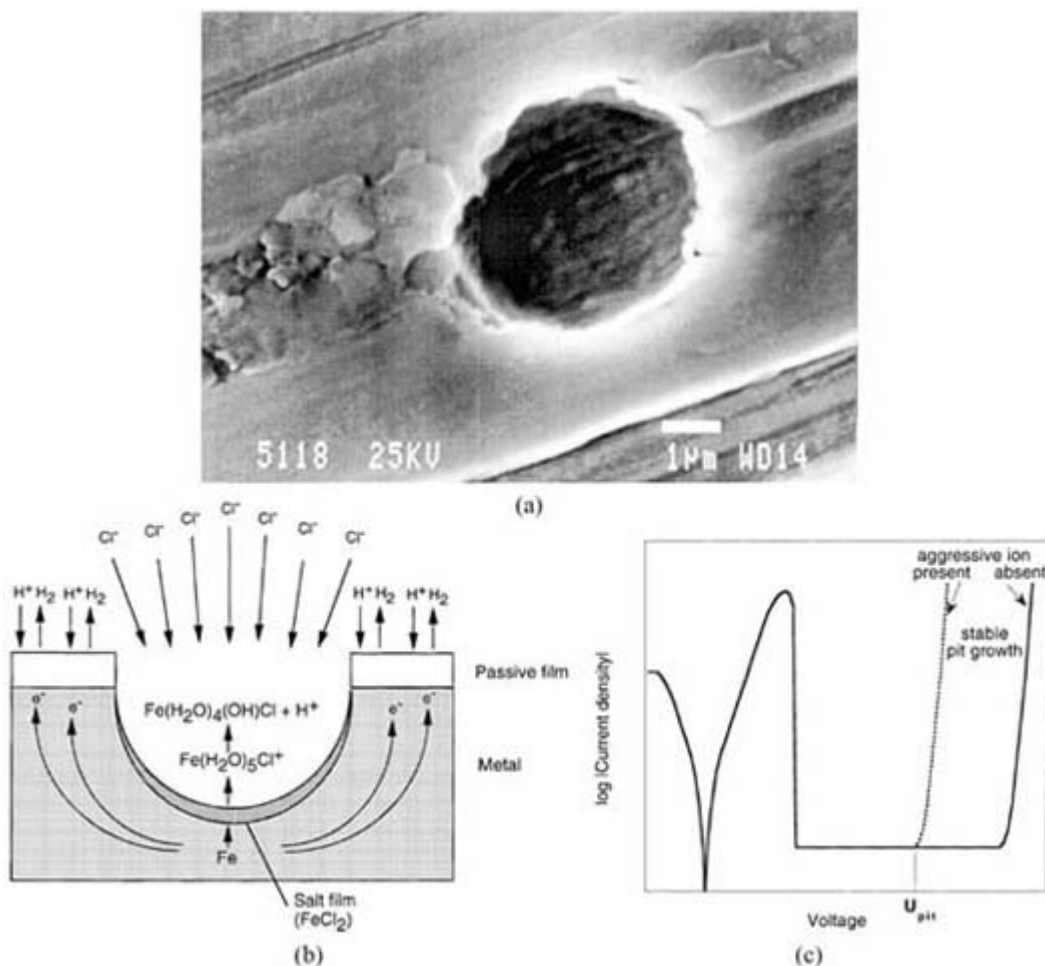


Figure C2.8.6. (a) SEM image of an early stage of pit formation on AISI304 steel in chloride containing solution. (b) Schematic cross section through an actively growing pit. Metal oxidation occurs at the pit base and the corresponding reduction on the passive film surrounding the pit. Acidification and an increase of the halogen ion concentration (due to migration) within the pit additionally accelerate dissolution. (c) Polarization curve of a passive metal showing localized breakdown of passivity and pit growth at U_{pit} (dashed line). The solid line represents a polarization curve of the same material in the absence of aggressive (pit-triggering) anions. In this case the current increase at higher anodic voltages indicates either transpassive oxide film dissolution or the onset of

oxygen evolution at the anodically polarized electrode.

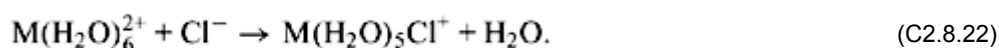
Pitting occurs with many metals in halide containing solutions. Typical examples of metallic materials prone to pitting corrosion are Fe, stainless steels and Al. The process is autocatalytic, i.e., by initial dissolution, conditions are established which further stimulate dissolution: inside the pit the metal (Fe in the example of figure C2.8.6 dissolves.

-16-

The M^{2+} species form an aquo-complex:



To maintain charge neutrality, additional halide ions (Cl^- in our example) have to migrate inside the pit thus increasing the local chloride concentration and a chloro-complex is formed.



The chloro-complex is in equilibrium with its hydroxo-chlorocomplex and H^+ .



For many metals the equilibrium lies strongly to the right hand side. Thus, within the pit the chloride concentration and the H^+ concentration both increase, further accelerating metal dissolution.

Generally, the following two stages of pitting are distinguished: pit initiation and pit growth. The reasons for the initiation of pits at distinct surface locations are manifold and can either be deterministic or stochastic in nature. They can be ascribed to bulk metal inhomogeneities (inclusions, precipitates, grain boundaries, dislocations etc.) or to properties of the passive film (thermally induced stochastic film rupture, electrostriction, local composition or structure variations). Initiation mechanisms assigning the key role to the passive film involve Cl^- penetration, local film thinning, or vacancy condensation; mechanisms focusing on the bulk metal ascribe the key role to preferential dissolution at inhomogeneities.

In an electrochemical polarization experiment on a passive system the onset of localized dissolution can be detected by a steep current increase at a very distinct anodic potential (the pitting potential, U_{pit})—see [figure C2.8.6\(c\)](#). This increase occurs far below either transpassive dissolution (oxide film dissolution due to the formation of soluble higher oxidation states (e.g. $Cr_2O_3 \rightarrow Cr^{6+} aq$) or the occurrence of oxygen evolution ($OH^- \rightarrow O_2$)).

In the potential range anodic to U_{pit} , stable pit growth occurs. The value of U_{pit} is shifted to lower anodic potentials with increasing temperature, increasing Cl^- concentration and decreasing pH, and is dependent on the presence of other anions in the electrolyte.

From an electrochemical viewpoint, stable pit growth is maintained as long as the local environment within the pit keeps the pit under active conditions. Thus, the effective potential at the pit base must be less anodic than the passivation potential (U_p) of the metal in the pit electrolyte. This may require the presence of voltage-drop (IR -drop) elements. In this respect the most important factor appears to be the formation of a salt film at the pit base. (The salt film forms because the solubility limit of e.g. $FeCl_2$ is exceeded in the vicinity of the dissolving surface in the highly Cl^- -concentrated electrolyte.)

In the potential range cathodic to U_{pit} , one frequently observes so-called metastable pitting. A number of pit growth events are initiated, but the pits immediately repassivate (an oxide film is formed in the pit) because the conditions within the pit are such that no stable pit growth can be maintained. This results in a polarization curve with strong current oscillations at $U < U_{pit}$.

Another type of localized corrosion closely related to pitting corrosion is crevice corrosion. This type of attack occurs preferentially in regions on the metal surface where mass transfer is limited (e.g., in narrow crevices or under deposits) and, hence, an increase in concentration of aggressive species (halides), combined with a pH decrease as discussed above and depletion of oxygen, can rapidly lead to activation of the surface in the crevice area. Metals which are susceptible to pitting corrosion also suffer from crevice corrosion. The presence of crevices on the surface often triggers localized corrosion already under conditions where stable pitting would not take place (e.g., with lower concentration of aggressive halides).

In all cases of localized corrosion, the ratio of the cathodic to the anodic area plays a major role in the localized dissolution rate. A large cathodic area provides high cathodic currents and, due to electroneutrality requirements, the small anodic area must provide a high anodic current. Hence, the local current density, i.e., local corrosion rate, becomes higher with a larger cathode/anode-ratio.

Localized corrosion is far more treacherous in nature and far less readily predictable and controllable than uniform corrosion and it is, moreover, capable of leading to unexpected damage with disastrous consequences, especially since inspection of corrosion damage is in many cases difficult.

Recently, the phenomenon of localized dissolution has attracted a great deal of interest in the field of semiconductor technology. This is due to the discovery of visible light emission from porous Si [119] which is formed by an electrochemical treatment of a Si surface in an HF-containing electrolyte [120]. It is interesting to note that the formation process is in many respects similar to pitting of metals [118] and that preferential triggering of the formation process at defects can be exploited to form highly defined localized dissolution [121].

C2.8.3.2 HIGH-TEMPERATURE OXIDATION AND CORROSION [122, 123, 124, 125, 126 AND 127]

So far, discussions have been limited to oxide film growth at low temperatures, where the model of Cabrera and Mott [58] usually applies. Oxide growth, controlled by the electric field across the film, follows an inverse logarithmic growth law (equation (C2.8.19)). At elevated temperature, scales can grow much thicker in water, air or oxygen, for example, or other more aggressive gases containing sulphur or chlorine. Mechanistically, the processes are similar to the passivation discussed earlier in terms of oxidation, reduction, ion transport and electron transfer, as outlined in figure C2.8.5(a). The main difference is that elevated temperatures promote ionic diffusion and, thus, oxide formation can proceed to a much greater extent than at low temperatures where only thin layers are formed by the high-field mechanism. The most common growth law observed at higher temperatures is the so-called parabolic rate law [128]:

$$x^2 = k_p t + C \quad (\text{C2.8.24})$$

where the rate of oxidation dx/dt is inversely proportional to the oxide thickness, x , and k_p is the parabolic rate constant. This indicates that a thermal diffusion process is rate controlling, with oxygen or cations or both diffusing

through a compact layer (figure C2.8.5(a)). Protective oxides are formed in this manner, the most important in practice being Cr_2O_3 , Al_2O_3 and SiO_2 . SiO_2 layers are particularly protective, but it is difficult to form continuous silica layers on silicon-containing steels. It is, however, straightforward to produce SiO_2 on silicon single crystals,

and the ability for silicon to form high-quality (amorphous) thermal oxide is to a large extent responsible for its successful application in MOS (metal/oxide/semiconductor) technologies [129].

The rate of diffusion through an oxide film depends on a number of factors, such as the temperature, oxygen partial pressure and structure of the oxide. At high temperatures (>0.7 of the melting point of the metal) lattice diffusion dominates through the crystalline oxide formed on a metal. However, at moderate temperatures diffusion via oxide grain boundaries is predominant. In this case, the rate of oxidation of a metal or alloy depends on the oxide grain size, which is often dictated by substrate grain orientation, surface pretreatment etc [130]. Deviation from parabolic oxidation behaviour is often observed and can be the result of the oxide grain size changing with time at a particular temperature. In this case, the number of oxide grain boundary 'easy diffusion paths' decreases with time, causing an apparent decrease in oxidation rate. If true parabolic behaviour is observed, then the change in oxidation rate with temperature will follow an Arrhenius equation; $k_p = Ae^{-\Delta E/PT}$. The log of the parabolic rate constant $\ln k_p$ is proportional to $1/T$, where T is the absolute temperature. The slope of such a plot yields the activation energy, ΔE , for oxidation.

If a compact film growing at a parabolic rate breaks down in some way, which results in a non-protective oxide layer, then the rate of reaction dramatically increases to one which is linear. This combination of parabolic and linear oxidation can be termed parilinear oxidation. If a non-protective, e.g. porous oxide, is formed from the start of oxidation, then the rate of oxidation will again be linear, as rapid transport of oxygen through the porous oxide layer to the metal surface occurs. Figure C2.8.7 shows the various growth laws. Parabolic behaviour is desirable whereas linear or 'breakaway' oxidation is often catastrophic for high-temperature materials.

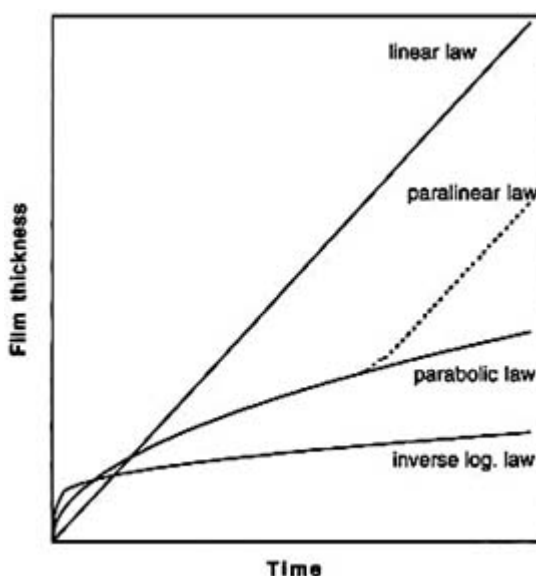


Figure C2.8.7. Principal oxide growth rate laws for low- and high-temperature oxidation: inverse logarithmic, linear, parilinear and parabolic.

High-temperature protective oxides often fail as the result of the development of stress during growth, which causes cracking and rupture of the scales leading to much faster metal degradation. Mechanisms for intrinsic growth stresses include the epitaxial relationship between the metal and scale [131], the volume change that occurs when a metal is converted into oxide by oxidation (the Pilling–Bedworth ratio) [132], compositional changes in either the metal or oxide and the influence of vacancies generated during oxidation [133]. Many discussions of the mechanisms of growth stresses have been published, e.g. [134, 135].

C2.8.4 SUPPRESSION OF CORROSION

C2.8.4.1 ELECTROCHEMICAL PROTECTION

Based on the polarization curves of [figure C2.8.4](#) there are several possibilities for reducing or suppressing the corrosion reaction. The main idea behind every case is to shift the corroding anode potential away from E_{corr} . This can be done in the following ways.

(i) *Cathodic protection* [[136](#), [137](#) and [138](#)]. By imposing a negative external voltage, the potential can be shifted cathodic to the corrosion potential (or better negative to the oxidation potential $E_{Fe/Fe^{2+}}$, to achieve thermodynamic stability).

Cathodic protection can also be achieved without the application of an external voltage by coupling with a less noble metal. The coupled metal has an E_{redox} that is negative to the material to be protected and thus becomes the anode which corrodes. The anode is therefore termed sacrificial.

(ii) *Anodic protection* [[139](#)]. If the material exhibits a passive behaviour (dotted line in [figure C2.8.4](#)), the potential of the corroding material can be anodically shifted into the passive range. This can be achieved by imposing a positive external voltage. The remaining corrosion current then depends on the quality of the passive film.

For practical applicability, several aspects have to be considered such as the anode material (sacrificial (e.g. zinc) or inert (e.g. Pt/Ti or graphite)), the conductivity of the medium and the current distribution. Cathodic protection is typically used for buried constructions (e.g. pipelines), off-shore structures or ship hulls.

C2.8.4.2 CHEMICAL INHIBITION [[140](#), [141](#), [142](#), [143](#), [144](#) AND [145](#)]

Corrosion suppression by inhibitors can be achieved by adding chemical species to the environment, which lead to a strong reduction of the dissolution rate. Depending on their specific action, corrosion inhibitors can be divided into the following groups.

(i) *Oxidizing inhibitors*. The idea is essentially the same as with anodic protection. If the material shows a passive range, the corrosion potential can be shifted into the passivity region by adding an additional redox couple to the electrolyte that possesses an E_{red} at potentials in the passive regime (oxidative inhibition).

-20-

The effectiveness of anodic methods can vary considerably and is mainly determined by the protective nature of the passive layer formed.

(ii) *pH modifiers*. The primary effect of pH modification can be deduced from the Pourbaix diagram ([figure C2.8.1](#)). The goal is to shift the pH into a region where thermodynamically, the formation of a passive layer is favoured as opposed to active dissolution. Often pH buffers are used, which keep the pH stable and thus hamper acidification, as discussed in the localized corrosion section.

(iii) *Surface blockers*. Type 1: the inhibiting molecules set up a geometrical barrier on the surface (mostly by adsorption) such as a variety of ionic organic molecules. The effectiveness is directly related to the surface coverage. The effect is a lowering of the anodic part of the polarization curve without changing the Tafel slope.

Type 2: the inhibiting species takes part in the redox reaction, i.e. it is able to react at either cathodic or anodic surface sites to electroplate, precipitate or electropolymerize. Depending on its 'activation' potential, the inhibitor affects the polarization curve by lowering the anodic or cathodic Tafel slope.

C2.8.4.3 SURFACE TREATMENTS [[137](#), [146](#), [147](#)]

For corrosion protection a large number of different types of surface treatment and coating have been developed, ranging from inorganic enamel coatings to organic coatings. In the following, the main two types of coating are

briefly discussed.

(i) *Paint and insulating coatings.* Coating the surface with some impermeable layer, such as paint, is frequently used due to the ease of application. The protection mechanism is simply to provide a physical barrier against metal dissolution. Unfortunately, this protection can fail disastrously if the coating is defective. Therefore, in many practical applications, a combination of an insulating coating and cathodic protection is employed.

(ii) *Deposition of a less noble metal (mostly by galvanizing).* The principle is again identical to cathodic protection. The coating has an E_{redox} that is negative to the material to be protected and the layer serves as a sacrificial anode. Therefore, this type of coating is not sensitive to defects, pinholes or mechanical damage during service. A typical example is galvanized steel (Zn layer on steel).

C2.8.4.4 PROTECTION AGAINST HIGH-TEMPERATURE CORROSION

Increasing operating temperatures result in increasing rates of corrosion and protective coatings are used to enhance component performance. These coatings serve as effective diffusion barriers between the oxidizing environment and the base alloy. In practice, corrosion-resistant coatings usually produce protective oxide scales consisting of thermally formed Cr_2O_3 , Al_2O_3 or SiO_2 films. The coating and the substrate should preferably have closely matched thermal expansion coefficients to prevent cracking during thermal cycling; the coating should also be able to withstand damage from impacts, erosion and abrasion.

-21-

C2.8.5 BRIEF OVERVIEW OF OTHER SPECIFIC CASES OF CORROSION

In the following, the most typical modes of corrosion—other than the above discussed uniform dissolution (active corrosion) and localized pitting and crevice corrosion (local active dissolution)—are briefly presented.

The paragraphs below are arranged in alphabetical order and are intended only as a short reference. For readers interested in a particular topic a few references are given which serve as a link for further reading. Generally, it should be noted that the separation of the categories below is to a large extent based on historic evolution rather than physicochemical mechanisms.

C2.8.5.1 ATMOSPHERIC CORROSION [148, 149, 150]

Atmospheric corrosion results from a metal's ambient-temperature reaction, with the earth's atmosphere as the corrosive environment. Atmospheric corrosion is electrochemical in nature, but differs from corrosion in aqueous solutions in that the electrochemical reactions occur under very thin layers of electrolyte on the metal surface. This influences the amount of oxygen present on the metal surface, since diffusion of oxygen from the atmosphere/electrolyte solution interface to the solution/metal interface is rapid. Atmospheric corrosion rates of metals are strongly influenced by moisture, temperature and presence of contaminants (e.g., NaCl, SO_2 , ...). Hence, significantly different resistances to atmospheric corrosion are observed depending on the geographical location, whether rural, urban or marine.

C2.8.5.2 CONTACT CORROSION = GALVANIC CORROSION [151, 152, 153, 154, 155 AND 156]

This type of corrosive attack occurs when dissimilar metals (i.e., with a different E_{redox}) are in direct electrical contact in corrosive solutions or atmospheres. Under such conditions, enhanced corrosion of the less noble part of the bimetallic couple takes place, whereas the corrosion rate of the more noble part of the couple is reduced or even completely suppressed (as in the case of corrosion suppression by cathodic protection). The difference in corrosion potential of the components of the couple provides the driving force for the corrosion reaction. However, to determine the kinetics of galvanic corrosion, knowledge of the nature and kinetics of the cathodic reaction at the

surface of the more noble metal as well as the nature and kinetics of the anodic reaction on the surface of the less noble metal are required. The nature and conductivity of the electrolyte solution determines the current and potential distribution: the larger the conductivity, the farther from the contact site the coupling action is experienced. A major factor in determining the danger of galvanic corrosion is the ratio of the area of the cathode and the anode. The higher the cathode area compared with the anode area, the larger the enhancement of dissolution of the less noble metal due to coupling.

C2.8.5.3 CAVITATION (CORROSION) [157, 158, 159 AND 160]

Cavitation damage is a form of deterioration associated with materials in rapidly moving liquid environments, due to collapse of cavities (or vapour bubbles) in the liquid at a solid–liquid interface, in the high-pressure regions of high flow. If the liquid in movement is corrosive towards the metal, the damage of the metal may be greatly increased (cavitation corrosion).

-22-

C2.8.5.4 CORROSION FATIGUE [161, 162 AND 163]

Corrosion fatigue is a type of failure (cracking) which occurs when a metal component is subjected to cyclic stress in a corrosive medium. In many cases, relatively mild environments (e.g., atmospheric moisture) can greatly enhance fatigue cracking without producing visible corrosion.

C2.8.5.5 DEALLOYING, SELECTIVE CORROSION

In certain alloys and under certain environmental conditions, selective removal of one metal (the most electrochemically active) can occur that results in a weakening of the strength of the component. The most common example is dezincification of brass [164, 165]. The residual copper lacks mechanical strength.

Another case of selective corrosion is the graphitization of grey cast iron, resulting in preferential removal of the metallic constituent, leaving graphite. Here again the physical form of the casting is maintained, but it is devoid of any mechanical strength.

C2.8.5.6 EROSION (CORROSION) [166, 167]

Erosion is the deterioration of a surface by the abrasive action of solid particles in a liquid or gas, gas bubbles in a liquid, liquid droplets in a gas or due to (local) high-flow velocities. This type of attack is often accompanied by corrosion (erosion–corrosion). The most significant effect of a joint action of erosion and corrosion is the constant removal of protective films from a metal's surface. This can also be caused by liquid movement at high velocities, and will be particularly prone to occur if the solution contains solid particles that have an abrasive action.

C2.8.5.7 FRETTING CORROSION [168]

Fretting corrosion is a form of damage which occurs at the interface of two closely fitting surfaces when they are subject to slight oscillatory slip and joint corrosion action. Almost all materials are subject to fretting and, hence, its incidence in vibrating machinery is high. The damage is mostly of a localized form and any debris which is generated (mostly oxide) has some difficulty escaping from the rubbing zone, and this can lead to an increase in stress.

C2.8.5.8 HYDROGEN EMBRITTLEMENT [169]

A process resulting in a decrease in toughness or ductility of a metal due to absorption of hydrogen. This atomic hydrogen can result, for instance, in the cathodic corrosion reaction or from cathodic protection.

C2.8.5.9 IMPINGEMENT ATTACK [158]

Localized erosion–corrosion caused by turbulence or impinging flow at certain points of the surface. In the majority of cases of impingement attack, a geometrical feature of the system results in turbulence at one or more parts of the surface.

-23-

C2.8.5.10 INTERGRANULAR CORROSION [170]

Corrosion damage due to enhanced dissolution in or adjacent to the grain boundaries of a metal, due to composition gradients between the grain boundary area and the bulk metal. An example is the intergranular attack of stainless steels, which can be explained by a chromium depletion. In a specific temperature region, carbon diffuses to the grain boundaries and reacts with chromium to form chromium carbides, thereby depleting the adjacent areas of chromium. Since stainless steels depend on chromium for corrosion resistance, the grain boundary areas become less resistant to corrosion and more susceptible to localized attack.

C2.8.5.11 MICROBIOLOGICALLY INDUCED CORROSION (MIC) [171, 172 AND 173]

Corrosion associated with the action of micro-organisms present in the corrosion system. The biological action of organisms which is responsible for the enhancement of corrosion can be, for instance, to produce aggressive metabolites to render the environment corrosive, or they may be able to participate directly in the electrochemical reactions. In many cases microbial corrosion is closely associated with biofouling, which is caused by the activity of organisms that produce deposits on the metal surface.

C2.8.5.12 STRAY-CURRENT CORROSION [138]

Corrosion due to stray current—the metal is attacked at the point where the current leaves. Typically, this kind of damage can be observed in buried structures in the vicinity of cathodic protection systems or the DC stray current can stem from railway traction sources.

C2.8.5.13 STRESS CORROSION CRACKING (SCC) [174, 175, 176, 177, 178 AND 179]

A process involving combined corrosion and straining of the metal due to residual or applied stresses. The occurrence of stress corrosion cracking is highly specific; only particular metal/environment systems will crack. The appearance of stress corrosion cracking may be either intergranular or transgranular in nature.

REFERENCES

- [1] Evans U R 1960 *The Corrosion and Oxidation of Metals* (London: Arnold)
 - [2] Behrens D (ed) 1987–1992 *Dechema Corrosion Handbook* (Weinheim: VCH)
 - [3] Atkinson J T N and VanDroffelaar H 1982 *Corrosion and Its Control—an Introduction to the Subject* (Houston, TX: NACE)
 - [4] Craig B D (ed) 1990 *Handbook of Corrosion Data* (Metals Park, OH: ASM International)
 - [5] Craig B D 1991 *Fundamental Aspects of Corrosion Films in Corrosion Science* (New York: Plenum)
 - [6] Dillon C P 1995 *Corrosion Resistance of Stainless Steels* (New York: Dekker)
-

- [7] Droffelaar H V and Atkinson J T N 1995 *Corrosion and its Control: an Introduction to the Subject* (Houston, TX: NACE)
- [8] Doring D D 1997 *Corrosion Atlas* (Amsterdam: Elsevier)
- [9] Fontana M G 1986 *Corrosion Engineering* (New York: McGraw-Hill)
- [10] Gellings P J 1976 *Introduction to Corrosion Prevention and Control for Engineers* Nijgh-Wolters-Noordhoff Universitaire Uitgevers Rotterdam
- [11] Greene R W (ed) 1986 *The Chemical Engineering Guide to Corrosion* (New York: McGraw-Hill)
- [12] Heitz E, Henkhaus R and Rahmel A 1992 *Corrosion Science: an Experimental Approach* (New York: Ellis Horwood)
- [13] Jones D A 1996 *Principles and Prevention of Corrosion* (Upper Saddle River, NJ: Prentice-Hall)
- [14] Kaesche H 1966 *Die Korrosion der Metalle* (Berlin: Springer)
- [15] Kowaka M 1990 *Metal Corrosion Damage and Protection Technology* (New York: Allerton)
- [16] Mansfeld F (ed) 1987 *Corrosion Mechanisms* (New York: Dekker)
- [17] Marcus P and Oudar J (ed) 1995 *Corrosion Mechanisms in Theory and Practice* (New York: Dekker)
- [18] Mattson E 1989 *Basic Corrosion Technology for Scientists and Engineers* (Chichester: Ellis Horwood)
- [19] Mercer A D 1990 *Corrosion in Seawater Systems* (New York: Ellis Horwood)
- [20] Orth H 1974 *Korrosion und Korrosionsschutz* (Stuttgart: MBH)
- [21] Pourbaix M 1973 *Lectures on Electrochemical Corrosion* (New York: Plenum)
- [22] Schweitzer P A 1998 *Encyclopedia of Corrosion Technology* (New York: Dekker)
- [23] Scully J C 1983 *Corrosion: Aqueous Processes and Passive Films* (London: Academic)
- [24] Scully J C 1990 *The Fundamentals of Corrosion* (Oxford: Pergamon)
- [25] Sedriks A J 1996 *Corrosion of Stainless Steels* (New York: Wiley).
- [26] Shreir L L, Jarman R A and Burstein G T (ed) 1994 *Corrosion* (Oxford: Butterworth-Heinemann)
- [27] Trethewey K R and Chamberlain J 1991 *Corrosion: for Students and Engineering* (Harlow: Longman)
- [28] Uhlig H H (ed) 1948 *The Corrosion Handbook* (New York: Wiley)
- [29] Uhlig H H 1963 *Corrosion and Corrosion Control—an Introduction to Corrosion Science and Engineering* (New York: Wiley)
- [30] Uhlig H H 1975 *Korrosion* (Berlin: Akademie)
- [31] Uhlig H H and Revie R W 1985 *Corrosion and Corrosion Control: an Introduction to Corrosion Science and Engineering* (New York: Wiley)
- [32] Landolt D 1993 *Corrosion et Chimie des Métaux* (Lausanne: Presses Polytechniques et Universitaires Romandes)

- [33] Parkins R N 1982 *Corrosion Processes* (London: Applied Science)
- [34] Atkins P W 1986 *Physical Chemistry* (Oxford: Oxford University Press)
- [35] Vetter K J 1961 *Elektrochemische Kinetik* (Berlin: Springer)
- [36] Bockris J O M and Reddy A K N 1970 *Modern Electrochemistry* (New York: Plenum)
- [37] Newman J 1991 *Electrochemical Systems* (Englewood Cliffs, NJ: Prentice-Hall)
- [38] Bard A J, Parsons R and Jordan J 1985 *Standard Potentials in Aqueous Solutions* (New York: Dekker)
- [39] Pourbaix M 1963 *Atlas d'Equilibres Électrochimiques* (Paris: Gautier-Villars)
- [40] Keddam M (ed) 1995 *Corrosion Mechanisms in Theory and Practice* ed P Marcus and J Oudar (New York: Dekker)
- [41] Levich V G 1962 *Physicochemical Hydrodynamics* (London: Prentice-Hall)
- [42] Albery W J and Hitchman M L 1971 *Ring Disc Electrodes* (Oxford: Clarendon)
- [43] Newman J 1967 *Advanced in Electrochemistry and Electrochemical Engineering* ed C W Tobias (New York: Wiley)
- [44] Diggle J W 1973 *Oxides and Oxide Films* (New York: Dekker)
- [45] Young L 1961 *Anodic Oxide Films* (London: Academic)
- [46] Frankenthal R P and Kruger J (eds) 1978 *Passivity of Metals* (Princeton, NJ: Electrochemical Society)
- [47] Froment M (ed) 1983 *Passivity of Metals and Semiconductors* (Amsterdam: Elsevier)
- [48] Heusler K (ed) 1995 *Passivity of Metals and Semiconductors* (Aedermannsdorf: TransTech)
- [49] *German–American Coll. on Electrochemical Passivation* 1989 *Corros. Sci.* **29**
- [50] Sato N and Hashimoto K (eds) 1990 *Passivation of Metals and Semiconductors* (Oxford: Pergamon)
- [51] Natishan P M, Isaacs H S, Janik-Czachor M, Macagno V A, Marcus P and Seo M (eds) 1998 *Passivity and its Breakdown* Proc. vol 97-26 (Pennington, NJ: Electrochemical Society)
- [52] McCafferty E and Brodd R J (Eds) 1986 *Surface, Inhibition and Passivation* Proc. vol 86-7 (Pennington, NJ: Electrochemical Society)
- [53] MacDougall B R, Alwitt R S and Ramanarayanan T A (eds) 1992 *Oxide Films on Metals and Alloys* Proc. vol 92-22 (Pennington, NJ: Electrochemical Society)
- [54] Hebert K R and Thompson G E (eds) 1994 *Oxide Films on Metals and Alloys VII* Proc vol 94-25 (Pennington, NJ: Electrochemical Society)
- [55] Bardwell J A (ed) 1996 *Surface Oxide Films* Proc vol 96-18 (Pennington, NJ: Electrochemical Society)
- [56] Chao C Y, Lin L-F and Macdonald D D 1981 *J. Electrochem. Soc.* **128** 1187
- [57] Chao C Y, Lin L-F and Macdonald D D 1982 *J. Electrochem. Soc.* **129** 1874
-

- [58] Cabrera N and Mott N F 1948 *Rep. Prog. Phys.* **12** 267
- [59] Vetter K J 1954 *Z. Electrochem. Ber. Bunsenges. phys. Chem.* **58** 230
- [60] Fehlner F P and Graham M J 1995 *Corrosion Mechanisms in Theory and Practice* ed P Marcus and J Oudar (New York: Dekker)
- [61] Kirchheim R 1987 *Electrochim. Acta* **32** 1619
- [62] Fehlner F P and Mott N F 1970 *Oxid. Met.* **2** 59
- [63] Wilmsen C W (ed) 1985 *Physics and Chemistry of III–V Compound Semiconductor Interfaces* (New York: Plenum)
- [64] Schmuki P, Sproule G I, Bardwell J A, Lu Z H and Graham M J 1996 *J. Appl. Phys.* **79** 7303
- [65] Heumann T and Rösener W 1955 *Z. Elektrochem.* **59** 722
- [66] Schmuki P, Virtanen S, Davenport A J and Vitus C M 1996 *J. Electrochem. Soc.* **143** 3997
- [67] Sieradzki K and Newman R C 1985 *J. Electrochem. Soc.* **133** 1979
- [68] Qian S, Newman R C, Cottis R A and Sieradzki K 1990 *J. Electrochem. Soc.* **137** 435
- [69] Revesz A G and Kruger J 1976 *Passivity of Metals* ed R P Frankenthal and J Kruger (Princeton, NJ: Electrochemical Society) p 137
- [70] Sakashita M and Sato N 1978 *Passivity of Metals* ed R P Frankenthal and J Kruger (Princeton, NJ: Electrochemical Society) p 479
- [71] Clayton C R and Lu Y C 1986 *J. Electrochem. Soc.* **133** 2465
- [72] Monnartz P 1911 *Metallurgie* **8** 161
- [73] Schmuki P, Virtanen S, Isaacs H S, Ryan M P, Davenport A, Böhni H and Stenberg T 1998 *J. Electrochem. Soc.* **145** 791
- [74] MacDougall B, Mitchell D F and Graham M J 1982 *Corrosion* **38** 85
- [75] Hashimoto K, Osada K, Masumoto T and Shimodaira S 1976 *Corros. Sci.* **16** 71
- [76] Mischler S, Vogel A, Mathieu H J and Landolt D 1991 *Corros. Sci.* **32** 925
- [77] Kirchheim R, Heine B, Fischmeister H, Hofmann S, Knotte H and Stolz U 1989 *Corros. Sci.* **29** 899
- [78] Castle J E and Qiu J H 1989 *Corros. Sci.* **29** 591
- [79] Haupt S and Strehblow H H 1989 *Corros. Sci.* **29** 163
- [80] Landolt D 1990 *Advances in Localized Corrosion* vol 9, ed H S Isaacs, U Bertocci, J Kruger and S Smialowska (NACE) p 25
- [81] Mitchell D F and Graham M J 1986 *J. Electrochem. Soc.* **133** 936
- [82] Hoppe H W and Strehblow H H 1989 *Surf. Interface Anal.* **14** 121
- [83] Haupt S and Strehblow H H 1987 *Langmuir* **3** 873

- [84] Maurice V, Yang W P and Marcus P 1998 *J. Electrochem. Soc.* **145** 909
- [85] Costa D, Yang W P and Marcus P 1995 *Mater. Sci. Forum* **185–188** 325
- [86] Marcus P and Olefjord I 1979 *Surf. Interface Anal.* **4** 29
- [87] Dongil K, Kagwade S V and Clayton C R 1998 *Surf. Interface Anal.* **26** 155
- [88] Clayton C R and Olefjord I 1995 *Corrosion Mechanisms in Theory and Practice* ed P Marcus and J Oudar (New York: Dekker) p 175
- [89] Olefjord I and Wegrelius L 1990 *Corros. Sci.* **31** 89
- [90] Ryan M P, Newman R C and Thompson G E 1995 *J. Electrochem. Soc.* **142** L177
- [91] Toney M F, Davenport A J, Oblonsky L J, Ryan M P and Vitus C M 1997 *Phys. Rev. Lett.* **79** 4282
- [92] Stimming U 1986 *Electrochim. Acta* **31** 415
- [93] Sunseri C, Piazza S, Paolo A D and DiQuarto F 1987 *J. Electrochem. Soc.* **134** 2410
- [94] Chen C T and Cahan B D 1982 *J. Electrochem. Soc.* **129** 17
- [95] Cahan B D and Chen C T 1982 *J. Electrochem. Soc.* **129** 474
- [96] Leitner K and Schultze J W 1988 *Ber. Bunsenges. Phys. Chem.* **92** 181
- [97] Koenig U and Schultze J W 1992 *Solid State Ion. Diffus. Reactions* **53–56** 255
- [98] Gerischer H 1989 *Corros. Sci.* **29** 257
- [99] Gorse D, Rondot B and Belo M d C 1990 *Corros. Sci.* **30** 23
- [100] Hakiki N E, Belo M d C, Simoes A M P and Ferreira M G S 1998 *J. Electrochem. Soc.* **145** 3821
- [101] Schmuki P and Boehni H 1992 *J. Electrochem. Soc.* **139** 1908
- [102] Rajeshwar K, Peter L M, Fujishima A, Meissner D and Tomkiewich M (eds) 1997 *Proc. Symp. on Photoelectrochemistry* (Pennington, NJ: Electrochemical Society)
- [103] Sato N 1998 *Electrochemistry at Metal and Semiconductor Electrodes* (Amsterdam: Elsevier)
- [104] Morrison S R 1980 *Electrochemistry at Semiconductor and Oxidized Metal Electrodes* (New York: Plenum)
- [105] Dean M H and Stimming U 1987 *J. Electroanal. Chem.* **228** 135
- [106] Boehni H 1987 *Langmuir* **3** 924
- [107] Zsklarska-Smialowska Z 1986 *Pitting Corrosion of Metals* (Houston, TX: National Association of Corrosion Engineers)
- [108] Strehblow H H 1995 *Corrosion Mechanisms in Theory and Practice* ed P Marcus and J Oudar (New York: Dekker) p 201
- [109] Baroux B 1995 *Corrosion Mechanisms in Theory and Practice* ed P Marcus and J Oudar (New York: Dekker)

[110] Frankel G S 1998 *J. Electrochem. Soc.* **145** 2186

[111] Hine F, Komai K and Yamakawa K (eds) 1988 *Localized Corrosion* (London: Elsevier)

- [112] Tousek J 1985 *Theoretical Aspects of the Localized Corrosion of Metals* (Rockport, MA: TransTech)
- [113] Boehni H 1987 *Corrosion Mechanisms* ed F Mansfeld (New York: Dekker)
- [114] Brown B F, Kruger J and Staehle R W (eds) 1974 *Localized Corrosion* (Houston, TX: NACE)
- [115] Frankenthal R P and Kruger J (eds) 1984 *Equilibrium Diagrams of Localized Corrosion Proc.* vol 84-9 (Pennington, NJ: Electrochemical Society)
- [116] Isaacs H S, Bertocci U, Kruger K and Smialowska S (eds) 1990 *Advances in Localized Corrosion* (Houston, TX: NACE)
- [117] Natishan P M, Kelly A G, Frankel G S and Newman R (eds) 1996 *Critical Factors in Localized Corrosion II Proc.* vol 95-15 (Pennington, NJ: Electrochemical Society)
- [118] Schmuki P, Lockwood D J, Bsiesy A and Isaacs H S (eds) 1997 *Pits and Pores: Formation, Properties, and Significance for Advanced Luminescent Materials Proc.* vol 97-7 (Pennington, NJ: Electrochemical Society)
- [119] Canham L T 1990 *Appl. Phys. Lett.* **57** 1046
- [120] Cullis A G, Canham L T and Calcott P D J 1997 *J. Appl. Phys.* **82** 909
- [121] Schmuki P, Erickson L E and Lockwood D J 1998 *Phys. Rev. Lett.* **80** 4060
- [122] Kofstad P 1988 *High Temperature Corrosion* (London: Elsevier)
- [123] Lai G Y 1990 *High-Temperature Corrosion of Engineering Alloys* (Metals Park, OH: ASM International)
- [124] Saito Y, Oenay B and Maruyama T (eds) 1992 *High Temperature Corrosion of Advanced Materials and Protective Coatings* (Amsterdam: North-Holland)
- [125] Shores D A, Rapp R A and Hou P Y (eds) 1997 *Proc. Symp. on Fundamental Aspects of High Temperature Corrosion* vol 96-26 (Pennington, NJ: Electrochemical Society)
- [126] Mrowec S and Werber T 1978 *Gas Corrosion of Metals* (Washington, DC: National Bureau of Standards)
- [127] Rapp R A (ed) 1983 *High Temperature Corrosion—NACE 6* (Houston, TX: NACE)
- [128] Wagner C 1933 *Z. Phys. Chem. B* **21** 25
- [129] Sze S M 1983 *VLSI Technology* (New York: McGraw-Hill)
- [130] Graham M J and Hussey R J 1995 *Oxid. Met.* **44** 339
- [131] Jaenicke W, Leistikow S and Sadler A 1964 *J. Electrochem. Soc.* **111** 1031
- [132] Pilling N B and Bedworth R E 1923 *J. Inst. Met.* **29** 529
- [133] Appleby W K and Tylecote R F 1970 *Corros. Sci.* **10** 325
- [134] Stringer J 1972 *Werkstoffe Korros.* **23** 747

- [135] Saunders S R J, Evans H E and Stringer J A (eds) *Workshop on Mechanical Properties of Protective Oxide Scales. Materials at High Temperatures* vol 12 (Teddington)
- [136] Baeckman W v, Schenk W and Prinz W 1997 *Handbook of Cathodic Corrosion Protection: Theory and Practice of Electrochemical Protection Processes* (Houston, TX: Gulf)
- [137] Bayliss D A and Chandler K A 1991 *Steelwork Corrosion Control* (London: Elsevier)

- [138] Ashworth V and Booker C J L (eds) 1985 *Cathodic Protection: Theory and Practice* (New York: Ellis Horwood)
- [139] Riggs O L and Locke C E 1981 *Anodic Protection: Theory and Practice in the Prevention of Corrosion* (New York: Plenum)
- [140] Collie M J (ed) 1980 *Corrosion Inhibitors: Developments since 1980* (Park Ridge, NJ: Noyes)
- [141] Kuznetsov Y I 1996 *Organic Inhibitors of Corrosion of Metals* (New York: Plenum)
- [142] Rozenfeld I L 1981 *Corrosion Inhibitors* (New York: McGraw-Hill)
- [143] Nathan C C (ed) 1973 *Corrosion Inhibitors* (Houston, TX: NACE)
- [144] Robinson J S 1979 *Corrosion Inhibitors—Recent Developments* (Park Ridge, NJ: Noyes)
- [145] Sastri V S 1998 *Corrosion Inhibitors* (Chichester: Wiley)
- [146] Leidheiser H (ed) 1981 *Corrosion Control by Organic Coatings* (Houston, TX: National Association of Corrosion Engineers)
- [147] Munger C G 1984 *Corrosion Prevention by Protective Coatings* (Houston, TX: National Association of Corrosion Engineers)
- [148] Dean S W and Rhea E C (eds) 1982 *Atmospheric Corrosion of Metals* (Philadelphia, PA: ASTM)
- [149] Ailor W H (eds) 1982 *Atmospheric Corrosion* (New York: Wiley)
- [150] Leygraf C 1995 *Corrosion Mechanisms in Theory and Practice* ed P Marcus and J Oudar (New York: Dekker)
- [151] Mansfeld F, Hestenberg D H and Kenkel J V 1974 *Corrosion* **30** 343
- [152] Reboul M C 1979 *Corrosion* **35** 423
- [153] Walker M S 1979 *Mater. Performance* **18** 9
- [154] Jones D A 1968 *Electrochem. Technol.* **6** 241
- [155] Mansfeld F 1971 *Corrosion* **27** 436
- [156] Mansfeld F 1977 *Corrosion* **33** 224
- [157] Brennen C E 1995 *Cavitation and Bubble Dynamics* (New York: Oxford University Press)
- [158] Hammit F G 1980 *Cavitation and Multiphase Flow Phenomena* (New York: McGraw-Hill)
- [159] Young F R 1989 *Cavitation* (London: McGraw-Hill)
- [160] Preece C M and Hansson I L H 1981 *Advances in the Mechanics and Physics of Surfaces* ed R M Latanision and R J Courtel, vol 1 (Chur: Harvard) p 199
-

- [161] Devereux O F, McEvily A J and Staehle R W (eds) 1972 *Corrosion Fatigue* (Houston, TX: NACE)
- [162] Crooker T W and Leis B N (eds) 1983 *Corrosion Fatigue: Mechanics, Metallurgy, Electrochemistry and Engineering* STP 801 (ASTM)
- [163] Congleton J and Craig I H 1982 *Corrosion Processes* ed R N Parkins (London: Applied Science) p 209
- [164] Heidersbach R H 1968 *Corrosion* **24** 38

- [165] Heidersbach R H and Verink E D 1972 *Corrosion* **28** 397
 - [166] Levy A 1995 *Solid Particle Erosion and Erosion–Corrosion of Materials* (Materials Park, OH: ASM International)
 - [167] Heitz E 1991 *Corrosion* **47** 135
 - [168] Waterhouse R B 1972 *Fretting Corrosion* (Oxford: Pergamon)
 - [169] Gibala R and Hehemann R F (eds) 1984 *Hydrogen Embrittlement and Stress Corrosion Cracking* (Metals Park, OH: American Society of Metals)
 - [170] Cihal V 1984 *Intergranular Corrosion of Steels and Alloys* (Amsterdam: Elsevier)
 - [171] Heitz E, Fleming H-C and Sand W (eds) 1996 *Microbially Influenced Corrosion of Materials* (Berlin: Springer)
 - [172] Borenstein S W 1994 *Microbiologically Influenced Corrosion Handbook* (Cambridge: Woodhead)
 - [173] Thierry D and Sand W 1995 *Corrosion Mechanisms in Theory and Practice* ed P Marcus and J Oudar (New York: Dekker)
 - [174] Staehle R W, Forty A J and Rooyen D v (eds) 1969 *Fundamental Aspects of Stress Corrosion Cracking* (Houston, TX: NACE)
 - [175] Jones R H (ed) 1993 *Stress-Corrosion Cracking* (Materials Park, OH: ASM)
 - [176] McEvily A J (ed) 1990 *Atlas of Stress-Corrosion and Corrosion Fatigue Curves* (Materials Park, OH: ASM)
 - [177] Speidel M O, Denk J and Scarlin B 1991 *Stress Corrosion Cracking and Corrosion Fatigue of Steam-Turbine Rotor and Blade Materials* (Luxembourg: Commission of the European Communities)
 - [178] Yahalom J and Aladjem A (eds) 1980 *Stress Corrosion Cracking* (Tel-Aviv: Freund)
 - [179] Newman R C 1995 *Corrosion Mechanisms in Theory and Practice* ed P Marcus and J Oudar (New York: Dekker)
-

C2.9 Tribology

Andrew J Gellman

C2.9.1 INTRODUCTION

The term tribology translates literally into ‘the study of rubbing’. In modern parlance this field is held to include four phenomena: adhesion, friction, lubrication and wear. For the most part these are phenomena that occur between pairs of solid surfaces in contact with one another or separated by a thin fluid film. *Adhesion* describes the resistance to separation of two surfaces in contact while *friction* describes their tendency to resist shearing. *Lubrication* is the phenomenon of friction reduction by the presence of a fluid (or solid) film between two surfaces. Finally, *wear* describes the irreversible damage or deformation that occurs as a result of shearing or separation.

Tribological phenomena have been known to mankind since prehistorical times when friction between wooden sticks was used to produce fire. The first historical record of tribology described the use of lubrication in the

construction of Egyptian temples as far back as 2400 BC. The earliest scientific studies of tribological phenomena are attributed to da Vinci in the 15th century, Amontons in the 17th and Coulomb in the 18th. More recently, the application of modern methods of physics and chemistry to tribology is usually credited to Bowden and Tabor for work done in the post World War II era [1, 2]. In the past decade there has been a great deal of progress in the understanding of tribological phenomena. This progress has been catalysed by the development and application of a number of experimental and theoretical methods that allow study of tribological phenomena at an unprecedented level of detail [3, 4]. This effort has been further motivated by the development of several ‘high tech’ devices that have pushed the need for tribology into extreme conditions and environments. Examples include: the lubrication of ceramics for high temperature applications, lubrication of the surfaces of hard disks for data storage [5] and microelectromechanical systems (MEMS) [6]. These applications place unprecedented demands on the performance of tribological systems. Furthermore, there is increasing interest in solving tribological problems that arise in such diverse environments such as the human body (joints), vacuum (satellite components) and the Earth’s mantle (tectonic motion and earthquakes).

The scope of this entry includes a description of tribological phenomena and the modern tools that are spurring developments in our understanding of tribology. The goal is to provide the reader with a basic understanding of the concepts, an understanding of their limitations and a perspective on the breadth and scope of phenomena that are included under the umbrella of tribology.

C2.9.2 PHYSICAL DESCRIPTION OF TRIBOLOGICAL PHENOMENA

A typical physical process involving friction occurs in three steps: the formation of a contact between two solid surfaces, the shearing of those surfaces and the separation of those surfaces ([figure C2.9.1](#)). A device used to measure friction includes the elements shown in [figure C2.9.1](#) and measures friction forces (F) and normal forces (N) as extension (or compression) of springs during sliding. The phenomenological measure of friction is in terms of a friction coefficient given by

-2-

$$\mu = \frac{F}{N}.$$

The proportionality between normal and friction forces is observed for many systems but is not founded in any basic physics. Much work in the field of tribology has been devoted to rationalizing the implication that the friction coefficient does not depend upon the apparent contact area between the two solid surfaces. For this reason the field of contact mechanics has always been intimately linked to tribology.

Figure C2.9.1 Schematic representation of the steps involved in a tribological process: (a) contact between surfaces, (b) shearing under a constant normal force and (c) separation against adhesive forces. In the absence of gravity the normal (N) and friction (F) forces are measured by the extension or compression of the springs.

C2.9.2.1 CONTACT OF SOLID SURFACES

With very few exceptions it is possible to produce surfaces of materials that are atomically smooth across macroscopic length scales. The protrusions that exist on the surfaces of all common materials are called asperities. As two surfaces are brought together, the peaks of asperities come into contact first. If the two solids were nondeformable there would be at most three contact points spanning the entire apparent contact area and supporting the normal force. In reality, the normal forces between two solids cause both elastic and plastic deformation of the regions around the initial contact points such that the two solids come into contact across a finite area.

Modelling of the true contact area between surfaces requires consideration of the deformation that occurs at the peaks of asperities as they come into contact with mating surfaces. Purely elastic contact between two solids was first described by H Hertz [7]. The Hertzian contact area (A_H) between a sphere of radius r and a flat surface compressed under normal force N is given by

$$A_H = \pi(r\kappa N)^{2/3}$$
$$\kappa = \frac{3}{4} \left(\frac{1 - \nu_1^2}{E_1} + \frac{1 - \nu_2^2}{E_2} \right)$$

where E_1, E_2 are the elastic moduli, and ν_1, ν_2 are the Poisson numbers for the sphere and the surface. The friction force during shearing is proportional to the contact area; however, under conditions of purely elastic Hertzian contact it cannot be proportional to the normal force. This is inconsistent with the empirical observation of a coefficient of friction. The most important correction to the Hertz expression is due to Johnson, Kendall and Roberts (JKR) who included the fact that the surface tensions of two solids will contribute to the contact area [8]. The JKR expression for the contact area between a sphere and a flat surface is

$$A_{JKR} = \pi(r\kappa)^{2/3} (N + 3\pi r\gamma + \sqrt{6\pi r\gamma N + 9\pi^2 r^2 \gamma^2})^{2/3}.$$

In this expression $\gamma = \gamma_1 + \gamma_2 - \gamma_{12}$ where γ_1 and γ_2 are the surface tensions of the sphere and flat, respectively, and γ_{12} is the surface tension of the interface between them. The JKR expression recognizes the fact that even in the absence of a normal force ($N = 0$) surface tension will cause some elastic deformation of the surfaces producing a finite contact area. This fact alone renders the concept of a coefficient of friction meaningless since it implies that there is some finite friction force between solids even under zero normal force.

In reality most solids in contact under macroscopic loads undergo irreversible plastic deformation. This is caused by the fact that at high normal forces the stresses in the bulk of the solid below the contact points exceed the yield stress. Under these conditions the contact area expands until the integrated pressure across the contact area is equal to the normal force. Since the pressure is equal to the yield strength of the material σ , the plastic contact area is given by

$$A_p = \frac{N}{\sigma}.$$

Thus, under conditions of plastic deformation the real area of contact is proportional to the normal force. If the shear force during sliding is proportional to that area, one has the condition that the shear force is proportional to the normal force, thus leading to the definition of a coefficient of friction.

Determining the contact area between two rough surfaces is much more difficult than the sphere-on-flat problem and depends upon the morphology of the surfaces [9]. One can show, for instance, that for certain distributions of asperity heights the contact can be completely elastic. However, for realistic morphologies and macroscopic normal forces, the contact region includes areas of both plastic and elastic contact with plastic contact dominating.

Sliding of two solid surfaces in contact is induced by the application of a shear force. As the shear force spring in [figure C2.9.1](#) is stretched at a constant velocity, the shear force on the interface increases until sliding begins. This process is illustrated in [figure C2.9.2\(a\)](#) in which the shear force increases until a critical shear stress is reached. At that point sliding begins and the shear stress at the interface often drops to some constant value. The critical shear force (F_c) needed to induce sliding is often used to define a static coefficient of friction

$$\mu_s = \frac{F_c}{N}$$

while the shear force during sliding defines a dynamic coefficient of friction

$$\mu_d = \frac{F}{N}.$$

A second type of dynamics is commonly observed during sliding: stick–slip motion. This is responsible for the generation of acoustic emission, e.g., the squealing of chalk on a blackboard or brakes in a car. The dynamics of stick–slip motion are illustrated in [figure C2.9.2 b](#)) which shows the shear stress at the interface going through periodic oscillations. It is important to point out that the origins of stick–slip behaviour lie in the dynamics of the entire system and not simply the properties of the interface. In the device of [figure C2.9.1](#) stick–slip motion will always occur at low sliding speeds or for shear springs with low spring constants [10].

The role of solid and liquid lubricants at solid–solid interfaces is to reduce friction forces during sliding. Liquid lubrication is commonly described as occurring in three regimes that depend upon the normal forces and sliding speeds of two surfaces in contact. At low normal force and high sliding speeds liquid films can completely separate two surfaces, preventing solid–solid contact. Under these conditions, usually referred to as *hydrodynamic* lubrication, the frictional forces between two surfaces are determined by the rheological properties of the thin fluid film that separates them. As the normal force is increased and the sliding speed decreased the interface enters the *boundary* regime of lubrication. The surfaces have deformed under the high normal forces and are thought to be separated by monomolecular films of adsorbed molecules. These are typically surfactant-like species that are added to lubricant fluids for just this purpose. Under even higher normal forces the interface enters the *extreme pressure* regime during which direct solid–solid contact occurs. The surfaces are deformed even further and high rates of wear are observed exposing clean solid surfaces. Lubricant fluids usually contain extreme pressure additives which can react with clean exposed surfaces under high pressure and high temperature conditions to form thin solid films with low shear yield strengths. It is these thin solid films that provide lubrication. As implied by the discussion above, lubricant fluids are often very complicated mixtures containing as many as ten or 20 additives, each of which serves a specific purpose in reducing friction and wear of solid surfaces in sliding contact [2].

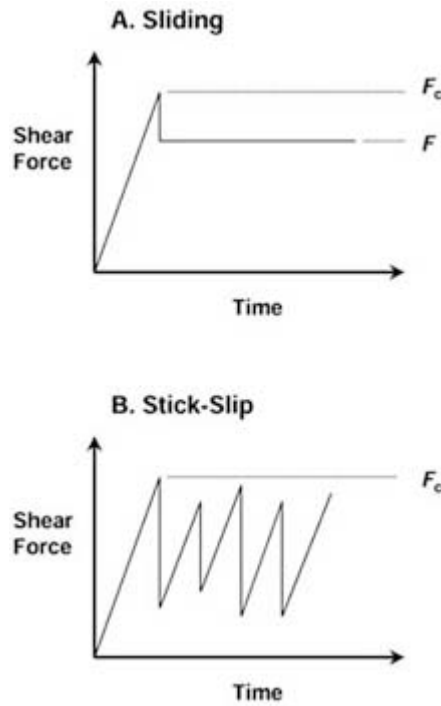


Figure C2.9.2 Shear force *versus* time during (a) sliding and (b) stick–slip motion. The motion of the surface beneath the sliding block of [figure C2.9.1](#) is at constant velocity.

C2.9.2.3 SEPARATION OF SOLID SURFACES

The separation of two surfaces in contact is resisted by adhesive forces. As the normal force is decreased, the contact regions pass from conditions of compressive to tensile stress. As revealed by JKR theory, surface tension alone is sufficient to ensure that there is a finite contact area between the two at zero normal force. One contribution to adhesion is the work that must be done to increase surface area during separation. If the surfaces have undergone plastic deformation, the contact area will be even greater at zero normal force than predicted by JKR theory. In reality, continued plastic deformation can occur during separation and also contributes to adhesive work.

C2.9.2.4 ENERGY DISSIPATION MECHANISMS

The friction between two surfaces in sliding contact often manifests itself experimentally as a force. At a fundamental level, however, it is more relevant to think of friction in terms of energy dissipation [11]. In a completely conservative world the lateral translation of one object over another does not require work. Nonetheless, in the presence of friction, work is done. If one considers two solids in contact as a thermodynamic system, then sliding is an adiabatic process in which work is done to the system through the shear spring in [figure C2.9.1](#) increasing its internal energy. The increase in internal energy can be considered to take two forms: thermal energy and potential energy. The thermal energy manifests itself as an increase in temperature. The potential energy increase is in the form of structural or even chemical changes to the system resulting from sliding. Examples include changes in the surface area due to

deformation, the creation of bulk defects in solids and changes in the composition of the system due to chemical reactions occurring during sliding or separation.

One of the primary goals of current research in the area of tribology is to understand how it is that the kinetic energy of a sliding object is converted into internal energy. These dissipation mechanisms determine the rate of energy flow from macroscopic motion into the microscopic modes of the system. Numerous mechanisms can be

and have been suggested: sliding causes excitation of atomic motions at the interface; bulk deformation along slip planes which leads to excitation of atomic motions and formation of defects; excitation of long wavelength motions of the solids can lead to acoustic emission; excitation of electronic motion in the solids can lead to friction due to electronic resistance. Understanding the relative importance of such energy dissipation mechanisms is one of the goals of tribological research. This is an extremely difficult challenge, since their relative importance depends upon the properties of the materials and the characteristics of the motions of the system.

C2.9.3 MODERN METHODS OF TRIBOLOGY

Progress in the understanding of tribology fundamentals has been accelerated over the past decade by the development of both experimental and theoretical tools for its study. While there have been many ideas put forward to describe the sources of friction, one of the primary impediments to progress has been a lack of reproducible measurements of friction under well defined conditions. It is possible to find tabulated lists of the coefficients of friction between various materials sometimes reported to three significant figures of accuracy. However, friction forces are sensitive to so many variables that it is not at all clear that these numbers are useful or even reproducible. As a simple example, friction between metals is influenced by adsorbed molecular films of a few monolayers thickness [12]. Without going to extraordinary lengths it is not possible to create perfectly clean metal surfaces and, as a result, the vast majority of friction measurements between metal surfaces are influenced by contaminant films. Even in the scientific literature it is difficult to find consistency among reported absolute values of friction between solids, undoubtedly because of lack of control over various characteristics of the sliding interface or the measurement mechanism itself.

C2.9.3.1 ATOMIC FORCE MICROSCOPY

The atomic force microscope (AFM) provides one approach to the measurement of friction in well defined systems. The AFM allows measurement of friction between a surface and a tip with a radius of the order of 5–10 nm (figure C2.9.3 a)). It is the true realization of a single asperity contact with a flat surface which, in its ultimate form, would measure friction between a single atom and a surface. The AFM allows friction measurements on surfaces that are well defined in terms of both composition and structure. It is limited by the fact that the characteristics of the tip itself are often poorly understood. It is very difficult to determine the radius, structure and composition of the tip; however, these limitations are being resolved. The AFM has already allowed the spatial resolution of friction forces that exhibit atomic periodicity and chemical specificity [3, 10, 13].

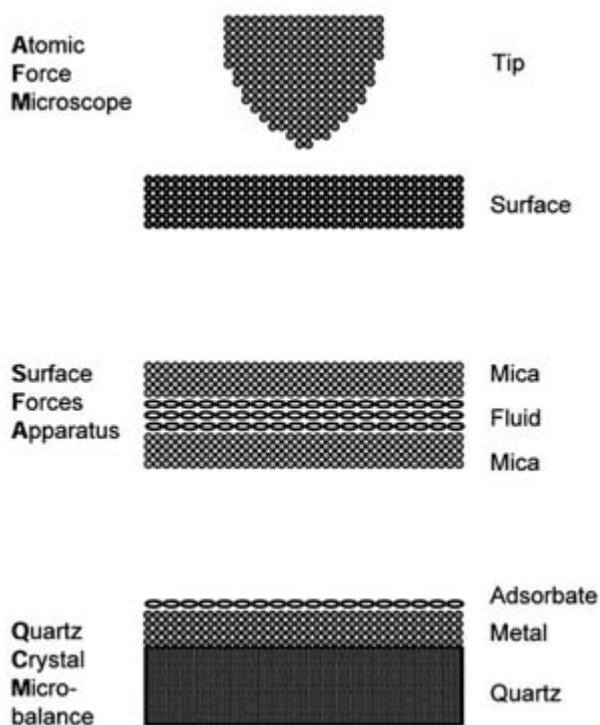


Figure C2.9.3 Schematic diagrams of the interfaces realized by (a) the atomic force microscope, (b) the surface forces apparatus and (c) the quartz crystal microbalance for achieving fundamental measurements of friction in well defined systems.

C2.9.3.2 SURFACE FORCES APPARATUS

The surface forces apparatus (SFA) measures forces between atomically flat surfaces of mica. Mica is the only material that can be prepared with surfaces that are atomically flat across square-millimetre areas. The SFA confines liquid films of a few molecular layers' thickness between two mica surfaces and then measures shear and normal forces between them (figure C2.9.3b)). In essence, it measures the rheological properties of confined, ultra-thin fluid films. The SFA is limited to the use of mica or modified mica surfaces but can be used to study the properties of a wide range of fluids. It has provided experimental evidence for the formation of layered structures in fluids confined between surfaces and evidence for shear-induced freezing of confined liquids at temperatures far higher than their bulk freezing temperatures [14, 15].

C2.9.3.3 MOLECULAR DYNAMICS

Molecular dynamics (MD) methods can be used to simulate tribological phenomena at a molecular level. These have been used primarily to simulate behaviour observed in AFM and SFA measurements. Such simulations are limited to short-timescale events, but provide a wealth of information and insight into tribological phenomena at a level of detail that cannot be realized by any experimental method. One of the most interesting contributions of molecular dynamics

is the demonstration that many of the predictions of continuum mechanics, such as those of Hertz, hold true down to length scales of the order of 10 nm, in spite of the obvious atomic coarseness on these length scales.

Investigations of confined fluids have revealed the layered structures observed experimentally by the SFA and shear-induced freezing/melting transitions that can occur during stick–slip motion [16, 17].

C2.9.3.4 QUARTZ CRYSTAL MICROBALANCE

Quartz crystal microbalances (QCMs) have been used to measure friction between thin films of adsorbed gases and gold surfaces. These measure both the resonant frequency and Q -factor of a quartz crystal oscillator coated with metal (figure C2.9.3(c)). The frequency is determined by the total mass of the system and is sensitive to the presence of fractions of a monolayer of adsorbates. The Q of the resonance serves to measure energy dissipation at the interface between the metal film and weakly adsorbed gases such as Ar, Xe and N_2 . One of the most important contributions of this measurement has been the observation of electronic effects in the friction between adsorbed N_2 and a Pb surface. These electronic effects manifested themselves as discontinuities in the energy dissipation as the Pb was heated and cooled through its superconducting phase transition [4].

C2.9.3.5 UHV SURFACE SCIENCE METHODS

Ultra-high vacuum (UHV) surface science methods allow preparation and characterization of perfectly clean, well ordered surfaces of single crystalline materials. By preparing pairs of such surfaces it is possible to form interfaces under highly controlled conditions. Furthermore, thin films of adsorbed species can be produced and characterized using a wide variety of methods. Surface science methods have been coupled with UHV measurements of macroscopic friction forces. Such measurements have demonstrated that adsorbate film thicknesses of a few monolayers are sufficient to lubricate metal surfaces [12, 18].

C2.9.4 OUTLOOK

Tribological problems will continue to plague society and will become more problematic as technology forces the development of mechanical systems that operate in increasingly extreme environments. These problems cannot be solved indefinitely using empiricism and adaptation of existing methods. This is spurring development of experimental methods for study of tribological phenomena in highly defined systems and at the atomic level. As a result, there is a very encouraging outlook for increased understanding of the mechanisms of adhesion, friction, lubrication and wear. Furthermore, this will have long term impact on the understanding of many other areas in which tribology plays an important but often unnoticed role. These include biological problems such as the adhesion of cells or the lubrication of joints and geological problems such as plate tectonics and earthquakes.

REFERENCES

- [1] Dowson D 1979 *History of Tribology* (London: Longman)
- [2] Bowden F P and Tabor D 1985 *The Friction and Lubrication of Solids* (Oxford: Clarendon)
- [3] Hähner G and Spencer N D 1998 Rubbing and scrubbing *Phys. Today* September, 22–7
- [4] Krim J K 1998 Fundamentals of friction *MRS Bull.* 20–1 (related articles in this issue of *MRS Bulletin* describe the status of several sub-disciplines of tribology)
- [5] Gellman A J 1998 Lubricants and overcoats for magnetic storage media *Curr. Opinion Colloid Interface Sci.* 3 368–72
- [6] Bhushan B (ed) 1998 *Tribology Issues and Opportunities in MEMS* (Dordrecht: Kluwer)
- [7] Hertz H 1886 *J. Reine Angew. Math.* 92 156
- [8] Johnson K L, Kendall K and Roberts A D 1971 Surface energy and the contact of elastic solids *Proc.R.Soc. A* 324 301–13
- [9] Greenwood J A 1992 Contact of rough surfaces *Fundamentals of Friction: Macroscopic and Microscopic Processes (NATO ASI Series E220)* eds I L Singer and H M Pollock (Dordrecht: Kluwer) pp 37–56
- [10] Persson B N J 1997 *Sliding Friction—Physical Principles and Applications* (New York: Springer)

- [11] Tabor D 1992 Friction as a dissipative process *Fundamentals of Friction: Macroscopic and Microscopic Processes (NATO ASI Series E220)* eds I L Singer and H M Pollock (Dordrecht: Kluwer) pp 3–24
- [12] McFadden C F and Gellman A J 1998 Metallic friction: the effect of molecular adsorbates *Surf. Sci.* **409** 171–82
- [13] Carpick R W and Salmeron M 1997 Scratching the surface: fundamental investigations of tribology with atomic force microscopy *Chem. Rev.* **97** 1163–94
- [14] Israelachvili J N, Homola A M and McGuiggan P M 1988 Dynamical properties of molecularly thin liquid films *Science* **240** 189–91
- [15] Granick S 1991 Motions and relaxations of confined liquids *Science* **253** 1374–9
- [16] Landman U, Luedtke W D and Ringer E M 1992 Molecular dynamics simulations of adhesive contact formation and friction *Fundamentals of Friction: Macroscopic and Microscopic Processes (NATO ASI Series E220)* eds I L Singer and H M Pollock (Dordrecht: Kluwer) pp 463–508
- [17] Cieplak M, Smith E D and Robbins M O 1994 Molecular origins of friction: the force on adsorbed layers *Science* **256** 1209
- [18] McFadden C F and Gellman A J 1995 Ultra-high vacuum boundary lubrication of the Cu–Cu interface by 2,2,2-trifluoroethanol *Langmuir* **11** 273–80
-

FURTHER READING

Bowden F P and Tabor D 1985 *The Friction and Lubrication of Solids* (Oxford: Clarendon)

Singer I L and Pollock H M (eds) 1992 *Fundamentals of Friction: Macroscopic and Microscopic Processes (NATO ASI Series E220)* (Dordrecht: Kluwer)

Persson B N J 1997 *Sliding Friction—Physical Principles and Applications* (New York: Springer)

Persson B N J and Tosatti E (eds) *Physics of Sliding Friction (NATO ASI Series E311)* (Dordrecht: Kluwer)

-1-

C2.10 Surface electrochemistry

Hans-Henning Strehblow and Dirk Lützenkirchen-Hecht

C2.10.1 INTRODUCTION

The structure and the composition of the electrode/electrolyte interface is of great scientific and technological interest. It is generally accepted that the adsorption of anions and cations from the solution is the initial step for many important electrochemical processes such as oxide formation, pitting corrosion, electrocatalysis and metal or semiconductor deposition. Classical electrochemical techniques, which are in principle based on potential and current measurements, are able to reveal a detailed description of the electrochemical interface. For example, the kinetics of oxide or chloride film formation on several different metals have been investigated with a high accuracy, and different growth modes can easily be identified [1, 2 and 3]. In some cases, also a microscopically detailed picture of the electrode surface can be derived. For example, each single crystal metal surface in a well defined electrolyte shows a typical cyclic voltammogram which can be used for a simple control of the single crystal surface preparation (as an example see [4, 5 and 6]); in contrast to the large experimental expenditures which are necessary for a conventional surface structure analysis e.g. with low energy electron diffraction (LEED) or grazing incidence x-ray diffraction. For the underpotential deposition (upd) of copper on stepped platinum surfaces, a clear influence of the copper coverage on the hydrogen adsorption reaction was found; even small amounts of adsorbed Cu on the Pt surface are able to block the hydrogen adsorption reaction [7, 8]. Due to the fact that this reaction preferentially takes place at Pt step sites, one can directly deduce that the blocking of hydrogen

adsorption is a consequence of the selective adsorption of Cu at step sites [8].

In general, however, the electrochemical methods cannot provide structural details of the electrode surface such as the binding geometry or the valency of adsorbates. In addition, they can hardly address the question of surface water, the charge distribution in the electrochemical double layer and the structure of the electrode and their changes with potential. Although in principle possible, it is often very difficult to calculate precise values for the coverage of adsorbates from electrochemical measurements, for example in the presence of coadsorbed species (see, e.g., [9]) or if a partial charge transfer from the adsorbate to the electrode occurs (see, e.g., [10]). Therefore, a large variety of different techniques have been introduced to the investigation of electrochemical interfaces in the past. *In situ* techniques, such as ultraviolet (UV) and visible reflectance, infrared (IR) and Raman spectroscopy, scanning tunnelling and atomic force microscopy (STM and AFM), surface x-ray scattering (SXS) and x-ray absorption spectroscopy (XAS), as well as *ex situ* techniques, among them electron diffraction methods (LEED and RHEED), x-ray and UV photoelectron spectroscopy (XPS and UPS), Auger electron spectroscopy (AES) and ion scattering spectroscopy (ISS), revealed a comprehensive picture—detailed on an atomic level—of the electrode surface in contact with the electrolyte and have furthered our understanding of electrochemical processes in many ways. In this contribution, we will report on recent results obtained with *in situ* and *ex situ* techniques with the focus of the present article on adsorption phenomena and underpotential deposition.

-2-

C2.10.2 ADSORPTION

During the past decades, anionic and cationic adsorption on metal electrodes have been intensively investigated. Especially the electrosorption of halides (Cl^- , Br^- , I^-) on noble metals (Au, Ag, Pt) were in the focus of interest. Several different techniques such as electrochemical methods including the electrochemical quartz microbalance technique and capacity measurements [3, 11], radiochemistry and radiotracer experiments [12, 13], SXS [14, 15], electroreflectance spectroscopy [16, 17 and 18], second harmonic generation (SHG) [19, 20], LEED and electron spectroscopies (XPS, UPS and AES) [21, 22, 23, 24 and 25], electrode resistivity measurements [26, 27] and STM [6, 28, 29] have been successfully applied. Well ordered phases which can be commensurate or incommensurate with the underlying noble metal surface were found (e.g. [6, 29, 30, 31 and 32]); in some cases several different adsorbate structures can be observed simultaneously (see, e.g., [29, 33]). As an example, results obtained recently for the adsorption of bromine on Ag (100) with *in situ* SXS are presented in figure C2.10.1 [15]. The absence of translational invariance perpendicular to a single crystal surface causes scattering rods between the three dimensional bulk reflections [34, 35]. SXS comprises the measurement of the intensity distribution of the scattered x-rays in reciprocal space; the comparison of the measured diffraction patterns with atomic models enables the accurate determination of adsorbate structures such as the adsorbate coverage, nearest neighbour bond distances and numbers. Although each element has a characteristic backscattering power, the type of adsorbate can also be identified; this has been successfully used for the investigation of the structures of surface water on silver [36].

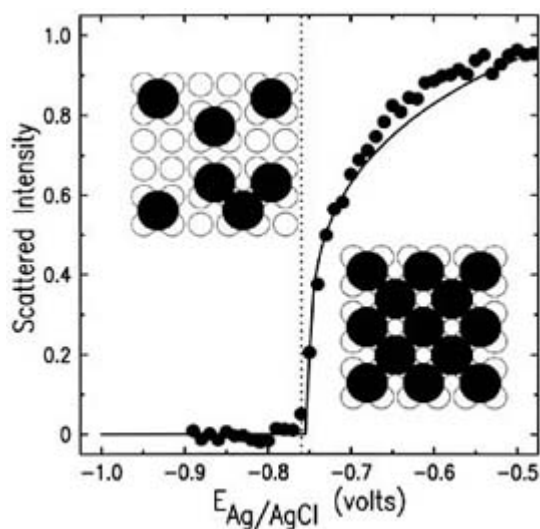


Figure C2.10.1. Potential dependence of the scattering intensity of the (1,0) reflection measured *in situ* from Ag (100)/0.05 M NaBr after a background correction (dots). The solid line represents the fit of the experimental data with a two dimensional Ising model with a critical exponent of 1/8. Model structures derived from the experiments are depicted in the insets for potentials below (left) and above (right) the critical potential (from [15]).

-3-

The Ag (100) surface is of special scientific interest, since it reveals an order–disorder phase transition which is predicted to be second order, similar to the two dimensional Ising model in magnetism [37]. In fact, the steep intensity increase observed for potentials positive to ~ -0.76 V against Ag/AgCl for the (1,0) reflection, which is forbidden by symmetry for the clean Ag(100) surface, can be associated with the development of an ordered $(\sqrt{2} \times \sqrt{2})R45^\circ$ -Br lattice, where the bromine is located in the fourfold hollow sites of the underlying fcc (100) surface; this structure is depicted in the lower right inset in figure C2.10.1 [15].

Below the critical potential, the scattered intensity is zero. Obviously, the experimental data are in good agreement with the curve calculated using a critical exponent of 1/8 as predicted by the Ising model. The measurement of further scattering peaks gives additional information about the position of the adsorbed bromine ions for potentials cathodic to the critical potential. The determined positions are consistent with the assumption of a lattice gas adsorption at the fourfold sites, however with a large lateral displacement, especially for low coverages. The observations can be explained by unbalanced Br–Br interactions and filling of the unoccupied area with surface water [15]. While similar results were obtained for Cl^- on Ag(100) [15], no order-disorder transition, but several different commensurate and incommensurate surface structures were found for bromine adsorption on Au (100) [38] although both surfaces are in the same universality class [37].

Besides adsorption on noble metal surfaces, however, also less noble metals like copper have been investigated in halide containing electrolytes during recent years. In figure C2.10.2 the cyclic voltammogram of a (111)-oriented Cu single crystal in 10 mM HCl is given together with two STM micrographs; the latter were recorded at potentials below and above the distinct oxidation and reduction peaks, respectively [39].

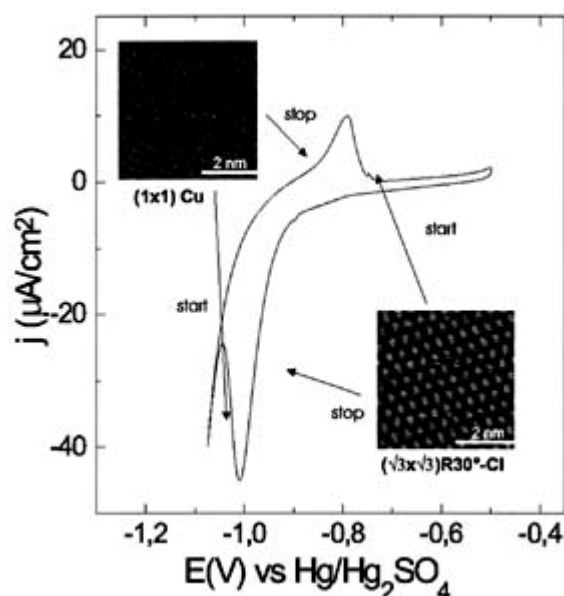


Figure C2.10.2. Cyclic voltammogram of Cu(111)/10 mM HCl and *in situ* measured STM micrographs revealing the bare Cu(111) surface (-1.05 V, left) and the $(\sqrt{3} \times \sqrt{3})R30^\circ\text{-Cl}$ adsorbate superstructure (-0.6 V, right) (from [39]).

-4-

Two clearly different surface structures are visible. Due to the lack of chemical information, however, the relation of these structures to the Cu(111) substrate or the presence of an ordered array of adsorbate molecules is very difficult for the following reasons. Firstly, it is well known from the literature that a surface reconstruction—i.e. a rearrangement or relaxation of the outer atomic layers—can be induced by a change of the electrode potential and charge; the reconstruction and its lifting are in general accompanied by small oxidation and reduction current densities (see, e.g., [40]). For example, the reconstruction observed during negative charging of Au(111) electrodes has been explained by an increasing compressive stress induced by an increasing density of s-p electrons (see [10] and references therein). Secondly, if the presence of an adsorbate is likely or expected, its nature cannot be clarified by STM experiments alone. For example, the first ordered adsorbate structures found for Au(111) in sulfate solutions were interpreted as a hydrogen bonded bisulfate layer [41], while later on, additional FTIR and radiochemical studies proved the presence of adsorbed sulfate rather than bisulfate [42, 43]. In addition, the long range ordered structure observed during the underpotential deposition of copper on gold was imaged by STM long before its composition was known: though anions were believed to be invisible to STM, these structures were ascribed to the metal deposit [44, 45 and 46] while later on it was shown by XAS [47] and SXS [48] that they are related to coadsorbed anions (see below). Furthermore, a surface reconstruction can also be affected by ionic adsorption [14, 20, 28, 40, 49], which can additionally complicate the interpretation of STM micrographs. In the presented study, *ex situ* XPS and ion scattering experiments were performed after the controlled emersion of the electrodes from the solution and their transfer to a UHV system in order to clarify the situation [39]. It has to be mentioned that UHV techniques such as LEED, XPS, UPS and ISS have been successfully applied for the *ex situ* investigation of electrode surfaces since the construction of well suited transfer cells in the 1980s [50, 51, 52, 53 and 54]; a review about this topic was given by Kolb [24]. For Cu(111) in dilute HCl solutions, only very low Cl^- surface concentrations were found for potentials negative to -0.9 V against $\text{Hg}/\text{Hg}_2\text{SO}_4$ in the anodic direction and below -1.0 V in the cathodic scan direction, while positive to the mentioned potentials strong Cl signals were found with XPS and ISS [39]. Therefore, the STM investigations reveal the unreconstructed (1×1) Cu(111) structure and the $(\sqrt{3} \times \sqrt{3})R30^\circ\text{-Cl}$ adsorbate structure, respectively [39].

As a further example for the meaning of *ex situ* investigations of emersed electrodes with surface analytical techniques, results obtained for the double layer on polycrystalline silver in alkaline solutions are presented in figure C2.10.3. This system is of scientific interest, since thin silver oxide overlayers (thickness up to about 5 nm) are formed for sufficiently anodic potentials, which implies that the adsorption of anions, cations and water can be studied on the clean metal as well as on an oxide covered surface [55, 56]. For the latter situation, a changed

adsorption behaviour can be expected due to the semiconducting properties of Ag oxides; this effect is responsible e.g. for a significant enhancement of the catalytic activity for ethene epoxidation [57]. In addition, the oxide layer can be doped with anions or cations as found for copper in alkaline chloride solutions [58]. One great advantage of XPS in this case is the possibility to determine surface concentrations of all double layer constituents; in particular the contributions of surface water, adsorbed hydroxyl ions and possible oxide signals can be fully separated as shown in figure C2.10.3(a) for an Ag electrode emersed from an alkaline perchlorate solution. The surface concentrations of the adsorbed hydroxyl ions, surface water as well as the oxide layer thickness calculated from the corresponding O^{2-} signals are presented in figure C2.10.3(b) and figure C2.10.3(c). It should be mentioned that an oxide formation was found for potentials significantly lower than the Nernst potential of Ag_2O formation [56].

-5-

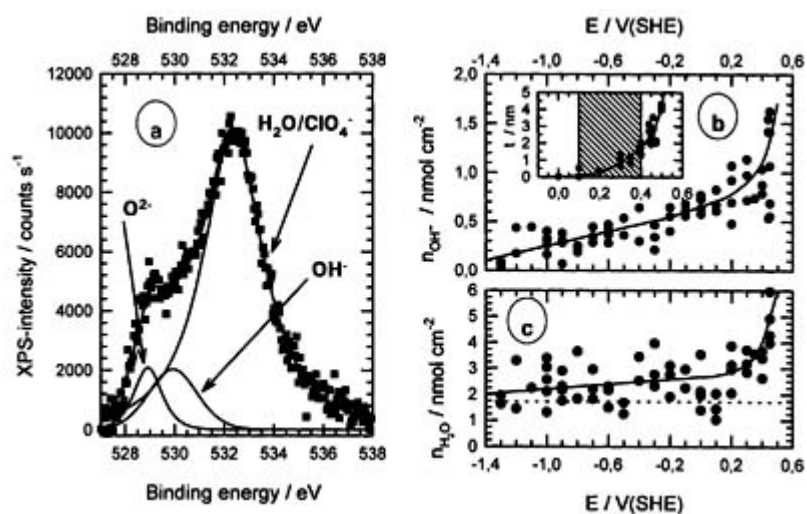


Figure C2.10.3. *Ex situ* investigation of the electrochemical double layer on Ag after hydrophobic emersion from 1 M NaClO₄ + 0.1 M NaOH. (a) Peak deconvolution of the XPS O1s signals after emersion at +0.2 V: A surface oxide as well as OH⁻ and water/perchlorate contributions can be identified. (b) Potential dependence of the OH⁻ surface concentration. The inset depicts the oxide thickness t determined from the O²⁻ signals. The hatched area visualizes the potential region where an underpotential Ag(I) oxide formation occurs. (c) Amount of adsorbed water calculated from the H₂O/ClO₄⁻ XPS signals; the perchlorate contributions were determined from the Cl 2p signals and subtracted accordingly [56]. The dashed horizontal line represents a monolayer coverage of H₂O.

In addition, the OH⁻ coadsorption significantly influences the adsorption of all other double layer components [55, 56]. For example, in chloride containing media, a significant reduction of the Cl⁻ surface concentration was found, while the concentration of cations and surface water was significantly increased compared to the emersion from acidic solutions [55]. Even the adsorption of the strongly binding iodine on silver is pH dependent [31]. The results can be explained by a specific adsorption of OH⁻ on silver [55, 56, 59]; this interpretation is consistent with the fact that appreciable amounts of adsorbed OH⁻ were also found for copper electrodes in alkaline solutions [60]. Due to the high bond strength between the OH⁻ ion and the Ag metal surface, OH⁻ is able to expel Cl⁻ from the inner Helmholtz plane [55], while it is not able to displace Br⁻ [59]. Br⁻ adsorption itself, however, suppresses Ag oxide formation [59].

The last example presented in this section deals with the pitting corrosion of Fe in ClO₄⁻ solutions. Perchlorate is less known as an aggressive ion but reveals some unique and remarkable characteristics with regard to pitting corrosion. For example, the critical pitting potential (1.46 V against a standard hydrogen electrode (SHE) for Fe/1 M NaClO₄) can be measured with an accuracy of less than 4 mV [61] which is very unexpected if compared to other aggressive ions such as Cl⁻ or Br⁻. In concentrated HClO₄, a slightly lower value of 1.37 V was found [62]. In figure C2.10.4 two Cl 2p XPS spectra obtained from an Fe electrode emersed from 1 M HClO₄ for a potential cathodic and anodic to the pitting potential E_p are compared. Obviously, only ClO₄⁻ species can be detected at a binding energy of ~208 eV (Cl 2p_{3/2} peak) for the passive iron ($E < E_n$), while clear and intense Cl⁻ contributions

were found at about 198 eV binding energy for $E > E_p$ [62]. The anionic fraction of Cl^- , i.e. the ratio of the Cl^- intensity and the sum of the chlorine and the perchlorate intensity ($I_{\text{Cl}^-} / (I_{\text{Cl}^-} + I_{\text{ClO}_4^-})$) is less than 5% for the passive Fe sample while a value of about 30% is

-6-

found for the corroded sample [62]. The experiments clearly suggest that perchlorate is decomposed at the oxide covered Fe surface for sufficiently anodic potentials, resulting in Cl^- ions which cause breakdown of passivity and finally lead to the localized corrosion phenomena. This observation is very special as a reduction of ClO_4^- is expected at more negative potentials for thermodynamic reasons. Presumably its decomposition is caused in the high field of the electrical double layer [62].

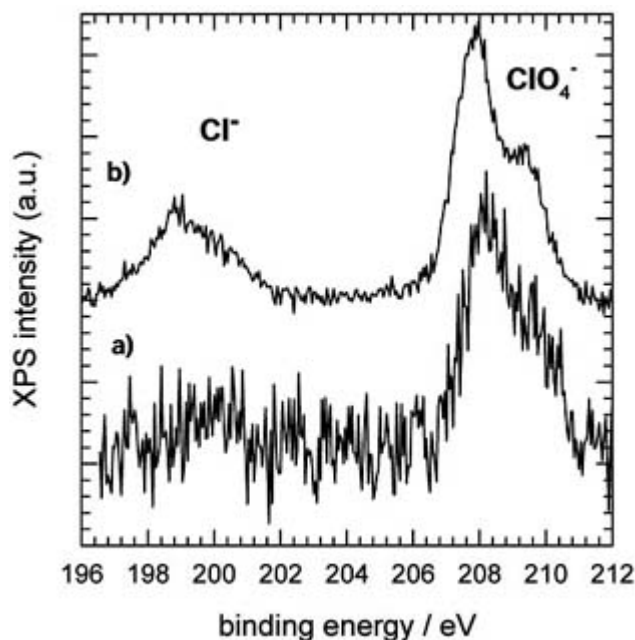


Figure C2.10.4. XPS Cl 2p signals of an iron specimen emersed from 1 M HClO_4 : (a) after passivation at 1 V (SHE); (b) after 2 minutes pitting corrosion at 1.5 V (SHE). Contributions of ClO_4^- at ~ 208 eV and Cl^- at ~ 198 eV are visible in different amounts.

C2.10.3 METAL MONOLAYER DEPOSITION

At potentials positive to the bulk metal deposition, a metal monolayer-or in some cases a bilayer-of one metal can be electrodeposited on another metal surface; this phenomenon is referred to as underpotential deposition (upd) in the literature. Many investigations of several different metal adsorbate/substrate systems have been published to date. In general, two different classes of surface structures can be classified: (a) simple superstructures with small packing densities and (b) close-packed (bulklike) or even compressed structures, which are observed for deposition of the heavy metal ions Tl, Hg and Pb on Ag, Au, Cu or Pt (see, e.g., [63, 64, 65, 66, 67, 68, 69 and 70]). In case (a), the metal adsorbate is very often stabilized by coadsorbed anions; typical representatives of this type are Cu/Au (111) (e.g. [44, 45, 71, 72 and 73]) or Cu/Pt(111) (e.g. [46, 74, 75, and 76]). It has to be mentioned that the two dimensional ordering of the Cu adatoms is significantly affected by the presence of coadsorbed anions, for example, for the upd of Cu on Au(111), the onset of underpotential deposition shifts to more positive potentials from SO_4^{2-} to Br^- and Cl^- [72].

-7-

STM measurements suggested for the bilayer formed by Cu and coadsorbed Cl^- either a (5×5) long range order similar to that of the (111) plane of CuCl or a (4×4) -based structure [46, 77, 78], while the bilayer formed with Cu and BR^- has always a (4×4) -based structure [78, 79]. Therefore, x-ray absorption spectroscopy (XAS) is a well suited technique for the investigation of these mixed overlayers since it probes the local atomic structure around an absorbing atom with a high precision [47, 80, 81]. Similar to SXS, different neighbouring atoms can easily be separated due to their different backscattering amplitudes. In addition, by the choice of the absorption edge, the central atom can also be selected. Usually, XAS spectra were recorded using photons polarized parallel to the surface in the past. However, the application of polarization resolved XAS is very promising for the investigation of highly anisotropic samples such as upd layers with coadsorbed anions [47, 82]: if the polarization of the x-rays and an investigated bond are aligned parallel to each other, the corresponding atom is visible in the XAS signal. However, with increasing bond angle, the respective contributions to the XAS decrease continuously until the atom is invisible for an angle of 90° . In [figure C2.10.5](#). Fourier transforms (FTs) of x-ray absorption spectra measured in the vicinity of the Cu K edge are displayed for several different anions using a polarization parallel to the sample surface ($E_{||}$) or parallel to the surface normal (E_{\perp}) [82]. Although the peaks in the FTs are generally shifted towards lower distances, these FTs can be interpreted as an approximation of the corresponding radial distribution functions. Due to the fact that neighbouring Cu atoms as well as the Au substrate, water and Cl, Br or S (from the sulfate) contribute to these FTs, the interpretation of the Ft data is not straightforward in the present situation, although one might expect stronger Cu–Cu peaks for $E_{||}$, while stronger Cu–Au and Cu–anion peaks are likely for E_{\perp} . Therefore, XAS multiple scattering calculations (FEFF 6.01 code [83, 84]) were performed for several different model structures of the bilayer structure: besides the Cu adatoms, up to 200 Au and anion atoms were included in these model clusters [47, 82]. For coadsorbed Cl^- , the best fit was obtained for a structural model which is very similar to the (111) plane of a CuCl crystal, in which the Cu adatoms are placed in registry with the top layer of Cl^- , while they are out of registry with the gold(111) substrate. However, the copper adatoms have to be moved out of the high symmetry positions in order to obtain better agreement between experiment and simulation [84], indicating that the bilayer is characterized by a large static disorder with a broad distribution of Cu–Cl, Cu–Au and Cu–Cu distances rather than by a single set of bond length and bond angles [47, 82].

For coadsorbed sulfate, the well ordered overlayer consists of a honeycomb ($\sqrt{3} \times \sqrt{3}$) $R30^\circ$ lattice of Cu adatoms, which corresponds to a coverage of $2/3$ [80, 81]. The sulfate anions occupy the empty centres of the honeycomb [81] and the fit of the polarization dependent XAS data is best if three oxygen ions from the sulfate are directed towards the the empty centres of the honeycomb [82]. These results are in accordance with those of an SXS study [85] and a quantum statistical model [86] and imply that STM and AFM images obtained from this system are rather images of sulfate than of upd Cu. For Cu upd on Pt(100), an ordered Br layer on top of a pseudomorphic Cu (1×1) layer was found [87]; the Cu monolayer grows with enhanced kinetics compared to the halide free solutions [48, 76, 87, 88]. At this point it should be mentioned briefly that only a few efforts have been made to deposit metal monolayers on semiconducting substrates to date (for references, see, e.g., [89, 90, 91 and 92]), despite the technological importance of the electrochemical metallization and the contamination of semiconductors with metal ions.

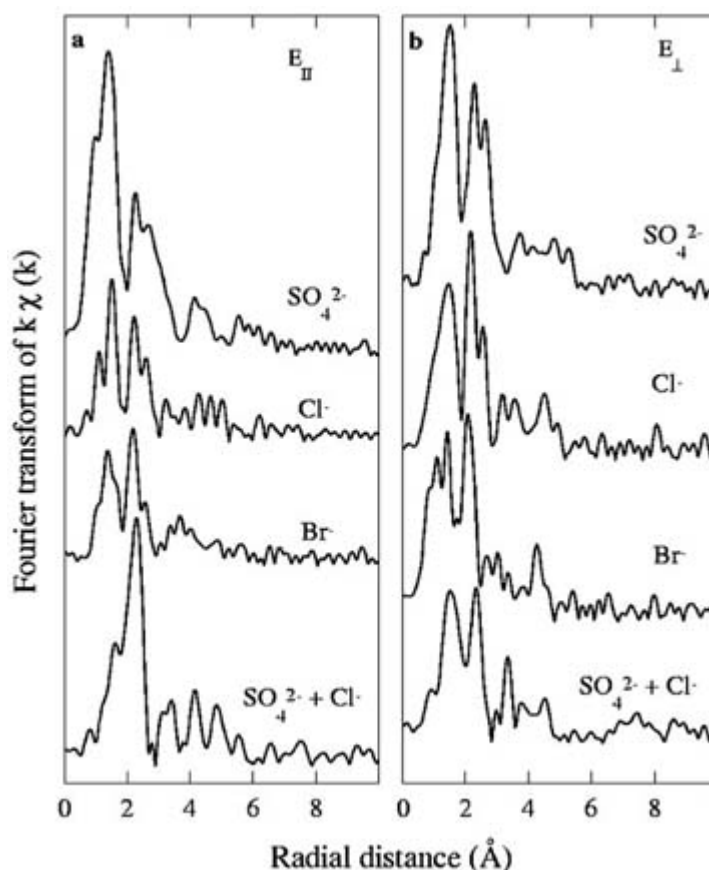


Figure C2.10.5. Magnitude of the Fourier transform of the k -weighted absorption fine structure $k\chi(k)$ measured at the Cu K edge for the underpotential deposition of Cu/Au(111) from 0.1 M $\text{KClO}_4 + 10^{-3}$ M $\text{HClO}_4 + 5 \times 10^{-5}$ M Cu $(\text{ClO}_4)_2 + 10^{-3}$ M potassium salt of sulfate, chloride, bromide and a mixture of sulfate and chloride, for polarization of the x-rays parallel to the sample surface (E_{\parallel}) or parallel to the surface normal (E_{\perp}) (from [81]).

C2.10.4 CONCLUSIONS

The presented examples clearly demonstrate that a combination of several different techniques is urgently recommended for a complete characterization of the chemical composition and the atomic structure of electrode surfaces and a reliable interpretation of the related results. Structure sensitive methods should be combined with spectroscopic and electrochemical techniques. Besides *in situ* techniques such as SXS, XAS and STM or AFM, *ex situ* vacuum techniques have proven their significance for the investigation of the electrode/electrolyte interface.

-9-

ACKNOWLEDGMENTS

The authors would like to thank Jacek Lipkowski, Ben Ocko, Klaus Wandelt and their coworkers for the provision of their data for publication in this article.

REFERENCES

- [1] Alonso C, Salvarezza R C, Vara J M and Arvia A J 1990 The mechanism of silver (I) oxide formation on polycrystalline silver in alkaline solution. Determination of nucleation and growth rates *Electrochim. Acta* **35** 489–96
- [2] Lohrengel M M 1993 Thin anodic oxide layers on aluminum and other valve metals: high field regime *Mater. Sci. Eng. R*

- [3] Jovic B M, Jovic V D and Drazic D M 1995 Kinetics of chloride ion adsorption and the mechanism of AgCl layer formation on the (111), (100) and (110) faces of silver *J. Electroanal. Chem.* **399** 197–206
- [4] Kolb D M and Schneider J 1986 Surface reconstruction in electrochemistry: Au(100)-(5 × 20), Au(111)-(1 × 23) and Au(110)-(1 × 2) *Electrochim. Acta* **31** 929–36
- [5] Uosaki K, Shen Y and Kondo T 1995 Preparation of a highly ordered Au(111) phase on a polycrystalline gold substrate by vacuum deposition and its characterization by XRD, GISXRD, STM/AFM and electrochemical measurements *J. Phys. Chem.* **99** 14 117–14 122
- [6] Itaya K 1998 *In situ* scanning tunneling microscopy in electrolyte solutions *Prog. Surf. Sci.* **58** 121–247
- [7] Nishihara C and Nozoye H 1995 Underpotential deposition of copper on Pt(S)-[n(111) × (100)] electrodes in sulfuric acid solution *J. Electroanal. Chem.* **386** 75–82
- [8] Nishihara C and Nozoye H 1995 Influence of underpotential deposition of copper with submonolayer coverage on hydrogen adsorption at the stepped surfaces Pt(955), Pt(322) and Pt(544) in sulfuric acid solution *J. Electroanal. Chem.* **396** 139–42
- [9] Taguchi S and Aramata A 1995 Voltammetric study of underpotential deposition (upd) of Zn²⁺ ions on Pt(111): effect of adsorbed anion *J. Electroanal. Chem.* **396** 131–7
- [10] Lipkowski J, Shi Z, Chen A, Pettinger B and Bilger C 1998 Ionic adsorption at the Au(111) electrode *Electrochim. Acta* **43** 2875–88
- [11] Birss V I and Smith C K 1987 The anodic behaviour of silver in chloride solutions—I. The formation and reduction of thin silver chloride films *Electrochim. Acta* **32** 259–68
- [12] Horanyi G and Rizmayer E M 1984 Radiotracer study of anion adsorption at silver electrodes in acidic medium *J. Electroanal. Chem.* **176** 339–48
- [13] Kolics A, Thomas A E and Wieckowski A 1996 ³⁶Cl-labelled and electrochemical study of chloride adsorption on a gold electrode from perchloric acid media *J. Chem. Soc. Faraday Trans.* **92** 3727–36
- [14] Wang J, Ocko B M, Davenport A J and Isaacs H I 1992 *In situ* diffraction and reflectivity studies of the Au(111)/electrolyte interface: Reconstruction and anion adsorption *Phys. Rev B* **34** 10 321–38
- [15] Ocko B M, Wang X J and Wandlowski Th 1997 Bromide adsorption on Ag(001): A potential induced two-dimensional Ising order–disorder transition *Phys. Rev. Lett.* **79** 1511–14
- [16] Adzic R R, Yeager E and Cahan B D 1977 Specular reflectance studies of bromine adsorption on gold *J. Electroanal. Chem.* **85** 267–76
- [17] Kolb D M and Franke C 1982 Surface states at the metal–electrolyte interface *Appl. Phys A* **49** 379–87

- [18] Franke C, Piazza G and Kolb D M 1989 The influence of halide adsorption on the electronic surface states of silver electrodes *Electrochim. Acta* **34** 67–73
- [19] Shi Z, Lipkowski J, Mirwald S and Pettinger B 1996 Electrochemical and second harmonic generation study of bromide adsorption at the Au(111) surface *J. Chem. Soc. Faraday Trans.* **92** 3737–46
- [20] Pettinger B, Lipkowski J and Mirwald S 1995 *In situ* SHG studies of adsorption induced surface reconstruction of Au(111)-electrodes *Electrochim. Acta* **40** 133–42
- [21] Kolb D M, Rath D L, Wille R and Hansen W N 1983 An ESCA study on the electrochemical double layer of emersed electrodes *Ber. Bunsenges. Phys. Chem.* **87** 1108–11 131
- [22] Salaita G N, Lu F, Laguren-Davidson L and Hubbard A T 1987 Structure and composition of the Ag(111) surface as a function of electrode potential in aqueous halide solutions *J. Electroanal. Chem.* **229** 1–17
- [23] Baltruschat H, Martinez M, Lewis S K, Lu F, Song D, Stern D A, Datta A and Hubbard A T 1987 Structure and composition of the Pt(s) [6(111) × (111)] step terrace surface vs. PH and potential in aqueous Br-solutions: Studies by LEED and Auger spectroscopy *J. Electroanal. Chem.* **217** 111–20
- [24] Kolb D M 1987 UHV techniques in the study of electrode surfaces *Z. Phys. Chem. NF* **154** 179–99
- [25] Hubbard A T 1988 Electrochemistry at well characterized surfaces *Chem. Rev.* **88** 633–56
- [26] Tucceri R I and Posadas D 1990 The effect of surface charge on the surface conductance of silver in surface inactive electrolytes *J. Electroanal. Chem.* **283** 159–66

- [27] Körwer D, Schumacher D and Otto A 1991 Resistance changes of thin film electrodes of silver *Ber. Bunsenges. Phys. Chem.* **95** 1484–8
- [28] Bittner A M, Wintterlin J, Beran B and Ertl G 1995 Bromine adsorption on Pt(111), (100), and (110)—an STM study in air and in electrolyte *Surf. Sci.* **335** 291–9
- [29] Tanaka S, Yau S-L and Itaya K 1995 *In situ* scanning tunneling microscopy of bromine adlayers on Pt(111) *J. Electroanal. Chem.* **396** 125–30
- [30] Ocko B M, Wang J and Watson G M 1994 The structure and electrocompression of electrodeposited iodine monolayers on Au(111) *J. Phys. Chem.* **98** 897–906
- [31] Yamada T, Ogaki K, Okubo S and Itaya K 1996 Continuous variation of iodine adlattices on Ag(111) electrodes: *In situ* STM and *ex situ* LEED studies 1996 *Surf. Sci.* **369** 321–35
- [32] Magnussen O M, Wang J X, Adzic R R and Ocko B M 1996 *In situ* x-ray diffraction and STM studies of bromide adsorption on Au(111) electrodes *J. Phys. Chem.* **100** 5500–8
- [33] Inukai J, Osawa Y, Wakisaka M, Sashikata K, Kim Y-G and Itaya K 1998 Underpotential deposition of copper on iodine-modified Pt(111): *In situ* STM and *ex situ* LEED studies *J. Phys. Chem.* **102** 3498–505
- [34] Feidenhans'l R 1989 Surface structure determination by x-ray diffraction *Surf. Sci. Rep.* **10** 105–88
- [35] Robinson I K and Tweet D J 1992 Surface x-ray diffraction 1992 *Rep. Prog. Phys.* **55** 599–651
- [36] Toney M F, Howard J N, Richter J, Borges G L, Gordon J G, Melroy O R, Wiesler D G, Yee D and Sorensen L B 1995 Distribution of water molecules at Ag(111)/electrolyte interface studied with surface x-ray scattering *Surf. Sci.* **335** 326–32
- [37] Persson B N J 1992 Ordered structures and phase transitions in adsorbate layers *Surf. Sci. Rep.* **15** 1–135
- [38] Ocko B M, Wang X J, Adzic R and Wandlowski Th 1998 Surface x-ray scattering studies of Electrosorption *Synchrotron Radiat. News* **11** 23–30
- [39] Wohlmann B, Park Z, Krufft M, Stuhlmann C and Wandelt K 1998 An *in situ* and *ex situ* study of chloride adsorption on Cu(111) electrodes in dilute HCl solutions 1998 *Colloids Surfaces A* **134** 15–19

-11-

- [40] Kolb D M 1996 Reconstruction phenomena at metal–electrolyte interfaces *Prog. Surf. Sci.* **51** 109–73
- [41] Magnussen O M, Hageboeck J, Hotlos J and Behm R J 1992 *In situ* scanning tunneling microscopy observations of a disorder–order phase transition in hydrogensulphate adlayers on Au(111) *Faraday Discuss.* **94** 329–38
- [42] Shi Z, Lipkowski J, Gamboa M, Zelenay P and Wieckowski A 1994 Investigations of SO_4^{2-} adsorption at the Au(111) electrode by chronocoulometry and radiochemistry *J. Electroanal. Chem.* **366** 317–26
- [43] Eden G J, Gao X and Weaver M J 1994 The adsorption of sulphate on gold(111) in acidic aqueous media: Adlayer structural interferences from infrared spectroscopy and scanning tunneling microscopy *J. Electroanal. Chem.* **375** 357–66
- [44] Magnussen O M, Hotlos J, Nichols R J, Kolb D M and Behm R J 1990 Atomic structure of Cu adlayers on Au(100) and Au(111) electrodes observed by *in situ* scanning tunneling microscopy *Phys. Rev. Lett.* **64** 2929–32
- [45] Hachiya T, Honobo T and Itaya K 1991 Detailed underpotential deposition of copper on gold (111) in aqueous solutions *J. Electroanal. Chem.* **315** 275–91
- [46] Behm R J, Hotlos J and Magnussen O M 1995 Effect of trace amounts of Cl in Cu underpotential deposition on Au(111) in perchlorate solutions: An *in situ* scanning tunneling microscopy study *Surf. Sci.* **335** 129–44
- [47] Wu S, Lipkowski J, Tylliszczak T and Hitchcock A P 1997 Early stages of copper electrocrystallization: electrochemical and *in situ* x-ray absorption fine structure studies of coadsorption of copper and chloride at the Au(111) electrode surface *J. Phys. Chem. B* **101** 10 310–22
- [48] Tidswell I M, Lucas C A, Markovic N M and Ross P N 1995 Surface structure determination using anomalous x-ray scattering: Underpotential deposition of copper on Pt(111) *Phys. Rev. B* **51** 10 205–8
- [49] Ocko B M, Magnussen O M, Wang J, Adzic R R, Shi Z and Lipkowski J 1994 A critical comparison of electrochemical and surface x-ray scattering results at the Au(111) electrode in KBr solutions *Electroanal. Chem.* **376** 35–9
- [50] O'Grady W E, Woo M Y C, Hagans P L and Yeager E 1997 Electrode surface studies by LEED–Auger *J. Vac. Sci. Technol.* **14** 365–8
- [51] Felter T E and Hubbard A T 1979 LEED and electrochemistry of iodine on Pt(100) and Pt(111) single crystal surfaces

- [52] Neff H, Foditsch W and Kötzt R 1984 An electrochemical preparation chamber for the Kratos ES 300 electron spectrometer *J. Electron. Spectrosc. Relat. Phenom.* **33** 171–4
- [53] Haupt S, Collisi U, Speckmann H D and Strehblow H-H 1985 Specimen transfer from the electrolyte to the UHV in a closed system and some examinations of the double layer on Cu *J. Electroanal. Chem.* **194** 179–90
- [54] Richarz F, Wohlmann B, Vogel U, Hoffschulz H and Wandelt K 1995 Surface and electrochemical characterization of PtRu alloys *Surf. Sci.* **335** 361–71
- [55] Hecht D and Strehblow H-H 1997 XPS investigations of the electrochemical double layer on silver in alkaline chloride solutions *J. Electroanal. Chem.* **440** 211–17
- [56] Lützenkirchen-Hecht D and Strehblow H-H 1998 Surface analytical investigations of the electrochemical double layer on silver electrodes in alkaline media *Electrochim. Acta* **43** 2957–68
- [57] Haul R, Hoge D, Neubauer G and Zeeck U 1982 Ethene epoxidation on silver oxide surface layers *Surf. Sci.* **122** L622–8
- [58] Modestov A D, Zhou G-D, Ge H-H and Loo B H 1995 A study by voltammetry and the photocurrent response method of copper electrode behavior in acidic and alkaline solutions containing chloride ions *J. Electroanal. Chem.* **380** 63–8
- [59] Lützenkirchen-Hecht D and Strehblow H-H 1998 Bromide adsorption on silver in alkaline solution: A surface analytical study *Ber. Bunsenges. Phys. Chem.* **102** 826–32
- [60] Härtinger S, Pettinger B and Doblhofer K 1995 Cathodic formation of a hydroxyl adsorbate on copper (111) electrodes in alkaline electrolyte *J. Electroanal. Chem.* **397** 335–8

- [61] Strehblow H-H and Titze B 1977 Pitting potentials and inhibition potentials of iron and nickel for different aggressive and inhibiting anions *Corr. Sci.* **17** 461–72
- [62] Prinz H and Strehblow H-H 1998 Investigations on pitting corrosion of iron in perchlorate electrolytes *Corr. Sci.* **40** 1671–83
- [63] Melroy O R, Toney M F, Borges G L, Samant M G, Kortright J B, Ross P N and Blum L 1989 An *in situ* grazing incidence x-ray scattering study of the initial stages of electrochemical growth of lead on silver(111) *J. Electroanal. Chem.* **258** 403–14
- [64] Müller U, Carnal D, Siegenthaler H, Schmidt E, Lorenz W J, Obretenov W, Schmidt U, Staikov G and Budevski E 1992 Superstructures of Pb monolayers electrochemically deposited on Ag(111) *Phys. Rev. B* **46** 12 899–901
- [65] Toney M F, Gordon J G, Samant M G, Borges G L, Melroy O R, Yee D and Sorensen L B 1992 Underpotentially deposited thallium on silver (111) by *in situ* surface x-ray scattering *Phys. Rev. B* **45** 9362–74
- [66] Chen C-H, Washburn N and Gewirth A A 1993 *In situ* atomic force microscope study of Pb underpotential deposition on Au(111): Structural properties of the catalytically active phase *J. Phys. Chem.* **97** 9754–60
- [67] Carnal D, Oden P I, Müller U, Schmidt E and Siegenthaler H 1995 *In situ* STM investigation of Tl and Pb underpotential deposition on chemically polished Ag(111) electrodes *Electrochim. Acta* **40** 1223–35
- [68] Brisard G, Zenati E, Gasteiger H A, Markovic N M and Ross P N 1995 Underpotential deposition of lead on copper (111): A study using a single crystal rotating ring disk electrode and *ex situ* low-energy electron diffraction and Auger electron spectroscopy *Langmuir* **11** 2221–32
- [69] Adzic R R, Wang J X, Magnussen O M and Ocko B M 1996 The structure of Tl adlayers on the Pt(111) electrode surface: effects of solution pH and bisulphate adsorption 1996 *J. Phys. Chem.* **100** 14 721–5
- [70] Li J and Abruna H D 1997 Coadsorption of sulphate/bisulphate anions with Hg cations during Hg underpotential deposition on Au (111): An *in situ* x-ray diffraction study *J. Phys. Chem. B* **101** 244–52
- [71] Blum L, Abruna H D, White J, Gordon J G, Borges G L, Samant M G and Melroy 1986 Study of underpotentially deposited copper on gold by fluorescence detected surface EXAFS *J. Chem. Phys.* **85** 6732–8
- [72] Shi Z, Wu S and Lipkowski J 1995 Coadsorption of metal atoms and anions: Cu UPD in the presence of SO₄ Cl and Br *Electrochim. Acta* **40** 9–15
- [73] Gordon J G, Melroy O R and Toney M F 1995 Structure of metal–electrolyte interfaces: copper on gold(111), water on silver(111) *Electrochim. Acta* **40** 3–8
- [74] Kolb D M, Kötzt R and Yamamoto K 1979 Copper monolayer formation on platinum single crystal surfaces: Optical and

electrochemical studies *Surf. Sci.* **87** 20–30

- [75] Kolb D M 1988 Structural investigations of electrode surfaces *Ber. Bunsenges. Phys. Chem.* **92** 1175–87
- [76] Markovic N M, Lucas C A, Gasteiger H A and Ross P N 1997 The structure of adsorbed bromide concurrent with the underpotential deposition (upd) of Cu on Pt(111) *Surf. Sci.* **372** 239–54
- [77] Batina N, Will T and Kolb D M 1992 Study of the initial stages of copper deposition by in situ STM *Faraday Discuss.* **94** 93–106
- [78] Matsumoto H, Inukai J and Ito M 1994 Structures of copper and halides on Pt(111), Pt(100) and Au(111) electrode surfaces studied by *in situ* scanning tunneling microscopy *J. Electroanal. Chem.* **379** 223–31
- [79] Ikemiya N, Miyaoka S and Hara S 1994 Observation of the Cu(1 × 1) adlayer on Au(111) in a sulfuric acid solution using atomic force microscopy *Surf. Sci.* **311** L641–8
- [80] Shi Z and Lipkowski J 1994 Coadsorption of Cu^{2+} and SO_4^{2-} at the Au(111) electrode *J. Electroanal. Chem.* **365** 303–9
- [81] Wu S, Lipkowski J, Tyliczszak T and Hitchcock H P 1995 Effect of anion adsorption on early stages of copper electrocrystallization at Au(111) surface *Prog. Surf. Sci.* **50** 227–36
- [82] Tyliczszak T, Hitchcock A, Wu S, Chen A, Szymanski G and Lipkowski J 1998 X-ray absorption studies of mixed overlayers formed by copper adatom co-adsorbed with anions at the Au(111) electrode surface *Synchrotron Radiat. News* **11** 31–8

-13-

- [83] Rehr J J, Mustre de Leon J, Zabinski S I and Albers R C 1991 Theoretical x-ray absorption fine structure standards *J. Am. Chem. Soc.* **113** 5135–40
- [84] Rehr J J, Albers R C and Zabinski S I 1992 High order multiple scattering calculation of x-ray absorption fine structure *Phys. Rev. Lett.* **69** 3397–400
- [85] Toney M F, Howard J N, Richer J, Borges G L, Gordon J G, Melroy O R, Yee D and Sorenson L B 1995 Electrochemical deposition of copper on a gold electrode in sulfuric acid: Resolution of the interfacial structure *Phys. Rev. Lett.* **75** 4472–5
- [86] Huckaby D A and Blum L 1991 A model for sequential first-order phase transitions occurring in the underpotential deposition of metals *J. Electroanal. Chem.* **315** 255–61
- [87] Markovic N M, Grgur B N, Lucas C A and Ross P N 1998 Upd of Cu on Pt(100): effects of anions on adsorption isotherms and interface structures *Electrochim. Acta* **44** 1009–17
- [88] Markovic N M, Gasteiger H A and Ross P N 1995 Copper electrodeposition on Pt(111) in the presence of chloride and (bi)sulphate: Rotating ring–Pt(111) disk electrode studies *Langmuir* **11** 4098–108
- [89] Zegenhagen J, Kazimirov A, Scherb G, Kolb D M, Smilgies D-M and Feidenhans'l R 1996 X-ray diffraction study of a semiconductor/electrolyte interface: n-GaAs(001)/H₂SO₄(:Cu) 1996 *Surf. Sci.* **352–354** 346–51
- [90] Koinuma M and Uosaki K 1996 Atomic structure of bare p-GaAs(100) and electrodeposited Cu on p-GaAs (100) surfaces in H₂SO₄ solutions: An AFM study *J. Electroanal. Chem.* **409** 45–50
- [91] Vereecken P M, Vanden Kerchove F and Gomes W P 1996 Electrochemical behaviour of (100) GaAs in copper(II) containing solutions *Electrochim. Acta* **41** 95–107
- [92] Scherb G, Kazimirov A, Zegenhagen J, Lee T L, Bedzyk M J, Noguchi H and Uosaki K 1998 *In situ* x-ray standing wave analysis of electrodeposited Cu monolayers on GaAs(001) *Phys. Rev. B* **58** 10 800–5

FURTHER READING

Gewirth A A and Siegenthaler H (eds) 1995 *Nanoscale Probes of the Solid/Liquid Interface (NATO ASI Series 288)* (London: Kluwer)

A survey of applications of scanning probes to electrochemical problems.

Gewirth A A and Niece B K 1997 Electrochemical applications of *in situ* scanning probe microscopy *Chem. Rev.* **97** 1129–62

Up to date summary of scanning probe studies with many literature and examples.

Abruna H D 1991 *Electrochemical Interfaces: Modern Techniques for In Situ Interface Characterization* (New York: VCH)

Comprehensive introduction into *in situ* techniques for the investigation of the electrochemical interface.

Melendres C A and Tadjeddine A (eds) 1994 *Synchrotron Techniques in Interfacial Electrochemistry (NATO ASI Series C432)* (Dordrecht: Kluwer)

Survey of the application of synchrotron radiation to the investigation of electrochemical problems.

Himpel F J, Akatsu H, Carlisle J A, Sutherland D G J, Jimenez I, Terminello L J, Jia J J, Callcott T A, Samant M G,

Stöhr J, Ederer D L, Perera R C C, Tong W and Shunh D K 1995 Surface and interface analysis at 3rd generation light sources *Prog. Surf. Sci.* **50** 37–51

Brief description of new possibilities for surface investigations with highly brilliant synchrotron x-ray sources.

-1-

C2.11 Ceramic processing

Kevin G Ewsuk

C2.11.1 INTRODUCTION

Ceramics represent a unique class of materials that are distinguished from common metals and plastics by their: (1) high hardness, stiffness, and good wear properties (i.e. abrasion resistance); (2) ability to withstand high temperatures (i.e. refractoriness); (3) chemical durability; and (4) electrical properties that allow them to be electrical insulators, semiconductors, or ionic conductors. Ceramics can be broken down into two general categories, traditional and advanced ceramics. Traditional ceramics include common household products such as clay pots, tiles, pipe, and bricks; porcelain china, sinks, and electrical insulators; and thermally insulating refractory bricks for ovens and fireplaces. Advanced ceramics, also referred to as ‘high-tech’ ceramics, include products such as spark-plug bodies, piston rings, catalyst supports, and water-pump seals for automobiles; thermally insulating tiles for the space shuttle; sodium vapour lamp tubes in street lights; and the capacitors, resistors, transducers, and varistors in the solid-state electronics we use daily.

The major differences between traditional and advanced ceramics are in the processing tolerances and cost. Traditional ceramics are manufactured with inexpensive raw materials, are relatively tolerant of minor process deviations, and are relatively inexpensive. Advanced ceramics are typically made with more refined raw materials and processing to optimize a given property or combination of properties (e.g. mechanical, electrical, dielectric, optical, thermal, physical, and/or magnetic) for a given application. Advanced ceramics generally have improved performance and reliability over traditional ceramics, but are typically more expensive. Additionally, advanced ceramics are typically more sensitive to the chemical and physical defects present in the starting raw materials, or those that are introduced during manufacturing.

In general, ceramic manufacturing involves creating fine particle size powders, forming powders into a particulate compact, and heat treating (i.e. sintering) that compact to produce a cohesive body with the desired microstructure and properties for a given application. Because powder systems have a relatively large total surface area for their mass, surfaces and interfaces are very important in ceramic processing. From the perspective of a physical chemist, ceramic processing involves understanding and controlling the physical chemistry of surfaces and interfaces [1, 2].

Initially in ceramic powder processing, particle surfaces are created that increase the surface energy of the system. During shape forming, surface/interface energy and interparticle forces are controlled with surface active additives.

Ultimately, the surface energy is used to produce a cohesive body during sintering. As such, surface energy, which is also referred to as surface tension, γ , is obviously very important in ceramic powder processing. Surface tension causes liquids to form spherical drops, and allows solids to preferentially adsorb atoms to lower the free energy of the system. Also, surface tension creates pressure differences and chemical potential differences across curved surfaces that cause matter to move.

The Laplace equation, which defines the pressure difference, ΔP , across a curved surface of radius, r ,

-2-

$$\Delta P = \gamma(1/r_1 + 1/r_2) \quad (C2.11.1)$$

has been characterized as the fundamental equation of capillarity [1]. In ceramic processing, the pressure associated with surface tension and capillary forces contribute to, among other things, particle clustering (i.e. agglomeration) and rearrangement, to the migration of liquids through pores during mixing, shape forming, and drying, and to pore shrinkage during sintering.

The equilibrium vapour pressure, P , over a curved surface is defined by the Kelvin equation

$$\ln P/P_0 = 2\gamma\Omega/rkT \quad (C2.11.2)$$

where P_0 is the equilibrium vapour pressure over a planar surface, Ω is the molecular volume of the condensed phase, k is Boltzmann's constant, and T is the absolute temperature. Because the chemical potential difference, $\Delta\mu$, between a curved and flat surface is related to the vapour pressures over those respective surfaces,

$$\Delta\mu = kT \ln P/P_0 \quad (C2.11.3)$$

chemical potential is also related to surface curvature:

$$\Delta\mu = \mu - \mu_0 = 2\gamma\Omega/r. \quad (C2.11.4)$$

The chemical potential of a curved surface is extremely critical in ceramic processing. It determines reactivity, the solubility of a solid in a liquid, the rate of liquid evaporation from solid surfaces, and material transport during sintering.

This chapter will describe some of the basic unit processes in ceramic manufacturing, and will touch on the pertinence of the physical chemistry of surfaces in selected unit processes. For a more comprehensive review of ceramics and ceramic processing, the reader is referred to other sources [3, 4 and 5].

C2.11.2 POWDER PROCESSING

Ceramic manufacturing involves multiple unit process steps ranging from raw materials beneficiation to finish machining (figure C2.11.1). Ceramics are fabricated using raw materials, typically in powder form, that are generally beneficiated to improve their handling and processability. The desired size and shape ceramic component is produced by consolidating powder in a process that generally involves a forming pressure. Ultimately, this powder compact is heat-treated (i.e. sintered) to form a cohesive body.

-3-

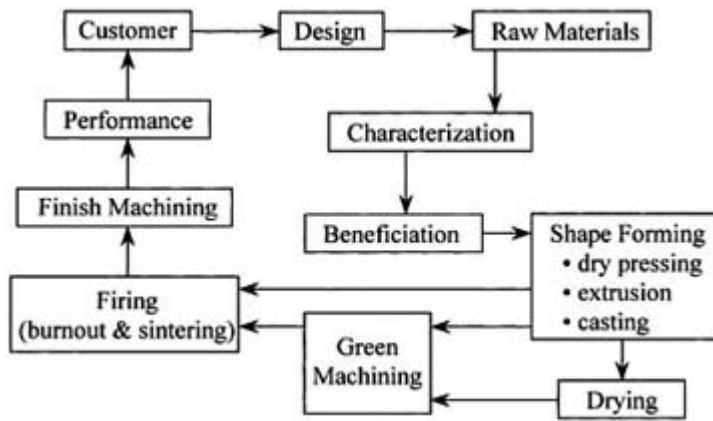


Figure C2.11.1. A flow chart summarizing the ceramic design and manufacturing process.

The fabrication of an alumina spark-plug body is a good example of ceramic manufacturing. The manufacturing process begins with an alumina powder (figure C2.11.2) comprised of individual alumina particles of the desired size distribution. To enhance densification, precursors of CaO , MgO , and SiO_2 are typically mixed with the alumina powder to produce several weight per cent of a glass phase during sintering. This mixture is then transformed into a slurry of ceramic particles dispersed in water, which is subsequently granulated with an organic binder by spray drying. Spray drying produces larger clusters of particles called agglomerates or granules that have improved powder flow, packing, and formability (figure C2.11.3). These granules are then pressed and machined to produce a powder compact of the desired size and shape that is held together by the organic additives. Finally, this powder compact is sintered to produce a dense ceramic spark-plug body (figure C2.11.4). The mechanical and electrical properties of this body are determined by the microstructure of the polycrystalline alumina ceramic produced on sintering (figure C2.11.5).

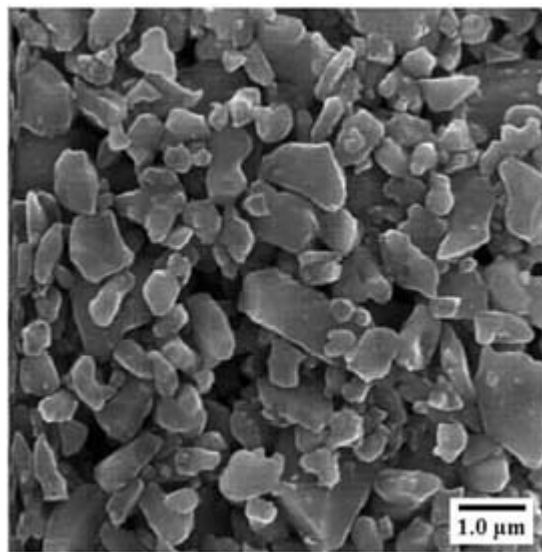


Figure C2.11.2. A scanning electron micrograph showing individual particles in a polycrystalline alumina powder.

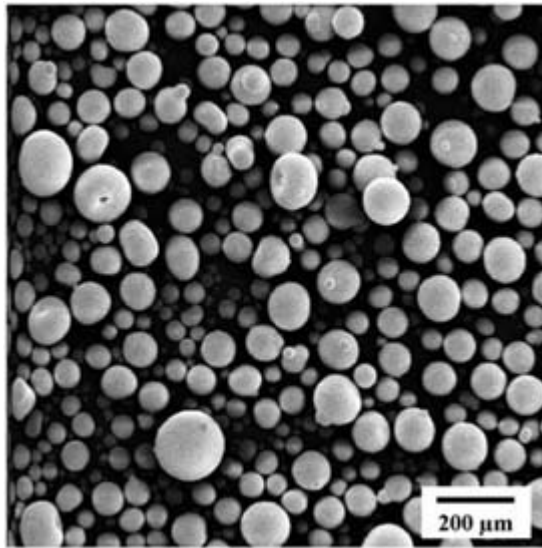


Figure C2.11.3. A scanning electron micrograph of the spherical alumina granules produced by spray drying a ceramic slurry. The granules are comprised of individual alumina particles, sintering additives, and an organic binder.



Figure C2.11.4. A commercial spark plug with its electrically insulating ceramic body comprised of alumina and glass (white portion).

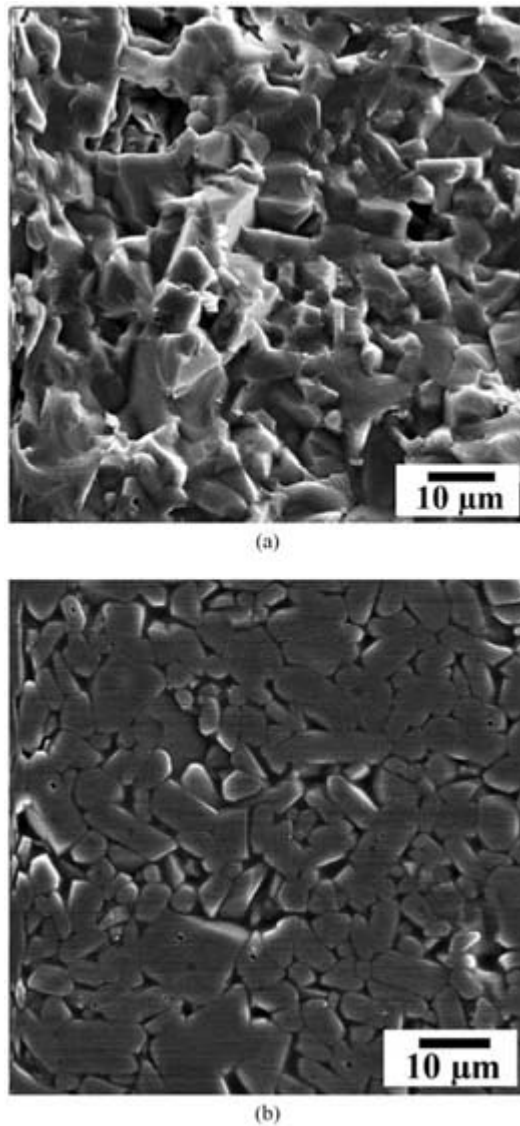


Figure C2.11.5. Scanning electron micrographs showing the microstructure of an alumina ceramic spark-plug body (a) fracture surface and (b) polished and thermally etched cross section.

There is unquestionably a substantial engineering component in manufacturing ceramics. There is also a very critical scientific component that involves understanding and controlling the physical chemistry of surfaces. Not only are a number of different unit process steps required to manufacture a ceramic, but each unit process has its own set of requirements for optimization. Often, the requirements to optimize one step are diametrically opposed to those for another unit process. This necessitates compromise in order to optimize the complete manufacturing process. For example, while a fine-particle-size powder provides a high surface area and driving force for sintering, electrostatic attraction and van der Waals forces promote agglomeration of fine particles and make them difficult to mix, pack, and compact. As a compromise, a practical lower limit of $\sim 0.1 \mu\text{m}$ diameter particles, is typical in advanced ceramic powder processing.

C2.11.2.1 RAW MATERIALS

Ceramic processing generally starts with ceramic powders that range from relatively impure, naturally occurring clays, to ultra-high-purity, controlled morphology powders. Inexpensive, mined raw materials are typically used to manufacture traditional, high-volume production, ceramics [6, 7, 8, 9, 10, 11 and 12]. Chemically synthesized ceramic powders, which are often considerably more expensive, are used to manufacture high-cost, lower-volume,

advanced ceramics [13, 14, 15, 16, 17, 18, 19 and 20].

Naturally occurring ceramic raw materials such as silica (SiO_2 , or quartz), silicates (e.g. talc) and aluminosilicates (e.g. clays) are generally mined from the Earth's surface. In contrast, nanometre size, controlled chemistry ceramic powders are produced from high-purity, specialty chemicals: the precipitation of solids from liquid solutions is one example. Precipitation occurs by a combination of nucleation and growth, both of which occur to lower the free energy of the system. The system's desire to minimize its surface energy per unit volume determines the shape of the precipitate. Most powders used in the manufacture of advanced ceramics fall somewhere between these two extremes. For example, reasonably pure ceramic powders can be formed by reacting constituent oxide powders and/or salts at an elevated temperature (i.e. calcining). Barium titanate, BaTiO_3 , which is used to make capacitors in solid-state electronics, can be produced by mixing BaCO_3 with TiO_2 and calcining. High-surface-area fine powders promote rapid and complete reaction of the constituent powders to produce the desired compound.

C2.11.2.2 BENEFICIATION

Beneficiation is the process or processes whereby the chemical and/or physical properties and characteristics of a raw material are modified to make it more processable. Particle-size reduction (i.e. comminution) using mechanical energy may be the most common process [21, 22, 23, 24, 25, 26, 27 and 28]. Crushing, grinding, and/or milling create new surfaces by breaking down aggregates (i.e. clusters of tightly bound particles) and by fracturing particles. Comminution produces the desired size distribution powder for subsequent processing.

After comminution, soluble impurities either inherent to the raw materials or introduced during processing can be extracted by washing (e.g. with water), followed by filtration [6, 23]. Chemical leaching and magnetic separation are also used to purify raw materials. In a more specialized process, a frothing agent can be used to promote differential adsorption of impurity particles onto gas bubbles to separate out the desired product [23].

C2.11.2.3 GRANULATION

In dry powder processing, after the desired particle size and chemistry are obtained, the powder is generally granulated. Powders comprised of micrometre-size particles are difficult to handle due to interparticle forces. Granulation transforms individual particles into agglomerates with controlled size, shape, and strength, to improve the flow, packing, and compaction behaviour of powders in ceramic processing [29, 30 and 31]. Granules are formed by spraying a liquid or a binder solution directly into a tumbling powder, or by spray drying a slurry in a heated chamber. In the former, granules form under the influence of the capillary forces between the liquid–solid (particle) interfaces. In spray drying, a combination of (liquid) surface tension and interparticle forces produce ~50–300 μm diameter granules. Due to liquid surface tension, the atomization of a slurry produces spherical droplets that subsequently dry to form spherical granules. Capillary forces hold the individual particles together within the granule during the drying, while van der Waals forces and bonds from the organic additives adsorbed onto particle surfaces hold the dry granule together.

C2.11.2.4 FORMING ADDITIVES

Immediately prior to, during, and/or immediately following granulation, forming additives or processing aids are commonly added to a ceramic powder to enhance processing [32, 33, 34, 35, 36 and 37]. Organic additives adsorb onto the surfaces of ceramic particles to modify surface energy and particle–particle interactions. Two common additives used in ceramic processing are binders and lubricants. Organic binders, which are also referred to as coagulants and flocculants, are polymer molecules or colloids that adsorb onto particle surfaces and promote interparticle bridging (i.e. flocculation). The main purpose of a binder is to provide strength to the powder compact after shape forming, which may be necessary for subsequent handling and/or green machining. Binders are used extensively in dry powder pressing operations, and are also added to extrusion bodies and to pastes.

Lubricants are added to lower interfacial frictional forces between individual particles and/or between particles and forming die surfaces to improve compaction and ejection (i.e. extraction of the pressed compact from the forming die). Individual particle surfaces can be lubricated by an adsorbed film that produces a smoother surface and/or decreases interparticle attraction. Forming (die) surfaces can be lubricated by coating with a film of low-viscosity liquid such as water or oil.

B1.20.3 SHAPE FORMING

Ceramic forming typically involves using pressure to compact and mould particles to the desired size and shape. Ceramics can be formed from slurries, pastes, plastic bodies (i.e. such as a stiff mud), and from wet and dry powders.

C2.11.3.1 SLURRIES

In preparation for the shape forming process, ceramic particles can be dispersed in a liquid. The dispersion of solid particles in a liquid is known as a slurry, and is often referred to as a suspension or a dispersion. Forming a slurry involves (1) wetting the solid particle surface with the liquid (i.e. replacing the solid–vapour interfacial area with solid–liquid interfaces), (2) breaking down agglomerates, and (3) controlling particle surface charge to prevent flocculation or reagglomeration [38, 39, 40, 41 and 42]. To optimize dispersion and stability, dispersants (also known as deflocculants or anticoagulants) are often added to slurries [33, 36]. Deflocculants prevent dispersed particles from reagglomerating in a slurry by keeping particle–particle separation distances sufficiently large such that the short range van der Waals attractive forces that will hold particles together are negligible. Particle separation is maintained by the steric effect of the preferential adsorption of large deflocculant molecules on the particle surfaces. Electrostatic stabilization is achieved through the use of the electrical double layer that forms around particles such that neighbouring particles are repelled from one another by like (negative or positive) surface charges. Deflocculation by electrostatic stabilization is common in clay slurries as well as with ceramic particles dispersed in polar liquids (e.g. water).

The use of acids and bases to control interparticle forces in oxide (ceramic)–water suspensions is an example of electrostatic stabilization. Hydroxylated oxide surfaces react with H^+ (acid) or OH^- (base) by surface ionization to become positively charged (low pH) or negatively charged (high pH), respectively. The like-charged particle surface layers repel neighbouring particles and stabilize the solution. A stable dispersion is produced by progressively adding

acid or base to a system to increase the particle surface charge such that the long-range electrostatic repulsive forces dominate over the short-range van der Waals attractive forces. Conversely, an oxide slurry can be induced to flocculate by adjusting the system pH to the point where there is no net charge on a particle's surface. The pH at which this occurs is defined as the point of zero charge (PZC) for the oxide. At the PZC, the electrostatic repulsive forces are eliminated and the van der Waals attractive forces take over, causing flocculation. Surfactants or wetting agents offer another means to improve dispersion in a slurry. By reducing the surface tension of a liquid, a wetting agent decreases the solid–liquid interfacial energy, making it more favourable for the liquid to coat the solid particles. A low-surface-tension surfactant also makes a good antifoam agent.

C2.11.3.2 CASTING

Slurry or slip casting provides a relatively inexpensive way to fabricate uniform-thickness, thin-wall, or large cross section shapes [43, 44, 45, 46, 47 and 48]. For slip casting, a slurry is first poured into a porous mould. Capillary suction then draws the liquid from the slurry to form a higher solids content, close-packed, leather-hard cast on the inner surface of the mould. In a fixed time, a given wall thickness is formed, after which the excess slurry is drained.

Electrophoretic deposition (EPD) is another method of casting slurries. EPD is accomplished through the controlled migration of charged particles under an applied electric field. During EPD, ceramic particles typically deposit on a mandrel to form coatings of limited thickness, or thin tubular shapes such as solid $\beta'''' - \text{Al}_2\text{O}_3$ electrolytes for sodium–sulfur batteries.

C2.11.3.3 DRYING

After casting, the residual liquid in the ceramic part must be removed by drying [49, 50, 51, 52, 53 AND 54]. It is important to achieve relatively uniform drying throughout the body in order to avoid the excessive differential (capillary) stresses and stress gradients that can result in drying cracks (i.e. like those formed in a mud puddle) and warping (i.e. like that seen in lumber on drying). Air drying by convection and conduction is the most common means of drying ceramic ware, whereby drying occurs by liquid evaporation at the drying front. Initially, the drying front starts at the ware surface and then moves into the part. During drying, liquid migrates to the drying front by capillary flow, chemical diffusion, and/or thermal diffusion at a rate determined by the permeability of the ware. The size of the porosity in the ceramic body, the viscosity and surface tension of the liquid, and the moisture gradient from inside the body to the drying front determine the permeability. When liquid migration cannot keep pace with the evaporation process, the drying front moves from the surface into the body, where drying continues by evaporation from the menisci of the liquid within the pores. Large pores and interstices are emptied in preference to smaller pores, and the large capillary stresses produced as the menisci recede into fine pores can result in cracking. The finer the particles, the greater the drying shrinkage, and the greater the capillary stresses during drying.

Stresses during drying can be minimized by controlled humidity drying, by supercritical drying, or by freeze drying. Controlled humidity drying utilizes a high-humidity atmosphere during the critical, initial stage of drying to maintain a liquid film on the (solid) surface of the ware (i.e. a solid–liquid interface). Supercritical drying is accomplished by heating ware under pressure in an autoclave until the liquid becomes a supercritical fluid (i.e. both a liquid and a vapour simultaneously), after which drying can be accomplished by isothermal depressurization to remove the vapour. Supercritical drying is often used to avoid generating catastrophic capillary stresses during the drying of fine-pore materials such as gels. Freeze drying makes use of freezing and sublimation to minimize drying stresses due to capillarity. In freeze drying, the temperature of the ware is initially decreased to below the freezing point of the liquid.

-9-

Then the pressure is reduced to transform the frozen liquid to a vapour and to remove it. Freeze drying is commonly used to make powders that are not agglomerated.

C2.11.3.4 POWDER PRESSING

Powder pressing may be the most common method of forming ceramic components [55, 56, 57, 58 and 59]. Dry pressing, also referred to as mechanical pressing, is an economical, yet versatile technique for fabricating small, relatively simple-shape powder compacts. Automated dry pressing, which is used extensively in the production of pharmaceutical tablets, is capable of producing 5000 ceramic parts per minute. Dry pressing involves compacting a, typically granulated, ceramic powder between two plungers in a die cavity. Friction between the powder and the die walls must be controlled during forming to minimize pressing pressure gradients that can create defects in the form of density gradients and/or cracking in a pressed powder compact. Friction can be controlled using lubricants during forming or through the design of the die (i.e. materials and geometry).

C1.11.4 THERMAL PROCESSING

Generally, the last step in ceramic component manufacturing is thermal processing [60, 61, 62 and 63]. This is the stage where the weakly-bound particulate body produced during shape forming is heat treated to produce a

cohesive body with the desired properties for its end-use application. Thermal consolidation, which is more commonly referred to as ‘firing’, typically involves two steps, burnout and sintering. Generally, both are accomplished in a single firing process with burnout preceding sintering.

C2.11.4.1 BURNOUT

The burnout stage involves eliminating the organic processing aids and any residual organic impurities or water prior to sintering [60, 61, 62 and 63]. Minor concentrations of residual liquid used in forming, and physically adsorbed moisture on particle surfaces can be eliminated on heating to ~200 °C. Most organic binders used in ceramic forming are physically adsorbed onto particle surfaces, and can be burned out by heating to ~500 °C. Clays such as kaolin must be heated to 700 °C to liberate the water of crystallization and produce the desired dehydrated aluminosilicate phase for subsequent processing. The decomposition of constituents such as sintering aids, which may be added in the form of a salt precursor, may require temperatures up to ~900 °C. Temperatures in excess of 1000°C may be required to completely eliminate chemically adsorbed water on fine-particle surfaces.

C2.11.4.2 RAW MATERIALS

Sintering involves the densification and microstructure development that transforms the loosely bound particles in a powder compact into a dense, cohesive body [60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72 and 73]. The end-use properties of a finished ceramic are largely dependent on the degree of densification achieved during sintering, and on the microstructure produced; consequently, sintering is one of the most critical steps in ceramic processing. Sintering, which is often considered to be synonymous with densification, is usually accomplished by heating a powder compact to approximately two-thirds of its melting temperature for a given time. Sintering can also occur by subjecting a powder compact to externally applied pressure, or heat and pressure simultaneously (e.g. hot pressing and hot isostatic

pressing). A ceramic densifies during sintering as the porosity (i.e. void space) between the solid particles is reduced in size with time. Concurrently, the cohesiveness of the body increases as interparticle contact (i.e. grain boundary) area increases during sintering.

(a) Driving force for sintering. Ceramic powder compacts sinter as a result of the thermodynamic driving force to minimize the Gibbs’ free energy, G , of a system [61, 62, 63 and 64, 74]. This includes minimizing the volume, interfacial, and surface energy in the system. In a powder compact, excess free energy is present primarily in the form of surface or interfacial energy (i.e. liquid–vapour and/or solid–vapour interfaces) associated with porosity. Under the influence of elevated temperature and/or pressure during sintering, atoms migrate to thermodynamically more stable positions within a powder compact. Material transport is driven by the chemical potential difference that exists between surfaces of dissimilar curvature within the system. Physically, in a particulate system, atoms or ions move from higher energy convex (i.e. as viewed from the particle centre out) particle surfaces to lower energy concave particle surfaces to decrease the curvature and chemical potential gradients in the system.

Material transport can occur by solid-state, liquid-phase, and/or vapour-phase mechanisms. For polycrystalline ceramics, material transport commonly occurs as ions diffuse through the volume, along grain boundaries (i.e. particle–particle intersections) and on particle surfaces (figure C2.11.6.). Additionally, ions can vaporize from, and subsequently recondense onto, particle surfaces (i.e. evaporation–condensation). A powder compact will densify (i.e. undergo volume contraction) when material transport occurs in a manner that allows particle centres to approach during sintering. Material transport by volume and grain boundary diffusion can result in densification. Material transport that changes the geometry of the system without densification is termed coarsening. Grain growth is perhaps the most prevalent form of coarsening during sintering. Coarsening can occur when material is transported by volume diffusion, surface diffusion, or evaporation–condensation.

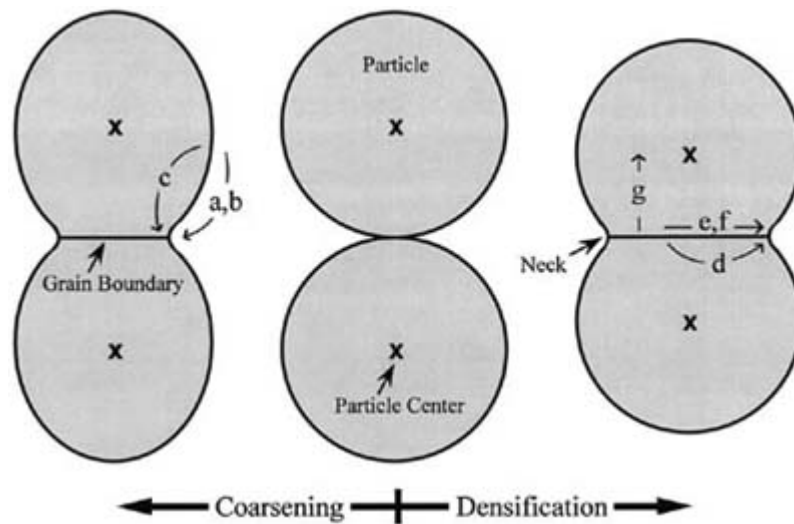


Figure C2.11.6. The classic two-particle sintering model illustrating material transport and neck growth at the particle contacts resulting in coarsening (left) and densification (right) during sintering. Surface diffusion (a), evaporation~condensation (b), and volume diffusion (c) contribute to coarsening, while volume diffusion (d), grain boundary diffusion (e), solution~precipitation (f), and dislocation motion (g) contribute to densification.

-11-

(b) *Densification and microstructure development.* Microstructurally, material transport during sintering manifests itself as interparticle pore shrinkage, grain boundary formation, a decrease in the total volume of the system (i.e. densification), and an increase in the average size of the particles that make up the compact (i.e. grain growth) [61, 62, 63 and 64]. Interparticle contacts flatten, the curvature within the system decreases, and the surface area and free energy of the system decrease during sintering

The ideal sintering process can be divided into three basic stages [74]. Initially, material is transported from convex particle surfaces to the pore~grain boundary intersection to form necks between adjacent particles. As this occurs, grain boundaries grow to create a three-dimensional array of approximately cylindrical, interconnected (i.e. continuous) pore channels at three grain junctions throughout the compact. These pore channels shrink in diameter during intermediate-stage sintering. Ultimately, because of Rayleigh instability (i.e. the critical ‘cylinder’ length to diameter ratio), the channels pinch off to form approximately spherical, isolated (i.e. closed) pores at four grain junctions within the ceramic matrix. The radial shrinkage of closed pores and the growth of larger grains at the expense of smaller ones constitute final-stage sintering.

Sintering phenomena are generally similar in real powder compacts; however, factors including surface energy anisotropy and packing heterogeneities in real systems can contribute to heterogeneous (i.e. non-uniform) densification and microstructure development. To circumvent this problem, minor concentrations of select chemicals, referred to as sintering aids or dopants, are commonly added prior to sintering. These chemical impurities preferentially segregate to high-energy crystallographic planes to decrease the crystalline anisotropy in the system to provide improved control over microstructure development during sintering. MgO-doped Al₂O₃ is the classic example in ceramics [71]. Impurity segregation to high-energy grain boundaries will also produce lower-energy interfaces that reduce the overall driving force for material transport during sintering.

(c) *Solid-state sintering.* Ceramics can be densified by solid-state [71, 72, 73, 75], liquid-phase [76], and viscous [77] sintering. Solid-state sintering refers to the process whereby densification occurs by solid-state, diffusion-controlled material transport. Densification occurs as higher-energy, solid~vapour (i.e. pore) interfaces are replaced by lower energy, solid~solid (i.e. grain boundary) interfaces. The change in free energy associated with the elimination of porosity, which drives densification, can be approximated by

$$dG = \gamma_{ss} dA_{ss} - \gamma_{sv} dA_{sv}. \quad (C2.11.5)$$

After the pore surfaces are eliminated and densification is complete, grain growth can further reduce the free energy of the system by reducing the amount of high-energy, solid–solid interfacial area. The change in free energy associated with the elimination of particle–particle interfaces, which drives grain growth, can be approximated by

$$dG = -\gamma_{ss} dA_{ss}. \quad (\text{C2.11.6})$$

Because densification occurs via the shrinkage of thermodynamically unstable pores, densification and microstructure development can be assessed on the basis of the dihedral angle, θ , formed as a result of the surface energy balance between the two solid–vapour and one solid–solid interface at the pore–grain boundary intersection [61, 78, 79 and 80],

$$\theta = 2 \cos^{-1} \left(\frac{\gamma_{ss}}{2\gamma_{sv}} \right) \quad (\text{C2.11.7})$$

-12-

where γ_{ss} and γ_{sv} are the solid–solid and solid–vapour interfacial energies, respectively (figure C2.11.7). Pore shrinkage and densification are favoured by a dihedral angle that is greater than the geometric dihedral angle, Ψ , of the regular polyhedron whose number of sides equals the coordination number of the pore. Theoretically, for the ideal four-sided pore present during final-stage sintering, θ must be greater than 70.4° (i.e. the geometric dihedral angle for a tetrahedron) to achieve 100% theoretical density during sintering. A larger dihedral angle will be required to eliminate larger pores (i.e. surrounded by more grains) formed as a result of packing defects. The larger the dihedral angle, the larger the intergranular pores that can be eliminated during sintering and the greater the surface tension driving force for pore shrinkage. Thermodynamics and/or kinetics limit the shrinkage of pores trapped within grains (i.e. intragranular porosity) and pores above a critical size [78, 79 and 80].

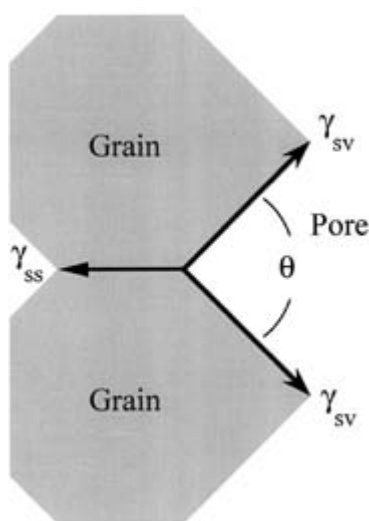


Figure C2.11.7. An illustration of the equilibrium dihedral angle, θ , formed by the balance of interfacial energies at a pore–grain boundary intersection during solid-state sintering.

(d) *Liquid-phase sintering.* To promote faster densification at lower temperatures, relatively small concentrations of chemical additives, referred to as sintering aids, are commonly used to create a liquid phase during sintering. Traditional liquid-phase sintering involves heating and melting crystalline solids to form a eutectic liquid during sintering [63, 76]. The requirements for liquid-phase sintering are that the liquid wets the solid particles, there is sufficient liquid present, and that the solid is soluble in the liquid. The concentration of the liquid and the solubility of the solid in the liquid (i.e. reactivity) increase dramatically with increasing temperature above the eutectic temperature.

Liquid-phase sintering is significantly more complex than solid-state sintering in that there are more phases, interfaces, and material transport mechanisms to consider. In general, densification will occur as long as it is

energetically favourable to replace liquid–vapour (subscript lv), solid–solid (subscript ss), and solid–vapour (subscript sv) interfaces with solid–liquid (subscript sl) interfaces during sintering:

$$dG = \gamma_{sl} dA_{sl} - (\gamma_{lv} dA_{lv} + \gamma_{ss} dA_{ss} + \gamma_{sv} dA_{sv}). \quad (C2.11.8)$$

Densification during liquid-phase sintering occurs in three stages. Initially, liquid forms at particle intersections and redistributes throughout the particulate mass under the influence of the capillary action. Shear stresses due to the

-13-

capillary pressure imbalance on different particles (e.g. different size particles) result in particle rearrangement to improve packing, and contribute to initial-stage densification. Solution–precipitation controls densification during intermediate-stage sintering. Material dissolves from higher energy, convex particle surfaces and migrates to lower energy, pore surfaces where it precipitates. This process is sometimes referred to as grain accommodation, because individual grains will actually change shape to fill void space. Densification by solution–precipitation continues until a rigid, three-dimensional skeletal structure is formed. The transition to final-stage liquid-phase sintering occurs when closed pores are formed at a compact relative density of ~90%. Final-stage liquid-phase sintering, as with solid-state sintering, is characterized by the shrinkage of isolated pores and by grain growth.

In liquid-phase sintering, densification and microstructure development can be assessed on the basis of the liquid contact or wetting angle, ϕ , formed as a result of the interfacial energy balance at the solid–liquid–vapour intersection as defined by the Young equation:

$$\phi = \cos^{-1} \left(\frac{\gamma_{sv} - \gamma_{sl}}{\gamma_{lv}} \right) \quad (C2.11.9)$$

where γ_{sl} and γ_{lv} are the solid–liquid and liquid–vapour interfacial energies, respectively (figure C2.11.8). A low contact angle favours liquid wetting of particle surfaces and densification during liquid-phase sintering. Theoretically, ϕ must be less than 60° to achieve 100% of the theoretical density.

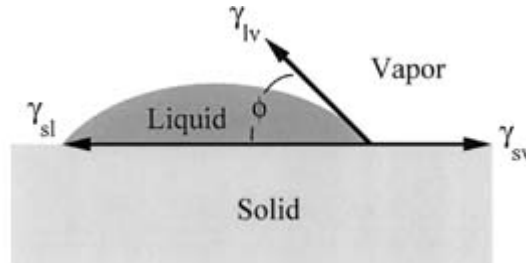


Figure C2.11.8. An illustration of the equilibrium contact (i.e. wetting) angle, ϕ , formed by the balance of interfacial energies for a liquid (sessile) drop on a flat solid surface.

(e) *Pressure sintering.* Pressure sintering employs the simultaneous use of both pressure and temperature during sintering to effect densification. Externally applied pressure on a powder compact increases the compressive stress at particle contacts, increasing the chemical potential gradient and the driving force for material transport relative to conventional sintering [63, 79, 80]. During conventional solid-state sintering, the driving force (DF) for final-stage pore closure at any given time, t , is determined by the surface energy, γ_{sv} of the sintering material and the radius, r , of the pore:

$$DF(\gamma)_t = 2\gamma_{sv}/r_t. \quad (C2.11.10)$$

During pressure sintering, interparticle compressive stress, approximated by the externally applied stress σ_a and normalized by the relative density of the compact ρ , supplements the surface tension driving force for pore shrinkage:

$$DF_t = DF(\gamma)_t + \sigma_a/\rho_t. \quad (C2.11.11)$$

-14-

As such, under an externally applied pressure a powder compact can densify faster and/or at a lower temperature during sintering. Pressure sintering is generally used to densify materials that are difficult or impossible to densify conventionally, and to produce dense fine-grain-size ceramics without the use of sintering aids. Additionally, the increased driving force for material transport and densification makes it possible to eliminate larger pores during pressure sintering, which can contribute to improved performance and reliability.

(f) *Sintering atmosphere.* The sintering atmosphere plays an important role in determining how a material densifies, the ultimate density achieved, and the end-use properties of the finished ceramic [63]. In particular, in complex electronic ceramics like lead zirconate titanate, where the phases present during and after sintering are strongly dependent on oxygen stoichiometry, the sintering atmosphere can determine if the system densifies by solid-state or liquid-phase sintering, and what the resultant electrical properties are [61]. In addition, if gas from the sintering atmosphere becomes trapped in closed pores during final-stage sintering and cannot readily diffuse through the system, it will impede and ultimately limit densification [81]. The pressure, P , of the gas trapped within a closed pore will counteract the surface tension driving force to shrink the pore during sintering:

$$DF_t = DF(\gamma)_t - P_t. \quad (C2.11.12)$$

Trapped gas in closed pores often limits densification when sintering with a liquid or viscous (glass) phase because rapid material transport through the liquid often results in pore closure early in the sintering process.

C2.11.5 SUMMARY

The manufacture of ceramics starts with the constituent raw materials and carries through to thermal consolidation. Intermediate processing steps include raw material beneficiation, shape forming, and pre-sinter thermal processing. Surfaces are created, modified, and eliminated during ceramic powder processing. Optimizing ceramic manufacturing requires understanding and controlling the physical chemistry of surfaces and interfaces during the various unit process steps. The control and utilization of surface energy and surface curvature are critical. Surface tension creates pressure gradients that contribute to the agglomeration and rearrangement of particles in powders, to the migration of liquids during mixing, shape forming, and drying, and to pore shrinkage during sintering. Chemical potential gradients associated with surface curvature determine the solubility of any particles in liquids, control the rate of evaporation from solid surfaces, and drive material transport during sintering. In combination with a strong engineering component, robust ceramic processing requires understanding and controlling the physical chemistry of surfaces.

ACKNOWLEDGMENTS

The author thanks Dr James Voigt and Dr Donald Ellerby of Sandia National Laboratories for their technical review of this article, and Dr Ellerby for providing the SEM micrographs herein.

-15-

REFERENCES

- [1] Adamson A W 1976 *Physical Chemistry of Surfaces* 3rd edn (New York: Wiley)

- [2] Reed J S 1995 *Introduction to the Principals of Ceramic Processing* 2nd edn (New York: Wiley) pp 18–27
- [3] Reed J S 1995 *Introduction to the Principals of Ceramic Processing* 2nd edn (New York: Wiley)
- [4] Richerson D W 1992 *Modern Ceramic Engineering: Properties, Processing, and Use in Design* 2nd edn (New York: Dekker)
- [5] Ewsuk K G 1993 Ceramics (processing) *Kirk–Othmer Encyclopedia of Chemical Technology* 4th edn, vol 5 (New York: Wiley) pp 603–33
- [6] Reed J S 1995 *Introduction to the Principals of Ceramic Processing* 2nd edn (New York: Wiley) pp 35–53
- [7] Richerson D W 1992 *Modern Ceramic Engineering: Properties, Processing, and Use in Design* 2nd edn (New York: Dekker) pp 373–81
- [8] Norton F H 1974 *Elements of Ceramics* 2nd edn (Reading, MA: Addison-Wesley) pp 24–71
- [9] Kingery W D 1960 *Introduction to Ceramics* (New York: Wiley) pp 15–31
- [10] Jones J T and Berard M F 1972 *Ceramics: Industrial Processing and Testing* (Ames, IA: The Iowa State University Press) pp 14–38
- [11] Ewsuk K G 1993 Ceramics (processing) *Kirk–Othmer Encyclopedia of Chemical Technology* 4th edn, vol 5 (New York: Wiley) pp 603–7
- [12] Brownell W E 1976 *Structural Clay Products, Applied Mineralogy* vol 9 (New York: Springer) pp 43–60
- [13] Reed J S 1995 *Introduction to the Principals of Ceramic Processing* 2nd edn (New York: Wiley) pp 54–66
- [14] Johnson D W Jr 1981 Nonconventional powder preparation techniques *Am. Ceram. Soc. Bull.* **60** 221–4, 243
- [15] Johnson D W Jr 1987 Innovations in ceramic powder preparation *Ceramic Powder Science, Advances in Ceramics* vol 21, ed G L Messing *et al* (Westerville, OH: The American Ceramic Society) pp 3–19
- [16] Rhodes W H and Natansohn S 1989 Powders for advanced structural ceramics *Am. Ceram. Soc. Bull.* **68** 1804–12
- [17] Anderson H, Kudas T T and Smith D M 1989 Vapor phase processing of powders; plasma synthesis and aerosol decomposition *Am. Ceram. Soc. Bull.* **68** 996–1000
- [18] Ganguli D and Chatterjee M 1997 *Ceramic Powder Preparation Handbook* (Norwell, MA: Kluwer)
- [19] Voigt J A 1993 Powder and precursor preparation by solution techniques *Characterization of Ceramics* ed R E Loehman (Greenwich, CT: Butterworth-Heinemann) pp 1–27
- [20] McColm I J and Clark N J 1988 *Forming, Shaping and Working of High Performance Ceramics* (New York: Chapman and Hall) pp 60–140
- [21] Reed J S 1995 *Introduction to the Principals of Ceramic Processing* 2nd edn (New York: Wiley) pp 313–33

- [22] Richerson D W 1992 *Modern Ceramic Engineering: Properties, Processing, and Use in Design* 2nd edn (New York: Dekker) pp 381–96
- [23] Norton F H 1974 *Elements of Ceramics* 2nd edn (Reading, MA: Addison-Wesley) pp 55–71
- [24] Jones J T and Berard M F 1972 *Ceramics: Industrial Processing and Testing* (Ames, IA: The Iowa State University Press) pp 20–38
- [25] Hogg R 1981 Grinding and mixing of nonmetallic powders *Am. Ceram. Soc. Bull.* **60** 206–11, 220
- [26] Greskovich C 1976 Milling *Ceramic Fabrication Processes, Treatise on Materials Science and Technology* vol 9, ed F F Y Wang (New York: Academic) pp 15–33

- [27] Somasundaran P 1978 Theories of grinding *Ceramic Processing Before Firing* ed G Y Onoda Jr and L Hench (New York: Wiley) pp 105–23
- [28] Ewsuk K G 1993 Ceramics (processing) *Kirk–Othmer Encyclopedia of Chemical Technology* 4th edn, vol 5 (New York: Wiley) pp 607–10
- [29] Reed J S 1995 *Introduction to the Principals of Ceramic Processing* 2nd edn, (New York: Wiley) pp 378–90
- [30] Richerson D W 1992 *Modern Ceramic Engineering: Properties, Processing, and Use in Design* 2nd edn (New York: Dekker) pp 411–13
- [31] Ewsuk K G 1993 Ceramics (processing) *Kirk–Othmer Encyclopedia of Chemical Technology* 4th edn, vol 5 (New York: Wiley) p 610
- [32] Reed J S 1995 *Introduction to the Principals of Ceramic Processing* 2nd edn (New York: Wiley) pp 400–4
- [33] Reed J S 1995 *Introduction to the Principals of Ceramic Processing* 2nd edn (New York: Wiley) pp 135–208
- [34] Richerson D W 1992 *Modern Ceramic Engineering: Properties, Processing, and Use in Design* 2nd edn (New York: Dekker) pp 421–8
- [35] Morse T 1979 *Handbook of Organic Additives for Use in Ceramic Body Formulation* (Butte, MA: Montana Energy and MHD Research and Development Institute)
- [36] Ewsuk K G 1993 Ceramics (processing) *Kirk–Othmer Encyclopedia of Chemical Technology* 4th edn, vol 5 (New York: Wiley) pp 610–12
- [37] Shanefield D J 1996 *Organic Additives and Ceramic Processing: with Applications in Powder Metallurgy, Ink, and Paint* 2nd edn (Boston, MA: Kluwer)
- [38] Allen T 1981 *Particle Size Measurement* 3rd edn (New York: Chapman and Hall) pp 246–66
- [39] Nelson R D 1988 Dispersing powders in liquids *Handbook of Powder Technology* vol 7, ed J C Williams and T Allen (New York: Elsevier)
- [40] Brindley G W 1960 Ion exchange in clay minerals *Ceramic Fabrication Processes* ed W D Kingery (New York: Wiley) pp 11–23
- [41] Michaels A S 1960 Rheological properties of aqueous clay systems *Ceramic Fabrication Processes* ed W D Kingery (New York: Wiley) pp 23–31

- [42] Onoda G Y Jr 1978 The rheology of organic binder solutions *Ceramic Processing Before Firing* ed G Y Onoda Jr and L Hench (New York: Wiley) pp 235–51
- [43] Reed J S 1995 *Introduction to the Principals of Ceramic Processing* 2nd edn (New York: Wiley) pp 492–533
- [44] Richerson D W 1992 *Modern Ceramic Engineering: Properties, Processing, and Use in Design* 2nd edn (New York: Dekker) pp 444–78
- [45] Cowan R E 1976 Slip casting *Ceramic Fabrication Processes, Treatise on Materials Science and Technology* vol 9, ed F F Y Wang (New York: Academic) pp 153–71
- [46] Magid H S 1960 Controls required and problems encountered in production slip casting *Ceramic Fabrication Processes* ed W D Kingery (New York: Wiley) pp 40–5
- [47] St Pierre P D S 1960 Slip casting nonclay ceramics *Ceramic Fabrication Processes* ed W D Kingery (New York: Wiley) ch 5, pp 45–51
- [48] Ewsuk K G 1993 Ceramics (processing) *Kirk–Othmer Encyclopedia of Chemical Technology* 4th edn, vol 5 (New York: Wiley) pp 615–17

- [49] Reed J S 1995 *Introduction to the Principals of Ceramic Processing* 2nd edn (New York: Wiley) pp 545–58
- [50] Norton F H 1974 *Elements of Ceramics* 2nd edn (Reading, MA: Addison-Wesley) pp 114–25
- [51] Jones J T and Berard M F 1972 *Ceramics: Industrial Processing and Testing* (Ames, IA: The Iowa State University Press) pp 69–89
- [52] Brownell W E 1976 *Structural Clay Products, Applied Mineralogy* vol 9 (New York: Springer) pp 101–25
- [53] Brinker C J and Scherer G W 1990 Sol–gel science *The Physics and Chemistry of Sol–Gel Processing* (New York: Academic) pp 453–513
- [54] Ewsuk K G 1993 Ceramics (processing) *Kirk–Othmer Encyclopedia of Chemical Technology* 4th edn, vol 5 (New York: Wiley) pp 619–20
- [55] Reed J S and Runk R B 1976 Dry pressing *Ceramic Fabrication Processes, Treatise on Materials Science and Technology* vol 9, ed F F Y Wang (New York: Academic) pp 71–93
- [56] Thurnauer H 1960 Controls required and problems encountered in production dry pressing *Ceramic Fabrication Processes* ed W D Kingery (New York: Wiley) pp 62–70
- [57] Reed J S 1995 *Introduction to the Principals of Ceramic Processing* 2nd edn (New York: Wiley) pp 418–45
- [58] Richerson D W 1992 *Modern Ceramic Engineering: Properties, Processing, and Use in Design* (New York: Dekker) pp 429–43
- [59] Ewsuk K G 1993 Ceramics (processing) *Kirk–Othmer Encyclopedia of Chemical Technology* 4th edn, vol 5 (New York: Wiley) pp 613–14
- [60] Norton F H 1974 Firing ceramic ware *Elements of Ceramics* 2nd edn (Reading, MA: Addison-Wesley) pp 126–53

-18-

- [61] Ewsuk K G 1993 Consolidation of bulk ceramics *Characterization of Ceramics* ed R E Loehman (Greenwich, CT: Butterworth–Heinemann) pp 77–101
- [62] Reed J S 1995 *Introduction to the Principals of Ceramic Processing* 2nd edn (New York: Wiley) pp 583–619
- [63] Ewsuk K G 1993 Ceramics (processing) *Kirk–Othmer Encyclopedia of Chemical Technology* 4th edn, vol 5 (New York: Wiley) pp 620–7
- [64] Richerson D W 1992 *Modern Ceramic Engineering: Properties, Processing, and Use in Design* (New York: Dekker) pp 519–64
- [65] Kingery W D 1960 *Introduction to Ceramics* (New York: Wiley) pp 15–31
- [66] Kingery W D, Bowen H K and Uhlmann D R 1967 *Introduction to Ceramics* (New York: Wiley) pp 448–515
- [67] Herbert J M 1985 Ceramic dielectrics and capacitors *Electrocomponent Science Monographs* vol 6 (New York: Gordon and Breach) pp 63–94
- [68] Jones J T and Berard M F 1972 *Ceramics: Industrial Processing and Testing* (Ames, IA: The Iowa State University Press) pp 69–89
- [69] Brownell W E 1976 Structural clay products *Applied Mineralogy* vol 9 (New York: Springer) pp 126–64
- [70] McColm I J and Clark N J 1988 *Forming, Shaping and Working of High Performance Ceramics* (New York: Chapman and Hall) pp 208–310
- [71] Coble R L and Burke J E 1963 Sintering in ceramics *Progress in Ceramic Science* vol 3, ed J E Burke (New York: MacMillan) pp 197–251

- [72] Thümmeler F and Thomma W 1967 The sintering process *J. Inst. Metals* **12** 69–108
- [73] Burke J E and Rosolowski J H 1976 Sintering *Reactivity of Solids (Treatise on Solid State Chemistry vol 4)* ed N B Hannay (New York: Plenum) pp 621–59
- [74] Herring C 1949 Surface tension as a motivation for sintering *The Physics of Powder Metallurgy* ed W E Kingston (New York: McGraw-Hill) pp 143–79
- [75] Coble R L 1961 Sintering crystalline solids. I, intermediate and final state diffusion models *J. Appl. Phys.* **32** 787–92
- [76] German R M 1985 *Liquid Phase Sintering* (New York: Plenum)
- [77] Brinker C J and Scherer G W 1990 Sol–gel science *The Physics and Chemistry of Sol–Gel Processing* (New York: Academic) pp 675–742
- [78] Kingery W D and Francois B 1965 The sintering of crystalline oxides, I. Interactions between grains boundaries and pores *Sintering and Related Phenomena* ed G C Kuczynski, N A Hooton and C F Gibbon (New York: Gordon and Breach) pp 471–98
- [79] Ewsuk K G 1986 Final stage densification of alumina during hot isostatic pressing *PhD Thesis* The Pennsylvania State University
- [80] Ewsuk K G and Messing G L 1986 A theoretical and experimental analysis of final-stage densification of alumina during hot isostatic pressing *Hot Isostatic Pressing: Theories and Applications* ed R J Schaefer and M Linzer (Materials Park, OH: ASM International) pp 23–33
-

-19-

- [81] Ewsuk K G 1992 Effects of trapped gases on ceramic-filled-glass composite densification *Solid State Phenomena* vol 25–26, ed A C D Chaklader and J A Lund (Brookfield, VT: Trans-Tech) pp 63-72 (Proc. Sintering' 91)
-

-1-

C2.12 Zeolites

Andreas Kogelbauer and Roel Prins

C2.12.1 INTRODUCTION AND HISTORY

Compared to other crystalline inorganic oxides, zeolites represent a special class of materials. Their crystalline, microporous nature with well-defined pore dimensions in combination with high thermal stability, ion exchange and sorption capacity, as well as the ability to generate acidity has made them unique materials for practical applications. In recent decades, zeolites have gained tremendous importance both from an academic and an economic point of view. On the one hand, it is the versatility of zeolites that makes them such outstanding materials. They are well suited for a broad range of applications such as use as drying agents, use in gas separation processes, use as detergent additives and as catalysts (see [section C2.12.7](#)). On the other hand, the variety of structure types, the broad range of chemical modifications of the zeolite matrix, and the derived physico-chemical properties all carry a distinct fascination for the scientist which is much reflected in the ever growing number of

zeolite-related publications and people involved in zeolite research [1].

The term *zeolite* was coined by the Swedish mineralogist A F Cronstedt who in 1756 observed that certain minerals exhibited intumescence upon heating. Since they seemed to boil he referred to them as ‘boiling stones’ (Greek: *zein* = to boil, *lithos* = stone) [2]. Mineralogical studies that followed in the nineteenth century comprised mainly the determination of morphological, physical, and chemical properties of zeolitic minerals. The first systematic studies regarding the reversible hydration behaviour of zeolites and their chemical composition date back to the mid-1800s [3] and were followed by investigations regarding their adsorption properties around the beginning of this century. To explain the selective adsorption of small molecules in chabazite, McBain coined the term *molecular sieve* [4]. With the availability of x-ray diffraction around the 1920s, structure determination of zeolites became possible. Analcime was the first zeolite structure to be determined in 1930 [5]. At the same time, the hydrated aluminosilicate framework with loosely bonded alkali and earth alkali cations was established as the common criterion to distinguish zeolites chemically from other materials. About 1938, Barrer started the systematic investigation of the properties of natural zeolites: in particular, he applied physico-chemical principles and thereby put the study of zeolites on a firm scientific base. His investigations regarding zeolite synthesis led to the first reproducible and substantiated synthesis of zeolites in a laboratory environment [6]. The industrial use of zeolites followed shortly afterwards. Researchers at Union Carbide Corporation discovered the commercially important zeolite types A, X, and Y which were commercialized in 1954. Most prominent among those was the use of synthetic zeolite X as cracking catalyst by Mobil Oil in 1962. The use of template molecules for zeolite synthesis during the 1960s led to a variety of new synthetic structures with interesting properties. Ongoing synthesis efforts have led to more than 100 different synthetic structures known today. Zeolite nomenclature is based on a three-letter code assigned to the different structure types by the International Zeolite Association (IZA) [7] and compiled in the *Atlas of Zeolite Structure Types* [8]. It is, however, still common practice to use traditional designations as for instance in the case of X and Y zeolites which belong to the FAU structure type. Many of these traditional designations originate from the laboratory in which the materials were synthesized (e.g. ZSM for Zeolite Socony Mobil). For those materials whose framework topology has been confirmed, cross-references exist in the *Atlas of Zeolite Structure Types* enabling the assignment of traditional designations to the IZA structure type.

-2-

C2.12.2 COMPOSITION AND STRUCTURE OF ZEOLITES

The traditional definition of a zeolite refers to microporous, crystalline, hydrated aluminosilicates with a three-dimensional framework consisting of corner-linked SiO_4 or AlO_4 tetrahedra, although today the definition is used in a much broader sense, comprising microporous crystalline solids containing a variety of elements as tetrahedral building units. The aluminosilicate-based zeolites are represented by the empirical formula



in which M represents a cation of valence n , and $x \geq 1$ since no Al–O–Al bonds are permitted in a zeolite according to Loewenstein’s rule [9]. The latter states that the ratio of silicon-to-aluminium must be equal to or greater than one due to local charge restrictions. The SiO_4 tetrahedra are charge balanced but each AlO_4 tetrahedron carries a formal charge of -1 due to the $+3$ charge of the aluminium atom (see figure C2.12.1). Cations M are therefore required for balancing the lattice charge. These cations are rather mobile if the zeolite is in a hydrated state and, therefore, they can be easily exchanged. Typically sodium, potassium or organic tetralkyl ammonium is present as a monovalent charge-compensating cation in synthetic zeolites. Besides cations from the alkaline and alkaline earth series, transition metal cations are also frequently found in zeolites from natural sources. One of the outstanding properties of zeolites derives from the exchange of the charge-balancing cations by protons which can be attained by treatment in dilute mineral acids or by the exchange of ammonium cations that are subsequently thermally decomposed to yield ammonia and surface bonded protons (see figure C2.12.2). These protons are acidic which makes zeolites solid Brønsted acids with an acid strength comparable to that of 70% sulphuric acid [10]. The concentration of these acid sites increases with the aluminium concentration in the zeolite lattice and is in the range of 10^{-3} – 10^{-4} mol per gram zeolite.

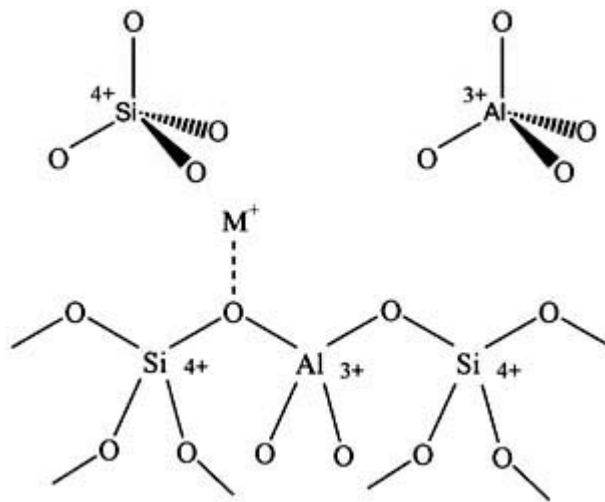


Figure C2.12.1. Origin of ion exchange capacity in zeolites. Since every oxygen atom contributes one negative charge to the tetrahedron incorporated in the framework, the silicon tetrahedron carries no net charge while the aluminium tetrahedron carries a net charge of -1 which is compensated by cations M.

-3-

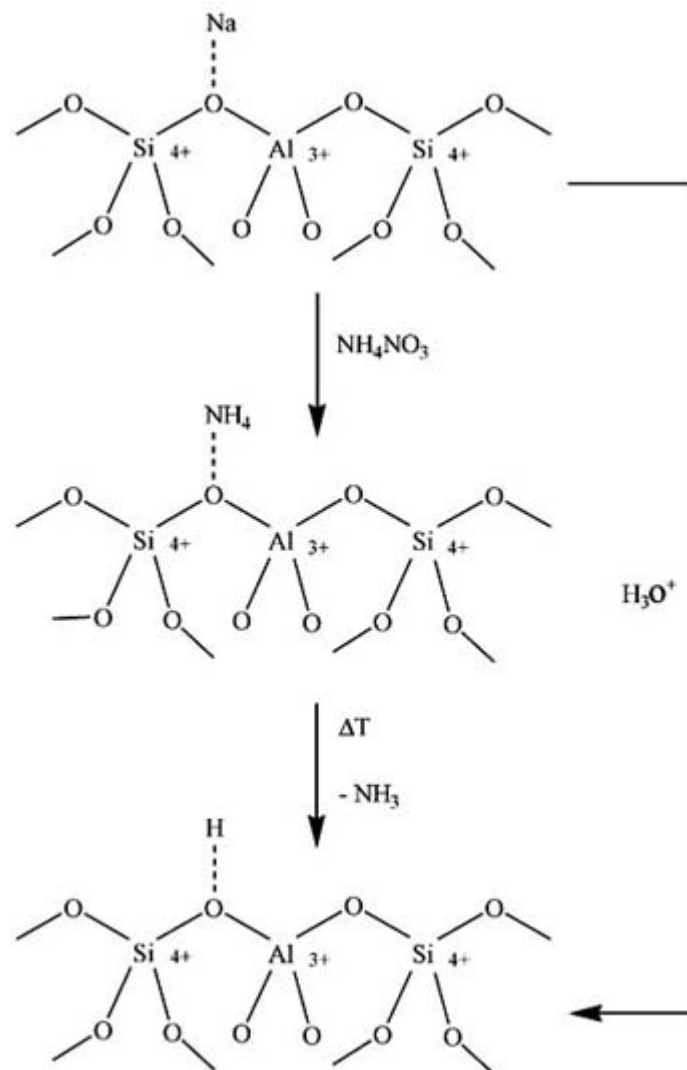


Figure C2.12.2. Formation of Brønsted acid sites in zeolites. Aqueous exchange of cation M with an ammonium salt yields the ammonium form of the zeolite. Upon thermal decomposition ammonia is released and the proton remains as charge-balancing species. Direct ion-exchange of M with acidic solutions is feasible for high-silica zeolites.

The connection of the primary building unit, the tetrahedron, through oxygen bridges leads to the secondary building units (SBU), some of which are illustrated in [figure C2.12.3](#) [11]. The way of depicting SBUs and whole zeolite structures, as in [figure C2.12.3](#), is common practice in zeolite science; only the central atoms of the tetrahedra (T-atoms) are drawn and lines represent oxygen bridges between tetrahedra. The pore openings that result from the arrangement of the primary building units are only referred to by the number of T-atoms; that is, a four-membered ring actually consists of four T-atoms and four bridging oxygen atoms in alternating arrangement. The prevalence of certain SBUs is used to classify zeolites but other ways of classifying framework topologies have also been developed [12](#).

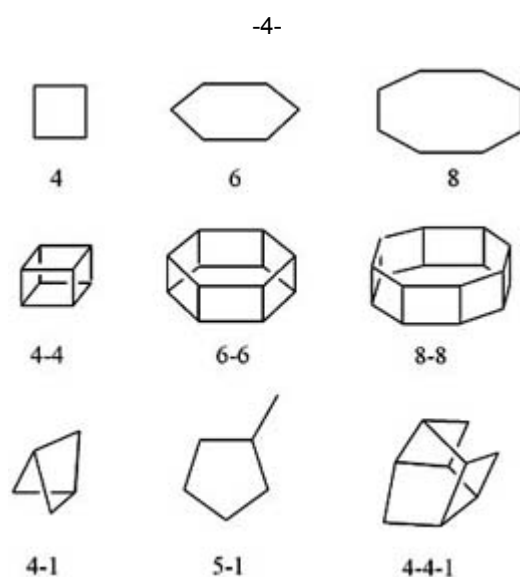


Figure C2.12.3. Secondary building units in zeolites. Each corner represents a T-atom (Si, Al) while the connecting lines represent oxygen bridges with the oxygen atom in the middle.

Many zeolite structures can be envisaged as being constructed from polyhedra that are obtained by appropriate arrangement of the SBUs. Some of the more common polyhedra are depicted in [figure C2.12.4](#). The formation of different zeolite structures from the same polyhedron, the sodalite cage, is demonstrated in [figure C2.12.5](#). By connection of two sodalite cages through one shared 4-ring, sodalite is obtained (IZA structure code SOD) whose largest pores are formed by 6-rings. The effective dimension of these pores is about 2.5 Å which is too small for any molecule of interest to penetrate into the zeolite micropores. Sodalite is therefore unimportant for technical applications. By connecting two sodalite cages with a double 4-ring prism zeolite A is obtained (IZA structure code LTA). The larger void that is formed by the specific arrangement of eight sodalite cages in zeolite A is called α -cage. It is accessible through 8-membered rings with a pore opening of 4.1 Å. The α -cage is therefore accessible to small molecules such as water which makes zeolite A an excellent drying agent. The pore diameter of zeolite A can further be varied between 3 and 5 Å by substituting different cations such as K, Na or Ca. These materials are commercially available as molecular sieve 3A, 4A and 5A. An even larger internal void space is obtained when sodalite cages are connected through double 6-ring prisms such as in the cubic faujasite (IZA structure code FAU). Zeolites X or Y are the synthetic equivalents and vary only in their Si/Al ratio. The pore opening of these zeolites is formed by 12-membered rings with diameters of 7.4 Å which are big enough even for larger organic molecules such as substituted aromatics. Reversing the stacking order results in hexagonal faujasite (IZA structure code EMT). This zeolite has distorted 12-membered rings as pore openings.

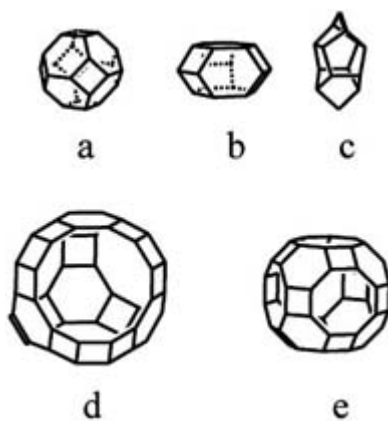


Figure C2.12.4. Typical polyhedra found in zeolites: (a) sodalite cage found in sodalite, zeolite A or faujasite; (b) cancrinite or ϵ -cage found in cancrinite, erionite, offretite or gmelinite; (c) the 5-ring polyhedron found in ZSM-5 and ZSM-11; (d) the large cavity of the faujasite structure; and (e) the α -cage forming the large cavity in zeolite A.

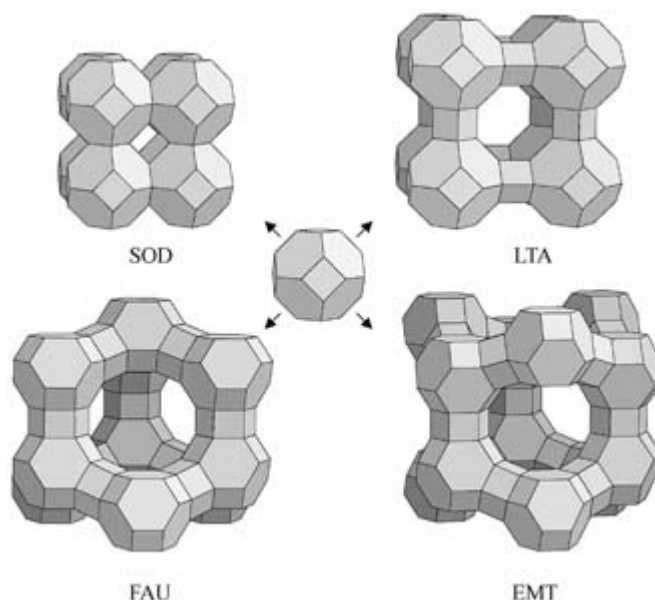


Figure C2.12.5. Different framework topologies based on the sodalite cage obtained through different connection patterns.

Another technically relevant class of zeolites is the pentasil group [13]. Their dominating structural building units are 5-membered rings and they contain less aluminium than the sodalite cage-based zeolites. The polyhedron shown in figure C2.12.4 can be arranged to form chains, as shown in figure C2.12.6, which build the basis for the ZSM-5 (IZA structure code MFI) and ZSM-11 (IZA structure code MEL) topologies. The pores of these zeolites, which are classified as medium pore zeolites, are formed by 10-membered rings. Small pore zeolites are in analogy those having 8-membered ring pores and large pore zeolites those with 12-membered rings and above. The micropores in ZSM-5 form a two-dimensional channel system in which straight channels are intersected by sinusoidal channels as depicted

in figure C2.12.7. In ZSM-11, perpendicular straight channels intersect each other. Transport of molecules, however, is possible in all three main crystallographic directions by moving from one channel system to the other. There are also zeolite structures with only mono- or two-dimensional channel systems such as L, ZSM-12, ferrierite and mordenite. Some zeolite structures exhibit small pore openings but larger internal voids (supercages) such as erionite or faujasite. A compilation of the structural and chemical characteristics of technically important

zeolite types is given in table C2.12.1.

Table C2.12.1. Characteristics of technically important zeolites.

Zeolite	IZA structure code	Typical unit cell composition	SiO ₂ /Al ₂ O ₃ range by synthesis	Dimensionality of channel system	Pore apertures (nm)
A	LTA	Na ₁₂ [(AlO ₂) ₁₂ (SiO ₂) ₁₂]27 H ₂ O	2.0–6.8	3	0.41
L	LTL	K ₉ [(AlO ₂) ₉ (SiO ₂) ₂₇]22 H ₂ O	6.0–10.0	1	0.71
X	FAU	Na ₈₆ [(AlO ₂) ₈₆ (SiO ₂) ₁₀₆]264 H ₂ O	2.0–3.0	3	0.74
Y	FAU	Na ₅₆ [(AlO ₂) ₅₆ (SiO ₂) ₁₃₆]250 H ₂ O	3.0–9.0	3	0.74
Mordenite	MOR	Na ₈ [(AlO ₂) ₈ (SiO ₂) ₄₀]24 H ₂ O	9.0–32	2 ^a	0.65×0.70 0.26×0.57
ZSM-5	MFI	(Na,TPA) ₃ [(AlO ₂) ₃ (SiO ₂) ₉₃]16 H ₂ O	30–∞	3	0.53×0.56 0.51×0.55;
Beta	BEA	(Na,TEA) ₅ [(AlO ₂) ₅ (SiO ₂) ₅₉]	20–∞	3	0.76×0.64 0.55×0.55

^a Free apertures in second channel system are too small for organic molecules to diffuse readily, making the channel system of mordenite essentially monodimensional.

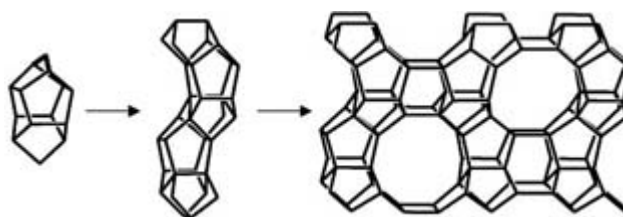


Figure C2.12.6. Framework topology of ZSM-5. The 5-ring polyhedron is connected into chains which form the ZSM-5 structure with the 10-membered openings of the linear channels.

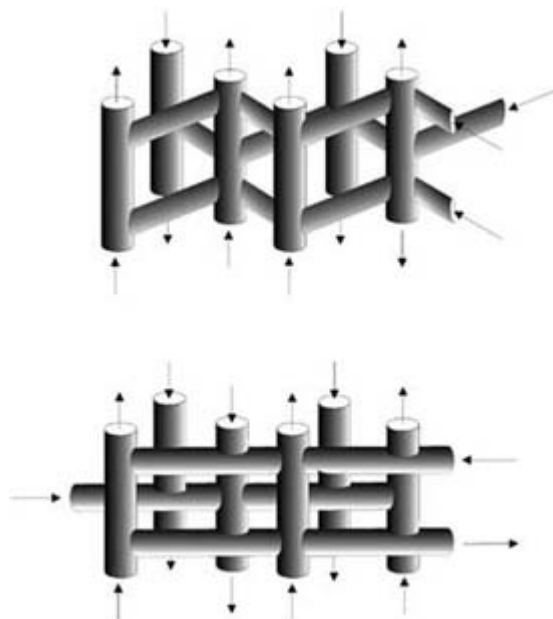


Figure C2.12.7. Channel system of MFI (top) and MEL (bottom). The linear channels are interconnected by zigzag channels in ZSM-5 while exclusively straight running channels are present in ZSM-11 – larger internal openings are present at the channel intersections – the arrows indicate the pathways for molecular transport through the channel system.

Progress in zeolite synthesis throughout the last decade has led to ultra-large pore zeolites, that is, zeolitic materials with pore sizes larger than those of 12-membered rings. These are very attractive and sought after materials because they can admit very bulky organic molecules typical of fine chemicals into the interior of the zeolite matrix where adsorption and catalysis can be carried out [14]. Very few structures have been synthesized so far which were initially based on aluminophosphates (AlPO_4 -8, 14-membered ring, $7.9 \times 8.7 \text{ \AA}$ pore dimension; VPI-5, 18-membered ring, 12.1 \AA pore diameter, see also below), or gallophosphates (cloverite, 20-membered ring with clover shape). Only recently true silicon-based zeolites with 14-membered rings have been synthesized (UTD-1 [15], CIT-5 [16]) that exhibit high thermal stability, especially when compared with the aluminium or gallium phosphates, and have comparable Brønsted acidity to other zeolites. Their synthesis, however, is based on the use of expensive or commercially unavailable template molecules.

-8-

The search for zeolitic materials with pore sizes larger than 10 \AA recently resulted in a new class of materials termed mesoporous molecular sieves (known by the designation M41S), which are prepared using surfactant micelles as templating agents [17]. The pore diameter of these materials is tunable in a wide range of about 30 to 100 \AA but the matrix consists of amorphous rather than crystalline silica walls. Incorporation of hetero-atoms (Al, Ti, B, Ni, Cr, Fe, Co, Mn) has been described, as was the synthesis of nonsiliceous materials such as oxides of W, Fe, Pb, Mo, and Sb [18]. Although these materials do not represent true zeolites, they are highly interesting materials which are commonly covered in the zeolite literature with great potential for shape-selective catalysis of bulky molecules.

Besides structural variety, chemical diversity has also increased. Pure silicon forms of zeolite ZSM-5 and ZSM-11, designated silicalite-1 [19] and silicalite-2 [20], have been synthesised. A number of other pure silicon analogues of zeolites, called porosils, are known [21]. Various chemical elements other than silicon or aluminium have been incorporated into zeolite lattice structures [22, 23]. Most important among those from an applications point of view are the incorporation of titanium, cobalt, and iron for oxidation catalysts, boron for acid strength variation, and gallium for dehydrogenation/aromatization reactions. In some cases it remains questionable, however, whether incorporation into the zeolite lattice structure has really occurred.

Compositional variety can also be achieved by ion exchange [24]. The cations are then located at the ion-exchange

positions rather than being incorporated in the zeolite lattice as oxygen tetrahedra. Ion exchange methods are used for the preparation of a number of commercially important zeolitic materials. Most important is the exchange for ammonium as discussed above because it represents the least damaging route for the preparation of the proton form of zeolites used in acid catalysis. Cs-exchange leads to zeolites that act as solid base catalysts. Using reducible metal salts for the ion exchange (e.g. $\text{Pt}(\text{NH}_3)_4^{2+}$) with subsequent reduction in hydrogen, metal-loaded zeolites with a high dispersion can be prepared which find applications in refinery processes as bifunctional catalysts (acidic and reducing functionality). Alternatively, the metal can be introduced using uncharged carbonyl complexes such as $\text{Ni}(\text{CO})_4$.

Additional to the aluminosilicate-based zeolites, a number of other crystalline microporous three-dimensional oxides have been synthesized [25]. Most prominent among these are the aluminophosphates (ALPO₄ series) [26, 27] whose framework is composed of strictly alternating $(\text{AlO}_4)^-$ and $(\text{PO}_4)^+$ tetrahedra. Since the pure ALPO₄ framework does not require charge-balancing cations, further compositional modifications are required to make use of these materials as catalysts. This is achieved by the partial or complete isomorphic substitution of framework phosphorus or aluminium by other elements during the synthesis. A variety of aluminophosphate derivatives have been synthesized in this way such as the SAPO (silicoaluminophosphate), MeAPO (metal aluminophosphate, Me=Mg, V, Cr, Mn, Fe, Co, Ni), ZnPO (zincphosphate), BePO (berylliumphosphate), GaPO₄ (galliumphosphate), and MeAPSO (metal silicoaluminophosphate) families. Although most elements substitute either for Al or P, in the case of silicon substitution of P and Al is possible. In SAPO-5 materials (IZA structure code AFI) silicon rich domains are simultaneously present besides SiAlP domains, the former being generated by substitution of Al-P pairs for silicon, the latter, which essentially carries all the Brønsted acidity, arising from substitution of phosphorus by silicon.

C2.12.3 SYNTHESIS OF ZEOLITES

Zeolites are the product of a hydrothermal conversion process [28]. As such they can be found in sedimentary deposits especially in areas that show signs of former volcanic activity. There are about 40 naturally occurring zeolite types. Types such as chabazite, clinoptilolite, mordenite and phillipsite occur with up to 80% phase purity in quite large

sedimentary deposits all over the world which makes mining economical. Due to the lack of purity and consistency in composition, the application of natural zeolites is rather limited and mainly adsorbent and ion exchange applications have been realized. Natural zeolites are, for instance, used as soil amendment, as cement additives or for the purification of municipal and nuclear wastewater [29].

For more demanding applications such as catalysis or the meaningful characterization of zeolitic phases by physico-chemical methods, only synthetic zeolites provide the required phase purity and compositional consistency. For their synthesis, hydrothermal conditions are commonly applied similar to those occurring during the formation of zeolitic phases in nature [23, 30]. The crystallization occurs from a gel formed from an aqueous silicate and aluminate solution at temperatures between 60 and 100 °C and atmospheric pressure, with higher temperatures and elevated pressures being occasionally required (e.g. for the synthesis of mordenite). Syntheses from clear solutions have been described but they are mainly of academic interest due to the low zeolite yields. More recently, new strategies toward zeolite synthesis have been developed which aim mainly at the formation of zeolite films and zeolitic membranes [31]. They are based on solid-state transformation, vapour-phase synthesis, secondary growth and casting of nanoparticles [32]. Synthesis routes that lead to binder-free materials are interesting from an applications point of view [33].

Since zeolites are metastable crystallization products they are subject to Ostwald's rule which states that metastable phases are initially formed and gradually transform into the thermodynamically most stable product. The least stable zeolitic phase (that with the lowest framework density) is therefore formed first and consumed with further synthesis time at the expense of a more stable phase due to a continuous crystallization/redissolution equilibrium.

The synthesis time is therefore of great importance. The primary condition for the crystallization of a zeolitic structure is a certain degree of supersaturation in the synthesis mixture leading to nucleation. If the degree of supersaturation is too high, rapid polycondensation occurs that does not permit the formation of highly organized crystals. The gel that is formed during this initial process, however, is sufficiently soluble to provide the right degree of supersaturation and zeolites can nucleate if other requirements such as the correct Si/Al ratio of the synthesis mixture or the presence of a specific template are fulfilled.

On a laboratory scale, hydrothermal synthesis is usually carried out in Teflon-coated, stainless-steel autoclaves under autogenous pressure. A typical synthesis mixture consists of up to four major constituents, a T-atom source (silicon and aluminium, other elements may also be incorporated as indicated above), a solvent (almost exclusively water), a mineralizer (OH^- , F^-), and a template. The T elements are usually provided as amorphous hydroxides, oxides or aluminosilicates, but alkoxy silicates may be used when high reactivity is required. Various solids such as precipitated gels, fumed silicas or clays may also be used as T-atom sources and the choice depends mainly on the desired reactivity. The mineralizer can be present in the T-atom source itself such as in aqueous silicate solutions. The primary function of the mineralizer is the dissolution of the T-atom source and the formation of the gel but it also assists in the quick equilibration of monomeric and polycondensated silicate species in the solution. Optimum concentration ranges exist for the mineralizing agent in dependence of the desired zeolite structure (e.g. aluminium rich zeolites crystallize preferentially at higher pH). Distinct differences in the resulting zeolites also result from the different pH ranges in which OH^- (strongly alkaline) and F^- (alkaline to slightly acidic) operate, the lower concentration of crystal defects in fluoride synthesis being one. The template generally assists in the formation of the solid by forming additional bonds with the zeolite (ionic, dipole, hydrogen bond or van der Waals interaction) and additionally exerts a structure directing effect. The most commonly encountered templates are either the hydrated alkaline or alkaline earth cations present in solution (e.g. Na^+ in the synthesis of zeolite X), or quaternary organic ammonium cations such as tetrapropyl ammonium (TPA) for ZSM-5 synthesis. Substituting TPA with tetrabutyl or tetraethyl ammonium reveals the structure

-10-

directing effect since ZSM-11 and zeolite beta are obtained instead of ZSM-5. Ongoing research has meanwhile identified a vast array of suitable templates and synthesis routes free of organic templates have been developed for zeolites that were formerly only attainable through template synthesis such as ZSM-5. As well as the factors discussed above, other variables such as the concentration and ratio of the constituents of the synthesis mixture, the preparation of the synthesis mixture, ageing, seeding, agitation during the synthesis, and crystallization time and temperature have a distinct influence upon the synthesis and need to be considered. The IZA maintains a collection of verified recipes [34] that are available through their web page [7].

C2.12.4 POST-SYNTHETIC MODIFICATION OF ZEOLITES

Only in very rare cases can zeolites be used directly in the form in which they were originally synthesized. For many larger-scale industrial applications, for instance, the synthetically obtained zeolite powders must be formed into larger attrition and crush-resistant particles using inorganic binder materials [33]. In most cases a thermal treatment in air (calcination) is at least required to decompose the organic template, to dehydrate the zeolite and to desorb impurities [35]. This holds particularly true if the proton form of a zeolite is desired from the ammonium form for acid catalysis.

The simplest and most commonly applied modification method is ion exchange [24]. By far the vast majority of studies regarding ion exchange of zeolites were carried out in aqueous solutions. Ion exchange processes are described by equilibrium ion-exchange isotherms; these relate the equivalent fraction of the ion in the zeolite to that in solution. Zeolites exhibit different selectivities for ion exchange depending upon factors such as the silicon-to-aluminium ratio of the zeolite, the size, charge and polarizability of the cation, the solvation medium and the size and stability of the solvation sphere. For example, the selectivity series for exchange of monovalent cations into NaY is $\text{Ag} \gg \text{Tl} > \text{Cs} > \text{Rb} > \text{K} > \text{Na} > \text{Li}$, lithium being the smallest cation but with the largest hydration sphere. The presence of ion exchange positions in small cages such as the sodalite cage, which is accessible through a 6-ring window with an effective pore diameter of about 2.5\AA , may hinder or exclude the exchange of certain cations

that are too bulky to penetrate it. As a consequence, the maximum theoretical exchange capacity, which is determined by the number of lattice aluminium atoms, is not attainable and only 63–65% ion exchange is observed (e.g. for Rb and Cs exchange of NaY). Another undesired effect that is frequently observed is the simultaneous exchange of protons, particularly in the case of transition metals which tend to hydrolyse in aqueous solution. Additionally, when di- and trivalent cations are exchanged they tend to hydrolyse and exchange as lower-valency hydroxylated ions; subsequent migration from the ion exchange positions and clustering upon calcination is frequently observed. More recently, solid-state ion exchange using fused salts has been successfully applied, a procedure in which the zeolite and the salt are ground and subsequently heated to elevated temperatures either under aerobic or anaerobic conditions [36]. For exchange processes that are difficult in aqueous solution, vapour phase exchange using salts that are volatile at elevated temperatures such as FeCl_3 and GaCl_3 has yielded a high degree of ion exchange [37].

Dealumination, the removal of aluminium from the zeolite framework, is generally applied for stabilizing zeolites and for the formation of mesopores which help in overcoming diffusional problems in the zeolite micropores. For instance ultrastable Y zeolite (USY), a major component of fluid-catalytic cracking (FCC) catalysts, is obtained by a twofold ammonium exchange with intermediate steam calcination leading to a dealuminated material that retains its structure upon heating up to 900 °C [38]. Due to the susceptibility of the Al–O bond to hydrolysis, one of the issues associated with the use of zeolites is their stability in acidic and basic media or under hydrothermal conditions as commonly experienced in speciality chemical manufacture or hydrocarbon conversion. During such treatment, Si–O–Al bonds are

broken and the aluminium is removed from its tetrahedral lattice position, leaving silanol nests as lattice defects (see figure C2.12.8). The extensive and uncontrolled removal of aluminium from the zeolite framework leads to the destabilization of the remaining framework and ultimately to complete structural collapse. A controlled dealumination treatment is the most successfully applied postsynthetic modification for stabilizing zeolites. The most commonly used dealumination methods are acid leaching, steaming, chemical treatment with silicon fluorides, and direct replacement of framework aluminium by means of gaseous SiCl_4 [24]. Combinations of these techniques have proven useful for the subsequent extraction of extra-framework species for example, steaming followed by acid leaching or extraction with complexing agents such as EDTA or oxalic acid. Inadvertent and thus undesired dealumination may occur during regular calcination of zeolites and during ion exchange in acidic medium, particularly in the case of zeolites with low lattice Si/Al ratio or a high concentration of lattice defects.

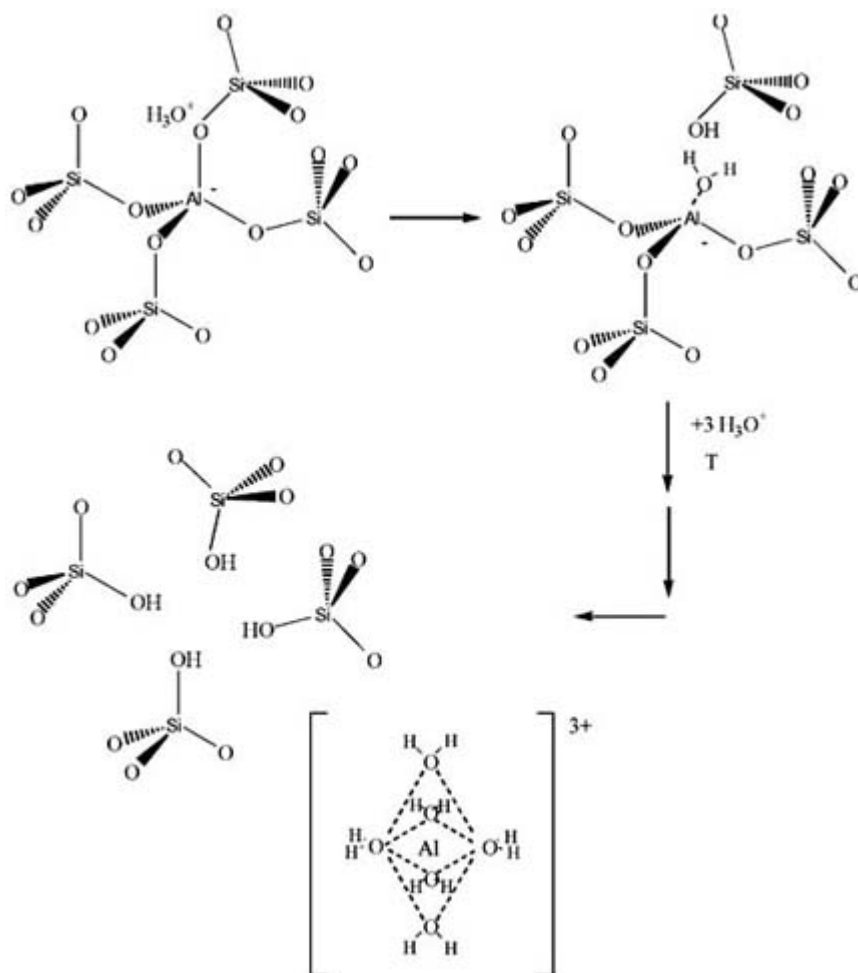


Figure C2.12.8. Schematics of the dealumination of zeolites. Water adsorbed on a Brønsted site hydrolyses the Al–O bond and forms the first silanol group. The remaining Al–O bonds are successively hydrolysed leaving a silanol nest and extra-framework aluminium. Aluminium is cationic at low pH.

-12-

Another important modification method is the passivation of the external crystallite surface, which may improve performance in shape selective catalysis (see C2.12.7). Treatment of zeolites with alkoxy silanes, SiCl_4 or silane, and subsequent hydrolysis or poisoning with bulky bases, organophosphorus compounds and arylsilanes have been used for this purpose [39]. In some cases, the improved performance was, however, not related to the masking of unselective active sites on the outer surface but rather to a narrowing of the pore diameters due to silica deposits.

C2.12.5 PHYSICAL AND CHEMICAL PROPERTIES OF ZEOLITES

Zeolites form small crystallites with an average size of about 1–2 μm . Specially modified synthesis methods have yielded crystals as small as 10 nm [40] and ranging up to millimetre size [41]. Crystal agglomeration and intergrowth are commonly observed. The density range of zeolites is from 1.9 to 2.3 g cm^{-3} . Due to the high porosity, specific surface areas are in the range of several 100 $\text{m}^2 \text{g}^{-1}$ with a micropore volume of the dehydrated zeolites in the range from 0.1 to 0.3 $\text{cm}^3 \text{g}^{-1}$. Due to the mobility of the hydrated cations, zeolites exhibit electrical conductivity. The sodium form of zeolites leads to a pH value between 9 and 12 in aqueous solution while the proton form reacts in an acidic manner. The lattice Si/Al ratio is the governing factor determining the overall physico-chemical properties. A schematic representation of the effect of the aluminium concentration on ion-exchange capacity, acid strength of the protons, resistance to acidic media, thermal stability and hydrophilicity is given in figure C2.12.9. Since every aluminium atom in the framework requires a charge balancing cation, the ion-

exchange capacity is proportional to the aluminium concentration. This also holds true if the cations are protons. The acid strength of each of these protons, however, increases with decreasing aluminium concentration and levels off at a constant value at a Si/Al ratio characteristic for each zeolite type. This effect is related to the mutual influence of aluminium atoms in close proximity to one another, referred to as next nearest neighbour (NNN) aluminium that are separated only by one SiO₄ tetrahedron [42]. For instance, in zeolite A with a Si/Al ratio of 1 only NNN aluminium is present. With decreasing aluminium concentration, aluminium atoms become more and more isolated in the silica matrix and only Si can be found in the NNN coordination shell. The overall acidity, the product of acid site concentration and acid strength, therefore goes through a maximum dependent on the Si/Al ratio. Since the Al–O bond is susceptible to hydrolysis in acidic medium, the acid and thermal stability are also higher when less aluminium is present in the zeolite lattice. The higher the aluminium concentration, the higher is the overall lattice charge leading to hydrophilic materials. The adsorption capacity of zeolite A decreases with decreasing polarity and polarizability of adsorbates. On the other hand, zeolites with a low aluminium concentration are increasingly more hydrophobic and selectively adsorb hydrocarbons over water.

-13-

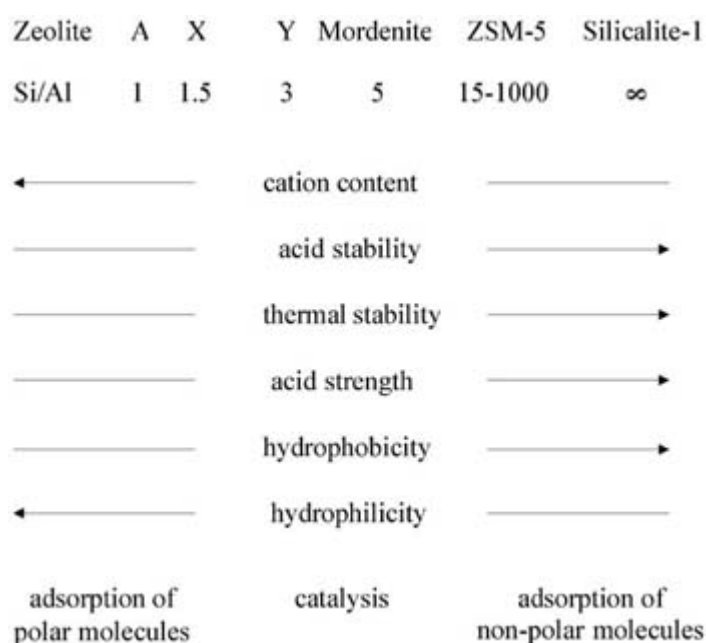


Figure C2.7.9. Effect of the Si/Al ratio of zeolites on their properties.

Factors other than the Si/Al ratio are also important. The alkali-form of zeolites, for instance, is *per se* not susceptible to hydrolysis of the Al–O bond by steam or acid attack. The concurrent ion exchange for protons, however, creates Brønsted acid sites whose AlO₄ tetrahedron can be hydrolysed (e.g. leading to complete dissolution of NaA zeolite in acidic aqueous solutions).

C2.12.6 CHARACTERIZATION OF ZEOLITES

Characterization of zeolites is primarily carried out to assess the quality of materials obtained from synthesis and postsynthetic modifications. Secondly, it facilitates the understanding of the relation between physical and chemical properties of zeolites and their behaviour in certain applications. For this task, especially, *in situ* characterization methods have become increasingly more important, that is, techniques which probe the zeolite under actual process conditions.

The first analytical tool to assess the quality of a zeolite is powder x-ray diffraction. A collection of simulated powder XRD patterns of zeolites and some disordered intergrowths together with crystallographic data is available from the IZA [43]. Phase purity and x-ray crystallinity, which is arbitrarily defined as the ratio of the intensity of

selected reflections in the diffractogram with respect to a standard material are determined in this way. Additional routine characterization techniques comprise elemental analysis for the determination of the chemical composition, N₂ and Ar adsorption for the determination of textural properties and scanning electron microscopy for the determination of the crystal size and morphology.

-14-

Solid-state nuclear magnetic resonance (NMR) spectroscopy applying magic angle spinning (MAS) and infrared (IR) spectroscopy have been traditionally used to characterize the zeolite itself. The widespread availability of modern Fourier transform spectrometers means that these instruments are becoming increasingly important in the study of dynamic processes occurring during adsorption and catalytic reactions [44]. *In situ* measurements of adsorbate interactions with precise control of temperature and partial pressure is considered state-of-the-art regarding IR spectroscopic measurements. The interesting spectral region is between 3800 and 1200 cm⁻¹ covering the stretching and deformation vibrations of the zeolite hydroxyl groups and most organic functional groups. Of particular importance is the differentiation of Lewis and Brønsted acid sites by means of pyridine adsorption. IR spectroscopy has meanwhile matured so far that *in situ* monitoring of the zeolite surface during catalytic reactions is possible. In MAS-NMR it is mainly ²⁷Al and ²⁹Si NMR that provides useful information about the local environment of aluminium and silicon in the zeolite matrix. For aluminium, differentiation between tetrahedral aluminium in the zeolite lattice and octahedral extra-framework aluminium is, in principle, possible. The number of AlO₄ tetrahedra directly linked to an SiO₄ tetrahedron can be determined from ²⁹Si NMR since different chemical shifts are observed for the corresponding Si nuclei. In the absence of large concentrations of silanol defects, which can be ascertained by ¹H cross-polarization measurements, the lattice Si/Al ratio can be determined from such data. Using *in situ* dehydration techniques, ¹H-NMR is also turning into a more commonly applied method providing quantitative information about the nature of zeolite hydroxyl groups. Recent developments have even led to the possibility of studying catalytic reactions on zeolites in continuous flow by means of NMR spectroscopy [45].

Alongside these techniques, microbalance measurements of adsorption capacities and kinetics, microcalorimetric measurements of adsorption processes and temperature-programmed desorption of base molecules have provided useful information about the thermochemistry of adsorption processes and the acidity characteristics of zeolites [46].

C2.12.7 APPLICATIONS OF ZEOLITES

The applications of zeolites can be divided into three major categories: ion exchange, adsorption and catalysis. The largest amount of zeolites is used in ion exchange applications while the largest value is derived from catalytic applications [1, 33].

The most important example for ion exchange applications is the use of zeolites as detergent additives for the removal of mainly Ca and partly Mg from washing waters. As an environmentally acceptable alternative, zeolites have taken over the traditional role of sodium polyphosphate which was a major contributor to the eutrophication of waters. Zeolite A is mainly used for this purpose and commercial syntheses have been optimized for the efficient preparation of large quantities from cheap, natural resources giving products with a homogeneous crystal size below 5 μm. The annual global production for this application has reached several hundred thousand tons [33]. Important natural zeolite-based ion exchange applications comprise the selective removal of ammonium from industrial and municipal wastewater and the removal of ¹³⁷Cs and ⁹⁰Sr from radioactive wastewater [29].

The excellent suitability of zeolites as adsorbents derives from three main characteristics: namely, a high intracrystalline void volume, a high electrostatic field and the molecular sieving effect. The high adsorption capacity of zeolites for water has already been mentioned, which is exploited when zeolites are used as static drying agents for refrigerants, in double glazing, or as additives in the manufacture of solvent-free polyurethanes. Industrial processes based on temperature-swing or pressure-swing adsorbents are applied for the desiccation of natural gas and cracking gas and the

purification of natural gas prior to liquefaction [33]. It is mainly the various ion-exchanged forms of zeolite A that are used for these applications.

Impurities such as hydrocarbons, carbon dioxide and water that are present in ppm levels can be successfully removed using molecular sieves. In particular, the presence of hydrocarbons is hazardous in cryogenic air separation plants since they form explosive mixtures with liquefied oxygen. Furthermore, the low degree of interaction of hydrogen with zeolites is exploited in the production of ultrahigh purity hydrogen where trace amounts of CO, CO₂, N₂, O₂, Ar, CH₄ and water are removed to ppm levels. Zeolites are also an excellent means of gas separation. The higher degree of interaction of N₂ with the cations in Ca-exchanged zeolite A or Li-exchanged zeolite X as compared to O₂ can be used to enrich oxygen up to 95 wt%. Another application, the selective removal of volatile organic compounds (VOC) from humid exhaust vents such as in commercially-used frying pans, derives from the selective adsorption of hydrocarbons on hydrophobic high-silica zeolites [47]. Molecular sieving effects, that is, selective adsorption due to size exclusion effects, are exploited in the separation of *n*-paraffins from iso-paraffins. Along the same lines ethylbenzene, *para*-ethyltoluene or *para*-xylene can be separated from their isomers.

Catalysis with zeolites has traditionally been located in the petroleum refining industry where zeolites have replaced the traditional amorphous silica–alumina as catalysts [48]. The zeolite catalysts (zeolite X at first and USY later) provided higher activity by orders of magnitude and also improved yields of motor gasoline. Their success as FCC catalysts led to the application in many refinery processes soon after. Over recent decades zeolites have moved increasingly into the manufacture of petrochemicals and base chemicals. The most recent trends are in an augmented use of zeolites for the production of fine and speciality chemicals [49, 50]. Additionally zeolites have shown remarkable potential in environmental catalysis, namely for exhaust purification (transition metal-exchanged zeolites) [51] or as replacements for traditionally used liquid catalysts such as sulphuric and hydrofluoric acid [50]. Alkylation, acylation, nitration or Beckmann rearrangement are some examples that are presently being studied.

The reasons for their excellent suitability as catalysts are multifaceted. Depending on their chemical state, zeolites can be used for acid catalysis, base catalysis, bifunctional catalysis (hydrogenation–dehydrogenation reactions coupled with acid catalysis), or redox reactions. As solid acids they provide a high acidity, both in terms of acid strength and acid site concentration, which is imperative for achieving high reaction rates in acid-catalysed reactions; corrosion is less of a concern compared to liquid acids. In this respect, theoretical work based on quantum chemical calculations has significantly contributed to the understanding of zeolite acidity in relation to carbenium ion chemistry [52, 53]. As inorganic solids, zeolites are easily separable by filtration facilitating product separation and permit fixed-bed flow-through operation. Further advantages arise from their excellent thermal stability and regenerability, properties sought for in heterogeneous catalysts. Some structures, like ZSM-5, show an exceptionally high intrinsic resistance to coke formation, which permits long reaction times between regeneration cycles. Of the utmost importance, however, is their ability to discriminate molecules based upon their size and shape, coined shape selectivity, which makes them attractive materials for intermediates and fine-chemicals synthesis. Following the original definition, shape selectivity can become apparent in three different ways (see figure [figure C2.12.10](#) [54]). It refers to the exclusion of molecules (reactant selectivity), the retardation of molecular transport within the zeolite pores depending on the molecular dimensions (product selectivity), or the confinement of the transition state for a certain reaction (transition state selectivity). A comprehensive discussion of the implications of shape selectivity was given recently [55].

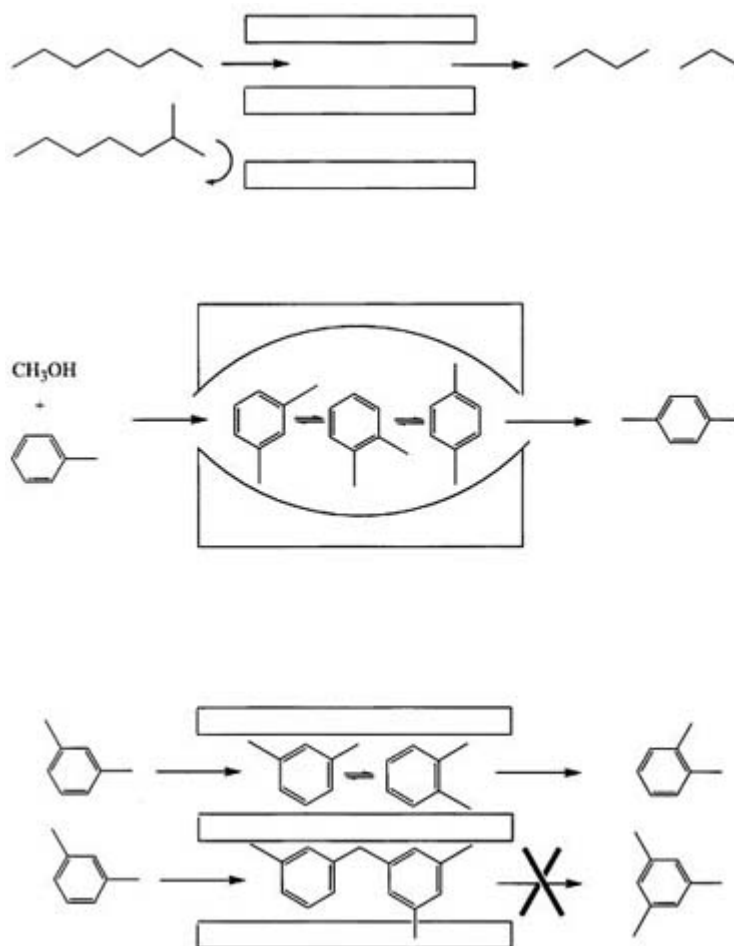


Figure C2.12.10. Different manifestations of shape-selectivity in zeolite catalysis. Reactant selectivity (top), product selectivity (middle) and transition state selectivity (bottom).

A typical example for reactant selectivity is the selectoforming process, the selective cracking of *n*-paraffins from reformates and naphthas using small-pore zeolites such as erionite. Only linear paraffins can penetrate into the zeolite micropore system and are converted while the desired branched and cyclic hydrocarbons remain unaffected. The same effect is being exploited in dewaxing processes and in the above mentioned *n*-paraffin/iso-paraffin separation by adsorption. The alkylation of toluene over ZSM-5 is an example of product selectivity. Due to its higher diffusivity, *para*-xylene can leave the pore system much more rapidly than the bulkier *ortho* and *meta* isomers which are continuously reequilibrated in the zeolite pores. Thereby, a higher yield of the *para*-isomer is obtained. During xylene isomerization, transition-state selectivity is manifested through the absence of trimethylbenzenes in the product which would originate from transalkylation reactions. The confined space in the zeolite pore prohibits the formation of the sterically demanding transition state for the bimolecular transalkylation.

Only a very few selected examples have been discussed. The number of processes based on shape-selective catalysis by zeolites is ever increasing, particularly in the field of speciality and fine chemicals and quite a few have been

commercialized. For a more comprehensive picture the reader is advised to consult the further reading suggestions.

REFERENCES

- [1] Moscou L 1991 The zeolite scene *Stud. Surf. Sci. Catal.* **58** 1–12
- [2] Cronstedt A F 1756 Observation and description of an unknown kind of rock to be named zeolites *Kongl Vetenskaps Akad. Handl. Stockholm* **17** 120–3
Sumelius I G 1992 Attempted translation of the original old-Swedish paper by Cronstedt *Molecular Sieves* ed M L Ocelli and H E Robson (New York: Van Nostrand Reinhold) pp 1–5
- [3] Damour A A 1942 Description de la faujasite, nouvelle espèce minérale *Ann. Mines* **4** 395–9
- [4] McBain J W 1932 Sorption by chabasite, other zeolites and permeable crystals *The Sorption of Gases and Vapors by Solids* (London: Routledge) pp 167–76
- [5] Taylor W H 1930 The structure of analcite ($\text{NaAlSi}_2\text{O}_6\cdot\text{H}_2\text{O}$) *Z. Kristallogr.* **74** 1–19
- [6] Barrer R M 1948 Syntheses and reactions of mordenite *J. Chem. Soc.* 2158–63
- [7] The IZA homepage on the World Wide Web is hosted by the Laboratory of Crystallography at the Swiss Federal Institute of Technology: <http://www.iza-online.org>
- [8] Meier W M, Olson D H and Baerlocher Ch 1996 *Atlas of Zeolite Structure Types* 4th revised edn (London: Elsevier)
- [9] Loewenstein W 1954 The distribution of aluminum in the tetrahedra of silicates and aluminates *Am. Mineral.* **39** 92–6
- [10] Xu T, Munson E J and Haw J F 1994 Toward a systematic chemistry of organic reactions in zeolites: *in situ* NMR studies of ketones *J. Am. Chem. Soc.* **116** 1962–72
- [11] Meier W M 1968 Zeolite structures *Molecular Sieves* (London: Society of Chemical Industry) pp 10–27
- [12] van Koningsveld H 1994 Structural subunits in silicate and phosphate structures *Stud. Surf. Sci. Catal.* **85** 35–76
- [13] Kokotailo G T and Meier W M 1980 Pentasil family of high silica crystalline materials *The Properties and Applications of Zeolites (Special Publication No. 33)* ed R P Townsend (London: The Chemical Society) pp 133–9
- [14] Casci J L 1994 The preparation and potential applications of ultra-large pore molecular sieves: a review *Stud. Surf. Sci. Catal.* **85** 329–56
- [15] Freyhardt C C, Tsapatsis M, Lobo R F, Balkus Jr K J and Davis M E 1996 A high-silica zeolite with a 14-tetrahedral-atom pore opening *Nature* **381** 295–8
- [16] Wagner P, Yoshikawa M, Lovallo M, Tsuji K, Tsapatsis and Davis M E 1997 CIT-5: a high-silica zeolite with 14-ring pores *Chem. Commun.* 2179–80
- [17] Kresge C T, Leonowicz M E, Roth W J, Vartuli J C and Beck J S 1992 Ordered mesoporous molecular sieves synthesized by a liquid-crystal template mechanism *Nature* **359** 710–12

- [18] Vartuli J C, Roth W J, Beck J S, McCullen S B and Kresge C T 1998 The synthesis and properties of M41S and related mesoporous materials *Molecular Sieves Science and Technology* vol 1, ed H G Karge and J Weitkamp (Berlin: Springer) pp 97–119
- [19] Flanigen E M, Bennet J M, Grose R W, Cohen J P, Patton R L, Kirchner R M and Smith J V 1978 Silicalite: a new hydrophobic crystalline silica molecular sieve *Nature* **271** 512–16
- [20] Bibby D M, Milestone N B and Aldridge L P 1979 Silicalite-2: a silica analogue of the aluminosilicate zeolite ZSM-11 *Nature* **280** 664–5
- [21] Gies H, Marler B and Werthmann U 1998 Synthesis of porosils: crystalline nanoporous silicas with cage- and channel-like void structures *Molecular Sieves Science and Technology* vol 1, ed H G Karge and J Weitkamp (Berlin: Springer) pp 35–64
- [22] Perego G, Millini R and Bellussi G 1998 Synthesis and characterization of molecular sieves containing transition metals in the framework *Molecular Sieves Science and Technology* vol 1, ed H G Karge and J Weitkamp (Berlin:

Springer) pp 187–228

- [23] Guth J-L and Kessler H 1999 Synthesis of aluminosilicate zeolites and related silica-based materials *Catalysis and Zeolites, Fundamentals and Applications* ed J Weitkamp and L Puppe (Berlin: Springer) pp 1–52
- [24] Kühn G H 1999 Modification of zeolites *Catalysis and Zeolites, Fundamentals and Applications* ed J Weitkamp and L Puppe (Berlin: Springer) pp 81–197
- [25] Schunk S A and Schuth F 1998 Synthesis of zeolite-like inorganic compounds *Molecular Sieves Science and Technology* vol 1, ed H G Karge and J Weitkamp (Berlin: Springer) pp 229–63
- [26] Szostak R 1998 Synthesis of molecular sieve phosphates *Molecular Sieves Science and Technology* vol 1, ed H G Karge and J Weitkamp (Berlin: Springer) pp 157–85
- [27] Martens J A and Jacobs P A 1999 Phosphate-based zeolites and molecular sieves *Catalysis and Zeolites, Fundamentals and Applications* ed J Weitkamp and L Puppe (Berlin: Springer) pp 53–80
- [28] Tschernich R W 1992 *Zeolites of the World* (Phoenix, AZ: Geoscience Press)
- [29] Colella C 1999 Natural zeolites in environmentally friendly processes and applications *Stud. Surf. Sci. Catal.* **125** 641–55
- [30] Jansen J C 1991 Synthesis of zeolites *Stud. Surf. Sci. Catal.* **58** 77–136
- [31] van Bekkum H, Geus E R and Kouwenhoven H W 1994 Supported zeolite systems and applications *Stud. Surf. Sci. Catal.* **85** 509–42
- [32] Mizumaki F 1999 Application of zeolite membranes, films and coatings *Stud. Surf. Sci. Catal.* **125** 1–12
- [33] Roland E and Kleinschmit P 1996 Zeolites *Ullman's Encyclopedia of Industrial Chemistry* (5th edn) vol A28, ed B Elvers and S Hawkins (Weinheim: VCH) pp 475–504
- [34] Robson H 1998 Verified syntheses of zeolitic materials *Microporous Mesoporous Mater.* **22**
- [35] Kouwenhoven H W and de Kroes B 1991 Preparation of zeolitic catalysts *Stud. Surf. Sci. Catal.* **58** 497–529
- [36] Karge H G 1997 Post-synthesis modification of microporous materials by solid-state reactions *Stud. Surf. Sci. Catal.* **105** 1901–48
- [37] Chen H Y and Sachtler W M H 1998 Activity and durability of Fe/ZSM-5 catalysts for lean burn NO_x reduction in the presence of water vapor *Catal. Today* **42** 73–83

- [38] McDaniel C V and Maher P K 1968 New ultra-stable form of faujasite *Molecular Sieves* (London: Society of Chemical Industry) pp 186–95
- [39] Szostak R 1991 Modified zeolites *Stud. Surf. Sci. Catal.* **58** 153–99
- [40] Madsen C and Jacobsen C J H 1999 Nanosized zeolite crystals—convenient control of crystal size distribution by confined space synthesis *Chem. Commun.* 673–4
- [41] Nadimi S, Oliver S, Kuperman A, Lough A, Ozin G A, Garces J M, Olken M M and Rudolf P 1994 Nonaqueous synthesis of large zeolite and molecular sieve crystals *Stud. Surf. Sci. Catal.* **84** 93–100
- [42] van Santen R A 1994 Theory of Brønsted acidity *Stud. Surf. Sci. Catal.* **85** 273–94
- [43] Treacy M M J, Higgins J B and von Ballmoos R 1996 *Collection of Simulated XRD Powder Patterns for Zeolites* 3rd revised edn (London: Elsevier)
- [44] Karge H G, Hunger M and Beyer H K 1999 Characterization of zeolites—infrared and nuclear magnetic resonance spectroscopy and x-ray diffraction *Catalysis and Zeolites, Fundamentals and Applications* ed J Weitkamp and L

Puppe (Berlin: Springer) pp 198–326

- [45] Hunger M and Horvath T 1995 A new MAS NMR probe for *in situ* investigations of hydrocarbon conversions on solid catalysts under continuous-flow conditions *Chem. Commun.* **1995** 1423–4
- [46] van Hooff J H C and Roelofsen J W 1991 Techniques of zeolite characterisation *Stud. Surf. Sci. Catal.* **58** 241–83
- [47] Fajula F and Plee D 1994 Application of molecular sieves in view of cleaner technology. Gas and liquid phase separations *Stud. Surf. Sci. Catal.* **85** 633–51
- [48] Blauwhoff P M M, Gosselink J W, Kieffer E P, Sie S T and Stork W H 1999 Zeolites as catalysts in industrial processes *Catalysis and Zeolites, Fundamentals and Applications* ed J Weitkamp and L Puppe (Berlin:Springer) pp 437–538
- [49] Hölderich W F and van Bekkum H 1991 Zeolites in organic syntheses *Stud. Surf. Sci. Catal.* **58** 631–726
- [50] Espeel P, Parton R, Toufar H, Martens J, Hölderich W and Jacobs P 1999 Zeolite effects in organic catalysis *Catalysis and Zeolites, Fundamentals and Applications* eds J Weitkamp and L Puppe (Berlin: Springer) pp 377–436
- [51] Traa Y, Burger B and Weitkamp J 1999 Zeolite-based materials for the selective catalytic reduction of NO_x with hydrocarbons *Microporous Mesoporous Mater.* **30** 3–41
- [52] Kazansky V B 1994 The catalytic site from a chemical point of view *Stud. Surf. Sci. Catal.* **85** 251–72
- [53] van Santen R A and Kramer G J 1995 Reactivity theory of zeolitic Brønsted acidic sites *Chem. Rev.* **95** 637–60
- [54] Csicsery S M 1976 Shape-selective catalysis *Zeolite Chemistry and Catalysis ACS Monograph* vol 171, ed J A Rabo (Washington, DC: American Chemical Society) pp 680–713
- [55] Weitkamp J, Ernst S and Puppe L 1999 Shape-selective catalysis in zeolites *Catalysis and Zeolites, Fundamentals and Applications* ed J Weitkamp and L Puppe (Berlin: Springer) pp 327–76

FURTHER READING

Breck D W 1974 *Zeolite Molecular Sieves: Structure, Chemistry, and Use* (New York: Wiley)

This is the first monograph that was devoted to structure, chemistry and use of zeolites. It reviews zeolite synthesis to 1973, gives a detailed structural description of synthetic and mineral zeolites, illustrates their physical properties and describes applications.

Barrer R M 1982 *Hydrothermal Chemistry of Zeolites* (London: Academic)

A complete survey to 1981 of low or medium Si/Al zeolite synthesis and transformations.

Jacobs P A and Martens J A 1987 *Synthesis of High-silica Aluminosilicate Zeolites* (Amsterdam: Elsevier)

A comprehensive summary of the preparation of high-silica zeolites.

Szostak R 1989 *Molecular Sieves, Principles of Synthesis and Identification* (New York: Van Nostrand Reinhold)

This book concentrates on synthesis and identification methods for molecular sieves including nonaluminosilicate molecular sieves and gives a good overview of structures and patented materials.

van Bekkum H, Flanigen E M and Jansen J C 1991 *Introduction to Zeolite Science and Practice* (Amsterdam: Elsevier)

A collection of detailed reviews covering all aspects of zeolite classification, synthesis, modification, characterization and applications.

Tschernich R W 1992 *Zeolites of the World* (Phoenix, AZ: Geoscience Press)

A book on natural zeolites, their classification, origin and use.

Dyer A 1988 *An Introduction to Zeolite Molecular Sieves* (Chichester: Wiley)

This monograph gives a good introduction into the fundamentals of zeolite science covering all aspects from classification over synthesis, characterization to applications.

Jansen J C, Stöcker M, Karge H G and Weitkamp J 1994 *Advanced Zeolite Science and Applications* (Amsterdam: Elsevier)

A compilation of lectures given at the 10th IZC Summer School covering specific advanced topics related to zeolite science (e.g. Rietveld refinement, host–guest interactions) and introducing novel applications such as dye-loaded zeolites for nonlinear optics, embedding of semiconducting clusters and zeolites in electrochemistry.

Karge H G and Weitkamp J 1998 *Molecular Sieves – Science and Technology* vol 1 *Synthesis* (Berlin: Springer)

This book focuses on various aspects of molecular sieve synthesis giving a broad overview of the different types of molecular sieves.

-21-

Weitkamp J and Puppe L 1999 *Catalysis and Zeolites: Fundamentals and Applications* (Berlin: Springer)

This book gives the most up-to-date account of the state of research and development regarding zeolite synthesis, modification, characterization and applications with a comprehensive list of cross references.

Chen N Y, Garwood W E and Dwyer F G 1989 *Shape Selective Catalysis in Industrial Applications* (New York: Dekker)

A very good description of processes, mainly in the oil refining and fuel upgrading sector, highlighting the impact zeolites have made on this industry.

-1-

C2.13 Plasma chemistry

Martin Schmidt and Kurt Becker

C2.13.1 INTRODUCTION

The plasma state is often referred to as the fourth state of matter [1]. It is characterized by the presence of free positive (and sometimes also negative) ions and negatively charged electrons in a neutral background gas. The charge carrier concentration can vary from 10^5 m^{-3} in a dilute interstellar plasma to 10^{28} m^{-3} in a dense stellar plasma. Most matter in the universe is found in the plasma state. Examples include the sun and other stars, interstellar matter and the terrestrial ionosphere. Naturally occurring plasmas on Earth are rare and include lightning and flames. Plasmas generated for technological applications include, among others, welding arcs, plasma torches, high-pressure lamps and the ignition spark in an internal combustion engine. In the efforts to solve the energy problem on Earth, magnetically confined plasmas in nuclear fusion reactors are one of several choices to achieve the extreme conditions under which nuclear fusion might occur [1]. This chapter focuses primarily on gaseous plasmas at pressures ranging from a fraction of an atmosphere to at most atmospheric pressure. Plasmas are generally created by supplying a sufficient amount of energy to a volume containing a neutral gas, so that free electrons and ions are generated from the atoms and molecules in the gas. The energy may be supplied in the form of electrical energy, heat, ultra violet radiation or particle beams. In technical plasma devices, the input energy is

generally supplied as electrical energy that causes the ignition of a gas discharge. Chemical reactions among the different neutral and ionic atomic and molecular species occur in this gaseous atmosphere (volume processes) and also at the surfaces that surround the plasma (surface or wall processes). The study and the technical utilization of these chemical reactions is referred to as plasma chemistry [2, 3, 4, 5, 6 and 7].

A brief description of a low-density non-equilibrium plasma is given followed by a review of its characteristic features and of the relevant collision processes in the plasma. Principles for the generation of plasmas in technical devices are discussed and examples of important plasma chemical processes and their technical applications are presented.

C2.13.2 THE CHARACTERIZATION OF PLASMAS

A plasma is a globally quasi-neutral system of free electrons and positive and negative ions in a neutral background gas consisting of atoms, molecules and free radicals (some of which may be in electronically and/or rotationally-vibrationally excited states) [8, 9]. Global quasi-neutrality means that the plasma contains overall an equal number of positive and negative charges. The electrons in the plasma have a mean kinetic energy that can range from less than 0.01 eV in an interstellar plasma to more than 10 keV in fusion plasmas. In most laboratory plasmas, the mean kinetic energy of the electrons is higher than the thermal energy corresponding to room temperature (0.025 eV) (figure C2.13.1). In some cases, the positive ions and the neutrals in the plasma also have temperatures significantly above room temperature. An additional criterion for the existence of a plasma (as opposed to a mere mixture of electrons, ions and neutrals) is that the charge carriers and their mutual electromagnetic interaction determine the properties of the system.

-2-

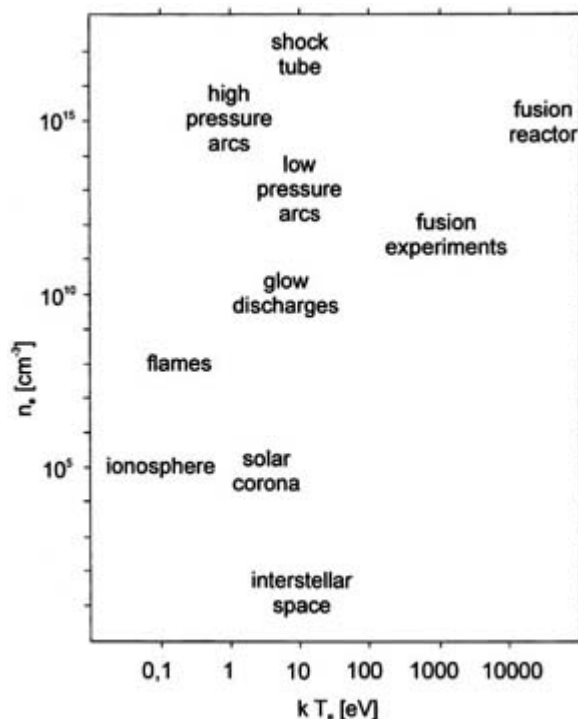


Figure C2.13.1. Electron energies and electron densities for different plasmas.

Any charge imbalance in a plasma (i.e. any local deviation from charge neutrality) results in a motion of the electrons that, in turn, leads to oscillations of the electrons with the electron plasma frequency ω_{pe} (Langmuir frequency)

$$\omega_{pl} = \sqrt{\frac{e_0^2 n_e}{\epsilon_0 m_e}}$$

where e_0 is the elementary charge, n_e is the electron density, ϵ_0 is the permittivity of free space, and m_e denotes the electron mass. Deviations from the global quasi-neutrality of the plasma are possible only locally in a small volume referred to as the Debye sphere whose radius is characterized by r_D , the Debye length:

$$r_D = \sqrt{\frac{\epsilon_0 k T_e}{e_0^2 n_e}}$$

where k is the Boltzmann constant and T_e refers to the temperature of the electrons. A plasma, as opposed to a mixture of electrons and ions, exists if the linear dimensions of the plasma (diameter, length) are large compared to the Debye length. The number of charge carriers in the Debye sphere amounts to 10^4 for an electron temperature of 10 000 K and an electron density of 10^{15} cm^{-3} . Electromagnetic forces in such systems are important if the plasma frequency is higher than the collision frequency of the charge carriers with the neutral particles in the plasma.

-3-

A non-thermal, non-equilibrium plasma is characterized by an electron temperature T_e much larger than the ion temperature T_i and the neutral gas temperature T_g ($T_e \gg T_i, T_g$). Typical non-thermal, non-equilibrium plasmas used in technological applications have electron temperatures of 10^4 to 10^5 K (corresponding to mean electron energies of about 0.5–5 eV). In plasmas that are in or near thermal equilibrium ('thermal' plasmas), the electron, ion and neutral temperatures are roughly equal. The degree of ionization α in a plasma is given by

$$\alpha = \frac{n_e}{n_e + n_0}$$

to denote the fraction of charge carriers (e.g. positive ions) in the plasma. In the above relation n_0 is the neutral gas density. Weakly ionized ('thin') plasmas have a degree of ionization in the range of 10^{-6} , whereas α approaches unity in fully ionized plasmas.

The generation of non-thermal plasmas by externally supplied electrical energy is possible because of the efficient interaction of the light electrons in the plasma with the external electric field. This results in a plasma with a high mean electron energy compared to the low energy of the near-thermal ions and neutrals. Energy transfer from the light electrons to the heavy particles in elastic collisions is negligible due to the difference in their masses. A selective energy transfer from the electrons to the heavy particles occurs *via* the various inelastic electron collision processes. In a molecular plasma, electron collisions will also lead to the formation of new species *via* dissociative processes. On the other hand, the energy gained by the ions in the external electric field is transferred efficiently to the gas molecules *via* elastic collisions. But this energy is generally small, so that the plasma will consist of a 'hot' electron gas and a 'cold' ionic and neutral gas.

The velocity distribution of the electrons in a plasma is generally a complicated function whose exact shape is determined by many factors. It is often assumed for reasons of convenience in calculations that such velocity distributions are Maxwellian and that the electrons are in thermodynamical equilibrium. The Maxwell distribution is given by

$$f(v) dv = 4\pi \left(\frac{m_e}{2\pi k T_e} \right)^{3/2} v^2 e^{-m_e v^2 / 2k T_e} dv$$

and the energy distribution with $kT_e = eU_e$ and $\frac{1}{2}mv^2 = eU$ is given by

$$h(U) dU = U^{1/2} \frac{2}{\sqrt{\pi} U_e^{3/2}} e^{-U/U_e} dU = U^{1/2} f(U) dU.$$

In plasmas, Maxwellian distributions are generally found only in cases where the energy exchange between the electrons is an important process. In most plasmas generated by external electrical fields, the observed velocity distributions deviate significantly from a Maxwellian distribution. The energy distribution function of the electrons is determined by their energy gain in the electric field and by their losses in elastic and inelastic collisions. The distribution function can be calculated using the Boltzmann equation [10, 11]. The shape of realistic electron energy distributions is often characterized by a lack of electrons with higher energies (figure C2.13.2), [12].

-4-

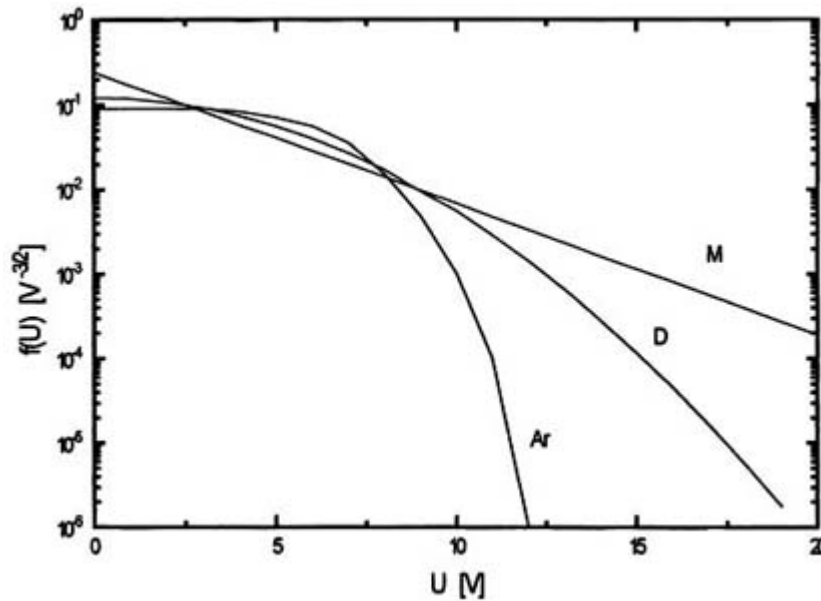


Figure C2.13.2. Electron energy distributions $f(U)$ for a mean electron energy of 4.2 eV, Maxwell distribution (M), Druyvesteyn distribution (D) and a calculated distribution (Ar) for an Ar plasma [12].

The transport of particles in the plasma is diffusive or convective for the neutrals, whereas the charge carriers move under the influence of the external and internal electric and magnetic fields. The drift velocity v of the charged particles is proportional to the electric field E :

$$v = \mu E$$

where μ denotes the mobility. The mobility is related to the diffusion coefficient D by the Einstein relation

$$\mu = e_0 D / kT.$$

The movement of the fast electrons leads to the formation of a space-charge field that impedes the motion of the electrons and increases the velocity of the ions (ambipolar diffusion). The ambipolar diffusion of positive ions and negative electrons is described by the ambipolar diffusion coefficient D_a :

$$D_a = \frac{D_i \mu_e + D_e \mu_i}{\mu_e + \mu_i}.$$

Non-thermal plasmas in contact with insulating walls (substrate) have an important property. The plasma with the hot electrons is positively charged relative to the wall (self-bias). A sheath with a positive space charge and an electric field is formed between the wall and the plasma. The hot electrons travel faster to the wall than the heavy

ions, but the two currents must be equal. This is achieved by the negative potential (10–20 V) of the wall which reflects the slow electrons and accelerates the ions towards the wall. The sheath potential V_S is determined for a planar surface by

-5-

$$V_S = \frac{kT_e}{2e} \ln \left(\frac{m_e}{2.3m_i} \right)$$

where m_i denotes the ion mass.

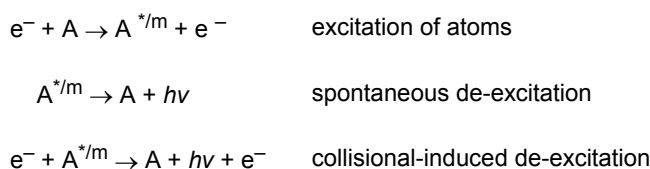
C2.13.3 COLLISION PROCESSES IN PLASMAS

Collision processes involving the different plasma components play an important role in non-thermal plasmas (table C2.13.1) [13, 14 and 15]. Electron collision processes are of particular importance because of the high temperature or high mean energy of the plasma electrons. Stepwise excitation and ionization, that is the excitation/ionization of an atom or molecule which is already in an excited or, in particular, in a metastable state, can occur with appreciable probability even though the concentration of excited/metastable species in a non-thermal plasma is generally low. The energy spacing between excited states is typically much smaller than the energy gap between the ground state and the first excited state. The number of low-energy electrons is typically much higher than the number of electrons with energies above about 10 eV and the excitation/ionization cross section out of an excited state is much larger than the cross section for excitation/ionization of ground-state species. This may result in rate coefficients (see below) for stepwise excitation/ionization that are quite large. Metastable species cannot decay *via* radiative dipole transitions to lower states. This results in a comparatively long lifetime of microseconds or even milliseconds for these species (compared to nanoseconds for excited states which can decay radiatively *via* dipole transitions). As a consequence, metastables can accumulate in the plasma and can be an efficient source of species for stepwise excitation/ionization processes or for super-elastic collisions in which the scattered electron gains energy. Ionization due to binary collisions involving metastable atoms or molecules is an efficient mechanism for charge carrier production. The generation of free radicals by electron collisions in molecular plasmas is an important precursor for plasma chemical reactions. Electron impact ionization is the fundamental process for sustaining a non-thermal plasma. Electron-impact-induced dissociation leading to the formation of free radicals is the most important reaction channel in plasma chemistry. In conventional chemistry, the formation of radicals is determined by the temperature of the entire system offering a different spectrum of secondary reactions from that resulting from radical production by electron collision in the cold neutral gas environment of a non-thermal plasma.

-6-

Table C2.13.1 Collision processes of electrons and heavy particles in non-thermal plasmas. The asterisk * denotes short-lived excited particles, the superscript m denotes long-lived metastable excited atoms or molecules.

Collisions of electrons



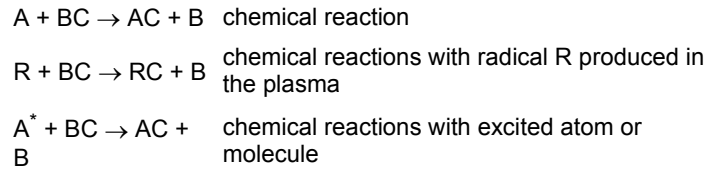
$e^- + A \rightarrow A^+ + 2e^-$	ionization of atoms
$e^- + AB \rightarrow AB^{*/m} + e^-$	excitation of molecules
$AB^* \rightarrow AB + h\nu$	spontaneous de-excitation
$e^- + AB^* \rightarrow AB + h\nu + e^-$	collisional-induced de-excitation
$e^- + AB^* \rightarrow A^{(*)} + B + e^-$	dissociation of molecules
$e^- + AB \rightarrow A + B^+ + 2e^-$	dissociative ionization
$e^- + A^{*/m} \rightarrow A + e^- + E_{kin}$	super-elastic collisions
$e^- + A^{*/m} \rightarrow A^{**} + e^-$	stepwise excitation
$e^- + A^{*/m} \rightarrow A^+ + e^-$	stepwise ionization
$e^- + A \rightarrow A^-$	attachment
$e^- + A^- \rightarrow A + 2e^-$	detachment
$e^- + A^+ \rightarrow A$	recombination
$e^- + A^+ + M \rightarrow A + M$	three-body collision recombination

Table continued on next page.

Table C2.13.1 Continued.

Collisions of heavy particles

$A^+ + B \rightarrow A + B^+$	charge transfer
$A^m + B \rightarrow A + B^+ + e$	Penning ionization
$A^m + A^m \rightarrow A + A^+ + e$	pair collision
$A^* + A \rightarrow A_2^+ + e$	Hornbeck-Molnar ionization
$A^+ + BC \rightarrow AC^+ + B$	ion-molecule reaction



The probability for a particular electron collision process to occur is expressed in terms of the corresponding electron-impact cross section σ which is a function of the energy of the colliding electron. All inelastic electron collision processes have a minimum energy (threshold) below which the process cannot occur for reasons of energy conservation. In plasmas, the electrons are not mono-energetic, but have an energy or velocity distribution, $f(v)$. In those cases, it is often convenient to define a rate coefficient k for each two-body collision process:

$$k = \int \sigma(v) v f(v) dv$$

where $\sigma(v)$ denotes the cross section (here written as a function of velocity rather than energy) and $f(v)$ represents the velocity distribution function of the electrons. Realistic plasmas typically exhibit complicated velocity distribution functions for the plasma electrons (see above). For the simple case of a Maxwellian velocity distribution of the electrons and a collision cross section, whose low-energy behaviour (in a limited range of impact energies E above the threshold energy E_{thr}) can be described by the expression

$$\sigma(E) = \pi r^2 (1 - E_{\text{thr}}/E)$$

the resulting rate coefficient has the form

$$k(T) = \pi r^2 (8kT_e/\pi m_e)^{1/2} \exp(-E_{\text{thr}}/kT_e).$$

This expression corresponds to the Arrhenius equation with an exponential dependence on the threshold energy and the temperature T_e . The factor in front of the exponential function contains the collision cross section and implicitly also the mean velocity of the electrons.

C2.13.4 PLASMA GENERATION

The usual means for the generation of technological plasmas is by supplying electrical energy to the gas in the plasma reactor (figure C2.13.3) and [figure C2.13.4](#). The electrons are accelerated in the external electric field and transfer energy by collisions with the other particles to the plasma. Depending on the time dependence of the sustaining external electric field, discharges are classified as direct current (DC) or alternate current (AC) discharges. The DC low-pressure normal glow discharge between two plane electrodes in a cylindrical glass tube is the prototypical DC discharge and has been studied extensively for about 100 years [16]. Such discharges exhibit characteristic luminous structures. The brightest part of the discharge is the negative glow, which is separated from the cathode by the cathode dark space (Crookes or Hittorf dark space). The large drop of the electrical potential in this cathode dark space is called the cathode fall. The positive column and the negative glow are separated by the

Faraday dark space. The positive column extends to the anode, which may be covered by the anode glow.

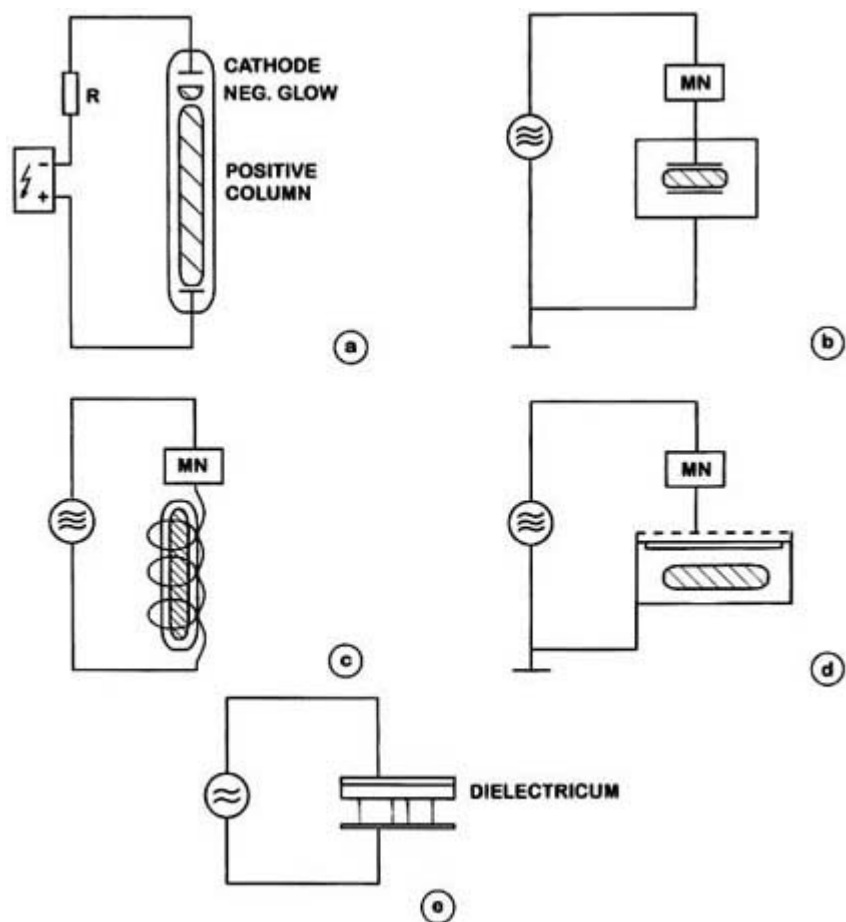


Figure C2.13.3. Schematic illustrations of various electric discharges: (a) DC-glow discharge, R denotes a resistor; (b) capacitively coupled RF discharge, MN denotes a matching network; (c), (d) inductively coupled RF discharge, MN denotes matching network; (e) dielectric barrier discharge.

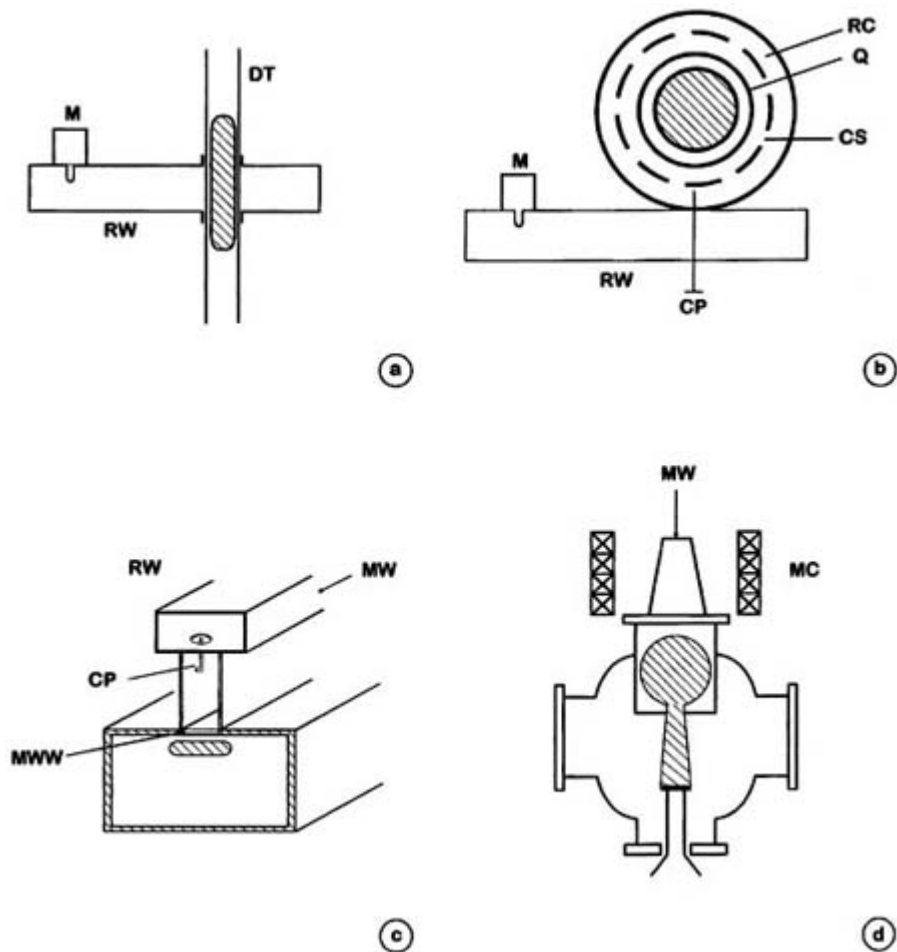


Figure C2.13.4. Schematic illustrations of selected microwave discharges: (a) the discharge tube DT is inserted through a rectangular wave guide RW parallel to the electric field, M denotes the magnetron; (b) plasma reactor with slots, CP denotes the coupling probe, RC refers to the ring circulator CS to the coupling slots and Q denotes the quartz tube [22]; (c) planar microwave plasma source: MW, microwave and MWW, microwave window [21]; (d) electron cyclotron plasma reactor, MC denotes the magnetic coils.

The microscopic processes in the DC glow discharge are fairly well understood [17]. Positive ions are accelerated by the high electric field in the cathode fall towards the cathode surface. The collisions of the energetic ions with the surface sputter neutral atoms and, most importantly, produce secondary electrons which are accelerated in the cathode fall. The energetic electrons transfer most of their energy in the cathode dark space and the negative glow to heavy particles in inelastic collisions by excitation and dissociation and create charge carriers by impact ionization. The discharge regions close to the cathode (cathode fall and negative glow) are necessary for establishing a self-sustaining glow discharge. A positive column is formed only under conditions where there is a long narrow separation between cathode and anode so that significant charge carrier losses to the surrounding discharge wall occur. The electrons, which have lost most of their energy in the negative glow, gain energy in the longitudinal electrical field of the positive

column. A steady-state electron energy distribution is formed that produces ions and electrons in sufficient numbers to balance the charge carrier losses at the wall.

Discharges sustained by time-varying electric fields are referred to as pulsed or AC discharges. High-frequency electromagnetic fields in the radio-frequency (RF) [8] or microwave [18] range are of particular interest for generating discharge plasmas for various plasma chemical applications.

In a capacitively coupled RF discharge, the electrodes are covered by sheath regions similar to the cathode dark space in the DC glow discharge. The bulk plasma occupies the region between the electrodes. The coupling capacitor between the RF generator and the powered electrode and the appropriate choice of different areas for the (smaller) powered electrode and (larger) grounded electrode leads to a negative DC potential between the plasma and the powered electrode, the so-called self-bias potential. The self-bias accelerates the ions near the powered electrode to energies of up to a few hundred electron volts. In contrast to the DC discharge, the electrodes of RF discharges must not be in conductive contact with the plasma. The most commonly used RF frequency is 13.56 MHz; the pressure range is between 0.1-100 Pa. Such discharges are successfully used for the plasma-assisted deposition of thin films, for plasma etching and for the sputtering of insulating materials.

The inductively coupled plasma [19] is excited by an electric field which is generated by an RF current in an inductor. The changing magnetic field of this inductor induces an electric field in which the plasma electrons are accelerated. The helicon discharge [20] is a special type of inductively coupled RF discharge.

Gas discharge plasmas have also been successfully excited by microwaves. Two characteristic features of the microwaves are (i) the fact that the wavelength is of the order of the dimensions of the plasma apparatus (the standard frequency of 2.45 GHz corresponds to a wavelength of 12.2 cm) and (ii) the short period of the exciting microwave field. Only electromagnetic waves with a frequency higher than the electron plasma frequency f_0 can penetrate into the plasma. Waves with lower frequencies than f_0 are reflected. The electron density which corresponds to the frequency f_0 is called the cut-off density; however, electromagnetic fields can penetrate through a skin depth into the plasma. This skin sheath permits a partial absorption of electromagnetic power even above the cut-off density. These facts limit the electron density for a 2.45 GHz excitation to $10^{11} - 10^{12} \text{ cm}^{-3}$.

Plasmas excited by microwaves may be produced in closed structures, in open structures and in resonance with a magnetic field. In closed structures, the plasma vessel is surrounded by metallic walls. Depending on the resonance conditions either multi-mode or single-mode cavities are used. Discharges in open structures include microwave torches, slow-wave structures and surfatrons in which the plasma is generated by the excitation of surface waves. Various types of slotted waveguides are applied successfully for the excitation of specially shaped microwave plasmas for technical applications [21, 22]. Such configurations produce plasmas with electron concentrations up to 10^{12} cm^{-3} in a broad pressure range.

Microwave discharges at pressures below 1 Pa with low collision frequencies can be generated in the presence of a magnetic field B where the electrons rotate with the electron cyclotron frequency. In a magnetic field of 875 G the rotational motion of the electrons is in resonance with the microwaves of 2.45 GHz. In such low-pressure electron cyclotron resonance plasma sources collisions between the atoms, molecules and ions are reduced and the formation of unwanted particles in the plasma volume ('dusty plasma') is largely avoided.

A special type of the RF discharge is the silent or dielectric barrier discharge [23] which can be operated at pressures

from 0.1 – 10 bar. Such a discharge was already used in 1857 by Siemens [24] for the production of ozone from air or oxygen. The silent discharge is generated between two electrodes with a dielectric barrier covering at least one electrode. The gas-filled gap is small, rarely exceeding a few millimetres. Voltages of 5 – 100 kV with frequencies from 50 Hz up to 1 MHz are necessary to sustain these discharges. The breakdown is connected with the formation of a large number of statistically distributed filaments (filament diameter 0.1 mm). The charge carriers from the plasma remain on the dielectric barrier and compensate the external electric field. Therefore, the lifetime of the filaments is very short (1 – 10 ns). The current density in one filament can be as high as $100 - 1000 \text{ A cm}^{-2}$ with electron densities in the range $10^{14} - 10^{15} \text{ cm}^{-3}$, and electron energies in the range 1 – 10 eV. Homogeneous dielectric barrier discharges are also observed. This type is called atmospheric pressure glow discharge [23].

Plasmas are used for the processing of solid surfaces such as in etching, cleaning and oxidizing and for the deposition of various thin films. In these applications chemical processes occur in the plasma volume as well as on the surfaces. The substrate may be positioned within the active plasma region or outside (remote plasma processing). Plasma chemistry is used for the treatment of gases to create and to change gaseous compounds in homogeneous gaseous reactions or heterogeneous surface processes. Several examples are discussed below.

C2.13.5.1 PLASMA SURFACE PROCESSES

(A) PLASMA ETCHING

Plasma etching is an important process in the manufacture of microelectronic devices [25]. It is used for pattern transfer during the fabrication of integrated circuits. Structured masks of photoresist determine where the etching should occur and where not. For structures that are small in comparison with the thickness of the mask, anisotropic etching using a non-thermal plasma is the only feasible technique. The removal (sputtering) of atoms or molecules from a solid surface is possible by momentum transfer from the heavy particles impinging on the surface and *via* collision cascades. Chemical surface reactions initiated by reactive particles from the plasma can lead to the formation of volatile reaction products. The nature of a gaseous reaction product formed in the surface reactions often determines the choice of etching gas used in a particular application (table C2.13.2, [3, 9]). For instance, fluorine-containing molecules are used in the etching of silicon because SiF₂, the main reaction product, is highly volatile. Hydrocarbons are etched using O₂ because of the benign by-products H₂O and CO₂. The etching of Al uses Cl-containing gases because of the high vapour pressure of AlCl₃. Ion impact leads to non-isotropic etching owing to the directed flow of the ions through the plasma sheath to the surface. Ion impact can also influence the formation of inhibitors, e.g. sidewall passivation (figure C2.13.5). Characteristic parameters to describe the etching process are the etch rate, the selectivity, the anisotropy, the uniformity of the process across the wafer surface and the possible material damage. Chemical reactions are isotropic and have a high etch rate and a high selectivity. Material damage is usually negligible. Physical sputtering processes are characterized by high anisotropy, low selectivity and low etch rates. The impact of energetic ions can damage the substrate. In these processes, the plasma serves two purposes. Firstly, it activates the often inert feed gases of the plasma gas mixtures through the formation of reactive radicals and, secondly, the ions are accelerated in the plasma sheath and impact on the substrate nearly perpendicularly to the surface and influence the surface processes.

Table C2.13.2 Gases for etching of various materials [3, 4].

Silicon	CF ₄ /O ₂ , CF ₂ Cl ₂ , CF ₃ Cl, SF ₆ /O ₂ /Cl ₂ , Cl ₂ /H ₂ /Cl ₂ F ₂ /CCl ₄ , C ₂ ClF ₅ /O ₂ , SiF ₄ /O ₂ , NF ₃ , CCl ₄ , C ₂ ClF ₅ /SF ₆ , C ₂ F ₆ /CF ₃ Cl, Br ₂ , CF ₃ Cl/Br ₂
SiO ₂	CF ₄ /H ₂ , C ₂ F ₆ , C ₃ F ₈ , CHF ₃ /O ₂
Si ₃ N ₄	CF ₄ /O ₂ /H ₂ , C ₂ F ₆ , C ₃ F ₈ , CHF ₃ , NF ₃ , CHF ₃ /O ₂
Organics, polymers	O ₂ , CF ₄ /O ₂ , SF ₆ /O ₂
Silicides	CF ₄ /O ₂ , NF ₃ , SF ₆ /Cl ₂ , CF ₄ /Cl ₂
Al	BCl ₃ , BCl ₃ /Cl ₂ , CCl ₄ /Cl ₂ /BCl ₃ , SiCl ₄ /Cl ₂
Cr	Cl ₂ , CCl ₄ /Cl ₂
Mo, Nb, Ta, Ti, W	CF ₄ /O ₂ , SF ₆ /O ₂ , NF ₃ /H ₂
Au	C ₂ Cl ₂ F ₄ , Cl ₂ , CClF ₃
GaAs	BCl ₃ /Ar, Cl ₂ /O ₂ /H ₂ , CCl ₂ F ₂ /O ₂ /Ar/He, CCl ₄ , PCl ₃ , HCl, Br ₂ , COCl ₂

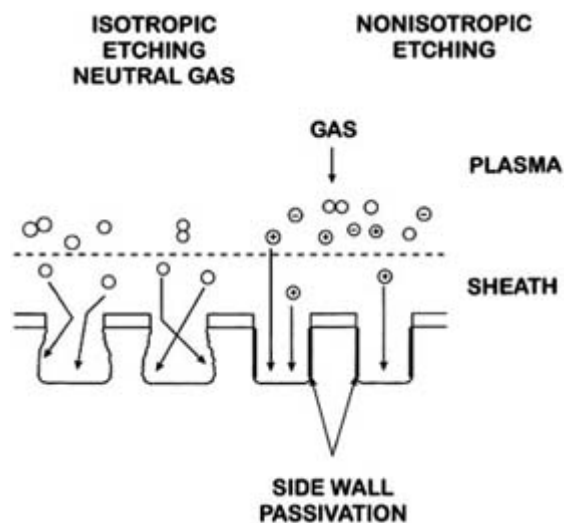
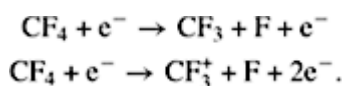


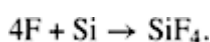
Figure C2.13.5. Schematic illustrations of isotropic etching by a neutral gas and anisotropic plasma etching.

-13-

As an example, we look at the etching of silicon in a CF₄ plasma in more detail. Flat Si wafers are typically etched using quasi-one-dimensional homogeneous capacitively or inductively coupled RF-plasmas. The important process in the bulk plasma is the formation of fluorine atoms in collisions of CF₄ molecules with the plasma electrons



The subsequent reaction of the F atoms with the silicon surface leads to the formation of the volatile product SiF₄:



The flux of F radicals to the wafer is nearly isotropic. Anisotropic etching is due to ions that are incident on the wafer essentially perpendicular to the surface (see above).

The sidewall protection by a thin polymeric film is an additional important process that occurs. The formation of F atoms in the gas phase leads to the formation of CF₃ radicals among other species. These radicals are the precursors for the deposition of a polymeric C_xF_y film on the substrate surface. The growth of this film is governed by a balance between the ion-induced fragmentation and the desorption of CF_x particles. The thickness of this film is typically 1–6 nm [26]. The F atoms do not react directly with the silicon surface. They diffuse through this film and then react. The volatile SiF₄ reaction products diffuse back into the gas phase after penetrating the thin film. The ion bombardment generates additional F atoms in the fragmentation of the C_xF_y film and increases the diffusion velocity of the F atoms and the SiF₄ molecules due to the energy deposited into the film. It is obvious that the ion current density at the sidewall of the trench is much smaller than the density at the bottom of the trench. This sidewall passivation mechanism is crucial for the success of anisotropic etching. The formation of sidewall layers also results in characteristic angular distributions of the ions at the wall. CF₃⁺ ions which have a high etch potential have a narrow angular distribution, whereas CF₂⁺ and CF⁺ ions, which are responsible for fluorocarbon layers, have a broad angular distribution [27]. The addition of oxygen to the CF₄ plasma leads to an increase of the etch rate, but decreases the anisotropy. This is understandable on the basis of the gas phase chemistry. CO₂ is formed, the CF_x concentration is reduced, and the formation of fluorine is enhanced. The decrease of the CF_x concentration impedes the formation of the protective sidewall layers.

A special case of plasma etching is the etching of hydrocarbons in an oxygen discharge. The removal of photoresist in oxygen-containing plasmas is a frequently employed process in the semiconductor industry. The cleaning of metallic work pieces covered with organic surface layers is another widely applied technique in industry. The oxygen plasma generates oxygen atoms that react with the surface contaminant to form volatile CO_2 and H_2O . Plasma cleaning causes low thermal stress of the surface. Ecologically and environmentally harmful solvents are largely avoided.

(B) SURFACE TREATMENT

Plasmas are successfully applied in surface oxidation at low substrate temperatures. In the plasma oxidation process the substrate is held at a floating potential in an oxygen plasma [9]. Plasma anodization is usually carried out with a positively biased substrate. The high oxidation rates of plasma anodization are a result of the high currents of electrons and negative ions to the substrate. High-quality oxide layers with excellent electrical properties, which cannot be achieved by standard thermal oxidation methods [28], are produced by a remote oxygen plasma and keeping the substrate surface at room temperature.

-14-

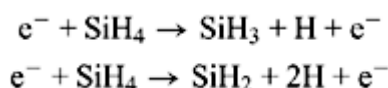
The surface properties of solids are controlled by the chemical structure of the surface layer: e.g., CH_3 and CF_3 groups at the surface produce a high-degree hydrophoby. Oxygen- and nitrogen-containing groups (e.g., $-\text{OH}$, $-\text{NH}_2$) produce a hydrophilic surface with high adhesion and unique properties for printing, painting, glueing and cell growth. Treatment of polymers in 'mild' (low ion energy) remote plasmas in selected gases provides highly selective surface properties. Biomaterials prepared by plasma chemical methods are sterile [29].

(C) THIN-FILM DEPOSITION

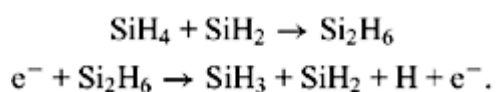
Thin films with a broad spectrum of properties can be deposited in plasma chemical processes [30, 31]. Hard coatings such as diamond films and TiN films, soft plasma polymer films, insulating SiO_x films, highly conducting Si films, anti-reflection coatings, semi-permeable membranes and very effective diffusion barriers can be deposited. Important parameters for the film deposition are (i) the nature of the precursor, (ii) the gas mixture and (iii) the selected plasma parameters. The advantage of plasma-assisted processes is the possibility to work with lower substrate temperatures than in pure chemical vapour deposition techniques, since the important reactions are initiated by energetic electrons. Plasma-deposited films have a high substrate adhesion because the substrate surface at the beginning of the deposition is activated by the plasma-surface interaction.

The deposition of amorphous hydrogenated silicon (a-Si:H) from a silane plasma doped with diborane (B_2H_6) or phosphine (PH_3) to produce p-type or n-type silicon is important in the semiconductor industry. The plasma process produces films with a much lower defect density in comparison with deposition by sputtering or evaporation.

The SiH_3 radical is the dominant growth precursor for the formation of the a-Si:H films in a low-temperature silane plasma [32]. Silane molecules are dissociated by energetic plasma electrons:



followed by the reactions



The SiH_3 radical physisorbs on the a-Si:H surface and recombines there with another SiH_3 radical to form disilane Si_2H_6 , or abstracts H from the surface to form a dangling bond and SiH_4 . The film growth is determined by the chemisorption of the SiH_3 radical on a free dangling bond site by formation of a Si-Si bond. The cross-linking of

neighbouring Si-H bonds leads to the elimination of H₂.

Admixtures of oxygen or oxidizing agents such as N₂O to the silane plasma enable the deposition of SiO₂ films. Other Si-containing compounds such as SiCl₄ or tetraethoxysilane (Si(OCH₂CH₃)₄) are used for plasma-enhanced SiO₂ deposition at lower temperatures [33].

The deposition of organic films by plasma polymerization is an important application of non-thermal plasmas [30]. Plasma polymers are formed at the electrodes and the walls of electrical discharges containing organic vapours. Oily products, soft soluble films as well as hard brittle deposits and powders are formed. The properties of plasma

-15-

polymers are similar to those of conventional polymers, but their structure is different. The often highly cross-linked material is not characterized by a mere repetition of the basic units. The properties of the polymers are essentially determined by the deposition conditions as well as by the plasma and by the gas flow rather than by the particular monomer. All organic compounds can be used as monomers or precursors for plasma polymerization. Functional groups such as double bonds are not necessary. Thus methane can be used as a precursor for plasma polymerization. Plasma polymer films have numerous advantages over conventional films: good adhesion on very different substrates, freedom from pinholes, good conformation to various substrate surfaces, a high degree of cross-linkage, chemical inertness and low levels of leachables [29].

The reaction mechanisms of plasma polymerization processes are not understood in detail. Poll *et al* [34] (figure C2.13.6) proposed a possible generic reaction sequence. Plasma-initiated polymerization can lead to the polymerization of a suitable monomer directly at the surface. The reaction is probably triggered by collisions of energetic ions or electrons, energetic photons or interactions of metastables or free radicals produced in the plasma with the surface. Activation processes in the plasma and the film formation at the surface may also result in the formation of non-reactive products.

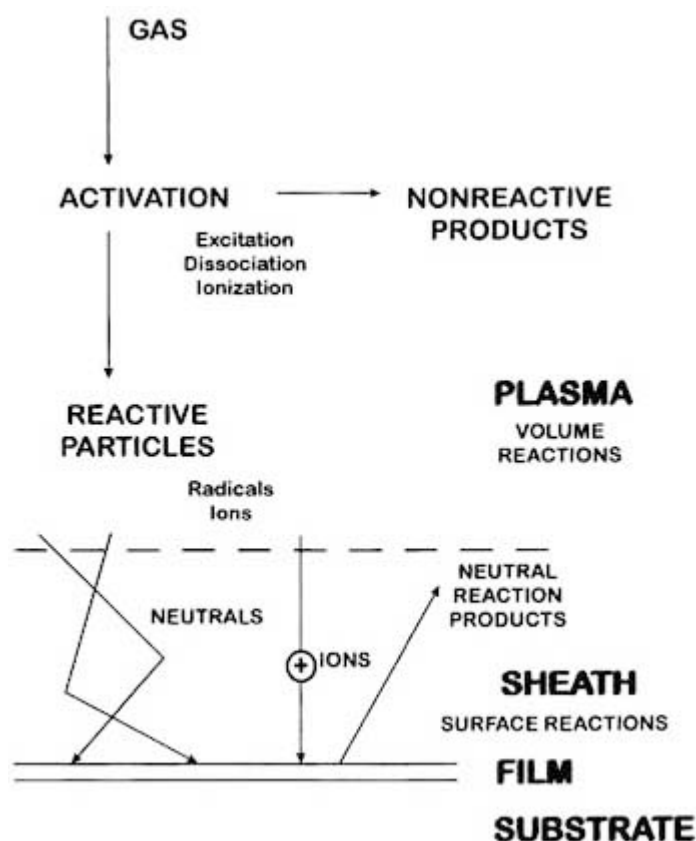


Figure C2.13.6. Schematic illustrations of plasma - assisted thin - film deposition.

An important and well studied example is the deposition of plasma-polymerized fluorinated monomer films [35]. Monomers are fluoroalkyls, fluorohydroalkyls, cyclo-fluoroalkyls, as well as unsaturated species. The actual

-16-

monomers are the CF_x radicals from which the polymer deposit is built. Electron impact produces the active species: ions, F atoms and F_2 molecules, and CF_x radicals. CF^+ and CF_2^+ ions are more effective for the polymer deposition than CF_3^+ ions (see above) [27]. The film growth occurs by the addition of CF_x radicals to previously activated sites. The activation of surface sites occurs *via* collisions of charged particles, *via* ion collisions on negatively charged substrates and *via* electron collisions on positively charged substrates. The fluorine-to-carbon ratio influences the fluorine content in the plasma which, in turn, determines whether polymerization or etching is the dominant process (figure C2.13.7, [36]). The admixture of oxygen enhances the fluorine concentration and thus the etching properties by reducing the recombination probability of F atoms by formation of CO, CO_2 , COF_2 . The addition of H_2 reduces the fluorine concentration *via* the formation of HF, thus enhancing the polymerization. The competition between etching and deposition is also influenced by other conditions such as the substrate bias.

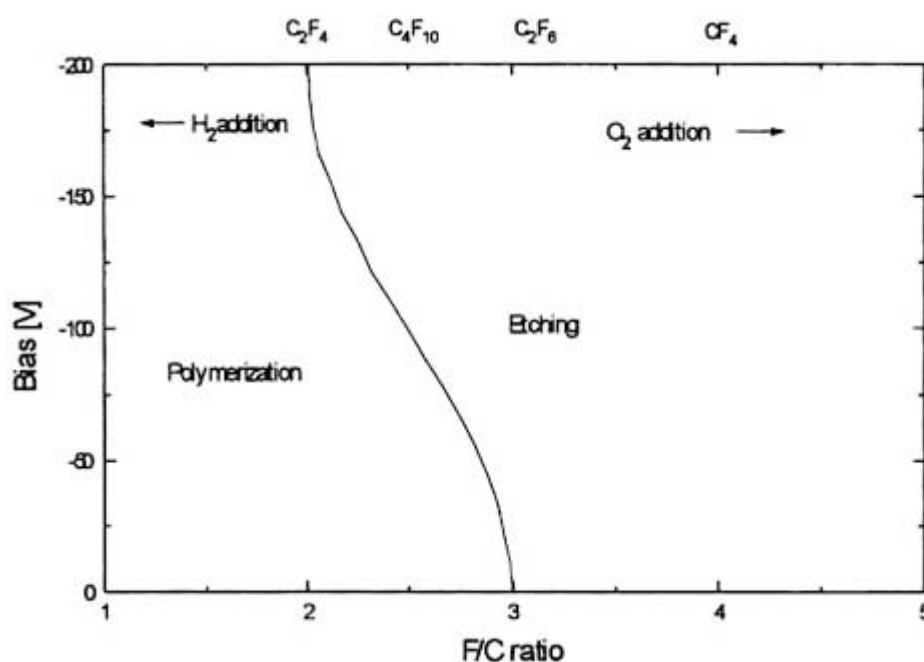
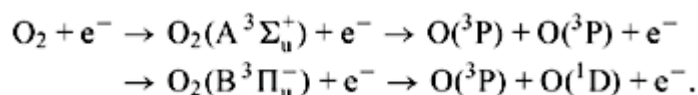


Figure C2.13.7. Change between polymerizing and etching conditions in a fluorocarbon plasma as determined by the fluorine-to-carbon ratio of chemically reactive species and the bias voltage applied to the substrate surface [36].

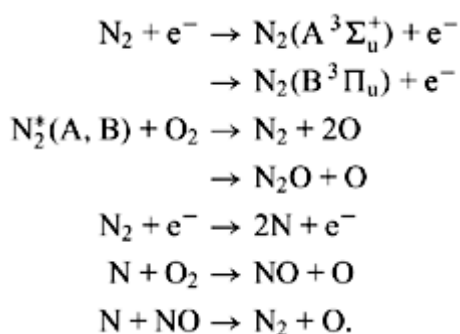
C2.13.5.2 PLASMA VOLUME PROCESSES

The chemical reactions in plasmas find applications in the generation or conversion of gaseous products, primarily *via* homogeneous reactions or in surface treatment and modification processes *via* heterogeneous reactions. A classical example for the production of a gaseous product is the ozone synthesis in dielectric barrier discharges. The electrons are the most important species for ozone formation [23]. A non-thermal plasma generated in a dielectric barrier discharge reactor at atmospheric pressure in pure oxygen causes a significant fraction of the oxygen molecules to be dissociated as the result of electron collisions.

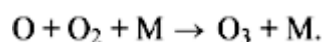
-17-



Nitrogen molecules, a major constituent of air, are excited by electron collisions and the excitation energy is transferred to the O₂ molecules, or the N₂ molecules may be dissociated and O atoms formed *via* the reactions



The ozone formation occurs in a three-body collision of O atoms with O₂ molecules:



The probability for three-body collisions increases with increasing pressure making the use of an atmospheric pressure plasma desirable. The above process is used worldwide for ozone production for water purification.

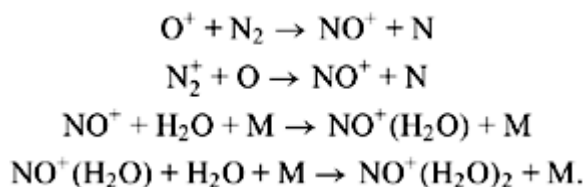
Pollution control such as the reduction of nitrogen oxides, halocarbons and hydrocarbons from flue gases [37] is another important field of plasma-assisted chemistry using non-thermal plasmas. The efficiency of plasma chemical reactions can be enhanced by introducing catalysts into the plasma [38, 39].

C2.13.5.3 PLASMA CHEMISTRY IN NATURE

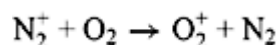
Naturally occurring plasma induced processes on Earth are not common. One example is the formation of nitrogen oxides in lightning. Another area of naturally occurring plasmas is the ionosphere [40]. The charge carriers are mainly produced by UV radiation from the Sun. The neutral gas composition in the ionosphere is determined by oxygen and nitrogen with smaller admixtures of water molecules, CO₂ and inert gases such as He and Ar. The most effective ionization process is the photoionization of O₂ and N₂ leading to the formation of atomic and molecular ions.

Mass spectrometric investigations of the ionosphere show an abundance of molecular ions such as NO⁺ and watercluster ions [41]. This is an indication of the result of ion-molecule reactions which change the chemical state of the ions in this plasma:

-18-



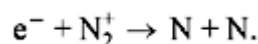
In addition, charge transfer processes such as



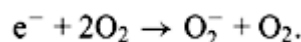
are efficient

Molecular ions have an important role in charge carrier losses in the ionosphere. The probability of electron-atom-

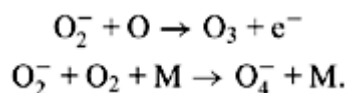
ion recombination processes is very small, because the frequency of the necessary three-body collisions (momentum conservation) at this low pressure is very low. Dissociative electron-molecular ion recombination processes are more effective. Here the two atoms that are formed ensure momentum conservation:



Negative ions [42] are the result of electron attachment processes such as



Loss processes for O_2^{-} ions are



The ozone formation in the atmosphere is induced by radiation and a result of three-body collisions of the oxygen atoms with O_2 molecules. This process requires a higher gas density and is, therefore, not efficient in the ionosphere.

C2.13.6 PLASMA MODELLING

C2.13.6.1 MICROSCOPIC KINETICS

Modelling plasma chemical systems is a complex task, because these systems are far from thermodynamical equilibrium. A complete model includes the external electric circuit, the various physical volume and surface reactions, the space charges and the internal electric fields, the electron kinetics, the homogeneous chemical reactions in the plasma volume as well as the heterogeneous reactions at the walls or electrodes. These reactions are initiated primarily by the electrons. In most cases, plasma chemical reactors work with a flowing gas so that the flow conditions, laminar or turbulent, must be taken into account. As discussed before, the electron gas is not in thermodynamic equilibrium

-19-

with the heavy particles. The velocity distribution function of the electrons is determined by the energy and momentum exchange in collisions with the heavy particles, in electron-electron collisions and by the energy gain in the electric field. It can be determined from the Boltzmann equation or by particle simulation methods (e.g., Monte Carlo, particle in cell). The concentrations of the various kinds of particles can be calculated by systems of balance equations including the generation and loss mechanisms of particles by electron impact, heavy-particle collisions and the interaction with surfaces. The reaction probabilities of molecules also depend on their electronic, vibrational and rotational states. Aside from the mathematical problems in solving a coupled system of balance equations, the collision cross sections, the natural life times of excited states, sticking and recombination coefficients at the wall etc must be known. Their knowledge, however, is limited in the case of most, if not all, realistic plasmas. Only a few very special cases can be solved, provided further simplifying assumptions are made [43, 44].

C2.13.6.2 MACROSCOPIC KINETICS

The concept of macroscopic kinetics avoids the difficulties of microscopic kinetics [46, 47]. This method allows a very compact description of different non-thermal plasma chemical reactors working with continuous gas flows or closed reactor systems. The state of the plasma chemical reaction is investigated, not in the active plasma zone, but

in the final state after all unstable reaction products have been converted into stable products. The investigation is restricted to the gross reaction; intermediate reaction steps are not considered. Chemical quasi-equilibrium states play an important role in this method. Quasi-equilibrium is reached in plasmas with high electron concentration (high-power plasma) after a comparatively short time, whereas this state is reached after a much longer time in a low-power plasma (low electron concentration). Experimentally determined reaction rates of the net reaction allow the computational treatment of the system. The specific energy $P/V\tau$ (where P is the power, V is the plasma volume and τ is the residence time of the gas in the plasma) is a particularly crucial parameter. Up until now, only simple systems have been successfully modelled.

C2.13.7 CONCLUSIONS

Non-thermal plasmas with their hot electron gas extend the realm of conventional chemistry in many interesting directions. New technical applications have emerged such as surface treatment of materials in plasma etching and thin-film deposition in the microelectronics industry, plasma chemical surface modifications to achieve a variety of desired properties (increased wettability, biocompatibility) and the deposition of various coatings to improve the hardness and the tribological and optical properties of materials. As successfully and widely used as plasma chemistry has been in surface processing applications, the applications of plasma chemical methods for producing gaseous products are limited. The ozone synthesis is a unique process in that category and has been well known for more than a century. The development of new processes for the synthesis or decomposition of gaseous compounds will require a broader as well as a more detailed knowledge of the processes in non-thermal plasma, particularly at higher gas densities, and will necessitate the study of the effect of catalysts on the plasma.

REFERENCES

- [1] Chen F F 1984 *Introduction to Plasma Physics and Controlled Fusion* (New York: Plenum)
- [2] Hollahan J R and Bell A T (ed) 1974 *Techniques and Applications of Plasma Chemistry* (New York: Wiley)
- [3] Boenig H V 1988 *Fundamentals of Plasma Chemistry and Technology* (Lancaster, PA: Technomic)
- [4] Polak L S and Lebedev Yu A (ed) 1998 *Plasma Chemistry* (Cambridge: Cambridge International Science)
- [5] McTaggart F K 1967 *Plasma Chemistry in Electrical Discharges* (Amsterdam: Elsevier)
- [6] Drost H 1978 *Plasmachemie* (Berlin: Akademie)
- [7] Rummel T 1951 *Hochspannungs-Entladungschemie und ihre industrielle Anwendung* (Munich: Oldenbourg)
- [8] Lieberman M A and Lichtenberg A J 1994 *Principles of Plasma Discharges and Materials Processing* (New York: Wiley)
- [9] Grill A 1994 *Cold Plasma in Materials Fabrication from Fundamentals to Applications* (New York: IEEE)
- [10] Allis W P 1956 Motions of ions and electrons *Handbuch der Physik* vol XXI, ed S Flügge (Berlin: Springer)
- [11] Shkarofsky I P, Johnston T W, and Bachynski M P 1966 *The Particle Kinetics of Plasmas* (Reading, MA: Addison-Wesley)
- [12] Winkler R 1972 Geschwindigkeitsverteilungsfunktion und Bilanzgrößen der Elektronen des anisothermen Argon-Plasmas bei Ionisierungsgraden von 10^{-9} bis 10^{-2} *Beitr. Plasma Phys.* **12** 193–211
- [13] Hasted J B 1964 *Physics of Atomic Collisions* (London: Butterworth)
- [14] McDaniel E W 1989 *Atomic Collisions Electron and Photon Projectiles* (New York: Wiley)

- [15] Christophorou L G (ed) 1984 *Electron–Molecule Interactions and Their Applications* (Orlando, FL: Academic)
- [16] Francis G 1956 The glow discharge at low pressure *Handbuch der Physik* vol XXII, ed S Flügge (Berlin: Springer)
- [17] Lister G G 1992 Low pressure gas discharge modelling *J. Phys. D. Appl. Phys.* **25** 1649–80
- [18] Ferreira C M and Moisan M (ed) 1993 Microwave discharges, fundamentals and applications *NATO ASI Series, Series B: Physics* vol 302 (New York: Plenum)
- [19] Hopwood J 1992 Review of inductively coupled plasmas for plasma processing *Plasma Sources Sci. Technol.* **1** 109–16
- [20] Chen F F 1995 Helicon plasma sources *High Density Plasma Sources* ed O Popov (Park Ridge, MD: Noyes)
- [21] Ohl A 1998 Fundamentals and limitations of large area planar microwave discharges using slotted waveguides *J. Physique IV* **8** Pr7 83–98
- [22] Korzec D, Werner F, Winter R and Engemann J 1996 Scaling of microwave slot antenna (SLAN): a concept for efficient plasma generation *Plasma Sources Sci. Technol.* **5** 216–34
- [23] Kogelschatz U, Eliasson B and Egli W 1995 Dielectric barrier discharges. Principles and applications *J. Physique IV* **7** C4 47–66

-21-

- [24] Siemens W 1857 über die elektrostatische Induction und die Verzögerung des Stromes in Flaschendrähnen *Poggendorfs Ann. Phys. Chem.* **102** 66–122
- [25] Sugawara M 1998 *Plasma Etching: Fundamentals and Applications* (Oxford: Oxford University Press)
- [26] Schaepekens M, Standaert T E F M, Rueger N R, Sebel P G M, Oehrlein G S and Cook J M 1999 Study of the SiO₂ - to - Si₃ N₄ etch selectivity mechanism in inductively coupled fluorocarbon plasmas and a comparison with the SiO₂ - to - Si mechanism *J. Vac. Sci. Technol A* **17** 26–37
- [27] Janes J 1993 Mass selected ion angular impact energy distributions at the powered electrode in CF₄ reactive-ion etching *J. Appl. Phys.* **74** 659–67
- [28] Bruno G, Capezuttuto P and Losurdo M 1995 On the use of the plasma in III–V semiconductor processing *Phenomena in Ionized Gases (ICPIG Hoboken, NJ, 1995 (AIP Conference Proceedings vol 22))* ed K H Becker, W E Carr and E E Kunhardt (Woodbury, NY: American Institute of Physics) pp 146–55
- [29] Ratner B D, Chilkoti A and Lopez G P 1990 Plasma deposition and treatment for biomaterial applications *Plasma Deposition, Treatment and Etching of Polymers* ed R d'Agostino (Boston, MA: Academic) pp 463–516
- [30] Yasuda H 1985 *Plasma Polymerization* (Orlando, FL: Academic)
- [31] Konuma M 1992 *Film Deposition by Plasma Techniques* (Berlin: Springer)
- [32] Perrin J, Leroy O and Bordage M C 1996 Cross-sections, rate constants and transport coefficients in silane chemistry *Contr. Plasma Phys* **36** 3–49
- [33] Charles C, Garcia P, Grolleau B and Turban G 1992 Mass spectrometric study of tetraethoxysilane and tetraethoxysilane oxygen plasmas in a diode type radio-frequency reactor *J. Vac. Sci. Technol A* **10** 1407–13
- [34] Poll H-U, Arzt M and Wickleder K-H 1976 Reaction kinetics in the polymerization of thin films on the electrodes of a glow-discharge gap *Eur. Polym. J* **12** 505–12
- [35] d'Agostino R, Cramarossa F and Fracassi F 1990 Plasma polymerization of fluorocarbons *Plasma Deposition, Treatment and Etching of Polymers* ed R d'Agostino (Boston, MA: Academic) pp 95–162

- [36] Coburn J W and Winters H F J 1979 Plasma etching, a discussion of mechanisms *J. Vac. Sci. Technol* **16** 391–403
- [37] Penetrante B M, Bardsley J N and Hsiao M C 1997 Kinetic analysis of non-thermal plasmas used for pollution control *Japan. J. Appl. Phys.* **36** 5007-17
- [38] Venugopalan M and Veprek S 1983 Kinetics and catalysis in plasma chemistry *Top. Curr. Chem* **107** 1–58
- [39] Kizling M B and Järas S G 1996 A review of plasma techniques in catalyst preparation and catalytic reactions *Appl. Catalysis A* **147** 1–21
- [40] Kelley M C 1989 *The Earth's Ionosphere: Plasma Physics and Electrodynamics* (San Diego, CA: Academic)
- [41] Bauer S J 1973 *Physics of Planetary Ionospheres* (Berlin: Springer)
- [42] Wisenberg J and Kockarts G 1980 Negative ion chemistry in the terrestrial D region and signal flow graph theory *J. Geophys. Res* **85** 4642–52
-

-22-

- [43] Storch D G and Kushner M J 1993 Destruction mechanisms for formaldehyde in atmospheric pressure low temperature plasmas *J. Appl. Phys.* **73** 51–5
- [44] Gortschakov G, Loffhagen D and Winkler R 1998 The homogeneity of a stabilized discharge-pumped XeCl* laser *Appl. Phys B* **66** 313–22
- [45] Gorse C and Capitelli M 1996 Non-equilibrium vibrational, electronic and dissociation kinetics in molecular plasmas and their coupling with the electron energy distribution function *NATO ASI Series C* **482** 437–49
- [46] Eremin E N 1968 *Elementy Gazovoi Electrochemii* (Moskva: Isdat. Moskovskovo Universiteta)
- [47] Rutscher A and Wagner H - E 1993 Chemical quasi-equilibria: a new concept in the description of reactive plasmas *Plasma Sources Sci. Technol* **2** 279–88
-

-1-

C2.14 Biophysical chemistry

J J Ramsden

C2.14.1 INTRODUCTION

Biophysical chemistry may be defined as the application of physical chemistry to biological systems. The underlying question (e.g.[1]) is whether the laws of physics and chemistry suffice to understand biology. This question, whose origins go back a long way and which encompasses notions of vitalism and so on, has still not been definitively answered. Nowadays it is formulated somewhat differently, typically as ‘What emergent properties are needed to characterize biological systems?’ Half a century ago the prevailing view was encapsulated

in a rather well-known statement of P A M Dirac: ‘The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble.’ Biology being even more complicated than chemistry, the difficulty was expected to be correspondingly greater. Nevertheless, as P W Anderson [2] has pointed out, biology is not applied chemistry, any more than chemistry is applied physics. Emergent properties are expected to arise through increases in scale and complexity, and there is no need to endow them with mystical attributes: one hopes to formulate them in mathematical terms just like the descriptions of simpler systems.

Biological systems are bewilderingly complex; the task of biophysical chemistry is to render this complexity intelligible. It is, challengingly, complexity of the most difficult kind, falling between the tractable extremes of elements few enough to be enumerated and analysed exactly, and huge crowds of similar or identical elements whose statistical properties are sufficient for understanding the behaviour of the whole. The molecules of biology are astonishingly diverse and even apparently very minor changes—the single amino acid substitution in haemoglobin causing sickle cell anaemia is a paradigmatic example—has consequences at the levels of the structure of the molecules, the shape of the erythrocytes containing haemoglobin, the health of the individual human being, and the biology of the population.

Most biochemists and molecular biologists make use of chemistry and physics in their investigations. Every time they run a ‘gel’—a kind of chromatography used to separate protein (or DNA) mixtures, and appearing in almost every molecular biological lecture or publication—certain assumptions are made relating the position of a protein on the gel to its molecular mass, and a few ‘standard’ proteins are used to calibrate the positions. Is the procedure reliable? Certainly, many factors other than mass intervene. For example, the protein MARCKS, $M_r \approx 30\,000$, appears near proteins almost three times heavier [3]. Why? Biophysical chemistry can provide the answer, and a correct apprehension of the reasons enhances the domain of application of the technique. Another example is the use of biosensors to investigate protein–protein association. Many investigators first coat the sensing platform with a thick dextran hydrogel, and covalently attach one of the pair of proteins whose association is to be investigated to the matrix. The hydrodynamics of the aqueous phase containing the binding partner are thereby greatly distorted compared with the free solution, and the measured binding rates tend to reflect highly retarded transport rather than affinity [1]. Even greater caution is required when using gene chips, two dimensional arrays of $\sim 10^4$ to 10^5 single stranded DNA fragments whose nucleic acid constitution and coordinates are, in principle, known. If it is desired to ascertain whether certain mutations are present in the gene of a patient, his or her DNA is isolated, separated into single-stranded molecules, labelled with a fluorescent marker, and brought into contact with a gene chip including sequences complementary to those sought. If they are present in the isolate, they will be bound specifically to the chip,

-2-

and after dissociating nonspecifically bound material, the remaining fluorescence is scanned and the corresponding sequences identified from their spatial coordinates. If it is desired to ascertain which genes in a given cell are expressed at a given moment, the messenger RNA (selected copies of the gene corresponding to those amino acid sequences, proteins, which will subsequently be synthesized—see also [section C2.14.8](#) is isolated from the cell, labelled and exposed to the gene chip. The pattern of residual fluorescence should bear some relation to the messenger RNA population in the cell. Potential sources of misinterpretation include: variable degrees of labelling; complementary binding hindered by the presence of the label, or by the immobilization of the DNA strand to the chip surface; insufficient time allowed for specific association, or for dissociation of nonspecifically bound material; binding of the same sequence to multiple sites, or of different sequences to the same site. Clearly the specificity of binding to a given site increases with increasing length of the immobilized DNA sequence, but then so does the likelihood of mistakes in the synthesis of the sequences on the chip. The pattern of fluorescent spots—their positions and intensities—may depend as much on biophysico-chemical details of the binding processes taking place at the molecular level as on the actual nucleic acid sequences themselves. Since the information potentially available from the pattern is huge, it must necessarily be analysed largely automatically, by algorithms which fix certain choices regarding the interpretation of the data, and hence it is important for their alternatives to be thoroughly investigated.

Classical biophysical chemistry has concentrated heavily on the elucidation of the structures of biomolecules and

biomolecular assemblies. Since there are already several excellent texts and reviews dealing with techniques such as x-ray diffraction, electron microscopy, scanning probe microscopies and so on, and since they are all dealt with elsewhere in this encyclopaedia, it is superfluous to discuss them here. A few structural methods whose domain of application has been so exclusively concerned with biological material that they have not warranted individual entries elsewhere will be discussed in [section C2.14.2](#).

Apart from the sheer complexity of the static structures of biomolecules, they are also rather labile. On the one hand this means that especial consideration must be given to the fact (for example in electron microscopy) that samples have to be dried, possibly stained, and then measured in high vacuum, which may introduce artifacts into the observed images [5]. On the other, apart from the vexing question of whether a protein in a crystal has the same structure as one freely diffusing in solution, the static structure resulting from an x-ray diffraction experiment gives few clues to the molecular motions on which operation of an enzyme depends [6].

Biology has been defined as the organization of parts and processes, their reciprocal interactions and directedness. Given instantaneous and perfect mixing, could not metabolic duties be accomplished via specific intermolecular interactions, at least on the level of an individual cell? The description and understanding of specificity is another great area of classical biophysical chemistry. Compared with the rather high level of development of structural elucidation, the field of biomolecular interactions has progressed less; they are often described using oversimplified models, e.g. exponential decay for the dissociation of two molecules, or of a molecule from a surface, despite the fact that empirically this law has been found to be an exception to the more general occurrence of time-dependent dissociation rate 'constants' ([section C2.14.4.3](#)). Specificity is dealt with in [section C2.14.6](#); one should keep in mind that mixing is neither instantaneous nor perfect, and the relative importance of directed transport over random diffusion within a cell is a major current preoccupation in biophysical chemistry.

Biological reactions *in vivo* rarely operate under conditions even remotely approaching those of reversibility. For a living organism, the *rate* of a process is usually more important than the attainment of equilibrium, and large driving

-3-

forces are employed to achieve a rapid rate, leading to profligate waste of free energy from the classical chemical thermodynamical viewpoint. The law of mass action has universal validity, but emphasis on the 'biochemical standard state' (in which all components are fully hydrated and dissolved at a concentration of 1 M in water, which itself has a concentration of 55.6 M), far removed from reality, is to the detriment of discussion of the actual concentrations of the molecules participating in biochemical pathways. Moreover, since biology is a realm of almost universally broken ergodicity, the premises of statistical mechanics approaches need to be carefully scrutinized; more emphasis on pathways would be appropriate.

The degree of coarse graining appropriate to solve a biological problem is an ever-present preoccupation [7]. When attempting to predict the three-dimensional structure of a protein, is it enough to consider the amino acids as structureless blobs, or do their individual atoms need to be included? The structure of large, elongated extracellular matrix proteins such as fibronectin can be successfully modelled even more coarsely as a string of beads [8], each bead comprising tens of amino acids. Coarse graining seems to be carried too far, however, when human brain activity is investigated *in vivo* using nuclear magnetic resonance (NMR) imaging. 'High spatial resolution' in the context of this technique means averaging over a volume of $\sim 10\text{mm}^3$, containing hundreds of thousands of neurons, each of which may be connected to thousands of other neurons elsewhere. It might be recalled that currently we are not even capable of giving a comprehensive description of the brain of the leech, with only a couple of dozen ganglions, each containing about 400 nerve cells. Moreover, although the NMR imaging technique is considered to be noninvasive, one may legitimately enquire whether a human being will, indeed can, behave naturally and normally when constrained to lie still in a narrow tube with his head tightly fixed.

A further problem confronting the biophysical chemist is that biological systems are usually highly compartmentalized, and the actual numbers of each active species taking part in a reaction within a given compartment may be very small. Hence there can be no question of passing to the thermodynamic limit. Fluctuations are expected to play a dominant, sometimes enabling, sometimes destructive role in biology, and the

way that living beings have evolved sufficient autonomy to deal with fluctuations is itself a fascinating area.

Compartmentalization not only leads to a greatly amplified role for the incoherent fluctuations of groups of molecules, but inevitably implies the possibility of interaction between those molecules and the compartment walls (often a bilayer lipid membrane), i.e. interfacial phenomena, added to which is the fact that the properties of the solvent (usually water) may be modified in the vicinity of the wall. In experiments now considered as classic, Kempner and Miller [9, 10] centrifuged *intact* cells and showed that virtually no enzymes were present in the soluble phase; the term cytosol refers to the solution, containing a wide variety of enzymes and other macromolecules, obtained from cell fractionation *after disruption*. Despite these and other observations, experiments to elucidate *in vitro* reactions between biomolecules under homogeneous conditions greatly preponderate over those in which the equivalent reaction is studied heterogeneously. Perhaps the recent introduction of some excellent new techniques for investigating biomolecular interactions at the solid–liquid interface [11] will start to shift this balance, complementing and ultimately replacing molecular biological techniques such as the two-hybrid system.

Once pairwise biomolecular interactions have been correctly characterized, the next step is to understand how they mesh together in a network. During the past couple of decades, molecular biology has been rather successful at identifying individual genes and their protein products, and obtaining valuable mechanistic insights through site-directed mutagenesis. A cell is not a mosaic of individual reactions, however, any more than a protein is a (linear) mosaic of amino acids. To establish that process X is catalysed by Xase, which may then be cloned, sequenced and

-4-

expressed, is only the beginning of understanding how the reactions are linked together in an elaborate network of control in which some gene products regulate the activity of others, and which is not merely hierarchical, but heterarchical. Moreover, it is essential to recognize that the intramolecular bonds which determine the structures of biological molecules are usually comparable in strength to the intermolecular interactions as well as interactions with the solvent (water). Hence the structures will in many cases be modified upon association, changing the affinity of the complex for a third and subsequent molecules, and hence a comprehensive catalogue of all the pairwise interactions in a cell will be a very incomplete representation of the network.

In order to meet the challenge of understanding how hundreds or even thousands of partly interrelated reactions fit together to form a coherent functioning whole, *systems theory* was developed in the 1960s. It was supposed to lead to general principles governing complex organizations such as the living cell. A considerable *œuvre* was achieved (e.g. [12]), but it must be conceded that it has had rather little influence on biological research. Perhaps it was too ambitious for the level of phenomenological knowledge extant when it was first developed, and since therefore no immediate application of its predictive power was possible, it fell into neglect. Moreover the theory quickly becomes intractable when nonlinear systems involving more than two elements are considered. After three decades of immensely diligent data-gathering, the situation may now be more propitious for the application of systems theory to biology, but to begin with a more modest programme may be in order, such as an investigation of the scaling laws of biological reactions, which are far more strongly nonlinear than those encountered in most nonbiological systems.

Acknowledgement of the existence of biological *systems* refocuses attention on the old question of whether new physico-chemical concepts are required in order to understand their working. Is regulation and control theory, as developed mainly in departments of engineering and electronics, adequate to describe biological systems, or is their complexity—multilevelled and on multiple time scales—sufficiently great to make them qualitatively different in some way? This question is still open. The old dream was to predict protein structure (and possibly even some dynamical properties) from gene sequence, and then function from structure. The first part looks close to being achieved, although many proteins are post-translationally modified (glycosylated, lipidated, etc) and these modifications, carried out by other proteins, are crucial for regulating the specific molecular interactions which play such an essential role in metabolism. But the notion of the ‘function’ of a particular molecule embodies all its interactions with others, and it is not clear that studying the genome alone will enable one to understand them. Put even more starkly, could one predict the existence of a central nervous system merely from studying the genome? The reductionist view is to seek a molecular explanation, but a list of all the molecules in a cell, even if their spatially varying concentrations are given, is not sufficient to understand how the whole cell works, let alone

multicellular organisms.

It is of course futile to attempt to cover all of biophysical chemistry within one short chapter. The emphasis will be on necessary concepts, especially where these diverge from the mainstream of physical chemistry; inevitably there will be gaps. Material which is well known and readily found in standard texts (e.g. [13]), or in other chapters of this encyclopaedia, will be dealt with cursorily, if at all, but new material, and topics less well known than they should be, are accorded a more detailed treatment. Given the central role of macromolecular interactions in biological systems, they are covered extensively ([section C2.14.6](#) and [section C2.14.7](#) and part of [section C2.14.3](#)), except for protein folding ([section C2.14.2](#)) since it is also the topic of chapter 2.5. Biological membranes and associated topics such as the conduction of the nervous impulse [14, 15] will not be discussed. Another omission is the interaction of light with biological molecules, because it does not seem that in essence they diverge significantly from photochemical processes in general, without excluding the possibility that photons may be involved in intercellular signalling. For the same reason, intra- and intermolecular electron transfer phenomena have also been omitted.

-5-

To some extent, the division into sections is arbitrary: for example, aspects of biological structure prediction ([section C2.14.2.2](#)) hinge on kinetic considerations ([section C2.14.3.5](#) and [section C2.14.4](#)). Most obviously in a developing organism, but actually throughout life, morphogenesis and regulation depend on proteins which are being synthesized at definite rates. The essence of enzyme function lies in the dynamical aspect of structure (conformational relaxation). Indeed, if one had to name a single principle characterizing biophysical chemistry and how it differs from the rest of physical chemistry, it would be the emphasis on kinetics, and an alternative definition of biophysical chemistry might be the science of kinetics and pattern. The reason for including pattern, i.e. spatial inhomogeneity, will become clear in [section C2.14.4.4](#).

C2.14.2 BIOLOGICAL STRUCTURE

Ever since it became apparent that many proteins can refold from a denatured random coil into a fully functional enzyme, implying that the amino acid sequence alone encodes the necessary information, this field has been divided into two: the experimental determination and the prediction of structure from sequence. It may be that one day it will be much easier and quicker to compute structures from sequences ([section C2.14.2.2](#)), but at present the experimental determination is more reliable. The choice is between classical methods capable of yielding three-dimensional coordinates of many or all the atoms in the molecule, which is however not under *in vivo* conditions; and a panoply of diverse methods capable of lower resolution, or of elucidating only one particular aspect of structure, but under physiological conditions. Nowadays it is realized that the determination of structure is a problem of inference using data from diverse sources which must be combined, and no one method is universally applicable. It is furthermore as well to remember that biological structure is set in a dynamical context, and that the characteristic patterns of biopolymer structural fluctuations are probably essential to understanding functional mechanisms; Ageno [16] gives an excellent example of the consequences of rapid transitions between different conformational states of DNA.

C2.14.2.1 STRUCTURE DETERMINATION

The classical methods of determining three-dimensional native structures of biopolymers, i.e. the spatial coordinates of all their atoms, are well documented by Cantor and Schimmel [13] and will not be reiterated in any detail here. The most comprehensively useful technique is x-ray diffraction (see chapter B1.9), for which the material must be prepared in crystalline form. The attainable resolution is strongly dependent on the size and quality of the crystal. Crystallization is still a highly empirical art. Intense x-ray sources (synchrotron radiation) enable smaller crystals to be used, although the rate of radiation damage is correspondingly faster. Until recently, membrane proteins, which may comprise about a third of the expressed protein repertoire of a typical cell, could not be crystallized in their native environment and their structure determination using x-rays was problematical; the use of cubic phase lipids offers a promising new route [17].

The main problem with x-ray (and neutron) diffraction is that the information it is made to yield is essentially

static. X-ray diffraction generates data on the millisecond time scale, whereas amino acid residues in a protein can rotate about 10^{11} times per second, and even proton exchange takes place on a submicrosecond scale. Hence an enormous amount of information is averaged out. Some attempts to quantify molecular motility have been made by analysing the temperature dependence of the broadening of the Debye–Waller factors [18], but large infrequent motions do not perceptibly contribute to the broadening. Since it is precisely these motions which may constitute the key part of enzyme action, their invisibility vitiates the structure → function path of inference. An equally serious problem is that many, or possibly most, proteins can exist in several stable conformational states and therefore possess the ability to

-6-

remember. This polymorphism, proper characterization of which may be essential to understand the functional mechanism of a protein, usually remains undetected by the classical methods, for the simple reason that in the final stage of numerically refining the atomic coordinates derived from the diffraction data a computer is programmed to find a single optimal structure, not the optimal mix of (unknown) structures, which is probably anyway indeterminate from the available information.

A further problem is the influence of the rather unusual—from the physiological viewpoint—salt conditions necessary for crystallization. It should not be presumed that proteins embedded in a crystal are in their most common native structure. It is well known that, with the exception of sodium or potassium chloride, which are not very useful for inducing crystallization, salts change key protein parameters such as the melting temperature [19].

A weakness with x-ray, but not neutron diffraction (although the latter is experimentally more difficult, mainly because neutrons are far harder to produce and focus than x-rays) is that the hydrogen atoms are invisible, and their positions must be inferred during numerical structure refinement. Given the primordial role of hydrogen bonding in determining biological structure, this omission is unfortunate.

The spatial arrangement of atoms in two-dimensional protein arrays can be determined using high-resolution transmission electron microscopy [20]. The measurements have to be carried out in high vacuum, but since the method is used above all for investigating membrane proteins, it may be supposed that the presence of the lipid bilayer ensures that the protein remains essentially in its native configuration.

Nuclear magnetic resonance spectroscopy (chapters B1.11, B1.12, B1.13 and B1.14) can provide cross-relaxation rates between two proton spins, from which a set of short range (up to ~ 5 Å) distance constraints can be generated [21], from which in turn a three dimensional conformation can be computed [22, 23]. The number of constraints is much larger than the number of degrees of freedom in the protein, which partly compensates for the limited accuracy of the constraints (uncertainties can be as much as ~ 1 Å, but renders the computational problem exceedingly difficult, comparable to the protein folding problem (section 2.5). The upper limit of protein molecular weight is about 200 000. Since the intrinsic time scale of NMR is about 1 ms, as in the case of x-ray diffraction, many conformations are averaged out. An attraction of the method is that the protein is not constrained in a crystal, but is presumably present in its native structure, although in order to measure signals of adequate intensity, rather high protein concentrations have to be used and there is a risk of aggregating the protein.

Apart from these mainstream methods enabling one to gain a comprehensive and detailed structural picture of proteins, which may or may not be in their native state, there is a wide variety of other methods capable of yielding detailed information on one particular structural aspect, or comprehensive but lower resolution information while keeping the protein in its native environment. One of the earliest of such methods, which has recently undergone a notable renaissance, is analytical ultracentrifugation [24], which can yield information on molecular mass and hence subunit composition and their association/dissociation equilibria (via sedimentation equilibrium experiments), and on molecular shape (via sedimentation velocity experiments), albeit only at solution concentrations of at least a few tenths of a gram per litre.

The new scanning probe microscopies (chapter B1.19) have been used enthusiastically by biologists almost since their invention, because biomolecules can be investigated in aqueous, physiological *milieux* at room temperature. Early hopes of using atomic force microscopy to sequence DNA, or scanning tunnelling microscopy to characterize

individual ion channels, have not been realized, however. While a degree of resolution close to that achievable with good electron microscopy has been reported for fairly rigid protein arrays, notably bacteriorhodopsin, biological samples are on the whole too sticky and labile to be successfully imaged at submolecular resolution. Even the smallest achievable probe–sample interaction forces deform the sample, assuming that it does not slide about the surface, and the use of specially fine tips (radius < 1 nm) is vitiated by the rapid accumulation of biological debris at the tip. A brighter future lies in the more easily attainable realm of imaging the two-dimensional arrangement of proteins at a surface, from which the radial distribution function, a rich source of information on short and long range intermolecular interactions, can be obtained. For this application, it is sufficient to image each protein as a featureless blob.

Circular dichroism has been a useful servant to the biophysical chemist since it allows the non-invasive determination of secondary structure (α -helices and β -sheets) in dissolved biopolymers. Due to the dissymmetry of these structures (containing chiral centres) they are biaxial and show circular birefringence. Circular dichroism is the Kramers–Kronig transformation of the resulting optical rotatory dispersion. The spectral window useful for distinguishing between α -helices and so on lies in the region 200–250 nm and hence is masked by certain salts. The method as usually applied is only semi-quantitative, since the measured optical rotations also depend on the exact amino acid sequence.

Another technique used for structural inference is dielectric dispersion in the frequency [25] or time [26] domains. The biopolymer under investigation must have a permanent dipole moment μ_0 . It is first dissolved in a dielectrically inert solvent, e.g. octanol, which may be considered to bear some resemblance to a biological lipid membrane, and then the complex impedance $\hat{\epsilon} = \epsilon' + i\epsilon''$ is measured over a range of frequencies f typically from a few kHz up to several tens of MHz. The dielectric dispersion arises through the rotational relaxation of the molecules. One or more Debye relaxation functions:

$$\epsilon' = \epsilon_{\infty} + \Delta\epsilon_0 / (1 + [f/f_0]^2) \quad (\text{C2.14.1})$$

and

$$\epsilon'' = \Delta\epsilon_0 [f/f_0] / (1 + [f/f_0]^2) \quad (\text{C2.14.2})$$

where $\Delta\epsilon_0$ is the dielectric relaxation amplitude (related to the square of the permanent dipole moment μ_0 [25]), are fitted to the data in order to determine the relaxation frequency f_0 , which is related to the rotational friction coefficient and hence to the shape of the molecule. Water contributes significantly to the impedance spectrum and its contribution must be carefully assessed and eliminated.

Careful measurement of the kinetics of association of a molecule with a surface can also yield structural information at this level of resolution [27], and lateral clustering and crystallization can also be deduced. This is described in more detail in [section C2.14.7.2](#).

Another method applicable to interfaces is the determination of the partial molecular area \bar{a} of a biopolymer partitioning into a lipid monolayer at the water–air interface using the Langmuir trough [28]. The first step is to record a series of pressure π –area (A) isotherms with different amounts n of an amphiphilic biopolymer spread at the interface.

A particular surface pressure is then chosen, and n plotted against the values of A at that pressure. From the

conservation of mass, we must have

$$n = \Gamma A + c_b V \quad (\text{C2.14.3})$$

where Γ is the surface concentration and c_b the concentration in the subphase bulk of volume V . Such plots yield straight lines from which Γ and c_b can be determined using equation (C2.14.3). A plot of Γ (from which \bar{a} can be deduced) versus π can then be constructed; typically such a plot has several features which can be assigned to conformational transitions.

A little known structural method for investigating protein motions is to measure the Rayleigh scattering of Mössbauer radiation (RSMR) [29]. A Mössbauer source moving with velocity $\pm v$ irradiates the sample, and both elastically and inelastically scattered radiation are measured as a function of the scattering angle θ . The energy spectrum (resonant peak position and linewidth) and the fraction of elastic scattering embody information on the dynamics of the protein. These measurements are especially valuable for characterizing the main types of movements in a protein, namely:

1. solid state motions (amplitudes $A_1 \sim 0.1\text{--}0.2 \text{ \AA}$ and correlation times $\tau_0 \sim 10^{-13} - 10^{-12} \text{ s}$);
2. large scale individual motions of small groups of atoms (amplitudes A_2 up to 0.5 \AA and correlation times $\tau_2 \sim 10^{-11} - 10^{-9} \text{ s}$);
3. complex cooperative motions of larger domains (amplitudes A_3 up to 1 \AA and correlation times $\tau_3 \sim 10^{-8} - 10^{-7} \text{ s}$);

C2.14.2.2 THE PREDICTION OF STRUCTURE

A native protein is folded from a linear chain comprising anything from about 30 to 2000 amino acids, mainly chosen (with differing probabilities) from a set of twenty or so different ones. The prediction of the stable, three-dimensional structure (or structures) of a biopolymer is an horrendously difficult problem. It is not even known if the stable structure corresponds to the global energy minimum; even if it does, the calculation of this minimum, of a rough energy surface with countless local minima, is an extremely difficult optimization problem. In any case, realistic estimates of the time needed to search through all possible conformations would exceed the lifetime of the universe, whereas proteins are known to fold spontaneously within typically a few seconds. This is sometimes referred to as the ‘Levinthal paradox’.

An excellent account of the statistical physics of polymer chains, with some consideration of biological macromolecules, is given by Lifschitz *et al* [30]. Much recent work on the protein folding problem seems to have been inspired by the concept of frustration in spin glasses [31]—whether the analogy is deep or superficial remains to be seen—and hence it has been proposed that proteins fold on a rough energy landscape [32], which implies certain generic features of the folding pathway, but whether these are sufficient to solve the folding problem is unclear. Perhaps the most delicate issue is the relative importance of local versus nonlocal (i.e. between residues distant from each other along the polypeptide chain) interactions [33]. From a practical viewpoint this approach has not led to a

successful algorithm for predicting structure. It is a mark of the lack of progress that much attention is now being devoted to expert algorithms based on the (now quite large) data bases of known structures and the sequences specifying them; they merely compare a new sequence of unknown structure with extant sequences whose structures are known. Of course a causal basis is lacking and no fundamental insight into the underlying mechanisms governing folding is gained.

A different approach is based on the realization that equilibrium thermodynamics cannot dictate a sequence of

events under time constraints unless the contributions to the thermodynamic potential themselves represent kinetic parameters [34]. Life as a whole is well characterized as a thermodynamic system operating under kinetic constraints; individual life is essentially transient, and metastable structures are nearly always good enough. This naturally leads to viewing expedience rather than equilibrium as the driving principle of folding, and the preeminence of the Lagrangian $\mathcal{L} (=T - V$ for conservative systems, where T and V are respectively the kinetic and potential energies), rather than the Hamiltonian $\mathcal{H} = T + V$. Minimization of the action (the integral of \mathcal{L}) is an inerrant principle for finding the correct solution of a dynamical problem [35]; the difficulty resides in the fact that there is no general recipe for constructing \mathcal{L} .

A solution leading to a successful algorithm was recently found for the folding of ribonucleic acid (RNA) [36]. Natural RNA polymers (figure C2.14.1) are mainly made up from four different ‘bases’, A, C, G and U. As with DNA, multiple hydrogen bonding favours the formation of G–C and A–U pairs [16, 37, 38] which leads to the appearance of certain characteristic structures. Loop closure is considered to be the most important folding event.



Figure C2.14.1. Diagram of a fragment of a folded RNA polymer, the Q β replicase MDV-1 [176]. Note the various structural features: stems closed with a loop (‘hairpins’), bows, and single strands.

V (the potential) is identified with the enthalpy, i.e. the number n of base pairings (contacts), and T corresponds to the entropy. At each stage in the folding process, as many as possible new favourable intramolecular interactions are formed, while minimizing the loss of conformational freedom (the principle of sequential minimization of entropy loss, SMEL). The entropy loss associated with loop closure is ΔS_{loop} (and the rate of loop closure $\sim \exp(\Delta S_{\text{loop}})$); the function to be minimized is $\exp(-\Delta S_{\text{loop}}/R)/n$ [36]. A quantitative expression for ΔS_{loop} can be found by noting that the N monomers in an unstrained loop ($N \geq 4$) have essentially two possible conformations, pointing either inwards or outwards. For loops smaller than a critical size N_0 , the inward ones are in an apolar environment, since the enclosed water no longer has bulk properties, and the outward ones are in polar bulk water; hence the electrostatic charges on

the ionized phosphate moieties of the bases will tend to point outwards. For $N < N_0$, $\Delta S_{\text{loop}} = -RN \ln 2$, and for $N > N_0$, the Jacobson–Stockmayer approximation based on excluded volume yields $\Delta S_{\text{loop}} \sim R \ln N$.

In the case of proteins, it is advantageous to make use of the fact that not all combinations of the two dihedral angles (the only degrees of freedom of the polypeptide chain) specifying the orientations of the $C_\alpha - C$ and $C_\alpha - N$ bonds are permitted [39]. Essentially there are just three basins of attraction, corresponding to left- and right-handed α -helices (compact conformations) and the β -sheet (extended conformation). Consensus sequences (runs of amino acid residues whose local conformations fall in the same basin) result in persistent, ultimately global structure being built up (the folding problem can be viewed as fixing the relation between local and global structures). As with RNA, the barriers are entropic, determined solely by loop closure (except where contacts have to be disassembled), and the SMEL principle applies. This approach has been successfully used to fold bovine pancreatic trypsin inhibitor [40].

C2.14.2.3 BIOMOLECULAR ASSEMBLIES

The modern era of biochemistry and molecular biology has been shaped not least by the isolation and characterization of individual molecules. Recently, however, more and more polyfunctional macromolecular complexes are being discovered, including nonrandomly codistributed membrane-bound proteins [41]. These are made up of several individual proteins, which can assemble spontaneously, possibly in the presence of a lipid membrane or an element of the cytoskeleton [42] which are themselves supramolecular complexes. Some of these complexes, e.g. snail haemocyanin [43], are merely assembled from a very large number of identical subunits; viruses are much larger and more elaborate; and we are still some way from understanding the processes controlling the assembly of the wonderfully intricate and beautiful structures responsible for the iridescent colours of butterflies and moths [44].

Specific intramolecular interactions (section C2.14.6) can be expected to play a role in the spontaneous assembly of the constituent elements, so that simply mixing the constituents will result in a correctly assembled structure provided the environment is right. One wonders whether stigmergic building algorithms [45] are involved, in which individual elements communicate only through the local environment. If one observes a nest of ants after a disturbance, the impression is one of haphazard movement as the ants drag exposed eggs hither and thither, but within a very short time they have all been moved to safety, without any hierarchical command system in operation.

C2.14.3 BIOLOGICAL EQUILIBRIUM

Ergodicity is generally broken in biological systems, and hence the standard notion of equilibrium is not very useful for solving biological problems; as already mentioned in section C2.14.2.2, most living systems operate under time constraints. Processes are therefore not infinitesimally slow and perfectly reversible: an organism is willing to sacrifice free energy in order to ensure that events take place rapidly and are consequently irreversible.

C2.14.3.1 TRANSFER AND STORAGE OF CHEMICAL POTENTIAL

Many biochemical reactions are involved in converting and storing energy, and the primary consideration is the chemical potential at which the product is recovered, rather than the yield. Consider the simple reaction



-11-

The rate of storage of chemical potential (in other words, the power P carried by the chemical reaction) is $j\mu_B$, where $j = db/dt = J_f - J_b$, the net flux per unit volume (here, as elsewhere, lower case letters denote concentrations, and subscripts f and b refer to the forward and backward reactions). If A and B are at equilibrium, $j = 0$; if the forward reaction proceeds at a finite rate, then $\mu_B < \mu_A$, but if j goes to its maximum value, $j = J_f$ (i.e. $J_b = 0$), $\mu_B = 0$ and no chemical potential is stored. At what intermediate flux is P optimal [46]? The change in chemical potential is given by the van't Hoff isotherm

$$\Delta\mu = \mu_B - \mu_A = -RT[\ln K - \ln(b^*/a^*)] \quad (C2.14.5)$$

where the stars serve as a reminder that the concentrations should be multiplied by activity coefficients, assumed to equal unity in the following discussion. Provided one is allowed to assume that the equilibrium constant $K = k_f/k_b$, a valid assumption if Boltzmann equilibrium is maintained in the steady state and hence a single rate coefficient correctly characterizes the reaction, then

$$\Delta\mu = RT \ln \frac{bk_b}{ak_f} = RT \ln \left[1 - \frac{j}{J_f} \right] \quad (\text{C2.14.6})$$

and the efficiency of free energy transfer is $(\mu_A + \Delta\mu)/\mu_A$. Writing P as $j(\mu_A + \Delta\mu)$, it is a simple matter to substitute in expression (C2.14.6) for $\Delta\mu$, differentiate with respect to j and set the derivative to zero, obtaining:

$$\frac{j_{\max}/J_f}{1 - j_{\max}/J_f} - \ln \left[1 - \frac{j_{\max}}{J_f} \right] = \frac{\mu_A}{RT}. \quad (\text{C2.14.7})$$

Once the reaction is accomplished, it is usually desirable to store the product, i.e. a further reaction



must follow (C2.14.4). Since all the reactions are reversible, at first sight it would appear impossible to store B indefinitely: the back reaction out of the store could only be prevented by an infinitely high energy barrier. To avoid this, the store is fitted with a door; only when this door is closed is the barrier infinitely high. Furthermore, the store has a variable volume such that the chemical potential is constant and independent of the amount of material contained.

When the door is open, the optimal net flux into the store is j_{\max} , given by equation (C2.14.7). It may be that the stochastically gated diffusion treated by Szabó *et al* [47], see also [48] is a good representation of typical biological storage reactions (C2.14.8).

C2.14.3.2 'MISSING ENTROPY'

The calorimetrically measured ΔH is usually assigned to the formation and breaking of chemical bonds. The equation

$$-RT \ln K = \Delta H - T \Delta S \quad (\text{C2.14.9})$$

linking enthalpy and entropy is, as has been aptly pointed out [49], 'as infallible as the laws of thermodynamics'; equally infallible statements can be generated by adding, *pace* Hall and Knight, some arbitrary quantity ξ to both terms on the right-hand side, which becomes $\Delta H + \xi - (T \Delta S + \xi)$. As Weber has pointed out [50], the separation of ΔH and ΔS depends upon specific hypotheses relating them, among which there is no *prima facie* unique choice, and one may legitimately enquire whether a proposed choice of ξ is appropriate. As hinted at by Planck [51], ΔH is a composite quantity comprising (a) the chemical bonding energy, and (b) the integral of the specific heat increments ΔC_p (defined as dQ/dT in a reversible change, where Q is the heat), i.e. $\int_0^T \Delta C_p dT$, obtainable by painstaking measurements of ΔC_p from 0 K to T . For ideal gases and other small entities, this integral is practically zero and can be neglected, but for biological macromolecules it could well be more significant than the chemical bonding energy. Since $C_p = (\partial H/\partial T)_p$, this assertion can easily be verified by enquiring whether ΔH varies with temperature for the reaction under consideration [52, 53]. Since $dQ = T dS$ in a reversible change, equation (C2.14.9) can be rewritten as [49]

$$-RT \ln K = \Delta H_0 - \Delta W \quad (\text{C2.14.10})$$

where the subscript 0 stands as a reminder that this is the (defined) heat of reaction at 0 K, and

$$\Delta W = T \int_0^T \frac{\Delta C_P}{T} dT - \int_0^T \Delta C_P dT \quad (\text{C2.14.11})$$

represents the work obtainable by conversion of ‘thermal’ heat expendable in the separation of chemically bonded atoms [49]. At all temperatures above absolute zero, the integral $\int_0^T \Delta C_P dT$ must be subtracted from the measured enthalpy in order to obtain the true heat of reaction.

This argument was pointed out almost 30 years ago by Benzinger [49], in a paper which referred to some still earlier work of his, and yet its implications, even more pertinent today given the wider use of calorimetry in molecular biology, still appear to be largely ignored, an exception being Weber’s work on the association enthalpy of protein subunits [50].

The ‘missing entropy’ $-\int_0^T (\Delta C_P/T) dT$ is proportional to the number of possible states. In typical biological macromolecules and also other nonergodic materials such as glasses and spin glasses, disorder is quenched: only one of the very large number of possible realizations occurs, and Nernst’s third law is violated [54].

The entropy of a solution is itself a composite quantity comprising: (i) a part depending only on the amount of solvent and solute species, and independent from what they are, and (ii) a part characteristic of the actual species (A, B, ...) involved (equal to zero for ideal solutions). These two parts have been denoted respectively cratic and unitary by Gurney [55]. At extreme dilution, (ii) becomes more or less negligible, and only the cratic term remains, whose contribution to the free energy of mixing is

-13-

$$-kT \ln W_{\text{config}} = kT (n_A \ln x_A + n_B \ln x_B + \dots) \quad (\text{C2.14.12})$$

where the n_A, x_A etc represent the numbers and mole fractions of A, B,....

C2.14.3.3 COOPERATIVE EFFECTS IN BINDING

Switching and control (‘signal transduction’) in biological systems as elsewhere usually strives to achieve a highly nonlinear response, which *inter alia* confers a certain immunity from noise onto the system. Cooperative binding is an easy way to achieve this end. Most signalling in biology is based on the binding of a ligand L to an unoccupied site S on a receptor B:



The ligand-receptor complex C has changed properties which typically allow it to undergo further, previously inaccessible reactions (e.g. binding to a DNA promoter sequence). The role of L is to switch B from one of its stable conformational states to another. The approximate equality of the intramolecular, molecule–solvent and L–B binding energies is an essential feature of such biological switching reactions. An equilibrium binding constant K_0 is defined according to the law of mass action:

$$K_0 = \frac{c}{s\ell}. \quad (\text{C2.14.14})$$

If there are n independent binding sites per receptor, conservation of mass dictates that $s = nb_0 - c$, where b_0 is the

total concentration of B, and the binding ratio $r = c/b_0$ (number of bound ligands per biopolymer) becomes

$$r = \frac{nK_0\ell}{1 + K_0\ell}. \quad (\text{C2.14.15})$$

Suppose now that the sites are not independent, but that addition of a second (and subsequent) ligand next to a previously bound one (characterized by an equilibrium constant K_1) is easier than the addition of the first ligand. In the case of a linear receptor B, the problem is formally equivalent to the one-dimensional Ising model of ferromagnetism, and neglecting end effects, one has [56]:

$$r = \frac{n}{2} \left(1 - \frac{1 - K_0\ell}{[(1 - K_0\ell)^2 + 4K_0\ell/q]^{1/2}} \right) \quad (\text{C2.14.16})$$

where the degree of cooperativity q is determined by the ratio of the equilibrium constants, $q = K_1 / K_0$. For $q > 1$ this yields a sigmoidal binding isotherm. Another interesting case, also yielding a sigmoidal relation between r and ℓ , is represented by the uncharged oligopeptide alamethicin which partitions as a monomer into bilayer lipid membranes and aggregates within the membrane [57].

-14-

If $q < 1$, then binding is anticooperative, for example when an electrically charged particle adsorbs at an initially neutral surface; the accumulated charge repels subsequent arrivals and makes their incorporation more difficult [58].

Another important noise-reduction mechanism is to incorporate a threshold into the responsive apparatus. Essentially this is why antibodies are multidentate, why serial triggering of antigen-presenting cells [59] is necessary, and so on. The benefits of a response threshold T in have been thoroughly investigated in the context of radiation detectors [60], and the argument can be adapted to biological detectors. Suppose that L ligands are incident on an area containing R receptors. The number arriving at any particular receptor will fluctuate around $\lambda = L/R$, the mean number of ligands per receptor, and assuming a Poisson distribution for these fluctuations, the expected number \bar{f} of activated receptors (i.e. those receiving T or more ligands) is fR , where

$$f = 1 - e^{-\lambda} \left(1 + \lambda + \frac{\lambda^2}{2!} + \dots + \frac{\lambda^{T-1}}{(T-1)!} \right). \quad (\text{C2.14.17})$$

\bar{f} is binomially distributed and its standard deviation σ is $[N\bar{f}(1 - \bar{f})]^{1/2}$. The least detectable signal L' must exceed L by a certain amount—let us suppose that the least detectable increment is σ (the argument remains unchanged if some other multiple of σ is taken) and it is given by the solution of

$$Rf' = Rf + \sigma \quad (\text{C2.14.18})$$

where f' is given by equation C2.14.17 with $\lambda = L' / R$. As an example, suppose that $R = 200$ and background $L = 400$. If $T = 1$, then 79 additional ligands are needed to engender a response; but if $T = 2$, only 53 are required. The optimal threshold depends on the expected background level. Note that the most perfect possible detector is still subject to a basic limitation due to the inherently fluctuating nature of the input; in the case of a Poisson process $\sigma = \sqrt{L} = 20$ ligands would be the smallest detectable increment.

C2.14.3.4 THE EXPERIMENTAL DETERMINATION OF ASSOCIATION (BINDING) CONSTANTS

The two main difficulties facing the experimenter are (i) how to detect binding, and (ii) how to ensure that the system under investigation is truly in an equilibrium state.

(i) Typically an atom or group of atoms is selected as a reporter whose measured property (e.g. intensity of a particular Raman line) is characteristic of the state of binding. One of the most popular reporters is the photoluminescence intensity at a certain wavelength, since it can be very easily measured using a commercial fluorimeter. Sometimes the intrinsic photoluminescence of an amino acid (tryptophan, tyrosine) can be used, but quantum yields are low and the sample has to be excited in the far ultraviolet. Hence a fluorescent group is often covalently bound to one of the reaction partners, although this may drastically change its binding properties, an obvious caveat all too often overlooked. Such assays have to be calibrated, typically by measuring the photoluminescence (or other property) at binding saturation [57], although since true saturation requires one partner to be in infinite excess more or less ingenious extrapolation procedures are required. Once this is done, the relationship linking photoluminescence intensity with intermediate degrees of binding must be established. These steps are not trivial and usually end up relying on assumptions (e.g. of linearity) which are far from incontrovertible.

-15-

The easiest way to accomplish (ii) is to dilute the supposedly equilibrium bound state and check that the predicted degree of dissociation takes place on the time scale of the experiment. This is often inconvenient in homogeneous assays, however. Another check is to incubate the ligand and receptor for different times τ : invariance of c with τ would be evidence of equilibrium having been reached, provided that the range of τ has been chosen judiciously. One may also compare the amount bound after adding successive small increments of ligand L to receptor B with the amount bound after having added L to B in a single large increment: if the system is in thermodynamic equilibrium, c should be path-independent.

The development of extremely sensitive microcalorimeters has popularized the ‘direct’ determination of reaction enthalpy ΔH simply by measuring the heat evolved upon mixing L and B. The free energy is determined from the equilibrium constant K deduced from fitting a plot of r versus μ to an appropriate isotherm, such as [equation \(C2.14.15\)](#). No labelling is required, but it is often awkward to properly establish the reversibility of the reaction, and the validity of the evaluation of K depends on the binding model assumed. A graver deficiency is the neglect of specific heat effects ([section C2.14.3.2](#)), manifested by temperature-dependent ‘standard’ reaction enthalpies. Here, as elsewhere, the very ease of experimentation can lead the investigator into error, and microcalorimetry is no exception.

Great interest has recently been developed in heterogeneous systems in which B is immobilized to a solid surface and ligand binding measured directly, e.g. using a quartz crystal microbalance (QCM) or an optical method such as optical waveguide lightmode spectroscopy (OWLS), ellipsometry or surface plasmon resonance (SPR) [61], i.e. the solid surface plays a dual role as both receptor and sensing platform. When carried out properly neither labelling of the participating molecules nor calibration of the response are required, and direct measurement of the reverse reaction can be accomplished with ease. These heterogeneous methods are discussed in more detail in [section C2.14.7.2](#).

Attempts have been made to combine the sensitive detection possibilities of heterogeneous systems with the more familiar (to non-electrochemists) homogeneous ones by recreating a quasihomogeneous environment between the solid–liquid interface proper (the sensing platform) and the bulk liquid phase by interposing a hydrogel (e.g. carboxydextran) fixed to the solid phase. The receptors are covalently attached to the hydrogel scaffold. A drawback to this procedure is that it has been found empirically that association kinetics measured with such hydrogels do not correspond to those expected from truly homogeneous systems, the reason being that mass transport within the hydrogel is sufficiently drastically retarded that binding becomes diffusion limited for all but the slowest reactions [4]. Furthermore, upon dissociation the ligand has an extremely high probability of rebinding before it can diffuse away from the receptor (see also [section C2.14.6.3](#)). Hence binding parameters derived from the kinetics do not reflect the true chemical affinity of the receptor for its ligand. This is apart from the fact that covalent immobilization to the hydrogel may inactivate the receptors, e.g. by involving amino acids on an epitope or at the active site. A good way to avoid these difficulties is to anchor B to a lipid bilayer [62].

C2.14.3.5 CONFORMATIONAL RELAXATION

Agno [63], Blumenfeld [64] and possibly others have emphasized that biological systems are constructions: a living cell is much closer to a mechanical clock than to a bowl of consommé. To characterize the latter, a statistical approach is adequate, in which the motions of an immense number of individual particles are subsumed into a few macroscopic parameters such as temperature and pressure. But one does not usually need to know the pressure when analysing the working of a clock. The energy contained in a given system can be divided into two categories: (a) the multitude of microscopic or thermal motions sufficiently characterized by the temperature and (b) the (usually small number of) macroscopic, highly correlated motions, whose existence turns the construction into a machine. The total energy

-16-

contained in the microscopic degrees of freedom may be far larger than those in the macroscopic ones, but nevertheless the microscopic energy can usually be successfully neglected in the analysis of a construction. In informational terms, the macrostates are remembered, but the microstates are not [65].

The question then is, to what degree can the microscopic motions influence the macroscopic ones: is there a flow of information between them [66]? Biological systems appear to be nonconservative *par excellence* and present at least the possibility that random thermal motions are continuously injecting new information into the macroscales. There is certainly no shortage of biological molecular machines for turning heat into correlated motion (e.g. [67] and section C2.14.5; note also [16]).

A construction makes use of only an insignificant fraction of the Gibbs canonical ensemble and hence is essentially out of equilibrium. This is different from thermodynamic nonequilibrium—it arises because the system is being investigated at time scales much shorter than those required for true statistical equilibrium. Such systems exhibit ‘broken ergodicity’ [68], as epitomized by a cup of coffee in a closed room to which cream is added and then stirred. The cream and coffee equilibrate within a few seconds (during which vast amounts of microinformation are generated within the whorled patterns); the cup attains room temperature within tens of minutes; and days may be required for the water in the cup to saturate the air in the room.

Broken ergodicity may be regarded as a generalization of broken symmetry, a concept introduced by Landau (see [69]) in the context of phase transitions, and which leads to a new thermodynamic quantity, the order parameter ξ whose value is zero in the symmetrical phase. ξ may be thought of as conferring a kind of generalized rigidity on a system [70], allowing an external force applied at one point to be transferred to another. Some protein molecules demonstrate this very clearly: flash photolysis of oxygenated haemoglobin causes motion of the iron core of the haem which results in (much larger) movement at the distant intersubunit contacts, leading ultimately to an overall change in the protein conformation involving hundreds of atoms.

In the case of enzymatic catalysis, it has been proposed that when substrate binds to the active site, local fast vibrational relaxation takes place on the picosecond time scale, but the active site is no longer in equilibrium with the rest of the molecule and the resulting strain modifies the energy surface on which the enzymatic reaction takes place [71, 72]. Subsequent conformational relaxation involves making and breaking a multiplicity of weak bonds, but at a slower rate than the reaction being catalysed. This description implies a definite and striking prediction: the reaction rate should exhibit an inverse Arrhenius temperature dependence, because increasing the temperature accelerates conformational relaxation, and hence shortens the time during which the strained molecule is able to accelerate the enzymatic reaction. Evidence for this mechanism is provided by the pulsed photolysis of carbonmonoxy (relaxed, R) haemoglobin at 532 nm. Time-resolved resonance Raman spectroscopy of aromatic amino acid residues associated with the intersubunit contact show that a strained tense (T) conformation (characterized by the tyr α_{42} – asp β_{99} intersubunit hydrogen bond and a close trp β_{37} – tyr α_{140} contact, indicated by an increased tyr 830/850 cm^{-1} Fermi doublet intensity ratio and a decreased trp 880 cm^{-1} band intensity respectively [73]) is produced within the 7 ns duration of the photolysis pulse. Strain is also inferred from the enhanced optical adsorption difference (compared with the difference between equilibrium T and R forms) at 315 nm (due to the trp β_{37} – tyr α_{140} contact) which appears on the nanosecond time scale. The enhancement then

relaxes with the same (microsecond) time constants [74] characteristic of tertiary structural changes in the vicinity of the haem iron which can be probed by the Soret band adsorption (R A Copeland, S Dasgupta, J J Ramsden, R H Austin and T G Spiro, unpublished observations). Another intriguing piece of evidence comes from direct observation of the adenosine triphosphate (ATP)-induced generation of mechanical force by immobilized myosin interacting with actin tethered to beads held in optical traps. Upon hydrolysis

-17-

of ATP, adenosine diphosphate (ADP) is released. Individual hydrolysis events could be monitored by microscopic observation of fluorescently labelled adenosine. The simultaneous monitoring of bead displacements due to the mechanical force exerted on the actin clearly showed that force is generated several hundred milliseconds after release of ADP [75].

C2.14.4 Kinetics It has already been emphasized (section C2.14.1, section C2.14.2.2 and section C2.14.3.1) that kinetics are of paramount importance in describing living systems [76]. The root of this may ultimately lie in the fact that whereas inanimate matter has endless time in which to undergo its transformations, mortal, animate matter is constantly racing against the clock.

C2.14.4.1 TRANSPORT IN BIOLOGICAL SYSTEMS

The determination of biological affinity by mixing two species and measuring their rates of association and dissociation presupposes that the contribution of transport to the association dynamics is precisely known. Well-defined hydrodynamic conditions are therefore a prerequisite for the experimental determination of affinities via rates.

In a homogeneous system, the rate of mixing is governed by Smoluchowski's equations [77], according to which the diffusion-limited association rate of S and L (equation (C2.14.13)), supposed uncharged, equals that of the flux and is

$$(C2.14.19)$$

where d and D are the molecular radii and diffusivities respectively. In the presence of an energy barrier characterized by an association (forward) rate coefficient k_f , one introduces a vicinal concentration (subscript v) and writes the rate as $dc/dt = k_{fv}$ per S, and the flux from the bulk to the vicinity of L as $4\pi(d_S + d_L)(D_S + D_L)(c - c_v)$, giving the familiar expression:

$$(C2.14.20)$$

The equivalent equations for heterogeneous and quasi-heterogeneous systems (the latter are small vesicles which can practically be handled as homogeneous systems, but which are nevertheless large enough to possess a macroscopic solid-liquid interface) are dealt with in section C2.14.7.

At first sight it seems that biochemical reactions taking place in the cytoplasm can be modelled homogeneously, but in fact the cytoplasm is a complex, highly viscous medium belonging to the class of soft matter or complex fluids, which bears little relation to the cytosol [78], and in which diffusion may be anomalous. It is a current experimental challenge to reconstruct the cytoplasm *in vitro* and systematically investigate biomolecular reaction kinetics in such media, although since the majority of intracellular reactions actually seem to take place at the solid-liquid interface [78], it is even more important to correctly apply the methods of heterogeneous kinetics to biochemical reactions.

-18-

An important aspect of biological transport is that nature makes extensive use of the reduction of dimensionality to speed up search and discovery (SD) (see also [section C2.14.6.2](#)). SD is enormously enhanced upon moving from three to two or one dimensions, because the spatial extent to be explored is drastically reduced. Affinity follows kinetics in being enhanced upon moving from three dimensions to two dimensions [79].

C2.14.4.2 SMALL SYSTEMS

Consider again the prototypical homogeneous reaction (C2.14.13), which Rényi has analysed in detail [80]. Taking the forward reaction only (other cases are also dealt with in [80]), and supposing that $k_f \ll 4 \pi (d_s + d_L)(D_S + D_L)$ (cf equation (C2.14.20)), then

$$\frac{dc}{dt} = k_f[\langle s \rangle \langle \ell \rangle + \Delta^2(\gamma_t)] = k_f \langle s \ell \rangle \quad (\text{C2.14.21})$$

where the angular brackets denote expected numbers, and γ_t is the number of C molecules created up to time t . The term $\Delta^2(\gamma_t)$ expresses the fluctuations in γ_t : $\langle \gamma_t^2 \rangle = \langle \gamma_t \rangle^2 + \Delta^2(\gamma_t)$: supposing that γ_t approximates to a Poisson distribution, then $\Delta^2(\gamma_t)$ will be of the same order of magnitude as $\langle \gamma_t \rangle$. The so-called kinetic mass action law (KMAL) putting $\langle s \rangle = s_0 - c(t)$ and so on, the subscript 0 denoting initial concentration at $t = 0$, is a first approximation in which $\Delta^2(\gamma_t)$ is supposed negligibly small compared to $\langle s \rangle$ and $\langle \ell \rangle$, implying that $\langle s \rangle \langle \ell \rangle = \langle s \ell \rangle$, whereas strictly speaking it is not since s and ℓ are not independent. The neglect of $\Delta^2(\gamma_t)$ is justified for molar quantities of starting reagents (except near the end of the process, when $\langle s \rangle$ and $\langle \ell \rangle$ become very small), but inconceivably so for reactions in minute subcellular compartments.

These number fluctuations, i.e. the $\Delta^2(\gamma_t)$ term, will constantly tend to be eliminated by diffusion. On the other hand, because of the correlation between s and ℓ , initial inhomogeneities in their spatial densities lead to the development of zones enriched in either one or the other faster than the enrichment can be eliminated by diffusion. Hence instead of L disappearing as t^{-1} (when $\ell_0 = s_0$), it is consumed as $t^{-3/4}$ [82], and in the case of a reversible reaction, equilibrium is approached as $t^{-3/4}$ [82] (charged particles are dealt with in [83]). Deviations from perfect mixing are more pronounced in dimensions lower than three.

C2.14.4.3 NONEXPONENTIAL DECAY AND ITS ORIGINS

The paradigmatical binding reaction (equation (C2.14.22)) is generally analysed as a second order forward reaction and a first order backward reaction, leading to the following rate law:

$$\frac{dc}{dt} = k_f s \ell - k_b c \quad (\text{C2.14.22})$$

supposing $\ell \approx \ell_v$. Despite its beguiling simplicity, this equation cannot, in general, be solved analytically, but a numerical solution is straightforward and can be fitted to experimental data to determine the forward and backward rate coefficients k_f and k_b . Ideally, the data collected should comprise both an association phase, during which S and L

are brought into contact, and a dissociation phase, in which C is diluted into a large volume of pure solvent, and the fitting carried out globally over both phases. It is unfortunate that many of the traditional binding assays used in biochemistry are awkward or impossible to apply to dissociation. This has led to an underappreciation of the fact

that simple Poisson dissociation (rate proportional to the amount remaining undissociated) giving familiar exponential decay is the exception rather than the rule in biomolecular interactions. This is very easy to demonstrate with a heterogeneous reaction such as the adsorption of serum albumin onto silica (e.g. [84, 85]): the dissociation rate coefficient is clearly time dependent. The amount of protein bound, $v(t)$, can be represented by the integral [86]

$$v(t) = k_f \ell \int_0^t \phi(t_1) Q(t, t_1) dt_1 \quad (\text{C2.14.23})$$

where ϕ is the fraction of unoccupied receptors. The memory kernel Q denotes the fraction of L bound at epoch t_1 which remain adsorbed at epoch t (if dissociation is indeed a first order (Poisson) process $Q(t) = \exp(-k_b t)$). A necessary condition for the system to reach equilibrium is

$$\lim_{t \rightarrow \infty} Q(t) = 0. \quad (\text{C2.14.24})$$

Processes of this type have been analysed [84, 85] by adding an irreversible step, either in parallel:

$$\frac{dc_{\text{irr}}}{dt} = k_{\text{irr}} s \ell \quad (\text{C2.14.25})$$

for which the memory function is [87]:

$$Q = \frac{k_{\text{irr}}}{k_f} + e^{-k_b t} \quad (\text{C2.14.26})$$

or in series

$$\frac{dc_{\text{irr}}}{dt} = k_{\text{irr}} c \quad (\text{C2.14.27})$$

for which the memory function is [86]

$$Q = \frac{k_{\text{irr}} + k_b e^{-(k_{\text{irr}} + k_b)t}}{k_{\text{irr}} + k_b} \quad (\text{C2.14.28})$$

to equation (C2.14.22), which is modified accordingly. Note that in neither of these cases does $\lim_{t \rightarrow \infty} Q(t) = 0$; the

systems do not reach equilibrium and the usual KMAL assumption that the equilibrium constant K can be equated to the quotient of the forward and backward rate coefficients does not apply; the backward reaction (dissociation) coefficient is time dependent and can be obtained from the quotient [86]

$$k_b(t) = \frac{\int_0^t \phi(t_1) Q'(t, t_1) dt_1}{\int_0^t \phi(t_1) Q(t, t_1) dt_1}. \quad (\text{C2.14.29})$$

The existence of multiple stable conformations with different affinities implies time-dependent dissociation: a molecule initially associated in a low-affinity conformation has the chance to switch to a high affinity one before dissociating. Even a single conformation may actually comprise many slightly different subconformations ('conformational substates', CS, possibly rotamers) separated by finite barriers [7, 40]. At some finite temperature the molecule will exist in several different CS (i.e. ergodicity is broken), each of which may be presumed to make a slightly different contribution to the rate of any process in which that conformation of the biopolymer participates. In biological (and inanimate glassy) systems relaxation is empirically often found to follow 'stretched exponential' (Kohlrausch) decay:

$$c(t)/c_0 = \exp[-(k_b t)^\beta], \quad 0 < \beta < 1. \quad (\text{C2.14.30})$$

If the contributions from the different CS are additive and relax in parallel,

$$\frac{c(t)}{c_0} = \int_0^\infty w(k_b) e^{-k_b t} dk_b \quad (\text{C2.14.31})$$

with which equation (C2.14.30) can be simulated, but unless there is some independent way of determining the weight distribution $w(k_b)$, its choice is arbitrary. Series relaxation avoids this difficulty [88]: relaxation on the n th level is only possible if certain elements in the $(n-1)$ th level satisfy some condition. For example, in the case of biopolymer relaxation (cf section C2.14.3.5) the condition might be that μ_{n-1} monomer units in level $n-1$ attain one particular state of their $2^{\mu_{n-1}}$ possible ones, giving an average relaxation time $\tau_n = 2^{\mu_{n-1}} \tau_{n-1}$.

C2.14.4.4 PATTERN FORMATION

Relative to the multicellular organism into which it develops, the fertilized egg is rather homogeneous, but within a few generations of cell division, an embryonic animal already shows remarkable spatial variation, which ultimately develops into the differentiated limbs and organs of the adult organism. The realization that diffusion and chemical reaction provides an adequate basis for the formulation of a mathematical model of morphogenesis dates back to a seminal paper published by Turing in 1952 [89]. The essential idea is that the initially homogeneous, stable state moves out of stability due to some random disturbance (not necessarily diffusive; wetting and percolation may also play a role [90]). The unstable state generates waves of morphogens, molecules capable of leading to the generation of differentiated forms.

The great complexity of morphogenesis makes it rather difficult to formulate a theory of the process beyond stating

the equations; particular cases have to be investigated with the help of numerical simulation. In this respect, cellular automata [91] seem to show great promise. An example is the model of neurogenesis in *Drosophila* recently described by Luthi *et al* [92]. The system is divided up into cells corresponding to the actual cellular divisions of the organism, each of which is initially assigned concentrations of a substance which promotes neuralization within the cell and inhibits it in the neighbours to which it is transmitted. Subsequent evolution is governed by plausible rules inspired by recent advances in the molecular biology of the developing embryo.

C2.14.5 BIOLOGICAL MACHINES

One of the most fascinating recent developments in biology has been the discovery of numerous highly complex biopolymer assemblies (see also section C2.14.2.3) such as the ribosome or the bacterial flagellum [93, 94 and 95], the envy of nanotechnologists seeking to miniaturize man-made mechanical devices (note that the word 'machinery' is also sometimes used to refer to multienzyme complexes such as the proteasome [96]), and an entire

organism might indeed be considered as a machine [63]. Even very complex processes such as mitosis can now be analysed in considerable biophysicochemical detail [97]. Mitosis serves as an exemplar of the process design which seems to epitomize much of life, the so-called S^3 architecture [97]: stochastic (influenced by noise and showing strong fluctuations); self-correcting (working by trial and error, with checkpoints and feedbacks to ensure efficient regulation); and synchronized (which may itself be self-correcting).

Mitosis is characterized [99] by steady elongation (at a velocity v_g , depending *inter alia* on the concentration of monomeric tubulin) of the microtubule (polymerized tubulin) filaments which search for, and ultimately mechanically separate freshly replicated DNA prior to cell division, punctuated by their abrupt shrinkage (with velocity v_s). This dynamic instability is characterized by length fluctuations of the order of the mean microtubule length, hinting at a phase transition. Let f_{gs} denote the frequency of switching from growth to shrinkage ('catastrophe'), and the reverse switching back to growth (in a different direction) by f_{sg} ('rescue'). When $v_g f_{sg} = v_s f_{gs}$, at which point the average tubule length $\bar{l} = v_g v_s / (v_s f_{gs} - v_g f_{sg})$ diverges, growth switches from unbounded (during the so-called interphase, between cell division) to bounded (during mitosis, when the microtubules have to find and grab chromosomes) [97]. The molecular origin of catastrophe and rescue lies in the fact that tubulin monomers can bind to guanosine triphosphate (GTP), and the complex can spontaneously assemble to form filaments. But the GTP slowly hydrolyses to guanosine diphosphate (GDP), thereby somewhat changing the tubulin conformation such that it prefers to be monomeric. The microtubule can only be disassembled from the end, however: a catastrophe occurs if the rate of GTP hydrolysis exceeds that of tubulin addition for a while. After a catastrophe, growth occurs in a new direction. This dynamic instability-based mechanism is an extremely efficient way to search a volume [100].

C2.14.5.1 THE GENERATION OF TRANSLATION AND ROTATION

The forces involved in muscle contraction [101] can now be directly scrutinized by attaching an actin filament to a small dielectric sphere which can be nanomanipulated using optical tweezers [102, 103] and bringing the filament into contact with myosin. Using similar techniques it has become possible to directly observe kinesin molecules moving along microtubules [104]: the kinesin is labelled with a fluorescent molecule and imaged with low background total internal reflexion fluorescence microscopy, sensitive enough to detect single molecules.

Apart from the development of imaging and force measurement devices, an important biophysico-chemical problem is

-22-

understanding how motion occurs within the framework of molecular interactions. The general concept is based on Brownian particles moving along a periodic but asymmetric ('sawtooth') potential, resulting in directed ('processive') motion [67] (see [95] and [105] for examples of rotatory motors). Either the potential or the force acting on the particle fluctuate. Beautiful experiments based on direct imaging of these molecular motions, and direct measurements of the fluctuating forces, have enabled this generic concept to be refined into realistic models whose parameters are closely bounded by the experimental observations [67]. Kinesin has two heads connected by a flexible hinge. At the start of a cycle, both heads are sitting in a potential well. ATP hydrolysis (see below; and cf the effect of GTP hydrolysis on tubulin, above, and the general mechanism of enzyme catalysis described in [section C2.14.3.5](#)) results in a conformational change of one head which consequently advances to the next well. The hinge is now strained, and during relaxation bringing the two heads together again it is more probable (because of the asymmetric shape of the wells) that the laggard head is dragged to the advanced one, rather than vice versa. This relaxation together with the release of hydrolysed ATP completes the cycle [67].

A vital biophysico-chemical problem is to understand how chemical energy (released by ATP or GTP hydrolysis [105], or by protons falling down an electrochemical potential gradient [95]) is converted into mechanical energy. The thermodynamic constraints on the energy requirements of biological machines have been set out by Gray [106]. The force F which has to be applied to a molecular lever requires accurate knowledge of its position x if reversible work is to be performed. Specifying the positional accuracy as Δx , the uncertainty principle gives the energy requirement as

$$\Delta E \geq hc/(4\Delta x) \quad (\text{C2.14.32})$$

and the uncertainty in the force generated at x is then

$$\Delta F = F(x) \pm \Delta x(dF/dx). \quad (\text{C2.14.33})$$

To compute the work W done by the system, equation (C2.14.33) is integrated over the appropriate x interval. The first term on the right-hand side yields the reversible work W_{rev} , and the second term yields $-\Delta x \sum_j |F_j - F_{j+1}|$ for any cycle involving j steps. The energy conversion factor ϵ is

$$\epsilon = W/(Q + \Delta E) \quad (\text{C2.14.34})$$

where Q is the net energy input during the cycle. With the help of inequality (C2.14.32) and defining two dimensionless quantities:

$$\alpha = hc \sum_j |F_j - F_{j+1}|/(4QW_{\text{rev}}) \quad (\text{C2.14.35})$$

and

-23-

$$z = \Delta E/Q \quad (\text{C2.14.36})$$

(the relative energy cost of control), one can write

$$\epsilon/\epsilon_{\text{rev}} \leq (1 - \alpha/z)/(1 + z) \quad (\text{C2.14.37})$$

where $\epsilon_{\text{rev}} = W_{\text{rev}}/Q$, the classical conversion factor. The maximum possible value of this ratio is obtained by substituting z by its optimal value z_{opt} , obtained from the turning point of equation (C2.14.33) and given by

$$z_{\text{opt}} = \alpha(1 + \sqrt{1 + 1/\alpha}) \quad (\text{C2.14.38})$$

which is

$$\left(\frac{\epsilon}{\epsilon_{\text{rev}}}\right)_{\text{max}} = \frac{1 - 1/(1 + \sqrt{1 + 1/\alpha})}{1 + \alpha(1 + \sqrt{1 + 1/\alpha})}. \quad (\text{C2.14.39})$$

If more energy than z_{opt} is used, then α decreases because of the energy cost of information; if less, then ϵ decreases because of the irreversibility (dissipation etc). For a macroscopic system these quantities are insignificant. But consider the myosin motor: taking $F_j \approx 2$ pN [102, 103], the displacement $x \approx 10$ nm [102], and $Q \approx 0.067$ aJ (the energy released by hydrolysing a single ATP molecule), then the energy cost of optimum control, Qz_{opt} , is equivalent to hydrolysing almost 150 ATP molecules (cf [107]) and $(\epsilon/\epsilon_{\text{rev}})_{\text{opt}} = 0.0033$. As with the

storage reaction discussed earlier ([section C2.14.3.1](#)), reversible operation is far from efficient. Chemical to mechanical conversion occurs at a finite rate which may be essentially uncontrolled, i.e. determined intrinsically. The parallel to the storage reaction may be developed further by noting that the duty ratio of a molecular machine (the fraction of time the motor spends attached and working) corresponds to the fraction of time the store is open.

C2.14.6 THE SPECIFICITY OF BIOMOLECULAR INTERACTIONS

The marvellous intricacy of a living organism could not function without the multitude of highly specific interactions which pervade almost every aspect of physiology. The concept of molecular recognition can be traced at least as far back as Fischer's lock and key mechanism for the recognition of its substrate by an enzyme [108]. Given perfect mixing along with specificity, cell physiology could presumably function in a structureless medium on the basis of concentration gradients and diffusion; real cells are internally structured, but even the transport of molecules and organelles along cytoskeletal filaments requires specific binding of molecules to their carriers, and the assembly of large multimolecular, multifunctional complexes ([section C2.14.2.3](#)) also requires specific recognition between the constituents.

From the viewpoint of biophysical chemistry, the main problems to be solved are: (i) what is the submolecular basis of

-24-

recognition and (ii) what is the required degree of specificity? Beyond these two, the study of specific interactions inevitably leads to questions about the origin of specificity and how it evolved from more primitive and presumably less specific interactions, but these lie beyond the scope of this chapter.

C2.14.6.1 THE SUBMOLECULAR BASIS OF RECOGNITION

One of the earliest successes of biophysical chemistry in the postwar era, which helped to lay the foundations of modern molecular biology, was the discovery of the base pairing mechanism in nucleic acids [16, 37], based on the hydrogen bond [38, 109] ([figure C2.14.2](#)), which could be considered to be the most important bond in biology; as well as providing the basis for molecular recognition and all that implies, it also gives water its unusual and vital properties. As well as ensuring the fidelity of DNA replication and its transcription into RNA, nucleotide base pairing also allows RNA to adopt the unique structure ([figure C2.14.1](#) and [section C2.14.2.2](#)) needed for its subsequent translation into protein [110]. Furthermore, hydrogen bonding determines the folding of a denatured protein in water into its three-dimensional conformation via hydrogen bonds between donor and acceptor groups on polar amino acids, and the inability of water to form hydrogen bonds with apolar amino acid residues, which drives as many of them as possible into the protein interior.

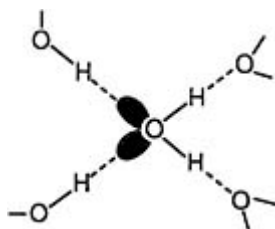


Figure C2.14.2. The hydrogen bond in water. The oxygen lone pairs (shaded blobs) are the donors, and the hydrogen atoms the acceptors [177, 178].

Given that the hydrogen bond is rather weak, with bond energies $\Delta E \sim$ a few kT , biological recognition has to be based on multiple interactions (cf [section C2.14.3.3](#)). If dissociation requires the simultaneous, independent rupture of all the interactions, then k_b (cf [equation \(C2.14.13\)](#)) $\sim \exp(-\nu \Delta E/kT)$, where ν is the number of bonds, even a few of which, taken together, thereby become equivalent to a single strong covalent bond.

Hydrogen donor/acceptor complementarity, complemented by electrostatic complementarity, although this appears to play the minor role, is the basis for a vast effort in computational drug design based on putative receptor structures mostly derived from x-ray crystallography [111]. Calculations based on static structures without allowing for subtle structural modifications of the binding partners following initial association have produced less than spectacular results; attempts are now being made to incorporate flexibility into the simulated molecules. The simple idea of docking taking place much as two rigid spacecraft interact is further complicated by the ubiquitous presence of water, itself a strongly hydrogen bonding molecule. Some interactions may involve expulsion of solvent. The omission of water in numerical simulations of docking is likely to be fatal for the accuracy and relevance of the results.

-25-

C2.14.6.2 THE REQUIRED DEGREE OF SPECIFICITY

SEARCH AND DISCOVERY (SD)

A somewhat naïve view is that affinity should be maximal for the molecule to be recognized, and zero for all other molecules. This strategy may not overall be the most efficient for recognition. Given the essential similarity of many biomolecules, the complete absence of attraction or even the presence of repulsion between any given pair of dissimilar molecules is likely to be rather exceptional, but nature can make good use of weak, ‘nonspecific’ attractive interactions. Consider the recognition of a particular DNA base sequence, say 10 base pairs long, by a protein. This type of process is very common and is the basis of transcription regulation and DNA restriction (scission of foreign DNA). If out of all 4^{10} possible sequences only the correct sequence has any affinity, then it can only be found by a tedious process of trial and error in three-dimensional space. If the protein has weak affinity for all of the DNA however, it can quickly bind anywhere on the molecule, and then execute a fast one-dimensional walk along it until the recognition site is found [112]. Searching in one or two dimensions is much more efficient than in three [113], and it has been experimentally demonstrated that a dimerization reaction has a much higher affinity at the two-dimensional solid–liquid interface than in three-dimensional bulk liquid [79].

A conceptually related effect occurs in immune recognition, when a ligand (antigen) present at the surface of an antigen presenting cell (APC) is bound by a T lymphocyte (TL). Binding triggers a conformational change in the receptor protein to which the antigen is fixed, which initiates further processes within the APC, resulting in the synthesis of more receptors, and so on. Apparently, effective stimulation of these further processes depends on *sustained* activation at the surface (*pace* the noise-reduction effect of a response threshold discussed in [section C2.14.3.3](#), and cf [88]). This can be accomplished with a few, or even only one TL, provided the affinity is not too high: the TL binds, triggers one receptor, then dissociates and binds anew to a nearby untriggered receptor (successive binding attempts in solution are highly correlated [114, 115 and 116]). This ‘serial triggering’ [59] can formally be described by:



(with rate coefficient k_a) where the starred R denotes an activated receptor, and



with rate coefficient k_d for dissociation of the ligand from the activated receptor, and the same rate coefficient k_a for reassociation of the ligand with an already activated receptor. The rate of activation (triggering) is $-dr/dt = -k_a r$, solvable by noting that $d\ell/dt = -k_a(r + r^*) + k_d r_L^*$. One obtains

$$(C2.14.42)$$

$$\ell(t) = \frac{k_a \tau}{1 - Y e^{-t/\tau}} + \frac{k_a(\ell_0 - r_0) - k_d - 1/\tau}{2k_a}$$

-26-

where $\tau = \{4\ell_0 k_a k_d + [k_a(\ell_0 - r_0) - k_d]^2\}^{-1/2}$ and $Y = (k_d + k_a[\ell_0 + r_0] - 1/\tau)/(k_d + k_a[\ell_0 + r_0] + 1/\tau)$, and the sought-for solution is then

$$r(t) = r_0 \exp \left[\ln \left(\frac{1 - Y e^{-t/\tau}}{1 - Y} \right) - \frac{t}{\tau} \right]. \quad (\text{C2.14.43})$$

CELLULAR RECOGNITION

Yet another example of SD is provided by the leukocytes which are constantly circulating in the bloodstream but do not normally interact with tissue. Venules of inflamed and infected tissue are dilated, however, which changes the hydrodynamic regimen and allows some leukocytes to come into contact with the venule wall (endothelium) [117]. The leukocytes are coated with glycoproteins which can interact with selectins (proteins able to selectively bind oligosaccharides) present in the outer membranes of the venule wall tissue. The combination of weakened hydrodynamic flow and weak selectin-glycoprotein bonds induces the leukocytes to roll along the venule wall [117], until they encounter integrins, another class of membrane-embedded proteins, which are able to interact more strongly with complementary molecules embedded in the leukocyte outer membrane. The leukocyte then stops, spreads out over the endothelium and penetrates between its constituent cells in order to search for, and destroy, pathogens.

THE IMMUNE REPERTOIRE

Antibodies binding to an antigen interact with a relatively small portion of the molecule. The number N of foreign antigens which must be recognized by an organism is very large, perhaps greater than 10^{16} , and there is a smaller number N' ($\sim 10^6$) of self-antigens which must *not* be recognized. Yet the immunoglobulin and T-cell receptors may only contain $n \sim 10^7$ different motifs. Recognition is presumed to be accomplished by a generalized lock and key mechanism involving complementary amino acid sequences. How large should the complementary region be, supposing that the system has evolved to optimize the task [118]? (A similar problem is posed by the olfactory system [119].) If P_S is the probability that a random receptor recognizes a random antigen, the value of its complement $P_F = 1 - P_S$ maximizing the product of the probabilities that each antigen is recognized by at least one receptor, and that none of the self-antigens is recognized, i.e. $(1 - P_F)^N P_F^{nN'}$, is:

$$P_F = \left(1 + \frac{N}{N'} \right)^{-1/n}. \quad (\text{C2.14.44})$$

Using the above estimates for n , N and N' , one computes $P_S \approx 2 \times 10^{-6}$. Suppose that the complementary sequence is composed of m classes of amino acids and that at least c complementary pairs on a sequence of s amino acids are required for recognition. Since the probability of a long match is very small, to a good approximation the individual contributions to the match can be regarded as being independent. A pair is thus matched with probability $1/m$, and mismatched with probability $1 - 1/m$. Starting at one end of the sequence, runs of c matches occur with probability m^{-c} , and elsewhere they are preceded by a mismatch and can start at $s - c$ possible sites. Hence

$$P_S = [(s - c)(m - 1)/m + 1]/m^c. \quad (\text{C2.14.45})$$

If $s \gg c > 1$, one obtains

$$c = \log_m [s(m - 1)/m] - \log_m P_S. \quad (\text{C2.14.46})$$

Supposing s to be a few tens, $m = 3$ (positive, negative and neutral residues), and again using the numbers given above (since they all enter as logarithms the exact values are not critical) one estimates $c \sim 15$, which seems to be in good agreement with observation.

C2.14.6.3 MODELLING INTERACTIONS WITH BROWNIAN DYNAMICS

The first predictions of antibody–antigen binding rates were made on the basis of the Smoluchowski [equation \(C2.14.20\)](#). Experimental work suggested a rate about 1000 times slower, which was understood to reflect the rather precise rotational alignment required for two molecules to dock specifically, since the area of the docking zone (epitope) is only a tiny fraction of the total surface area of the antibody. Careful calculations taking this into account indicated that the actual rates should be about a million times slower than those predicted from [equation \(C2.14.20\)](#), and that the experimentally measured rates were therefore a thousand times faster than expected. Two interpretations for the discrepancy were proposed: long range attractive forces steering the antigen to the complementary sequence on the antibody [120]; and the Franck–Rabinowitsch (cage) effect [114, 116]. It was a notable early achievement of Brownian dynamics (BD) [120, 121 and 122] to elucidate the conditions under which either or both apply. In these simulations, a large number of Brownian trajectories of the ligand are started on the surface of a sphere of radius b centred on the receptor. A fraction β terminate with a successful encounter, and the remainder reach the surface of a ‘quitting sphere’ of radius $q > b$. The bimolecular association rate coefficient is $\beta k_a(b)/[1 - \Omega(1-\beta)]$, where

$$k_a(b) = \left[\int_b^\infty \frac{\exp(\Delta G^{\text{IF}}/RT)}{4\pi D z^2} dz \right]^{-1}. \quad (\text{C2.14.47})$$

$\Delta G^{\text{IF}}(z)$ is the ligand–receptor interfacial interaction potential ([section C2.14.7.1](#), [equation \(C2.14.52\)](#), [equation \(C2.14.53\)](#) and [equation \(C2.14.54\)](#)), and Ω is the probability that a particle at q will return to b , equal to the ratio $k_a(b)/k_a(q)$ [121].

In favourable contrast to molecular dynamics, BD allows molecular movements of realistically long duration to be simulated. Nevertheless, the practical number of protein molecules which can be simulated is only two; since collective phenomena are often of crucial importance in determining the course of interaction events, other simulation techniques, such as cellular automata [115], need to be used to capture the behaviour of large numbers of particles.

C2.14.7 INTERFACIAL PHENOMENA

Interfaces play a predominant role in metabolic processes [78], as well as in immune recognition (e.g. T lymphocytes recognizing antigens on the surface of antigen presenting cells, [section c2.14.6.2](#)). Historically, the bulk of experimental

work on the kinetics of biomolecular interactions has dealt with homogeneous systems, however. Heterogeneous systems are in principle very attractive for investigating molecular interactions experimentally because the

contribution of transport to the kinetics can be precisely controlled, allowing diffusion and chemical effects to be separated. Several excellent experimental techniques for investigating heterogeneous binding kinetics *in situ* have been developed during recent years. Compared with the methods available for homogeneous systems, they are generally more sensitive and less cumbersome, have better time resolution, and do not require the use of labelled molecules [61]. The *in situ* capability arises because the solid part of the interface can play an additional role as a transducer for converting the number of bound molecules into an electrical or optical signal. The various techniques may be classified as follows:

- i. *electrochemical*: typically the change of electrode impedance due to the adsorption of biomolecules is monitored. Another possibility is to monitor the collective oscillations (surface plasmons) of the electrons in a thin metallic film [123]. Since these oscillations are excited at optical frequencies the measured surface plasmon resonance (SPR) may be considered as a hybrid opto-electronic technique. The oscillations are retarded by interactions with biomolecules adsorbed at the metal surface;
- ii. *mechanical*: the oscillation frequency of a quartz crystal is inversely proportional to the mass of biomolecules (and their shape and viscoelasticity) attached to the crystal surface;
- iii. *optical*: the reflectance change of the solid–liquid interface due to the formation of a thin film of biomolecules is monitored. Since they are the most sensitive, the most versatile (especially regarding possible choices of solid materials) and the most informative, optical methods have become rather popular.

The simplest approach conceptually is to directly measure the reflectance at different angles and fit the Fresnel equations to the data (scanning angle reflectometry, SAR) [124]. A thin film of adsorbed biomolecules needs at least two parameters to characterize it, its refractive index n_A and geometrical thickness d_A . Even though an adlayer composed of randomly adsorbed particles is nonuniform (heterogeneous on the nanometer scale), the uniform thin film approximation appears to yield satisfactory results [125], although the optically determined geometrical thickness depends on the refractive index profile perpendicular to the interface and may be smaller than the largest dimension of the adsorbed molecule. The number ν of adsorbed molecules per unit area can be calculated from the relation [126]

$$\nu = \frac{(n_A - n_C)d_A}{dn/dc} \quad (\text{C2.14.48})$$

where n_C is the refractive index of the solvent and dn/dc is the refractive index increment of the biomolecules in solution [127].

The Fresnel equations predict that reflexion changes the polarization of light, measurement of which forms the basis of ellipsometry [128]. Although more sensitive than SAR, it is not possible to solve the equations linking the measured parameters with n_A and d_A in closed form, and hence they cannot be solved unambiguously, although their product yielding ν (equation C2.14.48) appears to be robust.

The most recently introduced optical technique is based on the retardation of light guided in an optical waveguide when biomolecules of a polarizability different from that of the solvent they displace are adsorbed at the waveguide surface (optical waveguide lightmode spectroscopy, OWLS) [11]. It is even more sensitive than ellipsometry, and the mode

equations characterizing the phase velocities of the guided light can be solved analytically to yield n_A and d_A . All work reported so far appears to have been restricted to the measurement of two modes, which are sufficient to completely characterize uniform isotropic films, but in principle higher modes can also be measured, enabling more complex films to be characterized. The integrated optical interferometer [129], in which two orthogonally polarized modes interfere with one another (since each polarization interacts differently with the adsorbed biomolecules, the interference pattern shifts according to ν) is a further development which offers even higher sensitivity,

proportional to the path length over which interference and adsorption take place.

C2.14.7.1 BIOCOMPATIBILITY

The reactions of biopolymers at interfaces form the basis of some extremely important industrial processes. The primary process in all cases is the adsorption of biomolecules, usually proteins. If ultimately living cells are adsorbed, this always takes place onto a preadsorbed protein layer (which may be secreted by the cells themselves [130]). These processes can be classified into three categories:

- i minimal adsorption, as in the preparation of materials for surgical implants in contact with the blood (stents, replacement tubing, heart valves, biosensors, etc), filtration (including renal dialysis), storage of pharmaceuticals in solution, and antifouling paint for ships' hulls;
- ii maximal adsorption, mainly for surgical implants in contact with tissue (e.g. replacement bones) which should be mechanically firmly integrated into the host;
- iii variable adsorption, as in coatings for chromatographic separation materials.

Current emphasis is on the behaviour of proteins at the solid–liquid interface, but liquid–air and liquid–liquid interfaces, which were actually investigated much earlier [131], are still important.

Much of the science of biocompatibility can be reduced to the principles of how to determine the interfacial energies between biopolymer and surface. The biopolymer is considered to be large enough to behave as bulk material with a surface; since (for example) a water cluster containing only 15 molecules and with a diameter of 0.5 nm already behaves as a bulk liquid [132] it appears that most biological macromolecules can be considered to have surfaces. The interfacial energy ΔG^{IF} can be decomposed into three components:

$$\Delta G_{123}^{IF} = \Delta G_{123}^{(LW)} + \Delta G_{123}^{(da)} + \Delta G_{123}^{(el)} \quad (C2.14.49)$$

corresponding to the Lifschitz–van der Waals (LW), electron donor-acceptor (da) and electrostatic (el) interactions; subscripts 1, 2 and 3 refer to solid, solvent and biopolymer respectively.

A salient feature of natural surfaces is that they are overwhelmingly electron donors [133]. This is the basis for the ubiquitous 'hydrophilic repulsion' which ensures that a cell can function, since massive protein–protein aggregation and protein–membrane adsorption is thereby prevented. In fact, for biomolecule interactions under typical physiological conditions, i.e. aqueous solutions of moderately high ionic strength, the donor–acceptor energy dominates.

-30-

Tables of single substance surface tensions γ can be built up through the measurement of contact angles [134]; a few examples are collected in [table C2.14.1](#). These can be combined pairwise according to:

$$\gamma_{12}^{(LW)} = (\sqrt{\gamma_1^{(LW)}} - \sqrt{\gamma_2^{(LW)}})^2 \quad (C2.14.50)$$

and

$$\gamma_{12}^{(da)} = 2(\sqrt{\gamma_1^{\oplus}} - \sqrt{\gamma_2^{\oplus}})(\sqrt{\gamma_1^{\ominus}} - \sqrt{\gamma_2^{\ominus}}). \quad (C2.14.51)$$

The interfacial free energies for infinite parallel surfaces at contact are given by the relation [134]:

$$\Delta G_{123}^{(\text{LW or da})\parallel} = \gamma_{13}^{(\text{LW or da})} - \gamma_{12}^{(\text{LW or da})} - \gamma_{23}^{(\text{LW or da})} \quad (\text{C2.14.52})$$

and these can already provide an indication whether a surface is suitable for promoting or hindering biopolymer adsorption.

A next step is to consider the surface–particle distance z and curvature (interfacial radius R) dependence of the interactions [134], for which approximate expressions are:

$$\Delta G^{(\text{LW})} = 2\pi \ell_0^2 \Delta G^{(\text{LW})\parallel} R/z \quad (\text{C2.14.53})$$

where ℓ_0 is the equilibrium contact distance (distance of closest approach);

$$\Delta G^{(\text{el})} = 4\pi \varepsilon_0 \varepsilon \psi_1 \psi_3 \ln[1 + \exp(-\kappa z)] R \quad (\text{C2.14.54})$$

where the ψ are the electrostatic surface potentials (see [135] for an up-to-date discussion), and $1/\kappa$ is the Debye length; and

$$\Delta G^{(\text{da})} = 2\pi \chi \Delta G^{(\text{da})\parallel} \exp[(\ell_0 - z)/\chi] r \quad (\text{C2.14.55})$$

where χ is the decay length for the da interaction. These equations (equation C2.14.53), (equation C2.14.54) and (equation C2.14.55) are for a sphere approaching an infinite plane; for two spheres approaching each other the perfect energies must be halved. Their sum (C2.14.49) can be integrated to compute the association distance δ_a [136]:

-31-

$$\delta_a = \int_{\ell_0}^{\infty} [\exp(\Delta G(z)/kT) - 1] dz \quad (\text{C2.14.56})$$

whence the adsorption rate constant (cf k_f in section C2.14.4 can be computed:

$$k_a = D/\delta_a. \quad (\text{C2.14.57})$$

Cases are known in which the use of the single substance surface tensions leads to predictions at variance with observation. For example, using equations (C2.14.49), (C2.14.50), (C2.14.51), (C2.14.52), (C2.14.53), (C2.14.54) and (C2.14.55) and the data in table C2.14.1 the interfacial free energy between serum albumin and silica is predicted to be positive, and the protein should therefore be repelled, whereas as is well known it is strongly adsorbed. There are several possible reasons for discrepancies. One is that the characteristic length scale of the (macroscopic) surface tension is different from (probably larger than) the characteristic length for protein adsorption; possibly it is more appropriate to use the microsurface tension [137] for these calculations. Another is that the use of average curvatures, surface potentials and so on is too crude for biomolecules with their intricate surface topography. Finally, equation C2.14.49, equation C2.14.50, equation C2.14.51, equation C2.14.52, equation C2.14.53, equation C2.14.54 and equation C2.14.55 assume that the protein remains unchanged upon interaction with the surface. While this is likely to be true up to the moment of initial contact, native folded proteins are only marginally stable and intramolecular contacts maintaining the native structure may be substituted by molecule–surface contacts with

concomitant unfolding. Since [equation C2.14.56](#) and [equation C2.14.57](#) apply to the approach of the protein up to its initial contact with the surface, [equation C2.14.49](#), [equation C2.14.50](#), [equation C2.14.51](#), [equation C2.14.52](#), [equation C2.14.53](#), [equation C2.14.54](#) and [equation C2.14.55](#) should be valid for computing k^a [equation C2.14.56](#) – [equation C2.14.57](#), but a complete description of the adsorption process must take subsequent events into account. Folding is entropically costly since compact configurations are restricted to fairly small regions of the Ramachandran map [39], but in the native conformation this cost is outweighed by the enthalpy-losing intramolecular contacts. At a surface, however, the possibility of losing enthalpy by protein–surface contacts enables the molecule to adopt an extended, less entropically costly (since the corresponding Ramachandran map regions are large) configuration. This view is corroborated by observed changes in the optical rotation (circular dichroism) of protein solutions to which minute colloidal particles onto which the proteins can adsorb are added; the changes are consistent with varying degrees of loss of α -helical secondary structure [138]. There are still some puzzles, however: some protein–surface combinations appear to lead to an increase of α -helical structure [139], and differential scanning calorimetry of proteins in the absence and presence of minute colloidal particles [140] have shown that the temperature of the denaturation transition can be significantly lower for adsorbed proteins compared with the native dissolved state. Clearly the nature of the protein–surface contacts need more careful scrutiny: a complicating feature is that if conformational rearrangement does take place, it will usually lead to a biopolymer surface chemically different from that of the native conformation. For example, essentially no polar amino acids are found in the interior of a globular protein, and therefore almost any conformational rearrangement must result in the dilution of polar residues on the surface. The diminished protein surface polarity should result in stronger adsorption to apolar surfaces. A plethora of non-native adsorptive contacts constraining the polypeptide chain could in principle cost even more entropy than the native folded structure. Desolvation of the protein–solvent and protein–surface interfaces will also contribute to the free energy [141]. A further complication at all but the lowest coverages is that lateral interactions between adsorbed proteins will also affect ΔG^{IF} [142]. Far too few different proteins and surfaces have been investigated experimentally sufficiently carefully for reliable general conclusions to be drawn on these matters at present.

Table C2.14.1 Single substance macroscopic surface tensions/(mJ m⁻²) of various materials (data mostly from [134]).

Material	$\gamma^{(LW)}$	γ^{\oplus}	γ^{\ominus}
Biomaterials			
Cellulose	44	1.6	17
Dextran T-150	42	0	55
Fibrinogen	37	0.1	38
Immunoglobulin	34	1.5	50
Lecithin	29	2.7	60
Serum albumin	27	6.3	51
Synthetic polymers			
Nylon 6,6	36	0.02	22
Polyethylene	33	0	0
Polyethylene oxide	43	0	64
Polystyrene	42	0	1.1
Polyvinyl chloride	43	0.04	3.5
Teflon	18	0	0

Metal oxides			
SiO ₂	39	0.8	41
TiO ₂	42	0.6	46
ZrO ₂	35	1.3	3.6
Liquids			
Water	22	25.5	25.5
Ethanol	19	0	68
Chloroform	27	3.8	0
Hexadecane	28	0	0

-33-

Materials implanted in a living body should not engender an immune response. Even though the adsorption of a few pure proteins to interfaces under diverse conditions appears now to be reasonably well characterized, at least phenomenologically, adsorption from highly complex body fluids such as blood are only just beginning to be investigated quantitatively using the same methods applied successfully to pure materials [143]. Among the hundreds of different proteins in blood are enzymatic networks capable of triggering thrombus formation [144], or an immune response as soon as the adsorbed layer acquires certain, essentially still unknown, attributes. Furthermore, the implanted material must not corrode or disintegrate, releasing particles which could themselves engender an immune response. In fact, a truly biocompatible material needs to have an adaptive capability, and should thus qualify for the appellation ‘smart’, just as biological tissue itself does.

C2.14.7.2 KINETICS OF BIOPOLYMER ADSORPTION

The presence of the solid surface imposes new conditions onto the disposition of reactants, compared with the homogeneous case (section C2.14.4.1). Adsorption is often observed to approach a plateau, yet is irreversible with respect to dilution. The plateau must therefore arise because no more space is available for adsorption, rather than through a dynamic adsorption–desorption equilibrium and it can be inferred that the dissolved biopolymer does not adsorb to its preadsorbed congeners. This behaviour is by no means universal: it has been proposed that the plaques associated with spongiform encephalopathies arise through the native, normally soluble PrP^C protein being partially denatured upon contact with a surface to become the pathogenic PrP^{Sc} form, to which the PrP^C can adhere to form multilayers.

A common feature of biopolymer adsorption is that its rate is usually one to three orders of magnitude smaller than the diffusion-limited rate to a perfect sink:

$$(dv/dt)_{\max} = Dc_b/\delta \quad (\text{C2.14.58})$$

where C_b is the bulk dissolved concentration and δ the thickness of the diffusion boundary layer [145] (this can be quickly ascertained by drawing a tangent to a plot of v versus t at $t \rightarrow 0$ and comparing its slope with the right-hand side of equation C2.14.58. This implies the existence of an energy barrier characterized by a rate coefficient k_a (equations (C2.14.56) and (C2.14.57)). For adsorption to small particles (colloidal minerals, vesicles etc.) of radius R , the effective δ is given by [146]:

$$(\text{C2.14.59})$$

$$1/\delta_{\text{eff}} = 1/R + 1/\delta.$$

Even cursory inspection of typical (v,t) data shows that the evolution does not follow the single exponential approach to saturation implied by, for example, (equation C2.14.22) with initial concentrations $\lambda_0 \gg s_0$. Such data are sometimes described as ‘biphasic’, and one encounters attempts to fit and interpret them with two exponentials, even though there does not seem to be any theoretical justification for doing so. The basic kinetics of adsorption are described by:

$$dv/dt = k_a c_v \phi \tag{C2.14.60}$$

-34-

where c_v is the concentration in the vicinity of the surface and depends on c_b , k_a and hydrodynamic factors, i.e. D and δ [87], and ϕ is the fraction of the surface available for adsorption. The familiar Langmuir expression $\phi = 1 - \theta$, where θ is the fraction of surface occupied, introduced to describe the adsorption of gases onto metals, whose surface is assumed to consist of discrete, noninteracting sites larger than the ligand, indeed predicts an exponential approach to saturation. For most cases of practical interest in biocompatibility, however, the Langmuir assumptions are invalid: the surface is a continuum. The adsorption of one particle creates an *exclusion zone* around it, within which the centre of another particle may not adsorb, since particles cannot overlap. The exclusion zone has twice the diameter of the particle and its area is quadruple that of the particle, hence, for small θ , $\phi = 1 - 4\theta$. As coverage increases, exclusion zones begin to overlap and the factor -4θ overcompensates for the loss of available area; for the overlap of two exclusion zones, a factor proportional to θ^2 , and for the overlap of three exclusion zones, a factor proportional to θ^3 , must be added back [147], the proportionality constants depending on the shapes of the particles and whether lateral diffusion or desorption is allowed. This problem of random sequential adsorption (RSA) has been solved exactly in one dimension (useful for describing proteins adsorbing onto DNA) [148], and accurate interpolation formulae, for two dimensions are now available [149], which have been shown to describe experimental adsorption data very well [150]. The RSA process has infinite memory and the configurations generated are quite different in many respects from their equilibrium counterparts [151].

RSA has turned out to be an extremely useful formalism for making structural inferences from adsorption kinetics. Where pure random sequential adsorption is observed, the area a occupied per molecule can be determined (note that a is the constant of proportionality between θ and v). If this area depends on c_b , conformational rearrangement leading to spreading is inferred and its kinetics and magnitude can be determined [152]. Nucleation and growth of two dimensional islands also have a characteristic kinetic signature [153]. Occasionally, Langmuir adsorption is observed in protein adsorption onto a continuum [154], unambiguously implying that clustering or crystallization of the adsorbed biopolymer takes place, thereby annihilating the exclusion zones. Both kinetic parameters and the crystal unit cell dimensions can then be determined [154].

A simple mapping enables the RSA formalism to be applied to binding of a ligand L to receptors R embedded in a surface (the RSA-random site (RSA-RS) model) [149],

-35-

C2.14.8 BIOLOGICAL INFORMATION

Information, like energy, is an irreducible concept, but is surprisingly rarely mentioned in textbooks of biophysical chemistry (whereas bioenergetics has developed into a distinct field of its own). Yet information is maybe even more germane than energy to the very essence of life, starting with the DNA which, it is often stated, encodes our organism, initially via the amino acid sequence which encodes protein structure, and so on. The neologism ‘bioinformatics’ usually denotes the analysis of DNA sequences, in particular the comparison of sequences derived from different organisms but apparently encoding the same protein. That this is a very difficult task is well illustrated by the analysis of transcriptional promoter sequences (to which a protein must bind in order for transcription to be initiated), which have few discernible common features which could be used to identify them.

To get some flavour of the magnitude of the problem, consider that the familiar bacterium *E. coli* contains almost five million nucleic acid base pairs, which encode about 4000 genes and a comparable number of promoters. Eukaryotic organisms have much more DNA, of which only a small proportion appears to code for proteins. It is a puzzle that non-coding DNA (introns and intra-gene sequences) shows long range correlations whereas coding DNA does not [156].

Perhaps the central question in this field is whether the genome specifies the construction of the organism in a rather deterministic (and hierarchical, according to the homeotic gene concept [157]) fashion, or whether the genes merely specify rules with which the organism can be constructed, more in the spirit of the stigmergic building referred to previously [45], or the brain, in which it appears that connexions between specific cells are not preprogrammed, but grow according to an algorithm given genetically to select certain favourable system structures [158]; moving back a stage further, the genes could merely specify how to construct an algorithm for specifying the construction. It is actually difficult to establish the existence of a real command structure, and it is consequently legitimate to enquire whether so-called master genes are merely akin to the king who daily ordered the sun to set, and in the morning to rise again, and was considered by most of his subjects to be an omnipotent autocrat whose orders were infallibly obeyed. What is established is that genes are powerless in isolation: they specify protein, but the realization of the specifications (and indeed the synthesis of the genes themselves) itself involves (other) proteins. The scheme of organization thus appears to be heterarchical rather than hierarchical, rather like the brain [159].

Current views of metabolic regulation are largely inspired by the *lac* operon of *E. coli*, which was comprehensively described almost 40 years ago [160] and was for many years thereafter the sole exemplar discussed in textbooks. Much work has been dedicated to identifying and characterizing the molecules involved, but how all these elements fit together remains elusive [161], and the need to move beyond the treatment of individual elements in isolation, towards concepts such as distributive control and supramolecular organization has been stressed [162]. Systems theory [12, 163, 164] was originally developed to render tractable this jungle of complexity, but it no longer seems to be part of mainstream research in the field, possibly because of the insufficiently close collaboration of the different disciplines which would have been needed to ensure its successful application to biological problems [165]. More recently an approach based on analogies between genetic and electric circuits seems capable of yielding valuable insights [166, 167, 168].

Biological information is also concerned with the analysis of biological messages and their import. The fundamental premise of the protein-folding problem section C2.14.2.2 is that the full three-dimensional arrangement of the protein molecule can be predicted, given only the amino acid sequence, together with the solvent composition, temperature and pressure. One test of the validity of this premise is to compare the information content of the sequence with the information contained in the structure [169]. The former can be obtained from Shannon's formula:

-36-

$$H = - \sum_{r=1}^R p_r \log_2 p_r \quad (\text{C2.14.64})$$

where p_r is the probability of occurrence of the r th amino acid, and the latter can be quantified via the algorithmic complexity [169]; they are approximately 2 and 0.5 bits per amino acid respectively, i.e. the sequence contains somewhat more information than is required by the structure. This is a puzzle. Clearly discussion of information content needs to be complemented by an appraisal of the value of the information [65]. The discrepancy between the two measures hints at the protein indeed being an object whose structure in part reflects noise expressed macroscopically [66, 158].

Analysis of the global statistics of protein sequences has recently allowed light to be shed on another puzzle, that of the origin of extant sequences [170]. One proposition is that proteins evolved from random amino acid chains, which predict that their length distribution is a combination of the exponentially distributed random variable giving the intervals between start and stop codons, and the probability that a given sequence can fold up to form a compact

structure, which increases with sequence length. An alternative view is that modern proteins evolved from a small set of ‘starter’ sequences, but this does not provide a simple, natural explanation for the observed extant distribution in the way that the first proposition does [170].

Reference to gene chips as a tool for investigating the expression of messenger RNA (mRNA), the precursor to protein synthesis, has already been made [section C2.14.1](#). The alternative is to extract all the proteins in a cell and separate them (according to molecular weight and isoelectric point) using two-dimensional gel electrophoresis [171], after which their abundances may be quantified. This is a much more onerous procedure than the gene chip method, and has its own drawbacks, such as poor recovery of membrane proteins, but on the other hand the relationship between mRNA and protein abundance is complex, nor can the gene chip take account of the numerous post-translation modifications such as glycosylation.

The collection of proteins expressed in a cell is called its *proteome* (cf the genome, the collection of genes). Much effort has been expended on identifying the individual proteins separated by 2D gel electrophoresis, but this is rather like discussing an author’s use of particular words, for example when the authorship of a work is disputed: as Yule has pointed out in that context, such discussion gives not the slightest notion as to what the vocabulary is like *as a whole* [172]. It is therefore of great interest to examine the properties of the entire proteome, and such an investigation has yielded the curious result that the distribution of rates of protein synthesis p_r (or protein abundances) in prokaryotes follows the simple canonical law (scl):

$$p_r = P(r + \rho)^{-1/\theta} \tag{C2.14.65}$$

where r is the rank, and P , ρ and θ are parameters, remarkably well [173]. This might be regarded merely as a useful exercise in data reduction were it not for the fact that the simple canonical law is precisely the distribution expected for a communication system minimizing its energy expenditure, while constrained by the given amount of information which has to be conveyed, word by word.

The distribution also has a certain information content which can be calculated using equation (C2.14.64), and it turns

out that this is a rather low number: typically around 8 bits/protein [173], far less than the information contained in the sequence, let alone in the structure [169]. Does this mean that much of the sequence-structure information is actually irrelevant to the protein, which has ‘merely’ to fulfil a certain specified function? For an enzyme or a protein which has to recognize a substrate or a binding site, the conformation and side-chain chemistry of the catalytic or binding site may be far more crucial than the rest of the protein. But one should also bear in mind that proteins do not exist in isolation. Their expression (and folding, for about 10% of proteins) requires the presence of other proteins, and nearly all proteins interact with others, either to build up structures, or in subtle and complex signalling pathways. Hence the information content and value of an individual protein cannot be assessed in isolation, but must be evaluated in the context of the entire repertoire, just as the genome is not a mosaic of individual genes, each coding for a single protein or attribute, but a highly complex interconnected, possibly even heterarchical, network [174]; and just as a little ‘no’ in a long paragraph could be absolutely crucial to the import of the entire message.

C2.14.9 CONCLUDING REMARKS

The study of living organisms, although traditionally reserved for the biologist, is a field in which biology, chemistry and physics must work together in order to make real advances. Ageno [16] has further emphasized that to view this development as the ‘conquest’ of one discipline by another is quaintly outmoded: the classification of this or that field as part of a particular discipline is rather arbitrary and mainly of historical interest. The fusion of the work of the biologist, chemist and physicist is irremediable; one may call this fusion biological physics, or, more comprehensively, biophysical chemistry; a toolbox with the help of which a mathematical description of biological

phenomena can be given. Sometimes these descriptions will be caricatures of reality, but one hopes that they at least fulfil their purpose of capturing its essence.

It has become fashionable to prefix the names of disciplines with ‘bio’, as in biophysics, bioinformatics and so on, giving the impression that in order to deal with biological systems, a different kind of physics, or information science, is needed. But there is no imperative for this necessity. Biological systems are often very complex and compartmentalized, and their scaling laws may be different from those familiar in inanimate systems, but this merely means that different emphases from those useful in dealing with large uniform systems are required, not that a separate branch of knowledge should necessarily be developed.

Experimental work in the field is often burdened by tension between the small amounts of pure materials generally available for investigation (a problem compounded by their instability) and the complexity of biological systems and hence the large number of possible interpretations of data, which calls for more detailed investigation than otherwise. The frequently encountered, seemingly poor, reproducibility of experiments with biological materials and systems also calls for more experiments than would suffice for simpler systems. This ‘irreproducibility’ might well be a manifestation of the fact that the measured phenomena are the result of multiplicative rather than additive processes. For the latter, the sum and its distribution converge rapidly enough to their asymptotic values; for the former, a principle comparable to the central limit theorem is lacking; in fact the average value of the product diverges exponentially from the most probable value as the number of random variables contributing to it increases [175].

REFERENCES

- [1] Thompson D'A W 1942 *On Growth and Form* (Cambridge: Cambridge University Press)
- [2] Anderson P W 1972 More is different *Science* **177** 393–5
- [3] Manenti S, Sorokine O, Van Dorsselaer A and Taniguchi H 1992 Affinity purification and characterization of myristoylated alanine-rich protein kinase C substrate (MARCKS) from bovine brain *J. Biol. Chem.* **267** 22 310–15
- [4] Schuck P 1996 Kinetics of ligand binding to receptors immobilized in a polymer matrix, as detected with an evanescent wave biosensor. I. A computer simulation of the influence of mass transport *Biophys. J.* **70** 1230–49
- [5] Hillman H 1991 *The Case for New Paradigms in Cell Biology and in Neurobiology* (Lewiston: Edwin Mellen Press)
- [6] Welch E R (ed) 1986 *The Fluctuating Enzyme* (New York: Wiley)
- [7] Frauenfelder H 1984 From atoms to biomolecules *Helv. Phys. Acta* **57** 165–87
- [8] Rocco M, Infusini E, Daga M G, Gogioso L and Cuniberti C 1987 Models of fibronectin *EMBO J.* **6** 2343–9
- [9] Kempner E S and Miller J H 1968 The molecular biology of *Euglena gracilis* IV. Cellular stratification by centrifuging *Exp. Cell. Res.* **51** 141–9
- [10] Kempner E S and Miller J H 1968 The molecular biology of *Euglena gracilis* V. Enzyme localization *Exp. Cell. Res.* **51** 150–6
- [11] Ramsden J J 1993 Review of new experimental methods for investigating random sequential adsorption *J. Stat. Phys.* **73** 853–77
- [12] von Bertalanffy L 1993 *Théorie générale des systèmes* (Paris: Dunod)
- [13] Cantor C R and Schimmel P R 1980 *Biophysical Chemistry* Parts I, II and III (San Francisco: Freeman)
- [14] Hodgkin A L 1971 *The Conduction of the Nervous Impulse* (Liverpool: University Press)
- [15] Markin V S, Pastushenko V F and Chizmadzhev Y A 1987 *Theory of Excitable Media* (New York: Wiley)
- [16] Ageno M 1967 Linea di ricerca in fisica biologica *Accad. naz Lincei* **102** 3–58
- [17] Landau E M, Rummel G, Cowan-Jacob S W and Rosenbusch J P 1997 Crystallization of a polar protein and small molecules from the aqueous compartment of lipidic cubic phases *J. Phys. Chem. B* **101** 1935–7
- [18] Petsko G A and Ringe D 1984 Fluctuations in protein structure from x-ray diffraction *A. Rev. Biophys. Bioengng.* **13** 331–71

- [19] Cacace M G, Landau E M and Ramsden J J 1997 The Hofmeister series: salt and solvent effects on interfacial phenomena *Q. Rev. Biophys.* **30** 241–78
- [20] Henderson R 1995 The potential and limitations of neutrons, electrons and x-rays for atomic resolution microscopy of unstained biological molecules *Q. Rev. Biophys.* **28** 171–93
- [21] Güntert P 1998 Structure calculation of biological macromolecules from nmr data 1998 *Q. Rev. Biophys.* **31** 145–237
- [22] Crippen G M 1977 A novel approach to calculation of conformation: distance geometry *J. Comput. Phys.* **24** 96–107
- [23] Metzler W J, Hare D R and Pardi A 1989 Limited sampling of conformational space by the distance geometry algorithm: implications for structures generated from NMR data *Biochemistry* **28** 7045–52
- [24] Schuster T M and Toedt J M 1996 New revolutions in the evolution of analytical ultracentrifugation *Curr. Opinion Cell. Biol.* **6** 650–8
- [25] Schwarz G and Savko P 1982 Structural and dipolar properties of the voltage-dependent pore former alamthycin in octanol/dioxane *Biophys. J.* **39** 211–19
-

-39-

- [26] Ermolina I V, Fedotov V D and Feldman Yu D 1998 Structure and dynamic behaviour of protein molecules in solution *Physica A* **249** 347–52
- [27] Ramsden J J 1998 Towards zero-perturbation methods for investigating biomolecular interactions *Colloids Surfaces A* **141** 287–94
- [28] Schwarz G 1996 Peptides at lipid bilayers and at the air/water interface *Ber. Bunsenges. Phys. Chem.* **100** 999–1003
- [29] Goldanskii V I and Krupyanski Y F 1989 Protein and protein-bound water dynamics studied by Rayleigh scattering of Mössbauer radiation (RSMR) *Q. Rev. Biophys.* **22** 39–92
- [30] Lifschitz I M, Grosberg A Yu and Khokhlov A R 1978 Some properties of the statistical physics of polymer chains with volume interaction *Rev. Mod. Phys.* **50** 683–713
- [31] Anderson P W 1978 The concept of frustration in spin glasses *J. Less-Common Metals* **62** 291–4
- [32] Bryngelson J D, Onuchic J N, Succi N D and Wolynes P G 1995 Funnels, pathways, and the energy landscape of protein folding: a synthesis *Proteins* **21** 167–95
- [33] Simon I 1985 Investigation of protein refolding: a special feature of native structure responsible for refolding ability *J. Theor. Biol.* **113** 703–10
- [34] Fernández A 1990 Glassy kinetic barriers between conformational substates in RNA *Phys. Reine Lett.* **64** 2328–233
- [35] von Helmholtz H 1887 Über die physikalische Bedeutung des Princips der kleinsten Wirkung *J. Reine Angew. Math.* **100** 137–66, 213–222
- [36] Fernández A and Cendra H 1996 *In vitro* RNA folding: the principle of sequential minimization of entropy loss at work *Biophys. Chem.* **58** 335–9
- [37] Watson J D and Crick F H C 1953 A structure for deoxyribose nucleic acid *Nature* **171** 737–8
- [38] Watson J D and Crick F H C 1953 Genetical implications of the structure of deoxyribonucleic acid *Nature* **171** 964–7
- [39] Ramachandran G N and Sasisekharan V 1968 Conformation of polypeptides and proteins *Adv. Prot. Chem.* **23** 283–438
- [40] Fernández A and Colubri A 1998 Microscopic dynamics from a coarsely defined solution to the protein folding problem *J. Math. Phys.* **39** 3167–87
- [41] Damjanovich S, Gáspár R and Pieri C 1997 Dynamic receptor superstructures at the plasma membrane *Q. Rev. Biophys.* **30** 67–106
- [42] Forgacs G 1995 On the possible role of cytoskeletal filamentous networks in intracellular signalling *J. Cell. Sci.* **108** 2131–43
- [43] Decker H and Sterner R 1990 Hierarchien in der Struktur und Funktion von sauerstoffbindenden Proteinen *Naturwissenschaften* **77** 561–8
- [44] Ghiradella H 1991 Light and color on the wing: structural colors in butterflies and moths *Appl. Opt.* **30** 3492–500
- [45] Theraulaz G and Bonabeau E 1995 Coordination in distributed building *Science* **269** 686–8
- [46] Porter G 1983 Transfer and storage of chemical and radiation potential *J. Chem. Soc. Faraday Trans. 2* **79** 473–82
- [47] Szabó A, Shoup D, Northrup S H and McCammon J A 1982 Stochastically gated diffusion-influenced reactions *J. Chem. Phys.* **77** 4484–93

- [48] Zwanzig R 1990 Rate processes with dynamical disorder *Acc. Chem. Res.* **23** 148–52
- [49] Benzinger T H 1971 Thermodynamics, chemical reactions and molecular biology *Nature* **229** 100–2
- [50] Weber G 1995 van't Hoff revisited: enthalpy of association of protein subunits *J. Phys. Chem.* **99** 1052–9
- [51] Planck M 1897 *Vorlesungen über Thermodynamik* (Leipzig: von Veit)
- [52] Naghibi H, Tamura A and Sturtevant J M 1995 Significant discrepancies between van't Hoff and calorimetric enthalpies *Proc. Natl Acad. Sci. USA* **92** 5597–9

-40-

- [53] Haynes C A and Norde W 1995 Structural stabilities of adsorbed proteins *J. Colloid Interface Sci.* **169** 313–28
- [54] Alexander S 1998 What is a solid? *Physica A* **249** 266–75
- [55] Gurney R W 1953 *Ionic Processes in Solution* (New York: McGraw-Hill)
- [56] Schwarz G 1970 Cooperative binding to linear biopolymers *Eur. J. Biochem.* **12** 442–53
- [57] Schwarz G, Stankowski S and Rizzo V 1986 Thermodynamic analysis of incorporation and aggregation in a membrane *Biochim. Biophys. Acta* **861** 141–51
- [58] Ramsden J J and Máté M 1998 Kinetics of monolayer particle deposition *J. Chem. Soc. Faraday Trans.* **94** 783–8
- [59] Valitutti, Müller S, Cella M, Padovan E and Lanzavecchia A 1995 Serial triggering of many T-cells by a few peptide-MHC complexes *Nature* **375** 148–51
- [60] Zweig H J 1961 Theoretical considerations of the quantum efficiency of photographic detectors *J. Opt. Soc. Am.* **51** 310–19
- [61] Ramsden J J 1994 Experimental methods for investigating protein adsorption kinetics at surfaces *Q. Rev. Biophys.* **27** 41–105
- [62] Ramsden J J 1998 Biomimetic protein immobilization using lipid bilayers *Biosensors Bioelectronics* **13** 593–8
- [63] Ageno M 1992 *La 'Macchina' Batterica* (Rome: Lombardi)
- [64] Blumenfeld L A 1981 *Problems of Biological Physics* (Berlin: Springer)
- [65] Chernavsky D S 1990 Synergetics and information *Matematika Kibernetika* **5** 3–42
- [66] Shaw R 1981 Strange attractors, chaotic behaviour, and information flow *Z. Naturf. a* **36** 80–112
- [67] Derényi I and Vicsek T 1998 Realistic models of biological motion *Physica A* **249** 397–406
- [68] Palmer R E 1982 Broken ergodicity *Adv. Phys.* **31** 669–735
- [69] Landau L D and Lifschitz E M 1980 *Statistical Physics* 3rd edn (Oxford: Pergamon) part 1
- [70] Anderson P W and Stein D 1987 Broken symmetry, emergent properties, dissipative structures, life *Self-Organizing Systems* ed F E Yates (New York: Plenum) pp 445–57
- [71] Averbukh I Sh, Blumenfeld L A, Kovarsky V A and Perelman N F 1986 A model of the mechanism of enzyme action in terms of protein conformational relaxation *Biochim. Biophys. Acta.* **873** 290–6
- [72] Blumenfeld L A, Burbajev D S and Davydov R M 1986 Processes of conformational relaxation in enzyme catalysis *The Fluctuating Enzyme* ed E R Welch (New York: Wiley) pp 369–402
- [73] Dasgupta S, Copeland R A and Spiro T G 1986 Ultraviolet Raman spectroscopy indicates fast ($\ll 7$ ns) R→T-like motion in haemoglobin *J. Biol. Chem.* **261** 10 960–2
- [74] Hofrichter J, Sommer J H, Henry E R and Eaton W A 1983 Nanosecond absorption spectroscopy of haemoglobin *Proc. Natl Acad. Sci. USA* **80** 2235–9
- [75] Ishijima A, Kojima H, Funatsu T, Tokunaga M, Higuchi H, Tanaka H and Yanagida T 1998 Simultaneous observation of individual ATPase and mechanical events by a single myosin molecule during interaction with actin *Cell* **92** 161–71
- [76] Romanovsky J M, Stepanova N V and Chernavsky D S 1974 *Kinetische Modelle in der Biophysik* (Jena: Gustav Fischer)

- [77] von Smoluchowski M 1917 Versuch einer mathematischen Theorie der Koagulationskinetik kolloider Lösungen *Z. Phys. Chem.* **92** 129–68
- [78] Clegg J S 1984 Properties and metabolism of the aqueous cytoplasm and its boundaries *Am. J. Physiol.* **246** R133–R151
- [79] Ramsden J J and Grätzel M 1986 Formation and decay of methyl viologen radical cation dimers on the surface of colloidal CdS *Chem. Phys. Lett.* **132** 269–72
-

-41-

- [80] Rényi A 1953 Kémiai reakciók tárgyalása a sztochasztikus folyamatok elmélete segítségével *Magy. Tud. Akad. Mat. Kut. Int. Közl.* **2** 83–101
- [81] Zeldovich Ya B 1977 The role of thermal fluctuations of concentration in the kinetics of bimolecular reactions *Elektrokhimia* **13** 677–9
- [82] Zeldovich Ya B and Ovchinnikov A A 1978 Asymptotic form of the approach to equilibrium and concentration fluctuations *JETP Lett.* **26** 440–2
- [83] Oshanin G S, Ovchinnikov A A and Burlatsky S F 1989 Fluctuation-induced kinetics of reversible reactions *J. Phys. A : Math. Gen.* **22** L977–L982
- [84] Kurrat R, Ramsden J J and Prenosil J E 1994 Kinetic model for serum albumin adsorption: experimental verification *J. Chem. Soc. Faraday Trans.* **90** 587–90
- [85] Kurrat R, Prenosil J E and Ramsden J J 1997 Kinetics of human and bovine serum albumin adsorption at silica–titania surfaces *J. Colloid Interface Sci.* **185** 1–8
- [86] Talbot J 1996 Time dependent desorption: a memory function approach *Adsorption* **2** 89–94
- [87] Ramsden J J 1998 Kinetics of protein adsorption *Biopolymers at Interfaces* ed M Malmsten (New York: Dekker)
- [88] Palmer R G, Stein D L, Abrahams E and Anderson P W 1984 Models of hierarchically constrained dynamics for glassy relaxation *Phys. Rev. Lett.* **53** 958–61
- [89] Turing A M 1952 The chemical basis of morphogenesis *Phil. Trans. R. Soc. B* **237** 5–72
- [90] Forgacs G, Newman S A, Obukhov S P and Birk D E 1991 Phase transition and morphogenesis in a model biological system *Phys. Rev. Lett.* **67** 2399–402
- [91] Wolfram S 1983 Statistical mechanics of cellular automata *Rev. Mod. Phys.* **55** 601–44
- [92] Luthi P O, Preiss A, Chopard B and Ramsden J J 1998 A cellular automaton model for neurogenesis *Drosophila Physica D* **118** 151–60
- [93] Howard J 1997 Molecular motors *Nature* **389** 561–7
- [94] Namba K and Vonderviszt F 1997 Molecular architecture of the bacterial flagellum *Q. Rev. Biophys.* **30** 1–65
- [95] Caplan S R and Kara-Ivanov M 1993 The bacterial flagellar motor *Int. Rev. Cytol.* **147** 97–164
- [96] Baumeister W, Walz J, Zühl F and Seemüller E 1998 The proteasome: paradigm of a self-compartmentalizing protease *Cell* **92** 367–80
- [97] Leibler S 1996 Collective phenomena in mitosis *Physics of Biomaterials: Fluctuations, Self-assembly and Evolution* (Dordrecht: Kluwer) pp 135–51
- [98] Riste T and Sherrington D (eds) 1996 *Physics of Biomaterials: Fluctuations, Self-assembly and Evolution* (Dordrecht: Kluwer)
- [99] Mitchison T and Kirschner M 1984 Dynamic instability of microtubule growth *Nature* **213** 237–42
- [100] Holy T E and Leibler S 1994 Dynamic instability of microtubules as an efficient way to search in space *Proc. Natl Acad. Sci. USA* **91** 5682–5
- [101] Ishiwata S, Nishizaka T, Kato H, Tadakuma H, Iga T, Suzuki N, Miyata H and Kinoshita K 1996 Microscopic analysis of the nature of forces in a single actomyosin motor and its assemblage *Neurokhimiya* **13** 305–13
- [102] Yanagida T, Harada Y and Ishijima A 1993 Nanomanipulation of actomyosin molecular motors *in vitro*: a new working principle *TIBS* **18** 319–23

[103] Nishizaka T, Miyata H, Yoshikawa H, Ishiwata S and Kinosita K 1995 Unbinding force of a single motor molecule of muscle measured using optical tweezers *Nature* **377** 251–4

-42-

- [104] Vale R D, Funatsu T, Pierce D W, Romberg L, Harada Y and Yanagida T 1996 Direct observation of single kinesin molecules moving along microtubules *Nature* **380** 451–3
- [105] Noji H, Yasuda R, Yoshida M and Kinosita K 1997 Direct observation of the rotation of F_1 -ATPase *Nature* **386** 299–302
- [106] Gray B F 1975 Reversibility and biological machines *Nature* **253** 436–7
- Gray B F 1975 *Nature* **257** 72
- [107] Funatsu T, Harada Y, Tokunaga M, Saito K and Yanagida T 1995 Imaging of single fluorescent molecules and individual ATP turnovers by single myosin molecules in aqueous solution *Nature* **374** 555–9
- [108] Fischer E 1894 Einfluss der Configuration auf die Wirkung der Enzyme *Ber. dt Chem. Ges.* **27** 2985–93
- [109] Baker E N and Hubbard R E 1984 Hydrogen bonding in globular proteins *Prog. Biophys. Molec. Biol.* **44** 97–179
- [110] Fernández A 1991 Functional metastable structures in RNA replication *Physica A* **176** 499–513
- [111] Böhm H-J 1994 The development of a simple empirical scoring function to estimate the binding constant for a protein–ligand complex of known three-dimensional structure *J. Comp.-Aided Mol. Design* **8** 243–56
- [112] Ramsden J J and Dreier J 1996 Kinetics of the interaction between DNA and the type IC restriction enzyme StyR 124/3I *Biochemistry* **35** 3746–53
- [113] Adam G and Delbrück M 1968 Reduction of dimensionality in biological diffusion processes *Structural Chemistry and Molecular Biology* ed A Rich and N Davidson (San Francisco: Freeman)
- [114] Rabinowitch E 1937 Collision, coordination, diffusion and reaction velocity in condensed systems *Trans. Faraday Soc.* **33** 1225–33
- [115] Luthi P O, Ramsden J J and Chopard B 1997 The role of diffusion in irreversible deposition *Phys. Rev. E* **55** 3111–15
- [116] Caldin E F, de Forest L and Queen A 1990 Steric and repeated collision effects in diffusion-controlled reactions in solution *J. Chem. Soc. Faraday Trans.* **86** 1549–54
- [117] Bruinsma R 1996 Physical aspects of the adhesion of leukocytes *Physics of Biomaterials: Fluctuations, Self-Assembly and Evolution* (Dordrecht: Kluwer) pp 61–101
- [118] Percus J K, Percus O E and Perelson A S 1993 Predicting the size of the T-cell receptor and antibody combining region from consideration of efficient self–nonself discrimination *Proc. Natl Acad. Sci. USA* **90** 1691–5
- [119] Lancet D, Sadovsky E and Seidemann E 1993 Probability model for molecular recognition in biological receptor repertoires: significance to the olfactory system *Proc. Natl Acad. Sci. USA* **90** 3715–19
- [120] Kozack R E, d’Mello M J and Subramaniam S 1995 Computer modeling of electrostatic steering and orientational effects in antibody–antigen association *Biophys. J.* **68** 807–14
- [121] Madura J D, Davis M E, Gilson, M K, Wade R C, Luty B A and McCammon J A 1994 Biological applications of electrostatic calculations and Brownian dynamics simulations *Rev. Comput. Chem.* **5** 229–67
- [122] Northrup S H and Erickson H P 1992 Kinetics of protein–protein association explained by Brownian dynamics computer simulation *Proc. Natl Acad. Sci. USA* **89** 3338–42
- [123] Turbadar T 1959 Complete adsorption of light by thin metal films *Proc. Phys. Soc.* **73** 40–4
- [124] Schaaf P, Dejardin Ph and Schmitt A 1985 Réflectométrie appliquée aux interfaces diffuses: possibilités et limites de la technique *Rev. Phys. Appl.* **20** 631–40
- [125] Guemouri L, Ogier J and Ramsden J J 1998 Optical properties of protein monolayers during assembly *J. Chem. Phys.* **109**–8
- [126] de Feijter J A, Benjamins J and Veer F A 1978 Ellipsometry as a tool to study the adsorption behaviour of polymers at the air–water interface *Biopolymers* **17** 1759–72
-

-43-

- [127] Ball V and Ramsden J J 1998 Buffer dependence of refractive index increments of protein solutions *Biopolymers* **46** 489–92
- [128] Maternaghan T J and Ottewill R H 1974 An ellipsometric study of the adsorption of gelatin on silver bromide *J. Phot. Sci.* **22** 279–85
- [129] Stamm Ch and Lukosz W 1994 Integrated optical difference interferometer as biochemical sensor *Sensors Actuators B* **18–19** 183–7
- [130] Ramsden J J, Li S-Y, Heinzle E and Prenosil J E 1994 Kinetics of adhesion and spreading of animal cells *Biotechnol. Bioengng* **43** 939–45
- [131] Ramsden W 1903 Separation of solids in the surface layers of solutions and ‘suspensions’ *Proc. R. Soc.* **72** 156–64
- [132] Sinanoğlu O 1981 What size cluster is like a surface? *Chem. Phys. Lett.* **81** 188–90
- [133] van Oss C J, Giese R F and Wu W 1997 On the predominant electron donicity of polar solid surfaces *J. Adhesion* **63** 71–88
- [134] van Oss C J 1996 *Forces Interfaciales en Milieux Aqueux* (Paris: Masson)
- [135] Attard P 1996 Electrolytes and the electric double layer *Adv. Chem. Phys.* **92** 1–159
- [136] Spielman L A and Friedlander S K 1974 Role of the electrical double layer in particle deposition by convective diffusion *J. Colloid. Interface. Sci.* **46** 22–31
- [137] Sinanoğlu O 1981 Microscopic surface tension down to molecular dimensions and microthermodynamic surface areas of molecules or clusters *J. Chem. Phys.* **75** 463–8
- [138] Kondo A, Oku S and Higashitani K 1991 Structural changes in protein molecules adsorbed on ultrafine silica particles *J. Colloid Interface Sci.* **143** 214–21
- [139] Zoungrana T, Findenegg G H and Norde W 1997 Structure, stability and activity of adsorbed enzymes *J. Colloid Interface Sci.* **190** 437–48
- [140] Steadman B L, Thompson K C, Middaugh C R, Matsuno K, Vrona S, Lawson E Q and Lewis R V 1992 The effects of surface adsorption on the thermal stability of proteins *Biotech. Bioengng.* **40** 8–15
- [141] Haynes C A and Norde W 1994 Globular proteins at solid/liquid interfaces *Colloids Surf. B* **2** 517–66
- [142] Kondo A and Higashitani K 1992 Adsorption of model proteins with wide variation in molecular properties on colloidal particles *J. Colloid Interface Sci.* **150** 344–51
- [143] Kurrat R, Wälivaara B, Marti A, Textor M, Tengvall P, Ramsden J J and Spencer N D 1998 Plasma protein adsorption on titanium *Colloids Surf. B* **11** 187–201
- [144] Davie E W, Fujikawa K and Kisiel W 1991 The coagulation cascade: initiation, maintenance and regulation 1991 *Biochemistry* **30** 10 363–10 370
- [145] Levich V G 1962 *Physicochemical Hydrodynamics* (Englewood Cliffs, NJ: Prentice-Hall)
- [146] Spiro M and Freund P L 1983 Colloidal catalysis *J. Chem. Soc. Faraday Trans.* **1** 1649–58
- [147] Schaaf P and Talbot J 1989 Surface exclusion effects in adsorption processes *J. Chem. Phys.* **91** 4401–9
- [148] Rényi A 1958 Egy egydimenziós véletlen térkitöltési problémától *Magy. Tud. Akad. Mat. Kut. Int. Közl* **3** 109–25
- [149] Jin X, Talbot J and Wang N H L 1994 Analysis of steric hindrance effects on adsorption kinetics and equilibria *AIChE J.* **40** 1685–96
- [150] Ramsden J J 1993 Concentration scaling of protein deposition kinetics *Phys. Rev. Lett.* **71** 295–8
- [151] Evans J W 1993 Random and cooperative sequential adsorption 1993 *REv. Mod. Phys.* **65** 1281–1329
- [152] Van Tassel P R, Guemouri L, Ramsden J J, Tarjus G, Viot P and Talbot J 1998 A model for the influence of conformational change on protein adsorption kinetics *J. Colloid Interface Sci.* **207** 317–23

-
- [153] Csúcs G and Ramsden J J 1998 Generalized ballistic deposition of small buoyant particles *J. Chem. Phys.* **109** 779–81
- [154] Ramsden J J, Bachmanova G I and Archakov A I 1994 Kinetic evidence for protein clustering at a surface *Phys. Rev. E* **50** 5072–6

- [155] Jin X, Wang N H L, Tarjus G and Talbot J 1993 Irreversible adsorption on nonuniform surfaces: the random site model *J. Phys. Chem.* **97** 4256–8
- [156] Voss R F 1992 Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences *Phys. Rev. Lett.* **68** 3805–8
- [157] Gehring W J, Affolter M and Bürglin T 1994 Homeodomain proteins *A. Rev. Biochem.* **63** 487–526
- [158] érdi P and Barna Gy 1984 Self-organizing mechanism for the formation of ordered neural mappings *Biol. Cybernetics* **51** 93–101
- [159] Kaehr R and von Goldammer E 1998 . . . again computers and the brain *J. Mol. Electron.* **4** S31–S37
- [160] Jacob F and Monod J 1961 Genetic regulatory mechanisms in the synthesis of proteins *J. Mol. Biol.* **3** 318–56
- [161] Aon M A and Cortassa S 1995 Cell growth and differentiation from the perspective of dynamical organization of cellular and subcellular processes *Prog. Biophys. Molec. Biol.* **64** 55–79
- [162] Srere P 1994 Complexities of metabolic regulation *Trends Biochem. Sci.* **19** 519–20
- [163] Savageau M A, Voit E O and Irvine D H 1987 Biochemical systems theory and metabolic control theory: 1. Fundamental similarities and differences *Math. Biosci.* **86** 127–45
- [164] Savageau M A, Voit E O and Irvine D H 1987 Biochemical systems theory and metabolic control theory: 2. The role of summation and connectivity relationships *Math. Biosci.* **86** 147–69
- [165] Jones R W and Gray J S 1963 System theory and physiological processes *Science* **140** 461–6
- [166] McAdams H H and Shapiro L 1995 Circuit simulation of genetic networks *Science* **269** 650–6
- [167] Hlavacek W S and Savageau M A 1995 Subunit structure of regulator proteins influences the design of gene circuitry: analysis of perfectly coupled and completely uncoupled circuits *J. Mol. Biol.* **248** 739–55
- [168] Yuh C-H, Bolouri H and Davidson E H 1998 Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene *Science* **279** 1986–02
- [169] Dewey T G 1997 Algorithmic complexity and thermodynamics of sequence-structure relationships in proteins *Phys. Rev. E* **56** 4545–52
- [170] White S H 1994 The evolution of proteins from random amino acid sequences II. Evidence from the statistical distributions of the lengths of modern protein sequences *J. Mol. Evolution* **38** 383–94
- [171] VanBogelen R A and Olson E R 1995 Application of two-dimensional protein gels in biotechnology *Biotech. Ann. Rev.* **1** 69–103
- [172] Yule G U 1944 *Statistical Study of Literary Vocabulary* (Cambridge: Cambridge University Press)
- [173] Ramsden J J and Vohradský J 1998 Zipf-like behavior of prokaryotic protein expression *Phys. Rev. E* **58** 7777–80
- [174] Wright S 1982 Character change, speciation and the higher taxa *Evolution* **36** 427–43
- [175] Redner S 1990 Random multiplicative processes *Am. J. Phys.* **58** 267–73
- [176] Fernández A 1989 Correlation of pause sites in MDV-1 RNA replication with kinetic refolding of the growing chain *Eur. J. Biochem.* **182** 161–3
- [177] Symons M C R 1981 Water structure and reactivity 1981 *Acc. Chem. Res.* **14** 179–87
- [178] Allen L C 1975 A model for the hydrogen bond *Proc. Natl Acad. Sci. USA* **72** 4701–5

FURTHER READING

von Bertalanffy L 1993 *Théorie générale des systèmes*(Paris: Dunod)

A standard work defining the scope of systems theory

Blumenfeld L A 1981 *Problems of Biological Physics* (Berlin: Springer)

A very thoughtful and original work; one of the few to ask ‘awkward’ questions.

Cantor C R and Schimmel P R 1980 *Biophysical Chemistry* Parts I, II and III (San Francisco: Freeman)

The standard textbook of ‘classical’ biophysical chemistry. Very strong on methods for elucidating the structure of biomolecules

(Part II).

Mahler H R and Cordes E H 1971 *Biological Chemistry* 2nd edn (New York: Harper and Row)

Solid account of the molecules of life.

Musha T and Sawada Y (eds) 1994 *Physics of the Living State*(Tokyo: Ohmsha)

Covers some less-commonly discussed aspects of the field

Peyrard M (ed) 1995 *Nonlinear Excitations in Biomolecules*(Les Ulis: Editions de Physique)

Good descriptions of biological phenomena from a physico-chemical viewpoint.

van Oss C J 1996 *Forces Interfaciales en Milieux Aqueux* (Paris: Masson)

Refreshingly original approach to a topic of central importance in biology and biocompatibility.

Riste T and Sherrington D (eds) 1996 *Physics of Biomaterials: Fluctuations, Self-assembly and Evolution*(Dordrecht: Kluwer)

An excellent modern account of the exciting interface between biology, physics and chemistry.

Rosen R 1967 *Optimality Principles in Biology* (London: Butterworths)

A concise, still highly relevant account of a supremely important principle.

-1-

C2.15 Optoelectronics

William L Wilson

C2.15.1 INTRODUCTION

Optoelectronic technologies [1] encompass a wide variety of devices and structures used to generate, manipulate, detect and direct light signals. This interdisciplinary field, born at the intersection of microelectronics and optics, is poised to provide the primary building blocks of the communications and computing infrastructure of the 21st century. The development of high-bandwidth optical networks using wave division multiplexing [2], for example, requires a wide variety of components for network construction. The optical communications and optical networking revolution is being fought with fibre, optical routers, modulators and detectors. The development of these basic optoelectronic components will be key in determining the progression of communication technologies for years to come.

In this chapter we review the fundamental processes which allow us to define and control optical sources and signals. The basic mechanisms for generation, transmission and manipulation are described. A large number of detailed treatises have been published describing many of the phenomena covered here [3, 4, 5, 6 and 7]. Because of space limitations, we will offer rudimentary descriptions and insights of the subjects covered and will attempt to ensure that the references listed will allow exploration of the subject matter to whatever detail is desired. It is important to note that any description of optoelectronic technologies will have roots deep in optical physics, quantum mechanics and electromagnetic theory. Here only essential formulae are derived; detailed derivation of all the expressions can be found in the references cited. Our goal is to give the reader a flavour of the technology and hopefully to stir your imagination and interest in this exciting, evolving and rapidly expanding field.

C2.15.2 ELECTROMAGNETIC WAVES

In order to understand how light can be controlled, we must first review some of the basic properties of the electromagnetic field [8]. The electromagnetic theory of light is governed by the equations of James Clerk Maxwell. The field phenomena in free space with no sources are described by the basic set of relationships below:

$$\nabla \times \mathbf{H} = \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \quad (\text{C2.15.1})$$

$$\nabla \times \mathbf{E} = -\mu_0 \frac{\partial \mathbf{H}}{\partial t} \quad (\text{C2.15.2})$$

$$\nabla \cdot \mathbf{E} = 0 \quad (\text{C2.15.3})$$

$$\nabla \cdot \mathbf{H} = 0 \quad (\text{C2.15.4})$$

-2-

where \mathbf{E} and \mathbf{H} are the electric and magnetic fields respectively. Here the constants ϵ_0 and μ_0 are the electric permittivity and the magnetic permeability of free space.

The necessary boundary conditions required for \mathbf{E} and \mathbf{H} to satisfy Maxwell's equations give rise to the well known wave equation for the electromagnetic field:

$$\nabla^2 U - \frac{1}{c^2} \frac{\partial^2 U}{\partial t^2} = 0 \quad (\text{C2.15.5})$$

where $c = 1 / (\epsilon_0 \mu_0)^{1/2} = 3 \times 10^8 \text{ m s}^{-1}$, the speed of light in a vacuum.

This wave equation is the basis of all wave optics and defines the fundamental structure of electromagnetic theory with the scalar function U representing any of the components of the vector functions \mathbf{E} and \mathbf{H} . (Note that equation (C2.15.5) can be easily derived by taking the curl of equation (C2.15.1) and equation (C2.15.2) and substituting relations (C2.15.3) and (C2.15.4) into the results.)

Although a complete treatment of optical phenomena generally requires a full quantum mechanical description of the light field, many of the devices of interest throughout optoelectronics can be described using the wave properties of the optical field. Several excellent treatments on the quantum mechanical theory of the electromagnetic field are listed in [9].

In general, the wave equation describes the propagation characteristics of a disturbance through some transparent media, specifically the electromagnetic wave is just a subset of the physical phenomena which satisfy this relationship. The solutions of this equation with appropriate boundary conditions gives rise to all the behaviours we commonly associate with the wave properties of light. For monochromatic waves, all of the components of the electric field are harmonic functions of time and space. Electromagnetic waves are by definition *transverse*, i.e. the electric (\mathbf{E}) and magnetic (\mathbf{H}) field disturbances are orthogonal to the propagation direction (z), with the \mathbf{E} and \mathbf{H} fields orthogonal to each other (figure C2.15.1). The electric field is characterized by the amplitude (A), the wavelength (λ), the phase of the wave and the velocity of the wavefront. The plane wave is the most general and simplest example of a three-dimensional solution of the wave equation, in addition it provides a somewhat ideal input field for all optoelectronic applications. The wave has the form:

$$U(\mathbf{r}, t) = A e^{-i(\mathbf{k} \cdot \mathbf{r} - \omega t)} \quad (\text{C2.15.6})$$

where A is the amplitude of the field disturbance, \mathbf{k} is the propagation vector and has the magnitude $2\pi/\lambda$ and ω is the angular frequency, $\omega = 2\pi\nu$.

-3-

where A is the amplitude of the field disturbance, \mathbf{k} is the propagation vector and has the magnitude $2\pi/\lambda$ and ω is the angular frequency, $\omega = 2\pi\nu$.

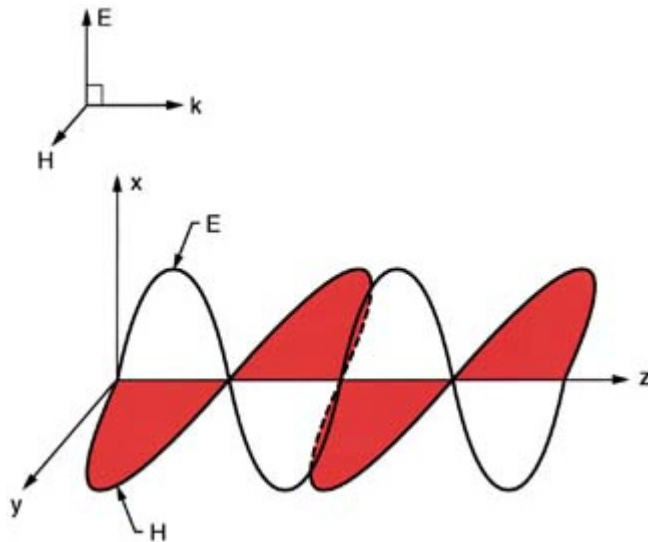


Figure C2.15.1. The transverse electromagnetic wave.

This basic equation describes waves, whose properties are related as follows:

$$\nu = \frac{c}{\lambda} \tag{C2.15.7a}$$

$$\frac{\omega}{k} = v_p = \frac{1}{\sqrt{\mu_0 \epsilon_0}} \tag{C2.15.7b}$$

where ν is the frequency of the wave and v_p is the phase velocity.

For the electromagnetic fields \mathbf{E} and \mathbf{H} the form of the waves of interest is

$$\begin{aligned} \mathbf{E}(\mathbf{r}, t) &= \mathbf{E}_0 e^{-i(\omega t - \mathbf{k} \cdot \mathbf{r})} \\ \mathbf{H}(\mathbf{r}, t) &= \mathbf{H}_0 e^{-i(\omega t - \mathbf{k} \cdot \mathbf{r})} \end{aligned} \tag{C2.15.8}$$

for many of the applications described here.

Above we described the nature of Maxwell's equations in free space in a medium, two more vector fields need to be

-4-

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} \tag{C2.15.9}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \quad (\text{C2.15.10})$$

$$\nabla \cdot \mathbf{D} = 0 \quad (\text{C2.15.11})$$

$$\nabla \cdot \mathbf{B} = 0 \quad (\text{C2.15.12})$$

with

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P} \quad (\text{C2.15.13})$$

$$\mathbf{B} = \mu_0 \mathbf{H} + \mu_0 \mathbf{M}. \quad (\text{C2.15.14})$$

These new quantities allow us to directly relate properties of the *media* to \mathbf{E} and \mathbf{H} . In essence they afford us the opportunity to quantify the field–matter interaction. The media response to the fields is described generally in terms of the polarization, \mathbf{P} and the magnetization, \mathbf{M} . (We note that in free space \mathbf{P} and $\mathbf{M} = 0$ and we recover [equation \(C2.15.1\)](#), [equation \(C2.15.2\)](#), [equation \(C2.15.3\)](#) and [equation \(C2.15.4\)](#) above.)

For isotropic media we will assume that \mathbf{P} is parallel to \mathbf{E} with the coefficient of proportionality independent of direction:

$$\mathbf{P} = \chi_e \mathbf{E} \quad (\text{C2.15.15})$$

where the constant χ_e is the electric susceptibility of the medium. The electric displacement is therefore proportional to \mathbf{E} :

$$\mathbf{D} = \varepsilon \mathbf{E} \quad (\text{C2.15.16})$$

where $\varepsilon = 1 + 4\pi\chi_e$ is the dielectric constant. This parameter relates the material properties of the media to the polarization generated through its interaction with the external field. This polarization becomes a source term in Maxwell's equations giving rise to new fields mediated *via* the material–field interaction [9]. Absorption and dispersion processes can be attributed to ε with

$$\hat{n} = \sqrt{\varepsilon} = n + iK \quad (\text{C2.15.17})$$

being the complex refractive index, where the real part is related to dispersive properties of the media and K , the absorption coefficient, is determined by the imaginary part of the polarization [8].

The last attribute of the electromagnetic field we need to discuss is wave polarization. The nature of the transverse field is such that the oscillating field disturbance (which is perpendicular to the propagation direction) has a particular orientation in space. The polarization of light is determined by the time evolution of the direction of the electric field

-5-

vector $\mathbf{E}(r,t)$. In our description, \mathbf{E}_0 is modulated by a phase factor which maps out the field oscillation in space. If z is the axis of propagation, we can define the amplitude factor as

$$\mathbf{E}_0 = E_x \hat{x} + E_y \hat{y} \quad (\text{C2.15.18})$$

where the unit vectors \hat{x} and \hat{y} are orthogonal and

$$E_x = e_x \exp(i\phi_x) \quad (\text{C2.15.19})$$

the amplitudes e_x with their phase factors ϕ map out the polarization vector in the x – y plane. The resultant \mathbf{E}_i ranges from linear polarized light, for ϕ_x or $\phi_y = 0$, through all possible combinations resulting in elliptical fields. In figure

C2.15.2 a right circularly polarized wave is illustrated. As the wave propagates, E_0 sweeps out a circle in the x - y plane. It is clear that, given a well characterized light source, there are many attributes we can attempt to control (wavelength, polarization, etc.); the question is how to generate well-characterized light?

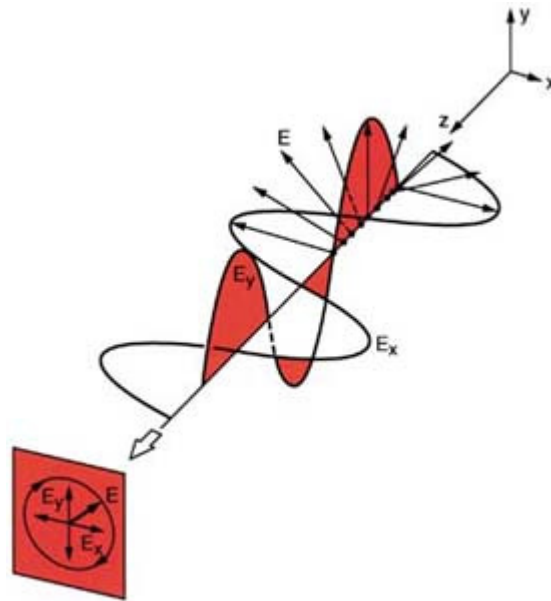


Figure C2.15.2. Right circularly polarized light. As the wave propagates the resultant E sweeps out a circle in the x - y plane.

C2.15.3 SOURCES: THE LASER

Given the general description of the electromagnetic field, let us explore the sources available for optoelectronics. The one primary light source for optoelectronic device and system architectures is the laser. The laser [10] is the source of choice simply because if we want to control light fields they need to be well defined at the start and the laser is the most

The acronym LASER (Light Amplification via the Stimulated Emission of Radiation) defines the process of amplification. For all intents and purposes this method was elegantly outlined by Einstein in 1917 [11] wherein he derived a treatment of the dynamic equilibrium of a material in a electromagnetic field absorbing and emitting photons. Key here is the insight that, in addition to absorption and spontaneous emission processes, in an excited system one can stimulate the emission of a photon by interaction with the electromagnetic field. It is this ‘stimulated’ emission process which lays the conceptual foundation of the laser.

The essential result of quantum theory [12] is that each physical system can be found upon measurement to be in one of a pre-determined set of energy states—the eigenstates of the system. These eigenstates [13] result from the solution of the Schrödinger equation for the system under study with the Hamiltonian [13] chosen to give the most complete characterization of the total energy of the system. Some classic analyses of generic systems [13] include the harmonic oscillator, the hydrogen atom and the hydrogen molecule ion problems. In each of these cases the solutions allow us to adequately predict the energetic processes for the complex systems mentioned. Let us assume we have a system described by Figure C2.15.3. Let us isolate two levels E_1 and E_2 . If the system is in E_2 , there is a finite probability per unit time that the system will decay to E_1 with the emission of a photon of energy $h\nu_{21}$. (The energy difference is $(E_n - E_{n-1}) = h\nu_{n,n-1}$.) This spontaneous emission is characterized by a lifetime for state E_2 . For an ensemble of the systems described above, we can represent the time rate of change of the population density as

$$-\frac{d\rho_2}{dt} = A_{21}\rho_2 = \rho_2/(\tau_{\text{spon}}) \quad (\text{C2.15.20})$$

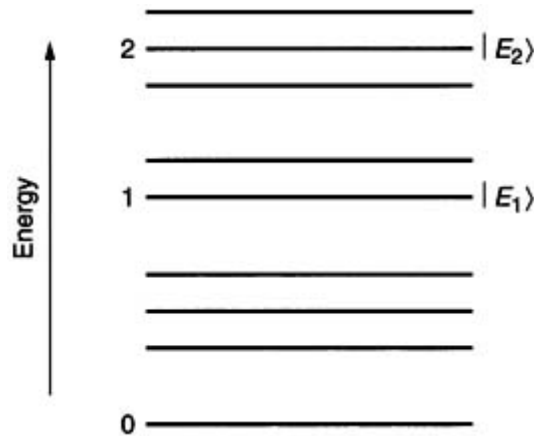


Figure C2.15.3. Generalized energy level diagram.

-7-

In the presence of an electromagnetic field of energy of about $h\nu_{21}$, our systems can undergo absorptive transitions from E_1 to E_2 , extracting a photon from the electric field. In addition, as described by Einstein, the field can induce emission of photons from E_2 to E_1 (given E_2 is occupied). Let the energy density of the external field be $E(\nu)$ then,

$$W_{21} = B_{21}E(\nu) \quad (\text{C2.15.21})$$

$$W'_{12} = B_{12}E(\nu) \quad (\text{C2.15.22})$$

where the B_{ij} are the Einstein coefficients for absorption and stimulated emission. The W_{ij} are the associated transition probabilities. In this picture the total transition rate would be

$$W'_{21} = B_{21}E(\nu) + A_{21} \quad (\text{C2.15.23})$$

where we have simply added the spontaneous contribution. In thermal equilibrium it can be shown that [1]

$$W_{\text{eq}} = \frac{\lambda^2 I}{8\pi n^2 h \nu \tau} g(\nu) \quad (\text{C2.15.24})$$

where W_{eq} is the equilibrium transition probability and $g(\nu)$ is a spectral lineshape function.

Using equation (C2.15.24), we can derive a general expression for the absorption coefficient for this simple two-level system:

$$\alpha(\nu) = (N_2 - N_1) \frac{c^2}{8\pi n^2 \nu^2 \tau} g(\nu). \quad (\text{C2.15.25})$$

If we substitute equation (C2.15.25) into Beer's law

$$I(l) = I(0) e^{\alpha(\nu)l} \quad (\text{C2.15.26})$$

it is clear that when the upperstate population exceeds that of the lower state there will be an exponential increase in the field intensity as the photon flux propagates through the active media. This ‘population inversion’ is the primary condition required for laser action. Since this is a non-equilibrium condition energy must be introduced into the system to reach this state, the process of ‘pumping’ is the introduction of the energy required to reach inversion and depends on the system in which we are attempting to obtain laser action. The first optical lasers were developed nearly 50 years after the introduction of the Einstein equations—direct evidence of the difficulty of creating inverted conditions. The key to laser design has been to find systems that can be efficiently pumped to produce gain.

It has been said that anything will lase if pumped with enough energy, but the efficiency of the pumping process is important for practical, economical devices. In this regard two-level lasers are of little interest because, except under extraordinary pumping conditions, one can only equalize the populations of the upper and lower levels. A three-level laser is illustrated in [figure C2.15.4\(a\)](#). The first solid-state laser ($\text{Cr}^{3+}:\text{AlO}_3$) ruby was of the three-level variety (b). The scheme works as follows. Atoms are pumped optically from state 1 to state 3. Non-radiative relaxation moves the

-8-

population from state 3 to state 2, creating (at sufficient pumping levels) a population inversion between level 2 and level 1. Lasing occurs on the 2–1 transition with the linewidth of the emission determined by the kinetics of the system and the resonator design.

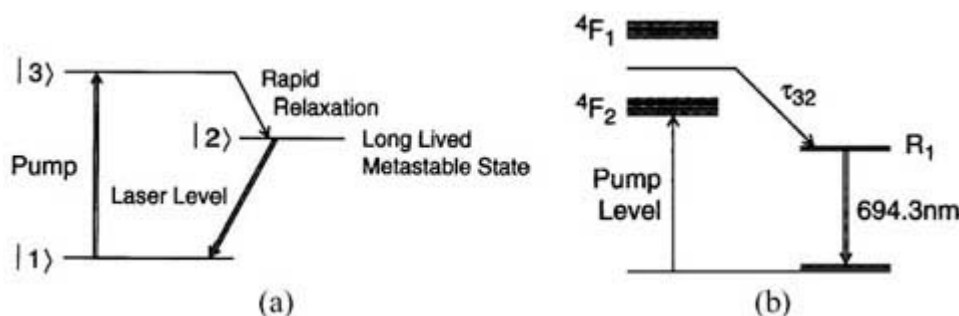


Figure C2.15.4. (a) A three-level laser energy level diagram and (b) the ruby system.

Four-level lasers offer a distinct advantage over their three-level counterparts, (figure C2.15.5). The $\text{Nd}^{3+}:\text{YAG}$ system is an excellent example of a four-level laser. Here the terminal level for the laser transition, |2>, is unoccupied thus resulting in an inverted state as soon as any atom is pumped to state 3. Solid-state systems based on this pumping geometry dominate the marketplace for high-power laser devices.

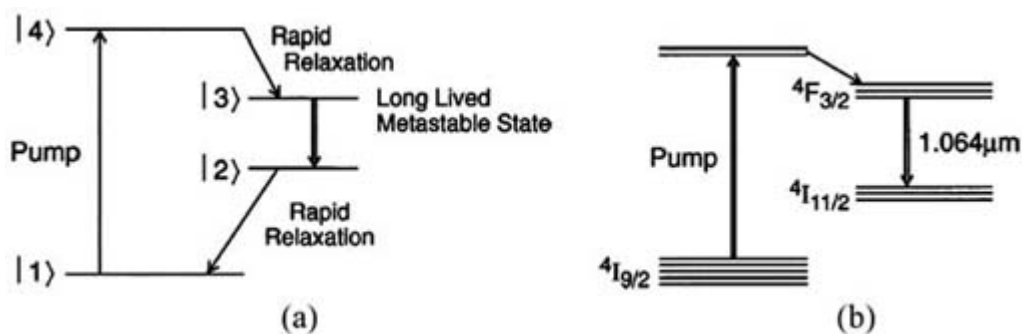


Figure C2.15.5. (a) A four-level laser energy level diagram and (b) the $\text{Nd}^{3+}:\text{YAG}$ system.

A wide variety of methods has been used to pump laser systems. Although optical pumping has been implied, there is an array of collisionally or electron impact pumped systems, as well as electrically pumped methods. The efficiency of the pumping cycle in many ways defines the utility and applications of each scheme. The first

material where optical laser action was observed was in the ruby system mentioned above. Here intense flashlamps were used to pump the system, which runs naturally in a pulsed mode. True continuous wave (CW) systems were first demonstrated in gaseous gain media.

The He-Ne laser system was the first efficient CW laser. It is still one of the most common systems in use today. Its level diagram is shown in [figure C2.15.6](#). Here a DC or RF discharge is used to excite the He⁺ ions, which in turn

collisionally excite Ne⁺ ions. Lasing occurs between several S and P bands with resonators designed to optimize the wavelength of interest.

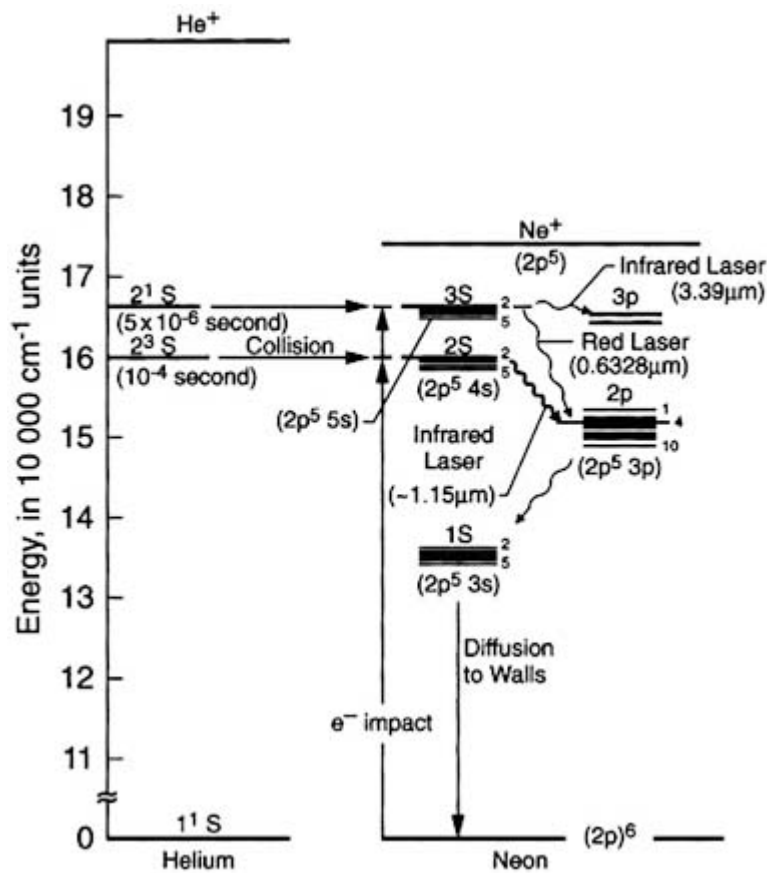


Figure C2.15.6. The He-Ne system.

For primary optoelectronic device applications the most relevant laser source is the semiconductor laser. A detailed analysis of semiconductor laser theory can be found in several references [6, 14] and is treated elsewhere in this volume, the basic operation of the lasers will be described qualitatively here. Semiconductors [15] are in complex multi-atom crystalline systems which can be characterized by dense bands of energy levels (derived from atom-atom interactions) separated by 'forbidden' gaps. These gaps are regions of phase space which do not match the boundary conditions required for electronic states. The simplest view of a semiconductor is as an ensemble of interacting atoms characterized by loosely bound valence electrons, coupled to a strong periodic potential derived from the atomic nuclei. The periodicity gives rise to the boundary conditions mentioned above, with the details of the energy levels determined by the atoms in the array and the specifics of the crystal structure. [Figure C2.15.7](#) shows a generic energy level diagram for metals, semimetals, insulators, and semiconductors. Intrinsic semiconductors have full valence bands and generally have band gaps that range from 0.1 to 4 eV. Absorption and

emission occurs *via* promotion of excited electrons from occupied to unoccupied band states. The key property of these materials is that the number of these states can be modified thereby changing the electronic properties of these materials. It is the ability to change the

-10-

electronic properties of these materials easily that has led to their extensive use as materials for complex electronic device fabrication. The addition ('doping') of electron-deficient or electron-rich atoms to the lattice can greatly modify the populations of the electrons and holes. Clearly, one now has another degree of freedom to adjust when attempting to reach inversion. The ability to create 'unoccupied' (hole) and/or 'occupied' (electron) states allows for 'chemical' pumping in addition to any other scheme designed. The p-n junction laser is a perfect example of this exploitation, (figure C2.15.8). Here a p-type material, (excess 'holes') is joined with a n-type material, (excess electrons), figure C2.15.8(a). At the junction a 'depletion layer' is created by the internal electric fields in the material which results in potential barriers that spatially localize each of the carriers, figure C2.15.8(b). When the structure is positively biased these barriers are lowered, allowing charge injection into the depleted region resulting in radiative recombination (figure C2.15.8(c)). This electric pumping process is extremely efficient and results in the low-current, high-output devices that are common today.

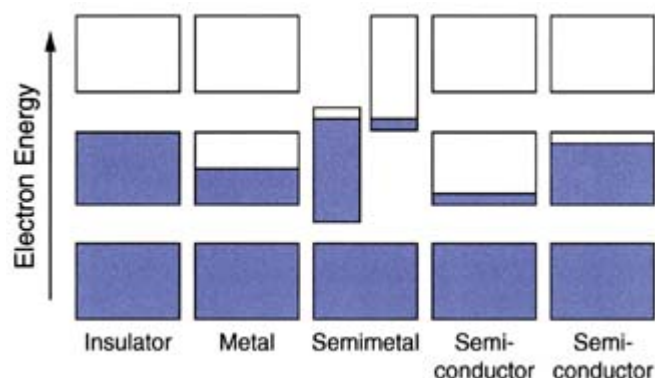


Figure C2.15.7. Generic band diagrams: insulator, metal, semimetal, and semiconductor.

We do not have space, nor is it appropriate to review all laser types and modes of operation, (the references included will afford the reader ample opportunity to survey the field). For reference we include a table giving an overview of the common laser types and their modes of operation (table C2.15.1). In general, pulsed laser output results from *Q*-switching or mode locking the devices. In both of these cases the kinetics of the optical system and the configuration of the optical resonator define the modulation frequency limits. In a *Q*-switched system a controllable loss is introduced into the resonator, allowing the steady-state population inversion to reach a level far above that achieved by conventional pumping. When this additional loss is removed the system begins oscillations at a point well above the threshold, lasing occurs with a rapid depletion of the gain, which eventually turns off the oscillation. The pulses generated have widths typically in the range of tens to hundreds of nanoseconds, with repetition rates of 10^4 – 10^5 pulses s^{-1} .

-11-

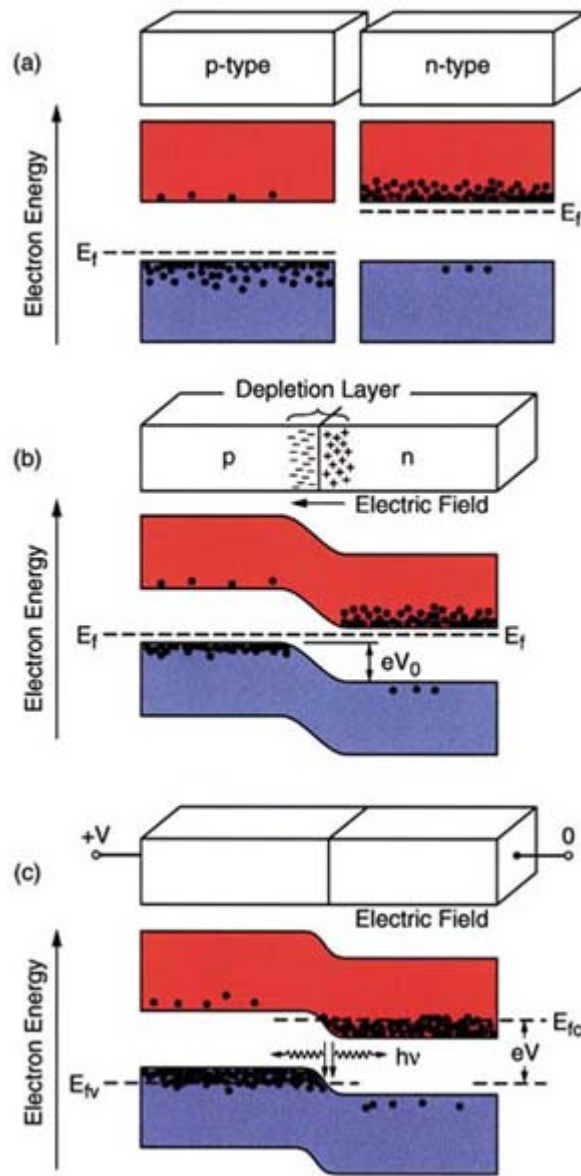


Figure C2.15.8. The p–n junction: (a) p-type and n-type materials, (b) depletion layer formation at the p–n interface or ‘junction’ and (c) p–n junction laser action.

Table C2.15.1 Common laser sources (s denotes solid-state lasers and g denotes gaseous lasers).

Laser type	Lasing wavelength λ	Efficiency η (%)	Mode of operation	Typical output power
ArF excimer (g)	193 nm	1	Pulsed	500 mJ
KrF excimer (g)	248 nm	1	Pulsed	500 mJ
He–Cd (g)	442 nm	0.1	CW	10 mW
Ar ⁺ (g)	514 nm	0.05	CW	10 W
He–Ne	633 nm	0.05	CW	1 mW

Kr ⁺ (g)	647 nm	0.01	CW	500 mW
Dye laser (R6G)	550–650 nm	0.005	CW or pulsed	100 mW
Ruby (Cr ³⁺) (s)	694 nm	0.1	Pulsed	5 J
Ti ³⁺ :Al ₂ O ₃ (s)	650–1180 nm	0.01	CW	1–10 W
Nd ³⁺ :glass (s)	1064 nm	1	Pulsed	10–50 J
Nd ³⁺ :YAG (s)	1064 nm	0.5	CW or pulsed	10–30 W
KF colour centre (s)	1.25–1.45 μm	0.005	CW	500 mW
CO ₂	10.6 μm	10–20	CW	100 W

Another method for producing pulsed laser output is longitudinal ‘mode locking’. Here, the natural longitudinal modes of a laser resonator are phase locked, resulting in wavepacket formation. This method takes advantage of the coupled gain characteristics of laser modes in optical resonators. It is the ability to dynamically control the phase relationships between the lasing modes in the cavity that makes this phenomenon possible. The width of this pulse is determined by the gain bandwidth, with the limits defined by the uncertainty principle. Mode locking is actually achieved by the intra-cavity modulation of the optical gain of the laser at the round trip frequency of the resonator. This frequency is $c/2l$, where c is the speed of light and l is the cavity length of the laser. This modulation induces sidebands which couple the gain of adjacent cavity modes.

The importance of laser light, in brief, is that its base characteristics, coherence, spectral and polarization purity, and high brilliance allow us to manipulate its properties. Gain switching [1, 10] and mode locking [16] are prime examples of our ability to very specifically control the laser output. It is easy to see why lasers are the ideal sources for optoelectronic applications.

C2.15.4 NONLINEAR OPTICS

The high-field output of laser devices allows for a wide variety of ‘nonlinear’ interactions [17] between the radiation field and the matter. Many of the initial relationships can be derived using engineering principles by simply expanding the media polarizability in a Taylor series in powers of the electric field:

$$P = \epsilon_0 \chi E + P_{NL} \quad (C2.15.27)$$

where

$$P_{NL} = 2\chi^{(2)} E^2 + 4\chi^{(3)} E^3 + \dots \quad (C2.15.28)$$

A wide variety of useful phenomena that allow the manipulation of the wavelength, amplitude and phase of the optical fields are mediated by χ , the first- and higher-order susceptibilities. In essence, the $\chi^{(n)}$ represent the complex interactions of the electric fields with the nonlinear media. They determine the explicit interaction of the quantum mechanical system (the propagation medium) with the quantized radiation field. Our engineering approach is a precursor to the well known semi-classical approach, where the radiation field is treated classically and the media quantum mechanically [18]. Assuming the electric dipole interaction represents the dominant contribution to the interaction Hamiltonian, the macroscopic polarization of the material of interest is the expectation value of the dipole matrix element scaled by the volume,

$$\mathbf{P}(r, t) = \frac{\langle E_a | \hat{d} | E_b \rangle}{\Delta V}$$

where the interaction Hamiltonian is

$$\hat{H}_i = -\hat{d}\mathbf{E}(r, t) \quad (\text{C2.15.30})$$

(here $\mathbf{E}(r, t)$ represents all external driving fields). This quantity incorporates all matter–field interactions. A perturbative expansion of this system allows correlation of the terms of equation (C2.15.29) to the n th order, with the terms generated from the simple Taylor series expansion described initially. This analysis gives an atomic/molecular basis for the susceptibilities, allowing greater insight to the nonlinear processes observed. An excellent treatment of this analysis is found in [18]. For our purposes the simple view that the intense fields drive a nonlinear oscillation of the polarizable electronic states of the materials is sufficient.

We will obtain a flavour of the nonlinear phenomena by exploring the processes generated via the matter–field

-14-

interactions to second order. The polarization to second order in the electric field is

$$\begin{aligned} \mathbf{P}^{(2)}(r, t) = \frac{\epsilon_0}{(2\pi)^2} \sum_{i,k} \int_{-\infty}^{+\infty} d\omega_1 \\ \int_{-\infty}^{+\infty} d\omega_2 \chi_{i,j,k}^{(2)}(\omega : \omega_1, \omega_2) \mathbf{E}_j(\omega_1) \mathbf{E}_k(\omega_2) e^{i(\omega_1 + \omega_2)t} \end{aligned} \quad (\text{C2.15.31})$$

where the various contributions arise through the permutation of the indices j and k . The most encountered spectral component of $\chi^{(2)}$ arises from a two-photon coupling of a single electric field, resulting in the generation of a second harmonic of the field. Second harmonic generation (SHG) [19], discovered soon after the laser, is an essential wavelength conversion tool utilized throughout laser physics and engineering. The relevant terms in the polarization are of the form

$$\mathbf{P}_i^{(2)}(\omega, z) = \epsilon_0 \sum_{i,j,k} \chi_{i,j,k}^{(2)}(\omega : 2\omega, -\omega) \mathbf{E}_j(2\omega, z) \mathbf{E}_k^*(\omega, z) \quad (\text{C2.15.32})$$

and

$$\mathbf{P}_i^{(2)}(2\omega, z) = \epsilon_0 \sum_{i,j,k} \chi_{i,j,k}^{(2)}(2\omega : \omega, \omega) \mathbf{E}_j(\omega, z) \mathbf{E}_k^*(\omega, z). \quad (\text{C2.15.33})$$

$\mathbf{P}_i^{(2)}$ will of course be the source term in the wave equation. It is clear that for SHG the generated polarization scales as $|\mathbf{E}(\omega)|^2$. In general, the intensity scales with the incident power per cross sectional area. To maximize the SHG output, the interaction length of the ω and 2ω fields needs to be as long as possible, or power begins to be converted back to ω . In most materials, dispersion and diffraction effects limit the conversion efficiency. A wide variety of techniques have been developed to solve this problem. The most widespread is the use of uniaxial nonlinear crystals for wavelength conversion. In these systems (which have orientation dependent indices of refraction) the crystals are cut such that the propagating second harmonic and fundamental wavelength traverse the media at the same speed, thus resulting in optimal conversion. This is commonly known as phase matching. (A detailed analysis of the phase matching arrangements can be found in [20].) In addition, several waveguide and fibre SHG devices have been developed [21]. Another second-order process of great utility is optical rectification [22] (the coupling and generation of DC fields using an optical field). An outgrowth of this process is the ‘electro-optic effect’, which allows the manipulation of optical radiation with strong DC fields in appropriate media. In

these materials, the change in index can be written as

$$\Delta n \approx \frac{\chi^{(2)}}{n} E(0) = -\frac{1}{2} n^3 \tilde{r} E(0) \quad (\text{C2.15.34})$$

where r is the Pockels coefficient. This effect allows for electric field induced phase shifting of the optical fields. The nonlinear process allows the direct, rapid, and efficient coupling of electrical RF signals to high-frequency optical fields, making the Pockels effect an essential tool for high-frequency modulation.

-15-

C2.15.5 OPTICAL LIGHT GUIDES

Traditionally, light signals are directed using lenses, mirrors, and optical prisms. This free-space guiding of light, although exceedingly useful, is not very robust for optical transmission over long distances. The push to transmit ‘information’ optically drove the development of optical conduits to transmit light signals from place to place. The development of guided-wave optics has been key to the advances of optical communication, leading to the optoelectronics revolution. In addition, the concentration of intense radiation afforded by these guides has impacted every area from laser design to nonlinear optical devices and has opened the door for the development of *integrated optical devices*. In this section we will outline the basics of guided-wave optics. As in [1], we will start with the planar-mirror waveguide as an introduction to the essential concepts, and then move on to dielectric structures.

Consider a light conduit constructed using two planes of parallel mirrored surfaces (figure C2.15.9). Assuming the mirrors are lossless, a light ray at an appropriate angle θ will propagate along the conduit axis, reflecting without loss of energy. (Note that waveguides of this type are not made in practice due to the difficulty of fabricating mirrors with low enough losses.) Consider launching a monochromatic wave into the guide with wavelength $\lambda = \lambda_0/n$ (where n is the refractive index between the plates). Given our basic assumption of lossless surfaces, the guide constrains the propagating modes to those that maintain the same transverse distribution at all distances along the waveguide axis. To fulfill this requirement *any launched wave must remain unchanged after two reflections*. In addition there are a limited number of angles which satisfy these conditions for this system, and they must satisfy the relationship

$$\sin \theta_m = m \frac{\lambda}{2d} \quad (\text{C2.15.35})$$

where $m = 1, 2, 3, \dots$. The guided-wave modes are composed of two distinct plane waves at $\pm\theta$. We define the propagation constant of the m th mode as

$$\beta_m^2 = k^2 - \frac{m^2 \pi^2}{d^2}. \quad (\text{C2.15.36})$$

There is a maximum number of modes possible, defined by the range of accessible angles. For $\sin \theta < 1$, the maximum allowed value of modes is the greatest integer smaller than $1/(\lambda/2d)$ or

$$M \cong \frac{2d}{\lambda}. \quad (\text{C2.15.37})$$

As shown, the number of modes increases with increasing mirror separation and decreasing wavelength. If $2d/\lambda$ is less than one, $M = 0$ and no self-consistent modes are supported. $2d$ represents the cut-off wavelength of the guide. It is the longest wavelength supported by the guide. It is clear that if the spacing is adjusted properly, M can be set to one and only a single mode will be supported. For completeness, we note TM (transverse magnetic) and TE (transverse electric) mode distributions define the direction of \mathbf{E} of the propagating field as shown in [figure](#)

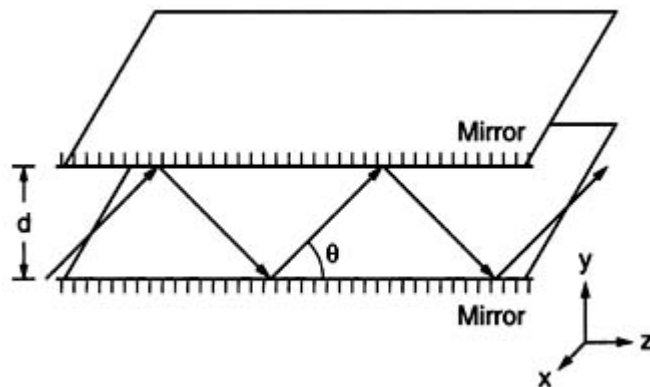


Figure C2.15.9. The planar-mirror light guide.

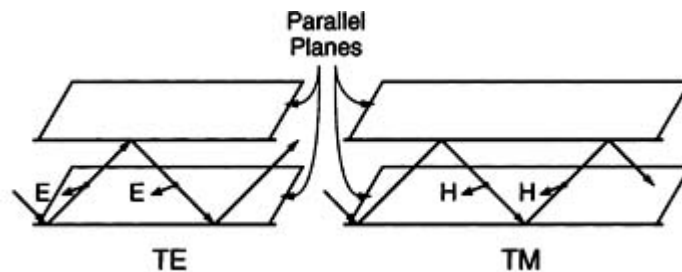


Figure C2.15.10. Orientation of the TE and TM modes.

C2.15.6 THE DIELECTRIC WAVEGUIDE

Optical conduits as described above are generally not practical. The most common waveguide is the slab dielectric waveguide. In these devices, a high-transmission material is surrounded by a media of a lower refractive index. The light is guided into the device by total internal reflection. The basic structure is shown in [figure C2.15.11](#). As with our initial description, light rays making an angle θ with the z -axis experience multiple total internal reflections at the interfaces, provided that θ is *smaller than the complement of the critical angle*. The slab boundaries define all of the properties of the guide. As before, rays making angles larger than the complement of the critical angle refract, losing a fraction of their optical power at each reflection and eventually vanishing (the unguided waves of [figure C2.15.11](#)). The detailed analysis of the waveguide modes requires a full solution of Maxwell's equations both inside and outside the high-index core with the appropriate boundary conditions. Such an analysis, which is beyond the scope of this review, can be found in [1, 23]. We will summarize the results here. As with our first analysis, a twice-reflected wave undergoes a phase shift that must be zero or a multiple of 2π to be self-consistent. The number of TE modes allowed is

$$M \cong \frac{\sin \theta_c}{\lambda/2d} \tag{C2.15.38}$$

where M is increased to the nearest integer, More generally

$$M = 2 \frac{d}{\lambda_0} NA \quad (C2.15.39)$$

where the numerical aperture $NA = (n_1^2 - n_2^2)^{1/2}$. The NA is the sine of the angle of acceptance of rays from air into the core of the slab. When $\lambda/2d > \sin \theta_c$ or $(2d/\lambda_0)NA < 1$ the waveguide supports a single mode. The TE and TM modes in a dielectric planar waveguide are as shown in [figure C2.15.12](#). The possible modes can be characterized by a propagation constant, β where

$$\beta_m = nk_0 \cos \theta_m. \quad (C2.15.40)$$

One can of course fabricate two-dimensional waveguides. These devices confine light in two transverse directions (x and y). An important example of two-dimensional waveguides is the optical fibre, which we will treat directly. Generally, two-dimensional waveguides are of the channel variety. An array of two-dimensional waveguide geometries is shown in [figure C2.15.13](#). So far, we have not considered modal interactions in guides, that is, the coupling of light from one mode to another, or the energy transfer between modes. The mode coupling of light is an important tool in optoelectronics. Although a full treatment of this process is beyond the scope of this chapter, we will describe one relatively simple device as an example and leave it to the reader to survey the references for greater detail on a wide variety of structures. It should be stated that the design and fabrication of these devices is an exciting area of current research.

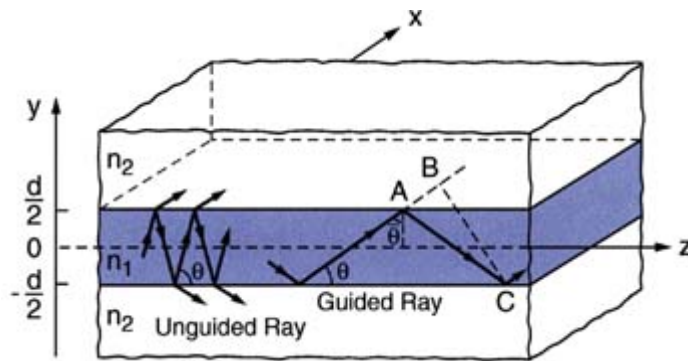


Figure C2.15.11. The dielectric waveguide.

-18-

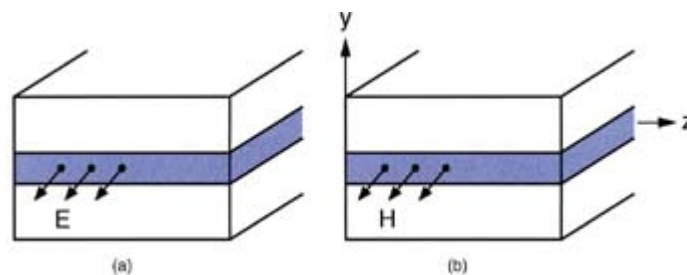


Figure C2.15.12. (a) TE and (b) TM modes for the dielectric planar waveguide.

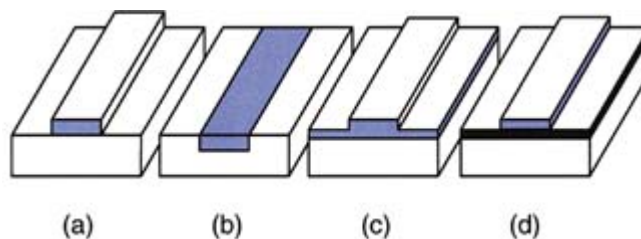


Figure C2.15.13. Two-dimensional waveguide configurations, the darker shading indicates the different indexes.

Our model device will be the directional coupler ([figure C2.15.14](#)). The basic function of the structure is to couple two optical inputs to two optical outputs. Consider two single-mode waveguides on the same substrate, as shown in the figure. Using a simple-coupled mode model, if the guides are non-interacting, the amplitudes of the input fields as a function of propagation distance are

$$\frac{dA_1}{dz} = -i\beta A_1(z) \quad (\text{C2.15.41a})$$

$$\frac{dA_2}{dz} = -i\beta A_2(z) \quad (\text{C2.15.41b})$$

with solutions

$$A_i(z) = A_i(0) e^{i\beta z} \quad (\text{C2.15.42})$$

where $i = 1$ and 2 respectively. (Note that the modes propagate unchanged.) If the two guides are brought close enough together, the evanescent fields of the two modes interact, allowing energy exchange. One can define a coupling constant κ that characterizes the perturbative interaction of the modes. Under these conditions equation (C2.15.41a) and equation (C2.15.41b) become

$$\frac{dA_1}{dz} = -i\beta A_1(z) - i\kappa A_2(z) \quad (\text{C2.15.43a})$$

$$\frac{dA_2}{dz} = -i\beta A_2(z) - i\kappa A_1(z). \quad (\text{C2.15.43b})$$

If the guides are lossless (here we ignore bending losses) and we only launch a field into guide 1 (i.e. $A_1 = 1, A_2 = 0$)

$$A_1(z) = \cos(\kappa z) \quad (\text{C2.15.44a})$$

$$A_2(z) = \sin(\kappa z). \quad (\text{C2.15.44b})$$

The optical power $|A_i|^2$ oscillates between the guides depending upon the propagation distance. Clearly, by controlling z , β and κ , a wide variety of passive devices (beam splitters, combiners, attenuators, and interferometers) can be readily constructed.

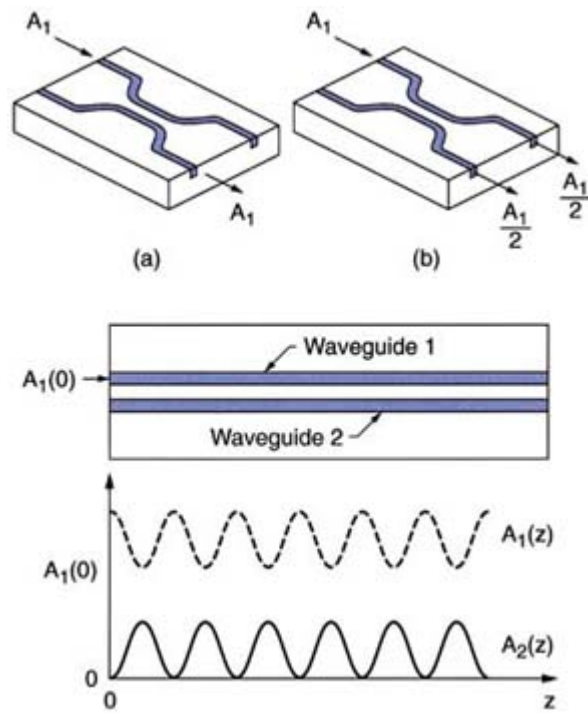


Figure C2.15.14. The directional coupler.

C2.15.7 THE OPTICAL FIBRE

The optical fibre [4, 24] is the most extensively used optical waveguide. The cylindrical step index structure is composed of an ultra-pure, extremely-low-loss, high-index core surrounded by a slightly lower index cladding. The fibre properties are characterized by the relative diameters of the cladding and the core, $2a/2b$ (where a and b are the core and cladding radii), and the delta of the two materials (figure C2.15.15). The delta, Δ , is defined as

$$\Delta = \frac{n_1 - n_2}{n_1} \quad (\text{C2.15.45})$$

where n_1 and n_2 are the core and cladding indexes of refraction, respectively. (Typical deltas range from 0.001 to 0.02.) An optical field is guided in the fibre core just as in the dielectric waveguide *via* the total internal reflection at the core-cladding boundary. Again, rays propagate if their angle of incidence is less than the complement of the critical angle. The fibre input is defined in terms of the numerical aperture NA , with

$$\theta_a = \sin^{-1} NA \quad (\text{C2.15.46})$$

where θ_a is the fibre acceptance angle. The NA can also be written in terms of the Δ of the fibre:

$$NA = (n_1^2 - n_2^2)^{1/2} \cong n_1(2\Delta)^{1/2}. \quad (\text{C2.15.47})$$

Another important characterization parameter for fibres is the normalized frequency V :

$$V = k_0 a (n_1^2 - n_2^2)^{1/2} = k_0 NA. \quad (\text{C2.15.48})$$

Note that here $k_0 = 2\pi/\lambda$ and a is the core radius. The parameter V determines the number of modes supported by

the fibre design, therefore defining the cut-off frequency of the fibre. For $V < 2.405$, only a single mode is supported. Although written here very simply, the V parameter is exactly derived from solution of the complex eigenvalue problem of the weakly guiding fibre [25]. A full solution of the Helmholtz equation representation of the Maxwell's equations in cylindrical coordinates must be obtained in the core and cladding, with the appropriate boundary conditions. The Bessel function solutions of this problem give rise to the characteristic equations for V . In practice, many of the solutions are obtained graphically. For large V there are a large number of roots to the characteristic equations, allowing for a large number of propagating modes. In this limit, the number of modes is

$$M \approx \frac{4}{\pi^2} V^2 \tag{C2.15.49}$$

for $V \gg 1$. Under these conditions the propagation constant can be found to be

$$\beta_{l,m} \cong n_1 k_0 \left[1 - \frac{(l+2m)^2}{M} \Delta \right] \tag{C2.15.50}$$

-21-

and the group velocity of the (l, m) mode is

$$v_{l,m} \cong c_1 \left[1 - \frac{(l+2m)^2}{M} \Delta \right]. \tag{C2.15.51}$$

This spread in velocity is called ‘modal dispersion’ and is the principle limit to the use of multimode fibres for long-distance transmissive applications.

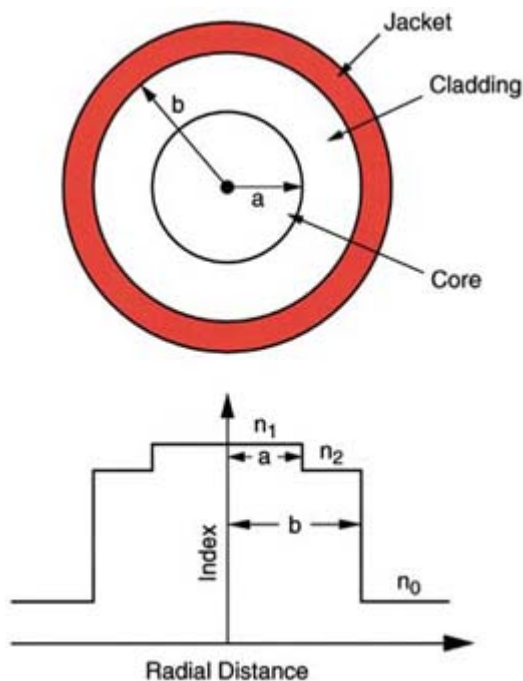


Figure C2.15.15. The structural profile of a step index fibre.

As described above for small a and NA , a fibre is single mode if $V < 2.405$. Here only one mode, with one group velocity, is possible. This lack of ‘modal dispersion’ is why single-mode fibre dominates transport media in long-haul communication systems.

Graded index structures allow greater control over fibre characteristics. In these structures, the core has a variable

refractive index, highest in the centre and gradually decreasing until it reaches that of the cladding at the core–cladding interface. The result is that the phase velocity gradually increases with spatial position in the core. Properly designed structures can greatly minimize the differences in the group velocity of the fibre modes, limiting modal dispersion.

For optical transmission, the parameters of greatest importance are attenuation (i.e. loss) and ‘material’ dispersion. In effect they define the limits of the optical communication system. Loss, due to absorption and scattering, limits the lengths between the transmission nodes. In transmission quality fibre, the loss is in units of decibels per kilometre.

(Attenuation of less than 0.2 dB km^{-1} is common for telecom quality fibre.) Since modal dispersion can be greatly mitigated by fibre design, real material dispersion is of greatest consequence. Silica is a dispersive media. There is a wavelength dependence to the index of refraction, as a result an optical pulse of finite bandwidth, $\delta\lambda$, spreads as it propagates along the core axis. This spread limits the spacing of successive pulses and hence the maximum transmission frequency. Figure C2.15.16 shows a plot of the dispersion for silica as a function of wavelength within the transmission window. Because the zero dispersion point is at approximately $1.31 \mu\text{m}$, this wavelength has become one of the base telecom transmission bands. The other key telecom transmission wavelength, $1.54 \mu\text{m}$, is near the loss minimum of the fibre.

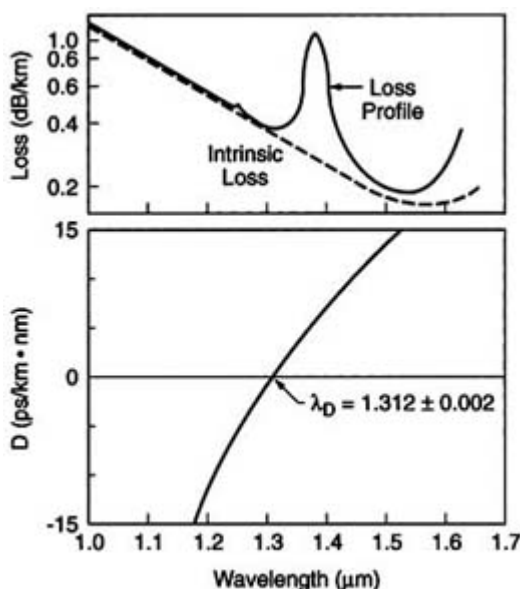


Figure C2.15.16. Wavelength dependent loss (upper) and dispersion (lower) in a silica fibre.

Nonlinear dispersion becomes relevant at sufficient pulse powers. In some fibre structures the interplay between the nonlinear dispersion and the group velocity dispersion can be used to produce non-dispersive waves called solitons. Solitons, although beyond the scope of this treatment, may revolutionize the communication systems of the future. A full treatment of soliton theory can be found in [4, 26].

C2.15.8 OPTICAL MODULATION AND DETECTION

We will complete our survey of optoelectronics with a brief discussion of optical modulation and optical detection. These two categories of devices are important because they define the speed of the optoelectric link. The ability to generate and detect high-frequency signals determines the ultimate limits on any optical circuit. Transmission modulation can be accomplished two ways: (a) direct modulation, switching of an optical source and (b) external ‘shuttering’ of a CW optical source. Although both are adequate, the latter allows for greater flexibility and

The most useful direct modulation technique is the current gain switching of semiconductor laser devices. This technique is unique to semiconductor sources, nearly all other lasers are modulated externally. In these devices the excitation current of the laser is modulated, resulting in modulated gain and therefore modulated output power. A detailed analysis of this process is found in [27]. Simply put, an oscillating current of the form

$$I = I_0 + i_m e^{i\omega_m t} \tag{C2.15.52}$$

is applied to the device. Since the laser output power is

$$P_0 \approx \frac{\eta(I - I_{th})}{\lambda} \tag{C2.15.53}$$

there will be a modulation in the output power. Careful laser design can result in modulation frequencies in the hundreds of megahertz to gigahertz range. In general, the limitation of the high-frequency response of these laser devices is due to the broadening of linewidth and spectral output due to a change of the material absorption, gain and index of refraction as a function of the carrier density. This ‘chirping’ of the laser output bounds the modulation frequencies possible. To surpass this limitation external modulation of CW lasers is employed, allowing modulation frequencies greater than 10 GHz.

A wide variety of external modulators are used in practice. Electro-optic modulators can produce amplitude, frequency, or phase modulation utilizing the Pockels effect (mentioned when we studied nonlinear optical phenomena above). By polarizing the input wave and setting the electro-optic device between crossed polarizers. A controllable bias on the modulator determines the optical phase shift induced on the optical beam and therefore the output. Waveguide electro-optic devices are also of great interest. Single-mode waveguide devices, such as couplers and interferometers, were mentioned earlier and are becoming of great importance in the communications industry. Waveguides fabricated using electro-optic materials such as LiNbO₃ can be made into active devices. The Pockels effect allows dynamic index switching, enabling the modulation of an optical input. A diagram of an optical Mach–Zender modulator is shown in figure C2.15.17. In addition, external electroabsorption has achieved wide success in communication links. Here the process that limits the direct modulation of a semiconductor laser, i.e. the change (shift) of the absorption spectrum with carrier density, is used for high-speed modulation. A beneficial side effect is that fully integrated laser/modulator packages can be fabricated.

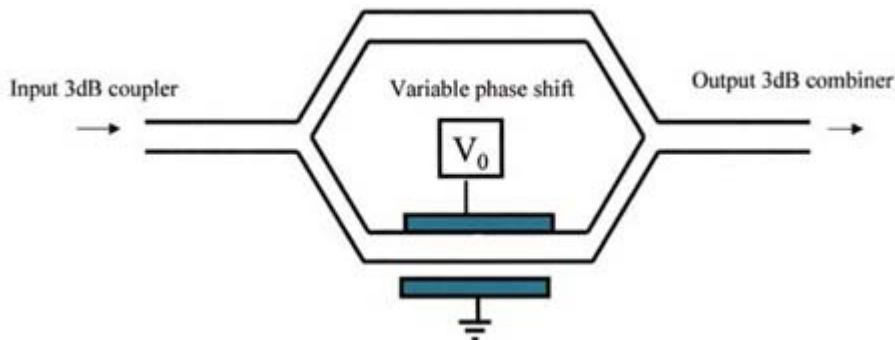


Figure C2.15.17. The Mach–Zender modulator. A 3 dB coupler splits the input wave into the two arms of the device. The output 3 dB combiner recombines half the wave with its phase-shifted counterpart. By adjusting V_0 the output transmission can be rapidly modulated.

Optical detectors generally fall into two broad categories: thermal and photoelectric detectors. Thermal devices operate by converting the photon detected into heat with the heat evolution used to quantify the emissive output. Photoelectric devices convert the photons directly to electrons or other charge carriers, with the subsequent current proportional to the photon input. In general thermal devices are slow and not easily integrated into complex devices. The dominant detectors for optoelectronic applications are of the photoelectric variety. Again, two general categories of devices are prevalent: those that rely on the photoelectric effect and those that are based on photoconductivity. Photodetectors, based on the photoelectric effect, are known as phototubes. In these devices a photoemissive material is configured as a cathode. The input radiation induces the ejection of a photoelectron from the cathode, which is accelerated towards the anode (which is at a higher electric potential), generating a current. Often secondary emission devices (called dynodes) are used to produce a large amplification of the photocurrent. These dynodes are electron multipliers with successive stages, resulting in large current gains. Phototubes are commonly used as high-gain (10^7), high-sensitivity detectors for a wide variety of applications.

For optoelectronic applications photoconductive devices are more common. Semiconductor photoconductives tend to be inexpensive and can be easily integrated with other components. In these devices photo-illumination above the bandgap of the material generates charge carriers, increasing the conductivity of the semiconductor and resulting in a photocurrent. For example in a p–n junction device, photons absorbed in the depletion layer generate electrons and holes under the influence of an electric field, this directly generates the photocurrent described. If a large reversed bias is placed across the junction, the large field produced can accelerate photogenerated carriers with enough kinetic energy to excite additional carriers by impact ionization. This ‘avalanche’ effect results in a dramatic improvement in detector sensitivity. Avalanche photodiodes have become critical devices for many photonic applications.

C2.15.9 OPTICAL COMMUNICATIONS

The primary driver for the expansion of optoelectronic technologies is optical communications [2]. It was realized in the second-half of the 20th century that an increase of several orders of magnitude in bandwidth would be possible if optical waves were used as the carrier for telephone signals. The basic configuration of an optical communication

system is shown in figure C2.15.18. All the components described in this review are used, and, in many cases, were developed to complete the optical link. The primary application that drove optoelectronics research through the 1980s and early 1990s was long-distance business; the goal being to send gigabit data rate information efficiently over very long distances (thousands of kilometres). The need to go farther and faster led to the development of extremely-low-loss optical fibre, ultra-high-speed modulators and high-sensitivity avalanche photodetectors. This important innovation enabled all-optical transmission. The development of the optical fibre amplifier revolutionized optical communication, allowing digital optical signals to be transmitted vast distances before being converted back to electronic pulses, therefore greatly simplifying long-haul links. (These devices amplify the optical data signals via stimulated emission.)

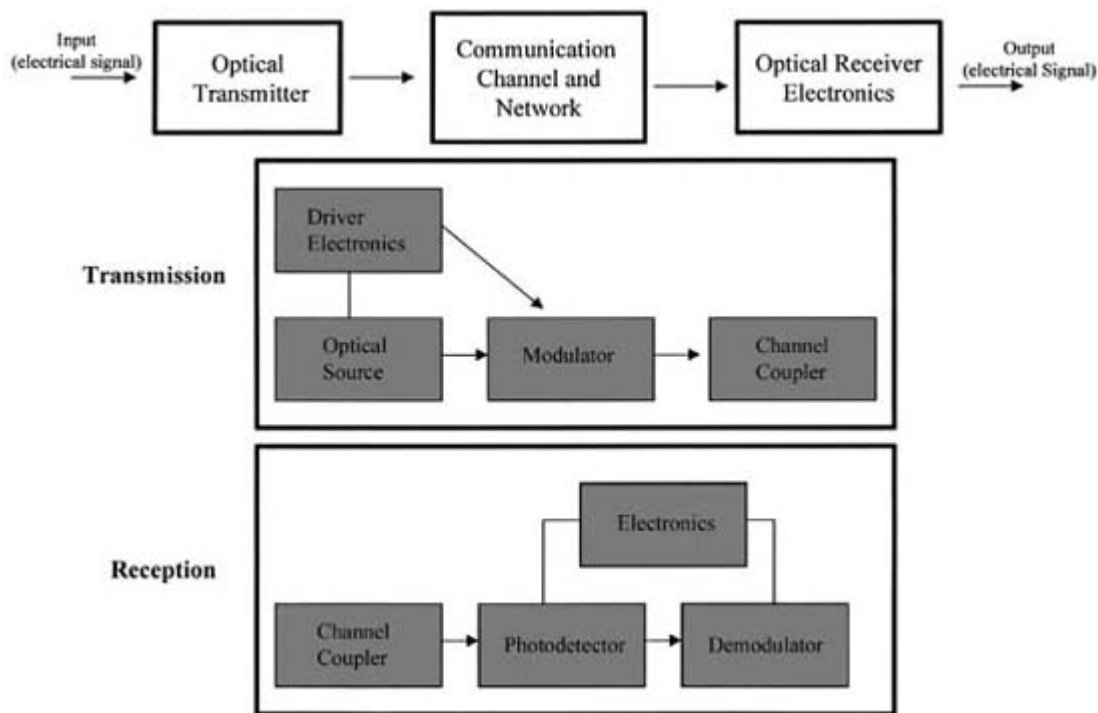


Figure C2.15.18. The basic components of an optical communications link.

During the last 5–10 years optical networks of all varieties have become a very important part of the World’s communications infrastructure for data transmission. The new focus has become the transmission of multi-gigabit information over moderate distances at *low cost*. These requirements are profoundly affecting the design criteria for optoelectronics devices and components. In addition, coarse and dense wavelength division multiplexed systems are being developed to increase bandwidth. The essence of wavelength division multiplexing is the simultaneous transmission of many optical data channels each at a different wavelength on a single optical fibre. Systems with as many as 40, 1–10 Gbit optical channels are becoming commercially available. Low-cost components will be the key factor in determining how rapidly this technology will be deployed throughout the communications infrastructure.

C2.15.10 CONCLUSION

In this chapter we have laid out much of the underlying basis for a wide variety of optoelectronic devices and structures. I hope that it is clear from the text that optoelectronics is optical physics in action. While optical technologies are beginning to mature, much new work is needed to provide new sources and devices for next-generation photonic applications. The introduction of organic electronic devices [28] and novel photonic bandgap materials [29] will no doubt add fuel to the fire. I expect these technologies to redefine computing and communication well into the 21st century.

REFERENCES

- [1] Saleh B E A and Teich M C 1991 *Fundamentals of Photonics* (New York: Wiley–Interscience)
- Yariv A 1967 *Quantum Electronics* (New York: Wiley)
- Yariv A 1997 *Optical Electronics and Modern Communications* 5th edn (New York: Oxford University Press)

- [2] Palais J C 1992 *Fiber Optic Communication* 3rd edn (Englewood Cliffs, NJ: Prentice Hall)
Tanenbaum A S 1996 *Computer Networks* 3rd edn (Upper Saddle River, NJ: Prentice Hall)
- [3] Hecht E 1998 *Optics* 3rd edn (New York: Addison-Wesley)
- [4] Agrawal G P 1989 *Nonlinear Fiber Optics, Quantum Electronics: Principles and Applications* (New York: Academic)
- [5] Das P K 1990 *Lasers and Optical Engineering* 2nd edn (New York: Springer)
- [6] Agrawal G P and Dutta N K 1988 *Long-Wavelength Semiconductor Lasers* (New York: Dekker)
- [7] Bucher P N and Cotter D 1990 *The Elements of Nonlinear Optics* (New York: Cambridge University Press)
- [8] Jackson J D 1962 *Classical Electrodynamics* (New York: Wiley)
Born M and Wolf E 1970 *Principles of Optics* (Oxford: Pergamon)
- [9] Cohen-Tannoudji C, Dupont-Roc J and Grynberg G 1989 *Photons and Atoms* (New York: Wiley–Interscience)
Cohen-Tannoudji C, Dupont-Roc J and Grynberg G 1992 *Atom–Photon Interactions* (New York: Wiley–Interscience)
- [10] Schawlow A L and Townes C H 1958 Infrared and optical lasers *Phys. Rev.* **112** 1940
Maiman T H 1960 Stimulated optical radiation in ruby *Nature* **187** 493
Milonni P W and Eberly J H 1988 *Laser* (New York: Wiley)
- [11] Einstein A 1917 Zur Quantentheorie der Strahlung *Phys. Z.* **18** 121
- [12] Allen L and Eberly J H 1975 *Optical Resonance in Two-level Atoms* (New York: Wiley)

- [13] Cohen-Tannoudji C, Diu B and Laloe F 1977 *Quantum Mechanics* vol 1 (New York: Wiley–Interscience)
Cohen-Tannoudji C, Diu B and Laloe F 1977 *Quantum Mechanics* vol 2 (New York: Wiley–Interscience)
- [14] Thomson G H B 1981 *Physics of Semiconductor Lasers* (New York: Wiley)
- [15] Kittel C 1986 *Introduction of Solid State Physics* 6th edn (New York: Wiley)
- [16] Siegman A E 1986 *Lasers* (Mill Valley, CA: University Science Books)
- [17] Blomenbergen N 1991 *Nonlinear Optics* (Reading, MA: Addison-Wesley)
- [18] Shubert M and Wilhelm B 1986 *Nonlinear Optics and Quantum Electronics* (New York: Wiley)
- [19] Franken P A, Hill A E, Peters C W and Weinreich G 1961 Generation of optical harmonics *Phys. Rev. Lett.* **7** 118
- [20] Maker P D, Terhune R W, Nisenoff M and Savage C M 1962 Effects of dispersion and focusing on the production of optical harmonics *Phys. Rev. Lett.* **8** 21
Giordimaine J A 1962 Mixing of light beams in crystals *Phys. Rev. Lett.* **8** 19
- [21] Stolen R H and Tom H W K 1987 *Opt. Lett.* **12** 585
- [22] Shen Y R 1984 *The Principles of Nonlinear Optics* (New York: Wiley)
- [23] Lotspeich J F 1975 Explicit general eigenvalue solutions for dielectric slab waveguides *Appl. Opt.* **14** 327

- [24] Cheo P K 1985 *Fiber Optics and Optoelectronics* (Englewood Cliffs, NJ: Prentice Hall)
- [25] Gloge D 1971 Weakly guiding fibers *Appl. Opt.* **10** 2252
Gloge D 1971 Dispersion in weakly guiding fibers *Appl. Opt.* **10** 2442
- [26] Lamb G L Jr 1980 *Elements of Soliton Theory* (New York: Wiley)
Hasegawa A 1989 *Optical Solitons in Fiber* (Berlin: Springer)
Mollenauer L F, Stolen R H and Gordon J P 1980 *Phys. Rev. Lett.* **45** 1095
- [27] Lau K T, Bar-Chaim N, Ury I and Yariv A 1983 Direct amplitude modulation of semiconductor GaAs lasers up to X-band frequencies *Appl. Phys. Lett.* **43** 11
Lau K T, Harder Ch and Yariv A 1983 Direct modulation of semiconductor at $f > 10$ GHz *Appl. Phys. Lett.* **44** 273
- [28] Lovinger A J and Rothberg L J 1996 *J. Mater. Res.* **11** 1581
Katz H E 1997 *J. Mater. Chem.* **7** 369
Bao Z, Feng Y, Dodabalapur A, Raju V R and Lovinger A J 1997 *Chem. Mater.* **9** 1299
- [29] Joannopoulos J D, Meade R D and Winn J N 1995 *Photonic Crystals* (Princeton, NJ: Princeton University Press)

-28-

FURTHER READING

Saleh B E A and Teich M C 1991 *Fundamentals of Photonics* (New York: Wiley-Interscience)

Yariv A 1997 *Optical Electronics and Modern Communications* 5th edn (New York: Oxford University Press)

Singh J *Optoelectronics: An Introduction to Materials and Devices* (New York: McGraw-Hill)

Chuang S L 1995 *The Physics of Optoelectronic Devices* (Wiley series in pure and applied optics)

AgraWal G P 1997 *Fiber-Optic Communication Systems* (New York: Wiley-Interscience)

-1-

C2.16 Semiconductors

Henryk Temkin and Stefan K Estreicher

I think there is a world market for maybe five computers.

Thomas Watson Sr, IBM Chairman, 1943

The history of semiconductor devices can be traced back to the paper of Braun, published in 1874, describing rectifying behavior of a contact [1]. However, for many years semiconductors were considered too difficult a subject and the science of semiconductors began only during World War II.

The physics of semiconductors was understood rather quickly but the materials were far too poor for practical applications. The first high-purity semiconductor grown in large quantities was Ge. Si emerged only in the mid 1950s as the semiconductor of choice. Compound semiconductors, such as GaAs, began to play a role in the mid-1970s. These developments are presented in a number of recent review articles [2, 3, 4 and 5].

Since the early days of this field, the scientific and technological advances have been chasing each other. Advances in technology enabled implementation of new ideas, which in turn suggested new applications. In the last 50 years this process has resulted in a series of unprecedented advances that have transformed our society. There is, as of yet, no sign of a slowdown.

The present article reviews basic concepts of semiconductor physics and devices with emphasis on current problems. Further details can be found in the references.

C2.16.1 INTRODUCTION

Semiconductors are a class of materials whose conductivity, while highly pure, varies with temperature as $\exp(-E_g/k_B T)$, where E_g is the size of a forbidden energy gap. The conductivity of semiconductors can be made to vary over orders of magnitude by *doping*, the intentional introduction of appropriate impurities. The range in which the conductivity of Si can be made to vary is compared to that of typical insulators and metals in [figure C2.16.1](#).

In an *intrinsic* semiconductor, the conductivity is limited by the thermal excitation of electrons from a filled *valence band* (VB) into an empty *conduction band* (CB), across a forbidden energy *gap* of width E_g . The process leaves *holes* (h^+) in the VB and *electrons* (e^-) in the CB, and both of these *charge carriers* participate in the conduction.

In an *extrinsic* semiconductor, the conductivity is dominated by the e^- (or h^+) in the CB (or VB) provided by shallow *donors* (or *acceptors*). If the dominant charge carriers are negative (electrons), the material is called n type. If the conduction is dominated by holes (positive charge carriers), the material is called p type.

-2-

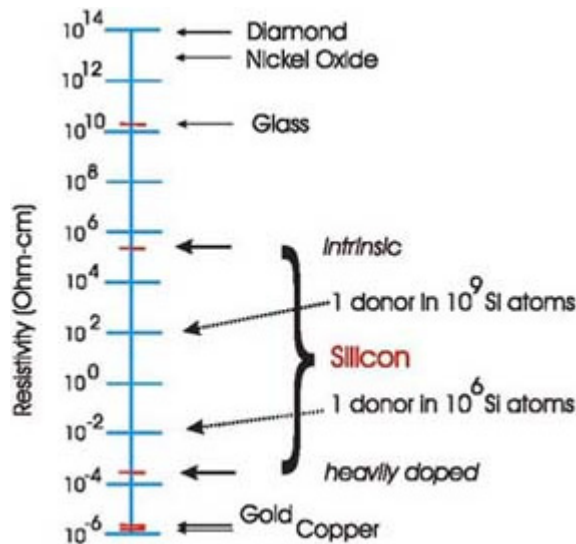


Figure C2.16.1. A nomogram comparing electrical resistivity of pure (intrinsic) and doped Si with metals and insulators.

C2.16.2 MATERIALS

There are hundreds of semiconductor materials, but silicon alone accounts for the overwhelming majority of the applications world-wide today. The families of semiconductor materials include tetrahedrally coordinated and mostly covalent solids such as group IV elemental semiconductors and III–V, II–VI and I–VII compounds, and their ternary and quaternary alloys, as well as more exotic materials such as the adamantane, non-adamantane and organic semiconductors. Only the key features of some of these materials will be mentioned here. For a more complete description, the reader is referred to specialized publications [6, 7, 8 and 9].

C2.16.2.1 ELEMENTAL SEMICONDUCTORS

The ‘group IV’ semiconductor materials are fourfold coordinated covalent solids from elements in column IV of the periodic table. The elemental semiconductors are diamond, silicon and germanium. They crystallize in the diamond lattice.

Diamond may never be used to make devices because it is nearly impossible to make it sufficiently n type, that is to obtain high electron concentration. Substitutional B is a good shallow acceptor, and interstitial Li has been reported to produce some n type conductivity.

Silicon is used in many forms, from high-purity thin films to bulk material, which may be crystalline, multi- or polycrystalline and amorphous (usually hydrogenated). Silicon is the material discussed the most in this article. Substitutional B and P are the most common (of many) shallow acceptors and donors, respectively.

-3-

Germanium is very similar to Si, but its band gap is too small for many practical applications. Large crystals of ultra-high-purity Ge have been grown for use as gamma-ray detectors. In such crystals, the net concentration of electrically active centres is incredibly low, of the order of 10^{12} cm^{-3} . Isotopically pure Ge crystals have been grown as well [10].

C2.16.2.2 COMPOUND SEMICONDUCTORS

There is a great number of mostly covalent and tetrahedral binary IV–IV, III–V, II–VI and I–VII semiconductors. Most crystallize in the zincblende structure, but some prefer the wurtzite structure, notably GaN [11, 12]. While the bonding in all of these compounds (and their alloys) is mostly covalent, some ionic character is always present because of the difference in electron affinity of the constituent atoms.

C2.16.2.3 IV – IVS

The ionic character of compounds of C, Si, and Ge [13] ranges from a few percent (SiGe) to as much as 16% (SiC). In addition to compounds, many alloys can be grown, such as $\text{C}_x\text{SiGe}_{1-x}$, where x is of the order of 0.02 to 0.04. Compounds such as $\text{Si}_x\text{Ge}_{1-x}$ are used because their gap can be made to vary from the Ge to the Si value.

Compounds and alloys of group IV elements normally have the zincblende structure. A notable exception is SiC which can crystallize in hundreds of polytypes that differ in the way the Si–C units are stacked along the c -axis of the crystal. For example, the *zincblende* structure (3C) has the sequence ABC–ABC–... with cubic symmetry, and the *wurtzite* structure (2H) is hexagonal with the sequence AB–AB–... (figure C2.16.2). The lowest-energy structure of SiC is 6H, with sequence ABCACB–ABCACB–.... In all these polytypes, each atom is fourfold coordinated and makes (almost or exactly) tetrahedral angles with its neighbours. In addition to the cubic and hexagonal polytypes, many other structures of SiC exist with rhombohedral or trigonal symmetries.

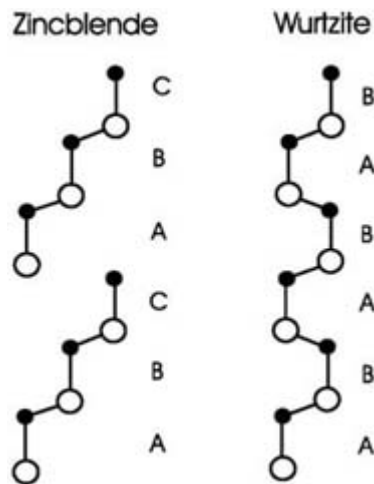


Figure C2.16.2. The sequence of atoms in the two polytypes of SiC, zincblende and wurtzite, along the *c*-direction. The zincblende lattice has perfect tetrahedral angles.

-4-

C2.16.2.4 III – VS

III–V compound semiconductors with precisely controlled compositions and gaps can be prepared from several material systems. Representative III–V compounds are shown in the gap–lattice constant plots of figure C2.16.3. The points representing binary semiconductors such as GaAs or InP are joined by lines indicating ternary and quaternary alloys. The special nature of the binary compounds arises from their availability as the substrate material needed for epitaxial growth of device structures.

Figure C2.16.3. A plot of the energy gap and lattice constant for the most common III–V compound semiconductors. All the materials shown have cubic (zincblende) structure. Elemental semiconductors, Si and Ge, are included for comparison. The lines connecting binary semiconductors indicate possible ternary compounds with direct gaps. Dashed lines near GaP represent indirect gap regions. The line from InP to a point marked * represents the quaternary compound InGaAsP, lattice matched to InP.

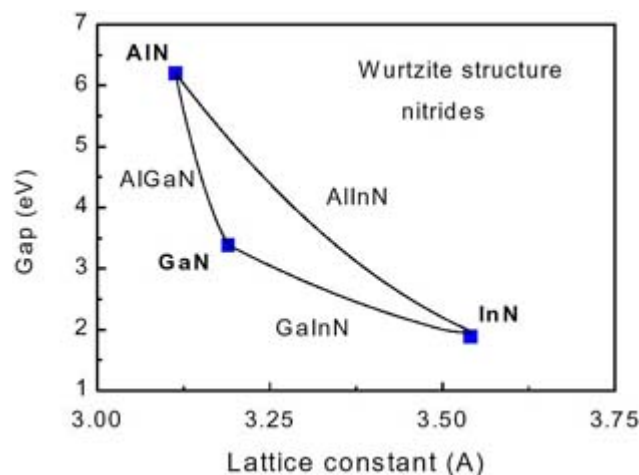


Figure C2.16.4. A plot of the energy gap and lattice constant for large-gap nitrides. These materials have wurtzite structure.

Ternary and quaternary semiconductors are theoretically described by the *virtual crystal approximation* (VCA) [7]. Within the VCA, ternary alloys with the composition $AB_{1-x}C_x$ are considered to contain two sublattices. One of them is occupied only by atoms A, the other is occupied by atoms B or C. The second sublattice consists of virtual atoms, represented by a weighted average of atoms B and C. Many physical properties of ternary alloys are then expressed as weighted linear combinations of the corresponding properties of the two binary compounds. For example, the lattice constant d dependence on composition is written as:

$$d = (1 - x)d_{AB} + xd_{AC}.$$

This approximation, known as *Vegard's law*, accurately describes the average lattice constant (but not the microscopic structure!) of most ternary compounds. However, the expression for the gap must be modified by the inclusion of a quadratic term

$$E_g(x) = (1 - x)E_g(AB) + xE_g(AC) - \Omega x(1 - x)$$

where the *bowing coefficient* Ω is positive. The term $-x(1 - x)$ is due to the random distribution of atoms B and C within their sublattice. It represents the probability of finding sequences of atoms B–A–C, or C–A–B, in the random alloy [14].

Some semiconductors with compositions close to $AB_{0.5}C_{0.5}$ are known to become ordered. This results in changes in the gap, and electrical and optical properties, compared to random alloys of the same composition.

Two of the material systems shown in [figure C2.16.3](#) are of particular importance. These are the ternary compounds formed from group III elements such as Al and Ga in combination with As and quaternary compounds formed from Ga and In in combination with As and P [8, 15]. Ternary $Al_xGa_{1-x}As$ grown on GaAs is the best known of the general class of compounds $A_x^{III}B_{1-x}^{III}C^V$. Quaternary $Ga_xIn_{1-x}As_{1-y}P_y$ grown on InP is representative of the general class $A_x^{III}B_{1-x}^{III}C_y^VD_{1-y}^V$. The lattice constants, gaps, indices of refraction and most other parameters of these materials depend on their composition x and y .

$Al_xGa_{1-x}As$ grown on GaAs is used for the preparation of *light-emitting diodes* (LEDs), *injection lasers* and bipolar *transistors*. The lattice constants of GaAs (0.565 nm) and AlAs (0.566 nm) are almost identical. Aluminum atoms can be substituted for Ga atoms in the GaAs lattice to form $Al_xGa_{1-x}As$ without significant change in the lattice constant. It is thus possible to vary the gap from $E_g(\text{GaAs}) = 1.43$ eV to $E_g(\text{AlAs}) = 2.16$ eV simply by adjusting the Al fraction x in the epitaxial layer. This feature of $Al_xGa_{1-x}As$ is quite unique among the compound semiconductors.

An even wider range of gaps can be reached with *lattice-mismatched* structures of $Ga_xIn_{1-x}As$ grown on GaAs. This ternary system is shown in [figure C2.16.3](#) by the line joining GaAs and InAs. The thickness of defect-free layers of $Ga_xIn_{1-x}As$ is limited by the biaxial compressive strain arising from the lattice mismatch with the substrate. Structures based on $Ga_xIn_{1-x}As$ are thus implemented in the form of *quantum wells*, typically less than 10 nm thick.

The usual acceptor and donor dopants for $Al_xGa_{1-x}As$ compounds are elements from groups II, IV and VI of the periodic table. Group II elements are acceptors and group VI elements are donors. Depending on the growth conditions, Si and Ge can be either donors or acceptor, i.e. *amphoteric*. This is of special interest in LEDs.

Quaternary $Ga_xIn_{1-x}As_{1-y}P_y$ grown on InP is of major importance to fibre-optic communications. In quaternary compounds, both the gap and the lattice constant can be tailored by changing the chemical composition. In thick layers, in order to avoid the generation of strain-induced defects, care must be taken in adjusting the ratio of x and y to maintain the lattice-matched composition $x = 2.2y$. The available gaps range from 1.34 eV in InP to ~0.75 eV in

lattice-matched $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$. A particularly interesting feature of this system is the high reliability of LEDs and diode lasers [16].

While quaternary layers and structures can be exactly lattice matched to the InP substrate, strain is often used to alter the gap or carrier transport properties. In $\text{Ga}_x\text{In}_{1-x}\text{As}$ or $\text{Ga}_x\text{In}_{1-x}\text{As}_{1-y}\text{P}_y$ grown on InP, strain can be introduced by moving away from the lattice-matched composition. In sufficiently thin layers, strain is accommodated elastically, without any change in the in-plane lattice constant. In this material, strain can be either compressive, with the lattice constant of the layer trying to be larger than that of the substrate, or tensile.

C2.16.2.5 III – V NITRIDES

Figure C2.16.2 shows the gap–lattice constant plots for the III–V nitrides. These compounds can have either the wurtzite or zincblende structures, with the wurtzite polytype having the most interesting device applications. The large gaps of these materials make them particularly useful in the preparation of LEDs and diode lasers emitting in the blue part of the visible spectrum. Unlike the smaller-gap III–V compounds illustrated in figure C2.16.3 single crystals of the nitride binaries of AlN, GaN and InN can be prepared only in very small sizes, too small for epitaxial growth of device structures. Substrate materials such as sapphire and SiC are used instead.

There is also a possibility of preparing mixed III–V nitride alloys, e.g. $\text{GaAs}_{1-y}\text{N}_y$, connecting the two sets of semiconductor materials. Their gap dependence on composition is the subject of active research.

C2.16.3 GENERAL PROPERTIES OF SEMICONDUCTORS

C2.16.3.1 ENERGY BANDS

The optical and electrical characteristics of semiconductors are conveniently described by energy level diagrams [17, 18, 19 and 20]. Electrons in atoms are restricted to sets of discrete energy states, separated by gaps in which electrons are not allowed. In solids, formed by bringing isolated atoms close together, the allowed energy levels of discrete atoms spread into essentially continuous energy *bands*. Two such bands are of particular importance in semiconductors: the highest-lying filled VB and the lowest-lying empty CB. The VB and CB are separated by the energy gap E_g .

In a defect-free, undoped, semiconductor, there are no energy states within the gap. At $T = 0$ K, all of the VB states are occupied by electrons and all of the CB states are empty, resulting in zero conductivity. The thermal excitation of electrons across the gap becomes possible at $T > 0$ and a net electron concentration in the CB is established. The electrons excited into the CB leave empty states in the VB. These holes behave like positively charged electrons. Both the electrons in the CB and holes in the VB participate in the electrical conductivity.

Calculated plots of energy bands as a function of wavevector k , known as band diagrams, are shown in figure C2.16.5 for Si and GaAs. Semiconductors can be divided into materials with *indirect* and *direct* gaps. In direct-gap

semiconductors (represented by GaAs) the minimum energy in the CB and the maximum in the VB occur for the same value of k , namely $k = 0$, the Γ point. This is not the case in indirect materials (represented by Si) in which the maximum of the VB occurs at $k = 0$ but the minimum of the CB at $k \neq 0$. This difference has profound consequences for the rates of electron–hole *recombination* across the gap.

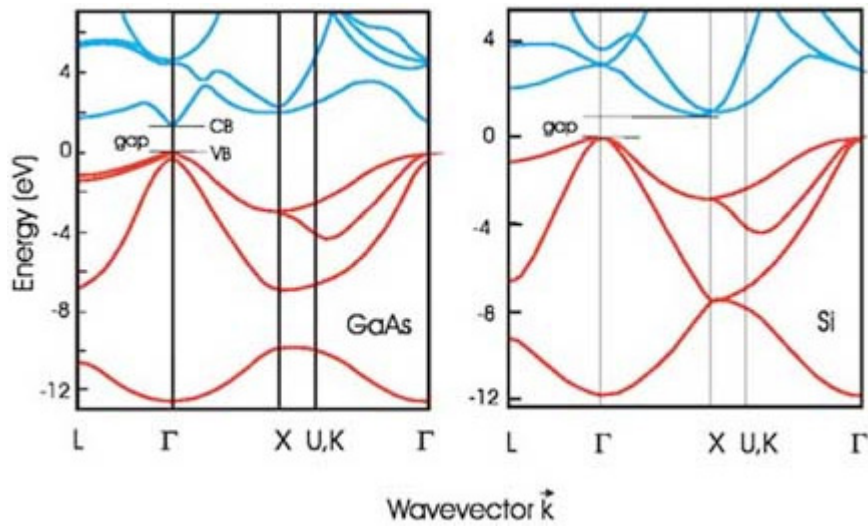


Figure C2.16.5. Calculated plots of energy bands as a function of wavevector k , known as band diagrams, for Si and GaAs. Indirect (Si) and direct (GaAs) gaps are indicated. High-symmetry points of the Brillouin zone are indicated on the wavevector axis.

All technologically important properties of semiconductors are determined by defect-associated energy levels in the gap. The conductivity of pure semiconductors varies as $\sigma \sim \exp(-E_g/k_B T)$, where E_g is the gap. In most semiconductors with practical applications, the size of the gap, $E_g \sim 1\text{--}2$ eV, makes the thermal excitation of electrons across the gap a relatively unimportant process. The introduction of shallow states into the gap through *doping*, with either donors or acceptors, allows for large changes in conductivity (figure C2.16.1). The donor and acceptor levels are typically a few meV below the CB and a few tens of meV above the VB, respectively. The depth of these levels usually scales with the size of the gap (see below).

C2.16.3.2 ELECTRICAL PROPERTIES

The application of a small external electric field E to a semiconductor results in a net average velocity component of the carriers (electrons or holes) called the *drift velocity*, v_d . The coefficient of proportionality between E and v_d is known as the carrier *mobility* μ . At higher fields, where the drift velocity becomes comparable to the thermal velocity of the carriers (which is about 10^7 cm s⁻¹ in Si at room temperature), the carriers decelerate by scattering with charged impurities and *lattice vibrations* (phonons and local vibrational modes). The simple linear relationship between E and v_d no longer applies. Thus the low-field mobility μ describes the mean free time between collisions. It depends strongly on the *effective mass* (see below) of the carriers, the purity of the semiconductor and the temperature. The effective mobility μ of carriers in a semiconductor reflects the contributions of various scattering mechanisms and is written as $1/\mu = 1/\mu_{\text{lattice}} + 1/\mu_{\text{impurity}}$. The temperature dependence of the mobility reflects these individual

contributions. For instance, since $\mu_{\text{lattice}} \sim T^{-3/2}$ and $\mu_{\text{impurity}} \sim T^{+3/2}$, most semiconductors show a peak in the mobility measured as a function of temperature [21]. In some of the modern semiconductor structures, it is possible to essentially eliminate μ_{impurity} and thus reach very high, metallic, mobilities at low temperatures [22].

In terms of the carrier mobility, the electrical conductivity σ of an n type semiconductor can be written as

$$\sigma = ne\mu$$

where n is the conduction electron density and e the electron charge. Since n is a strong (exponential) function of temperature, σ varies with temperature both through n and μ .

In thermal equilibrium at the temperature T , the distribution of electrons in the band is given by the Fermi–Dirac distribution function $f(E) = [1 + e^{(E-E_f)/kT}]^{-1}$, where k is the Boltzmann constant. The function $f(E)$ describes the probability that a state with an energy E is occupied at the temperature T . The quantity E_f , called the *Fermi level*, denotes the energy level with the occupation probability 1/2 at $T = 0$. At $T = 0$, all the available states below E_f are filled, $f(E < E_f) = 1$, all the states above E_f are empty, $f(E > E_f) = 0$. At $T = 0$, the Fermi level coincides with the chemical potential.

Instead of plotting the electron distribution function in the energy band diagram, it is convenient to indicate the position of the Fermi level. In a semiconductor of high purity, the Fermi level is close to mid-gap. In p type (n type) semiconductors, it lies near the VB (CB). In very heavily doped semiconductors the Fermi level can move into either the CB or VB, depending on the doping type.

The distributions of states in the CB and VB are described by the effective *density of states*. The concentration of electrons in the CB can be calculated as $n = \int_{E_c}^{\infty} f(E) N(E) d(E)$, where $f(E)$ is the Fermi distribution and $N(E) d(E)$ is the density of states between E and $E + dE$. A simpler way of calculating n is to represent all the electron states in the CB by an effective density of states N_c at the energy E_c (band edge). The electron density is then simply $n = N_c f(E_c)$.

Most of our ideas about carrier transport in semiconductors are based on the assumption of diffusive motion. When the electron concentration in a semiconductor is not uniform, the electrons move (*diffuse*) under the influence of concentration gradients, giving rise to an additional contribution to the current. In this motion, electrons also undergo collisions and their temporal and spatial distributions are described by the *diffusion equation*. The proportionality constant between the flux and the concentration gradient is called *diffusivity*, D ($\text{cm}^2 \text{s}^{-1}$). Diffusivity and mobility are related by Einstein's relationship $D = (kT/q)\mu$, where q is the carrier charge. In the context of diffusive motion, it is also straightforward to introduce the carrier lifetime τ as the average time between collisions and define a diffusion length $L = \sqrt{D\tau}$. This concept is particularly useful to describe *minority carriers*. The minority carrier diffusion length varies from the sub-micrometre in heavily doped semiconductors to tens of micrometres in high-purity materials.

C2.16.3.3 RECOMBINATION

In n type semiconductors, electrons are the *majority carriers*. Holes will also be present through accidental incorporation of acceptor impurities or, more importantly, through the intentional creation of electron-hole pairs. Holes in n type and electrons in p type semiconductors are *minority carriers*.

There are many ways of increasing the equilibrium carrier population of a semiconductor. Most often this is done by generating electron–hole pairs as, for instance, in the process of absorption of a photon with $h\omega \geq E_g$. Under reasonable levels of illumination and doping, the generation of electron–hole pairs affects primarily the minority carrier density. However, the excess population of minority carriers is not stable: it gradually disappears through a variety of *recombination* processes in which an electron in the CB fills a hole in a VB. The excess energy E_g is released as a photon or phonons. The former case corresponds to a *radiative* recombination process, the latter to a *non-radiative* one. The radiative processes only rarely involve direct recombination across the gap. Usually, this type of process is assisted by shallow defects (impurities). Non-radiative recombination involves a defect-related deep level at which a carrier is trapped first, and a second transition is needed to complete the process.

Radiative recombination of minority carriers is the most likely process in direct gap semiconductors. Since the carriers at the CB minimum and the VB maximum have the same momentum, very fast recombination can occur. The radiative recombination lifetimes in direct semiconductors are thus very short, of the order of the ns. The presence of deep-level defects opens up a non-radiative recombination path and further shortens the carrier lifetime.

The situation is very different in indirect gap materials where phonons must be involved to conserve momentum. Radiative recombination is inefficient, resulting in long lifetimes. The minority carrier lifetimes in Si reach many ns, again in the absence of defects. It should be noted that long minority carrier lifetimes imply long diffusion lengths. Minority carrier lifetime can be used as a convenient quality benchmark of a semiconductor.

C2.16.4 DEFECTS AND IMPURITIES

Intrinsic defects (or ‘native’ or simply ‘defects’) are imperfections in the crystal itself, such as a vacancy (a missing host atom), a self-interstitial (an extra host atom in an otherwise perfect crystalline environment), an anti-site defect (in an AB compound, this means an atom of type A at a B site or vice versa) or any combination of such defects. *Extrinsic* defects (or impurities) are atoms different from host atoms, trapped in the crystal. Some impurities are intentionally introduced because they provide charge carriers, reduce their lifetime, prevent the propagation of dislocations or are otherwise needed or useful, but most impurities and defects are not desired and must be eliminated or at least controlled.

The presence of defects and impurities is unavoidable. They are created during the growth or penetrate into the material during the processing. For example, in a crystal grown from the melt, impurities come from the crucible and the ambient, and are present in the source material. Depending on factors such as the pressure, the pull rate and temperature gradients, the crystal may be rich in vacancies or self-interstitials (and their precipitates).

After the growth, virtually all the processing steps create defects and/or add impurities. Ion implantation and electron irradiation create vacancy–self-interstitial pairs (*Frenkel pairs*). Wet or dry etching, the deposition of organic masks, metallic contacts, or other surface layers, furnace or rapid thermal anneals and other processes also result in defects or impurities penetrating into the material or diffusing through the bulk.

Experimentally, local vibrational modes associated with a defect or impurity may appear in infra-red absorption or Raman spectra. The defect centre may also give rise to new photoluminescence bands and other experimentally observable signature. Some defect-related energy levels may be visible by deep-level transient spectroscopy (DLTS) [23].

C2.16.4.1 NOMENCLATURE

Most electrical and optical properties of semiconductors are determined by the impurities and defects they contain. The underlying reason for this is that any kind of imperfection in the crystal means a different local potential and therefore new energy eigenvalues (*energy levels*) and eigenfunctions associated with the defect or impurity and its immediate environment. If the new levels are in the gap, as is often the case, they give rise to some electrical activity. The energy levels can be *shallow* (near a band edge) or *deep* (far from a band edge). ‘Hyper-deep’ refers to a localized level in the VB.

A gap level is called an *acceptor level* if the defect is neutral when the state is empty (no electron). It is called a *donor level* if the defect is neutral when the state is occupied (one electron). The former is often labelled (0 / –) and the latter (+ / 0), where the first (second) sign refers to the charge of the defect when no electron (one electron) is present. Double or triple acceptor and donor levels are similarly labelled.

Common terminology used to characterize impurities and defects in semiconductors includes point and line defects, complexes, precipitates and extended defects. These terms are somewhat loosely defined, and examples follow.

A *point defect* refers to a localized defect (such as a monovacancy) or impurity (such as interstitial O). This includes any relaxation and/or distortion of the crystal around it. Many point defects are now rather well understood, especially in Si, thanks to a combination of experiments providing information of microscopic nature

(such as electron paramagnetic resonance, local vibrational mode spectroscopy, or photoluminescence) and ‘*ab initio*’ or ‘first-principles’ theory (which means that no experimentally adjusted parameters are used). Tremendous progress has been achieved since the mid-1980s in the theory of defects in semiconductors, to the point where nearly quantitative predictions are common.

A combination of a small number of point defects is called a *complex*. Examples are the boron–hydrogen pair in Si, a small cluster of vacancies such as the hexavacancy, or a C–C pair. *Precipitates* refer to more complicated and larger aggregates of impurities or defects, such as the O-related thermal donors in Si, or metallic impurities trapped at some internal void. Such precipitates can be of substantial size, involving anywhere from a handful to thousands of atoms. They can be permanent sinks for specific impurities (such as SiO₂ precipitates in Si) or serve as sources of impurities or defects, such as the {311} platelets of self-interstitials in Si. Such defects are difficult to study. The number of degrees of freedom and local minima of a multi-dimensional potential energy surface render the theorist’s task very challenging. Experimentalists must deal with complicated spectra and often broad lines. As a result, only a few small complexes are well understood, and almost nothing microscopic is known about large complexes and precipitates. For example, despite almost 50 years of experimental studies and theoretical modelling, the structure and nature of O-related thermal donors in Si is still unknown [24, 25].

Extended defects range from well characterized dislocations to grain boundaries, interfaces, stacking faults, etch pits, D-defects, misfit dislocations (common in epitaxial growth), blisters induced by H or He implantation etc. Microscopic studies of such defects are very difficult, and crystal growers use years of experience and trial-and-error techniques to avoid or control them. Some extended defects can change in unpredictable ways upon heat treatments. Others become gettering centres for transition metals, a phenomenon which can be desirable or not, but is always difficult to control. Extended defects are sometimes cleverly used. For example, the ‘smart-cut’ process relies on the controlled implantation of H followed by heat treatments to create blisters. This allows a thin layer of clean material to be lifted from a bulk wafer [26].

Point defects and complexes exhibit *metastability* when more than one configuration can be realized in a given charge state. For example, neutral interstitial hydrogen is metastable in many semiconductors: one configuration has H at a relaxed bond-centred site, bound to the crystal, and the other has H atomic-like at the tetrahedral interstitial site.

Bistability refers to defects which have different configurations in different charge states. The transition to a bistable state can be induced by the exposure to band-gap light for example. One example is the EL2 centre in GaAs [27]. A defect or impurity is called *negative U* when the energy gained by pairing two electrons (or holes) at the defect is greater than the Coulomb repulsion between them. The best known example is the vacancy (V) in Si which is stable in the spin singlet ++, 0 and – – charge states but unstable in the spin doublet + and – states [28].

The energetics associated with a metastable and/or bistable defect are often described using a *configuration diagram*. It is a semi-quantitative plot of the energy against a global coordinate which combines the position of the impurity and all the relaxations and distortions of the crystal, which can be substantial. The configuration diagram in figure C2.16.6 was obtained from muon spin rotation (μ SR) data in Si [29], and relates to the states of muonium (Mu), a light isotope of hydrogen. This example illustrates configuration diagrams, acceptor and donor levels, metastability, bistability and negative-*U* behaviour.

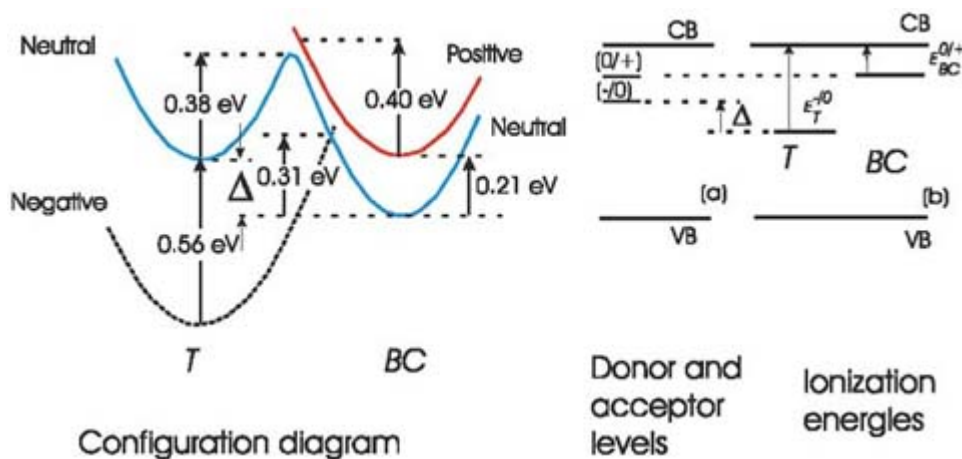


Figure C2.16.6. The energy states of a metastable and bistable muonium in Si are illustrated in a *configuration diagram*. It plots the defect energy as a function of a coordinate which combines position and all the relaxations and distortions of the crystal. The specific example, discussed in the text, illustrates acceptor and donor levels, metastability, bistability and negative- U [50] behaviour.

The configuration diagram consists of three (potential) energy curves associated with the three charge states of hydrogen in Si. The two types of minima correspond to the impurity at the tetrahedral interstitial (T) and at the bond-centred (BC) sites, respectively. There is little relaxation of the host crystal when hydrogen is at the T site, but the BC site is a minimum of the energy only after an Si–Si bond relaxes substantially to become the bridged Si–H–Si bond. The H_T^- state exists only in n type Si and is stable only at the T site. The H_{BC}^+ state is observed predominantly in p type Si, and is only stable at the BC site. The neutral state is found in two configurations, at the BC site (stable) and the T site (metastable). The energy difference between these two sites is shown as Δ , the value of which is estimated at a few tenths of an eV. The μ SR data show that all three charge states coexist at the microsecond time scale above room

-12-

temperature and there is experimental evidence that this is the case for H as well. The energy differences marked in the figure were either measured directly (ionizations) or deduced from fitting the data.

The ionization energies and impurity levels are shown in the flat-band figure next to the configuration diagram. The donor level (+ / 0) corresponds to the ionization energy $H_{BC}^0 \rightarrow H_{BC}^+$ since the transition occurs between essentially identical configurations. The ionization energy measured by μ SR is very close to the donor level obtained for hydrogen by DLTS [30], 0.175 ± 0.005 eV.

The situation is more complicated for the acceptor level, as the measured ionization energy corresponds to a transition from the stable state H_T^- to the metastable state H_T^0 . However, the acceptor level corresponds to the energy difference between stable states $H_T^- \rightarrow H_{BC}^0$ that is the ionization energy corrected by Δ . The μ SR data therefore imply that muonium (or hydrogen) is negative U if $\Delta < 0.35$ eV, and positive U otherwise. The energy difference U between the acceptor and donor levels is in any case quite small, whether positive or negative. Note that quantum effects (such as the zero-point energy) do play a role in this case, as the impurity (a muon of hydrogen) is very light.

C2.16.4.2 SHALLOW LEVELS

Shallow impurities have energy levels in the gap but very close to a band. If an impurity has an empty level close to the VB maximum, an electron can be thermally promoted from the VB into this level, leaving a hole in the VB. Such an impurity is a *shallow acceptor*. On the other hand, if an impurity has an occupied level very close to the CB minimum, the electron in that level can be thermally promoted into the CB where it participates in the conductivity. Such an impurity is a *shallow donor*.

There are other, more exotic, possibilities. For example, if a defect has an empty level near the CB, an electron may become trapped in it. This localized electron may in turn bind a hole in a loose orbit, forming a *bound exciton*.

Shallow donors (or acceptors) add new electrons to the CB (or new holes to the VB), resulting in a net increase in the number of a particular type of charge carrier. The implantation of shallow donors or acceptors is performed for this purpose. But this process can also occur unintentionally. For example, the precipitation around 450°C of interstitial oxygen in Si generates a series of shallow double donors called thermal donors. As-grown GaN crystal are always heavily n type, because of some intrinsic shallow-level defect. The presence and type of new charge carriers can be detected by Hall effect measurements.

Since shallow-level impurities have energy eigenvalues very near those of the perfect crystal, they can be described using a perturbative approach first developed in the 1950s and known as *effective mass theory* (EMT). The idea is to approximate the band nearest to the shallow level by a parabola, the curvature of which is characterized by an effective mass parameter m^* .

The simplest example is that of the shallow P donor in Si. Four of its five valence electrons participate in the covalent bonding to its four Si nearest neighbours at the substitutional site. The energy of the fifth electron which, at ~0 K, is in an energy level just below the minimum of the CB, is approximated by $\hbar^2 k^2 / 2m^*$ plus the screened Coulomb attraction to the P^+ ion, $e^2 / \epsilon r$, where ϵ is the dielectric constant or the frequency-dependent dielectric function. The Schrödinger equation for this electron reduces to that of the hydrogen atom, but m^* replaces the electronic mass and ϵ screens the Coulomb attraction.

-13-

In Si, the binding energy is reduced to ~20–40 meV independently of the shallow donor. The solution further yields a hydrogenic series of levels analogous to the 1s, 2sp etc states of atomic hydrogen. This series bridges the shallow level of the donor and the CB minimum, facilitating the thermal ionization of the electron into the CB. This ionization, which begins below liquid nitrogen temperature, is what provides free electrons to the CB.

The Bohr radius is very large, 3–5 nm, and the shallow impurity wavefunction extends over a large portion of the crystal. Doping up to the ‘metallic limit’ consists in implanting a sufficiently high concentration of donors so that the shallow-donor wavefunctions overlap, creating a half-filled impurity band in which the electrons move freely.

C2.16.4.3 DEEP LEVELS

If the level(s) associated with the defect are deep, they become electron–hole recombination centres. The result is a (sometimes dramatic) reduction in carrier lifetimes. Such an effect is often associated with the presence of transition metal impurities or certain extended defects in the material. For example, substitutional Au is used to make fast switches in Si. Many point defects have deep levels in the gap, such as vacancies or transition metals. In addition, complexes, precipitates and extended defects are often associated with recombination centres. The presence of grain boundaries, dislocation tangles and metallic precipitates in poly-Si photovoltaic devices are major factors which reduce their efficiency.

Deep-level defects cannot be described by EMT or be viewed as simple perturbations to the perfect crystal. Instead, the full crystal-plus-defect problem must be solved and the geometries around the defect optimized to account for lattice relaxations and distortions. The study of deep levels is an area of active research.

In order to remove the unwanted electrical activity associated with deep-level impurities or defects, one can either physically displace the defect away from the active region of the device (*gettering*) or force it to react with another impurity to remove (or at least change) its energy eigenvalues and therefore its electrical activity (*passivation*).

Gettering is a black art. It consists in forcing selected impurities (typically, transition metals) to diffuse toward unimportant regions of the device. This is often done by creating precipitation sites and performing heat treatments. The precipitation sites range from small oxygen complexes to layers such as an Al silicide. The formation of such a

metallization injects vacancies into the bulk and they enhance the diffusivity of some transition metals. Phosphorus gettering occurs during the heat treatment that follows the implantation of a heavily doped n^+ layer in Si. However, boron-rich buried layers and H-induced platelets are efficient gettering sites as well, because transition metals are much more stable at such defective regions than dissolved in the bulk. Recent success in achieving copper contacts on Si involved creating a gettering layer (TiN) in the subsurface region to precipitate Cu and keep it out of the active region of the device.

Passivation involves mostly the use of hydrogen [31]. H is a rapid diffuser in most semiconductors and is unavoidable. It may be present in an ambient (water vapour for example) and in many processing steps such as the deposition of Schottky contacts or antireflection coatings, the use of organic masks etc. Hydrogen diffuses and traps at a range of impurities and defects. The trapping always involves some covalent interaction, a change in the configuration and electronic structure of the complex and a shift in its energy eigenvalues. Passivation results when an energy level shifts from the gap into a band. The thermal stability of most H–impurity pairs is normally rather low (a few hundred degrees Celsius at the most), but that of H–defect pairs tends to be higher. This is why H is used to passivate grain boundaries and other defects in poly-Si solar cells for example. In GaN, hydrogen passivates the shallow Mg acceptor with an

-14-

unusually high thermal stability. In order to obtain p type GaN, the material is often annealed at high temperatures to dissociate the {H, Mg} pairs and diffuse H out of the crystal.

C2.16.4.4 DIFFUSION

In addition to the configuration, electronic structure and thermal stability of point defects, it is essential to know how they diffuse. A variety of mechanisms have been identified. The simplest one involves the diffusion of an impurity through the interstitial sites. For example, copper in Si diffuses by hopping from one tetrahedral interstitial site to the next via a saddle point at the hexagonal interstitial site.

However, most impurities and defects are Jahn–Teller unstable at high-symmetry sites or/and react covalently with the host crystal much more strongly than interstitial copper. The latter is obviously the case for substitutional impurities, but also for interstitials such as O (which sits at a relaxed, puckered bond-centred site in Si), H (which bridges a host atom–host atom bond in many semiconductors) or the self-interstitial (which often forms more exotic structures such as the ‘split- $\langle 110 \rangle$ ’ configuration). Such point defects migrate by breaking and re-forming bonds with their host, and phonons play an important role in such processes.

The vacancy is very mobile in many semiconductors. In Si, its activation energy for diffusion ranges from 0.18 to 0.45 eV depending on its charge state, that is, on the position of the Fermi level. While the equilibrium concentration of vacancies is rather low, many processing steps inject vacancies into the bulk: ion implantation, electron irradiation, etching, the deposition of some thin films on the surface, such as Al contacts or nitride layers etc. Such non-equilibrium situations can greatly affect the mobility of impurities as vacancies flood the sample and trap interstitials.

Self-interstitials are also mobile. In Si, the activation energy for diffusion is believed to be of the order of the eV, but this drops to zero when minority charge carriers are present. This is probably due to *recombination-enhanced diffusion*, a process in which the defect itself is a recombination centre for electrons and holes. The recombination releases an energy about equal to size of the gap and this energy is used to propel the impurity above a diffusion barrier. Since many processing steps which inject self-interstitials also inject minority carriers, self-interstitials tend to diffuse excessively fast during the processing step and react with impurities and defects. They often kick out substitutional impurities, thus transforming a slow-diffusing impurity into a rapidly diffusing one. Important examples include boron in Si and the 3d transition metals which diffuse much faster as interstitial than substitutional impurities.

More exotic diffusion processes have been identified, although they may not be fully understood. One example is the substantial enhancement [25] of the diffusivity of interstitial O by H, resulting in the increased formation rate of

C2.16.5 STRUCTURES AND DEVICES

C2.16.5.1 P – N JUNCTION

In order to obtain appreciable conductivities, semiconductors must be doped with small amounts of selected impurities. It is possible to switch the doping type from n to p type, or vice versa, either during the growth of a crystal or by the selective introduction of impurities after the growth. The boundary region between the p type and n type regions is

-15-

called a *p-n junction* [17, 32, 33, and 34].

The detailed spatial distribution of carriers in the immediate vicinity of the p–n junction is very important. Some of the majority carriers at the junction neutralize each other. This results in a thin region depleted of free carriers, known as the *space-charge region*. In this region, negatively ionized acceptors on the p side repel the mobile electrons from the n side of the junction. Similarly, the mobile holes from the p side are repelled by the positively ionized donors on the n side. The result is a built-in electric field which inhibits carrier diffusion across the p–n junction. The electrostatic potential, or *contact potential*, associated with this field bends the conduction and valence bands in the space-charge region by an amount called the *barrier height*, as is illustrated in figure C2.16.7.

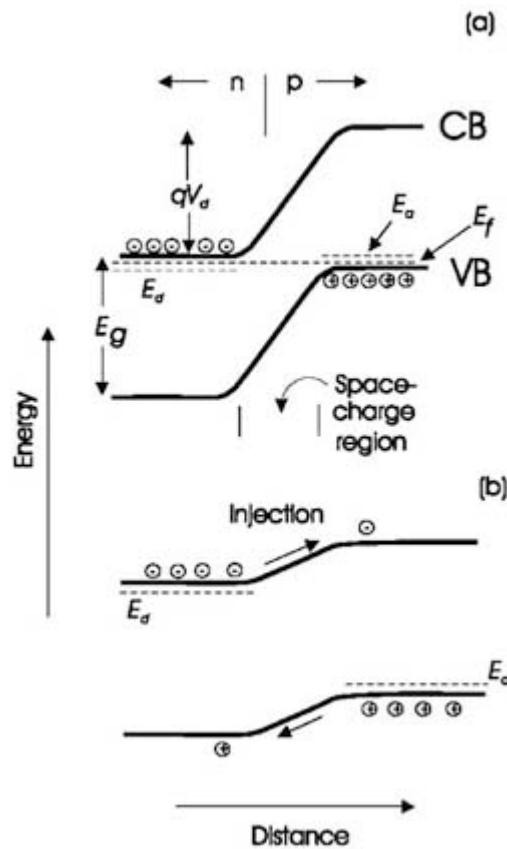


Figure C2.16.7. A schematic energy band diagram of a p–n junction without external bias (a) and under forward bias (b). Electrons and holes are indicated with – and + signs, respectively. It should be remembered that the energy of electrons increases by moving up, holes by moving down. Electrons injected into the p side of the junction become minority carriers. Approximate positions of donor and acceptor levels and the Fermi level, are indicated.

V_d is the built-in potential of the p–n junction.

A current flow across the p–n junction can be accomplished only by the application of an external voltage which opposes the contact potential and reduces the barrier height. This voltage, called a *forward bias*, supplies electrons at the n and holes at the p contact. Since the barrier height is similar in magnitude to the gap energy, the external bias needs to be fairly small, on the order of the gap energy divided by the electron charge.

-16-

Under a forward bias, the majority carriers cross the p–n junction and become minority carriers, e.g. holes on the n side of the junction. These holes rapidly come into thermal equilibrium with the crystal and reach the energy approximately equal to the energy of the valence band edge on the n side of the junction (it should be remembered that electrons moving up on the energy band increase their energy while holes increase their energy by moving down). Similarly, the minority electrons reach equilibrium at the p side of the junction. However, the minority carriers are not in thermodynamic equilibrium with the majority carriers and must give up their excess energy. A hole on the n side recombines with a majority electron, which then loses an energy about equal to E_g . The minority electron loses a similar energy by recombining with a majority hole.

The distributions of excess, or *injected*, carriers are indicated in band diagrams by so-called quasi-Fermi levels for electrons (E_{fn}) or holes (E_{fp}). These functions describe steady state concentrations of excess carriers in the same form as the equilibrium concentration. In equilibrium we have $E_{fn} = E_{fp} = E_f$.

C2.16.5.2 PHOTODETECTORS AND SOLAR CELLS

The current–voltage characteristic of an ideal p–n junction is $I = I_s[\exp(qV/mkT) - 1]$, where q is the electron (hole) charge, m is the ideality factor, $1 \leq m \leq 2$, and I_s the saturation current. Under applied reverse bias, the current through the junction is limited to I_s . Illuminating a reverse-biased diode with photons of energy greater than E_g results in the generation of photocarriers. The photocurrent is proportional to the intensity of the incident light. In low-doped diodes, I_s can be quite low and even small amounts of light can be detected. The diode is being operated as a *photodetector*. Very sophisticated detector structures have been designed, particularly those relying on high-field *avalanche multiplication* of photocarriers [35].

The p–n junction diode can also be used to convert optical energy directly to electrical power, without external power supplies. Absorption of a photon with $E > E_g$ produces an e^-h^+ pair. The internal electric field of the p–n junction separates the carriers; the e^- and h^+ move toward metallic contacts on opposite sides of the cell. The resulting photocurrent is sent to an external load. The maximum power delivered to a load is obtained under small forward bias. In Si cells, the largest voltage output produced in the open circuit mode (i.e. with $I = 0$) is about 0.7 V. Si solar cell power efficiencies as high as 24% have been reported, close to the theoretical limit of 32%. The power generated depends on the design of the diode itself and a match to the electrical load. The gap of the cell's semiconductor must match the solar spectrum as closely as possible and the structure of the gap should allow for efficient absorption of the solar photons. Perfection of the semiconductor material is also very important for high efficiency. Electrons and holes generated far from the electrodes must be extracted from the bulk of the cell and this requires long minority carrier diffusion lengths.

C2.16.5.3 LIGHT EMITTING DIODES

Light is generated in semiconductors in the process of radiative recombination. In a direct semiconductor, minority carrier population created by injection in a forward biased p–n junction can recombine radiatively, generating photons with energy about equal to E_g . The recombination process is *spontaneous*: individual electron–hole recombination events are random and not related to each other. This process is the basis of LEDs [36].

LEDs can be now fabricated in all primary colours and with efficiencies much higher than those of light bulbs. Red LEDs, based on InGaAlP, are sufficiently bright to be used in traffic lights and automobile brake lights. Blue LEDs, based on GaInN, have become commercially available in the last few years. White light can be produced by blue LEDs

pumping appropriate phosphor materials. Inexpensive and very reliable infrared LEDs are common in communication systems.

C2.16.5.4 BIPOLAR TRANSISTORS

The bipolar junction transistor (BJT) consists of three layers doped n–p–n or p–n–p that constitute the emitter, base and collector, respectively. This structure can be considered as two back-to-back p–n junctions. Under normal operation, the emitter–base junction is forward biased to inject minority carriers into the base region. For example, the n type emitter injects electrons into a p type base. The electrons in the base, now minority carriers, diffuse through the base layer. The base–collector junction is reverse biased and its electric field sweeps the carriers diffusing through the base into the collector. The BJT operates by transport of minority carriers, but both electrons and holes contribute to the overall current.

A band diagram of a biased n–p–n BJT is shown in figure C2.16.8. Under forward bias, electrons are injected from the n type emitter, giving rise to the current $I_{E,n}$ flowing into the p type base. Some of the carriers injected into the base recombine in the base or at the surface. This results in a reduction of the base current by I_r , the lost recombination current, and the base current becomes $I_B = I_{E,n} - I_r$. At the same time, holes are injected from the base into the emitter giving rise to the $I_{E,p}$ component of the current. The two components, I_r and $I_{E,p}$, must be minimized since they reduce the collector current $I_C = (I_{E,n} - I_{E,p}) - I_r$. The ‘current gain’ of the transistor is the ratio I_C/I_B . Since the base current is much smaller than the collector current, large gains are possible.

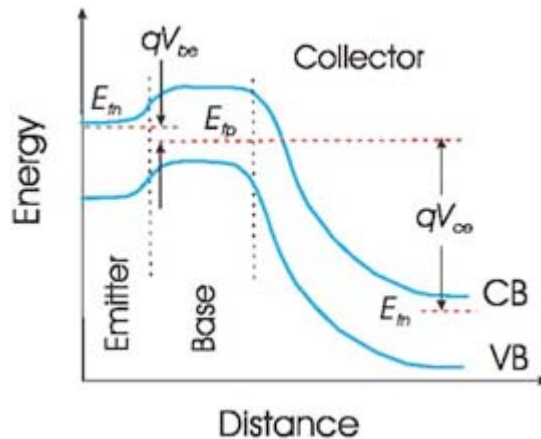


Figure C2.16.8. Schematic energy band diagram for an n–p–n bipolar junction transistor. Positions of quasi-Fermi levels and bias voltages are indicated.

In a BJT, the width of the base region must be smaller than the minority carrier diffusion length, and the two junctions strongly influence each other. Small increases in the current flowing through the first junction result in large increases in the collector current. The ratio of the current at the collector to the base current is the current gain. In Si-based BJTs, the gain is limited by the injection of the base majority carriers into the emitter. In order to maintain an adequate current gain, the doping level of the base must be lower than that of the emitter. Unfortunately, this results in higher base resistance and larger *RC* time constant of the transistor.

Some of these problems are avoided in *heterojunction* bipolar transistors (HBTs) [37, 38], the majority of which are based on III–V compounds such as GaAs/AlGaAs. In an HBT, the gap of the emitter is larger than that of the base. The conduction and valence band offsets that result from the matching up of the two different materials at the heterojunction prevent or reduce the injection of the base majority carriers into the emitter. This permits the use of

the highly doped, very thin bases needed to achieve high gain and high speed of operation at the same time. The conduction band offset at the emitter–base interface gives rise to the possibility of injecting carriers with high kinetic energy and *ballistic* transport in the base.

C2.16.5.5 METAL – OXIDE – SEMICONDUCTOR TRANSISTORS

Metal–oxide–semiconductor field-effect transistors (MOSFETs) are the basic devices of modern electronics [32, 33 and 34]. They can be produced in large numbers with very reproducible characteristics and their low power consumption makes large-scale integrated circuits possible. They are majority carrier devices and thus relatively insensitive to materials defects. A schematic cross section of a MOSFET is shown in [figure C2.16.9](#). The heart of the device is the metal–oxide–semiconductor (MOS) *gate* structure that controls the current flow in the channel. In normal operation of an *n-channel enhancement mode* device, the gate bias is used to produce a conducting channel between two contact regions known as the *source* and *drain*. A positive bias attracts electrons into the channel; a negative bias repels them. For low source–drain bias, the conductivity of the channel is proportional to the gate voltage, the channel width is uniform from source to drain and the transistor behaves as a voltage-controlled resistor. As the drain bias approaches a threshold value (V_p) the voltage drop across the oxide, in the vicinity of the drain, decreases. This results in a non-uniform channel width, lower at the drain, and overall higher resistance. For a drain bias greater than V_p , the channel is pinched off at the drain. Increased source–drain voltage does not produce any incremental source–drain current. At this point, the slope of the output current versus source–drain bias curve is zero. This bias region is known as the *saturation* region. In this region, small changes in the gate voltage produce large changes in the drain current and the device is said to have large *transconductance*.

The performance of MOSFETs depends critically on the gate dimension measured along the channel (the *gate length*), and the gate oxide thickness. Higher transconductance is obtained for shorter gate length and thinner oxides. State-of-the-art devices use gate length shorter than 200 nm and gate oxides thinner than 10 nm.

The carriers in the channel of an enhancement mode device exhibit unusually high mobility, particularly at low temperatures, a subject of considerable interest. The source–drain current is carried by electrons attracted to the interface. The ionized dopant atoms, which act as fixed charges and limit the carriers' mobility, are left behind, away from the interface. In a sense, the source–drain current is carried by the two-dimensional (2D) electron gas at the Si–gate oxide interface.

Electron effective masses are much smaller in compound semiconductors and quantum effects in 2D gas much more pronounced. It is also easier to engineer the 2D electron (or hole) gas by the use of *heterostructures* and *modulation doping*. The channel of a compound semiconductor FET can be formed by two layers with different gaps, for instance GaAs and InGaAs [39]. An undoped *spacer layer* may be introduced at the interface. The electrons associated with the donors in the wider-gap layer see lower-lying energy states in the adjacent narrower-gap material and transfer to it. The electrons and positively charged donors become spatially separated, effectively eliminating ionized impurity scattering in the narrower-gap channel. High-mobility electrons can be then maintained with high sheet charge densities down to very low temperatures.

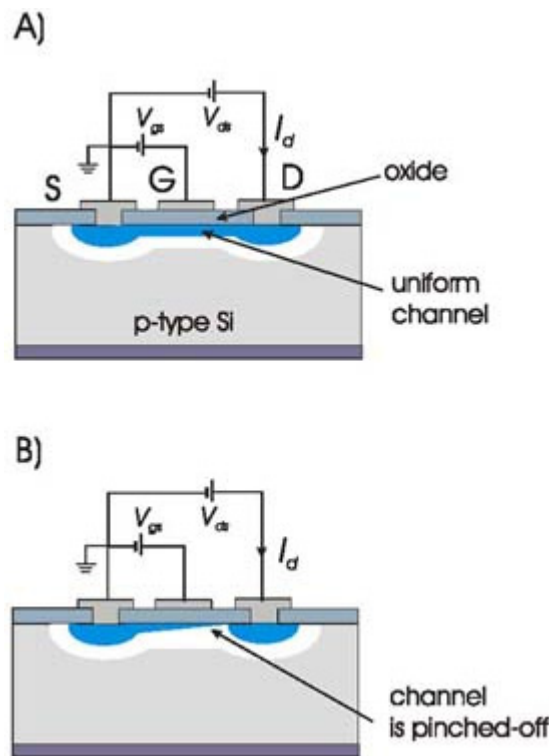


Figure C2.16.9. Schematic cross-section and biasing of a metal–oxide–semiconductor transistor. A uniform conducting channel is induced between source (S) and drain (D) for $V_{gs} > V_t$. Voltage V_{gs} is applied between the gate (G) and the source. Part (A) shows the channel for $V_{ds} < V_{gs} - V_t$; the transistor acts as a triode. The source–drain voltage is increased in part (B) to $V_{ds} > V_{gs} - V_t$. The channel is now pinched off at the drain side and I_d is saturated. This is the proper regime of operation of the MOS transistor.

C2.16.5.6 HETEROSTRUCTURES

In the p–n junction illustrated in [figure C2.16.7](#) both sides are made of the same semiconductor, and therefore have the same energy gap. Such a junction is called a *homojunction*. In a homojunction, the minority carriers are free to move away from the junction by a few diffusion lengths. It is therefore very difficult to achieve high carrier densities. Significantly higher carrier densities can be obtained by introducing an energy barrier at, or very near, the p–n junction. The energy barrier arises when two different semiconductors (therefore with different gaps) are joined [[8](#), [40](#)]. Barriers in the CB and VB, called the *band-edge discontinuities* ΔE_c and ΔE_v , are formed by changing the composition of the semiconductor layers. The junction of a small- with a large-gap semiconductor is called a *single heterojunction*. It confines one type of minority carrier (electrons or holes) to the p–n junction region.

A more effective carrier confinement is offered by a *double heterostructure* in which a thin layer of a low-gap material is sandwiched between larger-gap layers. The physical junction between two materials of different gaps is called a *heterointerface*. A schematic representation of the band diagram of such a structure is shown in [figure C2.16.10](#). The electrons, injected under forward bias across the p–n junction into the lower-bandgap material, encounter a potential barrier ΔE_c at the p–p junction which inhibits their motion away from the junction. The holes see a potential barrier of

ΔE_v at the n–p heterointerface which prevents their injection into the n region. The result is that the injected minority carriers are confined to the thin narrow-bandgap region. If this region is thinner than the average *diffusion*

length, very high densities of injected carriers can be obtained in a forward-biased diode. Heterojunctions are the basic structures of LEDs, semiconductor lasers and HBTs.

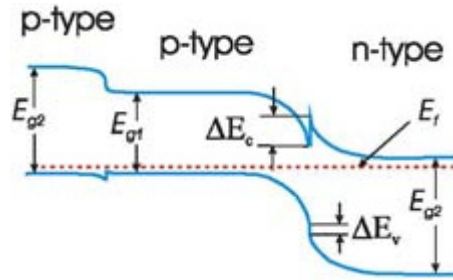


Figure C2.16.10. Band diagram of a double p-p-n double heterostructure without the external bias. The gap E_{g1} is smaller than E_{g2} . Conduction and valence band discontinuities are indicated. Structures of this type are used in LEDs and diode lasers.

C2.16.5.7 QUANTUM WELL STRUCTURES

Advances in epitaxial crystal growth methods make it possible to prepare heterostructures with essentially arbitrary thickness of the small-gap layer. When the thickness of this layer is reduced to dimensions of the order of 10 nm (between 20 and 30 atomic planes) a quantum mechanical description of the confined carriers is needed. Such heterostructures are called *quantum wells* [41, 42].

In quantum wells, Heisenberg's uncertainty principle requires an increase in the carrier energy over the equilibrium energy of the bulk semiconductor. The confined carriers are allowed only a few discrete states, with energies inversely proportional to the carrier's effective mass and the square of the well width. The incorporation of quantum wells into the material has a number of subtle consequences. The VB in direct bulk semiconductors is degenerate at the Γ point (see figure C2.16.5, resulting in two types of holes with the same energy, the *heavy holes* and *light holes*. The effective mass of the light hole is similar to that of the electron, while that of the heavy hole is typically ten times larger than that of the electron. In quantum wells, the degeneracy of the two hole bands is lifted. The energy shift due to quantum well confinement is larger for the light holes. This has important consequences for quantum well lasers and modulation-doped FETs.

C2.16.5.8 DIODE LASERS

The light emitted in the spontaneous recombination process can leave the semiconductor, be absorbed or cause additional transitions by stimulating electrons in the CB to make a transition to the VB. In this *stimulated recombination* process another photon is emitted. The rate of stimulated emission is governed by a detailed balance between absorption, and spontaneous and stimulated emission rates. Stimulated emission occurs when the probability of a photon causing a transition of an electron from the CB to VB with the emission of another photon is greater than that for the upward transition of an electron from the VB to the CB upon absorption of the photon. These rates are commonly described in terms of Einstein's A and B coefficients [8, 43]. For semiconductors, there is a simple condition describing the carrier density necessary for stimulated emission, or lasing. This carrier density is known as

the *threshold density*. There is also a corresponding *threshold current density* that has to be supplied to the p-n junction. Lasing can start when the density of electrons injected into the CB exceeds the hole density in the VB, a condition of *population inversion*. It occurs when the separation of the quasi-Fermi levels for holes (E_p) and electrons (E_n) is greater than the energy of the emitted photon,

$$E_n - E_p > h\omega$$

and the photon energy $h\omega$ must be at least equal to E_g . Thus, in semiconductor lasers, stimulated emission occurs between distributions of states in the conduction and valence bands. In most other lasers, such as gas or glass lasers, this transition occurs between discrete energy levels.

The emission wavelength of the laser is directly related to the size of the gap. The early lasers were based on GaAs and emitted therefore in the near infrared. Lasers based on InGaAsP produce light between 1.3 and 1.55 μm , specifically tailored to optical fibre communications [44]. Ongoing advances in GaN-based materials are resulting in lasers emitting in the blue [11]. We thus have a very wide range of gaps and emission wavelengths at our disposal.

A diode laser requires the generation of spatially localized high concentrations of minority carriers, a medium to provide the gain and a way of providing feedback to the stimulated emission [43, 44 and 45]. The medium is the semiconductor heterostructure arranged to help to confine the carriers and the photons to the same region of space. Light is generated by a p–n junction which injects electrons from the valence to the conduction band and thus provides the population inversion. This is followed by electron–hole recombination and the emission of light. Further recombination can be stimulated by light already present in the medium. The optical feedback is arranged by forming a cavity with two mirrors parallel to each other. Light generated within the cavity is then partially reflected back into the crystal. Such mirrors can be formed in most compound semiconductors by simply cleaving both ends of the heterostructure wafer.

Figure C2.16.11 illustrates the evolution of the threshold current density of diode lasers with the structure of the recombination region within the p–n junction, known as the *active* region. The early diode lasers were based on GaAs homojunctions. Their large threshold current densities resulted from poor carrier confinement and large effective active region thickness. This is because the diffusion length of electrons is fairly long in most of the semiconductors discussed here, of the order of several micrometres. A very high threshold current density limits operation to short pulses and cryogenic temperatures.

-22-

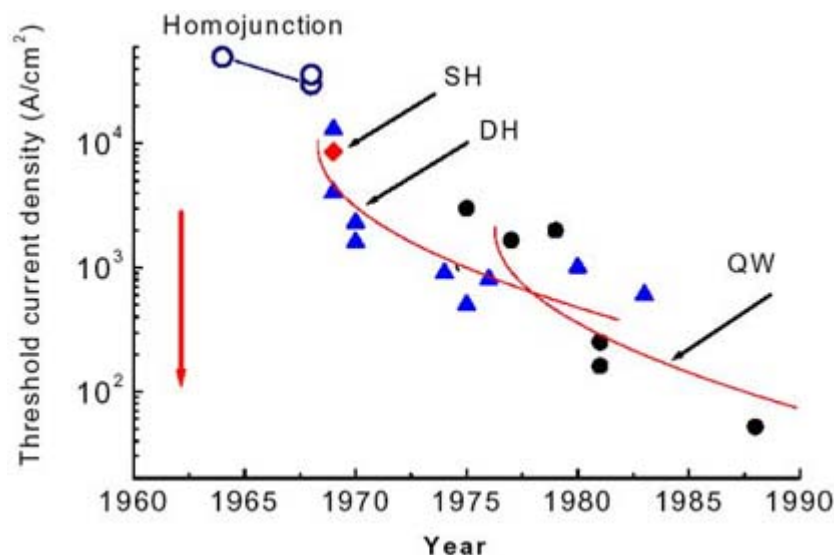


Figure C2.16.11. Changes in the threshold current density of diode lasers resulting from new structure concepts. A homojunction diode laser was first demonstrated in 1962. SH and DH stand for single and double heterostructure, respectively. The best laser performance is now obtained in quantum well (QW) lasers.

In a heterostructure laser, the active region can be defined by epitaxial layers and made considerably thinner. In GaAs/Al_xGa_{1-x}As heterostructures, the active region can be made as thin as 100 nm, and the threshold current density drops to less than 0.5 kA cm⁻². Such lasers readily achieve continuous operation at room temperature and are capable of high power output.

A logical consequence of this trend is a *quantum well laser* in which the active region is reduced further, to less than 10 nm. The 2D carrier confinement in the wells (formed by the CB and VB discontinuities) changes many basic semiconductor parameters, in particular the density of states in the CB and VB, which is greatly reduced in quantum well lasers. This makes it easier to achieve population inversion and results in a significant reduction in the threshold carrier density. Indeed, quantum well lasers are characterized by threshold current densities lower than 100 A cm^{-2} .

The history of the diode laser illustrated in figure C2.16.11 shows the interplay of basic device physics ideas and technology. A new idea often does not produce a better device right away. It requires a certain leap of faith to see the improvement potential. However, once the belief exists, the technology can be developed to demonstrate its validity. In the case of diode lasers, the better technology was invariably associated with improved epitaxial growth.

C2.16.6 OUTLOOK

We are aware of the dangers inherent in predicting the future. It is much safer to summarize the present and to extrapolate to the near term only.

The speed and general performance of semiconductor electronics have been doubling and the cost halving every 18

-23-

months for the last 50 years, a phenomenon known as Moore's law [46]. It is remarkable that this rate of advances still holds today, even as the active volume of devices is becoming so small that quantum effects are critical. The semiconductor industry has listed the expected challenges and specific goals in a 'road-map' extending well into the next decade [47]. The gate length of MOS transistors will be measured in nanometres, not in micrometres. The thickness of gate dielectrics is expected to drop to less than 20 nm. This reduction in size demands a much better microscopic understanding of materials and processes.

Atomic-scale devices already projected pose design challenges at the quantum mechanical level. The framework of quantum computing is now being discussed in research laboratories [48, 49].

In additions to improvements in Si, a variety of devices based on compound semiconductors can be expected. Blue lasers with high brightness and long operating lifetimes already exist in the laboratory. LEDs are likely to be used for all lighting purposes. The bandwidth of optical communications will continue to increase with ever faster semiconductor lasers.

There appears to be a world market for an infinite number of computers and other electronic devices.

REFERENCES

- [1] Braun F 1874 Ueber die Stromleitung durch Schwefelmetalle *Ann. Phys. Chem.* **153** 556–62
- [2] Fowler A B 1993 A semicentury of semiconductors *Phys. Today* October, p 59
- [3] Seitz F 1995 Research on silicon and germanium in World War II *Phys. Today* January, p 22
- [4] Riordan M and Hoddeson I 1997 The Moses of Silicon Valley *Phys. Today* December, p 42
- [5] Ross I M 1997 The Foundation of the Silicon Age *Phys. Today* December, p 34
- [6] Berger L I 1997 *Semiconductor Materials* (Boca Raton, FL: Chemical Rubber Company)

- [7] Cen A-B and Sher A 1995 *Semiconductor Alloys* (New York: Plenum)
 - [8] Casey H C Jr and Panish M B 1978 *Heterostructure Lasers* (New York: Academic)
 - [9] Madelung O (ed) 1996 *Semiconductors—Basic Data* (Berlin: Springer)
 - [10] Haller E E 1995 Isotopically engineered semiconductors *Appl. Phys. Rev. J. Appl. Phys.* **77** 2857
 - [11] Pankove J I and Moustaka T D 1997 *Gallium Nitride (Semiconductors and Semimetals 50)* (Boston: Academic)
 - [12] Willardson R K and Weber E R (eds) 1998 *Gallium Nitride II (Semiconductors and Semimetals 57)* (Boston: Academic)
 - [13] Phillips J C 1973 *Bonds and Bands in Semiconductors* (San Diego: Academic)
 - [14] Tsao J Y 1992 *Materials Fundamentals of Molecular Beam Epitaxy* (New York: Academic)
-

-24-

- [15] Panish M B and Temkin H 1993 *Gas Source Molecular Beam Epitaxy* (Berlin: Springer)
- [16] Chu S N G 1993 Long wavelength laser diode reliability and lattice imperfections *MRS Bull.* **17** 43
- [17] Grove A S 1967 *Physics and Technology of Semiconductor Devices* (New York: Wiley)
- [18] Harrison W A 1979 *Solid State Theory* (New York: Dover)
- [19] Seeger K 1973 *Semiconductor Physics* (New York: Springer)
- [20] Yu P Y and Cardona M 1995 *Fundamentals of Semiconductors* (Berlin: Springer)
- [21] Stillman G E and Wolfe C M 1976 *Thin Solid Films* **31** 69
- [22] Störmer H L 1983 *Surf. Sci.* **132** 519
- [23] Lang D V 1979 *Thermally Stimulated Relaxation in Solids (Topics in Applied Physics 31)* ed P Braunlich (Berlin: Springer), p 93
- [24] Jones R 1996 *Early Stages of Oxygen Precipitation in Silicon* (Dordrecht: Kluwer)
- [25] Shimura F (ed) 1994 *Oxygen in Silicon (Semiconductors and Semimetals 42)* (Boston: Academic)
- [26] Bruel M 1998 The history, physics, and applications of the smart-cut process *MRS Bull.* **23** 35, 1998
- [27] Queisser H J 1998 Defects in semiconductors: some fatal, some vital *Science* **281** 945
- [28] Watkins G D 1986 *Deep Centers in Semiconductors* ed S T Pantelides (New York: Gordon and Breach)
- [29] Hitti B, Kreitzman S R, Estle T L, Bates E S, Dawdy M R, Head T L and Lichti R L 1999 *Phys. Rev. B* **59** 4918
- [30] Bonde Nielsen K, Bech Nielsen B, Hansen J, Andersen E and Andersen J U 1999 *Phys. Rev. B* **60** 1716
- [31] Estreicher S K 1995 *Mater. Sci. Eng. R* **14** 319
- [32] Sze S M 1969 *Physics of Semiconductor Devices* (New York: Wiley-Interscience)
- [33] Streetman B G 1990 *Solid State Electronic Devices* (Englewood Cliffs, NJ: Prentice Hall)
- [34] Campbell S A 1996 *The Science and Engineering of Microelectronic Fabrication* (New York: Oxford University Press)
- [35] Kaneda T 1985 *Lightwave Communications Technology (Semiconductor and Semimetals 22D)* ed W T Tsang (Orlando, FL: Academic)

- [36] Saleh B E A and Teich M C 1991 *Photonics* (New York: Wiley)
 - [37] Asbeck P M 1990 *High Speed Semiconductor Devices* ed S M Sze (New York: Wiley)
 - [38] Shur M 1996 *Introduction to Electronic Devices* (New York: Wiley)
 - [39] Pearson S J and Shah N J 1990 *High Speed Semiconductor Devices* ed S M Sze (New York: Wiley)
 - [40] Margaritondo G (ed) 1988 *Electronic Structure of Semiconductor Heterojunctions* (Milan: Kluwer)
 - [41] Bastard G 1988 *Wave Mechanics Applied to Semiconductor Heterostructures* (New York: Halsted)
-

-25-

- [42] Zory P S Jr (ed) 1993 *Quantum Well Lasers* (Boston: Academic)
 - [43] Coldren L A and Corzine S W 1995 *Diode Lasers and Photonic Integrated Circuits* (New York: Wiley)
 - [44] Agrawal G P and Dutta N K 1993 *Semiconductor Lasers* (New York; Van Nostrand)
 - [45] Chuang S L 1995 *Physics of Optoelectronic Devices* (New York; Wiley)
 - [46] Moore G 1980 VLSI, what does the future hold? *Electron. Aust.* **42** 14
 - [47] The National Technology Roadmap for Semiconductors, sponsored by the Semiconductor Industry Association (SIA) and published by Sematech, Inc. The 1997 version can be viewed electronically at <http://notes.sematech.org>
 - [48] Landauer R 1991 Information is physical *Phys. Today* May, p 23
 - [49] Preskill J 1999 Battling decoherence: the fault-tolerant quantum computer *Phys. Today* June
-

-1-

C2.17 Nanocrystals

Vicki L Colvin and Dan M Mittleman

C2.17.1 INTRODUCTION

Throughout most of the twentieth century, the size of a solid has been an uninteresting parameter in the developing areas of solid-state physics and chemistry. Millimetre-thick gold wires have the same colour, conductivity and melting point as gold coins; even in state-of-the-art microelectronics, where structural features with sub-micrometre sizes have become typical, semiconductors and metals have the same behaviour as measured in their macroscopic counterparts. These facts are not surprising, as most material properties, such as conduction and colour, emerge from interactions between, at most, hundreds of unit cells, each less than a nanometre in size. Thus size is not an important factor in understanding and controlling crystalline solids unless the length of solid shrinks to the nanometre length scale. Of course, this is the relevant size range for the next generation of microelectronics, and is evidently an important length scale for many biological systems. These factors have spurred the development of the field of nanoscience, which is the study of the influence of size on the properties of solids. Innumerable

studies of the properties of nanocrystalline materials have amply demonstrated that size really does matter; it can be a powerful parameter in the systematic study of bulk behaviour, as well as in the design of new materials with unique and special properties.

For the purposes of this review, a nanocrystal is defined as a crystalline solid, with feature sizes less than 50 nm, recovered as a purified powder from a chemical synthesis and subsequently dissolved as isolated particles in an appropriate solvent. In many ways, this definition shares many features with that of ‘colloids’, defined broadly as a particle that has some linear dimension between 1 and 1000 nm [1]; the study of nanocrystals may be thought of as a new kind of colloid science [2]. Much of the early work on colloidal metal and semiconductor particles stemmed from the photophysics and applications to electrochemistry. (See, for example, the excellent review by Henglein [3].) However, the definition of a colloid does not include any specification of the internal structure of the particle. Therein lies the crucial distinction: in nanocrystals, the interior crystalline structure is of overwhelming importance. Nanocrystals must truly be ‘little solids’ (figure C2.17.1), with internal structures equivalent (or nearly equivalent) to that of bulk materials. This is a necessary condition if size-dependent studies of nanometre-sized objects are to offer any insight into the behaviour of bulk solids.

The definition above is a particularly restrictive description of a nanocrystal, and necessarily limits the focus of this brief review to studies of nanocrystals which are of relevance to chemical physics. Many nanoparticles, particularly oxides, prepared through the sol-gel method are not included in this discussion as their internal structure is amorphous and hydrated. Nevertheless, they are important nanomaterials; several textbooks deal with their synthesis and properties [4, 5]. The material science community has also contributed to the general area of nanocrystals; however, for most of their applications it is not necessary to prepare fully isolated nanocrystals with well defined surface chemistry. A good discussion of the goals and progress can be found in references [6, 7, 8 and 9]. Finally, there is a rich history in gas-phase chemical physics of the study of clusters and size-dependent evaluations of their behaviour. This topic is not addressed here, but covered instead in [chapter C1.1](#), Clusters and nanoscale structures, in this same volume.

-2-

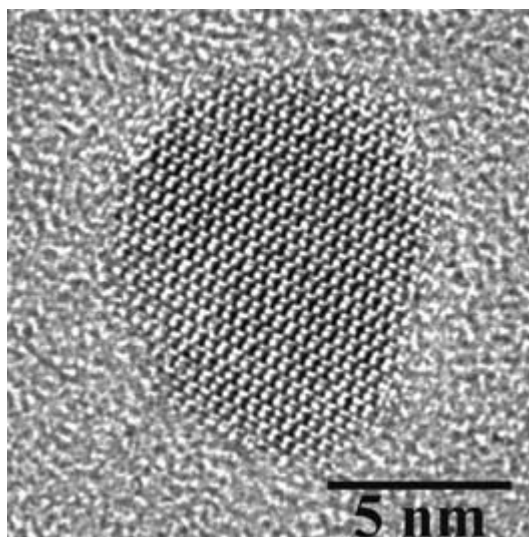


Figure C2.17.1. Transmission electron micrograph of a TiO_2 (anatase) nanocrystal. The mottled and unstructured background is an amorphous carbon support film. The nanocrystal is centred in the middle of the image. This microscopy allows for the direct imaging of the crystal structure, as well as the overall nanocrystal shape. This titania nanocrystal was synthesized using the nonhydrolytic method outlined in [79].

We begin our discussion of nanocrystals in this chapter with the most challenging problem faced in the field: the preparation and characterization of nanocrystals. These systems present challenging problems for inorganic and analytical chemists alike, and the success of any nanocrystal synthesis plays a major role in the further quantitative study of nanocrystal properties. Next, we will address the unique size-dependent optical properties of both metal and semiconductor nanocrystals. Indeed, it is the striking size-dependent colours of nanocrystals that first attracted

the interest of chemical physicists. Finally, the thermodynamic properties of nanocrystals will be reviewed. The melting point reduction and unusual structural metastability observed in solids of confined size are important results in understanding the physics of these systems.

C2.17.2 PREPARATION OF NANOCRYSTALS

Obtaining high-quality nanocrystalline samples is the most important task faced by experimentalists working in the field of nanoscience. In the ideal sample, every cluster is crystalline, with a specific size and shape, and all clusters are identical. While such uniformity can be expected from a molecular sample, nanocrystal samples rarely attain this level of perfection; more typically, they consist of a collection of clusters with a distribution of sizes, shapes and structures. In order to evaluate size-dependent properties quantitatively, it is important that the variations between different clusters in a nanocrystal sample be minimized, or, at the very least, that the range and nature of the variations be well understood.

Reaching the goal of the ideal nanocrystal sample is not an easy task. There are few commercial sources for nanocrystals, and the chemical reactions used to make them can require involved synthetic methodology. On the other hand, the last decade has seen enormous progress in this area and many solids have now been prepared in the

-3-

nanocrystalline phase. For the most well studied materials, particle size distributions of less than 5% on the diameter are routinely obtained [10, 11 and 12]; more typically, size distributions of 10–20% are reasonable. Note that polydispersity for a colloidal system, referred to here as σ , is defined as the standard deviation of the particle diameters divided by average particle diameter.

While size distribution is important, control over the nanocrystal surface is equally important. The best nanocrystal syntheses provide avenues for nanocrystals to be purified, collected as powders, and then redissolved in appropriate solvents. This requires control over the surface chemistry, in order to control the solubility of the nanocrystals. Such flexibility allows these materials to be structurally characterized and then assembled into a wide variety of configurations for further experiments.

There are many ingenious and successful routes now developed for nanocrystalline synthesis; some rely on gas phase reactions followed by product dispersal into solvents [7, 9, 13, 14 and 15]. Others are adaptations of classic colloidal syntheses [16, 17, 18 and 19]. Electrochemical and related template methods can also be used to form nanostructures, especially those with anisotropic shapes [20, 21, 22 and 23]. Rather than outline all of the available methods, this section will focus on two different techniques of nanocrystal synthesis which together demonstrate the general strategies.

C2.17.2.1 REACTIONS OCCURRING IN RESTRICTED ENVIRONMENTS

A logical departure point for the synthesis of nanocrystals is to view the problem as one of limiting the growth of a bulk crystal; this is challenging because the large surface free energy of a nanocrystal makes it a metastable system, highly prone to fusion and aggregation. One particularly elegant and versatile solution is to precipitate solids inside reactors which are themselves of nanometre scale [24, 25, 26, 27 and 28]. Such nanometre-scale reaction environments can be formed by mixing water, surfactant and oil to create inverse micelles [29, 30]. These water pools have diameters that can be tuned from 5 to ~60 Å in radius by varying the water/surfactant molar ratio, thus providing a direct avenue for size control [31, 32 and 33]. Ionic salts can be solvated in the inverse micelles; when such solutions are exposed to a counterion which forms an insoluble solid with the original salt, small crystals of the solid form within the micelle environment. This process, referred to as arrested precipitation, can be used to grow nanocrystals with sizes which are roughly equivalent to the original micellar size. Surface control can be achieved by adding organic ‘capping’ agents to the final solutions. Depending on the solubility and capping group affinity, the nanocrystals may directly precipitate out of the micellar solution [34, 35] or may be recovered as a powder after the micelle phase is disrupted by the addition of an alcohol [36, 37]. The advantage of this approach is

that it is relatively easy, and can be applied to many different materials including metals [38, 39, 40 and 41], ceramics [42, 43, 44, 45, 46 and 47] and some semiconductors [34, 35, 36 and 37, 48]. Shape control is also possible in surfactant-based syntheses of metal nanocrystals [49, 50 and 51]. Cizeron *et al* [52] use surfactants to control not only particle size, but more critically the growth rates of different crystal faces, producing highly anisotropic nanoneedles with aspect ratios exceeding 100:1 (figure C2.17.2).

-4-

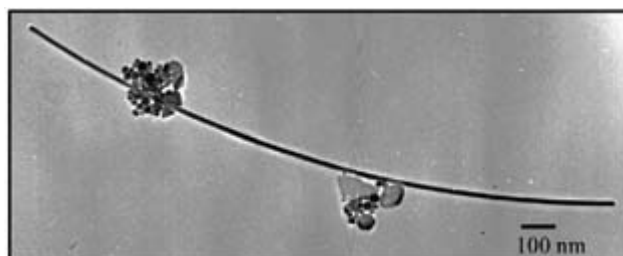


Figure C2.17.2. Transmission electron micrograph of a gold nanoneedle. Inverse micelle environments allow for a great deal of control not only over particle size, but also particle shape. In this example, gold nanocrystals were prepared using a photolytic method in surfactant-rich solutions; the surfactant interacts strongly with areas of low curvature, thus continued growth can occur only at the sharp tips of nanocrystals, leading to the formation of high-aspect-ratio nanostructures [52].

Though easy to implement and quite versatile, inverse micelle reactions have not generally produced the highest-quality nanocrystals. In such an environment the reaction necessarily occurs at room temperature. While crystalline metals are easily formed at these low temperatures, nanoparticles of semiconductors and oxides are often highly defective or even amorphous. This problem can be addressed to some extent by refluxing the nanocrystals in a high boiling point solvent after separating them from the micelle phase [35]. Another common problem with nanocrystals produced by this method is their relatively poor size distributions. Post-processing treatments, such as size-selective precipitation or filtration, are generally employed if nanocrystal monodispersity is of great importance [53, 54]. Pileni *et al* [54] recently used size-selective precipitation to narrow the size distribution of silver nanocrystals prepared in an inverse micelle reaction from 37% to 15% polydispersity. Perhaps the most severe limitation of the inverse micelle approach is that it can only form nanocrystals whose precursors are stable and solvated by water. More covalent semiconductors like silicon and gallium arsenide, as well as many II–VI materials like CdSe, require reactive organometallic precursors. Thus, alternative reactions which proceed in dry, organic solvents are necessary.

C2.17.2.2 PRECURSOR APPROACHES TO NANOCRYSTAL SYNTHESIS

This class of methods differ from inverse micelle methods in that the reactions are completed in organic solvents. Such solvents permit the reactions to proceed at much higher temperatures, leading to nearly perfect crystalline solids [55]. In addition, the use of organic solvents permits nanocrystals to be prepared from a wide variety of molecular precursors under oxygen-free and water-free conditions. Metal [12, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65 and 66], semiconductor [67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77 and 78] and ceramic nanocrystals [79, 80] have been generated using this basic strategy. These stringent controls over reaction conditions are important in the synthesis of covalent semiconductors, which require reactive organometallic starting materials. This can also be an advantage in the preparation of metal nanocrystals free from oxide or hydroxide contamination. Precursor methods have also shown remarkable success in producing highly monodisperse ($\sigma < 5\%$) nanocrystals, especially for II–VI semiconductor nanocrystals [10, 81]. For these reasons, this particular approach to nanocrystal synthesis is becoming a popular strategy despite the fact that it requires more involved synthetic methodology.

The first step in designing a precursor synthesis is to pick precursor molecules that, when combined in organic solvents, yield the bulk crystalline solid. For metals, a usual approach is to react metal salts with reducing agents to produce bulk metals. The main challenge is to find appropriate metal salts that are soluble in an organic phase.

One prevalent strategy for this involves the use of a phase transfer agent, such as tetraoctyl ammonium bromide, to bring gold and silver salts into an organic phase [12, 56, 57, 58, 59, 60, 61, 62 and 63]. Reduction of the metal salts occurs readily with the addition of sodium borohydride. Formation of the bulk solid is prevented by the saturation of the organic phase with capping molecules, which serve to limit crystallite growth [82]. Gold nanocrystals with average diameters of 2.5 nm and 20% polydispersity are the crude product of this reaction; fractional crystallization of these samples, however, can effectively size-separate these nanocrystals and provide samples with less than 1% polydispersity [12]. Rather than rely on phase transfer techniques, which do require exposure of the organic phase to water, gold complexes soluble in organic phases can be prepared directly. Upon heating, these decompose to form gold clusters [64, 65]. The extension of these general strategies to more reactive transition metals such as titanium and copper is also possible. For these systems, LiAlH_4 or hydrotriorganoborates are used to reduce metal salts in solvents such as tetrahydrofuran, producing nanocrystalline metals of a variety of sizes stabilized in organic phases. Bonneman *et al* [66] provide a comprehensive review of these versatile precursor strategies for nanocrystalline metals.

For covalent semiconductors the problem of creating any crystalline material in liquids at relatively low temperatures is much more challenging. The precursors are generally chosen to provide by-products that are stable as well as volatile at the reaction temperatures, thus providing a driving force for the reaction. For example, in one of the early reactions for making nanocrystalline GaAs, GaCl_3 and $\text{As}(\text{SiMe}_3)_3$ were used as precursors, since the silyl halide by-product was stable and volatile yielding a pure product of crystalline GaAs at temperatures as low as 240°C [78]. Since then, a variety of III–V semiconductor nanocrystals, including InP, have been produced through similar high-temperature reactions between reactive precursors [71, 72, 73 and 74, 76, 77, 83]. A different precursor strategy was employed for the formation of silicon and germanium nanocrystals. Here, the metal halides were reduced by alkali metals at high temperatures forming nanocrystals and salts as a by-product [67, 68 and 69].

Control over the size of the semiconductor nanocrystals formed in these reactions is possible, though the rationalization of the size control is not always straightforward. First, these strategies require the use of a strongly stabilizing solvent, such as tri-octyl phosphine oxide, which is thought to slow crystal growth because of its strong interactions with the growing crystallite surfaces. In perhaps the best characterized reaction, the formation of II–VI semiconductor nanocrystals, crystal growth is limited by a rapid quenching of the reaction temperature [10, 84, 85 and 86]. Particle nucleation occurs during a fixed period of time after injection of the cadmium precursor into hot mixtures of tri-octyl phosphine oxide and tri-octyl phosphine. Growth of these nuclei is highly temperature dependent, and the final nanocrystal size can be controlled by the reaction time at elevated temperatures, as well as the ratio of Cd:Se [11]. Typical size distributions formed by these reactions can be $\sigma < 5\%$ in the II–VI material systems [10, 81]. The III–V materials can achieve $\sigma = 20\%$ [87], and the group IV semiconductor nanocrystals achieve $\sigma \sim 30\%$ [69].

C2.17.2.3 OTHER CHEMICAL APPROACHES TO NANOCRYSTAL SYNTHESIS

The methods described above are by no means the only strategies for creating crystals of restricted size. Colloidal methods for creating nanoparticles are still used widely today, especially to make gold and semiconductor nanoparticles [16, 17]. These reactions have the advantage of providing nanoparticles in aqueous solutions, the ideal environment for electrochemical applications. Also, while gas phase clusters are not considered nanocrystals by the stringent definition given previously, it is possible to form nanocrystals in the gas phase and subsequently collect and disperse them into solvents. Such methods have been applied to the production of silicon [13, 88] as well as metal [89] nanocrystals.

C2.17.2.4 NANOCRYSTAL ASSEMBLY

In solution, nanocrystals are ideal spectroscopic samples; however many of their most important properties can only be realized when they are assembled into more complex structures. One way of building complex structures is to rely on the inherent tendency for monodisperse spheres to crystallize. Figure C2.17.3 shows the hexagonal close-

packed ordering of monodisperse silica nanoparticles; such crystallization does not require any inter-particle interaction, but occurs when highly uniform objects are driven to a minimum volume packing. This ordering has been observed in many nanocrystalline systems when crystallite size distributions fall below 5%; three-dimensional supercrystals of nanocrystals are the result [90, 91 and 92]. Crystallization can also be induced at air–liquid interfaces using Langmuir–Blodgett techniques. Thin monolayers of particle arrays thus formed can be transferred to any surface [93, 94]. Finally, high-coverage monolayers of nanocrystals or spatially patterned sub-monolayers can be formed on a variety of flat solid surfaces [14, 95, 96, 97, 98, 99, 100 and 101].

Figure C2.17.3. Close-packed array of sub-micrometre silica nanoparticles. When nanoparticles are very monodisperse, they will spontaneously arrange into hexagonal close-packed structure. This scanning electron micrograph shows an example of this for very monodisperse silica nanoparticles of ~250 nm diameter, prepared in a thin-film format following the techniques outlined in [236].

An equally important challenge for nanocrystal assembly is the formation of specific nanocrystal arrangements in solution. By using complementary DNA strands as tethers, Mirkin *et al* [102, 103] formed aggregates of gold nanocrystals with specific sizes; Alivisatos *et al* also used DNA to structure semiconductor nanocrystal molecules, though in this case the molecules contained only a few nanocrystals placed controlled distances from each other [104, 105 and 106]. The potential applications of biomolecular techniques to this area of nanoscience are immense, and the opportunities have been reviewed in several recent publications [107, 108, 109 and 110].

C2.17.3 CHARACTERIZATION OF NANOCRYSTALS

The goal of any nanocrystal characterization is to identify the position of every atom in a single nanocrystal, as well as

-7-

to determine the distribution of sizes and shapes present in a sample. Such precise characterization is possible in the case of some smaller metal and semiconductor clusters. These systems are amenable to new types of mass spectroscopy which can provide molecular weights of species up to 55 000 atomic mass units, thus permitting molecular composition and size to be precisely determined [111, 112, 113 and 114]. These systems are the exception, however, not the rule. More typically, nanocrystalline samples contain crystallites with thousands of atoms, molecular weights in excess of 100 000 amu, and diameters larger than 1 nm.

The problem of characterizing these larger nanocrystals is challenging and is, in its own right, an area of research. Some of the simplest characterization techniques are indirect: for example, the mean size of semiconductor nanocrystals can often be inferred from the sample's optical properties. These inferences typically rely on very simple models for the nanocrystal properties, and thus provide only very rough sizing estimates and no information about structure. Classic analytical methods of polymer and colloid science, such as chromatography and light scattering, can also be used to evaluate nanocrystalline samples [13, 115, 116, 117 and 118]. While they provide a measure of particle size and shape in solution, their applications to inorganic nanocrystals have been limited, in part because they provide little direct structural information. The most successful tools in this area have been adapted from material science and inorganic chemistry. These direct structural methods, when used together, provide a complete picture of the nanocrystal interior and its surface.

C2.17.3.1 MICROSCOPIES

In many ways the nanocrystal characterization problem is an ideal one for transmission electron microscopy (TEM). Here, an electron beam is used to image a thin sample in transmission mode [119]. The resolution is a sensitive function of the beam voltage and electron optics; a low-resolution microscope operating at 100 kV might

have a 2–3 Å resolution while a high-voltage machine designed for imaging can have a resolution approaching 1 Å. Since nanocrystalline samples range from ten to hundreds of angstroms in size, this type of microscopy allows both the interior crystal structure and the overall particle shape to be measured.

A single TEM picture of a nanocrystalline sample can provide an enormous amount of information ([figure C2.17.1](#)). Low-resolution TEM can also be used to determine sample distributions and shapes ([figure C2.17.2](#)) ([figure C2.17.4](#)) and ([figure C2.17.5](#)) [120]. Higher-resolution images show the discrete nature of the crystalline interior of nanoparticles and can detect the presence of certain crystalline defects ([figure C2.17.1](#)) and ([figure C2.17.6](#)) ; the Fourier transform of such images ([figure C2.17.6](#)) left panel) provides a measure of the lattice spacing and crystallographic parameters. Direct electron diffraction data can also be collected on fields of particles to verify the phase of the nanocrystal ([figure C2.17.7](#)).

-8-

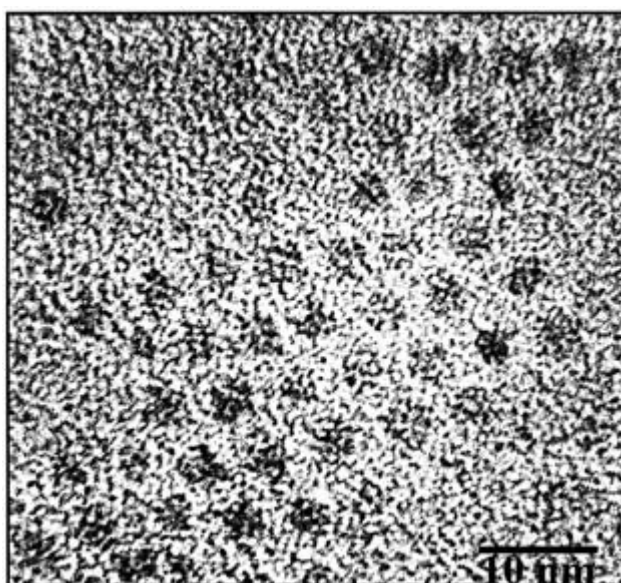


Figure C2.17.4. Transmission electron micrograph of a field of ZrO_2 (tetragonal) nanocrystals. Lower-resolution electron microscopy is useful for characterizing the size distribution of a collection of nanocrystals. This image is an example of a typical particle field used for sizing purposes. Here, the nanocrystalline zirconia has an average diameter of 3.6 nm with a polydispersity of only 5% [80].

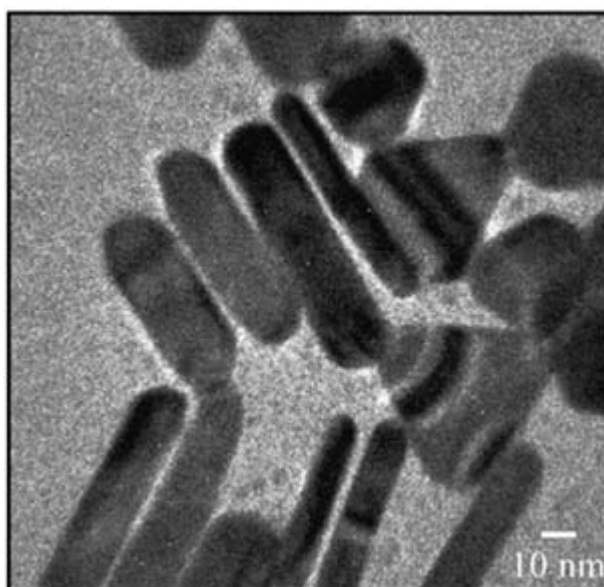


Figure C2.17.5. Transmission electron micrograph of a field of anisotropic gold nanocrystals. In this example, a lower magnification image of gold nanocrystals reveals their anisotropic shapes and faceted surfaces [36].

-9-

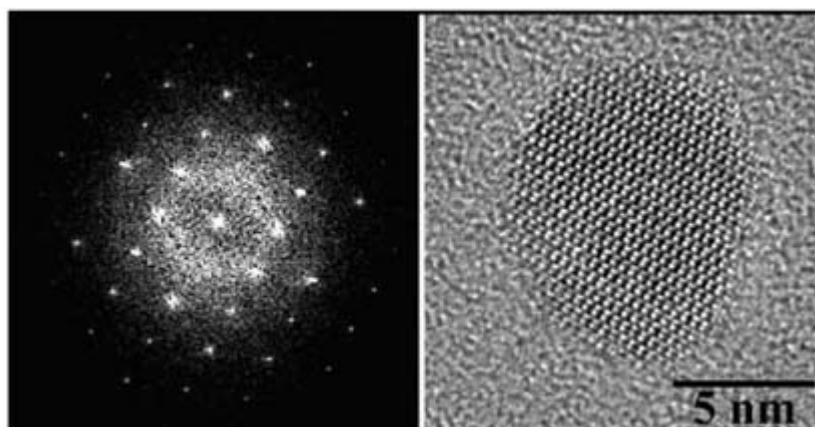


Figure C2.17.6. Transmission electron micrograph and its Fourier transform for a TiC nanocrystal. High-resolution images of nanocrystals can be used to identify crystal structures. In this case, the image of a nanocrystal of titanium carbide (right) was Fourier transformed to produce the pattern on the left. From an analysis of the spot geometry and spacing, one can determine that the nanocrystal is oriented with its [100] zone axis parallel to the viewing direction [217].

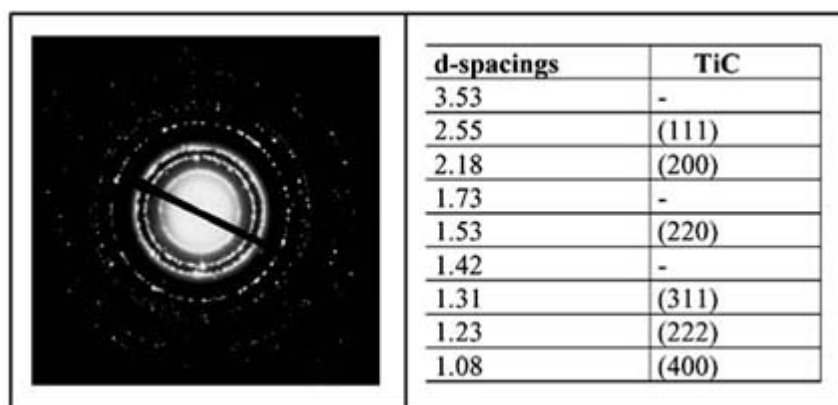


Figure C2.17.7. Selected area electron diffraction pattern from TiC nanocrystals. Electron diffraction from fields of nanocrystals is used to determine the crystal structure of an ensemble of nanocrystals [119]. In this case, this information was used to evaluate the phase of titanium carbide nanocrystals [217].

More sophisticated analyses of high-resolution TEM images can provide even deeper insight into subtle structural aspects of nanocrystals [121, 122 and 123]. Simulation of high-resolution images can, in principle, provide data concerning whether the average bond length in a nanocrystal is uniform or variable within the nanocrystal interior [124]. Another exciting prospect is the use of TEM to provide direct information about nanocrystal surfaces, including reconstructions and dynamic motions of atoms at surfaces [125]. In one case high-resolution studies of a gold nanocrystal surface led to the identification of a 2×1 surface reconstruction [126, 127].

Other forms of microscopy have been used to evaluate nanocrystals. Scanning electron microscopy (SEM), while having lower resolution than TEM, is able to image nanoparticles on bulk surfaces, for direct visualization of

-10-

nanocrystals in larger assemblies (figure C2.17.3). McEuen *et al* [128, 129] used a field emission SEM to detect the

presence or absence of single nanoparticles in small electrodes made lithographically; Bawendi *et al* [91] used SEM to determine the three-dimensional structure of arrays of CdSe clusters. Scanning tunnelling microscopy (STM) has also been successfully applied to metal nanocrystals; in these cases STM can be used both to image nanocrystals and to study their electrical properties [111, 130, 131, 132 and 133]. The use of scanning force microscopies for nanocrystal characterization is not as common, due to the relatively large probe sizes ($d > 50$ nm) required for force microscopy [134, 135]; however, the height resolution of force microscopy is quite good, of the order of 0.1 Å. The height of nanocrystals on surfaces has been used as a metric for nanocrystal diameter [58, 136]. Another difficulty with force microscopy is that lateral forces are large, and the probe can move nanocrystals unless they are tightly bound to the underlying surface. Intermittent contact-mode, or Tapping ModeTM, reduces these lateral forces and is a more appropriate choice for nanocrystal sizing.

C2.17.3.2 X-RAY DIFFRACTION

Although microscopic methods provide a direct visualization of nanocrystal samples, the images alone provide a misleading view of a nanocrystalline sample. Unreacted molecular species as well as small amorphous particles are difficult to see in many microscopies, yet they can comprise a large fraction of a supposedly nanocrystalline sample. In such low-purity samples, x-ray diffraction (XRD) studies would show no distinct crystalline features; highly pure nanocrystalline samples, in contrast, provide strong crystalline reflections (figure C2.17.8). The positions of these peaks provide an accurate fingerprint of the crystal structure of the nanocrystal. Such an unambiguous determination of nanocrystal structure is especially important, as many nanocrystalline materials adopt metastable crystal structures distinct from the bulk solid (figure C2.17.9) [137, 138].

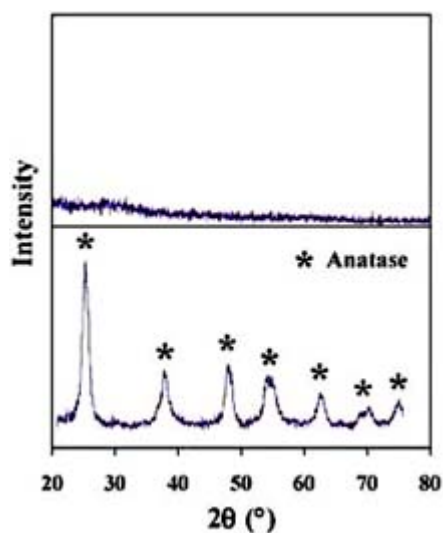


Figure C2.17.8. Powder x-ray diffraction (PXRD) from amorphous and nanocrystalline TiO₂ nanocrystals. Powder x-ray diffraction is an important test for nanocrystal quality. In the top panel, nanoparticles of titania provide no crystalline reflections. These samples, while showing some evidence of crystallinity in TEM, have a major amorphous component. A similar reaction, performed with a crystallizing agent at high temperature, provides well defined reflections which allow the anatase phase to be clearly identified.

The breadth of the peaks in an x-ray diffractogram provide a determination of the average crystallite domain size, assuming no lattice strain or defects, through the Debye–Scherr formula:

$$t = \frac{0.9\lambda}{B \cos \theta_B}$$

where t is the thickness of the crystal, λ is the wavelength of the x-rays, B is the full width at half maximum of the

diffraction peak, and θ_B is the Bragg angle of the peak [139, 140]. Figure C2.17.9 shows an example of the line broadening observed in nanocrystalline samples of different size. It is especially valuable to compare the domain sizes determined from XRD with the sizing found in TEM. Good agreement between these two measurements is conclusive evidence that the nanocrystals are free from crystalline defects. More extensive analysis and simulations of the XRD patterns of nanocrystals can provide information on nanocrystal defect density and type as well as the presence and distribution of strain in the nanocrystal lattice [141].

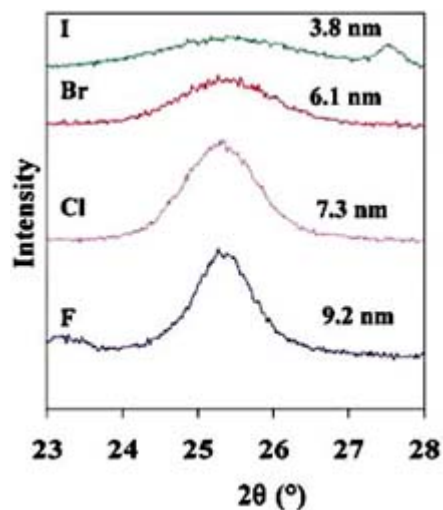


Figure C2.17.9. Size-dependent changes in PXRD linewidths. PXRD can be used to evaluate the average size of a sample. In these cases, different samples of nanocrystalline titania were analysed for their grain size using the Debye–Scherr formula. As the domain size increases, the widths of the diffraction peaks decrease.

C2.17.3.3 SURFACE CHARACTERIZATION

The characterization of nanocrystal surfaces is a crucial issue, as surface chemistry and defects can dominate nanocrystal properties. A common question concerning nanocrystal surface is the nature of the bonding at the crystal–organic interface. X-ray photoemission (XPS) is a useful method for not only identifying the atomic composition of a sample, but also for determining the oxidation states of the nanocrystal atoms [84, 142]. In CdSe nanocrystals, for example, XPS indicates the presence of both selenium bound to cadmium as well as oxidized selenium at the surface [84]; similar XPS studies of gold nanocrystals have indicated that, even when passivated by thiols, samples only contain gold in the zero-valent state. Information concerning the dynamics and also coverage of the organic groups at

the nanocrystal surface can be found using solution state NMR [58, 143, 144]. The size of nanocrystals reduces the rotational averaging in the liquid phase; however proton signals from capping groups can be detected in many systems. Vibrational spectroscopies have also been applied to the problem of organic group geometry and coverage, especially in metal systems [57, 59]. Also, extended x-ray absorption fine structure (EXAFS) and other x-ray absorption spectroscopies have been applied to nanocrystal surface characterization [145, 146]. These methods are particularly powerful, as they can, in principle, measure the spatial distribution of bond lengths within a nanocrystal.

C2.17.4 OPTICAL PROPERTIES OF NANOCRYSTALS

The striking size-dependent colours of many nanocrystal samples are one of their most compelling features; detailed studies of their optical properties have been among the most active research areas in nanocrystal science. Evidently, the optical properties of bulk materials are substantially different from those of isolated atoms of the

same material. In principle, one can describe the optical properties of a nanometre-sized object as an intermediate between these two limiting cases, with a continuous evolution from atomic to bulk properties with increasing particle size. Practically speaking, the most common approaches have used the bulk optical behaviour as a starting point for predicting spectra; successive modifications are made to these descriptions in order to account for a variety of size-dependent effects. Schemes of this nature are necessarily limited, both because the size-dependent modifications are usually treated in a perturbative or approximate fashion, and because the modified theories do not, in general, converge to the atomic limit as the size of the nanoparticle decreases. Nonetheless, such approaches have enjoyed a great deal of success in predicting the optical properties of a wide variety of nanoparticles.

This section will outline the simplest models for the spectra of both metal and semiconductor nanocrystals. The work described here has illustrated that, in order to achieve quantitative agreement between theory and experiment, a more detailed view of the molecular character of clusters must be incorporated. The nature and bonding of the surface, in particular, is often of crucial importance in modelling nanocrystal optical properties. While this section addresses the linear optical properties of nanocrystals, both nonlinear optical properties and the photophysics of these systems are also of great interest. The reader is referred to the many excellent review articles for more in-depth discussions of these and other aspects of nanocrystal optical properties [147, 148, 149, 150, 151, 152, 153 and 154].

C2.17.4.1 THE OPTICAL PROPERTIES OF SEMICONDUCTOR NANOCRYSTALS

One of the most striking features of semiconductor nanocrystal samples is their vivid colours, which vary with the mean particle size. These changes in the optical properties are not due to structural changes in the material; rather they are a reflection of the fact that the nature of the excitations which determine the colour of these solids is perturbed by the change in the size of the system. In many bulk semiconductors, the characteristic size of the lowest-lying optical excitations is much larger than the size of a unit cell [155]. As the system shrinks, this excitation is confined to a region which is smaller than its natural size, and the associated electronic energy levels shift as a result. Figure C2.17.10 shows typical absorption spectra of spherical CdSe semiconductor nanocrystals, for several different mean particle sizes. Evidently, the lowest-lying excitation shifts continuously to higher energy as the size of the particle decreases. Indeed, it is possible to follow the evolution of this state continuously from very large (nearly bulk-like) crystallites down to only ~ 10 Å radius. In bulk semiconductors, this lowest energy absorption feature corresponds to an excited electron and hole which are bound via the Coulomb interaction, known as an exciton. Thus, the corresponding excitation in the nanocrystals is generally assumed to be of the same character, although perturbed by the confinement. In CdSe, for example, the bulk exciton is 57 Å in radius [155]; thus, when CdSe nanocrystals shrink

-13-

to or below this characteristic length scale, their optical properties show an exquisite sensitivity to nanocrystal size (figure C2.17.10). This basic observation of a blue shift in the absorption energy of the excitonic transition in smaller clusters has been observed in many different semiconductor nanocrystals [72, 156, 157].

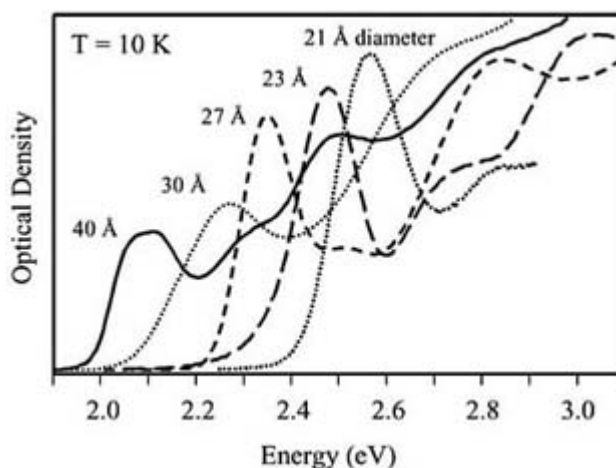


Figure C2.17.10. Optical absorption spectra of nanocrystalline CdSe. The spectra of several different samples in the visible and near-UV are measured at low temperature, to minimize the effects of line broadening from lattice vibrations. In these samples, grown as described in [84], the lowest exciton state shifts dramatically to higher energy with decreasing particle size. Higher-lying exciton states are also visible in several of these spectra. For reference, the band gap of bulk CdSe is 1.85 eV.

An explanation for these size-dependent optical properties, termed ‘quantum confinement’, was first outlined by Brus and co-workers in the early 1980s, [156, 158, 159, 160 and 161] and has formed the basis for nearly all subsequent discussions of these systems. Though recent work has modified and elaborated on this simple model, its basic predictions are surprisingly accurate. The energy of the lowest-lying exciton state is given by the following simple formula:

$$E(R) = E_g + \frac{\hbar^2 \pi^2}{2\mu R^2} - \frac{1.8e^2}{\epsilon R}.$$

Here, E_g and ϵ are the band gap energy and the dielectric constant of the bulk semiconductor, and μ is the reduced mass of the exciton system, $1/\mu = 1/m_e + 1/m_h$. The second term, proportional to $1/R^2$, arises from a simple quantum confinement effect, just as in the well known particle-in-a-box problem of undergraduate quantum mechanics. An additional adjustment to the energy arises from a Coulombic effect. Because the electron and hole are forced to occupy a small volume, the wavefunctions of these two oppositely charged particles overlap spatially to a greater degree than in the bulk crystal. The third term, proportional to $1/R$, accounts for this enhanced overlap. [Figure C2.17.11](#) illustrates this result for several different semiconductor materials.

-14-

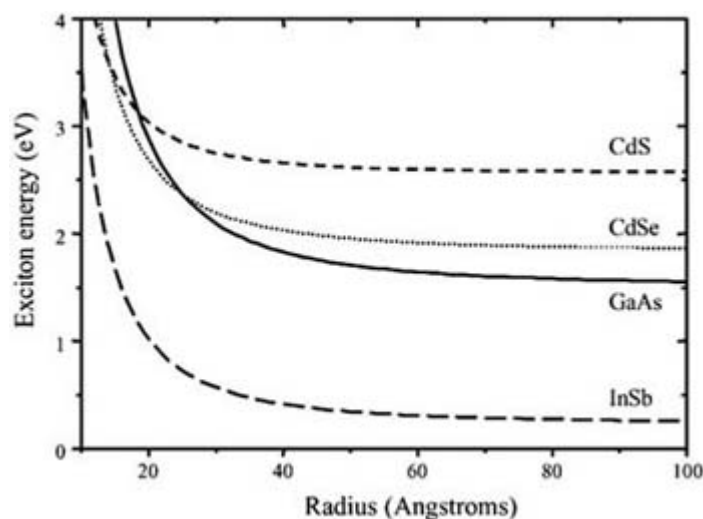


Figure C2.17.11. Exciton energy as a function of particle size. The Brus formula is used to calculate the energy shift of the exciton state as a function of nanocrystal radius, for several different direct-gap semiconductors. These estimates demonstrate the size below which quantum confinement effects become significant.

[Figure C2.17.12](#) depicts a comparison between experimentally determined exciton energies and those predicted by the Brus model, for CdSe nanocrystals. Given the level of approximation, the agreement is surprising. Nonetheless, the simple theory clearly overestimates the energies for the smallest crystallites. Recent work, both experimental and theoretical, has shown that the main deficiency of the quantum confinement model is that it fails to include molecular-level detail. For example, the Brus model assumes that the confining potential is spherically symmetric and infinitely high. High-resolution electron microscopy studies [162] as well as theoretical calculations [163] have suggested that the lattice structure and facets of nanocrystals lower the particle symmetry leading to intrinsic shifts in the confinement energies. Dielectric spectroscopy [164], as well as Stark absorption spectroscopy [165, 166]

have demonstrated that the ground state of the CdSe nanocrystals has a large static dipole moment. These measurements support the notion that the departure from perfect spherical symmetry can strongly influence the energies. Interactions between internal and surface states have also been shown to strongly influence the dynamics of the excited state, on time scales ranging from femtoseconds [167, 168] to nanoseconds [169, 170].

-15-

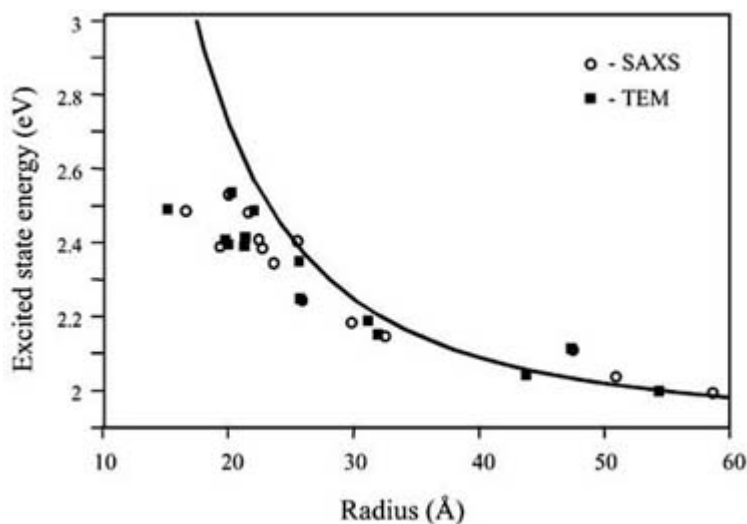


Figure C2.17.12. Exciton energy shift with particle size. The lowest exciton energy is measured by optical absorption for a number of different CdSe nanocrystal samples, and plotted against the mean nanocrystal radius. The mean particle radii have been determined using either small-angle x-ray scattering (open circles) or TEM (squares). The solid curve is the predicted exciton energy from the Brus formula.

While the Brus formula can be used to locate the spectral position of the excitonic state, there is no equivalent *a priori* description of the spectral width of this state. These bandwidths have been attributed to a combination of effects, including inhomogeneous broadening arising from size dispersion, optical dephasing from exciton–surface and exciton–phonon scattering, and fast lifetimes resulting from surface localization [167, 168, 170, 171]. Due to the complex nature of these line shapes, there have been few quantitative calculations of absorption spectra. This situation is in contrast with that of metal nanoparticles, where a more quantitative level of prediction is possible.

C2.17.4.2 THE OPTICAL PROPERTIES OF METAL NANOCRYSTALS

Like semiconductor nanocrystals, solutions of metal nanocrystals exhibit striking size-dependent colours, a fact which has fascinated scientists and artists alike for centuries. Gold colloids, in particular, have been used since the Middle Ages as colouring pigments for paints and especially stained glass. Faraday was the first to recognize their metallic character in 1857, and he remarked on their vivid colours. This colour arises from a sharply peaked resonance in the visible region of the spectrum, which occurs at much lower energy than in the bulk metal. Both the spectral position and width of this resonance are observed to vary with the size of the metal nanoparticle, suggesting, as in the semiconductor nanocrystals, that a fundamental excitation of the system is influenced by the restricted size.

The optical properties of metal nanoparticles have traditionally relied on Mie theory, a purely classical electromagnetic scattering theory for particles with known dielectrics [172]. For particles whose size is comparable to or larger than the wavelength of the incident radiation, this calculation is rather cumbersome. However, if the scatterers are smaller than ~10% of the wavelength, as in nearly all nanocrystals, the lowest-order term of Mie theory is sufficient to describe the absorption and scattering of radiation. In this limit, the absorption is determined solely by the frequency-dependent dielectric function of the metal particles and the dielectric of the background matrix in which they are

embedded. So, the size dependence of the optical properties enters only through the size-dependent dielectric function of the nanoparticles [172].

In the optical range, the dielectric function of a metal generally can be divided into two distinct contributions, arising from the intraband and the interband coupling. The former is generally described using a Drude formalism, whereas the latter is often described empirically, and differs dramatically for different metals. In nanometre-sized metals, the interband transitions, as in for example the 5d-to-conduction band transitions in gold, are generally assumed to be independent of the size of the crystallite. In contrast, the inelastic scattering time, which appears as a phenomenological parameter in the Drude model, is assumed to have a strong contribution from surface scattering [173]. A modified scattering rate of the form $\Gamma = \Gamma_0 + Av_F/R$ is often used [172, 174]. Here Γ_0 is the scattering rate in the bulk metal, v_F is the Fermi velocity and A is a constant of order unity. This form expresses the limitations on the mean free path of the free electrons as a result of the confining geometry.

Figure C2.17.13 presents a model calculation of the absorption of gold nanocrystals, using the formalism outlined above. The qualitative result is that, as metal colloids become smaller, the primary absorption peak shifts to lower energy, and broadens significantly. The peak shifts predicted are small, of the order of 0.1 eV for a 2 nm gold crystallite. In contrast, the peak widths are far more sensitive to size. This simple theory, and its variations, have been successful at explaining many experimental observations, especially for clusters greater than 3 nm in size [154]. This success is not surprising, given that the fundamental assumption has been the insensitivity of the interband transitions to the size of the crystallite. Unlike in the case of the semiconductor nanocrystals, where the transition involved highly delocalized exciton states, these band-to-band transitions involve the localized d-shell electrons of the metal. As a result, a clear quantum confinement effect has generally not been expected in metal nanocrystals. Recently, extremely detailed analyses of the optical spectra of highly monodisperse gold colloids suggest that some features of these spectra could be attributed to quantum confinement, in the form of a size-dependent interband dielectric function [111, 112]. Further work on the nonlinear optical properties of these materials may permit a more precise quantification of these effects.

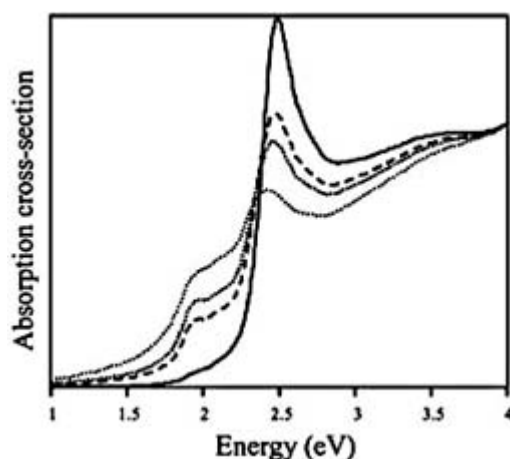


Figure C2.17.13. A model calculation of the optical absorption of gold nanocrystals. The formalism outlined in the text is used to calculate the absorption cross section of bulk gold (solid curve) and of gold nanoparticles of 3 nm (long dashes), 2 nm (short dashes) and 1 nm (dots) radius. The bulk dielectric properties are obtained from a cubic spline fit to the data of [237]. The small blue shift and substantial broadening which result from the mean free path limitation are

clearly evident.

Some of the most interesting recent work in the optical properties of nanocrystals involves the study of single nanocrystals rather than ensembles, using near-field optical techniques. These relatively new optical methods can

be used to address individual nanoparticles, rather than ensembles. This can potentially eliminate from consideration the effects of sample inhomogeneity and, in addition, can give rise to new and interesting optical effects. Studies of both metal [175] and semiconductor [176, 177] nanocrystals have been reported.

C2.17.5 THERMODYNAMIC PROPERTIES OF NANOCRYSTALS

In order to use any material for commercial purposes, it is important to understand its phase behaviour. Bulk gold, for example, melts at 1065°C and thus would be an unwise choice as an electrical interconnect in a high-temperature environment; many ceramics can crack during thermal processing due to solid–solid phase transformations that lower the volume of the crystal [178]. The extrapolation of such bulk phase behaviour to the properties of nanocrystals is not a straightforward problem. Nanocrystals are intrinsically metastable materials which, given the right circumstances, would fuse to create bulk crystals. Indeed, metal nanocrystals prepared on surfaces under high-vacuum conditions do spontaneously fuse into larger grains [179, 180]. On the other hand, solutions of nanocrystals stabilized with organic agents can exist for months or even years with unchanging sizes. Evidently, the metastability of nanocrystals is a sensitive function of their surface bonding, but the nanocrystal surface affects the crystallite in two distinct ways. First, surface atoms can make up 5–40% of the mass of a nanocrystal, and thus contribute significantly to the overall thermodynamic properties of the material. Second, the nanocrystal surface chemistry can raise the activation barriers for many thermodynamically favoured processes. Thus, it is important to consider the surface in both the kinetic and thermodynamic treatments of phase behaviour.

This section will describe the current status of research in two different aspects of nanocrystal phase behaviour: melting and solid–solid phase transitions. In the case of melting, thermodynamic considerations of surface energies can explain the reduced melting point observed in many nanocrystals. Strictly thermodynamic models, however, are not adequate to describe solid–solid phase transitions in these materials.

C2.17.5.1 NANOCRYSTALS AT HIGH TEMPERATURES

In a classic study, Buffat *et al* used electron microscopy at high temperature to measure the melting point of gold nanocrystals as a function of their size [181]. They observed that smaller nanocrystals have lower melting temperatures; this size-dependent melting point varies as roughly $1/R$. Thus, a gold nanocrystal of 5 nm diameter, for example, melts at 885°C, 180°C lower than bulk gold. Since this original study, this behaviour has been observed in both metal and semiconductor nanocrystals [182, 183, 184, 185, 186, 187, 188, 189 and 190], using techniques ranging from electron microscopy [191, 192 and 193] to nanocalorimetry [194].

The simplest approach to understanding the reduced melting point in nanocrystals relies on a simple thermodynamic model which considers the volume and surface as separate components. Whether solid or melted, a nanocrystal surface contains atoms which are not bound to interior atoms. This raises the net free energy of the system because of the positive surface free energy, but the energetic cost of the surface is higher for a solid cluster than for a liquid cluster. Thus the free-energy difference between the two phases of a nanocrystal becomes smaller as the cluster size

decreases and a reduction in the melting point is observed [181, 187, 193]. A variety of more elaborate theories which treat the nanocrystal surface structure, as well as the exact treatment of lattice strain, can provide quantitative agreement with measurements [195, 196 and 197].

While the variation in the melting temperature of nanocrystals can be explained using classic thermodynamic arguments, the process by which nanocrystals, or even bulk solids, undergo melting is an active area of research. [198, 199] Nanocrystals offer ideal systems with which to explore the role of kinetics in melting. Nanocrystals embedded in solid media have been observed to superheat, existing as solids well above their melting point [200, 201]. The extent of the superheating and its dependence on heating rate are sensitive to the nanocrystal size. One explanation is that long-range density fluctuations are responsible for inducing melting in solids [202]; in

nanocrystals these modes may be restricted. Another important issue in the mechanism of bulk melting is the role of surfaces and defects [192, 203, 204]. Several studies suggest that surface melting occurs at temperatures below the bulk melting point. This phenomenon would have important consequences for the thermal stability of nanocrystals as it could lead to shape changes in isolated nanocrystals and fusion in tightly packed nanocrystal arrays [179, 180, 205, 206, 207, 208 and 209].

Melting is only one of many processes that nanocrystals can undergo when they are heated. Temperature-induced phase transitions are equally important in nanocrystals, especially in covalent materials such as oxides [210]. Unlike melting and the solid–solid phase transitions discussed in the next section, these phase changes are not reversible processes: they occur because the crystal structure of the nanocrystal is metastable. For example, titania made in the nanophase always adopts the anatase structure. At higher temperatures the material spontaneously transforms to the rutile bulk stable phase [211, 212 and 213]. The role of grain size in these metastable–stable transitions is not well established; the issue is complicated by the fact that the transition is accompanied by grain growth which clouds the interpretation of size-dependent data [214, 215 and 216]. *In situ* TEM studies, however, indicate that the surface chemistry of the nanocrystals play a crucial role in the transition temperatures [217, 218].

C2.17.5.2 NANOCRYSTALS AT HIGH PRESSURE

The ability to control pressure in the laboratory environment is a powerful tool for investigating phase changes in materials. At high pressure, many solids will transform to denser crystal structures. The study of nanocrystals under high pressure, then, allows one to investigate the size dependence of the solid–solid phase transition pressures. Results from studies of both CdSe [219, 220, 221 and 222] and silicon nanocrystals [223] indicate that solid–solid phase transition pressures are elevated in smaller nanocrystals.

The observation of elevated transition pressures in nanocrystals may at first appear to contradict the observation that the melting temperatures of these same systems are reduced. This is not a contradiction, though, because the important variable in both phase changes is the *difference* in the free energies of the two states. If the high-pressure crystal structure (rock-salt in the case of the II–VI nanocrystals) has a higher surface free energy than the low-pressure phase (wurtzite for II–VI nanocrystals), then the transition pressure will be elevated in small clusters for strictly thermodynamic reasons. This explanation is consistent with the observed elevation in the phase transition pressures observed in smaller nanocrystals, and was the first model proposed to describe this high-pressure behaviour [219].

More recently, studies of the hysteresis of these phase transitions have illuminated the importance of kinetic factors in solid–solid phase transitions [224]. The change between crystal structures does not occur at the same point when pressure is increasing, as when it is decreasing; the difference between this ‘up-stroke’ and ‘down-stroke’ pressure

provides a measure of the activation energy required for the transition. These experiments have been performed both as a function of cluster size as well as temperature for II–VI nanocrystals [224]. As the nanocrystal size increases, a larger number of atoms in the crystallite interior must move during the phase transition, so the kinetic barrier increases. However, in sufficiently large nanocrystals, there is a larger volume available and defect-nucleation is a viable pathway. This lowers the kinetic barrier as the bulk limit is approached. For more complete reviews of studies of pressure-dependent properties in nanocrystals see [225, 226].

C2.17.6 CONCLUSIONS

This review has covered many of the essential features of the physical chemistry of nanocrystals. Rather than provide a detailed description of the latest and most detailed results concerning this broad class of materials, we have instead outlined the fundamental concepts which serve as departure points for the most recent research. This necessarily limited us to a discussion of topics that have a long history in the community, leaving out some of the new and emerging areas, most notably nonlinear optical studies [152] and magnetic nanocrystals [227]. Also, the

study of the electrical transport behaviour of nanocrystals and nanocrystal assemblies is an area of growing interest. Both single-electron transport and collective transport through organized assemblies [128, 228, 229 and 230] promise to be of great scientific as well as technological importance. Finally, we note that we did not discuss the many possible applications for nanocrystals. These little solids offer many unique and tunable features which promise to make them important materials in areas as diverse as microelectronics [64, 231, 232, 233 and 234] and biotechnology [107, 235]. Finally, we note that a search of one particular on-line scientific abstract database, using the keyword 'nanocrystal', demonstrates a monotonic increase in the number of publications per year, each year since 1993. This highly unscientific measure is nonetheless quite satisfying, as it testifies to the health and vitality of this exciting field of chemical physics.

ACKNOWLEDGMENTS

We would like to acknowledge A P Alivisatos and C V Shank, as well as all past and current members of the Alivisatos group for their insight and help with many of the ideas and data presented in this chapter. We would also like to acknowledge Steve Robertson for inspired assistance with the references and figures.

REFERENCES

- [1] Himenez P C and Rajagopalan R 1997 *Principles of Colloid and Surface Chemistry* (New York: Dekker)
 - [2] Ostwald W 1917 *Theoretical and Applied Colloid Chemistry* (London: Chapman and Hall)
 - [3] Henglein A 1988 Mechanism of reactions on colloidal microelectrodes and size quantization effects *Top. Curr. Chem.* **143** 115
 - [4] Brinker C J and Scherer G W 1990 *Sol-gel Science* (London: Academic)
-
- 20-
- [5] Bergna H E 1994 *The Colloid Chemistry of Silica* (Washington, DC: American Chemical Society)
 - [6] Rittner M N and Abrham T 1998 Nanostructured materials: an overview and commercial analysis *COM* 36
 - [7] Siegel R W 1994 Nanostructured materials: mind over matter *Nanostruct. Mat.* **4** 121
 - [8] Siegel R W 1996 Creating nanophase materials *Sci. Am.* **275** 74
 - [9] Siegel R W 1996 Gas phase synthesis and mechanical properties of nanomaterials *Analysis* **24** M10
 - [10] Murray C B, Norris D J and Bawendi M G 1993 Synthesis and characterization of nearly monodisperse CdE (E = S, Se, Te) semiconductor nanocrystallites *J. Am. Chem. Soc.* **115** 8706
 - [11] Peng Z G, Wickham J and Alivisatos A P 1998 Kinetics of II–VI and III–V colloidal semiconductor nanocrystal growth: focusing of size distributions *J. Am. Chem. Soc.* **120** 5343
 - [12] Whetten R L *et al* 1996 Nanocrystal gold molecules *Adv. Mater.* **8** 428
 - [13] Littau K A *et al* 1993 A luminescent silicon nanocrystal colloid via a high temperature aerosol reaction *J. Phys. Chem.* **97** 1224
 - [14] Andres R P *et al* 1996 Self-assembly of a two-dimensional superlattice of molecularly linked metal clusters *Science* **273** 1690

- [15] Mahoney W and Andres R P 1995 Aerosol synthesis of nanoscale clusters using atmospheric arc evaporation *Mat. Sci. Eng. A—Struct. Mat.* **204** 160
- [16] Vossmeier T *et al* 1994 CdS nanoclusters: synthesis, characterization, size dependent oscillator strength, temperature shift of the excitonic transition energy and reversible absorbance shift *J. Phys. Chem.* **98** 7665
- [17] Mews A *et al* 1994 Preparation, characterization and photophysics of the quantum dot quantum well system CdS/HgS/CdS *J. Phys. Chem.* **98** 934
- [18] Turkevich J, Stevenson P C and Hillier 1951 A study of the nucleation and growth processes in the synthesis of colloidal gold *Trans. Faraday Soc.* **11** 55
- [19] Chow M K and Zukoski C F 1994 Gold sol formation mechanism: role of colloidal stability *J. Coll. Int. Sci.* **165** 97
- [20] Hornyak G L and Martin C R 1997 Optical properties of a family of Au-nanoparticle containing alumina membranes in which the nanoparticle shape is varied from needle-like (prolate) to pancake like (oblate) *Thin Solid Films* **303** 84
- [21] Li W, Virtanen J A and Penner R M 1995 Self-assembly of n-alkanethiolate monolayers on silver nanostructures: protective encapsulations *Langmuir* **11** 4361
- [22] Martin B R *et al* 1999 Orthogonal self-assembly on colloidal gold–platinum nanorods *Adv. Mater.* **11** 1021
- [23] Yu Y *et al* 1997 Gold nanorods: electrochemical synthesis and optical properties *J. Phys. Chem. B* **101** 6661
- [24] Lianos P and Thomas J K 1986 Cadmium sulfide of small dimensions produced in inverted micelles *Chem. Phys. Lett.* **125** 299
- [25] Fendler J 1987 Membrane mimetic chemistry *Chem. Rev.* **87** 871
- [26] Wilcoxon J P, Williamson R L and Baughman R 1993 Optical properties of gold colloids formed in inverse micelles *J. Chem. Phys.* **98** 9933
- [27] Pileni M P 1997 Nanosized particles made in colloidal assemblies *Langmuir* **13** 3266

- [28] Pileni M P 1993 Reverse micelles as microreactors *J. Phys. Chem.* **97** 6961
- [29] Shioi A and Harada M 1996 Model for the geometry of surfactant assemblies in the oil-rich phase of Winsor II microemulsions *J. Chem. Eng. Japan* **29** 95
- [30] Jin J M, Parbhakar K and Dao L H 1997 Model for water-in-oil microemulsions: surfactant effects *Phys. Rev. E* **55** 721
- [31] Giustini M *et al* 1996 Microstructure and dynamics of the water-in-oil CTAB/n-pentanol/n-hexane/water microemulsions: a spectroscopic and conductivity study *J. Phys. Chem.* **100** 3190
- [32] Dunn C M, Robinson B H and Leng F J 1990 Photon-correlation spectroscopy applied to the size characterization of water-in-oil microemulsion systems stabilized by aerosol-OT; effect of change in the counterion *Spectrochim. Acta. A* **46** 1017
- [33] Zulauf M and Eicke H 1979 Inverted micelles and microemulsions in the ternary system H₂O/aerosol-OT-isooctane as studied by photon correlation spectroscopy *J. Phys. Chem.* **83** 480
- [34] Steigerwald M L *et al* 1988 Surface derivatization and isolation of semiconductor cluster molecules *J. Am. Chem. Soc.* **110** 3046

- [35] Kortan A R, Hull R and Opila R L 1990 Nucleation and growth of CdSe on ZnS quantum crystallite seeds and vice versa in inverse micelle media *J. Am. Chem. Soc.* **112** 1327
- [36] Cizeron J and Pileni M P 1997 Solid solution of Cd ZnS nanosized particles: photophysical properties *J. Phys. Chem. B* **101** 8887
- [37] Levy L *et al* 1997 Three dimensionally diluted magnetic semiconductor clusters CdMnS with a range of sizes and compositions: dependence of spectroscopic properties on the synthesis mode *J. Phys. Chem. B* **101** 9153
- [38] Qi L, Ma J and Shen J 1996 Synthesis of copper nanoparticles in nonionic water-in-oil microemulsions *J. Colloid Interface Sci.* **186** 498
- [39] Duxin N *et al* 1997 Nanosized Fe-Cu-B alloys and composites synthesized in diphasic systems *J. Phys. Chem. B* **101** 8907
- [40] Selvan S T *et al* 1998 Gold-polypyrrole core-shell particles in diblock copolymer micelles *Adv. Mater.* **10** 132
- [41] Manna A *et al* 1997 Synthesis and characterization of hydrophobic, aptotically-dispersible silver nanoparticles in Winsor Type II microemulsions *Chem. Mater.* **9** 3032
- [42] Feltin N and Pileni M P 1997 New technique for synthesizing iron ferrite magnetic nanosized particles *Langmuir* **13** 3927
- [43] Joselvich E and Willner I 1994 Forming nanophase TiO₂ in microemulsions *J. Phys. Chem.* **98** 7628
- [44] Chhabra V *et al* 1995 Synthesis, characterization, and properties of microemulsion mediated nanophase TiO₂ particles *Langmuir* **11** 3307
- [45] Qi L *et al* 1996 Preparation of BaSO₄ nanoparticles in non-ionic W/O microemulsions *Coll. & Surf.* **108** 117
- [46] Peres-Durand S, Rouviere J and Guizard C 1995 Sol-gel processing of titania using reverse micellar systems as reaction media *Coll. & Surf.* **98** 270
- [47] Petit C and Pileni M P 1997 Nanosized cobalt boride particles: control of the size and properties *J. Magn. Magn. Mater.* **166** 82
- [48] Motte L and Pileni M P 1998 Influence of length of alkyl chain used to passivate silver sulfide nanoparticles on two- and three-dimensional self-organization *J. Phys. Chem. B* **102** 4104

- [49] Tanori J and Pileni M P 1997 Control of the shape of copper metallic particles by using a colloidal system as template *Langmuir* **13** 639
- [50] Pileni M P *et al* 1998 Template design of microreactors with colloidal assemblies: control of the growth of copper metal rods *Langmuir* **14** 7359
- [51] Esumi K, Matsuhisa K and Torigoe K 1995 Preparation of rodlike gold particles by UV irradiation using cationic micelles as templates *Langmuir* **11** 3285
- [52] Cizeron J, Robertson S T and Colvin V L 1999 Preparation of gold nanoneedles, in preparation
- [53] Emory S R and Nie S 1998 Screening and enrichment of metal nanoparticles with novel optical properties *J. Phys. Chem.* **102** 493
- [54] Taleb A, Petit C and Pileni M P 1997 Synthesis of highly monodisperse silver nanoparticles from AOT reverse micelles: a way to 2D and 3D self-organization *Chem. Mater.* **9** 950
- [55] Haber J A, Gunda N V and Buhro WE 1998 Nanostructure by design: solution phase processing routes to nanocrystalline metals, ceramics, intermetallics and composites *J. Aerosol Sci.* **29** 637

- [56] Brust M *et al* 1994 Synthesis of thiol-derivatised gold nanoparticles in a two-phase liquid-liquid system *J. Chem. Soc. Chem. Comm.* 801
- [57] Brust M *et al* 1995 Synthesis and reactions of functionalized gold nanoparticles *J. Chem. Soc. Chem. Comm.* 1655
- [58] Terril R H *et al* 1995 Monolayers in three dimensions: NMR, SAXS, thermal and electron hopping studies of alkanethiol stabilized gold clusters *J. Am. Chem. Soc.* **117** 12 537
- [59] Hosteler M J *et al* 1996 Monolayers in three dimensions: synthesis and electrochemistry of w-functionalized alkanethiolate stabilized gold cluster compounds *J. Am. Chem. Soc.* **118** 4212
- [60] Wang Z L 1998 Structural analysis of self-assembling superlattices *Adv. Mater.* **10** 13
- [61] Leff D V, Brandt L and Heath J R 1996 Synthesis and characterization of hydrophobic organically-soluble gold nanocrystals functionalized with primary amines *Langmuir* **12** 4723
- [62] Kang S Y and Kim K 1998 Comparative study of dodecanethiol derivatized nanoparticles prepared in one and two phase systems *Langmuir* **14** 226
- [63] Hostetler M J, Stokes J J and Murray R W 1996 Infrared spectroscopy of three-dimensional self-assembled monolayers: n-alkanethiolate monolayers on gold cluster compounds *Langmuir* **12** 3604
- [64] Schon G and Simon U 1995 A fascinating new field in colloidal science: small ligand stabilized metal clusters and their possible applications in microelectronics *Colloid Polym. Sci.* **273** 202
- [65] Schmid G *et al* 1981 $\text{Au}_{33}(\text{PC}_6\text{H}_5)_3\text{Cl}_6$: a gold cluster of an exceptional size *Chem. Ber.* **114** 3634
- [66] Bonnemann H *et al* 1996 Nanoscale colloidal metals and alloys stabilized by solvents and surfactants: preparation and use as catalyst precursors *J. Organometall. Chem.* **520** 143
- [67] Heath J R and LeGoues F K 1993 A liquid solution synthesis of single crystal germanium quantum wires *Chem. Phys. Lett.* **208** 263
- [68] Heath J R 1992 A liquid solution phase synthesis of crystalline silicon *Science* **258** 1131
-

- [69] Heath J R, Shiang J J and Alivisatos A P 1994 Germanium quantum dots: optical properties and synthesis *J. Chem. Phys.* **101** 1607
- [70] Buhro W E, Hickman K M and Trentler T J 1996 Turning down the heat on semiconductor growth—solution chemical syntheses and the solution-liquid-solid mechanism *Adv. Mater.* **8** 685
- [71] Micic O I and Nozik A J 1996 Synthesis and characterization of binary and ternary III–V quantum dots *J. Lum.* **70** 95
- [72] Micic O I *et al* 1994 Synthesis and characterization of InP quantum dots *J. Phys. Chem.* **98** 4966
- [73] Micic O I *et al* 1995 Synthesis and characterization of InP, GaP and GaIn_2P quantum dots *J. Phys. Chem.* **99** 7754
- [74] Micic O I *et al* 1996 Highly efficient band-edge emission from InP quantum dots *Appl. Phys. Lett.* **68** 3150
- [75] Trentler T J *et al* 1997 Solution-liquid-solid growth of indium phosphide fibers from organometallic precursors: elucidation of molecular and non-molecular components of the pathway *J. Am. Chem. Soc.* **119** 2172
- [76] Kher S S and Wells R L 1996 Synthesis and characterization of colloidal nanocrystals of capped gallium arsenide *Nanostruct. Mater.* **7** 591
- [77] Douglas T and Theopold K H 1991 Molecular precursors for indium phosphide and synthesis of small III–V semiconductor clusters in solution *Inorg. Chem.* **30** 594

- [78] Olshavsky M A, Goldstein A N and Alivisatos A P 1990 Organometallic synthesis of GaAs crystallites exhibiting quantum confinement *J. Am. Chem. Soc.* **112** 9438
- [79] Trentler T J *et al* 1999 Synthesis of TiO₂ nanocrystals by nonhydrolytic solution-based reactions *J. Am. Chem. Soc.* **121** 1613
- [80] Trentler T J *et al* 1999 Size-controlled, tetragonal nanocrystalline zirconia, in preparation
- [81] Peng X *et al* 1997 Epitaxial growth of highly luminescent CdSe/CdS core/shell nanocrystals with photostability and electronic accessibility *J. Am. Chem. Soc.* **119** 7019
- [82] Leff D V *et al* 1995 Thermodynamic control of gold nanocrystal size: experiment and theory *J. Phys. Chem.* **99** 7036
- [83] Trentler T J *et al* 1995 Solution-liquid-solid growth of crystalline III–V semiconductors: an analogy to vapor-liquid-solid growth *Science* **270** 1791
- [84] Bowen-Katari J E, Colvin V L and Alivisatos A P 1994 X-ray photoelectron spectroscopy of CdSe nanocrystals with applications to studies of the nanocrystal surface *J. Phys. Chem.* **98**
- [85] Dabbousi B O *et al* 1997 (CdSe)ZnS core-shell quantum dots: synthesis and characterization of a size series of highly luminescent nanocrystallites *J. Phys. Chem. B* **101** 9463
- [86] Peng X G *et al* 1997 Epitaxial growth of highly luminescent CdSe/CdS core/shell nanocrystals with photostability and electronic accessibility *J. Am. Chem. Soc.* **119** 7019
- [87] Guzelian A A *et al* 1996 Synthesis of size-selected surface passivated InP nanocrystals *J. Phys. Chem.* **100** 7212
- [88] Jasinski J M and LeGoues F K 1991 Photochemical preparation of crystalline silicon nanoclusters *Chem. Mater.* **3** 989
- [89] Bowles R S *et al* 1981 Generation of molecular clusters of controlled size *Surf. Sci.* **106** 117
-

- [90] Peschel S and Schmid G 1995 First steps towards ordered monolayers of ligand stabilized gold clusters *Angew. Chem. Int. Ed. Engl.* **34** 1442
- [91] Murray C B, Kagan C R and Bawendi M G 1995 Self-organization of CdSe nanocrystallites into three-dimensional quantum dot superlattices *Science* **270** 1335
- [92] Taleb A, Petit C and Pileni M P 1998 Optical properties of self-assembled 2D and 3D superlattices of silver nanoparticles *J. Phys. Chem. B* **102** 2214
- [93] Mayya K S *et al* 1998 On the deposition of Langmuir–Blodgett films of Q-state CdS nanoparticles through electrostatic immobilization at the air-water interface *Thin Solid Films* **312** 300
- [94] Dabbousi B O *et al* 1994 Langmuir–Blodgett manipulation of size-selected CdSe nanocrystallites *Chem. Mater.* **6** 216
- [95] Colvin V L, Goldstein A N and Alivisatos A P 1992 Semiconductor nanocrystals covalently bound to metal surfaces using self-assembled monolayers *J. Am. Chem. Soc.* **114** 5221
- [96] Vossmeier T, Delonno E and Heath J R 1997 Light directed assembly of nanoparticles *Angew. Chem. Int. Ed. Engl.* **36** 1080
- [97] Ohara P C, Heath J R and Gelbart W N 1997 Self-assembly of submicrometer rings of particles from solutions of nanoparticles *Angew. Chem. Int. Ed. Engl.* **36** 1078
- [98] Ohara P C *et al* 1995 Crystallization of opals from polydisperse nanoparticles *Phys. Rev. Lett.* **75** 3466

- [99] Murthy S, Wang Z L and Whetten R L 1997 Thin films of thiol-derivatized gold nanocrystals *Phil. Mag. Lett.* **75** 321
- [100] Chung S W, Markovich G and Heath J R 1998 Fabrication and alignment of wires in two dimensions *J. Phys. Chem. B* **102** 6686
- [101] Grabar K C *et al* 1996 Kinetic control of interparticle spacing in Au colloid-based surfaces-rational nanometer scale architecture *J. Am. Chem. Soc.* **118** 1148
- [102] Mirkin C A *et al* 1996 A DNA-based method for rationally assembling nanoparticles into macroscopic materials *Nature* **382** 607
- [103] Mucic R C *et al* 1998 DNA-directed synthesis of binary nanoparticle network materials *J. Am. Chem. Soc.* **120** 12 674
- [104] Peng X, Wilson T E and Alivisatos A P 1997 Synthesis and isolation of a homodimer of cadmium selenide nanocrystals *Angew. Chem. Int. Ed. Engl.* **36** 145
- [105] Loweth C J *et al* 1999 DNA-based assembly of gold nanocrystals *Angew. Chem. Int. Ed.* **38** 1808
- [106] Peng X *et al* 1999 Synthesis and isolation of a homodimer of cadmium selenide nanocrystals *Angew. Chem. Int. Ed. Engl.* **36** 145
- [107] Storhoff J J *et al* 1998 One-pot colorimetric differentiation of polynucleotides with single base imperfections using gold nanoparticle probes *J. Am. Chem. Soc.* **120** 1959
- [108] Storhoff J J and Mirkin C A 1999 Programmed materials synthesis with DNA *Chem. Rev.* **99** 1849
- [109] Elghanian R *et al* 1999 Selective colorimetric detection of polynucleotides based on the distance-dependent optical properties of gold nanoparticles *Science* **277** 1078
- [110] Merkle R C 1999 Biotechnology as a route to nanotechnology *Trends Biotechnol.* **17** 271
-

-25-

- [111] Schaaff T G *et al* 1997 Isolation of smaller nanocrystal Au molecules: robust quantum effects in the optical spectra *J. Phys. Chem. B* **101** 7885
- [112] Alvarez M M *et al* 1997 Optical absorption spectra of nanocrystal gold molecules *J. Phys. Chem. B* **101** 3706
- [113] Lover T *et al* 1997 Functionalization and capping of a CdS nanocluster: a study of ligand exchange by electrospray mass spectrometry *Chem. Mater.* **9** 1878
- [114] Lover T *et al* 1997 Electrospray mass spectrometry of thiophenolate-capped clusters of CdS, CdSe and ZnS and cadmium and zinc thiophenolate complexes: observation of fragmentation and metal, chalcogenide and ligand exchange processes *Inorg. Chem.* **36** 3711
- [115] Mattoussi H *et al* 1998 Properties of CdSe nanocrystal dispersions in the dilute regime: structure and interparticle interactions *Phys. Rev. B* **58** 7850
- [116] Mattoussi H *et al* 1996 Characterization of CdSe nanocrystalline dispersions by small angle x-ray scattering *J. Chem. Phys.* **105** 9890
- [117] Wilcoxon J P and Craft S A 1997 Liquid chromatographic analysis and characterization of inorganic nanoclusters *Nanostruct. Mater.* **9** 85
- [118] Wei G T, Liu F K and Wang C R C 1999 Shape separation of nanometre gold particles by size-exclusion chromatography *Anal. Chem.* in press
- [119] Williams D B and Carter C B 1996 vols 1–3 (New York: Plenum)

- [120] Wang Z L *et al* 1998 Bundling and interdigitation of adsorbed thiolate groups in self-assembled nanocrystal superlattices *J. Phys. Chem. B* **102** 3068
- [121] Kirkland E J 1998 *Advanced Computing in Electron Microscopy* (New York: Plenum)
- [122] Hashimoto H *et al* 1980 Direct observations of the arrangement of atoms around stacking faults and twins in gold crystals and the movement of atoms accompanying their formation and disappearance *Japan. J. Appl. Phys.* **19** L1
- [123] Marks L D 1994 Experimental studies of small particle structures *Rep. Prog. Phys.* **57** 603
- [124] Marks L D, Heine V and Smith D J 1984 Direct observation of elastic and plastic deformations at Au(111) surfaces *Phys. Rev. Lett.* **52** 656
- [125] Bovin JO, Wallenburg R and Smith D J 1985 Imaging of atomic clouds outside the surfaces of gold crystals by electron microscopy *Nature* **317** 47
- [126] Marks L D and Smith D J 1983 Direct surface imaging in small metal particles *Nature* **303** 316
- [127] Marks L D and Smith D J 1985 Direct atomic imaging of solid surfaces IV. Dislocations on Au(100) *Surf. Sci.* **157** L367
- [128] Klein D L *et al* 1997 A single-electron transistor made from a cadmium selenide nanocrystal *Nature* **389** 699
- [129] Klein D L *et al* 1996 An approach to electrical studies of single nanocrystals *Appl. Phys. Lett.* **68** 2574
- [130] Pontifex G H *et al* 1991 STM imaging of small metal particles formed in anodic oxide pores *J. Phys. Chem.* **95** 9989
- [131] Durston P J *et al* 1997 Scanning tunnelling microscopy of ordered coated cluster layers on graphite *Appl. Phys. Lett.* **71** 2940
-

- [132] Chemseddine A, Jungblut H and Boulmaz S 1996 Investigation of the nanocluster self-assembly process by scanning tunneling microscopy *J. Phys. Chem.* **100** 12 546
- [133] Grabar K C *et al* 1997 Nanoscale characterization of gold colloid monolayers—a comparison of four techniques *Anal. Chem.* **69** 471
- [134] Bottomley L A, Coury J E and First P N 1996 Scanning probe microscopy—review *Anal. Chem.* **68** 185R
- [135] Nick L, Lammel R and Fuhrmann J 1995 Latex characterization by atomic force microscopy *Chem. Eng. Technol.* **18** 310
- [136] Schleef D *et al* 1997 Radial-histogram transform of scanning probe microscopy images *Phys. Rev. B* **55** 2535
- [137] Mchale J M *et al* 1996 Surface energies and thermodynamic stability in nanocrystalline aluminas *Science* **277** 788
- [138] Kitakami O *et al* 1997 Size effect on the crystal phase of cobalt fine particles *Phys. Rev. B* **56** 13 849
- [139] Cullity B D 1978 *Elements of X-ray Diffraction* (Reading, MA: Addison-Wesley)
- [140] Suryanarayana C and Norton M G 1998 *X-Ray Diffraction: A Practical Approach* (New York: Plenum)
- [141] Bawendi M G *et al* 1989 X-ray structural characterization of larger CdSe semiconductor clusters *J. Chem. Phys.* **91** 7282

- [142] Vanderputten D *et al* 1996 Angle resolved x-ray photoelectron spectroscopic experiments on the full series of molecular $[\text{Au}_{55}(\text{PR}_3)_{12}\text{Cl}_6]$ clusters *J. Chem. Soc. Dalton Trans.* **8** 1721
- [143] Sachleben J R *et al* 1998 Solution-state NMR studies of the surface structure and dynamics of semiconductor nanocrystals *J. Phys. Chem. B* **102** 10 117
- [144] Sachleben J R *et al* 1992 NMR studies of the surface structure and dynamics of semiconductor nanocrystals *Chem. Phys. Lett.* **198** 431
- [145] Marcus M A *et al* 1991 Structure of capped CdSe clusters by EXAFS *J. Phys. Chem.* **95** 1572
- [146] Rockenberger J *et al* 1998 The contribution of particle core and surface to strain, disorder and vibrations in thiocapped CdTe nanocrystals *J. Chem. Phys.* **108** 7807
- [147] Alivisatos A P 1996 Perspectives on the physical chemistry of semiconductor nanocrystals *J. Phys. Chem.* **100** 13 226
- [148] Brus L E 1993 *NATO ASI School on Nanophase Materials* ed G C Hadjipanayis (Dordrecht: Kluwer)
- [149] Alivisatos A P 1996 Semiconductor clusters, nanocrystals and quantum dots *Science* **271** 933
- [150] Heath J R and Shiang J J 1998 Covalency in semiconductor quantum dots *Chem. Soc. Rev.* **27** 65
- [151] Brus L 1998 Chemical approaches to semiconductor nanocrystals *J. Phys. Chem. Solids* **59** 459
- [152] Brus L 1991 Quantum crystallites and nonlinear optics *Appl. Phys. A* **53** 465
- [153] Bawendi M G, Steigerwald M L and Brus L E 1990 The quantum mechanics of larger semiconductor clusters ('quantum dots') *Ann. Rev. Phys. Chem.* **41** 477
- [154] Kreibig U and Genzel L 1985 Optical absorption of small metallic particles *Surf. Sci.* **156** 678

- [155] Pankove J I 1971 *Optical Processes in Semiconductors* (New York: Dover)
- [156] Chestnoy N, Hull R and Brus L E 1986 Higher excited electronic states in clusters of ZnSe, CdSe, and ZnS: spin-orbit, vibronic and relaxation phenomena *J. Chem. Phys.* **85** 2237
- [157] Guzelian A A *et al* 1997 Colloidal chemical synthesis and characterization of InAs nanocrystal quantum dots *Appl. Phys. Lett.* **69** 1432
- [158] Rossetti R, Nakahara S and Brus L E 1983 Quantum size effects in the redox potentials, resonance Raman spectra and electronic spectra of CdS crystallites in aqueous solution *J. Chem. Phys.* **79** 1086
- [159] Rossetti R *et al* 1984 Size effects in the excited electronic states of small colloidal CdS crystallites *J. Chem. Phys.* **80** 4464
- [160] Brus L E 1984 Electron-electron and electron-hole interactions in small semiconductor crystallites: the size dependence of the lowest excited electronic state *J. Chem. Phys.* **80** 4403-9
- [161] Brus L 1986 Zero-dimensional 'excitons' in semiconductor clusters *IEEE J. Quantum Electron.* **22** 1909
- [162] Shiang J J *et al* 1996 Symmetry of annealed wurtzite CdSe nanocrystals: assignment to the C_{3v} point group *J. Phys. Chem.* **100** 13 886
- [163] Leung K, Pokrant S and Whaley K B 1998 Exciton fine structure in CdSe nanoclusters *Phys. Rev. B* **57** 12 291
- [164] Blanton S A *et al* 1997 Dielectric dispersion measurements of CdSe nanocrystals colloids: observations of a permanent dipole moment *Phys. Rev. Lett.* **79** 865

- [165] Colvin V L and Alivisatos A P 1992 CdSe nanocrystals with a dipole moment in the excited state *J. Chem. Phys.* **97** 730
- [166] Colvin V L, Cunningham K L and Alivisatos A P 1994 Electric field modulation studies of optical absorption in CdSe nanocrystals: dipolar character of the excited state *J. Chem. Phys.* **101** 7122
- [167] Mittleman D M *et al* 1994 Quantum size dependence of femtosecond electronic dephasing and vibrational dynamics in CdSe nanocrystals *Phys. Rev. B* **49** 14 435
- [168] Banin U *et al* 1997 Quantum confinement and ultrafast dephasing dynamics in InP nanocrystals *Phys. Rev. B* **55** 7059
- [169] Bawendi M G *et al* 1992 Luminescence properties of CdSe quantum crystallites: resonance between interior and surface localized states *J. Chem. Phys.* **96** 946
- [170] Bawendi M G *et al* 1990 Electronic structure and photoexcited carrier dynamics in nanometre size CdSe clusters *Phys. Rev. Lett.* **65** 1623
- [171] Cerullo G, De Silverstri S and Banin U 1999 Size-dependent dynamics of coherent acoustic phonons in nanocrystal quantum dots *Phys. Rev. B* **60** 1928
- [172] Bohren C F and Hoffman D R 1983 *Absorption and Scattering of Light by Small Particles* (New York: Wiley)
- [173] Link S and El-Sayed M A 1999 Size and temperature dependence of the plasmon absorption of colloidal gold nanoparticles *J. Phys. Chem. B* **103** 4212
- [174] Kriebig U and Fragstein C V 1969 The limitation of electron mean free path in small silver particles *Z. Physik* **224** 307
- [175] Klar T *et al* 1998 Surface-plasmon resonances in single metallic nanoparticles *Phys. Rev. Lett.* **80** 4249
-

- [176] Brus L E and Trautman J K 1995 Nanocrystals and nano-optics *Phil. Trans. R. Soc. A* **353** 313
- [177] Nirmal M *et al* 1996 Fluorescence intermittency in single CdSe nanocrystals *Nature* **383** 802
- [178] Mackenzie J D and Ulrich D R 1988 *Ultrastructure Processing of Advanced Ceramics* (New York: Wiley-Interscience)
- [179] Olynick D L, Gibson J M and Averback R S 1998 Impurity-suppressed sintering in copper nanophase materials *Phil. Mag. A* **77** 1205
- [180] Flueli M, Buffat P A and Borel J P 1988 Real time observation by high resolution electron microscopy (HREM) of the coalescence of small gold particles in the electron beam *Surf. Sci.* **202** 343
- [181] Buffat P and Borel J P 1976 Size effect on the melting temperature of gold particles *Phys. Rev. A* **13** 2287
- [182] Castro T *et al* 1990 Size-dependent melting temperature of individual nanometre-sized metallic clusters *Phys. Rev. B* **42** 8548
- [183] Valov P M and Leiman V I 1997 Size effects in the melting and crystallization temperatures of copper chloride nanocrystals in glass *JETP Lett.* **66** 510
- [184] Goldstein A N, Colvin V L and Alivisatos A P 1991 Observation of melting in 30 angstrom diameter CdS nanocrystals *Mater. Res. Soc. Symp. Proc.* **206** 271
- [185] Goldstein A N, Echer C M and Alivisatos A P 1992 Melting in semiconductor nanocrystals *Science* **256** 1425
- [186] Goldstein A N 1996 The melting of silicon nanocrystals: submicrometre thin-film structures derived from nanocrystal precursors *Appl. Phys. A* **62** 33

- [187] Peppiat S J 1975 The melting of small particles II. Bismuth *Proc. R. Soc.* **354** 401
- [188] Peppiatt S J and Sambles J R 1975 The melting of small particles. I. Lead *Proc. R. Soc.* **345** 387
- [189] Allen G L *et al* 1986 Small particle melting of pure metals *Thin Solid Films* **144** 297
- [190] Pocza J F, Barna A and Barna P B 1969 Formation processes of vacuum deposited indium films and thermodynamical properties of submicroscopic particles observed by *in situ* electron microscopy *J. Vac. Sci. Technol.* **6** 472
- [191] Kofman R *et al* 1989 Solid-liquid transition of metallic clusters: occurrence of surface melting *Physica A* **157** 631
- [192] Kofman R *et al* 1994 Surface melting enhanced by curvature effects *Surf. Sci.* **303** 231
- [193] Allen G L, Gille W W and Jesser W A 1980 The melting temperature of microcrystals embedded in a matrix *Acta Metall.* **28** 1695
- [194] Lai S L *et al* 1996 Size-dependent melting properties of small tin particles: nanocalorimetric measurements *Phys. Rev. Lett.* **77** 99
- [195] Ross J and Andres R P 1981 Melting temperature of small clusters *Surf. Sci.* **106** 11
- [196] Borel J P 1981 Thermodynamical size effect and the structure of small clusters *Surf. Sci.* **106** 1
- [197] Skripov V P, Koverda V P and Skokov V N 1981 Size effect on melting of small particles *Phys. Status Solidi A* **66** 109
- [198] Boyer L L 1985 Theory of melting based on lattice instabilities *Phase Trans.* **5** 1
- [199] Cotteril R M J 1980 The physics of melting *J. Crystal Growth* **48** 582
-

- [200] Blackman M, Peppiat S J and Sambles J R 1972 Superheating of bismuth *Nature Phys. Sci.* **239** 61
- [201] Stella A *et al* 1996 Comparative study of thermodynamic properties of metallic and semiconducting nanoparticles in a dielectric matrix *Mater. Res. Soc. Symp. Proc.* **400** 161
- [202] Tolla F D, Ercolessi F and Tosatti E 1995 Maximum overheating and partial wetting of nonmelting solid surfaces *Phys. Rev. Lett.* **74** 3201
- [203] Kofman R *et al* 1991 Melting of non-spherical ultrafine particles *Z. Phys. D* **20** 267
- [204] Lutsko J F *et al* 1989 Molecular-dynamic study of lattice-defect-nucleated melting in metals using an embedded-atom-method potential *Phys. Rev. B* **40** 2841
- [205] Wang Z L *et al* 1998 Shape transformations and surface melting of cubic and tetrahedral platinum nanocrystals *J. Phys. Chem. B* **102** 6145
- [206] Harfenist S A and Wang Z L 1999 High-temperature stability of passivated silver nanocrystal superlattices *J. Phys. Chem. B* **103** 4342
- [207] Medeiros G *et al* 1998 Shape transition of germanium nanocrystals on a silicon (001) surface from pyramids to domes *Science* **279** 353
- [208] Lewis L J, Jensen P and Barrat J L 1997 Melting, freezing and coalescence of gold nanoclusters *Phys. Rev. B* **56** 2248
- [209] Zeng P *et al* 1998 Nanoparticle sintering simulations *Mater. Sci. Eng. A* **252** 301
- [210] Barsoum M 1997 *Fundamentals of Ceramics* (New York: McGraw-Hill)

- [211] Hahn H, Logas J and Averback R S 1990 Sintering characteristics of nanocrystalline TiO₂ *J. Mater. Res.* **5** 609
- [212] Iida Y *et al* 1997 *In situ* anti-Stokes Raman monitoring of gel-to-anatase phase transformation of titania *Appl. Spectrosc.* **51** 673
- [213] Kumar K P, Keizer K and Burggraaf A J 1994 Stabilization of the porous texture of nanostructured titania by avoiding a phase transformation *J. Mater. Sci. Lett.* **59**
- [214] Edelson L H and Glaeser A M 1988 Role of particle substructure in the sintering of monosized titania *J. Am. Ceram. Soc.* **71** 225
- [215] Penn R L and Banfield J F 1999 Formation of rutile nuclei at anatase (112) twin interfaces and the phase transformation mechanism in nanocrystalline titania *Am. Miner.* **84** 871
- [216] Zhang H and Banfield J F 1999 New kinetic model for the nanocrystalline anatase-to-rutile transformation revealing rate dependence on number of particles *Am. Miner.* **84** 528
- [217] Agrawal A, Cizeron J and Colvin V L 1998 *In situ* high-temperature transmission electron microscopy observations of the formation of nanocrystalline TiC from nanocrystalline anatase (TiO₂) *Microsc. Microanal.* **4** 269
- [218] Yin J S and Wang Z L 1997 *In situ* structural evolution of self-assembled oxide nanocrystals *J. Phys. Chem. B* **101** 8979
- [219] Haase M and Alivisatos A P 1992 Arrested solid-solid phase transition in 4-nm-diameter CdS nanocrystals *J. Phys. Chem.* **96** 6756
- [220] Tolbert S H and Alivisatos A P 1993 Size dependence of the solid-solid phase transition in CdSe nanocrystals *Z. Phys. D* **26** 56

-30-

- [221] Tolbert S H and Alivisatos A P 1994 Size dependence of a first order solid-solid phase transition: the wurtzite to rock salt transformation in CdSe nanocrystals *Science* **265** 373
- [222] Tolbert S H and Alivisatos A P 1995 The wurtzite to rock salt structural transformation in CdSe nanocrystals under high pressure *J. Chem. Phys.* **102** 4642
- [223] Tolbert S H *et al* 1996 Pressure-induced structural transformation in Si nanocrystals: surface and shape effects *Phys. Rev. Lett.* **76** 4384
- [224] Chen C C *et al* 1997 Size dependence of structural metastability in semiconductor nanocrystals *Science* **276** 398
- [225] Alivisatos A P 1997 Scaling law for structural metastability in semiconductor nanocrystals *Ber. Bunsenges Phys. Chem.* **101** 1573
- [226] Herhold A B *et al* 1999 Structural transformations and metastability in semiconductor nanocrystals *Phase Trans.* **68** 1
- [227] Sun S and Murray C B 1999 Synthesis of monodisperse cobalt nanocrystals and their assembly into magnetic superlattices *J. Appl. Phys.* **85** 4325
- [228] Chen S *et al* 1998 Gold nanoelectrodes of varied size: transition to molecular like charging *Science* **280** 2098
- [229] Shiang J J *et al* 1998 Cooperative phenomena in artificial solids made from silver quantum dots: the importance of classical coupling *J. Phys. Chem.* **102** 3425
- [230] Markovich G, Collier C P and Heath J R 1998 Reversible metal-insulator in ordered metal nanocrystal

monolayers observed by impedance spectroscopy *Phys. Rev. Lett.* **80**

- [231] Brus L 1993 Capped nanometre silicon electronic materials *Adv. Mater.* **5** 286
- [232] Heath J R *et al* 1998 A defect tolerant computer architecture: opportunities for nanotechnology *Science* **280** 1716
- [233] Colvin V L, Schlamp M C and Alivisatos A P 1994 Light-emitting diodes made from cadmium selenide nanocrystals and a semiconducting polymer *Nature* **370** 354
- [234] Schlamp M C, Peng X and Alivisatos A P 1997 Improved efficiencies in light emitting diodes made with CdSe(CdS) core/shell nanocrystals and a semiconducting polymer *J. Appl. Phys.* **82** 2345
- [235] Bruchez M *et al* 1998 Semiconductor nanocrystals as fluorescent biological labels *Science* **281** 2013
- [236] Jiang P *et al* 1999 Single-crystal colloidal multilayers of controlled thickness *Chem. Mater.* **11** 2132
- [237] Palik E D 1985 *Handbook of Optical Constants of Solids* (Orlando, FL: Academic)

-1-

C2.18 Etching and deposition

HP Gillis

C2.18.1 INTRODUCTION

Numerous technological processes involve removal of material from a solid surface by etching, or addition of new material to a solid surface (in this case called the substrate) by deposition. Both classes of reactions have long been studied in liquid solution [1]. Since about 1980 process demands of the microelectronics industry have stimulated development of both etching and deposition processes involving solid surfaces and gas phase reactants [2]. The general field of gas–surface reactions can be unified by classification into four groups, according to whether the products are volatile or involatile, and whether the reaction products incorporate atoms from the surface (see table C2.18.1). These reaction groups all share common fundamental concepts, and can be investigated by common techniques developed during the very active period of modern surface science since about 1975 [3]. These concepts and techniques have been introduced at several earlier points in this book. Regarding applications, the discussion of catalysis and corrosion in [chapter C2.7](#) and [chapter C2.8](#) respectively complement the present chapter.

Table C2.18.1. Classification of gas–solid reactions.

	Volatile products	Involatile products
Solid atoms incorporated	Etching	Corrosion
No atoms from solid	Catalysis	Deposition

Because surface chemical reactions occur at a localized geometrical interface between phases, transport processes are strongly coupled with these reactions. The overall reaction can be decomposed into a series of steps—attachment or ‘sticking’ of reactants to the surface; diffusion of reactants on the surface; formation of reaction product; disposition of reaction product—any one of which can be rate-limiting for the overall reaction. Fundamental progress in understanding etching and deposition reactions relies upon isolating one of these steps for investigation in simplified model circumstances, either theoretical or experimental. A comprehensive review published in 1994 summarizes such fundamental results from reactive molecular beam studies of both etching and deposition reactions. That review is highly recommended as an introduction to the field, since it also describes the fundamental concepts in adsorption dynamics and chemisorption, as well as the relevant experimental techniques [4].

In practical applications, gas–surface etching reactions are carried out in plasma reactors over the approximate pressure range 10^{-4} –1 Torr, and deposition reactions are carried out by molecular beam epitaxy (MBE) in ultrahigh vacuum (UHV: below 10^{-9} Torr) or by chemical vapour deposition (CVD) in the approximate range 10^1 – 10^3 Torr. These applied processes can be quite complex, and key individual reaction rate constants are needed as input for modelling and simulation studies—and ultimately for optimization—of the overall processes.

-2-

The objective of this chapter is to provide an introduction to etching and deposition for chemical physicists and physical chemists so they can select key fundamental questions in these complex processes for experimental or theoretical studies in much simpler model configurations. Since modern etching and deposition methods have developed mainly empirically, they are closely tied to and named after the experimental techniques used. We provide a brief introduction to this terminology, as well as a brief guide to the process literature, including review articles and advanced references. The bulk of this chapter is devoted to a discussion of selected themes and examples in etching and deposition where fundamental studies have already made advances. The goal of this discussion is to illustrate opportunities and approaches to these problems, not to provide a critical review of the present state of these fields.

C2.18.2 INDUSTRIAL IMPORTANCE OF ETCHING AND DEPOSITION

Etching (e.g., for decorating glass objects, for removing oxides and other impurities from the surfaces of metals) and deposition (e.g., for coating and passivating metallic surfaces) have a long and distinguished history in the chemical processing industries. The present discussion is limited to gas–surface reactions important in the microelectronics and optoelectronics industries.

C2.18.2.1 DRY ETCHING FOR PATTERN DEFINITION IN MICROELECTRONICS

The purpose of etching is to transfer features from a device design mask to an underlying film of device material, replicating accurately the cross-sectional profile of each feature. Various possible results are schematically illustrated in figure C2.18.1 where the mask is still in place after etching to show the connection between etch profile and the mask dimensions. In the earliest days of integrated-circuit fabrication, features 20–50 μm wide were transferred by wet etching into films about 1 μm thick, producing rounded, undercut or isotropic etching, illustrated in region A in the figure. As new design rules to increase the speed of devices and to pack devices more densely on the wafer required lateral dimensions below 1 μm , the errors due to undercut became intolerable and the need arose for a new method that would achieve anisotropic etching, to define straight sidewalls (B). Various implementations of plasma etching described below achieve this goal under proper conditions.

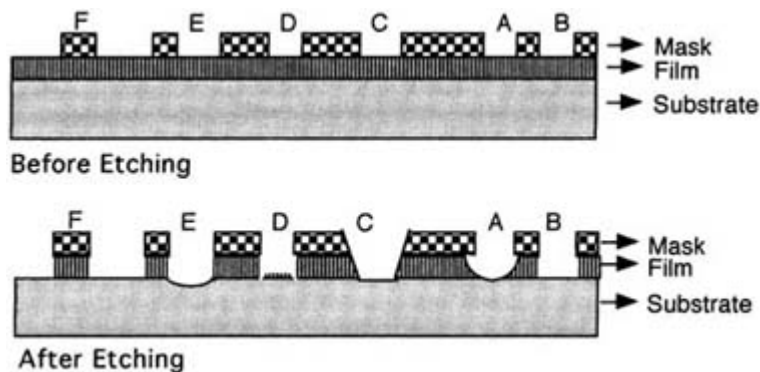


Figure C2.18.1. Schematic representation of various results of etching through a mask. The regions marked by letters are defined and described in the text.

-3-

All versions of plasma etching expose the sample and mask to ion bombardment in the presence of chemically reactive species. The interplay between these two components determines the net result of the etching process. Although the mechanism is still not understood in detail, it is recognized that ion-enhanced reactions at the bottom of the opening in the pattern (but not along the sidewalls) of the feature being transferred are essential for anisotropy. If ion bombardment overwhelms the chemical component of the etching process, the mask edges will be eroded and the profile will be overcut (C). Excessive ion component also produces trenching (D) because ions reflected from the sidewalls appear at the edge of the bottom and increase the ion flux (and the etch rate) in that region. Ideally, etch processes will be highly selective between materials; non-selectivity is illustrated in region E. In regions A–E small amounts of material are removed through the mask to define a recess into the film. In region F, much more material is removed to define a mesa in the film. Regions B and F illustrate the ideal features from which advanced devices are constructed.

Depending on conditions of ion energy and flux, temperature and chemical environment these features may receive ion bombardment damage while being etched. The term ‘etch damage’ includes effects such as creation of lattice defects or interstitials due to momentum transfer from the ions, ‘knock-on’ of various impurities into the substrate and electrical breakdown from non-uniform charge accumulations, all of which compromise the optical and electrical properties of the etched surface. Characterization of microstructural changes that constitute damage and development of new etching methods to reduce or eliminate damage, are major themes in present-day dry etching research. These newer methods of plasma etching are described below.

The most common form of anisotropic plasma etching, called reactive ion etching (RIE), is achieved by an AC glow discharge (usually at 13.56 MHz) between two metal electrodes in a reactive feed gas (figure (C2.18.2)(a)). RIE delivers ions to the sample with energy ~ 300 eV and therefore inflicts substantial damage, which must be removed in subsequent process steps. In electron cyclotron resonance (ECR) etching (figure (C2.18.2) (b)), microwave power is coupled into the cylindrical source chamber by an antenna and one or more electromagnet coils around the source generate the cyclotron resonance frequency. Coupling the excitation energy into the reactor by various types of inductive coil leads to inductively coupled plasma (ICP) etching, for which the electromagnets are not necessary. Both ECR and ICP can generate a very high-density plasma, in which arrival energy of the ions is controlled by independent RF electrical bias of the sample stage. Arrival energies can be controlled below ~ 50 eV, which is still sufficient to cause damage in many cases. A substantial literature exists on the design and performance of RIE, ICP and ECR systems [5]. Anisotropic etching is also achieved by chemically assisted ion beam etching (CAIBE) in which independent beams of ions and reactive species are directed to the sample (figure (C2.18.2) (c)). CAIBE provides much greater independent control over the identity, energy, direction of incidence and flux of both the ionic and reactive species than does RIE, ICP or ECR. Since the ion beams are extracted from broad-area plasma sources of the Kaufman type [6], ion energy must be ~ 200 eV to obtain useful current and ion-inflicted damage is common.

-4-

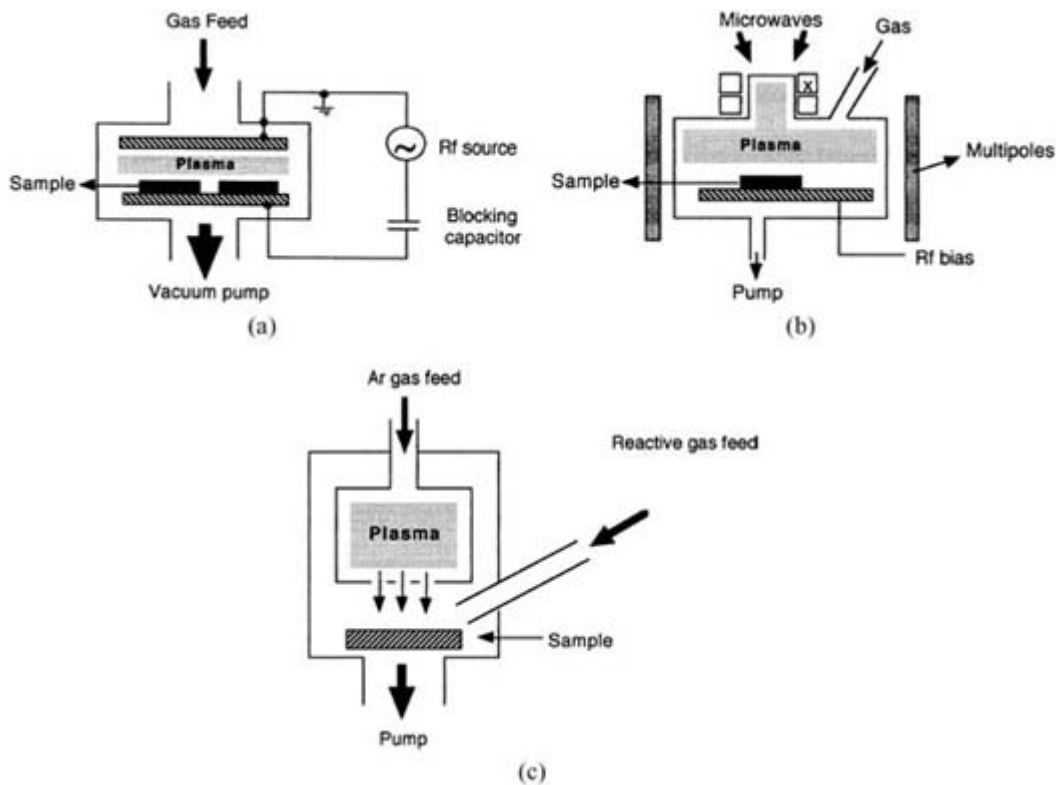


Figure C2.18.2. Schematic representations of various experimental configurations for plasma etching. (a) Reactive ion etching (RIE). (b) Electron cyclotron resonance etching (ECR). (c) Chemically assisted ion beam etching (CAIBE). The configurations are described in the text.

Dramatic progress has been achieved, largely through empirical engineering, in developing practical processes for etching silicon with fluorine species. Indeed, the ability to etch sub-micrometre features in Si CMOS device technology has played a key role in bringing the computer industry to its current highly developed state. Dry etching of III–V materials is less well developed, but is increasingly important due to the impact of III–Vs in the rapidly expanding areas of wireless and optical communication. The wide-bandgap group III nitride semiconductors, which are especially interesting as emitters in the blue region of the spectrum, present special challenges in dry etching [7].

C2.18.2.2 FUNDAMENTAL ISSUES IN DRY ETCHING

In order to design and optimize anisotropic dry etching processes, several issues must be understood:

- What is the mechanism by which ion-enhanced reactions give anisotropic etching?
- What are the underlying surface thermochemical reactions that are being enhanced?
- What controls the detailed evolution of the feature profile during etching? What is the relative importance of directed transport of neutral reactive species and of ion enhancement?
- What are the structure and composition of the surface during etching?
- How can damage be eliminated from the process?

Although substantial progress has been made, much work remains to be done, especially in the characterization and elimination of damage. Each of these issues invites study by methods of chemical physics and physical chemistry, especially when examined in a simplified model environment. Representative themes and examples, illustrating both progress achieved and remaining questions, are presented in [section C2.18.3](#).

C2.18.2.3 DEPOSITION PROCESSES IN MICROELECTRONICS AND OPTOELECTRONICS

While etching controls the lateral dimensions of microscopic devices in integrated circuits (figure C2.18.1), deposition controls their vertical dimensions, which are equally important in device function. Modern devices may comprise thin films of deposited material between 1 nm and 50 μm in thickness, which in turn may include numerous individual thinner layers. Each of these layers must be deposited with highly controlled and reproducible properties, including composition, thickness, strain, microstructure and surface morphology. Advanced optical and electronic devices from compound semiconductors require epitaxial growth, in which the crystalline orientation of the deposited film is registered with that of the substrate.

Numerous techniques have been developed for depositing films from vapours, ranging from straightforward evaporation to advanced chemical transport in which reactions are activated by heat, light or plasma. These have been surveyed in two comprehensive reviews [8, 9] and two popular interdisciplinary textbooks [10, 11]. The three most widely used chemically based techniques are:

- (1) *CVD*: gaseous reactants (precursors) delivered to a heated substrate in a flow reactor undergo thermal reaction to deposit solid films at atmospheric or reduced pressure, and volatile side products are pumped away. CVD is used for conductors, insulators and dielectrics, elemental semiconductors and compound semiconductors and is a 'workhorse' in the silicon microelectronics industry.
- (2) *Metallorganic chemical vapour deposition (MOCVD) or organometallic vapour phase epitaxy (OMVPE)*: in this variation of CVD, the precursors are organometallic compounds of the III–V or II–VI elements, a very large number of which are available. Because of chemical similarity in the various groups of elements, a wide range of compound semiconductor alloys can be produced, e.g. $\text{GaAs}_{1-x}\text{P}_x$, the basis of the familiar red light emitting diodes (LEDs) and lasers, and $\text{In}_x\text{Ga}_{1-x}\text{N}$, the basis of the new blue LEDs and lasers [12]. MOCVD is widely used for epitaxial growth of compound semiconductor heterostructures essential for optical and high-speed electronic devices.
- (3) *Metallorganic MBE (MOMBE)*: the 'solid source' Knudsen cells in conventional MBE are replaced with gaseous beams of organometallic precursors, directed toward a heated substrate in UHV. Compared to MOCVD, MOMBE eliminates gas phase reactions that may complicate the deposition surface reactions, and provides lower growth temperatures.

All three techniques have a vast and readily accessible literature, and are discussed regularly at numerous scientific conferences around the world.

C2.18.2.4 FUNDAMENTAL ISSUES IN CHEMICAL DEPOSITION OF THIN FILMS

The fundamental steps in CVD, MOCVD and MOMBE processes can be classified as follows [13]:

- (a) adsorption of precursors at the growth surface,
- (b) surface diffusion of precursors to growth sites,
- (c) surface reactions: incorporation of film constituents,
- (d) nucleation, followed by growth of film microstructure and topography,
- (e) desorption of by-products from surface reactions.

It is difficult to observe these surface processes directly in CVD and MOCVD apparatus because they operate at pressures incompatible with most techniques for surface analysis. Consequently, most fundamental studies have selected one or more of these steps for examination by molecular beam scattering, or in simplified model reactors from which samples can be transferred into UHV surface spectrometers without air exposure. Reference [4] describes many such studies. Additional themes and examples, illustrating both progress achieved and remaining questions, are presented in [section C2.18.4](#).

C2.18.3 SELECTED THEMES AND EXAMPLES IN ETCHING STUDIES

Plasma etching was introduced into silicon device fabrication technology in the middle 1970s in order to obtain the anisotropic pattern definition needed for device features smaller than 1 μm . Early processes used complex mixtures of fluorocarbon gases and additives selected to optimize anisotropy, selectivity and rate through time-consuming empirical studies. Almost immediately, model studies of essential features of the reactions in simpler environments were undertaken to obtain fundamental insights on which optimization could be rationally designed. Twenty-five years later, this is still a rich and productive field of research.

C2.18.3.1 EXPERIMENTAL STUDIES OF THERMAL ETCHING REACTION KINETICS AND DYNAMICS

It was quickly recognized that a purely thermal reaction between F atoms and Si surfaces was a key component of plasma etching. This reaction was studied in a series of UHV experiments where an effusive molecular beam of XeF_2 , which readily decomposed to give F atoms, was directed onto a Si surface. After some early controversies, careful mass spectrometric analysis in which the product flux was modulated demonstrated that the main reaction product was SiF_4 [14, 15]. Measured velocity and energy distributions of the SiF_4 products were non-Maxwellian for the temperature of the surface, showing excesses of both ‘cool’ and ‘hot’ molecules [16]. This suggests at least two modes of product desorption, both different from simple evaporation of weakly bound molecules. The overall dynamics of formation and desorption of SiF_4 is complicated, since the reaction occurs in a complex fluorosilyl ‘corrosion layer’ readily formed at the Si surface by the small and highly reactive F atoms [17]. This layer is described further in section C2.18.3.3. A series of studies by Engel and co-workers examined the effect on this reactive adlayer at the Si surface of using chlorine versus fluorine, and molecules versus atoms. In all cases coverage of etchant adequate to produce steady-state etching led to complex reactive adlayers [18].

In view of the complex kinetics and dynamics of the overall reaction, attention was directed to dynamical studies of individual steps, particularly the attachment of etchant molecules to surfaces. The emerging theme is that the sticking probability and the structure of the adsorbate layer depend strongly on the energy of the incoming molecules and on the structure of the surface, and that chemisorption can be highly site-selective. For example, at low incident kinetic energy where precursor-mediated chemisorption dominates, Cl_2 forms large islands of SiCl on $\text{Si}(111)-(7 \times 7)$, while at high incident energy direct activated chemisorption dominates, and Cl is adsorbed only at isolated sites [19].

-7-

The dangling bonds of a Si surface abstract one F atom from an incident F_2 molecule while the complementary F atom is scattered back into the gas phase [20]. This abstractive mechanism leads to F adsorption at single sites rather than at adjacent pairs of sites, as observed directly by scanning tunnelling microscopy [21]. Br atoms adsorb only to Ga atoms in the second layer of $\text{GaAs}(001)-(2 \times 4)$, where empty dangling bonds on the Ga atoms can be filled by electrons from the Br atoms [22].

C2.18.3.2 EXPERIMENTAL STUDIES OF ION-ENHANCED ETCHING REACTION KINETICS AND DYNAMICS

In what may be the single most influential experiment in the field of dry etching, John Coburn and Harold Winters of the IBM Almaden Research Centre in San Jose, CA, USA demonstrated that the rate of the F–Si thermal etching reaction is greatly enhanced if the Si surface is simultaneously bombarded with a beam of energetic Ar^+ ions and exposed to a flux of F atoms at thermal energies [23]: see figure C2.18.3. This phenomenon plausibly explained the origin of anisotropic etching in regions B and F of figure C2.18.1 since, under proper conditions of pressure and power, the electric fields in the plasma reactor could steer ions from the plasma onto the sample along its normal direction and enhance the rate only at the bottom of the feature being etched [24]. Ion-enhanced etching model studies, in which independent beams of energetic ions and electrically neutral reactive species were simultaneously directed onto the substrate in high vacuum or UHV with various *in situ* reaction diagnostic tools, constituted an important simplification of an essential step in plasma etching; many such studies followed the original example of Coburn and Winters.

The molecular mechanism of ion enhancement proved to be both subtle and complex. The complexity of the

intrinsic ‘corrosion layer’ in F–Si etching made it difficult to determine whether ion bombardment influenced the adsorption of reactant, formation of product or desorption of product [25]. Ion enhanced Cl–Si etching, which is not thermally spontaneous at room temperature, demonstrated that ion-induced recoil implantation of Cl into the Si lattice contributed substantially to the net reaction [26, 27 and 28]. Therefore, in steady-state ion-enhanced etching the ions contribute simultaneously to the formation and removal of the ‘corrosion layer’, and the only well defined experiment is to measure the kinetic energy distribution of each departing product species, as a function of ion beam energy and current density, by analogy with the sputtering of clean solids by inert ion beams [29]. Sputtering had already been explained by the linear cascade theory in which the incident ion transfers energy through a series of binary collisions with atoms in the solid as it penetrates beneath the surface [30]. The resulting fast recoils eventually cause atoms to be ejected from the solid with a kinetic energy distribution $\Phi(E) \propto E/(E+U_0)^3$ where U_0 is the surface escape energy, usually approximated as the sublimation energy of the solid [31]. Careful analysis of the data reviewed from several different ion-enhanced etching studies (Si(Cl₂; Ar⁺), Si(XeF₂; Ar⁺), Si(SF₆; Ar⁺), Si/SiO₂(Cl₂; Ar⁺), Si/SiO₂(XeF₂; Ar⁺)) showed that in all cases more than 80% of the molecules in each product species departed the surface with an energy distribution characteristic of a linear cascade [32, 33]. Each product species required a different value of U_0 ; the values determined were consistent with reasonable qualitative models of the ion-induced mixing layer or ‘corrosion layer’. To date no definitive quantitative description has been obtained, but the general trend is that in the energy range 0.5–2.5 keV ion-enhanced etching is dominated by cascade processes, the details of which vary with each gas–solid combination.

Model studies of the crystalline damage caused to the substrate by ion bombardment as a side effect of ion-enhanced etching relied upon Rutherford backscattering ion channelling measurements with glancing exit path geometry to enhance depth resolution of the damaged layer. During bombardment of Si with energies 250–1000 eV, argon and neon ions caused shallow damage (65–80 Å) while hydrogen ions caused damage 275–475 Å in depth depending on dose [34]. Progress in ion-enhanced etching up to 1992 has been reviewed by the founders of the field [35].

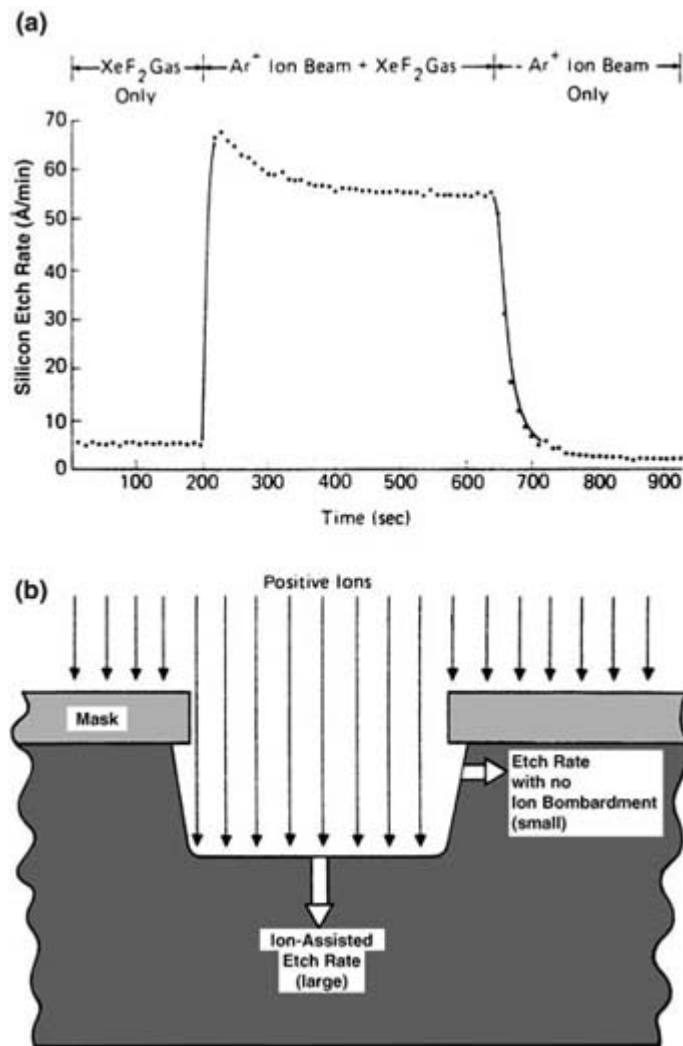


Figure C2.18.3. Relationship between ion-assisted etching and directionality in plasma etching. (a) Demonstration of the synergy between ion bombardment and reactive species during ion-assisted etching. (b) Ions incident on an etched feature. This situation prevails in glow discharges when the feature dimensions are much less than the plasma sheath thickness. Reproduced from [35]

More recent model experimental studies of ion-enhanced etching have emphasized the Si–chlorine combination with ion energies below 100 eV to simulate conditions in the newer ECR and inductively coupled high-density plasma etching systems which give highly anisotropic etching of high aspect ratio features with reduced damage in Si technology [36, 37 and 38]. As at higher ion energy, the collision cascade model describes the essential features of the process, with energy thresholds for removal of SiCl_x etch products lower than that of pure Si, in consequence of the reactive adlayer produced by chlorination of the surface.

C2.18.3.3 COMPOSITION AND STRUCTURE OF SURFACES EXPOSED TO ETCHANTS

The complexity of observed etching kinetics suggested that several species are present at the surface during steady-state reaction. Studies were begun to identify these species by surface analysis techniques to aid in clarifying the etching reaction mechanisms. This work has been developed farthest for Si–F etching, for which selected highlights are summarized as follows.

McFeely and co-workers used soft x-ray photoelectron spectroscopy (SXPS) to measure the changes in binding energies of Si(2p) levels after slight exposure to fluorine atoms via dissociative chemisorption of XeF₂ [39]. Using synchrotron radiation at 130 eV as the source enabled extreme surface sensitivity. Since this level is split into a

doublet by spin-orbit coupling, the $2p_{1/2}$ component was numerically removed from the data in order to simplify the spectrum and facilitate interpretation of peaks due to multiple species. [Figure C2.18.4](#) shows the results after exposing a clean Si(100)-(2 × 1) surface to 50 L of XeF₂. (The Langmuir (L), a convenient measure of exposing a surface to a steady background pressure of some gas for a specified time, is defined by 1 L = 1 × 10⁻⁶ Torr s. One Langmuir provides approximately one monolayer of adsorbate, if all molecules stick to the surface.) In addition to the peak for the bulk un-reacted Si, three new peaks appear with binding energy increases of approximately 1, 2 and 3 eV relative to the bulk peak. They have been assigned to SiF, SiF₂ and SiF₃ respectively. Examination of several other Si surfaces showed the same fluorosilyl species, independent of surface structure. The relative amounts of these species depended on surface structure in a manner rationalized by the number of dangling bonds available at each surface. Subsequent studies extended the exposure to 5 min at 5 × 10⁻² Torr of XeF₂, to examine conditions more typical of steady-state etching [40]. [Figure C2.18.5](#) shows the results, analysed as above. A new peak with binding energy approximately 4 eV higher than the peak for bulk Si is assigned to SiF₄, and the most abundant fluorosilyl species is SiF₃. These results are quite striking because they indicate that the etching reaction does not proceed by successive stripping of the outermost Si atomic layers, but rather by formation of a thick, highly fluorinated reaction layer—estimated to be about seven monolayers in thickness—in which the volatile reaction product SiF₄ can be trapped. Moreover, conversion of SiF₃ to SiF₄ appears to be a kinetic bottleneck in sequential fluorination reactions, but the complexity of the overall process prohibits identifying this step as rate-limiting.

-10-

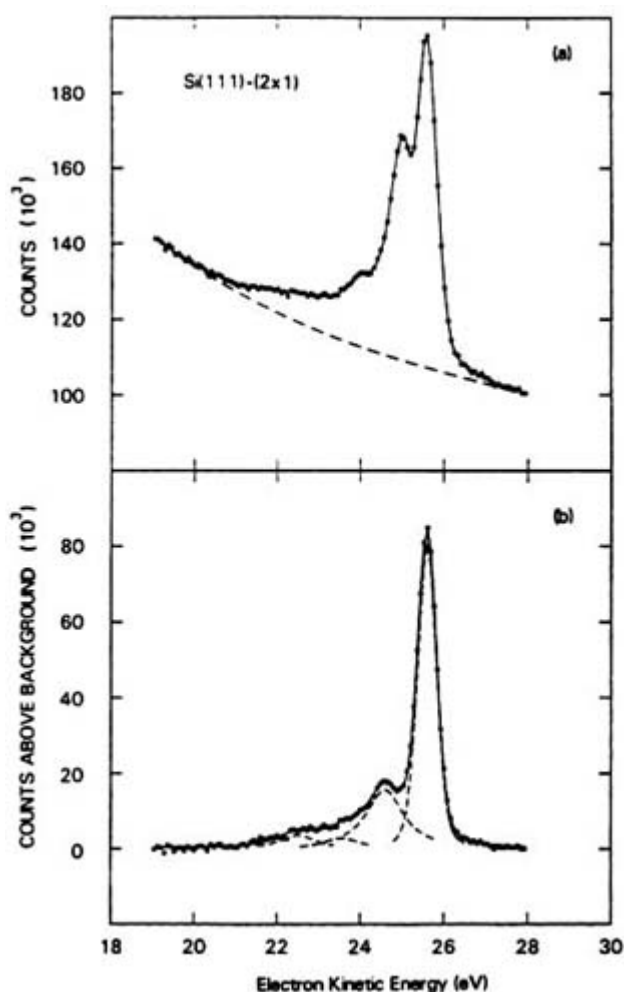


Figure C2.18.4. Upper panel shows the 2p photoemission spectrum of the Si(111)-(2 × 1) cleaved surface after exposure to approximately 50 L of XeF₂. The lower panel shows the 2p_{3/2} component of the spectrum after background subtraction. In addition to the unshifted Si(2p_{3/2}), there are three chemically shifted satellites

corresponding to SiF, SiF₂ and SiF₃ in order of increasing binding energy. The dashed curves show the separate components, and the solid curve shows the sum of the four dashed components. Reproduced from [39].

-11-

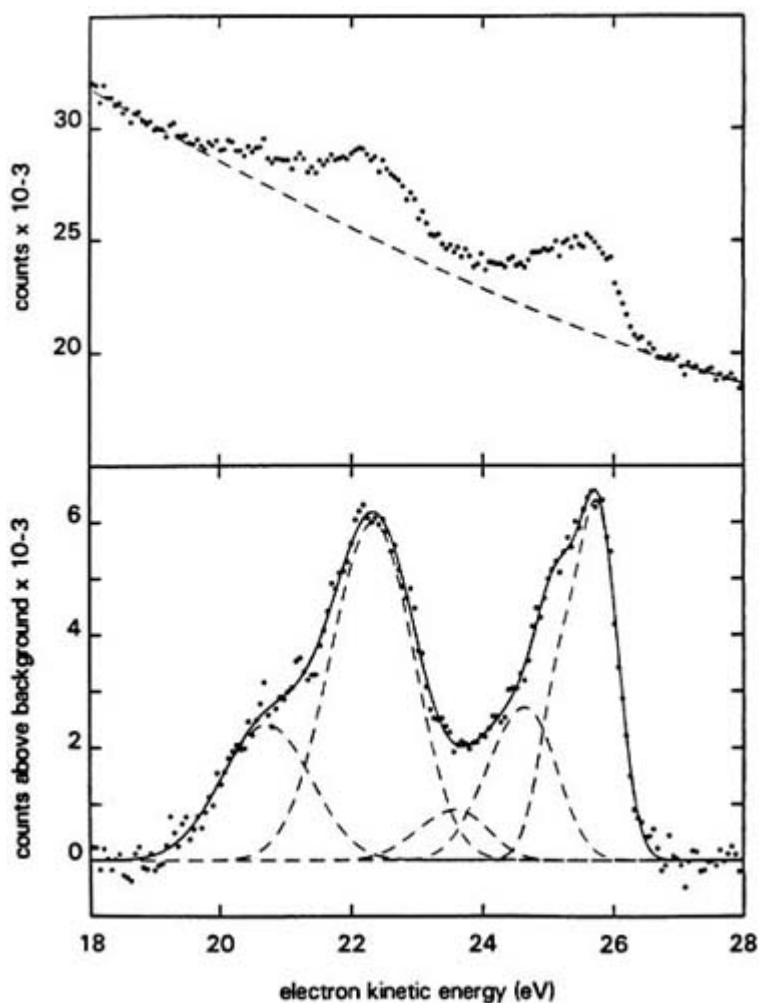


Figure C2.18.5. Si(2p) spectrum of Si(111) reacted with 5×10^{-2} Torr of XeF₂, using photon energy of 130 eV. The top panel shows the raw data and the fitted background. The bottom panel shows the spectrum after background has been subtracted and fitted into five components: bulk Si and the four fluorosilyl peaks. The solid curve is the sum of the individual dashed component curves. Reproduced from [40].

Yarmoff and co-workers continued and extended this study, supplementing the SXPS measurements with photon stimulated desorption to obtain greater depth analysis of the fluorosilyl layer, and determined the thickness and composition of the fluorosilyl layer as a function of XeF₂ exposure on a Si(111)-(7 × 7) surface [41]. As exposure increases, the concentration of SiF₃ at the surface increases, and buries a relatively constant concentration of SiF and SiF₂ that remains near the interface with the unreacted Si substrate. More precisely, the reaction layer proceeds through four regimes, as shown in [figure C2.18.6](#).

-12-

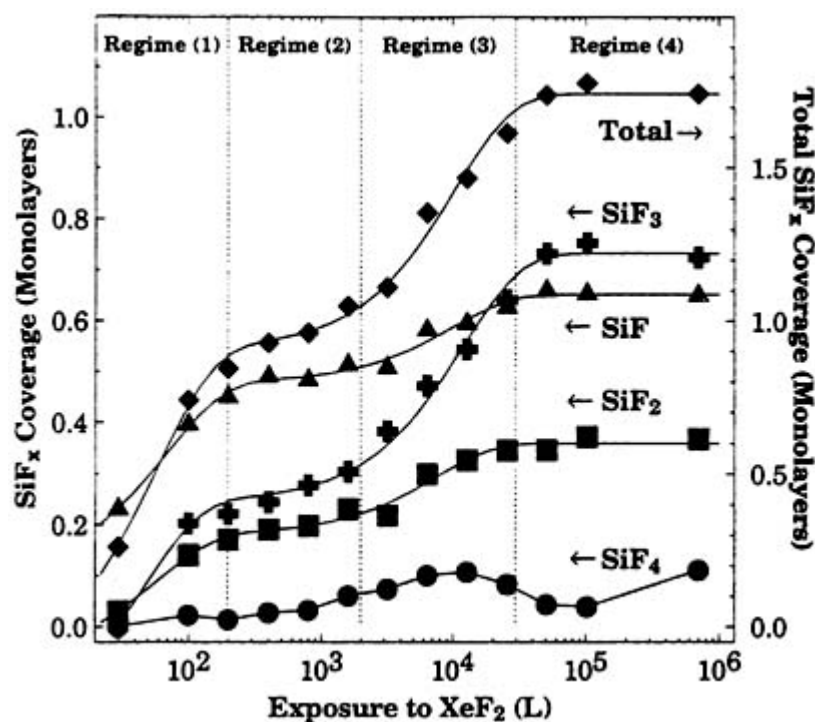


Figure C2.18.6. The coverages of fluorosilyl groups in the reaction layer shown as a function of exposure. The coverages refer to monolayers of SiF_x groups. The smooth curves are drawn through the data points. Reproduced from [41].

- (1) *Initial exposure regime:* This involves fluorination and etching of atoms in the 7×7 reconstruction, during which the destruction of the surface states in the reconstruction is correlated with creation of the species SiF_2 and SiF_3 .
- (2) *Quasi-equilibrium exposure regime:* After the 7×7 structure has been removed, quasi-equilibrium between etching and growth of the reaction layer is established. The reaction layer is about one monolayer thick, and contains primarily SiF . Defects form near the surface, partly from the large reaction exothermicity.
- (3) *Transition to steady-state etching:* The surface becomes sufficiently disordered to disrupt the quasi-equilibrium, and the reaction layer becomes a 'tree' structure of fluorosilyl chain structures terminated by SiF_3 groups.
- (4) *Steady-state etching:* Steady-state etching commences when the tree structure is fully developed.

The authors analyse these results in considerable detail, demonstrating that both the structure of the surface and steric interactions between F atoms on neighbouring SiF_3 groups influence the reaction progress.

Analysis by SXPS rationalized an earlier demonstration that heavily n-doped Si etches more rapidly with F atoms than does heavily p-doped Si, and produces less SiF_3 in the volatile etch products [42]. After exposure to sufficient XeF_2 to achieve steady-state etching, heavily p-Si(111) and n-Si(111) samples were analysed by SXPS as described above [43]. The p-type sample showed a much thicker fluorosilyl layer and substantially greater SiF_3 concentration. Formation of Si-F bonds involves considerable charge rearrangement, especially in converting the electron-depleted SiF_3 centre to SiF_4 . This conversion should be facilitated in n-type samples where the Fermi level lies in the high density of filled electronic states, but impeded in p-type samples. Thus, the SXPS data explain both the relative rates and the relative product distributions between heavily p- and n-doped samples.

Analysis by SXPS has provided insights into dry etching of III-V materials by the halogens. The general conclusions from a comprehensive review of the field are summarized as follows [44]. Molecular halogens attach by dissociative chemisorption, forming sequentially mono-, di- and tri-halides. The result is competition between etching and surface passivation, governed by temperature, surface stoichiometry and surface crystallinity. When

passivation occurs, the adsorbate usually forms an ordered overlayer. When etching occurs, there appears to be little preference between the III atom and the V atom; attachment occurs at whichever atoms are exposed at the surface.

Detailed atomic-level description of the etching mechanisms requires data not only on composition and electronic structure of the surface, as revealed by SXPS, but also on the atomic structure of the surface. The scanning tunnelling microscope has been used to demonstrate that purely thermal etching reactions depend on, and in turn influence, surface morphology [45, 46 and 47]. Since thermal barriers to adsorption of etchant molecules and to desorption of etch products depend on local structural features, various competing reaction pathways can be observed as a function of temperature to determine the dominant effects. This method holds promise for describing the detailed evolution of surface morphology during etching [48].

C2.18.3.4 THEORETICAL MODELS AND SIMULATIONS OF ETCHING REACTIONS

The method of molecular dynamics (MD), described earlier in this book, is a powerful approach for simulating the dynamics and predicting the rates of chemical reactions. In the MD approach most commonly used, the potential of interaction is specified between atoms participating in the reaction, and the time evolution of their positions is obtained by solving Hamilton's equations for the classical motions of the nuclei. Because MD simulations of etching reactions must include a significant number of atoms from the substrate as well as the gaseous etchant species, the calculations become computationally intensive, and the time scale of the simulation is limited to the order of 10^2 ps. Nonetheless, these simulations provide considerable insight into the early stages of the etching reaction.

The central ingredient in MD simulations is the interaction potential [49]. The first potential function used in MD of Si-F₂ etching reactions was developed by Stillinger and Weber by parameterizing an empirical functional form to fit data for bulk Si, gaseous F₂, gaseous SiF_x and gaseous Si₂F₆ [50]. MD simulations based on this potential predicted that etching would occur only at temperatures near the melting point of Si, and that molecular F₂ did not scatter on the clean Si(100) surface, both contrary to experimental results. Subsequent MD simulations with this same potential predicted that F atoms incident on Si(100) with kinetic energy 3 eV did react spontaneously [51].

A more accurate potential function for Si-F₂ etching was developed by Carter and co-workers by carrying out highly correlated *ab initio* quantum chemistry calculations of the interaction of F₂ with Si(100) (treated as a cluster) and fitting their results to the functional form of the Stillinger-Weber (SW) potential [52]. As a result, the SW Si-F bonding well became deeper and the non-bonding terms less repulsive. These changes enabled further MD calculations to observe buildup of a fluorosilyl layer, which is the first stage of etching [53, 54]. Running sufficient repetitions of the simulation to enable statistical predictions revealed the influence of surface steps and defects, surface coverage, molecular orientation and molecular internal energy on formation of the fluorosilyl layer [55, 56, 57 and 58]. Subsequent MD simulations of F₂ with Si(100) in which the original SW potential was compared directly with the reparametrized version due to Weakliem, Wu and Carter (WWC) demonstrated that while WWC was an improvement over SW, neither predicted adequate dissociative chemisorption when compared with experimental results [59]. More sophisticated *ab initio* calculations are therefore required to represent more accurately the interactions between F₂ and Si(100).

These theoretical descriptions of the thermal etching reaction between F₂ and Si(100) have been reviewed in some detail in the context of *ab initio* methods in surface chemistry [60].

MD simulations based on the SW potential have been applied to ion-enhanced etching systems; selected examples are Ar⁺ bombardment of Cl/Si [61], Ar⁺ bombardment of fluorosilyl/Si [62], Cl⁺ and F⁺ bombardment of Si [63], fluorosilyl ion bombardment of Si [64] and Cl⁺ bombardment of Si in the presence of a thermal background flux of Cl atoms [65]. General conclusions are as follows: (a) sputter yield in the presence of reactive species is enhanced over that for pure Si; (b) predicted reaction products agree qualitatively with experimental results; (c) surface is roughened during etching and (d) weakly bound product species are produced during ion bombardment and can be removed by thermal desorption. However, no universal mechanism for ion-enhanced etching has been identified

and the detailed evolution of the product distribution is not revealed. This probably originates partly in the short time scale of the simulations compared to the overall reaction time scale, and partly in limitations of the potential functions employed.

Perhaps greater success has appeared in simulations that emphasize the short-time consequences of binary collisions, namely energy loss from the ions and angular effects, to assess the effects of ion bombardment on the profile of the etched features [66, 67 and 68]. Earlier studies of molecular beam scattering of energetic F atoms with a fluorinated Si surface had demonstrated the influence of directed energy transport of neutral reactive species on evolution of the etch profile [69]. These various effects can be combined to simulate the buildup of non-uniform charge distributions on patterned surfaces, which have dramatic consequences both for the shape of the etched profile and for inflicting etch damage [70, 71 and 72]. The relationship between plasma process parameters, non-uniform charging, etched profile and etch damage are presently areas of intense research activity, discussed regularly at the continuing conference series International Symposium on Plasma- and Process-Induced Damage [73].

C2.18.3.5 MODEL STUDIES OF PHOTON- AND ELECTRON-ENHANCED ETCHING

In order to avoid ion bombardment damage while achieving anisotropy, alternate means of enhanced etching have been explored. In a seminal early study, Houle showed that cw bandgap excitation generated hot carriers that enhanced the etching of both n- and p-Si(100) by XeF₂ [74]. Later developments are summarized in a conference proceedings [75]. At present, photoelectrochemical etching is the only wet etching method for the wide-bandgap III-nitrides [76]; it has proved especially sensitive to defect structure in the films [77]. Coburn and Winters [23] and Veprek [78] examined electron-enhanced etching of Si. Gillis and co-workers developed low-energy electron enhanced etching (LE4), in which electrons with kinetic energy 1–15 eV and chemically reactive species at thermal velocities are delivered simultaneously to the surface. LE4 is accomplished either in UHV with separate beams of electrons and molecules [79] or in a DC plasma. In DC plasma it has produced excellent anisotropy and smooth surfaces and maintained stoichiometry of compound semiconductor surfaces when etching Si [80], GaAs [81] and GaN [82, 83 and 84]. Recently LE4 has been used to transfer a hexagonal array of 18 nm holes on a 22 nm lattice constant from a biologically derived pattern into Si(100) [85]. High-resolution cross-sectional transmission electron microscopy showed Si lattice fringes at the perimeter of the etched holes, demonstrating that LE4 does not inflict lattice displacement damage on the substrate.

C2.18.4 SELECTED EXAMPLES OF DEPOSITION STUDIES

Deposition by chemical reaction is a vast field that cannot be surveyed in the limited space here. Two particular examples have been selected because they illustrate the close relation between fundamental surface chemistry research

and process development. Moreover, both show great promise for nano-fabrication, where film thickness must be controlled at the atomic level.

C2.18.4.1 HOMOEPITAXY OF GALLIUM ARSENIDE BY ATOMIC LAYER EPITAXY

As outlined in section C2.18.2.4 above, nucleation is a key early step in film growth. This leads to micro-crystallites with uncontrolled boundaries, the coalescence of which may lead to high defect densities that require thermal annealing or other post-growth treatment to produce high-quality films. The method of atomic layer epitaxy (ALE) was developed to achieve the final crystal form immediately in compound materials by exposing the growing surface sequentially to reactants providing one of the constituent atoms [86]. ALE relies on self-limiting adsorption/reaction at each step to grow crystalline films layer by layer with precise control of thickness and superb uniformity of thickness.

Substantial work has been devoted to achieving such results for homoepitaxy of GaAs(100) by sequential reactions of the substrate with trimethylgallium (TMGa) $(\text{CH}_3)_3\text{Ga}$ and arsine AsH_3 in CVD reactors and in MOMBE configurations [87, 88]. Successful ALE growth (1 monolayer/cycle) was achieved in narrow regions of temperature and exposure. Surface science studies showed that in similar regions of temperature and exposure, TMGa dissociatively chemisorbed on the Ga-rich reconstructions of GaAs(100) to produce significant coverage of methyl groups that stabilized the complete Ga monolayer [89]. Outside these regions, the coverage of methyl groups was insufficient, and self-limiting deposition in the TMGa cycle was lost. Various model mechanisms were proposed and debated [90, 91]. A second factor influencing self-limiting deposition of Ga is the change in stoichiometry of the GaAs(100) surface as it evolves from As-rich to Ga-rich during the TMGa cycle. None of the known adsorbate-free reconstructions of the polar GaAs(100) surface are ideally terminated at one monolayer coverage, and therefore cannot support the ‘ideal’ ALE process. By viewing step edges as reservoirs where surface atoms may be added or removed in order to fill incomplete terminations, Creighton proposed the Ga-rich GaAs(100)-(1 × 2)- CH_3 reconstruction as the key participant in ALE [92]. This surface consists of 0.5 monolayer of CH_3 adsorbed on a full monolayer of dimerized Ga atoms. The adsorbate stabilizes the surface at 1 monolayer of Ga, and enables the self-limiting adsorption of Ga needed for ALE. Two candidate stabilized surfaces have been identified for the arsine cycle: the As-rich surface of GaAs(100) has a γ -(2 × 4) reconstruction terminated with 1 monolayer of As, and in the presence of adsorbed H atoms a c (2 × 8)/(2 × 4) reconstruction that can be saturated with 1 monolayer of As. ALE growth is controlled in a complex way by the competing reaction kinetics on the Ga-rich and As-rich surfaces. None of the proposed mechanisms explains all the experimental results. Progress until 1996 has been reviewed in moderate detail [93].

C2.18.4.2 DEPOSITION OF OXIDE FILMS BY ATOMIC LAYER PROCESSING

The Si/SiO₂ interface is crucial to the function of silicon devices. As the dimensions of these devices continue to shrink, the thickness of the oxide layers will be reduced to ~3 nm. The need will also arise for conformal and uniform deposition on three-dimensional structures with high aspect ratios, which cannot be achieved with the standard line-of-sight deposition methods. These demands for precise thickness control, uniformity and conformality can be met by a variation of ALE in which oxides are deposited by an alternating sequence of self-limiting reactions; one element of the oxide is deposited in each of the reactions [94]. Because these layers are amorphous rather than epitaxial with the substrate, the method is known as atomic layer processing (ALP). George and co-workers have studied the fundamental surface chemistry in two ALP reaction sequences in order to identify conditions under which they are self-limiting, and have used the results to deposit high-quality oxide films.

-16-

Deposition of SiO₂ has been achieved by exposing the substrate to the binary reaction sequence $\text{SiCl}_4 + 2\text{H}_2\text{O} \rightarrow \text{SiO}_2 + 4\text{HCl}$. This is divided into the following ‘half-reactions’ in which species at the surface are indicated by asterisks [95]:

- (A) $\text{Si-OH}^* + \text{SiCl}_4 \rightarrow \text{SiO-Si-Cl}_3^* + \text{HCl}$;
 (B) $\text{Si-Cl}^* + \text{H}_2\text{O} \rightarrow \text{Si-OH}^* + \text{HCl}$.

Surface species during the reaction were detected by Fourier transform infrared spectroscopy, using the Si–Cl stretch at 625 cm⁻¹ and the SiO–H stretch at 3740 cm⁻¹. During the (A) half-reaction, the Si–Cl peak increased while the SiO–H peak declined. The opposite behaviour was seen during the (B) half-reaction. Figure C2.18.7 shows the integrated absorbances of these two peaks *versus* time during the (A) and (B) half-reactions. These results clearly demonstrate that both of the binary reactions are complete and self-limiting at 600 K and 10 Torr; at lower temperatures the reactions did not go to completion. Temperature-programmed desorption of SiO₂ in UHV after various numbers of AB reaction cycles demonstrated that the growth rate was ~0.11 nm per cycle. Analysis by Auger electron spectroscopy (AES) showed a peak at 83 eV characteristic of the Si–SiO₂ interface, which decreased as the films grew thicker. The only AES peaks that increased with growth were those characteristic of stoichiometric SiO₂. This study not only produced chlorine-free, stoichiometric SiO₂ films by self-limiting reactions, but also provided detailed insight into the molecular mechanisms involved. A related study examined the fundamental surface chemistry in the deposition of stoichiometric Al₂O₃ by the self-limiting sequential reactions of trimethyl aluminium $(\text{CH}_3)_3\text{Al}$ and H₂O [96]. Deposition of Al₂O₃ on well characterized porous membranes

demonstrated conformal coating of the pore walls [97].

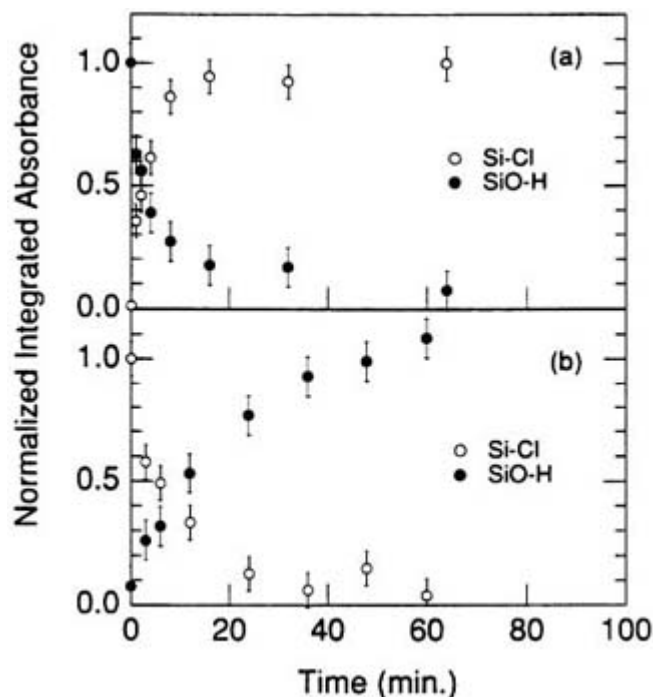


Figure C2.18.7. The integrated absorbance of the Si–Cl stretching vibration at 625 cm^{-1} and the SiO–H stretching vibration at 3740 cm^{-1} as a function of time during the (A) SiCl₄ and (B) H₂O half-reactions at 600 K and 10 Torr. Reproduced from [95].

-17-

Progress up to 1996 has been reviewed in moderate detail [98]. Subsequent developments have been summarized for Al₂O₃ [99] and for SiO₂, [100] and deposition of Si₃N₄ has been reported [101].

C2.18.5 CONCLUDING COMMENTS

The examples discussed in this chapter show a strong synergy between fundamental physical chemistry and device processing methods. This is expected only to become richer as shrinking dimensions place ever more stringent demands on process reliability. Selecting key aspects of processes for fundamental study in simpler environments will not only enable finer control over processes, but also enable more sophisticated simulations that will reduce the cost and time required for process optimization.

REFERENCES

- [1] Adamson A W 1997 *Physical Chemistry of Surfaces* 6th edn (New York: Wiley)
- [2] Vossen J L and Kern W (eds) 1991 *Thin Film Processes II* (San Diego, CA: Academic)
- [3] King D A and Woodruff D P (eds) 1988 Surface properties of electronic materials *The Chemical Physics of Solid Surfaces and Heterogeneous Catalysis* vol 5 (Amsterdam: Elsevier)
- [4] Yu M L and DeLouise L A 1994 Surface chemistry on semiconductors studied by molecular beam reactive scattering *Surf. Sci. Rep.* **19** 285–380

- [5] Lieberman M A and Lichtenberg A J 1994 *Principles of Plasma Discharges and Materials Processing* (New York: Wiley)
- [6] Harper J M E, Cuomo J J and Kaufman H R 1982 Technology and applications of broad-beam ion sources used in sputtering. Part II. Applications *J. Vac. Sci. Technol.* **21** 737–56
- [7] Gillis H P, Choutov D A and Martin K P 1996 The dry etching of Group III-Nitride wide bandgap semiconductors *J. Mater.* **48** 50–5
- [8] Bunshah R F (ed) 1982 *Deposition Technologies for Films and Coatings* (Park Ridge, NJ: Noyes)
- [9] Vossen J L and Kern W (eds) 1991 *Thin Film Processes II* (San Diego, CA: Academic)
- [10] Tu K-N, Mayer J W and Feldman L C 1992 *Electronic Thin Film Science for Electrical Engineers and Materials Scientists* (New York: Macmillan)
- [11] Smith D L 1997 *Thin-Film Deposition: Principles and Practice* (New York: McGraw-Hill)
- [12] Nakamura S, Senoh M, Nagahama S, Iwasa N, Matsushita I and Mukai T 1999 InGaN/GaN/AlGaIn-based LEDs and laser diodes *MRS Internet J. Nitride Semicond. Res.* **4S1** G1.1
- [13] Jensen K F and Kern W 1991 Thermal chemical vapor deposition *Thin Film Processes II* ed J L Vossen and W Kern (San Diego, CA: Academic) chapter III-1, pp 283–368
-

-18-

- [14] Winters H F and Houle F A 1983 Gaseous products from the reaction of XeF₂ with silicon *J. Appl. Phys.* **54** 1218–23
- [15] Houle F A 1986 A reinvestigation of the etch products of silicon and XeF₂: doping and pressure effects *J. Appl. Phys.* **60** 3018–27
- [16] Houle F A 1987 Dynamics of SiF₄ desorption during etching of silicon by XeF₂ *J. Chem Phys.* **87** 1866–72
- [17] Seel M and Bagus P S 1983 *Ab initio* cluster study of the interaction of fluorine and chlorine with the Si(111) surface *Phys. Rev.* **28** 2023–38
- [18] Engel T 1996 Fundamental aspects of the reaction of thermal and hyperthermal F, F₂, Cl, and Cl₂ with Si surfaces *Japan. J. Appl. Phys.* **35** 2403–9
- [19] Yan C, Jensen J A and Kummel A 1994 Large island formation versus single site adsorption for Cl₂ chemisorption onto Si(111)-(7 × 7) surfaces *Phys. Rev. Lett.* **72** 4017–20
- [20] Li Y L *et al* 1995 Experimental verification of a new mechanism for dissociative chemisorption: atom abstraction *Phys. Rev. Lett.* **74** 2603–6
- [21] Jensen J A, Yan C and Kummel A C 1995 Energy dependence of abstractive versus dissociative chemisorption of fluorine molecules on the silicon (111)-(7 × 7) surface *Science* **267** 493–6
- [22] Liu Yong, Komrowski A J and Kummel A C 1998 Site-selective reaction of Br₂ with second layer Ga atoms on the As-rich GaAs(001)-(2 × 4) surface *Phys. Rev. Lett.* **81** 413–16
- [23] Coburn J W and Winters H F 1979 Ion- and electron-assisted gas-surface chemistry—an important effect in plasma etching *J. Appl. Phys.* **50** 3189–96
- [24] Lieberman M A and Lichtenberg A J 1994 *Principles of Plasma Discharges and Materials Processing* (New York: Wiley) ch 1 and 2
- [25] Winters H F and Coburn J W 1985 Etching reactions at solid surfaces *Mater. Res. Soc. Symp. Proc.* **38** 189–200

- [26] Barish E L, Vitkavage D J and Mayer T M 1985 Sputtering of chlorinated silicon surfaces studied by secondary ion mass spectrometry and ion scattering spectroscopy *J. Appl. Phys.* **57** 1336–42
- [27] Mizutani T, Dale C J, Chu W K and Mayer T M 1985 Surface modification in plasma-assisted etching of silicon *Nucl. Instrum. Methods B* **7** 825–30
- [28] Sanders F H M, Kolfchoten A W, Dieleman J, Haring R A, Haring A and deVries A E 1984 Ion-assisted etching of silicon by molecular chlorine *J. Vac. Sci. Technol. A* **2** 487–91
- [29] Dieleman J, Sanders F M H, Kolfchoten A W, Zalm P C, deVries A E and Haring A 1985 Studies on the mechanism of chemical sputtering of silicon by simultaneous exposure to Cl_2 and low-energy Ar^+ ions *J. Vac. Sci. Technol. B* **3** 1384–91
- [30] Sigmund P 1969 Theory of sputtering. I. Sputtering yield of amorphous and polycrystalline targets *Phys. Rev.* **184** 383–416
- [31] Thompson M W II 1968 The energy spectrum of ejected atoms during the high energy sputtering of gold *Phil. Mag.* **18** 377–414
- [32] Zalm P C 1986 Ion-beam assisted etching of semiconductors *Vacuum* **36** 787–97
-

-19-

- [33] Zalm P C, Kolfchoten A W, Sanders F H M and Vischer P 1987 Surface processes in ion-induced etching *Nucl. Instrum. Methods B* **18** 625–8
- [34] Vitkavage D J, Dale C J, Chu W K, Finstad T G and Mayer T M 1986 Ion channeling studies of low energy ion bombardment induced crystal damage in silicon *Nucl. Instrum. Methods B* **13** 313–18
- [35] Winters H F and Coburn J W 1992 Surface science aspects of etching reactions *Surf. Sci. Rep.* **14** 161–269
- [36] Balooch M, Moalem M, Wang W and Hamza A V 1996 Low-energy Ar ion-induced and chlorine ion etching of silicon *J. Vac. Sci. Technol. A* **14** 229–33
- [37] Chang J P, Arnold J C, Zau G C H, Shin H-S and Sawin H H 1997 Kinetic study of low energy ion-enhanced plasma etching of polysilicon with atomic/molecular chlorine *J. Vac. Sci. Technol. A* **15** 1853–63
- [38] Levinson J A, Shaqfeh E S G, Balooch M and Hamza A V 1997 Ion-assisted etching and profile development of silicon in molecular chlorine *J. Vac. Sci. Technol. A* **14** 1902–12
- [39] McFeely F R, Morar J F, Shinn N D, Landgren G and Himpfel F J 1984 Synchrotron photoemission investigation of the initial stages of fluorine attack on Si surfaces: relative abundance of fluorosilyl species *Phys. Rev. B* **30** 764–70
- [40] McFeely F R, Morar J F and Himpfel F J 1986 Soft x-ray photoemission study of the silicon-fluorine etching reaction *Surf. Sci.* **165** 277–87
- [41] Lo C W, Shuh D K, Chakarian V, Durbin T D, Varekamp P R and Yarmoff J A 1993 XeF_2 etching of Si(111): the geometric structure of the reaction layer *Phys. Rev. B* **47** 15 649–59
- [42] Houle F A 1986 A reinvestigation of the etch products of silicon and XeF_2 : doping and pressure effects *J. Appl. Phys.* **60** 3018–27
- [43] Yarmoff J A and McFeely F R 1988 Effect of sample doping level during etching of silicon by fluorine atoms *Phys. Rev. B* **38** 2057–62
- [44] Simpson W C and Yarmoff J A 1996 Fundamental studies of halogen reactions with III-V semiconductor surfaces *Ann. Rev. Phys. Chem.* **47** 527–54
- [45] Villarrubia J S and Boland J J 1989 STM study of Si(111)-(7 × 7) exposed to Cl atoms *Phys. Rev. Lett.* **63** 306–9

- [46] Boland J J and Villarrubia J S 1990 Formation of Si(111)-(1 × 1) *Phys. Rev. B* **41** 9865–70
- [47] Patrin J C and Weaver J H 1993 Br₂ and Cl₂ adsorption and etching of GaAs(110) studied by use of scanning tunneling microscopy *Phys. Rev. B* **48** 17 913–21
- [48] Boland J J and Weaver J H 1998 A surface view of etching *Phys. Today* August, 34–40
- [49] Garrison B J and Srivastava D 1995 Potential energy surfaces for chemical reactions at solid surfaces *Ann. Rev. Phys. Chem.* **46** 373–94
- [50] Weber T A and Stillinger F H 1990 Dynamical branching during fluorination of the dimerized Si (100) surface: a molecular dynamics study *J. Chem Phys.* **92** 6239–45
- [51] Schoolcraft T A and Garrison B J 1991 Initial stages of etching of the Si{100}(2 × 1) surface by 3.0-eV normal incident fluorine atoms: a molecular dynamics study *J. Am. Chem. Soc.* **113** 8221–8

-20-

- [52] Wu C J and Carter E A 1992 Structures and adsorption energetics for chemisorbed fluorine atoms on Si (100)-2 × 1 *Phys. Rev. B* **45** 9065–81
- [53] Weakliem P C, Wu C J and Carter E A 1992 First-principles-derived dynamics of a surface reaction: fluorine etching of Si(100) *Phys. Rev. Lett.* **69** 200–3
- [54] Weakliem P C and Carter E A 1993 Surface chemical reactions studied via *ab initio*-derived molecular dynamics simulations: fluorine etching of Si(100) *J. Chem Phys.* **98** 737–45
- [55] Carter L E, Khodabandeh S, Weakliem P C and Carter E A 1994 First-principles-derived dynamics of F₂ reactive scattering on Si(100)-2 × 1 *J. Chem Phys.* **100** 2277–88
- [56] Carter L E and Carter E A 1994 Influence of single atomic height step surfaces on F₂ reactions with Si(100)-(2 × 1) *J. Vac. Sci. Technol. A* **12** 2235–9
- [57] Carter L E and Carter E A 1995 F₂ reaction dynamics with defective Si(100): defect-insensitive surface chemistry *Surf. Sci.* **323** 39–50
- [58] Carter L E and Carter E A 1996 *Ab initio*-derived dynamics for F₂ reactions with partially fluorinated Si(100) surfaces: translational activation as a possible etching tool *J. Phys. Chem.* **100** 873–87
- [59] Schoolcraft T A, Diehl A M, Steel A B and Garrison B J 1995 Molecular dynamics simulations of fluorine molecules interacting with a Si {100} (2 × 1) surface at 1000 K *J. Vac. Sci. Technol. A* **13** 1861–6
- [60] Radeke M R and Carter E A 1997 *Ab initio* dynamics of surface chemistry *Ann. Rev. Phys. Chem.* **48** 243–70
- [61] Feil H, Dieleman J and Garrison B 1993 Chemical sputtering of Si related to roughness formation of a Cl-passivated Si surface *J. Appl. Phys.* **74** 1303–9
- [62] Barone M E and Graves D B 1995 Chemical and physical sputtering of fluorinated silicon *J. Appl. Phys.* **77** 1263–74
- [62] Barone M E and Graves D B 1995 Chemical and physical sputtering of fluorinated silicon *J. Appl. Phys.* **77** 1263–74
- [63] Barone M E and Graves D B 1995 Molecular dynamics simulations of direct reactive ion etching of silicon by fluorine and chlorine *J. Appl. Phys.* **78** 6604–15

- [64] Helmer B A and Graves D B 1997 Molecular dynamics simulations of fluorosilyl ions with silicon *J. Vac. Sci. Technol. A* **15** 2252–61
- [65] Hanson D E, Voter A F and Kress J D 1997 Molecular dynamics simulation of reactive ion etching of Si by energetic Cl ions *J. Appl. Phys.* **82** 3552–9
- [66] Helmer B A and Graves D B 1998 Molecular dynamics simulations of Ar⁺ and Cl⁺ impacts onto silicon surfaces: distributions of reflected energies and angles *J. Vac. Sci. Technol. A* **16** 3503–14
- [67] Abrams C F and Graves D B 1998 Energetic ion bombardment of SiO₂ surfaces: molecular dynamics simulations *J. Vac. Sci. Technol. A* **16** 3006–19
- [68] Vyvoda M A, Lee H, Malyshev M V, Klemens F P, Cerullo M, Donnelly V M, Graves D B, Kornbilt A and Lee J T C 1998 Effects of plasma conditions on the shapes of features etched in Cl₂ and HBr plasmas. I. Bulk crystalline silicon etching *J. Vac. Sci. Technol. A* **16** 3247–58

-21-

- [69] Hwang G S, Anderson C M, Gordon M J, Moore T A, Minton T K and Giapis K P 1996 Gas–surface dynamics and profile evolution during etching of silicon *Phys. Rev. Lett.* **77** 3049–51
- [70] Hwang G S and Giapis K P 1999 Pattern-dependent charging in plasmas *IEEE Trans. Plasma Sci. Special Issue: Images in Plasma Science* **27** 102
- [71] Giapis K P and Hwang G S 1998 Pattern dependent charging and the role of electron tunneling *Japan. J. Appl. Phys.* **37** 2281
- [72] Hwang G S and Giapis K P 1998 The influence of surface currents on pattern-dependent charging and notching *J. Appl. Phys.* **84** 154
- [73] Complete information including the proceedings from previous conferences is available at the web site <http://www.p2id.org>
- [74] Houle F A 1989 Photochemical etching of silicon: the influence of photogenerated charge carriers *Phys. Rev. B* **39** 10 120–32
- [75] Dieleman J, Biermann U K P and Hess P (eds) 1995 Photon-assisted processing of surfaces and thin films *Proc. Symp. B E-MRS Conf. (1994) Appl. Surf. Sci.* **86** 543–81
- [76] Youtsey C, Adesida I and Bulman G 1997 Highly anisotropic photoenhanced etching of *n*-type GaN *Appl. Phys. Lett.* **71** linebreak 2151–4
- [77] Youtsey C, Adesida I, Romano L and Bulman G, Smooth photoenhanced wet etching of *n*-type GaN *Appl. Phys. Lett.* **72** 560–2
- [78] Veprek S and Sarott F A 1982 Electron-impact-induced anisotropic etching of silicon by hydrogen *Plasma Chem. Plasma Proc.* **2** 233–46
- [79] Gillis H P, Clemons J L and Chamberlain J P 1992 Low energy electron beam enhanced etching of Si(100)-(2 × 1) by molecular hydrogen *J. Vac. Sci. Technol. B* **10** 2729–33
- [80] Gillis H P, Choutov D A, Steiner P A IV, Piper J D, Crouch J H, Dove P M and Martin K P 1995 Low energy electron enhanced etching of Si(100) in hydrogen-helium DC plasma *Appl. Phys. Lett.* **66** 2475–7
- [81] Gillis H P, Choutov D A, Martin K P and Song L 1996 Low energy electron enhanced etching of GaAs(100) in chlorine-hydrogen DC plasma *Appl. Phys. Lett.* **68** 2255–7
- [82] Gillis H P, Choutov D A, Martin K P, Pearton S J and Abernathy C R 1996 Low energy electron enhanced etching of GaN/Si in hydrogen DC plasma *J. Electrochem. Soc.* **143** L251–4

- [83] Gillis H P, Choutov D A and Martin K P 1996 The dry etching of Group III-Nitride wide bandgap semiconductors *J. Mater.* **48** 50–5
- [84] Gillis H P, Choutov D A, Martin K P, Bremser M D and Davis R F 1997 Highly anisotropic, ultra-smooth patterning of GaN/SiC by low energy electron enhanced etching in DC plasma *J. Electron. Mater.* **26** 301–5
- [85] Winningham T A, Gillis H P, Choutov D A, Martin K P, Moore J T and Douglas K 1998 Formation of ordered nanocluster arrays by self-assembly on nano-patterned Si(100) surfaces *Surf. Sci.* **406** 221–8
- [86] Suntola S and Hyvarinen J 1985 Atomic layer epitaxy *Ann. Rev. Mater. Sci.* **15** 177–95
- [87] Nishizawa J, Kurabayashi T and Abe H 1987 Mechanism of surface reaction in GaAs layer growth *Surf. Sci.* **185** 249–68
- [88] Ozeki M, Usui A, Yoshinobu A and Nishizawa J (eds) 1994 *ALE-3: Proc. 3rd Int. Conf. on Atomic Layer Epitaxy (Sendai, Japan, May 1994)* *Appl. Surf. Sci.* **82/83**
-

-22-

- [89] Creighton J R, Lykke K R, Shamamian V A and Kay B D 1990 Decomposition of trimethylgallium on the gallium-rich GaAs(100) surface: implications for atomic layer epitaxy *Appl. Phys. Lett.* **57** 279–81
- [90] Yu M L 1993 A model for the atomic layer epitaxy of GaAs *Thin Solid Films* **225** 7–11
- [91] Creighton J R and Bansenauer B A 1993 The surface chemistry and kinetics of GaAs atomic layer epitaxy *Thin Solid Films* **225** 17–25
- [92] Creighton J R 1994 Surface stoichiometry and the role of adsorbates during GaAs atomic layer epitaxy *Appl. Surf. Sci.* **82/83** 171–9
- [93] George S M, Ott A W and Klaus J W 1996 Surface chemistry for atomic layer growth *J. Phys. Chem.* **100** 13 121–31
- [94] George S M, Sneh O, Dillon A C, Wise M L, Ott A W, Okada L A and Way J D 1994 Atomic layer controlled deposition of SiO₂ and Al₂O₃ using ABAB. . . binary reaction sequence chemistry *Appl. Surf. Sci.* **82/83** 460–7
- [95] Sneh O, Wise M L, Ott A W, Okada L A and George S M 1995 Atomic layer growth of SiO₂ using SiCl₄ and H₂O in a binary reaction sequence *Surf. Sci.* **334** 135–52
- [96] Dillon A C, Ott A W, Way J D and George S M 1995 Surface chemistry of Al₂O₃ deposition using Al(CH₃)₃ and H₂O in a binary reaction sequence *Surf. Sci.* **322** 230–42
- [97] Ott A W, McCarley K C, Klaus J W, Way J D and George S M 1996 Atomic layer controlled deposition of Al₂O₃ films using binary reaction sequence chemistry *Appl. Surf. Sci.* **106** 128–36
- [98] George S M, Ott A W and Klaus J W 1996 Surface chemistry for atomic layer growth *J. Phys. Chem.* **100** 13 121–31
- [99] Ott A W, Klaus J W, Johnson J M and George S M 1997 Al₂O₃ thin film growth on Si(100) using binary reaction sequence chemistry *Thin Solid Films* **292** 135–44
- [100] Klaus J W, Sneh O, Ott A W and George S M 1999 Atomic layer deposition of SiO₂ using catalyzed and uncatalyzed self-limiting surface reactions *Surf. Rev. Lett.* **6** 435–48
- [101] Klaus J W, Ott A W, Dillon A C and George S M 1998 Atomic layer controlled growth of Si₃N₄ films using sequential surface reactions *Surf. Sci.* **418** L14–19
-

C3.1 Transient kinetic studies

Robert A Goldbeck and David S Kliger

C3.1.1 INTRODUCTION AND HISTORICAL OVERVIEW

Transient, or time-resolved, techniques measure the response of a substance after a rapid perturbation. A swift 'kick' can be provided by any means that suddenly moves the system away from equilibrium—a change in reactant concentration, for instance, or the photodissociation of a chemical bond. Kinetic properties such as rate constants and amplitudes of chemical reactions or transformations of physical state taking place in a material are then determined by measuring the time course of relaxation to some, possibly new, equilibrium state. Determining how the kinetic rate constants vary with temperature can further yield information about the thermodynamic properties (activation enthalpies and entropies) of transition states, the exceedingly ephemeral species that lie between reactants, intermediates and products in a chemical reaction.

Relaxation kinetics may be monitored in transient studies through a variety of methods, usually involving some form of spectroscopy. Transient techniques and spectrophotometry are combined in time resolved spectroscopy to provide both the structural information from spectral measurements and the dynamical information from kinetic measurements that are generally needed to characterize the mechanisms of relaxation processes. The presence and nature of kinetic intermediates, metastable chemical or physical states not present at equilibrium, may be directly examined in this way.

The introduction in the 1920s of rapid mixing techniques to initiate chemical reactions first brought millisecond time resolution to the study of solution kinetics [1], overcoming the limitation of classical methods to reactions occurring in seconds or longer. Originally developed to use a continuous flow of reagents, the more reagent-conserving stopped flow approach is now commonly used and widely available in commercial instrumentation. Not only are bimolecular reactions studied by the rapid combination of reagents, but unimolecular reactions may also be initiated by rapid dilution, as in denaturant dilution studies of protein folding.

The millisecond barrier to fast kinetic studies was broken in the late 1940s and early 1950s by two developments: the flash photolysis method of Norrish and Porter [2] and the chemical relaxation techniques of Eigen [3], advances for which the three shared the 1967 Nobel Prize in chemistry. (The term relaxation techniques refers to kinetic methods in which a sudden change in an extensive parameter such as temperature, pressure, or electric field provides a perturbation from equilibrium small enough that any subsequent relaxation can be treated as a first order rate process, as discussed further below. This is distinguished from flash photolysis in which absorption of an optical photon creates a new physical or chemical state that is far from equilibrium.) The new techniques initially made possible transient kinetic studies of processes taking place on time scales as short as microseconds. The development of flash photolysis boosted the field of photochemistry tremendously by opening up transient photochemical and photophysical species such as free radicals and electronically excited states to direct observation and characterization. At the same time, the development of relaxation techniques opened up the field of fast solution kinetics by allowing researchers to directly follow the time evolutions of fast unimolecular and bimolecular reactions such as dissociations, isomerizations and near diffusion-controlled ionic association reactions.

The flash lamp technology first used to photolyse samples has since been superseded by successive generations of increasingly faster pulsed laser technologies, leading to a time resolution for optical perturbation methods that now extends to femtoseconds. This time scale approaches the ultimate limit on time resolution (Δt) available to flash photolysis studies, the limit imposed by chemical bond energies (ΔE) through the uncertainty principle, $\Delta E \Delta t \geq \frac{1}{2} \hbar$. Similarly, the most rapid relaxation method, temperature jumping by solvent absorption of a brief pulse of optical

or IR photons, is ultimately limited in time resolution by the energy redistribution processes, such as rotational and vibrational relaxations, leading to thermal equilibrium. These are of the order of picoseconds in condensed phases but can be much slower in the gas phase. The time scales applicable to some transient techniques are summarized in table C3.1.1. This article focuses on transient kinetic studies in the 10^{-9} to 1 s time regime. Ultrafast (femtosecond and picosecond) methods are covered elsewhere in the encyclopedia.

Table C3.1.1 Time-resolved methods and time scales.

Method	Time range (s)
Flow techniques	10^3 – 10^{-4}
Relaxation techniques	
Temperature jump	1 – 10^{-11}
Pressure jump	1 – 10^{-6}
Electric field jump	10^{-2} – 10^{-10}
EPR	10^{-5} – 10^{-10}
Flash photolysis	1 – 10^{-15}
Pulsed radiolysis	1 – 10^{-11}

C3.1.2 TIME RESOLVED PROCESSES

Fast transient studies are largely focused on elementary kinetic processes in atoms and molecules, i.e., on unimolecular and bimolecular reactions with first and second order kinetics, respectively (although conformational heterogeneity in macromolecules may lead to the observation of more complicated unimolecular kinetics). Examples of fast thermally activated unimolecular processes include dissociation reactions in molecules as simple as diatomics, and isomerization and tautomerization reactions in polyatomic molecules. A very rough estimate of the minimum time scale required for an elementary unimolecular reaction may be obtained from the Arrhenius expression for the reaction rate constant, $k = A e^{-E_a/RT}$. The quantity $k_B T/h$ from transition state theory provides an upper limit on the pre-exponential factor A that is of the order of 10^{13} s^{-1} , or a vibrational frequency, at room temperature. This leads to the estimate that a barrierless reaction can proceed over the course of tens or hundreds of femtoseconds. However, chemical reactions must often overcome a potential energy barrier associated with breaking bonds (while perhaps

-3-

forming others) and the addition of even a modest barrier slows the previous estimate considerably. An activation energy of only 23 kJ mol^{-1} (about 5% of a covalent bond energy), for example, will slow our hypothetical ultrafast reaction by four orders of magnitude and begin to bring it into the province of the fast kinetic methods discussed here. Moreover, entropic constraints often present in the transition state can further reduce the reaction rate constant by reducing A from the upper limit above.

The fastest bimolecular reactions are rate limited by the time it takes for reactants to diffuse toward one another. A typical diffusion-controlled bimolecular rate constant, k_D , is about $6 \times 10^9 \text{ l mol}^{-1} \text{ s}^{-1}$ for uncharged reactants dissolved in water at room temperature. (This is slower by a factor of 50 than the corresponding rate constant for binary collisions in a gas because of solvent viscosity.) We can define a pseudo-first-order rate constant, k_1 , for the bimolecular reaction of species A with B if one reactant, A for instance, is present in excess: $k_1 = k_D[A]$. This leads to an upper limit on the bimolecular reaction time constant of about half of a nanosecond for solution concentrations of A approaching 1 M, an estimate that will be proportionately slower for lower concentrations of

A. (The gas phase estimate is about 100 picoseconds for A at 1 atm pressure.) This suggests that the great majority of fast bimolecular processes, e.g., ionic associations, acid–base reactions, metal complexations and ligand–enzyme binding reactions, as well as many slower reactions that are rate limited by a transition state barrier can be conveniently studied with fast transient methods.

The absorption of a photon initiating photophysical and photochemical processes can itself be an extremely rapid event (as short as $\sim 10^{-15}$ s, for instance, given the $\sim 10^3$ cm^{-1} bandwidth typical of transitions in condensed phase polyatomic molecules and available in lasing media for ultrashort pulsed lasers). This has made light absorption a widely used trigger for fast kinetic studies. Ensuing photophysical and photochemical processes can take place on fast to ultrafast time scales. Unimolecular photophysical processes and their characteristic time scales include: fluorescence emission, 10^{-11} – 10^{-6} s; phosphorescence, 10^{-3} – 10^2 s; internal conversion (spin-conserving nonradiative relaxation) from higher excited states to the lowest excited state, 10^{-14} – 10^{-11} s; internal conversion from the lowest excited state to the ground state, 10^{-9} – 10^{-7} s; and intersystem crossing (spin-changing nonradiative relaxation) from the lowest excited singlet state to a triplet state, 10^{-11} – 10^{-8} s. Primary unimolecular photochemical processes, such as photodissociation into molecules, ions, or radicals, photo-isomerization or rearrangement and photoionization, proceed on excited state potential surfaces and typically are kinetically independent of temperature, as are primary bimolecular photochemical processes such as photodimerization, photoaddition, hydrogen atom abstraction and electron transfer to or from an excited acceptor or donor. (The kinetics of secondary, or dark, photochemical reactions proceeding from the products of primary photochemical processes will in general be dependent on temperature, however, as they take place on the ground state potential surface.) Primary photochemical processes generally compete kinetically with the photophysical processes of radiative and nonradiative relaxation as decay routes for the initially excited state. Additional bimolecular processes that may be generated by light excitation but which do not necessarily lead to permanent photoproducts include excimer and exciplex formation, the association of an excited species with a like or dissimilar ground state species, respectively, and quenching processes in which excitation energy is transferred to other species in either a contact or long-range interaction. As the above time scales suggest, current understanding of these photochemical and photophysical processes has benefited greatly from the application of fast time-resolved spectroscopic techniques.

-4-

C3.1.3 TRANSIENT SPECTROSCOPY

An apparatus for time-resolved spectroscopy can be schematically reduced to the basic elements shown in figure C3.1.1. The pump source in this figure is some device providing the perturbation needed to initiate changes in the sample to be studied. As discussed further below, this usually refers to a light source such as a laser when measurements are to be carried out on a fast time scale. However, this could refer to other types of perturbing device such as a stopped-flow apparatus, for example, to rapidly mix different reagents or a capacitor discharged across a sample cell to suddenly jump the sample temperature.

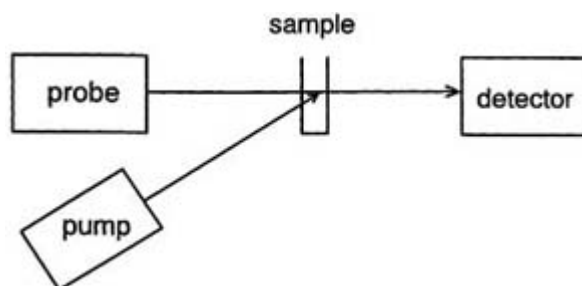


Figure C3.1.1. The basic elements of a time-resolved spectral measurement. A pump source perturbs the sample and initiates changes to be studied. Lasers, capacitive-discharge Joule heaters and rapid reagent mixers are some examples of pump sources. The probe and detector monitor spectroscopic changes associated with absorption, fluorescence, Raman scattering or any other spectral approach that can distinguish the initial, intermediate and final

states in a reaction.

A light source probes changes in the sample at various times after perturbation using some type of spectroscopy that can distinguish the initial reactant, intermediates and final product. A laser can be used to further excite the sample, producing fluorescence or Raman scattering that may be monitored as a function of time, for instance, or, alternatively, the absorption spectrum of the sample may be monitored using a variety of light sources that may be polarized or unpolarized, lasing or incoherent. Non-optical spectral techniques such as EPR [4] or NMR [5] can also be used to probe reaction dynamics, but in this chapter we will emphasize optical spectroscopies as these are most commonly used, particularly on fast time scales.

Finally, the detection system in figure C3.1.1 represents some device to detect the changes in the spectral properties of the probe beam caused by perturbing the sample. This is typically a photoelectric detector to record light coming from the sample, such as fluorescence or Raman scattering, or to record intensity changes of the probe light source in the case of absorption measurements. Probe and detection strategies can involve measurements made one wavelength at a time, using devices such as photomultipliers or photodiodes to record an intensity change as a function of time, or can involve multispectral measurements using photodiode arrays or charge-coupled devices to measure entire spectra at some specific time following application of the perturbation to the sample. In many cases, the goal is to obtain the time evolution of an entire spectrum monitoring some process of interest. This is accomplished with a single wavelength instrument by monitoring the time evolution of a signal at some wavelength and repeating this measurement at different wavelengths. Alternatively, with a multiwavelength instrument, one measures the spectrum at a specific time and repeats the spectral measurement at various times. It is also possible to accomplish this goal in a single measurement

-5-

using a streak camera, which records the spectral evolution over a range of wavelengths and delay times after a single perturbation. Streak cameras have been used mainly in ultrafast applications, however, as their high cost has tended to discourage applications to slower time regimes [6].

C3.1.4 RAPID MIXING

One can study a slow (minutes or longer) chemical reaction by mixing two chemicals in the sample compartment of a standard UV–vis spectrophotometer and measuring the spectrum as a function of time. Though perhaps not often thought of as such, this is a form of transient spectroscopy, albeit a slow one. To carry out such a measurement for a reaction which is complete on a time scale of milliseconds to seconds one needs to mix the chemicals and measure the spectra much more rapidly. For gases, this can be done by releasing reactants into a discharge flow apparatus, where they are mixed by diffusion and turbulence while being carried down a tube in an inert carrier gas such as helium. This is not, strictly speaking, a transient kinetic method, however, as the progress in time of the reaction is measured by the steady state detection of concentration as a function of distance travelled down the tube. For liquids, achieving satisfactory mixing times (not to mention conserving reactant) usually requires the use of a stopped-flow apparatus, as in figure C3.1.2 in which chemicals are rapidly forced into a sample cuvette by syringes whose plungers are quickly actuated at a specific time. A probe detection system is triggered immediately after the sample is mixed in this transient technique. Conductance may be monitored in the case of ionic solutions, whereas spectrophotometry provides a more general method for determining concentrations. Electronic detection methods provide the time resolution needed (oscilloscopes and transient recorders with response frequencies up to a few GHz are available) to monitor the conductance or spectral changes that accompany the reaction taking place. The rapid mixing approach is limited to the study of reactions taking place on time scales of milliseconds or longer simply because it takes this long for mixing to occur (although ultrarapid techniques have been developed to mix reactants on a 100 microsecond time scale [7]).

-6-

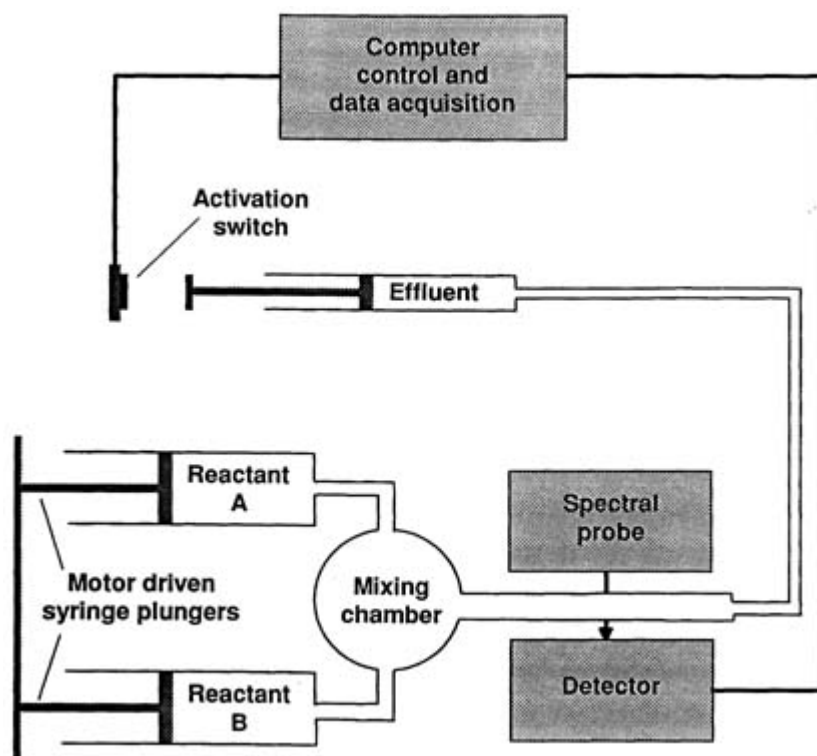


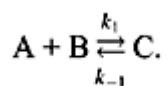
Figure C3.1.2. Stopped-flow apparatus with motor-driven syringes. Syringe plungers force the reactants A and B through a mixing chamber into a spectral cell. Kinetic data collection begins when the effluent syringe plunger is pushed out to contact an activation switch, about a millisecond after the initiation of mixing. (Adapted from Pilling M J and Seakins P W 1995 *Reaction Kinetics* (Oxford: Oxford University Press))

The example above of the stopped-flow apparatus demonstrates some of the requirements important for all forms of transient spectroscopy. These are the ability to provide a perturbation (pump) to the physicochemical system under study on a time scale that is as fast or faster than the time evolution of the process to be studied, the ability to synchronize application of the pump and the probe on this time scale and the ability of the detection system to time resolve the changes of interest.

C3.1.5 RELAXATION SPECTROSCOPY

How does one monitor a chemical reaction that occurs on a time scale faster than milliseconds? The two approaches introduced above, relaxation spectroscopy and flash photolysis, are typically used for fast kinetic studies. Relaxation methods may be applied to reactions in which finite amounts of both reactants and products are present at final equilibrium. The time course of relaxation is monitored after application of a rapid perturbation to the equilibrium mixture. An important feature of relaxation approaches to kinetic studies is that the changes are always observed as first order kinetics (as long as the perturbation is relatively small). This linearization of the observed kinetics means

that useful information about reactions involving higher order kinetic mechanisms may be obtained in a relatively simple manner. To see why this is so, consider the reaction:



The rate equation describing the kinetics of this reaction is

$$\frac{-d[A]}{dt} = \frac{-d[B]}{dt} = \frac{d[C]}{dt} = k_1[A][B] - k_{-1}[C].$$

If the equilibrium concentrations for A, B and C are a , b and c , respectively, the concentration changes resulting from the application of the perturbation will be

$$x = a - [A] = b - [B] = [C] - c$$

and we can then reduce the rate equation to

$$\frac{dx}{dt} = k_1(a - x)(b - x) - k_{-1}(c + x).$$

Expanding the right-hand side of this equation yields

$$k_1ab - k_1(a + b)x + k_1x^2 - k_{-1}c - k_{-1}x.$$

However, since the rate of change of all components is zero at equilibrium,

$$k_1ab - k_{-1}c = 0.$$

The perturbation being small, x^2 is negligible, so that

$$\begin{aligned}\frac{dx}{dt} &= -\{k_1(a + b) + k_{-1}\}x \\ &= -k_{\text{observed}}x.\end{aligned}$$

This shows that the observed rate for this process will follow first order kinetics, even though the reaction being studied is second order. Furthermore, both k_1 and k_{-1} may be determined by observing the kinetics at different starting concentrations that vary the quantity $(a+b)$.

C3.1.5.1 T JUMP

In a discharge T -jump apparatus, a capacitor is discharged across a sample cell containing conductive solution as shown in figure C3.1.3 in order to rapidly increase the temperature through Joule heating. Because equilibrium constants generally depend on temperature, the reaction mixture is rapidly triggered to change according to the kinetic scheme of the reaction under study. The change is given by van't Hoff's law:

$$\left(\frac{\partial \ln K}{\partial T}\right)_p = \frac{\Delta H^0}{RT^2}$$

which predicts that the equilibrium constant changes by about 2% per degree at room temperature for a reaction with a ΔH^0 value of 10 kJ mol⁻¹, for example.

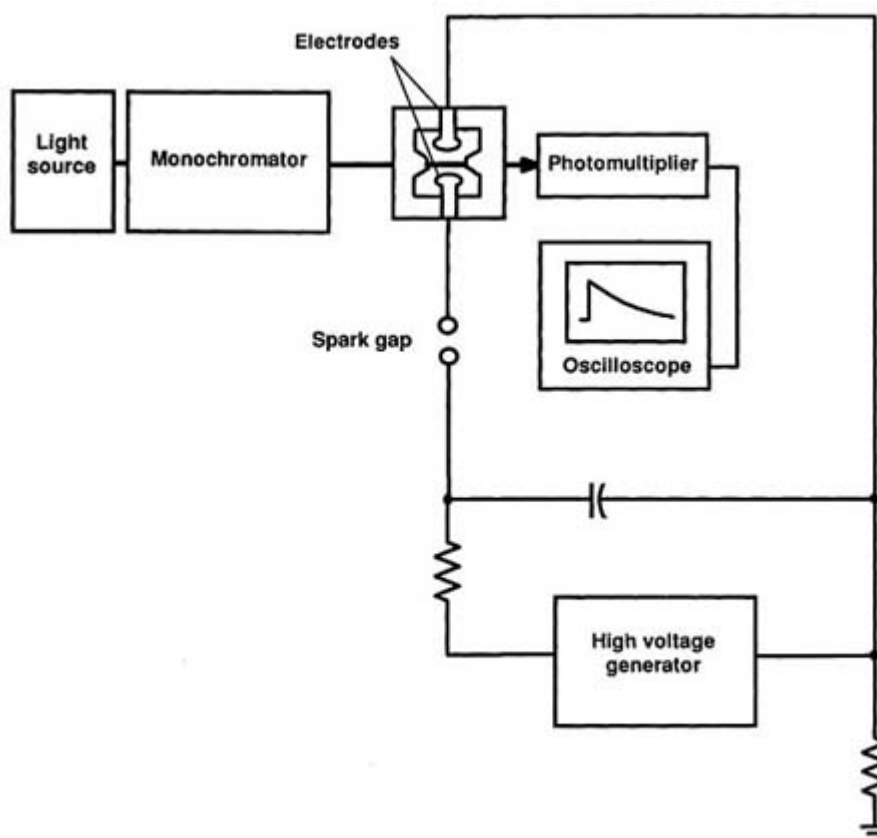


Figure C3.1.3. Schematic diagram of Joule heating T -jump apparatus for transient spectroscopy. (Adapted from French T C and Hammes G G 1969 *Methods Enzymol.* **16** 3.)

One can follow reactions of the order of microseconds or longer using a discharge T -jump. In a typical example, discharging 45 J of electrical energy into 10 cm³ of aqueous solution raises T by about 1 °C with an RC time constant of 1 μs for $R=20 \Omega$ (~0.5 M NaCl) and $C=0.1 \mu\text{F}$ (charged to 3×10^4 V).

-9-

C3.1.5.2 P JUMP

A sudden change in pressure can also be used to shift a chemical equilibrium, the change being given by the thermodynamic relation

$$\left(\frac{\partial \ln K}{\partial P}\right)_T = -\frac{\Delta V^0}{RT}$$

where ΔV^0 is the standard molar volume change for the reaction. This approach is less general than the T -jump method because relatively few reactions have significant volume changes. Examples are found, however, in the ionic association and dissociation reactions, in which solvation and electrostriction effects may produce a ΔV^0 of $\sim 10 \text{ cm}^3 \text{ mol}^{-1}$. In this case, a ΔP of 50 atm is required to produce a 2% change in K at room temperature. (Care must be taken in applying the constant T expression above for the change in K as it neglects the fact that a rapid P -jump is adiabatic and may be significantly non-isothermal, particularly in nonaqueous solvent.) Pressure jumps of the order of 10^2 atm with 10^{-5} s rise times may be obtained by rupturing a disc to suddenly admit a pressurized gas into a sample cell or, conversely, to suddenly release a pressurized gas from the cell [8, 9]. Differential P -jumps (~ 150 atm) can also be applied to study the kinetics of reactions at absolute pressures up to 2500 atm [10]. Because of the frequent need for high sensitivity in P -jump experiments, concentrations are often monitored using conductive measurements.

C3.1.5.3 E FIELD JUMP

The rapid application of a very high electric field ($|E| \sim 10^5 \text{ V cm}^{-1}$) can perturb chemical equilibria. This effect is described by the thermodynamic relation

$$\left(\frac{\partial \ln K}{\partial |E|} \right)_{P,T} = \frac{\Delta M^0}{RT}$$

where ΔM^0 is the change in standard molar polarization of the reaction. ΔM^0 is highest for reactions involving charge separations, e.g., ionic dissociations. Because of the resulting focus on conductive ionic reaction mixtures, pulsed fields of 10^{-6} s or shorter duration are typically used in order to limit Joule heating to acceptable levels. An interesting early application of the pulsed field technique was to measurement of the reaction rate constant for the prototypical proton transfer reaction $\text{H}^+ + \text{OH}^- \rightarrow \text{H}_2\text{O}$ [11]. The value measured, $1.3 \times 10^{11} \text{ M}^{-1} \text{ s}^{-1}$, is more than an order of magnitude faster than a typical diffusion controlled rate constant, reflecting the anomalously rapid diffusion of protons through water. Electric field jumps are also used to measure nanosecond electro-optic relaxation time constants for the dipole reorientations of biological macromolecules [12].

C3.1.6 FLASH PHOTOLYSIS

Laser-based pump strategies are generally necessary to study reactions taking place on time scales faster than microseconds. Lasers can be used to produce T -jumps on time scales faster than microseconds or to initiate reactions through rapid photochemical or photophysical processes. Lasers can also initiate ultrarapid mixing via a wide variety

-10-

of ‘caged’ compounds that release reactants upon photolysis [13]. Caged compounds contain a reactant moiety whose active site is blocked by a photolabile group. The cage can be rapidly photodissociated to produce a sample with reactants already microscopically mixed. One then studies the reaction on a time scale limited only by the time required for photodissociation and microscopic diffusion of reactants rather than by the time needed to macroscopically mix reagents. A similar approach can be used to rapidly change the pH of a solution to initiate a reaction or change a pH-dependent equilibrium. This can be done with compounds that exhibit dramatically different $\text{p}K_{\text{a}}$ values in their ground and excited states [14], such as sulphonated phenols [15, 16]. One can change the pH of a solution by several units within nanoseconds with such an approach. The acid–base chemistry involved is usually reversible, however, leading to eventual loss of the pH jump, often within a millisecond after excitation. Protons produced by irreversible photochemistry can provide more persistent jumps for use in single shot or flow experiments. The photoconversion of *o*-nitrobenzaldehyde to nitrosobenzoic acid, for example, produces a pH jump within a microsecond that persists for tens of milliseconds [17].

Lasers can also be used to produce T -jumps on time scales of picoseconds or longer in condensed phases such as liquid solutions. An intense pulsed laser, tuned to a solvent absorption, heats the solvent molecules on the time scale of the exciting laser pulse after rapid radiationless decay of the initial photoexcitation and rapid intra- and intermolecular energy transfer between the solvent and solute molecules’ degrees of freedom. Red or near-IR lasers are often frequency down-converted to reach the IR region where the solvent absorbs most efficiently. In aqueous solutions, for example, shifting the laser to the region around $1.5 \mu\text{m}$ results in strong absorption and temperature jumps of tens of degrees when the laser pulse is focused into a small volume. In a typical application, a 300 mJ pulse of $1.06 \mu\text{m}$ fundamental from a ns Nd:YAG laser is converted by a Raman shifter or OPO to a longer IR wavelength for efficient absorption by water. A conversion efficiency of 5%, for instance, produces 15 mJ of down-converted IR, which can be focused into a $1.5 \times 10^{-3} \text{ cm}^3$ absorbing volume to give a temperature jump of 10°C .

Reaction kinetics can be initiated most rapidly by the photoinitiation of a unimolecular reaction. With a sufficiently

fast excitation pulse, the speed of such reactions depends simply on intramolecular rates of energy transfer and the reaction dynamics themselves, precisely the properties one is often interested in studying. Researchers can study the photophysical properties of molecules by monitoring the spectra (e.g., absorption, emission or Raman) and dynamics of the excited states, or study the photochemical properties of excited states that decay through reactive channels such as isomerization, bond cleavage or ligand dissociation. It is also possible to initiate bimolecular photochemical reactions but these generally will occur on slower time scales involving the diffusion of reactant molecules to form reactive complexes.

The sensitivities of particular spectroscopic techniques to specific chemical features are described more fully in the next section. Perhaps the most common and versatile probes of reaction dynamics are time-resolved UV–vis absorption and fluorescence measurements. When molecules contain chromophores which change their structure directly or experience a change of environment during a reaction, changes in absorption or fluorescence spectra can be expected and may be used to monitor the reaction dynamics. Although absorption measurements are less sensitive than fluorescence measurements, they are more versatile in that one need not rely on a substantial fluorescence yield for the reactants, products or intermediates to be studied.

Unfortunately, the low resolution absorption spectra characteristic of condensed phase molecules at room temperature frequently do not provide a lot of information about the physicochemical nature of intermediates. Thus, time-resolved absorption measurements are often useful to initially characterize the kinetic characteristics of a reaction, but other spectroscopic methods may also be useful in probing more subtle or structure-specific mechanistic features. In the many cases in which one would like to obtain more information about the structural features of intermediates

-11-

than is available from absorption data, vibrational spectroscopies, including infra-red absorption measurements and Raman scattering, can be very useful. Often exquisitely sensitive to molecular structure, vibrational spectra contain much structural information, at least in principle, although their interpretation in terms of molecular structures may not always be straightforward in larger molecules because of spectral crowding.

It is also possible to obtain more structural information than is usually available from absorption data by making measurements with polarized light. Looking at linear dichroism (LD), the difference in absorption between linearly polarized light oriented parallel or perpendicular to a reference axis, as a function of time can provide detailed information about changes in orientation of that chromophore during the course of a reaction. An LD reference axis is determined in the molecular frame by the transition dipole of the chromophore used to photoinitiate the reaction, and in the laboratory frame by the polarization of the exciting laser. Measurements of differences in refractive index for parallel or perpendicularly polarized light (linear birefringence, LB) provide similar information and can sometimes be measured with greater sensitivity.

The use of circularly polarized light can further provide additional structural information. Time-resolved measurements of circular dichroism (CD), the difference in absorption intensity between left and right circularly polarized light, or optical rotatory dispersion (ORD), proportional to the difference in refractive index between circular polarizations, can provide information on kinetic changes in molecular structures exhibiting asymmetry. Changes in the helical content of proteins during the course of their reactions can be monitored by time-resolved CD measurements, for instance. It is also possible to induce a circular dichroism in a sample by the application of a magnetic field. Magnetic circular dichroism (and magnetic ORD) provides information complementary to that from natural CD (ORD) measurements, as discussed further below.

C3.1.7 SPECTROSCOPIC METHODS

C3.1.7.1 ULTRAVIOLET–VISIBLE ABSORPTION

Several strategies commonly used for time resolved optical absorption spectroscopy (TROA) are shown

schematically in [figure C3.1.4](#). Perhaps most common for microsecond to millisecond time-resolved measurements is the cw (continuous wave) probe approach ([figure C3.1.4\(a\)](#)). The probe can be a cw laser, provided a suitable wavelength is available, but cw xenon arc, tungsten filament or halogen lamps are most commonly used for UV–visible measurements, while glow bars have been used for IR measurements. Continuous probe sources offer the advantage of simplified timing—it is necessary to synchronize only the detection system and pump pulse. Two problems can offset this advantage, however. The first is the signal to noise ratio (S/N) of the data obtained with this method; the second is sample stability.

-12-

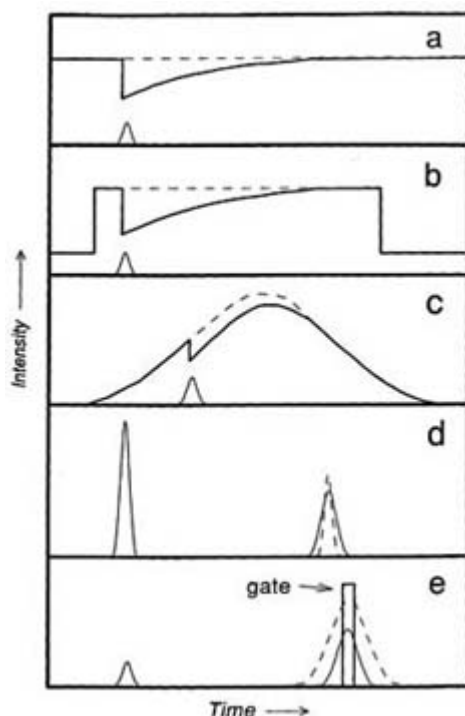


Figure C3.1.4. Schematic measurement traces depicting different probe strategies for transient spectroscopy and their typical time scales. Each panel represents the intensity of a probe beam against time. The small Gaussian at the bottom left of each panel represents the excitation (pump) pulse, defining the zero of time. (a) Strategy using a constant, cw light source as probe. The top line represents the probe light in the absence of laser pump excitation and the curved line represents a changing intensity after excitation of the sample. Time scales: milliseconds to seconds. (b) A cw arc lamp as in panel (a) augmented by a capacitive discharge across the arc to enhance intensity by a factor of 10–100. Time scales: microseconds to milliseconds. (c) A pulsed flashlamp is an alternative to the pulsed cw lamp in panel (b) that produces high power but low total energy from the probe light source, allowing for measurements on a faster time scale: nanoseconds to microseconds. (d) Pulsed lasers used for both pump and probe sources. Note that the pump pulse (left) is shown enlarged to emphasize the fact that the pump pulse must be much larger than the probe pulse. The probe pulse decreases in magnitude when the pump pulse excites the sample, creating a transient absorption. The probe pulse in the presence of the pump pulse is drawn wider only to make it easier to see in the figure. Time scales: nanoseconds to seconds. (e) With gated multichannel detection, the flashlamp probe in panel (c) is observed at a particular time delay after pump excitation, rather than monitored as a function of time as in single-wavelength detection. The probe pulse is diminished by the presence of transient absorption created by pump excitation, as measured by the spectra of the probe with and without pump sampled over a small range of delay times indicated by the ‘gate pulse’. The probe pulse and detector gate are kept overlapped as their joint time delay is varied to yield spectra as a function of time on time scales from nanoseconds to seconds.

The S/N of any light intensity measurement varies as the square root of the intensity (number of photons) produced by the source during the time of the measurement. The intensities typical of xenon arc lamps are sufficient for measurements of reasonable S/N on time scales longer than about a microsecond. However, a cw lamp will

produce few photons during measurement times faster than this and the signal will be noisy. This problem can be dealt with in

-13-

very stable samples by averaging many signals, but this approach is impractical for samples of moderate or high photolability. In fact, the cw probe approach can be problematic even with little signal averaging if the light source continues to irradiate a photolabile sample between data collection cycles. A modification of the cw probe approach improving the S/N of measurements on a nanosecond to microsecond time scale is shown in [figure C3.1.4\(b\)](#). The intensity of the cw lamp can be greatly increased (typically by a factor of several hundred) for times up to milliseconds by discharging a capacitor across an arc lamp just before firing the pump source [18].

The use of pulsed light sources such as lasers for both pump and probe, as shown in [figure C3.1.4\(c\)](#) [figure C3.1.4\(d\)](#) and [figure C3.1.4\(e\)](#) can achieve high S/N while avoiding photostability problems. The time delay between the probe and pump laser pulses can be varied over a wide range in measuring time-resolved absorption or emission. Sample photostability problems are avoided by proper adjustment of the (potentially very high) probe intensity. However, each single-wavelength, single-time measurement must be repeated at a number of wavelengths to obtain a spectrum. This process is further repeated at a number of time points to map out time-resolved spectra. This can again lead to sample deterioration in less photostable samples and can be tedious even for stable samples. A variation of this approach avoids the problem by using broad band pulsed light sources such as a laser-pumped dye, or 'soup' of laser dyes [19], or flashlamps as probe sources for multichannel measurements.

[Figure C3.1.4\(d\)](#) illustrates use of a pulsed xenon flashlamp probe source with a typical pulse width of several microseconds. This has the advantage of providing very high peak power for nanosecond measurements with high S/N, but low integrated intensity so that samples need not be exposed to excessive light from the probe source. For single-wavelength kinetic measurements, however, flashlamps have the disadvantage of being able to probe transient absorptions only for the duration of the flash, i.e., for microseconds or less.

All of the approaches described above can be used in a kinetic mode where the time evolution of absorption or emission signals are measured at one wavelength. As identifying transient intermediates often requires observing many wavelengths, an alternative is to replace photomultipliers or photodiodes with gated multichannel detectors such as intensified diode arrays ([figure C3.1.5](#)) or intensified CCDs to measure entire spectra. While these detectors can be used with cw sources, an optimal approach is to use gated multichannel detectors with flashlamps as depicted in [figure C3.1.4\(e\)](#) [20, 21].

-14-

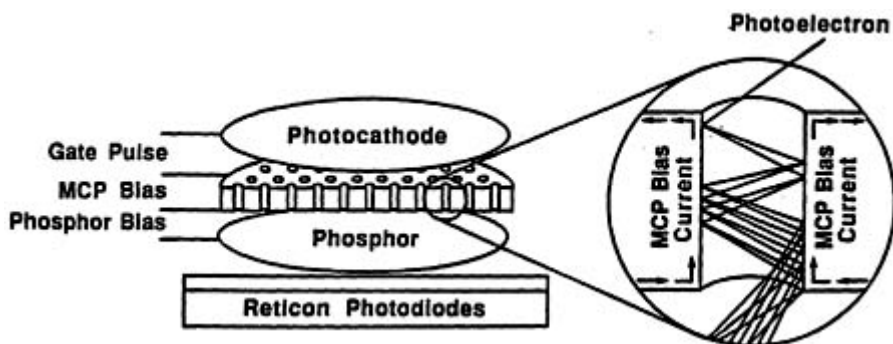


Figure C3.1.5. Schematic diagram of an intensifier-gated optical multichannel analyser (OMA) detector. The detector consists of a microchannel plate (MCP) image intensifier followed by a 1024-channel Reticon photodiode array. Light dispersed across the semitransparent photocathode ejects photoelectrons. These are accelerated toward the entrance of the microchannels by the gate pulse. The photoelectrons collide with the channel walls to produce secondary electrons, which are accelerated in turn by the MCP bias voltage to produce further collisions and electron multiplication. Electrons leaving the microchannels are further accelerated by the phosphor bias voltage,

about 6 kV, until they strike the phosphor and produce light. Several hundred photons are produced for every MCP electron, providing further gain. This light, intensified by a factor of 10^6 over the amount produced when the gate pulse is off, is detected by the photodiode array. (From Lewis J W, Yee G G and Kliger D S 1987 *Rev. Sci. Instrum.* **58** 939–44.)

Figure C3.1.6 schematically shows the use of a flashlamp probe source to efficiently measure entire spectra at a given delay time. The flashlamp output extends from the UV into the IR spectral region. Multichannel detectors can typically measure intensities simultaneously at 500 to 1000 wavelengths over this range. The multichannel detector is gated on at the peak intensity of the flashlamp to provide maximum S/N for sampling times as short as 2–5 ns. The probe light source and detector gate can be delayed together to measure spectra with constant S/N over a wide time range. It is still necessary to carry out experiments at multiple delay times with this method to obtain kinetic information, but delay times can often be logarithmically spaced, as shown in figure C3.1.7 so that fewer in number are needed than the number of wavelengths needed to accurately record the spectra of intermediates. This is particularly true when using global fitting approaches, described below, which extract the maximum amount of information from time-resolved spectra.

-15-

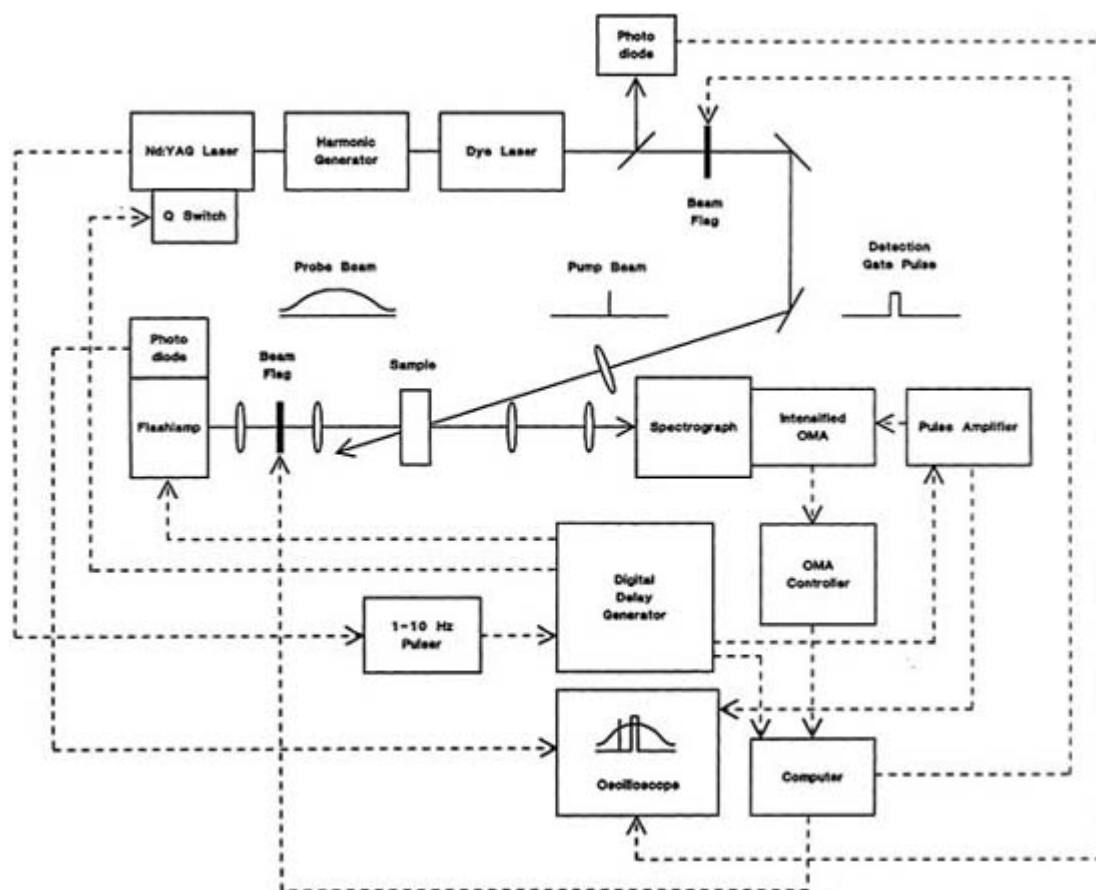


Figure C3.1.6. Block diagram for nanosecond absorption apparatus using multichannel detection. (From Goldbeck R A and Kliger D S 1993 *Methods Enzymol.* **226** 147–77.)

-16-

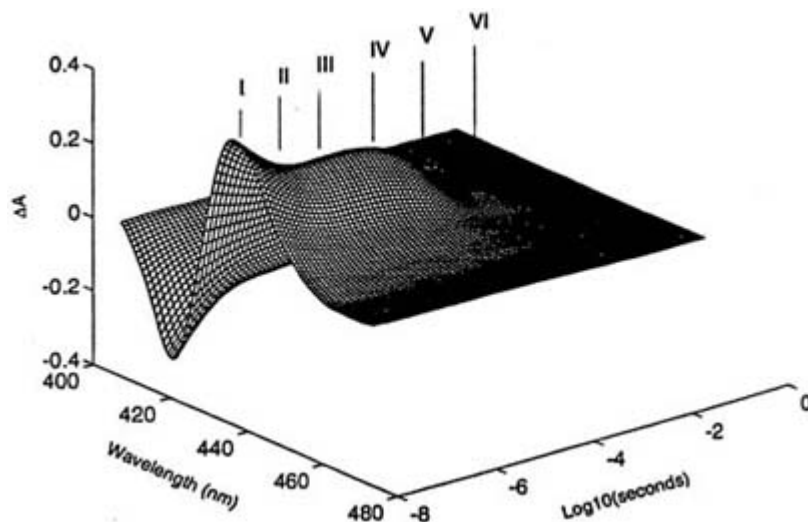


Figure C3.1.7. Time-resolved optical absorption data for the Soret band of photolysed haemoglobin–CO showing six first-order (or pseudo-first-order) relaxation phases, I–VI, on a logarithmic time scale extending from nanoseconds to seconds. Relaxations correspond to geminate and diffusive CO rebinding and to intramolecular relaxations of tertiary and quaternary protein structure. (From Goldbeck R A, Paquette S J, Björling S C and Kliger D S 1996 *Biochemistry* **35** 8628–39.)

C3.1.7.2 FLUORESCENCE

Fluorescence, the spontaneous emission of light in a spin-allowed transition from an excited state to a lower energy or ground state, can provide a very sensitive means for detecting the concentrations of atomic and molecular species in transient kinetic studies. Being a null measurement, the sensitivity can be extraordinarily high. Indeed, it is possible to detect the presence of a single molecule in solution through its laser-induced fluorescence (LIF) [22]. Sensitivity levels more typical of kinetic studies are of the order of 10^{10} molecules cm^{-3} . A schematic diagram of an apparatus for kinetic LIF measurements is shown in [figure C3.1.8](#). A limitation of this approach is that only relative concentrations are easily measured, in contrast to absorption measurements, which yield absolute concentrations. Another important limitation is that not all molecules have measurable fluorescence, as radiationless transitions can be the dominant decay route for electronic excitation in polyatomic molecules. However, the latter situation can also be an advantage in complex molecules, such as proteins, where a lack of background fluorescence allows the selective introduction of fluorescent chromophores as probes for kinetic studies. (Tryptophan is the only strongly fluorescent amino acid naturally present in proteins, for instance.)

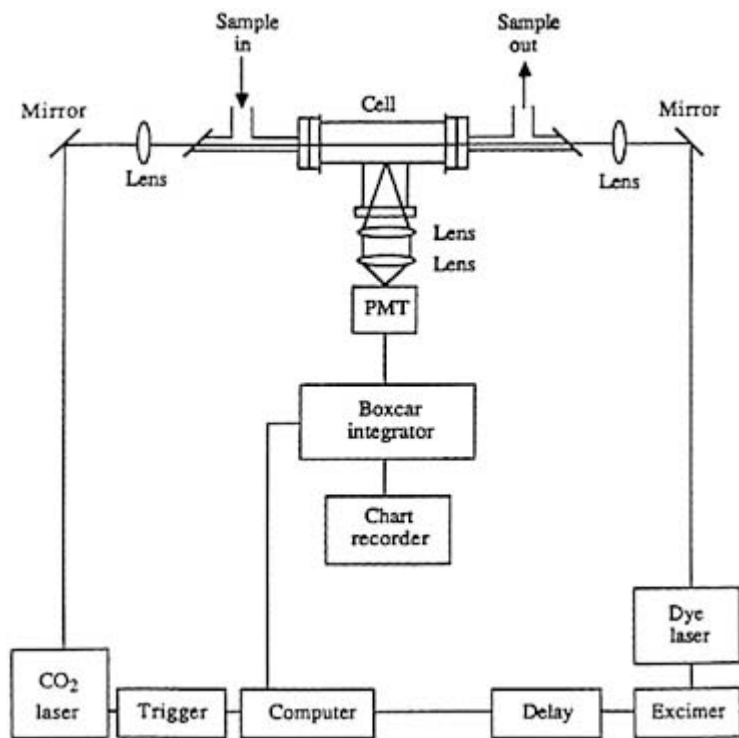


Figure C3.1.8. Schematic diagram of a transient kinetic apparatus using laser-induced fluorescence (LIF) as a probe and a CO₂ laser as a pump source. (From Steinfeld J I, Francisco J S and Hase W L 1989 *Chemical Kinetics and Dynamics* (Englewood Cliffs, NJ: Prentice-Hall).)

Determining a molecule's fluorescence lifetime (τ_F), typically of the order of nanoseconds for strong emitters, is frequently an object of transient kinetic study. A transient measurement of lifetime after excitation with a nanosecond flash lamp or pulsed laser (inexpensive subnanosecond pulsed nitrogen lasers are available for this purpose) can be accomplished by directly monitoring the time course of fluorescence intensity using a fast photomultiplier and transient recorder or boxcar integrator or, less directly, by measuring the statistical distribution of times between absorption and emission events (under low intensity illumination conditions) using a time-correlated single-photon counting apparatus [23, 24]. In general, the fluorescence lifetime can be shorter than the radiative lifetime (τ_R), given by the Einstein B coefficient for the emissive transition (usually estimated from the corresponding absorption band, as the same upper and lower states are usually connected by absorption and emission), because radiationless transitions and intermolecular excited state quenching can compete kinetically with emission in depopulating the excited state. A measurement of τ_F can thus be used to determine the total rate (k_{NR}) of nonradiative relaxation processes (internal conversion, intersystem crossing) and bimolecular quenching (k_Q): $k_{NR} + k_Q[Q] = 1/\tau_F - 1/\tau_R$, where $[Q]$ is the concentration of the quenching molecule.

Nonradiative relaxation and quenching processes will also affect the quantum yield of fluorescence, $\phi_F = k_R / (k_R + k_{NR} + k_Q[Q])$. Relative measurements of fluorescence quantum yield at different quencher concentrations are easily made in steady state measurements; absolute measurements (to determine k_{NR}) are most easily obtained by comparisons of steady state fluorescence intensity with a fluorescence standard. The usefulness of this situation for transient studies

of large molecules, such as biopolymers, lies in the ability to use the quasi-steady state fluorescence intensity of an embedded chromophore as a probe of dynamic changes in the solvent exposure of the chromophore (through changes in k_{NR}) or in long distance chromophore–chromophore quenching interactions (Förster energy transfer) over the 10^{-8} to 1s time regime. The rate of Förster (very weak coupling excitonic) energy transfer varies inversely as the sixth power of the chromophore to chromophore distance, making the observed fluorescence intensity a potentially sensitive ruler of intramolecular distances in static and time-resolved studies. Thus, for example, time-resolved measurements of tryptophan near-UV fluorescence intensity under steady state illumination can be used to

monitor conformational relaxations in aqueous proteins after rapid mixing with a denaturant that disrupts secondary and tertiary structure. As a protein's native structure unfolds, increased distance from a quenching group, such as the haem prosthetic group in haem proteins, can dramatically increase the fluorescence yield. In proteins lacking an internal quencher, increased exposure of tryptophan to the solvent enhances k_{NR} and reduces the fluorescence intensity.

C3.1.7.3 INFRARED ABSORPTION

Time-resolved spectroscopy in the IR region (TRIR) can give detailed information about structural changes in molecules by monitoring changes in the frequencies and absorption amplitudes of vibrational normal modes. The selection rule for vibrational transitions allows for those modes whose motions produce a change in electric dipole moment to be IR active. Changes in bond strengths or molecular symmetry accompanying intramolecular processes such as isomerizations, and the binding or dissociation of ligands in complexes, are examples of transient events that may be studied with TRIR. An apparatus for TRIR is shown schematically in [figure C3.1.9](#) [figure C3.1.10](#) shows the results of a TRIR study of the transient binding of CO to a copper atom in cytochrome *c* oxidase after photodissociation of the Fe–CO bond in this haem protein [25]. Evolution in the secondary structure of proteins may also be followed by measuring the TRIR of several vibrational bands arising from the peptide backbone: the N–H stretching vibration (3300 cm^{-1}), the C=O stretch, or amide I band ($1600\text{--}1700\text{ cm}^{-1}$) and the N–H bending vibration, or amide II band ($1520\text{--}1550\text{ cm}^{-1}$). TRIR of these bands has been used to directly monitor fast folding and unfolding reactions in the protein RNase A [26] and in small peptides [27], for example.

-19-

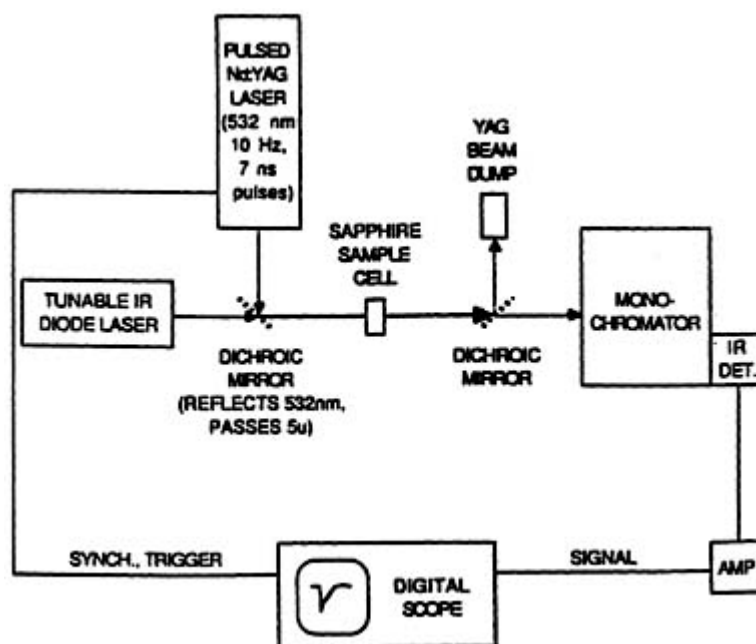


Figure C3.1.9. Block diagram for time-resolved infrared spectroscopy apparatus. (From Dyer R B, Einarsson Ó, Killough P M, López-Garriga J J and Woodruff W H 1989 *J. Am. Chem. Soc.* **111** 7657–9.)

-20-

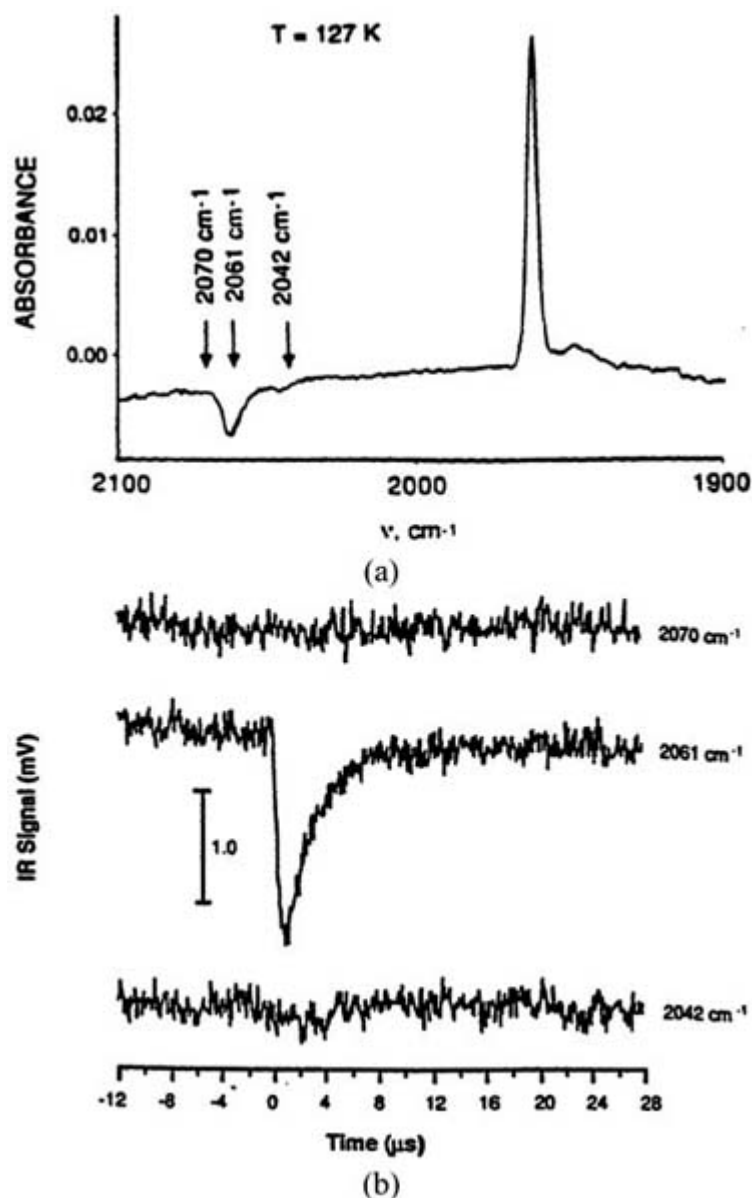


Figure C3.1.10. (a) Steady state IR difference spectrum (dark minus light) of cytochrome *c* oxidase CO complex measured at low temperature (127 K). This protein contains a copper atom situated immediately adjacent to a haem iron, the latter binding CO with high affinity at equilibrium. The band at 2061 cm^{-1} shows the presence of a Cu–CO bond in the intermediate species frozen immediately after photolysis (light) of the equilibrium, Fe–CO bonded protein (dark). (b) The time evolution of the Cu–CO bonded intermediate after room temperature photolysis of the Fe–CO protein complex. The CO is transferred to the Cu atom within femtoseconds of photodissociation of the Fe–CO bond. The Cu–CO bond then thermally dissociates with a time constant of about $1.5\text{ }\mu\text{s}$ and freed CO diffuses out of the protein. (The two frequencies outside the Cu–CO band provide control measurements.) (From Dyer R B, Einarsdóttir Ó, Killough P M, López-Garriga J J and Woodruff W H 1989 *J. Am. Chem. Soc.* **111** 7657–9.)

The absorption intensities characteristic of IR transitions, the intensities of IR light sources and the sensitivities of IR detectors are relatively low compared with those for visible and UV wavelengths. These factors present a challenge for experimentalists wishing to accumulate data with signal to noise ratios sufficient to resolve time-dependent changes. An additional factor is the wide presence of IR active transitions in polyatomic molecules. While on the one hand this constitutes one of the primary advantages of the TRIR technique, it means on the other hand that transient IR signals are often detected against interfering background absorptions from the solvent or from peripheral modes in large molecules. For example, water, an important solvent in TRIR studies of biological

molecules, is also a strong IR absorber. Differential TRIR techniques are often used to overcome interference from such background absorptions.

Instead of the blackbody-radiation light sources commonly used in dispersive IR spectrometers, time-resolved studies often use pulsed xenon flash lamps or tunable CW diode lasers, which concentrate IR output intensity in time or frequency space, respectively. TRIR has typically been measured in single-wavelength, kinetic mode, as IR sensitive OMAs are not yet widely available. Kinetic IR signals are collected with photoconductive detectors using materials such as indium antimonide (InSb) or mercury–cadmium–tellurium (HgCdTe). However, complete nanosecond TRIR spectra have been recorded using a dispersive scanning spectrometer with an HgCdTe detector [28], although the spectral accumulation times tend to be long. Another approach to spectral mode TRIR has been to combine Fourier transform techniques (FTIR) with time-resolved spectroscopy [29].

C3.1.7.4 RESONANCE RAMAN SCATTERING

The Raman effect, the inelastic scattering of photons resulting in frequency shifts that reflect the vibrational and rotational energies of the scattering molecule, is used in transient spectroscopy to obtain time-dependent vibrational information that often complements that obtained from TRIR. Raman scattering arises from changes in the polarizability of a molecule associated with vibrational (and rotational) motions, rather than from changes in the dipole moment itself, as is the case in IR absorption spectroscopy. Raman selection rules therefore can in general be different from those for absorption. The two methods thus complement one another, particularly in small molecules. This is most true for molecules with a centre of symmetry. In this case, the selection rules are exactly complementary—all Raman active transitions are IR inactive, and vice versa. Raman spectroscopy also offers an advantage for transient studies of solutes in solvents such as water that are not strongly Raman active, e.g. transient studies of aqueous biomolecules.

Bringing the frequency of the photon into near resonance with an electronic transition enhances the intensity of the inherently weak Raman scattering process. The ratio of Raman to Rayleigh scattering intensity, typically 10^{-9} to 10^{-6} , is increased in resonance Raman by one to two orders of magnitude to give ratios of 10^{-8} to 10^{-4} . Time-resolved resonance Raman spectroscopy (TR³) thus offers greater S/N ratios and higher time resolution for transient studies. It also offers greater specificity in the time-resolved vibrational spectroscopy of large molecules, as compared with TRIR spectroscopy, in that only the vibrational modes associated with the nuclear structure of the resonant chromophore are enhanced. This is a particularly important advantage in very large molecules, such as biopolymers, where many overlapping IR and Raman active modes may be present [30, 31 and 32].

Raman spectroscopy requires an intense, monochromatic light source. The field thus developed rapidly when lasers became commercially available in the 1960s. Nanosecond TR³ measurements are now performed with several configurations using pulsed or CW lasers. A pump–probe two-pulse method giving both kinetic and spectral information about dynamic processes is frequently used. Kinetic information is obtained in this method by varying the delay time between pump and probe, as provided for by a digital delay generator in the apparatus shown in

figure C3.1.11. The ability to choose different pump and probe wavelengths in this method reduces the likelihood of spectral artifacts. For nanosecond TR³ measurements, the pump source is typically a low repetition rate, high pulse-energy laser, often a Nd:YAG or excimer laser with a pulse duration of several nanoseconds. The probe light source in a Raman measurement must also be intense—intense enough to generate detectable Raman scattering, but not so intense as to photoinitiate changes in the sample (recall that the probe is in near resonance with an electronic transition of a photoreactive molecule). The same laser types used as pumps are also used as probe sources, although they are generally coupled with a dye laser, hydrogen, deuterium or methane Raman shifter or, more recently, a nonlinear harmonic generation crystal or an optical parametric oscillator (OPO) to increase the selection of available wavelengths.

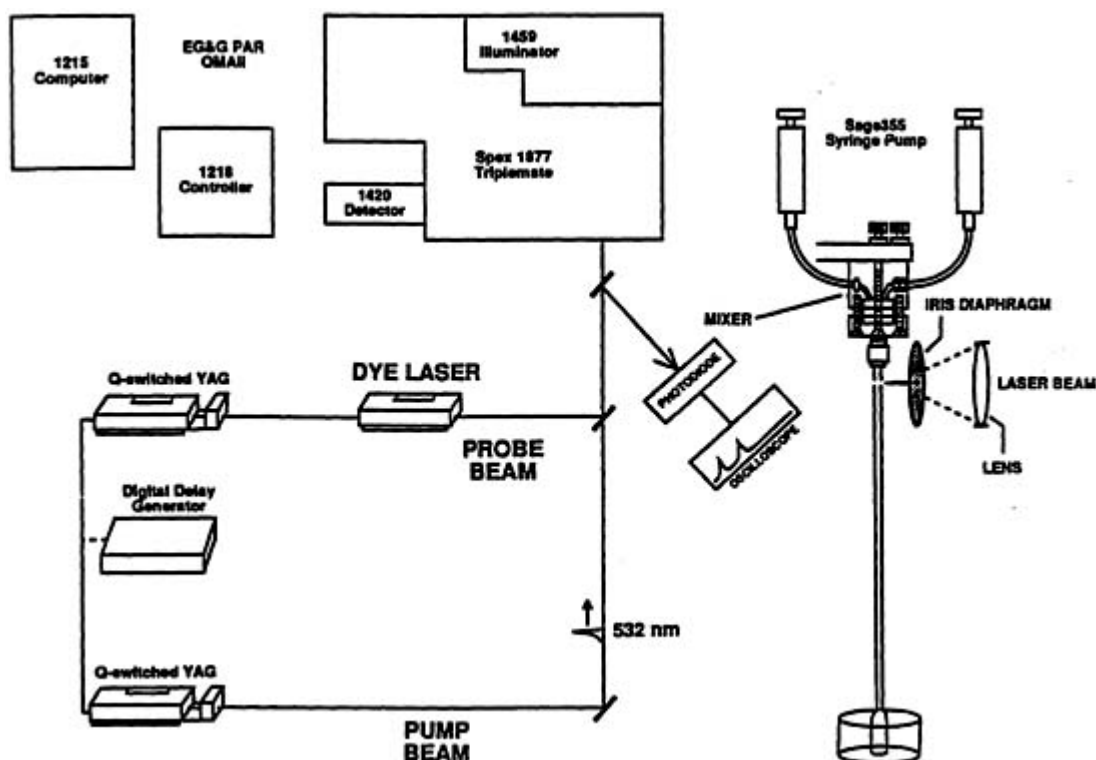


Figure C3.1.11. Apparatus for pump-probe time-resolved resonance Raman spectroscopy. (From Varotsis C and Babcock G T 1993 *Methods Enzymol.* **226** 409–31.)

Discriminating the small amount of light intensity in the Raman-shifted lines from the intense Rayleigh-scattered light of the probe laser places great demands on monochromator design, which must balance high wavelength resolution and stray-light rejection against signal throughput. Single, double and triple monochromators are used in TR³ measurements for correspondingly increased rejection of Rayleigh scattering. The addition of each grating typically introduces a throughput efficiency factor of roughly 30%, however, so that the overall efficiency of a triple monochromator is only about 3%. Single 0.75–1 m monochromators have output efficiencies ideal for TR³ experiments and may be used to measure Raman frequencies greater than 500 cm⁻¹. Notch filters combined with a single monochromator give Rayleigh rejection factors comparable to a double monochromator while maintaining greater throughput and are suitable for TR³ measurements at frequencies as low as 50 × 100 cm⁻¹ [30].

-23-

TR³ signals are preferably detected with a time-gated optical multichannel analyser (OMA), which allows the entire spectral change associated with each pump-probe pair to be recorded for a given time delay. Intensified diode array detectors, with optical quantum efficiencies of about 20%, can be gated with a high voltage pulse to capture the transient Raman signal of interest while suppressing interference from Rayleigh scatter and fluorescence falling outside the time window of the gate pulse. CCD (charge-coupled device) detectors are 2D arrays that have more recently become available for TR³ spectroscopy. They offer the advantages of low readout noise and higher quantum efficiencies in the IR region. Although also more sensitive to artifacts from cosmic rays, the narrow spikes these produce can be removed from affected data by using commercially available software.

C3.1.7.5 CIRCULAR DICHROISM AND OPTICAL ROTATORY DISPERSION

Circular dichroism (CD), the differential absorption of left *versus* right circularly polarized light, is the polarization spectroscopy perhaps best suited to detecting the presence of asymmetry in the structure or environment of molecular chromophores. Various time-resolved CD (TRCD) methods have been developed to take advantage of this sensitivity and obtain more detailed structural information about kinetic processes than is found from ordinary time-resolved absorption measurements [33]. Some examples of the processes studied with TRCD methods are: the effects of electronic excitation on the structure of chiral inorganic complexes, the changes in α -helical secondary

structure accompanying the folding reactions of proteins and the time evolutions of tertiary and quaternary structure in allosteric proteins, as reflected in the protein-environment-induced CDs of ‘reporter’ chromophores.

CD is a small effect. $\Delta\varepsilon/\varepsilon$, the ratio of the difference in circularly polarized extinction coefficients, $\Delta\varepsilon = \varepsilon_L - \varepsilon_R$, to total absorption, $\varepsilon = \frac{1}{2}(\varepsilon_L + \varepsilon_R)$, is typically only about 10^{-4} – 10^{-3} . Being so small, the measurement of CD with signal to noise ratios sufficient for static and time resolved studies requires special methods, each representing a different tradeoff between the factors such as time resolution, sensitivity to artifacts and experimental simplicity, that determine the method of choice for a particular kinetic study.

Kinetic CD measurements on slow time scales may be made on commercial CD instruments, which can be equipped for stopped-flow studies. Commercial CD instruments use rapid polarization modulation methods, introduced in the 1960s, and phase-locked detection to increase sensitivity. Linearly polarized light is passed through a photoelastic modulator (PEM), essentially a small quartz plate undergoing resonantly driven acoustic vibration, to produce time-varying elliptically polarized light cycling between left and right circular polarizations. The CD signal is detected as the AC component of the light intensity transmitted through the sample, normalized to the magnitude of the DC component and to a calibration factor (determined by measuring the CD of a standard substance) reflecting the relative gain of the AC and DC electronic amplification stages. Noise from instrumental sources, such as arc wander, containing frequency components lower than the modulator frequency is effectively filtered from the AC-modulated CD signal. However, the PEM resonant frequency, typically 1–100 kHz, not only sets an upper limit on the frequencies of the instrumental noise components suppressed, it also limits the maximum characteristic frequencies of the kinetic processes that may be studied. TRCD measurements on time scales faster than about a millisecond thus require unconventional methods.

An ellipsometric approach to CD measurements used for TRCD spectroscopy in the nanosecond regime is depicted schematically in [figure C3.1.12](#) [34]. The instrument used in this technique is based on a nanosecond Nd:YAG laser photolysis apparatus using a broad band microsecond xenon flash lamp as a probe. Rather than detecting the differential absorption of left and right circularly polarized light, the polarization state of elliptically polarized light is detected in this method, i.e., ellipsometry. The linearly polarized probe beam is passed through a strain plate, a fused

silica plate under slight mechanical compression, to produce highly eccentric elliptically polarized light. The CD signal is detected as the difference between right and left elliptically polarized light intensity transmitted through the sample and an analysing polarizer, normalized to the sum of the intensities and a proportionality factor determined by the pathlength and concentration of the sample and the magnitude of the strain plate retardance, δ . (No calibration against a CD standard is required in this method.) This can be written as $\Delta\varepsilon = (\delta/cI)/(I_{REP} - I_{LEP})/(I_{REP} + I_{LEP})$. The primary advantage of the near-null approach is that the signal is effectively amplified relative to the noise from instrumental sources by a factor inversely proportional to δ . The tradeoff for this increased sensitivity and time resolution is that more care must be taken to avoid potential interference from light scattering, fluorescence and optical artifacts—particularly those from linear birefringence present in the sample or optics—than is necessary for conventional PEM-based measurements. An instrument for nanosecond far-UV TRCD using this method is shown schematically in [figure C3.1.13](#).

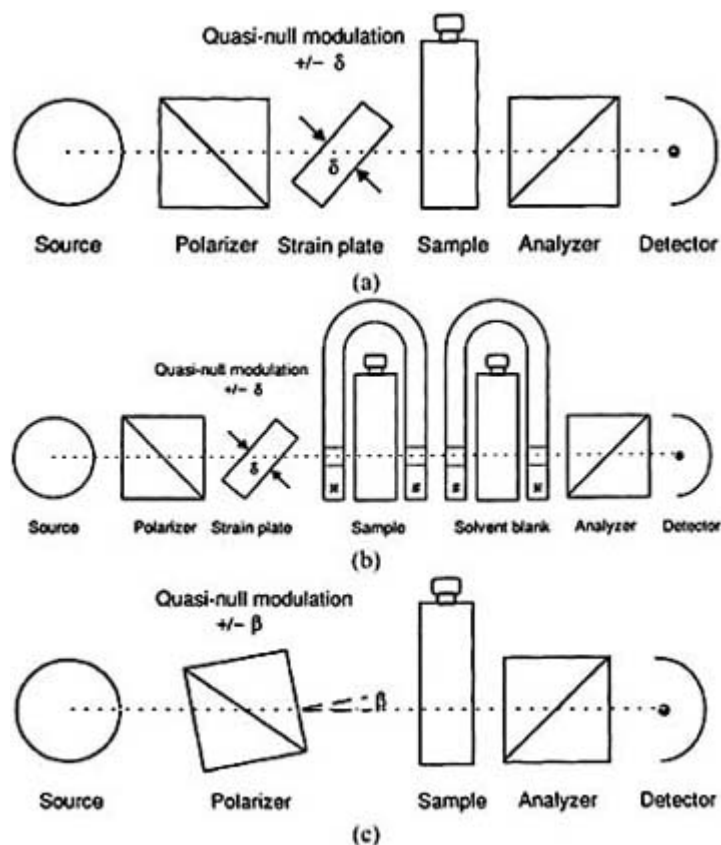


Figure C3.1.12. Schematic diagrams of optical configurations for quasi-null detection techniques used in transient kinetic studies of (a) CD, (b) MCD and (c) ORD/LD. Elliptical polarization is provided in (a) and (b) by a horizontal prism polarizer followed by a fused silica plate of linear retardance $\pm\delta$, magnitude $\sim 1^\circ$, induced by mechanical compression indicated by arrows. CD of the sample adds to or subtracts from the net ellipticity of the beam detected by the vertical analysing polarizer, giving rise to the differential signal shown in the text. A solvent blank in an opposed applied field cancels the Faraday rotation of solvent and cell in (b). The polarizer axis is rotated to $\pm\beta$ from horizontal in (c), where $\beta \sim 1^\circ$. ORD or LD of the sample adds to or subtracts from the net rotation of the beam detected by the analysing polarizer, giving rise to a differential signal. (From Chen E, Goldbeck R A and Klinger D S 1997 *Annu. Rev. Biophys. Biomol. Struct.* **26** 325–53.)

-25-

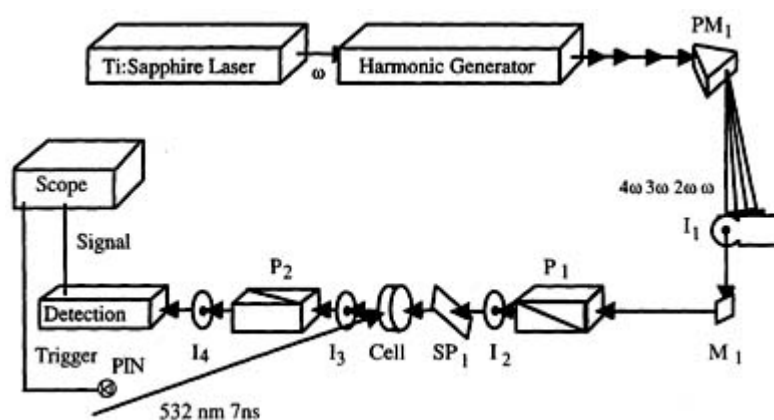


Figure C3.1.13. Experimental configuration for far-UV nanosecond CD measurements using a frequency-upconverted Ti:sapphire laser as a probe source. P_1 and P_2 are MgF_2 Rochon polarizers at cross orientations. SP_1 is a strained transparent plate with about 1° of linear birefringence for quasi-null ellipsometric CD detection. Prism PM_1 and the iris I_7 select the far-UV fourth harmonic of the argon laser-pumped Ti-sapphire laser's near-IR fundamental output to probe the ellipticity of the sample. A second laser beam at 532 nm is used to pump CD

transients in the sample. (From Goldbeck R A, Kim-Shapiro D B and Kliger D S 1997 *Annu. Rev. Phys. Chem.* **48** 453–79.)

Ultrafast TRCD has also been measured in chemical systems by incorporating a PEM into the probe beam optics of a picosecond laser pump–probe absorption apparatus [35]. The PEM resonant frequency is very low (1 kHz) in these experiments, compared with the characteristic frequencies of ultrafast processes and so does not interfere with the detection of ultrafast CD changes.

In principle, optical rotatory dispersion (ORD) and circular dichroism contain identical information about molecular structure and can be interconverted using the Kramers–Kronig integral transforms [36]. CD, being limited to absorption bands and easier to interpret than ORD, became the preferred approach to the study of optically active molecules after the development of PEM-based CD spectrometers. From an experimental point of view, however, optical rotation (OR), the rotation of the polarization plane of light by a chiral substance, is easier to measure than circular dichroism. This fact accounts for the historical importance of OR (the beginnings of chemical kinetics as a quantitative discipline, for instance, can be traced to the use of OR measurements to determine reaction rates for sucrose hydrolysis by Wilhemy in 1850 [37]) and for the recent development of rapid time-resolved ORD methods for kinetic studies. A near-null polarimetric ORD method has been incorporated into several generations of flash photolysis instruments developed over the past few decades for time-resolved applications extending into the nanosecond time regime [33]. (This method also doubles as a very sensitive technique for measuring linear dichroism in anisotropic samples and is useful for studies of orientational relaxation after laser photoselection [38].)

C3.1.7.6 MAGNETIC CIRCULAR DICHROISM

Magnetic circular dichroism (MCD) is independent of, and thus complementary to, the natural CD associated with chirality of nuclear structure or solvation. Closely related to the Zeeman effect, MCD is most often associated with orbital and spin degeneracies in chromophores. Chemical applications are thus typically found in systems where a chromophore of high symmetry is present: metal complexes, porphyrins and other aromatics, and haem proteins are

-26-

prominent examples. Time-resolved MCD (TRMCD) spectroscopy is best suited to the study of kinetic processes directly affecting chromophore electronic structure, e.g., spin, oxidation and ligation state changes in metal complexes and metalloporphyrins. TRMCD is measured by adding a magnet—permanent, electric or superconducting—to the TRCD instrumentation described above. This is straightforward to do for PEM-based instruments, such as the ultrafast MCD apparatus of Xie and Simon [39]. Ellipsometric TRMCD measurements, on the other hand, require an additional optical component to compensate for rotation of the probe beam's polarization orientation by the Faraday effect of the transparent solvent and cell windows in the magnetic field [33].

C3.1.8 ANALYSIS OF TIME-RESOLVED SPECTRAL DATA

Transient kinetic studies measure a time-resolved record of some property of the sample, such as absorption, emission or conductance, that can be analysed for its kinetic components. The data are usually stored as a digital computer file containing a linear array of observations against time. In the case of spectroscopic measurements, this may be generalized to a rectangular array of spectra against time. The particular form of the analysis that is applied to the data in order to obtain rate constants and amplitudes is determined by knowledge about, or assumption of, a particular kinetic mechanism. Determining an unknown mechanism is often an iterative process in which possible models are tested until the most parsimonious mechanism consistent with the data is found. This mechanism can frequently be assumed to involve only first order or pseudo-first order rate processes in fast kinetic studies. In this case, the analysis of single-wavelength absorption data may be as simple as a linear regression plot of \ln (absorption) against time, the slope of which provides the rate constant for a simple exponential decay. More complicated analysis, i.e., nonlinear least squares multi-exponential fitting of absorption against time, is required if

more than one process is present.

Multichannel time-resolved spectral data are best analysed in a global fashion using nonlinear least squares algorithms, e.g., a simplex search, to fit multiple first order processes to all wavelength data simultaneously. The goal in this case is to find the time-dependent spectral contributions of all reactant, intermediate and final product species present. In matrix form this is $\mathbf{A}(\lambda, t) = \mathbf{BC}$, where \mathbf{A} is the data matrix, rows indexed by wavelength and columns by time, \mathbf{B} contains spectra as columns and \mathbf{C} contains time-dependent concentrations of all species arranged in rows.

A general first order mechanism can be written symbolically in matrix form as

$$\frac{dc(t)}{dt} = \mathbf{K}c(t)$$

where \mathbf{K} is the matrix of rate constants and $c(t)$ is a column vector of time-dependent concentrations. A general solution for the concentrations (found using eigenvalue techniques, for instance) is

$$c_i(t) = \sum_j M_{ij} [\mathbf{M}^{-1} c(t=0)]_j e^{[\mathbf{M}^{-1} \mathbf{K} \mathbf{M}]_{jj} t}$$

where \mathbf{K} is diagonalized by a similarity transform with matrix \mathbf{M} . An efficient approach to fitting the kinetic mechanism represented by the elements of \mathbf{K} to the data in \mathbf{A} is to first apply singular value decomposition (SVD) to

-27-

\mathbf{A} , $\mathbf{A} = \mathbf{USV}^T$, where \mathbf{S} is a diagonal matrix of singular values, and \mathbf{U} and \mathbf{V} are orthonormal matrices containing in this case spectral and temporal information, respectively (superscript 'T' indicates matrix transpose) [40]. (SVD is very similar to principal component analysis [41, 42].) The data are then filtered of random noise by discarding singular values that fall below a value determined by the magnitude of the noise, leaving the truncated matrix \mathbf{S}_r . (The corresponding columns of \mathbf{U} and \mathbf{V} are also discarded to obtain \mathbf{U}_r and \mathbf{V}_r .) The singular values remaining correspond to those columns of \mathbf{U} and \mathbf{V} containing spectrokinetic information filtered of noise. Their number, r , is the effective rank of \mathbf{A} . Besides providing a convenient deconvolution of the data into spectral and temporal components and filtering these into a more compact representation, SVD also provides physical information in that r gives a lower bound on the number of independent spectrotemporal components, or kinetic species, present in the system. Although the individual spectra in \mathbf{U}_r and evolutions in \mathbf{V}_r do not generally correspond to those of physical species, observed rate constants can be efficiently fitted to the aggregate temporal information in \mathbf{V}_r . These constants correspond to the diagonal elements of $\mathbf{M}^{-1} \mathbf{K} \mathbf{M}$, which we will call \mathbf{K}_{obs} . In the simplest case, that of simple decays proceeding in parallel, $\mathbf{K} = \mathbf{K}_{\text{obs}}$, i.e., \mathbf{M} is the identity matrix, \mathbf{I} . The elements of \mathbf{C} are calculated from the solution for $c_i(t)$ using \mathbf{K}_{obs} and known, or trial, values of $c_i(t=0)$. The spectra of intermediates are then calculated from $\mathbf{B} = \mathbf{U}_r \mathbf{S}_r \mathbf{V}_r^T \text{pinv}(\mathbf{C})$, where pinv is the matrix pseudoinverse (essentially a least squares solution to the overdetermined inversion problem). Finally, any assumptions about initial concentrations can be checked by comparing \mathbf{B} against known model spectra. For more general first order kinetic mechanisms, \mathbf{M} is calculated from $\mathbf{M}^{-1} \mathbf{K} \mathbf{M} = \mathbf{K}_{\text{obs}}$ and knowledge of, or assumptions about, branching ratios or equilibrium constants for any back reactions that may be present. Such assumptions, if necessary, can be tested by comparing the calculated spectra in \mathbf{B} to known model spectra. (It may be expedient to obtain the observed rate constants, without determining a mechanism, by simply choosing all $c_i(t=0) = 1$ and setting $\mathbf{M} = \mathbf{I}$, in which case the spectra in \mathbf{B} are sometimes referred to as 'b-spectra' in the literature.) The derivation of first order mechanisms from transient spectral data is discussed in more detail in [43].

The Arrhenius relation given above for the temperature dependence of an elementary reaction rate is used to find the activation energy, E_a , and the pre-exponential factor, A , from the slope and intercept, respectively, of a (linear) plot of $\ln(k(T))$ against T^{-1} . The standard enthalpy and entropy changes of the transition state (at constant

temperature and pressure) can be found for most types of reaction from

$$\Delta H_0^\ddagger = E_a - RT$$

and

$$\Delta S_0^\ddagger = R \ln(Ah/k_B T) - R.$$

These expressions are modified in the case of non-unimolecular gas phase reactions to

$$\Delta H_0^\ddagger = E_a - RT(1 - \Delta n^\ddagger)$$

-28-

and

$$\Delta S_0^\ddagger = R \ln(Ah/k_B T) - R(1 - \Delta n^\ddagger)$$

where $\Delta n^\ddagger = P\Delta V^\ddagger/RT$, ΔV^\ddagger being the standard volume of activation. The free energy of activation is found from $\Delta G_0^\ddagger = \Delta H_0^\ddagger - T\Delta S_0^\ddagger$.

The presence of nonlinearity in an Arrhenius plot may indicate the presence of quantum mechanical tunnelling at low temperatures, a compound reaction mechanism (i.e., the reaction is not actually elementary) or the ‘unfreezing’ of vibrational degrees of freedom at high temperatures, to mention some possible sources.

REFERENCES

- [1] Hartridge H and Roughton F J W 1923 A method of measuring the velocity of very rapid chemical reactions *Proc. R. Soc. A* **104** 376–94
- [2] Norrish R G W and Porter G 1949 Chemical reactions produced by very high light intensities *Nature* **164** 658
- [3] Eigen M 1954 Methods for investigation of ionic reactions in aqueous solutions with half-times as short as 10^{-9} sec *Discuss. Faraday Soc.* **17** 194–205
- [4] Stehlik D and Möbius K 1997 New EPR methods for investigating photoprocesses with paramagnetic intermediates *Annu. Rev. Phys. Chem.* **48** 745–84
- [5] Liu X, Siegel D L, Fan P, Brodsky B and Baum J 1996 Direct NMR measurement of the folding kinetics of a trimeric peptide *Biochemistry* **35** 4306–13
- [6] Brückner V, Feller K-H and Grummt U-W 1990 *Applications of Time-Resolved Optical Spectroscopy* (New York: Elsevier)
- [7] Takahashi S, Yeh S R, Das T K, Chan C K, Gottfried D S and Rousseau D L 1997 Folding of cytochrome c initiated by submillisecond mixing *Nature Struct. Biol.* **4** 44–50.
- [8] Strehlow H and Becker M 1959 The pressure-jump method for the measurement of rates of ionic reactions *Z. Elektrochem.* **63** 457–61

- [9] Knoche W and Wiese G 1976 Pressure-jump relaxation techniques with optical detection *Rev. Sci. Instrum.* **47** 220–1
- [10] Quednau J and Schneider G M 1989 A new high-pressure cell for differential pressure-jump experiments using optical detection *Rev. Sci. Instrum.* **60** 3685–7
- [11] Eigen M and DeMaeyer L 1955 Kinetics of neutralization *Z. Elektrochem.* **59** 986–93
- [12] Porschke D and Obst A 1991 An electric field jump apparatus with ns time resolution for electro-optical measurements at physiological salt concentrations *Rev. Sci. Instrum.* **62** 818–20
-

-29-

- [13] Marriott G (ed) 1998 *Caged Compounds (Methods in Enzymology 291)* (New York: Academic)
- [14] Ireland J F and Wyatt P A H 1976 Acid–base properties of electronically excited states of organic molecules *Adv. Phys. Org. Chem.* **12** 131–221
- [15] Gutman M, Huppert D and Pines E 1981 The pH jump: a rapid modulation of pH of aqueous solutions by a laser pulse *J. Am. Chem. Soc.* **103** 3709–13
- [16] Gutman M 1986 Application of the laser-induced proton pulse for measuring the protonation rate constants of specific sites on proteins and membranes *Methods Enzymol.* **127** 522–38
- [17] Viappiani C, Bonetti G, Carcelli M, Ferrari F and Sternieri A 1998 Study of proton transfer processes in solution using the laser induced pH-jump: a new experimental setup and an improved data analysis based on genetic algorithms *Rev. Sci. Instrum.* **69** 270–6
- [18] Hodgson B W and Keene J P 1972 Some characteristics of a pulsed xenon lamp for use as a light source in kinetic spectrophotometry *Rev. Sci. Instrum.* **43** 493–6
- [19] Hofrichter J, Ansari A, Jones C M, Deutsch R M, Sommer J H and Henry E R 1994 Ligand binding and conformational changes measured by time-resolved absorption spectroscopy *Methods Enzymol.* **232** 387–415
- [20] Lewis J W, Yee G G and Kliger D S 1987 Implementation of an optical multichannel analyzer controller for nanosecond flash photolysis measurements *Rev. Sci. Instrum.* **58** 939–44
- [21] Lewis J W, Warner J, Einterz C M and Kliger D S 1987 Noise reduction in laser photolysis studies of photolabile samples using an optical multichannel analyzer *Rev. Sci. Instrum.* **58** 945–9
- [22] Goodwin P M, Ambrose W P and Keller R A 1996 Single-molecule detection in liquids by laser-induced fluorescence *Accounts Chem. Res.* **29** 607–13
- [23] Cline-Love L J and Shaver L A 1976 Time correlated single photon technique: fluorescence lifetime measurements *Anal. Chem.* **48** 370A–371A
- [24] Birch D J S and Imhof R E 1977 A single-photon counting fluorescence decay-time spectrometer *J. Phys. E: Sci. Instrum.* **10** 1044–9
- [25] Dyer R B, Einarsdóttir Ó, Killough P M, López-Garriga J J and Woodruff W H 1989 Transient binding of photodissociated CO to Cu_2^+ of eukaryotic cytochrome oxidase at ambient temperature. Direct evidence from time-resolved infrared spectroscopy *J. Am. Chem. Soc.* **111** 7657–9
- [26] Philips C M, Mizutani Y and Hochstrasser R M 1995 Ultrafast thermally induced unfolding of RNase A *Proc. Natl Acad. Sci. USA* **92** 7292–6
- [27] Williams S, Causgrove T P, Gilmanshin R, Fang K S, Callender R H, Woodruff W H and Dyer R B 1996 Fast events in protein folding: helix melting and formation in a small peptide *Biochemistry* **35** 691–7
- [28] Yuzawa T, Kato C, George M W and Hamaguchi H O 1994 Nanosecond time-resolved infrared spectroscopy with a dispersive scanning spectrometer *Appl. Spectrosc.* **48** 684–90
- [29] Siebert F Infrared spectroscopy applied to biochemical and biological problems *Methods. Enzymol.* **246**

- [30] Varotsis C and Babcock G T 1993 Nanosecond time-resolved resonance Raman spectroscopy *Methods Enzymol.* **226** 409–31
-

-30-

- [31] Friedman J M 1994 Time-resolved resonance Raman spectroscopy as probe of structure, dynamics, and reactivity in hemoglobin *Methods Enzymol.* **232** 205–31
- [32] Kincaid J R 1995 Structure and dynamics of transient species using time-resolved resonance Raman spectroscopy *Methods. Enzymol.* **246** 460–501
- [33] Goldbeck R A, Kim-Shapiro D B and Kliger D S 1997 Fast natural and magnetic circular dichroism *Annu. Rev. Phys. Chem.* **48** 453–79
- [34] Lewis J W, Tilton R F, Einterz C M, Milder S J, Kuntz I D and Kliger D S 1985 New technique for measuring circular dichroism changes on a nanosecond time scale. Application to (carbonmonoxy)myoglobin and (carbonmonoxy)hemoglobin *J. Phys. Chem.* **89** 289–94
- [35] Xie X and Simon J D 1989 Picosecond time-resolved circular dichroism spectroscopy: experimental details and applications *Rev. Sci. Instrum.* **60** 2614–27
- [36] Moscovitz A 1962 Theoretical aspects of optical activity *Adv. Chem. Phys.* **4** 67–112
- [37] Wilhemy L F 1850 Über das Gesetz, nach welchem die Einwirkung der Säuren auf den Rohrzucker stattfindet *Ann. Phys. Chem.* **81** 413–33, 499–526
- [38] Che D P, Shapiro D B, Esquerra R M and Kliger D S 1994 Ultrasensitive time-resolved linear dichroism spectral measurements using near-crossed linear polarizers *Chem. Phys. Lett.* **224** 145–54
- [39] Xie X L and Simon J D 1990 Picosecond magnetic circular dichroism spectroscopy *J. Phys. Chem.* **94** 8014–16
- [40] Hendler R W and Shrager R I 1994 Deconvolutions based on singular value decomposition and the pseudoinverse—a guide for beginners *J. Biochem. Biophys. Methods* **28** 1–33
- [41] Jolliffe I T 1986 *Principal Component Analysis* (New York: Springer)
- [42] Malinowski E R and Howery D G 1980 *Factor Analysis in Chemistry* (New York: Wiley)
- [43] Szundi I, Lewis J W and Kliger D S 1997 Deriving reaction mechanisms from kinetic spectroscopy. Application to late rhodopsin intermediates *Biophys. J.* **73** 688–702
-

FURTHER READING

Brückner V, Feller K-H and Grummt U-W 1990 *Applications of Time-Resolved Optical Spectroscopy* (New York: Elsevier)

A comprehensive review of fast and ultrafast time-resolved optical techniques.

Pilling M J and Seakins P W 1995 *Reaction Kinetics* (Oxford: Oxford University Press)

This kinetics text contains a comprehensive chapter on experimental techniques that overviews transient kinetic methods, as well as a chapter devoted to photochemistry.

Steinfeld J I, Francisco J S and Hase W L 1989 *Chemical Kinetics and Dynamics* (Englewood Cliffs, NJ: Prentice-Hall)

-31-

A kinetics text with a strong theoretical bent that overviews transient kinetic methods and discusses data analysis issues such as error propagation and sensitivity analysis.

Strehlow H and Knoche W 1977 *Fundamentals of Chemical Relaxation* Chemie, Weinheim and New York

A monograph on relaxation techniques intended to efficiently introduce newcomers to the field.

Bernasconi C F 1976 *Relaxation Kinetics* (New York: Academic)

A thorough treatment of the principles and experimental techniques of relaxation kinetics studies.

Rabek J F 1982 *Experimental Methods in Photochemistry and Photophysics* part 2 (New York: Wiley)

Comprehensively catalogues phenomena and techniques encountered in flash photolysis and fluorescence spectroscopies.

Bensasson R V, Land E J and Truscott T G 1993 *Excited States and Free Radicals in Biology and Medicine: Contributions from Flash Photolysis and Pulse Radiolysis* (Oxford: Oxford University Press)

Focuses on biological and biomedical applications of flash photolysis and pulse radiolysis methods.

Chen E, Goldbeck R A and Kliger D S 1997 Nanosecond time-resolved spectroscopy of biomolecular processes *Annu. Rev. Biophys. Biomol. Struct.* **26** 325–53

Reviews fast transient kinetic studies of biological molecules (excluding fluorescence studies).

-1-

C3.2 Electron transfer reactions

Gilbert C Walker and David N Bertan

C3.2.1 INTRODUCTION

C3.2.1.1 WHAT IS CHEMICAL ELECTRON TRANSFER?

Chemical reactions involve the redistribution of electronic charge and changes in chemical bonding. Chemical bonds can reorient, disconnect, and reform during chemical reactions. These bonding changes are coupled, in turn, to changes in the structure (orientation, hydrogen bonding, polarization etc.) of the surrounding solvent. Bonding changes can be subtle or dramatic. The subject of this chapter is the special class of reactions in which an electron is displaced by distances much larger than the length of a single chemical bond; such reactions are known as ‘electron-transfer reactions’.

Electron transfer reactions are conceptually simple. The coupled structural changes may be modest, as in the case of ‘outer-sphere’ electron transport processes. Other electron transfer processes result in bond formation or

cleavage—inner sphere electron transfer—and are more complex. Despite their apparent simplicity, outer-sphere electron-transfer reactions play a central role in chemistry, from current flow at electrodes to the early events in photosynthesis and to radiation damage in DNA. Theories that predict the relationship between chemical structure, solvation, spectroscopy and electron transfer rates were developed extensively over the last 50 years; they provide a valuable unifying thread in this field of research [1].

C3.2.1.2 THE DIVERSITY OF CHEMICAL ET SYSTEMS

Much of this chapter concerns ET reactions in solution. However, gas phase ET processes are well known too. See [figure C3.2.1](#). The ‘harpoon mechanism’ by which halogens oxidize alkali metals is fundamentally an electron transfer reaction [2]. One might guess, from this simple reaction, some of the structural parameters that control ET rates: relative electron affinities of reactants, reactant separation distance, bond length changes upon oxidation/reduction, vibrational frequencies, etc.

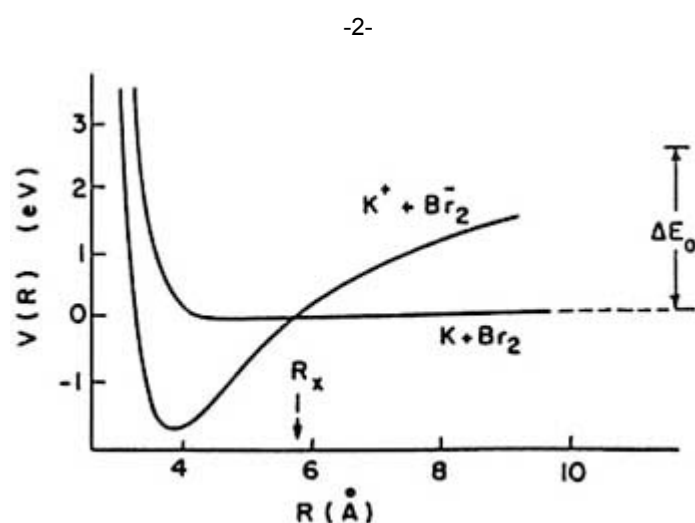


Figure C3.2.1. A slice through the intersecting potential energy curves associated with the $K+Br_2$ electron transfer reaction. At the crossing point between the curves (R_x), electron transfer occurs, thus ‘harpooning’ the Br_2 species, which is then associated with K^+ at shorter distances. From [2].

Much of the motivation for examining ET processes comes from biology. Many processes in bioenergetics involve transmembrane electron transport. The cascade of biological reactions that leads to an electrochemical gradient starts with electron transfer [3]. In the photosynthetic charge separation of purple bacteria, for example, a photoexcited state of a magnesium chlorophyll ‘special pair’ undergoes picosecond time-scale electron transfer to a pheophytin and then to a quinone [4]. These redox-active species are imbedded in a membrane-spanning protein whose structure was determined in 1984. These reactions are the subject of intense experimental and theoretical interest.

In the 1980s, considerable attention turned to ET reactions in fixed donor–acceptor geometries with the goal of understanding the control of the fixed distance biological reactions. Locking in the donor–acceptor separation distance simplifies the interpretation of measured ET kinetics [5], by removing uncertainties associated with intermolecular motion and docking. The simplest approach is to freeze donors and acceptors in a matrix [6], generating an ensemble of fixed distances. Simplifying the interpretation even further came from synthesizing covalently linked donor–bridge–acceptor molecules [7, 8] ([figure C3.2.2](#)). Indeed, unimolecular ET structures have been studied in considerable depth in both gas and solution-phase environments [9].

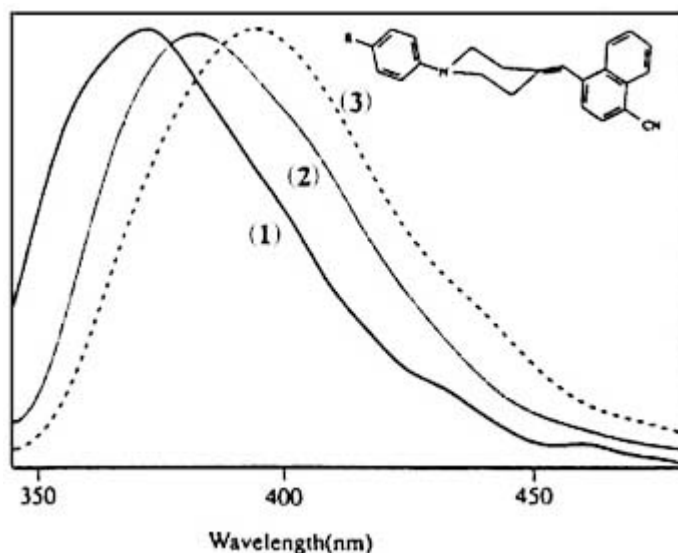


Figure C3.2.2. Long-range charge separation occurs from the S1 state of the rigidly bridged molecule above. This figure shows predominantly emission from the charge separated state. Absence of local emission from the S1 state indicates that ET occurs on a subnanosecond time-scale. Curves: (1) $R = -H$, (2) $R = -CH_3$, and (3) $R = -OCH_3$. From Wegewijs B and Verhoeven J W 1999 Long-range charge separation in solvent-free donor-bridge-acceptor systems *Adv. Chem. Phys.* **106** 248.

Electron transport processes at surfaces often involve electron-tunnelling transport. For example, in the scanning tunnelling microscope (STM) (see section B1.19 and figure C3.2.3), electrons flow between delocalized initial and final states. Depending on the experimental design, the tunnelling can proceed through vacuum or through attached atoms and molecules. In closely related photochemical experiments, electrons are driven from a delocalized electrode state to a localized molecular species or another electrode through an ‘insulating’ molecular monolayer [10].

In solid state materials, single-step electron transport between dopant species is well known. For example, electron-hole recombination accounts for luminescence in some materials [11]. Multistep hopping is also well known. Models for single and multistep transport are enjoying renewed interest in the context of DNA electron transfer [12, 13, 14 and 15]. Indeed, there are strong links between the ET literature and the literature of hopping conductivity in polymers [16].

-4-

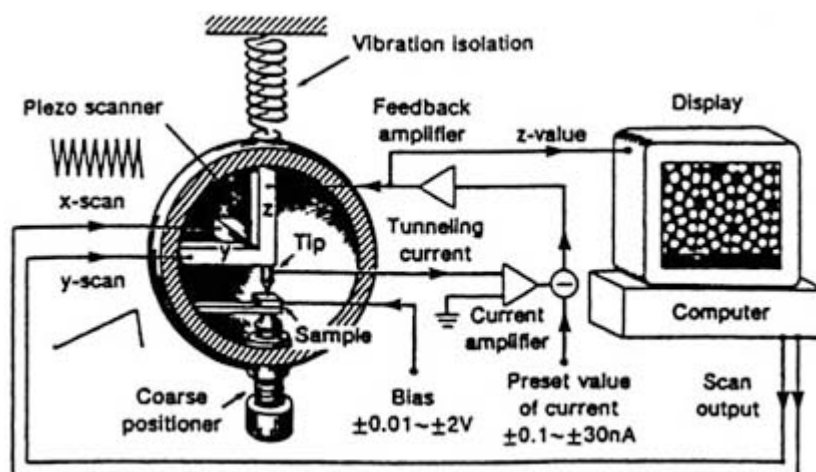


Figure C3.2.3. Schematic view of a scanning tunnelling microscope. From Chen C J 1993 *Introduction to Scanning Tunnelling Microscopy* (Oxford: Oxford University Press).

Consideration of the donor–bridge–acceptor systems just mentioned, in the presence of optical excitation, raises the question of whether the photoexcited state might have substantial charge transfer character. Indeed, excitation to charge-transfer excited states is observed in inorganic complexes; these transitions are known as metal–ligand, ligand–metal, and intervalence charge transfer bands. Intervalence bands are particularly diagnostic of bridge-mediated donor–acceptor interactions. The extent of excited state charge transfer is especially important in species such as the ‘special pair’ of chlorophylls in photosynthesis, as polarization of this excited state may influence its rate of ET [17].

Much of chemistry occurs in the condensed phase; solution phase ET reactions have been a major focus for theory and experiment for the last 50 years. Experiments, and quantitative theories, have probed how reaction-free energy, solvent polarity, donor–acceptor distance, bridging structures, solvent relaxation, and vibronic coupling influence ET kinetics. Important connections have also been drawn between optical charge transfer transitions and thermal ET.

C3.2.1.3 ET IN BIOLOGY

A substantial fraction of the named enzymes are oxido-reductases, responsible for shuttling electrons along metabolic pathways that reduce carbon dioxide to sugar (in the case of plants), or reduce oxygen to water (in the case of mammals). The oxido-reductases that drive these processes involve a small set of redox active ‘cofactors’, that is, small chemical groups that gain or lose electrons. These cofactors include iron porphyrins, iron–sulfur clusters and copper complexes as well as organic species that are ET active.

Many key protein ET processes have become accessible to theoretical analysis recently because of high-resolution x-ray structural data. These proteins include the bacterial photosynthetic reaction centre [18], nitrogenase (responsible for nitrogen fixation), and cytochrome *c* oxidase (the terminal ET protein in mammals) [19, 20]. Although much is understood about ET in these molecular machines, considerable debate persists about details of the molecular transformations.

C3.2.1.4 ET TECHNOLOGY: MEDICAL DIAGNOSTICS AND NANOSCALE ELECTRONICS

In addition to conventional applications in conducting polymers and electrooptical devices, a number of recent novel applications have emerged. Switching of DNA electron transfer upon single-strand/double-strand hybridization forms the basis for a new medical biosensor technology. Since the number of base pairs of length 20 is 4^{20} (or 10^{12}), a modest length DNA sequence, with ET properties that switch upon hybridization, can be employed to detect the presence of a complementary sequence. Many research efforts are engaged in devising ET-based biosensors to detect human disease and to sense the contamination of foodstuffs.

Small molecules are of the order of nanometres in linear dimensions. Conventional microelectronics technology employs features fully a hundred to a thousand times larger. As such, considerable interest is focused upon employing molecular species (together with and the rules of molecular ET) to devise ultra-small computing devices. Examples of recent advances include the demonstration of molecular-scale diodes (figure C3.2.4), prototypes for molecular scale memories, and single-electron devices [21, 22]. Remarkable physics arises in these devices. For example, in devices with dimensions of the order of the electron wavelength, conductivity is quantized; and current–voltage relations follow a stair-step pattern, rather than a simple linear relationship.

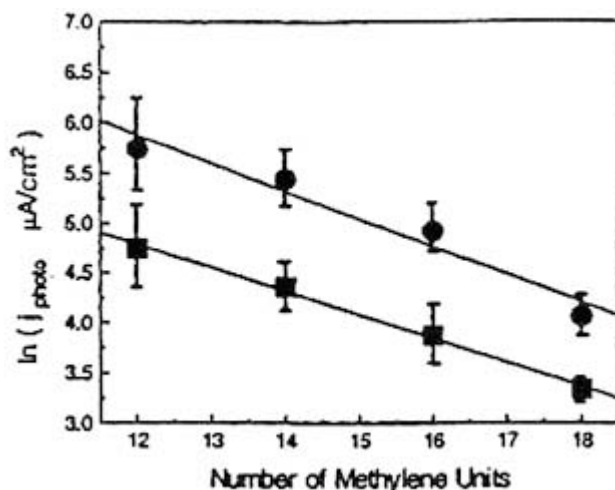


Figure C3.2.4. Plot of the log of photocurrent against number of methyl units in a alkylsilane based monolayer self-assembled on a *n* silicon electrode. The electrode is immersed in a solution with an electron donor. Best fits of experimental data collected at different light intensities: (○) 0.3 mW cm⁻²; (■) 0.05 mW cm⁻². From [10].

C3.2.2 ET THEORY AND EXPERIMENT

This section presents the basic theoretical principles of condensed phase electron transport in chemical and biochemical reactions.

-6-

C3.2.2.1 ADIABATIC ET THEORY

Modern electron transfer theory has its conceptual origins in activated complex theory, and in theories of nonradiative decay. The analysis by Marcus in the 1950s provided quantitative connections between the solvent characteristics and the key parameters controlling the rate of ET. The Marcus theory predicts an adiabatic bimolecular ET rate as

$$k_{\text{ET}} = A \exp\left(\frac{-\Delta G^*}{k_B T}\right) \quad (\text{C3.2.1})$$

where A is a collision frequency and ΔG^* is the free energy of activation for the donor–acceptor ET process when the redox species are in contact. Marcus theory leads to an ‘intersecting parabola model’ which results from the assumption that the distribution of nuclear configurations about the equilibrium is Gaussian in the distortion coordinate. In such a case, one finds that the activation energy is

$$\Delta G^* = \frac{(\Delta G + \lambda)^2}{4\lambda}. \quad (\text{C3.2.2})$$

Here λ is the ‘reorganization energy’ associated with the curvature of the reactant and product free energy wells and their displacement with respect to one another. Assuming a structureless polarizable medium, Marcus computed the solvent or outer-sphere component of the reorganization energy to be

$$\lambda_0 = (\Delta q)^2 \left(\frac{1}{\epsilon_0} - \frac{1}{\epsilon_s} \right) \left(\frac{1}{2a_1} + \frac{1}{2a_2} - \frac{1}{R} \right) \quad (\text{C3.2.3})$$

where Δq is the amount of charge transferred, ϵ_0 is the optical dielectric constant, ϵ_s is the static dielectric constant, a_1 is the donor radius, a_2 is the acceptor radius, and R is the donor–acceptor distance. Note that the outer-sphere reorganization energy is always positive, and grows with donor acceptor separation. For distances much larger than the donor/acceptor species size, the dependence of λ_0 on separation distance is weak. Reorganization energy is larger in polar solvents (compared to nonpolar solvents), where the difference between the optical and static dielectric constant (reciprocals) will be large. Reorganization energies are now computed routinely for much more complex media with contacts between regions with high and low dielectric constants.

C3.2.2.2 NONADIABATIC ET THEORY

In many instances the adiabatic ET rate expression overestimates the rate by a considerable amount. In some circumstances simply forming the the activated state geometry in the encounter complex does not lead to ET. This situation arises when the donor and acceptor groups are very weakly coupled electronically, and the reaction is said to be nonadiabatic. As the geometry of the system fluctuates, the species do not move on the lowest potential energy surface from reactants to products. That is, fluctuations into activated complex geometries can occur millions of times prior to a productive electron transfer event.

-7-

In this weakly coupled regime, ET in an encounter complex can be described approximately using a two-level system model [23]. As such, the time-dependent wave function is

$$\Psi(t) = i\Phi_A \sin\left(\frac{T_{DA}}{\hbar}t\right) + \Phi_D \cos\left(\frac{T_{DA}}{\hbar}t\right) \quad (\text{C3.2.4})$$

where Φ_D represents the donor wave function (acceptor for A) and T_{DA} is the donor–acceptor coupling. This coupling can be enhanced by mixing of the D and A states with each other via intervening bridge orbitals. Note that amplitude localized on donor or acceptor oscillates sinusoidally in time (neglecting relaxation processes) with a frequency determined by the strength of the donor–acceptor coupling, T_{DA} . Fermi’s golden rule of time-dependent perturbation theory can be used to compute the rate of ET based upon the short-time evolution of the system:

$$k_{\text{ET}} = \frac{2\pi}{\hbar} |T_{DA}|^2 \cdot \frac{1}{(4\pi\lambda k_B T)^{1/2}} \cdot \exp\left[-\frac{(\Delta G + \lambda)^2}{4\lambda k_B T}\right]. \quad (\text{C3.2.5})$$

Here we have treated the nuclear degrees of freedom classically as in the Marcus formulation [1].

C3.2.2.3 TUNNELLING BARRIERS

The new challenge that arises in making predictions of nonadiabatic electron transfer rates is to determine the electronic coupling element, T_{DA} . Simple orbital tunnelling analysis predicts that if (a) the donor is delocalized over N_D orbitals and the acceptor is delocalized over N_A orbitals, (b) the average of the donor and acceptor orbital energies are 2 eV removed from the bridging levels (based upon the electronic absorption properties of proteins), (c) the donor–acceptor distance is large (measured edge-to-edge between donor and acceptor in Å), and (d) the donor–acceptor interaction at ‘contact’ is 1 eV:

$$(\text{C3.2.6})$$

$$T_{\text{DA}}(eV) = \frac{1}{\sqrt{(N_A)}} \frac{1}{\sqrt{(N_D)}} (2.7) \exp[-0.72 R_{\text{DA}}] \propto \exp[-(\beta/2) R_{\text{DA}}].$$

This conjecture was made for protein ET systems in particular [24]. It was later found that the exponential decay constant, $\beta/2$, varies strongly as a function of bridging orbital symmetry and donor/acceptor energetics in both proteins and in smaller model compounds [25]. Considerable theoretical and experimental efforts have gone into determining this average decay parameter, and many studies of rigid donor–bridge–acceptor systems were motivated by a desire to address this question.

The distance decay of tunnelling through vacuum is more rapid than the decay for tunnelling through bond. Through space, the ‘barrier’ to tunnelling is essentially the binding energy of the donor/acceptor states. However, in a bridged system this barrier is generally much smaller, determined by the energy gap between the donor/acceptor states, the energies of the bridging orbitals, and the interaction strength among the orbitals. In large complex systems, the strength of this interaction is estimated from the tunnelling pathway model according to the formula [26]

$$T_{\text{DA}} = P \prod_i \prod_j \prod_k \varepsilon_i^{\text{bond}} \varepsilon_j^{\text{space}} \varepsilon_k^{\text{H-bond}} \quad (\text{C3.2.7})$$

where $\varepsilon_i^{\text{bond}} \approx 0.6$, $\varepsilon_j^{\text{space}} \approx 0.6 \exp[-1.7(R_{\text{bond}} - 1.4 \text{ \AA})]$ and $\varepsilon_k^{\text{H-bond}} \approx 0.36$. These factors are chosen in such a way that the ‘pathway product’ of equation (C3.2.7) joining the donor and acceptor sites in a protein or protein-protein complex is a maximum. Because there is a unique decay factor (ε) for each contact defined in a protein x-ray or NMR structure, the ‘strongest’ pathway can be determined by a relatively simple graph-search algorithm. The prefactor P can be chosen as in the square-barrier model described above (equation (C3.2.6)), or can be fitted to the experiment. Figure figure C3.2.5 shows the strongest pathways determined for a family of ruthenium modified cytochromes *c*. Pathway analysis also predicts that the average decay of coupling with distance depends upon protein secondary structure [26]. Ongoing studies utilizing modern quantum chemical methods are ‘summing up’ the large number of pathway contributions to the donor–acceptor coupling in an effort to make predictions of increasing quantitative reliability [27].

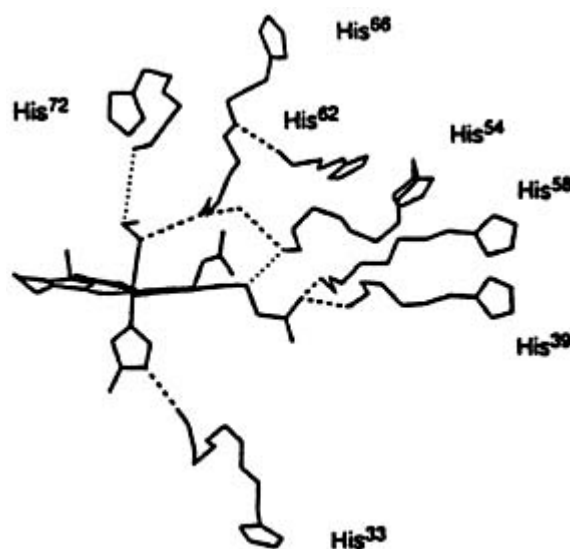


Figure C3.2.5. Strongest tunnelling pathways between surface histidines and the iron atom in cytochrome *c*. Steps in pathways are denoted by solid lines (covalent bonds), dashed lines (hydrogen bonds), and through-space contacts (dotted lines). Electron transfer distance to His 72 is 5 Å shorter than in His 66, yet the two rates are approximately

the same. The long-distance through space contact on the dominant pathway of His 72 accounts for this dramatic effect. From Langen R, Chang I J, Germanas J P, Richards J H, Winkler J R and Gray H B 1995 *Science* **268** 1733.

The pathway model makes a number of key predictions, including: (a) a substantial role for hydrogen bond mediation of tunnelling, (b) a difference in mediation characteristics as a function of secondary and tertiary structure, (c) an intrinsically nonexponential decay of rate with distance, and (d) pathway specific 'hot and cold spots' for electron transfer. These predictions have been tested extensively. The most systematic and critical tests are provided with ruthenium-modified proteins, where a synthetic ET active group can be attached to the protein and the rate of ET via a specific medium structure can be probed (figure C3.2.5).

-9-

The predictive power of pathway analysis is well illustrated with two of the Ru-modified systems of Gray and coworkers [29]. Consider, the His 72 and His 39 ruthenium-modified cytochromes *c* [28]. The ET rates in these proteins are about the same, despite the fact that the transfer distance is fully 5 Å shorter in the His 72 derivative. Average square barrier models with $\beta=1.4 \text{ \AA}^{-1}$ (equation (C3.2.6)) would predict the His 72 rate to be 1000 times faster. This equivalency of rates despite the great difference in distances is understood because the strongest pathways in the His 72 derivative contain a through-space tunnelling gap.

A large body of rate data in native and modified proteins was analysed recently in the context of the pathway model. Correcting for differences in activation free energies in different proteins (ΔG^*), rates fall into alpha-helical and beta-sheet 'zones' when plotting ET rate against distance in a wide range of native and modified proteins [29] (figure C3.2.6), consistent with one of the most fundamental predictions of the pathway model [26]. The average decay exponents in the two regions are approximately 1.1 \AA^{-1} (β -sheet) and 1.4 \AA^{-1} (α -helix), consistent with the pathway predictions.

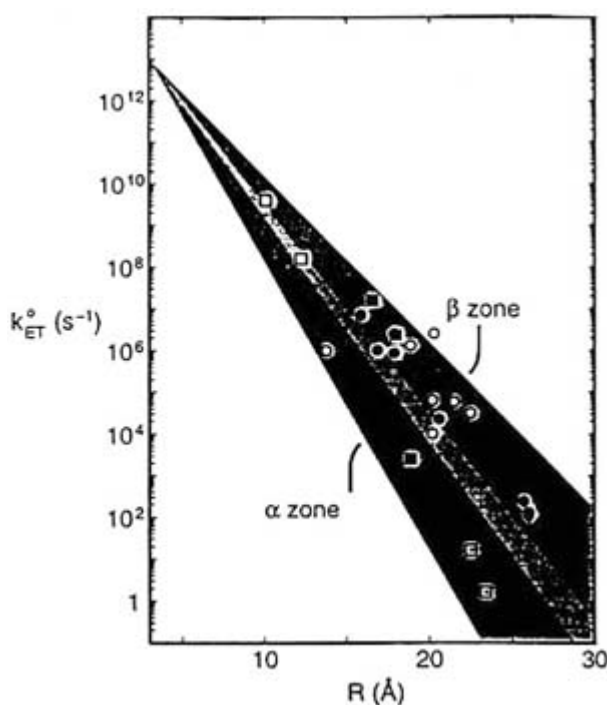


Figure C3.2.6. Zones associated with the distinctive decay of electronic coupling through α -helical against β -sheet structures in proteins. Points shown refer to specific rates in ruthenium-modified proteins and in the photosynthetic reaction centre. From Gray H B and Winkler J R 1996 *Electron transfer in proteins* *Ann. Rev. Biochem.* **65** 537.

In addition to testing predictions of the pathway model in proteins, experiments have also examined the prediction that the decay across a hydrogen bond (from heteroatom to heteroatom) should be about as costly as the decay across two covalent bonds. Indeed, by synthesizing a family of hydrogen bonded and covalently bonded systems with equal bond counts (according to this recipe), it was demonstrated that coupling across hydrogen bonded

contacts is about as favourable as across covalent bonds [30] (figure C3.2.7).

-10-

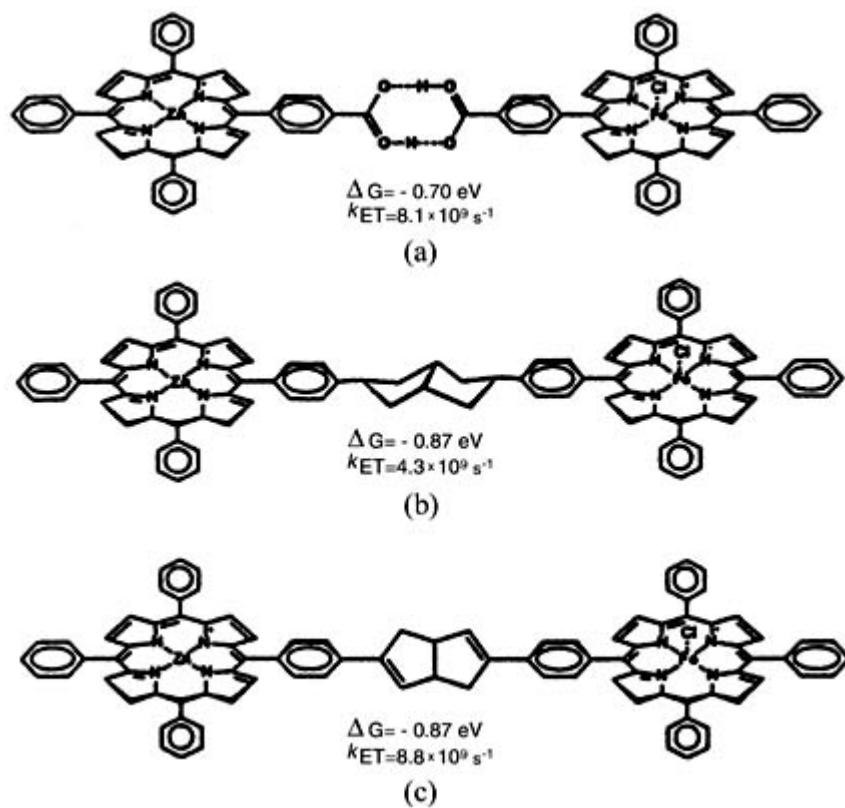


Figure C3.2.7. A series of electron transfer model compounds with the donor and acceptor moieties linked by (from top to bottom): (a) a hydrogen bond bridge; (b) all sigma-bond bridge; (c) partially unsaturated bridge. Studies with these compounds showed that hydrogen bonds can provide efficient donor–acceptor interactions. From Piotrowiak P 1999 Photoinduced electron transfer in molecular systems: recent developments *Chem. Soc. Rev.* **28** 143–50.

C3.2.2.4 BRIDGE ORBITAL SYMMETRY EFFECTS IN CHEMICAL SYSTEMS

The simplest theoretical orbital-based estimate of the coupling interaction, T_{DA} , is provided by the McConnell relation:

$$T_{\text{DA}} \propto \left[\frac{V}{\Delta E} \right]^N. \quad (\text{C3.2.8})$$

Rewriting T_{DA} as an exponential,

$$\beta = -\frac{1}{\alpha} \ln \left| \frac{V}{\Delta E} \right| \quad (\text{C3.2.9})$$

-11-

where V is the interaction between neighbouring bonds in the bridge, ΔE is the energy gap between the redox active D/A orbitals and the bridge mediating bonds, and a is the distance between the neighbouring bonds. In a system that mediates coupling by more than one set of orbitals, this expression can be generalized. Nevertheless, [equation \(C3.2.9\)](#) provides a rough means of describing the physical aspects of bridge mediation. (figure C3.2.8).

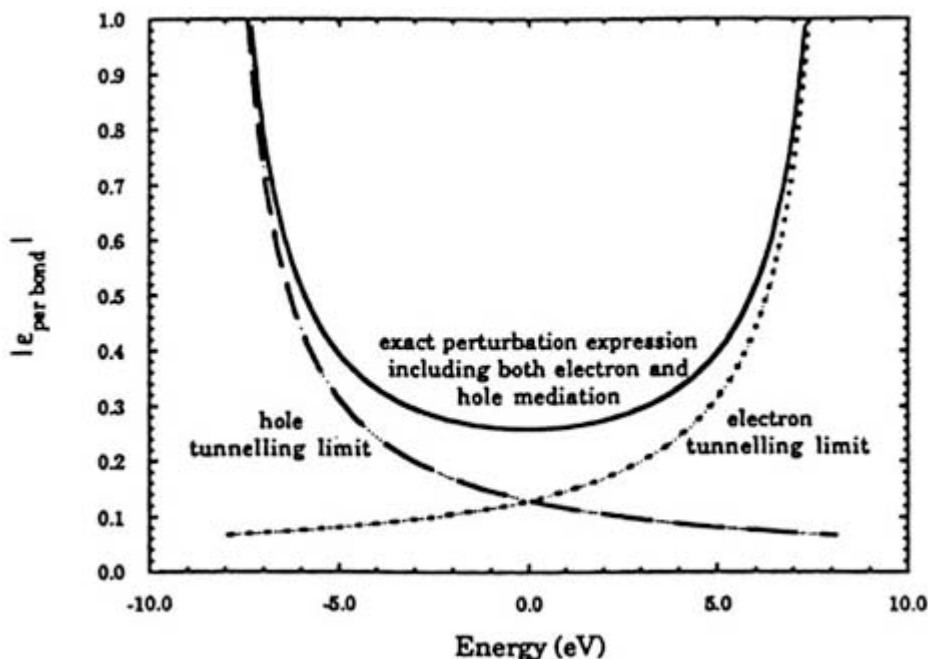


Figure C3.2.8. Dependence of the donor–acceptor coupling decay per bond, $|\epsilon \text{ per bond}|$, upon tunnelling energy. Average exponential decay parameter, β , is related to this decay per parameter by $\beta = (1/R_{\text{unit}}) \ln |\epsilon \text{ per bond}|$ for periodic bridges. R_{unit} is the spacing between repeating units in the bridge. Decay of the coupling with distance is softest ($|\epsilon \text{ per bond}|$ is closest to 1) for tunnelling energies near the frontier orbital energies of the bridge (which lie at -6 and $+7$ eV in this figure). From D N Beratan and J N Onuchic 1991 *Electron Transfer in Inorganic, Organic, and Biological Systems (Advances in Chemistry Series 228)* ed J R Bolton, N Mataga and G McLendon (Washington, DC: ACS Press).

Note that β is large in the limit that V is small comparable to ΔE . However, as the coupling strength increases or as the energy gap drops, β can become smaller. In fact, as the donor/acceptor states approach thermal energies (kT) of the bridge orbital energies, the golden rule ([equation C3.2.5](#)) treatments of the problem are no longer appropriate, and more elaborate models are called for. These models would include the possibility of multistate hopping and would need to consider the time-scales of hopping compared to the time-scales of thermal trapping on the individual sites.

The McConnell relation does not provide quantitative estimates of electronic propagation because (a) it does not include the influence of antibonding orbitals, (b) it neglects ‘through-space’ nonnearest neighbour interactions, and (c) it does not include contributions from multiple interfering coupling pathways.

It is now understood that inclusion of nearest and second-neighbour interactions is adequate for describing tunnelling interactions in many bridged systems [[31](#), [32](#)]. Moreover, tunnelling interactions have been dissected for various

bridge structures. This kind of analysis permits one to understand the nature of constructive and destructive interference interactions that arise in specific bridging species. The central nature of interference can be understood from the decay term $V/\Delta E$ in the McConnell relation. In a one-orbital model V is a negative number (a ‘resonance’ integral in the language of Hückel theory) but ΔE is positive (the D/A states lie at energies above the HOMO of the bridge). These V elements are readily calculated using quantum chemical methods ([figure C3.2.9](#)). As such, as

longer-length pathways are introduced, they will make contributions with sign that oscillates as their contributions drop in magnitude [31].

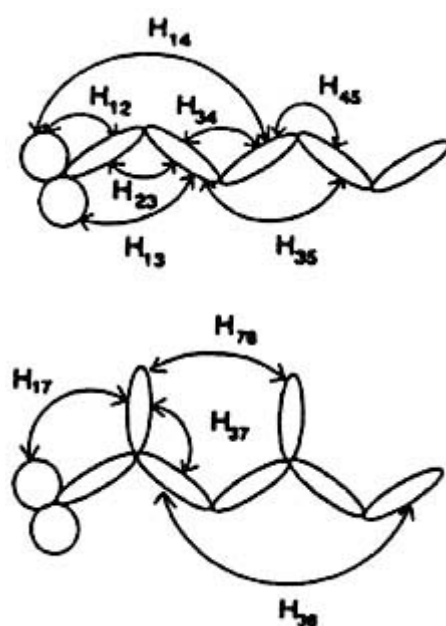


Figure C3.2.9. Both nearest neighbour and nonnearest neighbour coupling interactions mediate superexchange between the terminal pi-electron groups of rigid dienes with saturated bridging units. From [31].

Inclusion of coupling contributions from both bonding and anti-bonding orbitals give rise to a U-shaped dependence of coupling on D/A energetics (figure C3.2.8).

C3.2.2.5 INNER-SPHERE REORGANIZATION ENERGY

Whether adiabatic or nonadiabatic, it is the case that both solvent and intramolecular degrees of freedom respond to ET events. As such, the two rate expressions given above can be generalized such that

$$\lambda = \lambda_0 + \lambda_i \quad (\text{C3.2.10})$$

where λ_0 is the outer sphere (solvent) contribution to the reorganization energy and λ_i is the intramolecular or inner sphere contribution to the reorganization energy.

-13-

Rate formulations that treat the inner-sphere mode(s) quantum mechanically and the outer sphere modes classically are used rather widely. The rate expression for a single harmonic quantum mode is

$$k_{\text{ET}} = \frac{2\pi}{\hbar} |T_{\text{DA}}|^2 \cdot \frac{1}{(4\pi\lambda k_{\text{B}}T)^{1/2}} \sum \frac{e^{-S} S^n}{n!} \cdot \exp \left[\frac{(\Delta G + \lambda_0 + n\hbar\omega)^2}{4\lambda k_{\text{B}}T} \right] \quad (\text{C3.2.11})$$

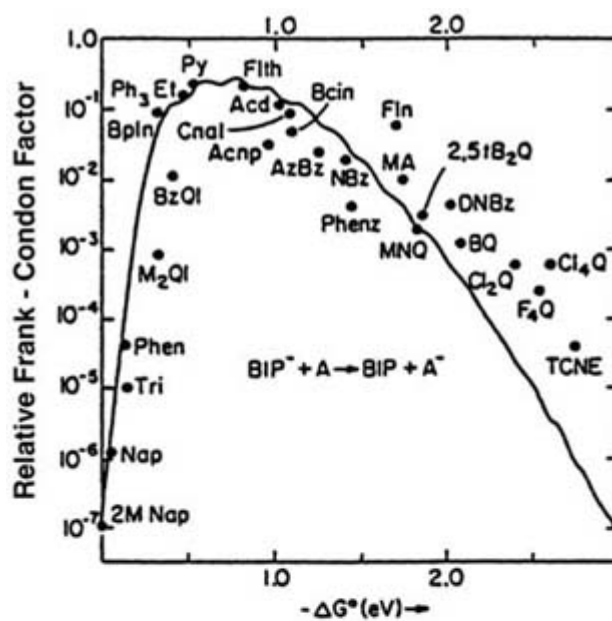
where S is the inner sphere reorganization energy divided by the energy of the vibrational quantum ($\hbar\omega$) and the other terms are as defined above. This expression is used for interpreting experimental rate data. The major qualitative effect of the quantum mode is to slow the drop off of the rate in the inverted region (where $-\Delta G > \lambda$ and see section C3.2.2.6). ET rates have been formulated to include the effects of multiple quantum modes, anharmonic

potentials; rates valid beyond the golden rule regime have been established as well [32].

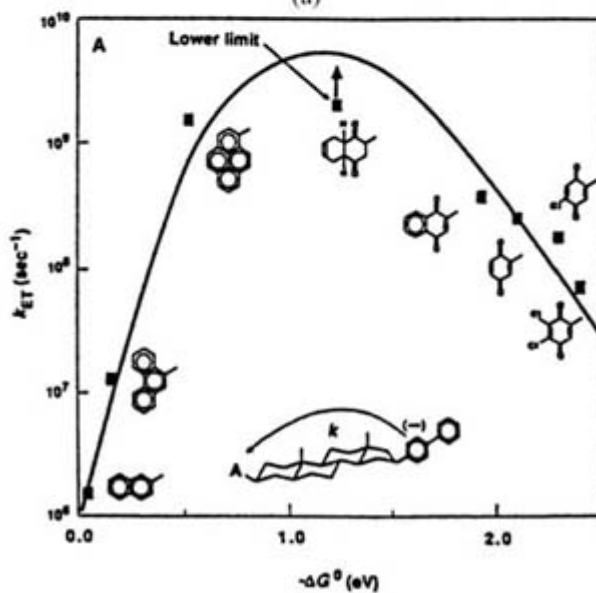
C3.2.2.6 FREE ENERGY – RATE RELATIONS

The form of the classical (equation C3.2.11) or semiclassical (equation C3.2.11) rate equations are ‘energy gap laws’. That is, the equations reflect a free energy dependent rate. In contrast with many physical organic reactivity indices, these rates are predicted to increase as $-\Delta G$ grows, and then to drop when $-\Delta G$ exceeds a critical value. In the classical limit, $\log(k_{\text{ET}})$ has a parabolic dependence on $-\Delta G$. When high-frequency chemical bond vibrations couple to the ET process, the dependence on $-\Delta G$ becomes asymmetrical, as mentioned above.

A tremendous effort was made in the 1980s to test the prediction of an inverted region. Covalently linked donor–acceptor species were constructed in such a way that the energies of the donor and acceptor groups (ionization potential of donor, electron affinity of acceptor) could be changed to vary ΔG (figure C3.2.10). Modelled loosely on photosynthetic ET systems, many of the structures contained porphyrin electron donors and quinone acceptors. Utilizing an essentially rigid chemical bridge removed ambiguity associated with a distribution of distances (and hence of T_{DA} values). Many practical limitations make mapping of the inverted region challenging, but there are now several examples of inverted behaviour for charge separation. Inverted behaviour was recently investigated in charge recombination reactions, that prove particularly amenable to study [33].



(a)



(b)

Figure C3.2.10.(a) Dependence of electron transfer rate upon reaction free energy for ET between biphenyl radical anions and various organic acceptors. Experiments were performed with the donors and acceptors frozen into organic (methyltetrahydrofuran) glasses. Parameters: 10^{-6} s; $\lambda_s=0.4$ eV; $\lambda_v=0.4$ eV; $\omega=1500$ cm^{-1} . From Miller J R 1987 *New J. Chem.* **11** 83. (b) Dependence of electron transfer rate upon reaction free energy for ET between biphenyl radical anions and various organic acceptors attached to each other by a rigid spacer measured methyl tetrahydrofuran solution. Parameters: $\lambda_s=0.75$ eV; $\lambda_v=0.45$ eV; $\omega=1500$ cm^{-1} ; $V=6.2$ cm^{-1} . From Closs G L and Miller J R 1988 Intramolecular electron transfer in organic molecules *Science* **240** 440.

C3.2.2.7 MARCUS CROSS-RELATION

A powerful application of outer-sphere electron transfer theory relates the ET rate between D and A to the rates of self exchange for the individual species. Self-exchange rates correspond to electron transfer in $D/D_{-}(k_{11})$ and $A/A_{+}(k_{22})$. These rates are related through the cross-relation to the D/A electron transfer reaction by the expression

where f_{12} is a value that is often of order unity [34].

The cross relation has proven valuable to estimate ET rates of interest from data that might be more readily available for individual reaction partners. Simple application of the cross-relation is, of course, limited if the electronic coupling interactions associated with the self exchange processes are drastically different from those for the cross reaction. This is a particular concern in protein/protein ET reactions where the coupling may vary drastically as a function of docking geometry.

C3.2.2.8 SOLVENT POLARITY-RATE RELATIONS

Electron transfer reaction rates can depend strongly on the polarity or dielectric properties of the solvent. This is because (a) a polar solvent serves to stabilize both the initial and final states, thus altering the driving force of the ET reaction, and (b) in a reaction coordinate system where the distance between reactants and products (DA and D^+A^-) is one, the strength of the solvent–electron coupling alters the curvature along a solvent nuclear coordinate of the diabatic representations for each electronic state. An incomplete but useful beginning point to understand this coupling is to think of the strength of the total dipolar field that a solvent, by reorganizing its nuclear charges, could project onto the axis of charge separation. It is easy to see that strongly polar solvents, like propylene carbonate or water, will exert a stronger coupling in this way than weakly polar solvents like toluene.

Formally, these effects enter the ET rate expressions by contributing to the reorganization energy, λ_0 , see equation (C3.2.3). The effects may be obtained either from spectroscopic parameters or from *a priori* models. The classical approach to obtaining λ_0 is to use dielectric continuum theory. In the case of two spherical reactants, the result appears in equation (C3.2.3) above. The validity of this approach has been tested often, though perhaps the best known examples are transition metal exchange reactions (figure C3.2.11). More recently, interest has focused on molecular representations of solvation shells and quantum effects on solvation rates that may be obtained from an analysis of the high-frequency solvent motions, particularly in the first solvent shell.

-16-

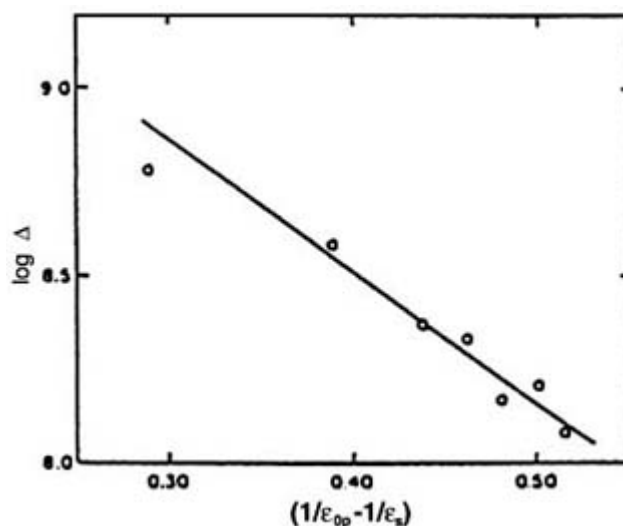


Figure C3.2.11. Log of the ET rate (Δ) against $(1/\epsilon_{op} - 1/\epsilon_s)$ for the bis(biphenyl) chromium⁺⁰ self-exchange reaction. From [34].

C3.2.2.9 INTERRELATIONSHIPS IN ET THEORY: OPTICAL VERSUS THERMAL ET

One of the most interesting aspects of the basic formulation of thermal electron-transfer theory (as portrayed in

figure [figure C3.2.10](#)) is that it is closely related to the Franck–Condon model for optical electronic transitions. This establishes a strong connection between spectroscopy and kinetics. Optical charge transfer spectra can be used to determine most (and in some cases all) of the required parameters for making rate-constant predictions using thermal electron-transfer theory [35]. λ and T_{DA} can be obtained from spectroscopic properties. This was first outlined in the late 1960s when it was shown that in a symmetric ($\Delta G=0$) system, $E_{op}=\lambda$. When T_{DA} is small and can be neglected, $E_{op}=\lambda+\Delta G_0$, in unsymmetrical systems. The use of the integrated oscillator strength of a charge transfer band to determine T_{DA} was also introduced. Using a Mullikan formalism, Hush showed that the electronic coupling element is related to the intensity of the charge transfer transition by

$$T_{DA} = \frac{(0.0206 \bar{\nu}_{\max} \epsilon_{\max} \Delta \bar{\nu}_{1/2})^{1/2}}{r_{DA}} \quad (\text{C3.2.13})$$

where $\bar{\nu}_{\max}$ and ϵ_{\max} are the frequency of the band maximum and the band width in wavenumbers, and r_{DA} is the donor–acceptor centroid distance in angstroms [36, 37]. This Mullikan–Hush expression has been extensively applied to outer sphere [38] and bridged ET systems [39].

C3.2.2.10 DYNAMICAL SOLVENT CONTROLLED RATES

Chemical changes are not irreversible unless there is some form of dissipation in the system. That is, the reaction free energy must be dispersed to a number of degrees of freedom distinct from the reaction coordinate. Models that include

-17-

dissipation predict that the reaction *mechanism* itself can be switched from adiabatic to nonadiabatic based upon the nature of the solvent relaxation. Solvent that relaxes ‘slowly’ induces multiple recrossings of the activated complex prior to relaxation away from the crossing region. The effect of this is to cause the onset of adiabatic-like behaviour earlier than might be anticipated. The rate expression that spans both dynamical regimes, assuming a Debye solvent with a single relaxation time, is

$$k_{ET} = \frac{k^{NA}}{1 + \kappa} \quad (\text{C3.2.14})$$

where the adiabaticity parameter is $\kappa = 4\pi T_{DA}^2 \tau / \hbar \lambda$, and k^{NA} is the nonadiabatic ET rate.

In Debye solvents, τ is the longitudinal relaxation time. The prediction that solvent polarization dynamics would limit intramolecular electron transfer rates was stated theoretically [40] and observed experimentally [41].

C3.2.2.11 VIBRATIONAL MODE COUPLING TO ET

The Franck–Condon principle reflected in the connection between optical and thermal ET also relates to the participation of high-frequency vibrational degrees of freedom. Charge transfer and resonance Raman intensity bandshape analysis has been used to determine effective vibrational and solvation parameters [42,43].

To make connection between the spectra and the ET process clearer, we note a simple model for the lineshape that includes a classical and a high-frequency degree of freedom. In this case the overall lineshape is

$$F(\nu) \propto \sum_n |\langle 0 | n \rangle| \cdot \exp \left[\frac{(\Delta G_n^0 + \lambda_0 - h\nu_{QM})^2}{4\lambda_0 k_B T} \right] \quad (\text{C3.2.15})$$

where λ_0 is the classical reorganization energy, and λ_1 is the reorganization energy in the quantum nuclear mode of frequency ν_{QM} . The free-energy difference between the charge-separated states then depends on the quantum number of the quantum nuclear mode, ($\Delta G_n^0 = (\Delta G^0 + n\hbar\nu_{QM})$). n is the index for the quantum mode.

This lineshape analysis also implies that electron-transfer rates should be vibrational-state dependent, which has been observed experimentally [44]. Spin-orbit relaxation has also been identified as an important factor in controlling the identity of both electron and vibrational-state distributions in radiationless ET reactions.

C3.2.2.12 COMPETITION BETWEEN INNER SPHERE AND OUTER SPHERE NUCLEAR POLARIZATION DYNAMICS

Early studies showed that the rates of ET are limited by solvation rates for certain barrierless electron transfer reactions. However, more recent studies showed that electron-transfer rates can far exceed the rates of diffusional solvation, which indicate critical roles for intramolecular (high frequency) vibrational mode couplings and inertial solvation. The interplay between inter- and intramolecular degrees of freedom is particularly significant in the Marcus inverted regime [45] (figure C3.2.12).

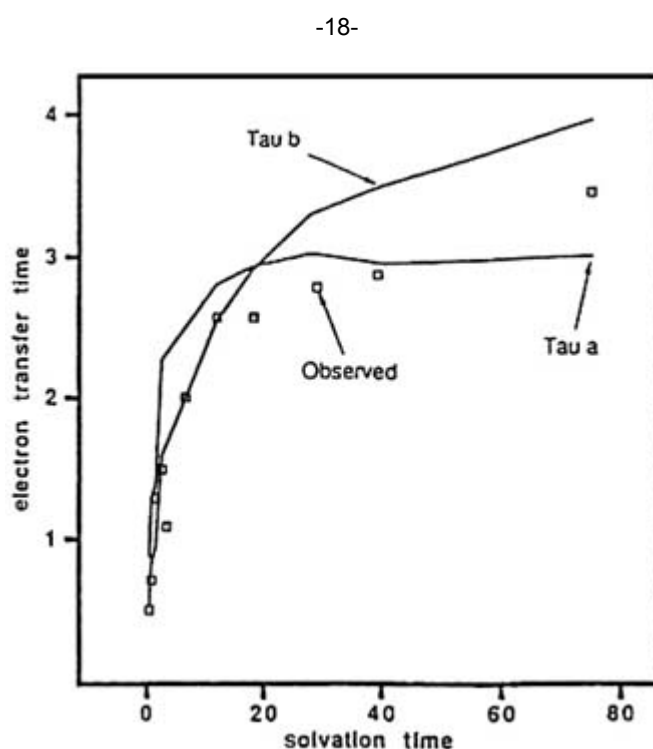


Figure C3.2.12. Experimentally observed electron transfer time in psec (squares) and theoretical electron transfer times (survival times, Tau a and Tau b) predicted by an extended Sumi-Marcus model. For fast solvents the survival times are a strong function of the characteristic solvent relaxation dynamics. For slower solvents the electron transfer occurs through the motion of intramolecular degrees of freedom. From [45].

C3.2.2.13 ORIENTATION IN INTERMOLECULAR ET

The electrostatic forces that control orientation also influence the D/A overlap and thus tunnelling probability. This balance is known to be manipulated in rather subtle ways in biological ET. In a small molecule system, the relative orientation of electron donor (dimethylaniline) and acceptor (coumarin 337) in a solvent/solute ET reaction was examined [46]. Figure figure C3.2.13 shows the time-dependent response that is measured at 500 nm after 400 nm electronic excitation of the coumarin. At early times, the transition moment probed at 500 nm belongs to coumarin. At later times after electron transfer from the dimethylaniline to the coumarin, the probed transition moment belongs to the dimethylaniline radical cation. The naïve expectation might be that the solvent molecules most stably oriented relative to the solute would preferentially undergo ET. In fact, it was found that the electron transfer occurred when the permanent dipole moments of the donor and acceptor were perpendicular, not antiparallel.

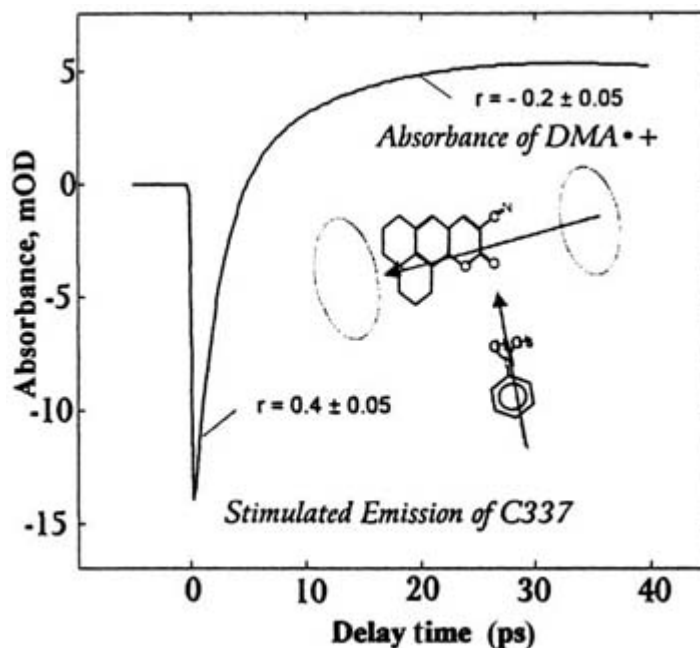


Figure C3.2.13. Orientation in a photoinitiated electron transfer from dimethylaniline (DMA) solvent to a coumarin solute (C337). Change in anisotropy, r , reveals change in angle between the pumped and probed electronic transition moments. From [46].

C3.2.2.14 ET IN PHOTOSYNTHETIC REACTION CENTRES

Electron transfer occurs in many steps in photosynthesis. Recently, considerable attention has been paid to the initial steps of charge separation. These initial steps occur in the picosecond time domain. A remarkable fact is that the photosynthetic reaction centre is highly symmetrical, with near C_2 symmetry in the cofactors but electron transfer occurs down only one path. It is believed that a slight asymmetry in the protein dielectric environment contributes to electron preference for one branch. Just how strongly the electron interacts with all of the nearby cofactors has also been an issue of interest; some argue that the electron passes along a transport chain, with identifiable populations at each cofactor, while others have argued that some cofactors participate only as superexchange mediators. Because the initial electron transfer processes are very fast (<3.5 ps in *Rhodobacter spheroides*) in photosynthetic reactions, these centres have also provided a testbed for analysing the role of nonequilibrium vibrations and their possible coherent coupling to electron transfer. Such coherences have been examined in smaller molecular systems [47, 48, 49 and 50] as well.

C3.2.2.15 LASER-FIELD DRIVEN ET COHERENCES

Calculations within the framework of a reaction coordinate degrees of freedom coupled to a bath of oscillators (solvent) suggest that coherent oscillations in the electronic-state populations of an electron-transfer reaction in a polar solvent can be induced by subjecting the system to a sequence of monochromatic laser pulses on the picosecond time scale. The ability to tailor electron transfer by such light fields is an ongoing area of interest [51] (figure C3.2.14).

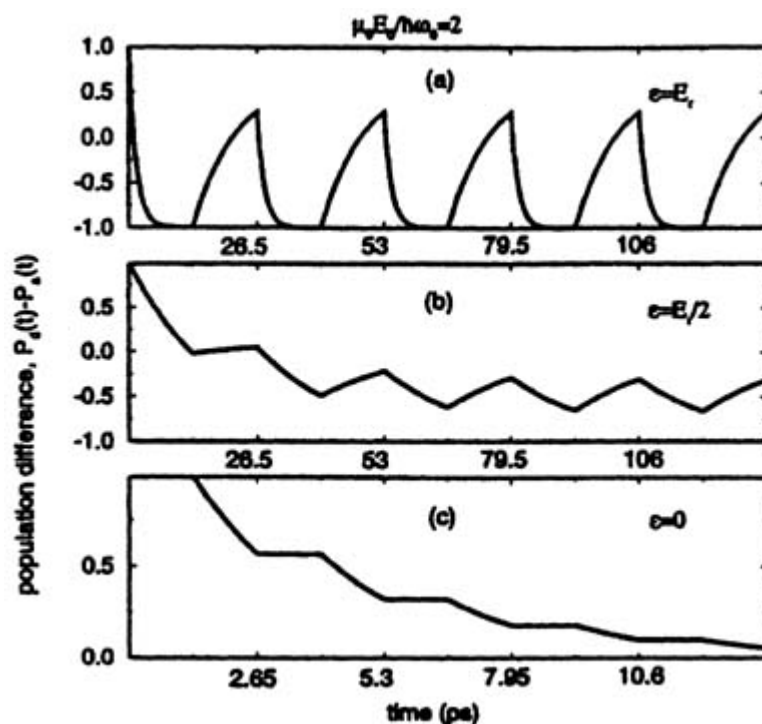


Figure C3.2.14. Electron population difference $x(t) = P_d(t) - P_a(t)$ for three electron transfer reactions in the presence of a pulsed laser field. Frequency of the field is tuned to solvent reorganization energy $\lambda = 1$ eV and the field strength is such that coupling potential (charge transfer dipole moment times the field strength) is twice the laser field frequency. (a) Activationless reaction, $\Delta G^0 = \lambda$, (b) reaction with $\Delta G^0 = \lambda/2$, and (c) symmetric electron transfer with no bias. From [51].

C3.2.3 APPLICATIONS IN COMPLEX SYSTEMS

This section describes the application of the theoretical principles described above to specific structures and processes of current interest in electron transfer research.

C3.2.3.1 DNA ELECTRON TRANSFER

Over the last decade attention has turned to the nature of electron transport in DNA [52]. Although DNA electron transfer has been of interest for a much longer time, new methodologies for attaching redox active donors and acceptors at defined positions in complex macromolecules has driven a resurgence of interest. One of the principle issues in this field concerns the mechanism of charge transfer, that is, whether the reactions proceed by long-range (single-step) tunnelling or multistep hopping. These mechanisms are summarized in figure figure C3.2.15.

-21-

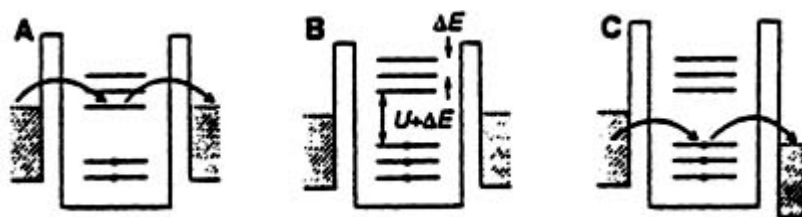


Figure C3.2.15. Schematic diagram showing (A) electron hopping between electron reservoirs *via* empty states of an intervening bridge, (B) tunnelling, and (C) hole hopping *via* filled states of an intervening bridge. From

Bockrath M, Cobden D H, McEuen P L, Chopra N G, Zettl A, Thess A and Smalley R E 1997 Single-electron transport in ropes of nanotubes *Science* **275** 1922–5.

β values reported for DNA electron transfer vary from 0.2 to 1.5 \AA^{-1} . Assuming that the reactions proceed by single-step tunnelling (see equation C3.2.5), explanations of the physical origin for this wide range of values include: (a) the stacking interactions (V of equation (C3.2.8)) might be highly variable because of variations in the nature of stacking or (b) changes in the energy denominator by changes in the average energetics of the redox active donor and acceptor orbitals. An alternative explanation for small apparent β values is that the process proceeds by multistep hopping (figure C3.2.15), where many very rapid short distance steps lead to a weak apparent distance dependence. Indeed, recent experiments involving oxidation of guanine, most likely fall in the regime of either tunnelling or multistate hopping, depending upon the details of the way in which the system is constructed.

Although the challenge of determining the distance dependence of ET in DNA seems academic, it is of considerably wider interest. There are now a number of companies developing medical diagnostic devices based upon changes in ET rates or electrode currents upon recognition and binding of single-stranded DNA. The sensitivity of these devices will depend upon how ET rates change upon oligomer binding, nature of base pair mismatches, and the solvation environment upon recognition.

C3.2.3.2 STM AND SINGLE MOLECULE CONDUCTIVITY

The invention of the scanning tunnelling microscope in the 1980s opened up new directions for electron transfer chemistry. The STM is discussed in detail in section B1.20 of this encyclopedia. Measurements of tunnelling current propagating through empty space and through various adsorbates provide a relatively direct probe of tunnelling propagation. Measurement of the current as a function of adsorbate layer thickness probes the decay parameter β directly (figure C3.2.16)). Single molecule bridged STM studies have mirrored results seen in solution chemistry: as the effective energy barrier is decreased and as the effective interaction between bridging units increases, β drops in size.

-22-

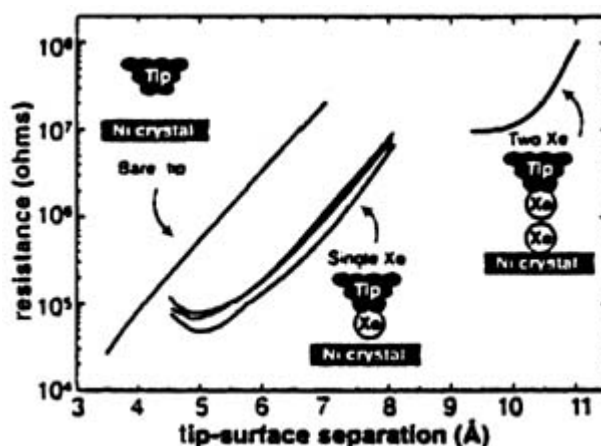


Figure C3.2.16. Dependence of measured resistance in an STM junction consisting of a ‘bare tip’ a tip with one Xe atom attached, and a tip with two Xe atoms. Note that the Xe atoms facilitate tunnelling (compared to empty space). From Yazdani A, Eigler D M and Lang N D 1996 Off resonance conduction through atomic wires *Science* **272** 1921–4.

An expression for the current across a molecular junction is developed by analogy with the description of unimolecular solution phase electron transfer. The conduction is written [20]

$$j = \frac{2\pi e}{\hbar} \sum_{i,f} f(E_i)[1 - f(E_f)]|T_{fi}|^2 \delta(E_i - E_f) \quad (\text{C3.2.16})$$

where f represents the fermi function and the sum is taken over states of the source electrode (i) and receiving electrode (f). When there are no molecular eigenstates of the molecule in the energy regime between the highest filled source electrode energy and lowest empty acceptor electrode, the conduction is limited by the tunnelling characteristics of the molecule. However, if eigenstates of the bridge fall in this gap, transport can involve a multistep hopping process. Indeed, as the voltage is varied and multiple eigenstates enter the gap, a molecular eigenstate staircase is seen in the current-applied voltage curves. Eigenstate staircases have been observed in carbon nanotube structures as well as in ‘break junctions’ bridged by molecules [21].

C3.2.4 FUTURE DIRECTIONS

We have surveyed the remarkable progress in the field of ET reactions, and have examined some of the key applications and successes of the theory. Many of the current frontiers of ET research lie in biological systems and in molecular-scale electronic devices.

C3.2.4.1 ENERGY FLOW INTO ET PROCESSES

The nitrogenase system reduces hundreds of millions of kilograms of nitrogen gas to ammonia each year, catalysing the reaction at ambient temperatures and atmospheric pressure. Nitrogenase consists of two proteins that contain

iron–sulfur and molybdenum–iron–sulfur clusters. Each (interprotein) electron transfer reaction is driven by the hydrolysis of two ATP molecules. The energy released as a consequence of ATP hydrolysis is about one-third of an electronvolt. A total of eight electrons must be delivered to the Mo–Fe protein to complete the catalytic cycle. It appears that the ATP consumed by nitrogenase drives substantial conformational changes that lead to protein–protein docking, an increase in ET driving force, and (following ET) protein–protein dissociation [53, 54]. Thus, ATP overcomes kinetic barriers to ET. This serves as an example of energy transduction from a storage species (ATP) to an ET process. The molecular mechanism of this process is the subject of great current interest. Equally challenging is to understand the way in which ET leads to transmembrane proton gradients, and how these gradients then drive the synthesis of ATP [3, 55] (figure C3.2.17)).

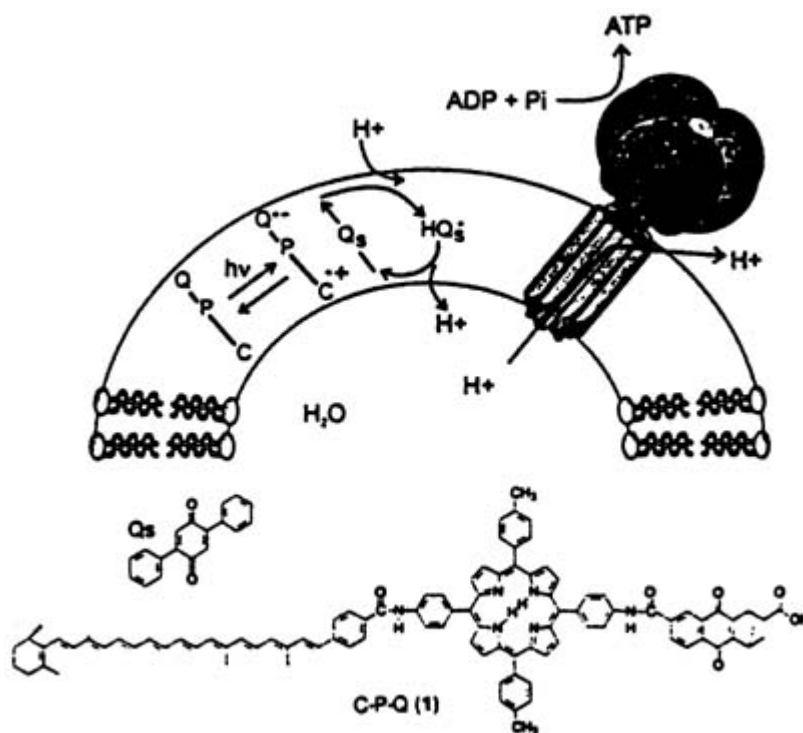


Figure C3.2.17. Diagram of a liposome-based artificial photosynthetic membrane showing the photocycle that pumps protons into the interior of the liposome and the CF_0F_1 -ATP synthase enzyme. From [55].

C3.2.4.2 PATHWAY FUNCTION

It is fairly clear that most biological electron transfer reactions fall in the weakly coupled regime (with a possible exception being the primary charge separation event in bacterial photosynthesis). However, there remains considerable debate concerning the functional significance of tunnelling pathways and of protein secondary and tertiary structure on the control of biological ET rates. The average decay constant for tunnelling through α -helix is predicted to be larger (rates drop more rapidly with distance) than for β -sheet structures. It was recently argued that this secondary structure effect is exploited to slow the rate of charge recombination in the photosynthetic reaction centre which is dominantly helical [56]. In contrast, the structure of cytochrome *c* oxidase appears to have direct beta strand-like pathways linking the redox centres to the oxygen reduction site. Since ET in this protein is thermal rather than

photochemical, and the reactions are exergonic, charge transfer recombination is not a competing reaction pathway.

With the emergence of increasingly high-resolution structural data for complex proteins, the influence of specific tunnelling pathways on protein function will be subjected to greater scrutiny. A particularly intriguing potential effect appears in the protein system that comprises the nitrogenase nitrogen fixing protein. When the two proteins dock to exchange electrons, it appears that there is a substantial structure change that ‘wires’ a tunnelling pathway. Following ET, the direct pathway is broken, electronically disconnecting the two redox centres. Since the protein dissociation reaction is particularly slow and the driving force of the reaction is not large, this disconnection could prove functionally important for localizing the electron at the proper site in the enzyme. Protein dynamics causes donor/acceptor interactions mediated by the protein to fluctuate. If fluctuations are sufficiently rapid, the root mean square coupling value should control the observed ET rate. A considerable open challenge is to understand what kinds of protein folds might have coupling interactions that are particularly sensitive to thermal fluctuations [57].

C3.2.4.3 INTERPRETING THE QUANTUM NATURE OF PROTEINS: REDUCED HAMILTONIANS AND CURRENT LOOPS

The complexity of protein structure motivates the development of new strategies for ‘information reduction’. One approach has been to devise hot and cold spot maps at the pathway level [26]. Other Hamiltonian based strategies have built ‘reduced’ Hamiltonians comprised of a subset of effective interactions that represent the interactions characteristics of the bridge [23] (figure C3.2.18). Density matrix approaches follow propagation of electron current density in two-level systems using equation C3.2.4, rather than tracking the wave function propagation. That is, the time domain rather than the energy domain quantum picture is analysed. Whereas electron amplitude decay and oscillation are seen in the wave function amplitude picture, vortices—closely associated with the nodal structure of the decaying wave function—appear in the current density maps [58].

-25-

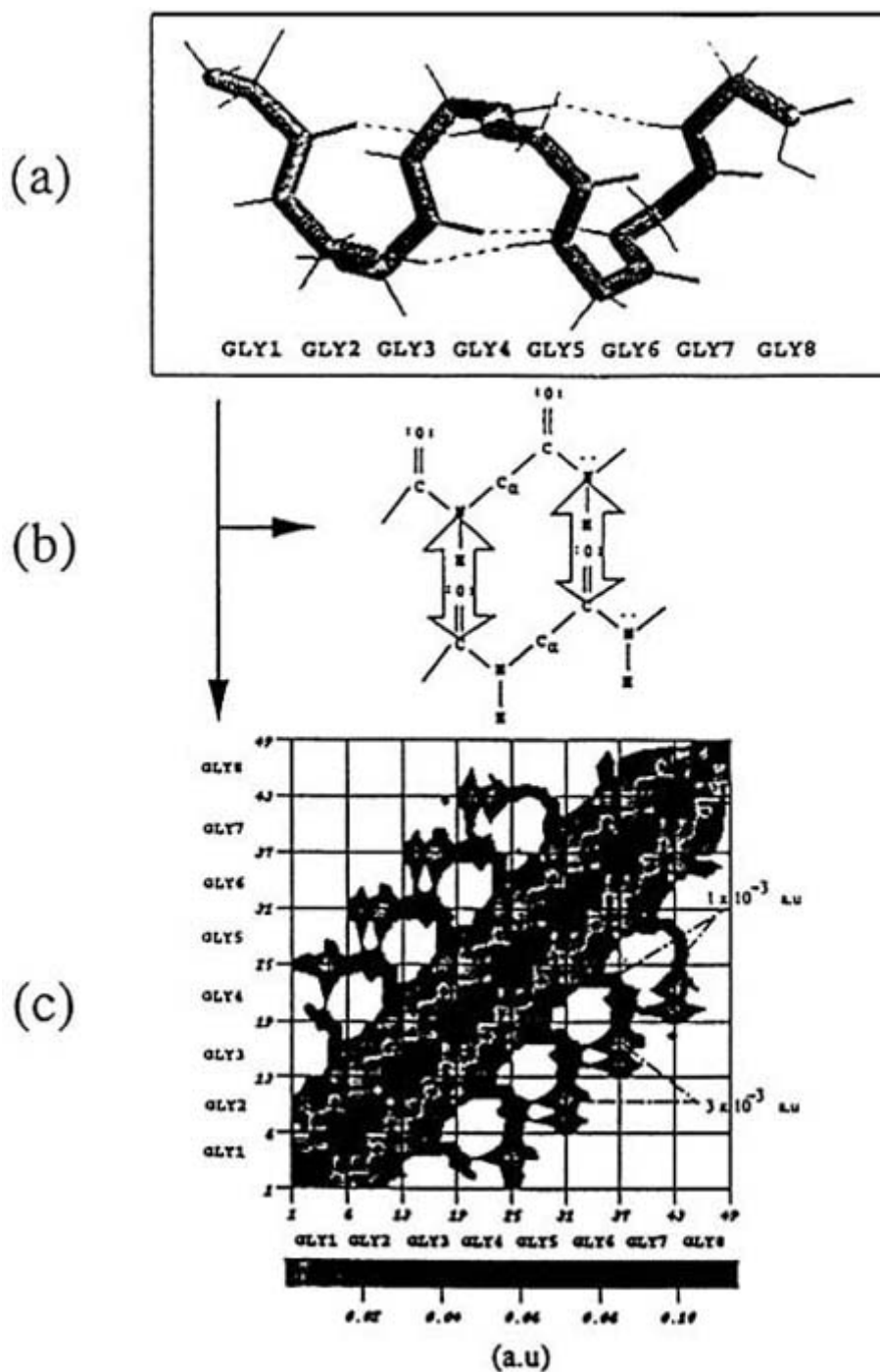


Figure C3.2.18.(a) Model α -helix, (b) hydrogen bonding contacts in the helix, and (c) schematic representation of the effective Hamiltonian interactions between atoms in the protein backbone. From [23].

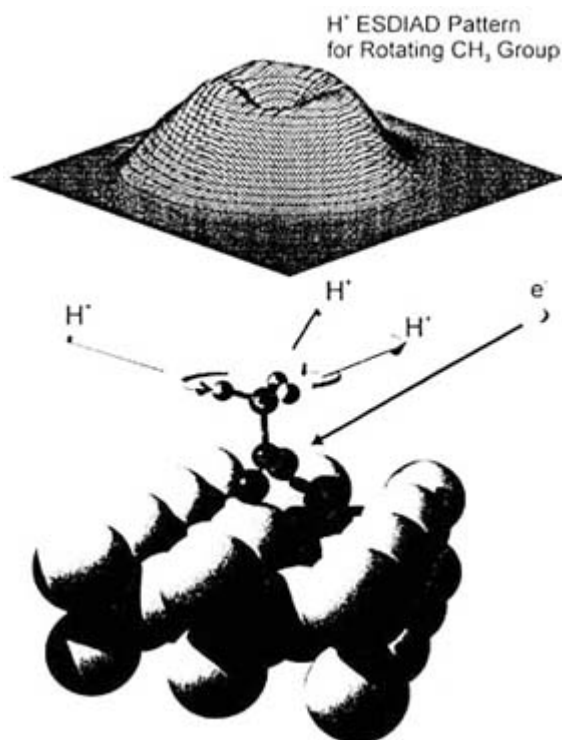


Figure C3.2.19. In this ESDIAD experiment where H^+ ions are produced and collected (see text), an adsorbed acetate species is excited by an incoming electron. H^+ ions are emitted in the direction of the C–H bond in the upward pointing $-CH_3$ group in the species. Circular symmetry of figure indicates that C–H bonds are spinning around the vertical axis in the acetate species. From Lee J G, Ahner J, Mocutt D, Denev S and Jates J T Jr 2000 *J. Chem. Phys.* **112** 335.

Not all processes that involve charge redistribution move charge between spatially well-localized regions. Electron scattering events fall into this regime and lie at the boundaries of the topics that we have discussed. Electron scattering processes are often used to practical advantage to probe the structure and dynamics of chemisorbed molecules, for example. One of these, ESDIAD (electron stimulated desorption ion angular distribution), invented in 1974 [59], may be used to observe the bond directions in chemisorbed species. This method uses electrons to ionize adsorbed molecules, making either positive or negative fragment ions. The ions escape from the molecule in directions closely similar to those of the chemical bonds that are being ruptured by the excitation event. By measuring the emission directions of the ion fragments, the characteristics of the ruptured chemical bond can be observed. In figure figure C3.2.19, an adsorbed acetate species is excited by an incoming electron [60]. H^+ ions are emitted in the direction of the C–H bond in the upward pointing $-CH_3$ group in the species. In the example shown here, millions of individual acetate species have been ionized, and the statistical distribution of the H^+ emission directions is shown by the volcano-shaped figure at the top. The circular symmetry of the figure indicates that the C–H bonds are spinning around the vertical axis in the acetate species, so that an almost equal probability of H^+ emission exists in all azimuthal directions. If the surface is cooled to very low temperatures, the rotation of the $-CH_3$ group ceases, and a multibeam H^+ pattern is observed. Measuring the temperature dependence of the beam pattern broadening into the volcano pattern allows one to measure the energy required to make the $-CH_3$ group spin. Such information is of importance in many technologies dependent upon molecular motions on surfaces, such as semiconductor device fabrication, corrosion inhibition, and heterogeneous catalysis.

REFERENCES

- [1] For recent reviews in this area, see
Bixon M and Jortner J (eds) 1999 Electron transfer-from isolated molecules to biomolecules, parts 1 and 2 *Adv. Chem. Phys.* **106** (parts A and B)
Chen P and Meyer T J 1998 Medium effects on charge transfer in metal complexes *Chem. Rev.* **98** 1439–78
Barbara P F, Meyer T J and Ratner M A 1996 Contemporary issues in electron transfer research *J. Phys. Chem.* **100** 13 148–68
Piotrowiak P 1999 Photoinduced electron transfer in molecular systems: recent developments *Chem. Soc. Rev.* **28** 143–50
- [2] Levine R D and Bernstein R B 1974 *Molecular Reaction Dynamics* (Oxford: Oxford University Press)
- [3] Cramer W A and Knaff D B 1990 *Energy Transduction in Biological Membranes* (New York: Springer)
- [4] Bendall D S (ed) 1996 *Protein Electron Transfer* (Oxford: Bios Scientific)
- [5] Marcus R A and Sutin N 1985 Electron transfers in chemistry and biology *Biochim. Biophys. Acta* **811** 265–322
- [6] Miller J R, Beitz J V and Huddleston R K 1984 Effect of free energy on rates of electron transfer between molecules *J. Am. Chem. Soc.* **106** 5057–68
- [7] Closs G L and Miller J R 1988 Intramolecular long distance electron transfer in organic molecules *Science* **240** 440–7
- [8] Fox M A and Chanon M (eds) 1988 *Photoinduced Electron Transfer* 4 vols (New York: Elsevier)
- [9] Jortner J and Bixon M (eds) 1999 *Adv. Chem. Phys.* **106** (parts I and II)
- [10] Gu Y, Akhremitchev B B, Walker G C and Waldeck D H 1999 Structural characterization and electron tunneling at n-Si/SiO₂/SAM/liquid interface *J. Phys. Chem. B* **103** 5220–6
- [11] Kittel C 1996 *Introduction to Solid State Physics* 7th edn (New York: Wiley)
- [12] Jortner J, Bixon M, Langenbacher T and Michel-Beyerle M E 1998 Charge transfer and transport in DNA *Proc. Natl Acad. Sci., USA* **95** 12 759–65
- [13] Henderson P T, Jones D, Hampikian G, Kan Y Z and Schuster G B 1999 Long-distance charge transport in duplex DNA: the phonon-assisted polaron-like hopping mechanism *Proc. Natl Acad. Sci., USA* **96** 8353–8
- [14] Ratner M 1999 Electronic motion in DNA *Nature* **397** 480–1
Lewis F D, Wu T F, Zhang Y F, Letsinger R L, Greenfield S R and Wasielewski M R 1997 *Science* **277** 673–6
Kelley S O and Barton J K 1999 Electron transfer between bases in double helical DNA *Science* **283** 375–81
Beratan D N, Priyadashy S and Risser S M 1997 *Chemistry and Biology* **4** 3–8
- [15] Meggers E, Michel-Beyerle M E and Giese B 1998 *J. Am. Chem. Soc.* **120** 12 950–5
- [16] Skotheim T A 1986 *Handbook of Conducting Polymers* vols 1 and 2 (New York: Dekker)
- [17] Walker G C, Maiti S, Cowen B R, Moser C C, Dutton P L and Hochstrasser R M 1994 Time resolution of electronic transitions of reaction centers in the infrared *J. Phys. Chem.* **98** 5778
- [18] Deisenhofer J, Epp O, Miki K, Huber R and Michel H 1984 X-ray structure analysis of a membrane–protein complex electron density map at 3 Å resolution and a model of the chromophores of the photosynthetic reaction center from *Rhodospseudomonas viridis* *J. Mol. Biol.* **180** 385–98
- [19] Schindelin N, Kisker C, Sehlessman J L, Howard J B and Rees D C 1997 Structure of ADP center dot AIF(4)(-)-stabilized nitrogenase complex and its implications for signal transduction *Nature* **387** 370–6
- [20] Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi H, Shinzawa-itoh K, Nakashima R, Yaono R and Yoshikawa S 1995 Structures of metal sites of oxidized bovine heart cytochrome c oxidase at 2.8 angstrom *Science* **269** 1069–74
-
- [21] Datta S 1997 *Electronic Transport in Mesoscopic Systems* (Cambridge: Cambridge University Press)
- [22] Ratner M and Jortner J 1997 *Molecular Electronics* (Malden, MA: Blackwell)
- [23] Skourtis S S and Beratan D N 1999 Theories of structure-function relationships for bridge-mediated electron transfer reactions *Adv. Chem. Phys.* **106** 377–452
- [24] Hopfield J J 1974 Electron transfer between biological molecules by thermally activated tunneling *Proc. Natl Acad.*

- [25] Beratan D N and Hopfield J J 1984 Calculation of electron tunneling matrix elements in rigid systems: mixed valence dithiaspirocyclobutane molecules *J. Am. Chem. Soc.* **106** 1584–94
- [26] Beratan D N, Betts J N and Onuchic J N 1991 Protein electron transfer rates set by the bridging secondary and tertiary structure *Science* **252** 1285–8
- [27] Kurnikov I V and Beratan D N 1996 *Ab initio* based effective Hamiltonians for long–range electron transfer: Hartree–Fock analysis *J. Chem. Phys.* **105** 9561–73
- [28] Langen R, Chang I-J, Germanas J P, Richards J H, Winkler J R and Gray H B 1995 *Science* **268** 1733
- [29] Gray H B and Winkler J R 1996 Electron transfer in proteins *Ann. Rev. Biochem.* **65** 537–61
- [30] DeRege P J F, Williams S A and Therien M J 1995 Direct evaluation of electronic coupling mediated by hydrogen bonds—implications for biological electron transfer *Science* **269** 1409–13
- [31] Jordan K D and Paddon-Row M N 1992 Long range interactions in a series of rigid nonconjugated dienes *J. Phys. Chem.* **96** 1188–96
- [32] Newton M D 1999 Electron transfer from isolated molecules to biomolecules *Advanced Chemical Physics* vol 106, ed J Jortner and M Bixon (New York: Wiley) pp 303–75
- [33] Gould I R, Ege D, Mattes S L and Farid S 1987 Return electron transfer with geminate radical ion pairs – observation of the Marcus inverted region *J. Am. Chem. Soc.* **109** 3794–6
- [34] Marcus R A and Sutin N 1985 *Biochim. Biophys. Acta* **811** 265
- [35] Ulstrup J 1979 *Charge Transfer Processes in Condensed Media* (Berlin: Springer)
- [36] Hush N S 1967 Intervalence-transfer absorption. Part 2. Theoretical considerations and spectroscopic data *Prog. Inorg. Chem.* **8** 391
- [37] Elliott C M, Derr D L, Matyushov D V and Newton M D 1998 Direct experimental comparison of the theories of thermal and optical electron-transfer: studies of a mixed-valence dinuclear iron polypyridyl complex *J. Am. Chem. Soc.* **120** 11 714–26
- [38] Chen P and Meyer T J 1998 Medium effects on charge transfer in metal complexes *Chem. Rev.* **98** 1439–78
- [39] Creutz C and Taube H 1973 Binuclear complexes of ruthenium amines *J. Am. Chem. Soc.* **95** 1086
- [40] Zusman L D 1980 *Chem. Phys.* **49** 295
- [41] Kosower E M and Huppert D 1983 Solvent motion controls the rate of intramolecular electron transfer *Chem. Phys. Lett.* **96** 433–5
- [42] Walker G C, Barbara P F, Doorn S K, Dong Y and Hupp J T 1991 Ultrafast measurements on direct photoinduced electron transfer in a mixed-valence complex *J. Phys. Chem.* **95** 5712
- [43] Wang C, Mohney B K, Williams R, Hupp J T and Walker G C 1998 Solvent control of vibronic coupling upon intervalence charge transfer excitation of $(\text{NC})_5\text{FeCNRu}(\text{NH}_3)_5^-$ as revealed by resonance Raman and near-infrared absorption spectroscopies *J. Am. Chem. Soc.* **120** 5848–9
- [44] Spears K G, Wen X and Zhang R 1996 Electron transfer rates from vibrational quantum states *J. Phys. Chem.* **100** 10 206–9
- [45] Barbara P F, Walker G C and Smith T P 1992 Vibrational modes and the dynamic solvent effect in electron and proton transfer *Science* **256** 975–81

- [46] Wang C, Akhremitchev B B and Walker G C 1997 Fem to second infrared and visible spectroscopy of photoinduced intermolecular electron transfer dynamics and solvent–solute reaction geometries: coumarin 337 in dimethylaniline *J. Phys. Chem. A* **101** 2735–8
- [47] Vos M H, Jones M R, Hunter C N, Breton J, Lambry J C and Martin J L 1996 Femtosecond spectroscopy and vibrational coherence of membrane-bound RCs of *Rhodobacter sphaeroides* genetically modified at positions M210 and L181 *The Reaction Center of Photosynthetic Bacteria—Structure and Dynamics* ed M E Michel-Beyerle (Berlin: Springer) pp 271–80
- [48] Hayashi M, Yang T-S, Yu J, Mebel A, Chang R, Lin S H, Rubtsov I V and Yoshihara K 1998 Vibronic and vibrational coherence and relaxation dynamics in the TCNE–HMB complex *J. Phys. Chem. A* **102** 4256–65
- [49] Jean J M, Fleming G and Friesner R 1992 Application of a multilevel redfield theory to electron transfer in condensed

phases *J. Chem. Phys.* **96** 5827

- [50] Wynne K and Hochstrasser R M 1999 Coherence and adiabaticity in ultrafast electron transfer *Adv. Chem. Phys.* **107** (Electron transfer from isolated molecules to biomolecules) part 2, 263–309
- [51] Evans D G, Coalson R D, Kim H J and Dakhnovskii Y 1995 Inducing coherent oscillations in an electron transfer dynamics of a strongly dissipative system with pulsed monochromatic light *Phys. Rev. Lett.* **75** 3649
- [52] See minireviews in *J. Biol. Inorg. Chem.* 1998 **3**
- [53] Howard J B and Rees D C 1996 Structural basis of biological nitrogen fixation *Chem. Rev.* **96** 2965–82
- [54] Seefeldt L C and Dean D R 1997 Role of nucleotides in nitrogenase catalysis *Acc. Chem. Res.* **30** 260–6
- [55] Steinberg-Yfrach G, Rigaud G-L, Durantini E N, Moore A L, Gust D and Moore T A 1998 Light driven production of ATP catalysed by F0F1–ATP synthase in an artificial photosynthetic membrane *Nature* **392** 479–82
- [56] Ramirez B E, Malmstrom B G, Winkler J R and Gray H B 1995 The currents of life: the terminal electron-transfer complex of respiration *Proc. Natl Acad. Sci., USA* **92** 11 949–51
- [57] Balabin I A and Onuchic J N 1998 A new framework for electron-transfer calculations—beyond the pathways-like models *J. Phys. Chem. B* **10** 7497–505
- [58] Daizadeh I, Guo J-X and Stuchebrukhov A 1999 Vortex structure of the tunneling flow in long-range electron transfer reactions *J. Chem. Phys.* **110** 8865–8
- [59] Czyzewski J J, Madey T E and Yates J T Jr 1974 *Phys. Rev. Lett.* **32** 777
- [60] Lee J-G, Ahner J, Mocutta D, Denev S and Yates J T Jr 2000 Thermal excitation of rotation of methyl group in chemisorbed acetate on Cu(1 1 0) *J. Chem. Phys.* **112** 3351
-

-1-

C3.3 Energy transfer in gases

George W Flynn

C3.3.1 INTRODUCTION

Almost all aspects of the field of chemistry involve the flow of energy either within or between molecules. Indeed, the occurrence of a chemical reaction between two species implies the availability of some minimum amount of energy in the reacting system. The study of energy transfer processes is thus a topic of fundamental importance in chemistry. Energy transfer in gases is of particular interest partly because very sophisticated methods have been developed to study such events and partly because gas phase processes lend themselves to very complete and detailed theoretical analysis.

In the gas phase molecules are generally separated by distances large compared to their diameters, and energy exchange between two different molecules occurs only when they collide, much like billiard balls on a pool table, or two trucks in a road accident such as that depicted in [figure C3.3.1](#). Nevertheless, despite these large separations, gas molecules collide at a very high rate. The formula for the collision rate, z , of a single molecule is [1]

$$z = \pi \sigma^2 \langle u \rangle n$$

where σ is the molecular diameter, $\langle u \rangle$ the mean speed and n the density of molecules. To get some feeling for the size of z , try to imagine a single nitrogen molecule in the atmosphere right in front of your nose. If you blink your eye, that molecule undergoes more than a billion collisions in the time it takes for your eye to open and close! Modern techniques for the study of such processes generally focus on single collision events. This would require a time resolution of better than a billionth of a second to investigate a nitrogen molecule in the atmosphere. Fortunately, for gases, the density n can be easily controlled thereby reducing z substantially. For the collisional energy transfer processes described in this chapter, pressures of 10^{-5} atmospheres are typically used, thereby

reducing the mean time between collisions to an experimentally manageable time scale of a few hundred thousandths of a second.

While collisions between atoms are very simple, since only translational motion is of importance, collisions between molecules are more complicated because of the internal degrees of freedom. For example, in a relatively simple collision between an argon atom and a linear CO₂ molecule, the atom has three translational degrees of freedom while the molecule has three translational, two rotational and four vibrational degrees of freedom. In principle, energy exchange among all of these degrees of freedom must be accounted for in any complete description of such a collision. Loss or gain of vibrational energy in molecular collisions is of special interest in chemistry because all reactions ultimately involve the breaking and making of chemical bonds, a process that is accelerated by putting energy into the vibrational degrees of freedom of a molecule.

-2-



Figure C3.3.1 A collision between a milk truck and a bread truck showing the well ordered truck contents at the top, the ‘scattering event’ in the middle and the post crash scrambling of the truck contents at the bottom.

C3.3.1.1 UNIMOLECULAR REACTIONS AND ENERGY TRANSFER

Of particular importance is loss of energy from molecules with ‘chemically significant’ amounts of vibrational energy. These are systems in which the molecule has sufficient energy to rupture a chemical bond. Chemical reactions of such highly vibrationally excited molecules, which normally take place on the ground electronic state potential energy surface, are often described by the Lindemann unimolecular reaction mechanism in which a substrate, S, is excited by collisions to S*, a level with energy sufficient to cause bond breaking or molecular rearrangement [2, 3 and 4]. S* is thus said to have a ‘chemically significant’ amount of energy. For large molecules, the time scale for decomposition of S* is sufficiently long that further collisions with the bath molecules can cause deactivation of the excited substrate, thus quenching the reaction process. The overall mechanism can be summarized by the equations





where B is a generalized representation for a bath molecule and P is the product, chemically distinct from S. Once the

-3-

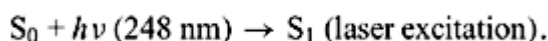
molecule is excited in step (C3.3.1), the rate of production of product is a competition between unimolecular breakup of S* in step (C3.3.2) and quenching in step (C3.3.3). The importance of vibrational energy loss in step (C3.3.3) is therefore paramount in determining the overall efficiency with which S is converted to product via this mechanism. It would be difficult to exaggerate the importance of this quenching step, since an enormous number of thermal chemical reactions proceed via this mechanism or some variant of it.

One of the complexities that arises in studying the vibrational energy loss from highly excited molecules is the very, very large number of vibrational states (e.g. 10^{15} vibrational states per cm^{-1}) in molecules of even moderate size at chemically significant energies (100–400 kJ mol^{-1}). Because of this, directly probing the vibrational states of S*, with even the highest resolution laser devices, is essentially impossible. Nevertheless, such collisions can be monitored in great detail by using a simple trick provided that the bath molecule B is relatively small. This trick amounts to realizing that the collision of S* with B can be viewed through the ‘eyes’ of the bath molecule B [5, 6]. When B is a small molecule with well resolved and assigned vibrational and rotational spectroscopic transitions, more information about the quenching process (C3.3.3) can be obtained from probing the bath B than from probing the donor S*. If we return for a moment to the collision between the bread truck and the milk truck of figure C3.3.1 a typical approach for the police to use in reconstructing such an accident is to take pictures of the post-collision scene to establish the speed and position of the two trucks before the initial ‘scattering event’ occurred. The more detail available in the post-collision picture, the better the chance of accurately reconstructing the collision event. In principle, the condition and position of either truck is sufficient to establish most of the details of the collision.

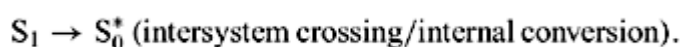
C3.3.2 EXPERIMENTAL APPROACH

C3.3.2.1 GENERAL SCHEME

Experiments of the type described above rely on the availability of both high resolution and very intense laser sources throughout the ultraviolet, visible and infrared spectral ranges [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17 and 18]. Figure C3.3.2 shows an energy level diagram and excitation scheme for a typical molecule. A short pulse of ultraviolet light from a laser excites a medium size molecule such as benzene or hexafluorobenzene to its first electronically excited singlet state,



The natural processes of intersystem crossing and internal conversion will quickly (e.g. 50 ns) carry the molecule from this excited electronic surface to the ground electronic surface *without a collision*,



The result is the preparation of molecules with a well defined energy E that is essentially pure vibrational energy in S_0 , the ground electronic state. (Laser excitation adds only $h/2\pi$ to the total angular momentum of the molecule and produces no increase in the translational energy.) This highly excited molecule S_0^* of energy E is our donor with chemically significant amounts of vibrational energy. Preparation of donor species in this way provides molecules

with

-4-

very well characterized energy and sets the origin of time at the laser pulse so that subsequent collisions can be studied if they occur on a time scale long compared to the laser pulse width and the molecular internal-conversion/intersystem-crossing time. Typical laser pulse widths are 10–30 ns while collision times can be controlled with pressure to take place on a time scale of several microseconds. (At a pressure of 20 mTorr, the mean time between collisions is about 4 μ s.)

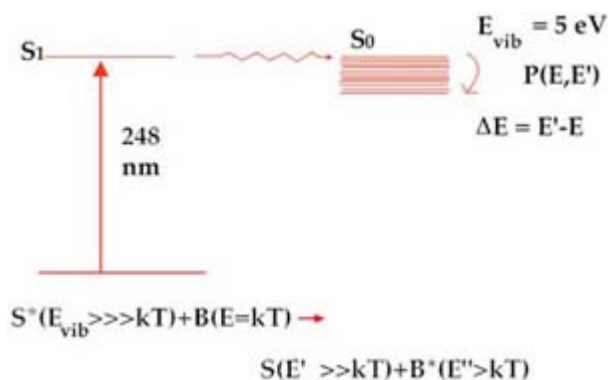
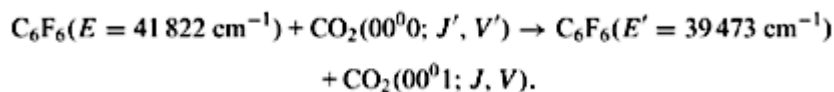


Figure C3.3.2. A simple diagram showing a clean method for preparing molecules with a specific and large amount of vibrational energy $E_{vib} = 5 \text{ eV}$ (roughly 460 kJ mol^{-1}). A pulse from an excimer laser excites an allowed ($S_0 \rightarrow S_1$) electronic transition in the molecule. Intersystem crossing and internal conversion, occurring collision-free in the molecule, carry it onto the ground electronic surface. Collisions of the hot donor molecule, S^* , with bath molecules, B , depicted by the equation at the bottom of the figure, cause the molecule to lose energy with a probability $P(E, E')$, where $\Delta E = E' - E$.

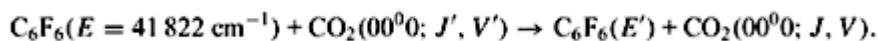
Once prepared in S^*_0 with well defined energy E , donor molecules will begin to collide with bath molecules B at a rate determined by the bath-gas pressure. A typical process of this type is the collision between a C_6F_6 molecule with approximately 5 eV ($40\,000 \text{ cm}^{-1}$ or 460 kJ mol^{-1}) of internal vibrational energy and a CO_2 molecule in its ground vibrationless state 00^0_0 to produce CO_2 in the first asymmetric stretch vibrational level 00^0_1 [11, 12 and 13]. This collision results in the loss of approximately $\Delta E = 2349 \text{ cm}^{-1}$ of internal energy from the C_6F_6 ,



J and V represent the rotational angular momentum quantum number and the velocity of the CO_2 , respectively. The hot, excited C_6F_6 donor can be produced via absorption of a 248 nm excimer-laser pulse followed by rapid internal conversion of electronic energy to vibrational energy as described above. Note that the result of this collision is to produce one quantum of vibrational energy in the CO_2 00^0_1 state with a corresponding loss of the same amount of energy from the internal vibrational degrees of freedom of the donor.

Actually, collisions in which the bath becomes vibrationally excited are relatively rare, occurring with a typical probability of 1% per gas-kinetic collision [6, 8, 11 and 13]. More common are processes that produce rotational and translational excitation in the bath acceptor while leaving the molecule in its ground (vibrationless) 00^0_0 state,

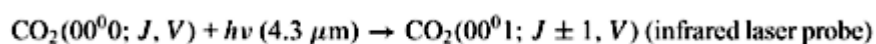
-5-



Here the energy loss $\Delta E = E' - E$ from the donor shows up as translational energy through a change in the bath molecule velocity (V' going to V) and as rotational energy through a change in the rotational-angular-momentum quantum number (J' going to J).

C3.3.2.2 INFRARED DIODE LASER PROBING

To ‘view’ these collisions through the eyes of the bath acceptor molecule, we need only to probe the $\text{CO}_2(00^00; J, V)$ molecules exiting the collision. Much like the police visiting the post-collision-accident scene ([figure C3.3.1](#)) we ‘take a picture’ of the $\text{CO}_2(00^00; J, V)$ with a high resolution camera. Our camera is an infrared laser with sufficient resolution to identify all of the quantum numbers for molecular vibration of CO_2 , the rotational angular momentum quantum number J and the recoil velocity V . This probe, or picture-taking, step can be represented by the equation



where light from a narrow-band-infrared-diode laser (linewidth of 0.0003 cm^{-1}) operating at a wavelength of approximately $4.3\ \mu\text{m}$ is used to sense the arrival of molecules in the $00^00; J$ state by observing or probing the fully allowed infrared transition, $00^00; J \rightarrow 00^01; J \pm 1$, in CO_2 . Such is the resolution of these laser devices that essentially any state of the CO_2 bath molecule produced in the collision process can be probed and the population of each quantum state measured.

Even more remarkable is the fact that these infrared diode lasers have sufficient resolution to measure the Doppler lineshape for the $00^00; J \rightarrow 00^01; J \pm 1$ transition of the recoiling molecules, hence providing a probe of the CO_2 molecular recoil velocity, V . The absorption frequency for a molecule moving with a component of velocity V_z parallel to the direction of propagation of the infrared laser beam is $\nu = \nu_0(1 \pm V_z/c)$, where ν_0 is the absorption frequency for the molecule at rest and c is the speed of light. The \pm sign determines whether the molecule is travelling in the same or opposite direction as the light beam. For an isotropic distribution of molecules whose translational motion is at equilibrium at a temperature T , there is a range of V_z given by the Boltzmann distribution. The corresponding spread of absorption frequencies arising from this velocity distribution gives rise to an inhomogeneously broadened spectral line shape that can be described by a Gaussian function whose full width at half height is given by [19]

$$\Delta\nu = 2(3.581 \times 10^{-7})\nu_0(T/M)^{1/2}$$

where M is the molecular weight of the molecule, ν_0 the absorption frequency for a molecule at rest and T the absolute temperature. [Figure C3.3.3](#) shows the relative Doppler lineshapes for a CO_2 molecule at a temperature of 300 K corresponding to a moderate velocity spread along the laser probe direction, and 3000 K corresponding to a rather large velocity spread along the laser probe direction. Also shown for comparison is the laser line width that is much narrower than even the 300 K Doppler line shape for room-temperature CO_2 . For an experiment conducted at room temperature, bath molecules start with a 300 K Doppler lineshape corresponding to a Boltzmann velocity distribution at this temperature, but collisions with hot donors produce recoiling bath molecules that have velocities comparable to those typical for temperatures in the 1500–4000 K range. [Figure C3.3.3](#) indicates that such linewidths can easily be measured with modern, commercially available, high-resolution infrared lasers. While there is no *a priori* reason to

expect that molecules scattered by collisional energy transfer will have an isotropic Boltzmann distribution of velocities, experiments performed under simple gas-bulb conditions with molecules initially thermalized at

temperatures in the 200–400 K range have so far found that the spectral absorption lineshapes for recoiling molecules in these studies are Gaussian within experimental error and, therefore, can be characterized by a temperature T [5, 9, 12, 16].

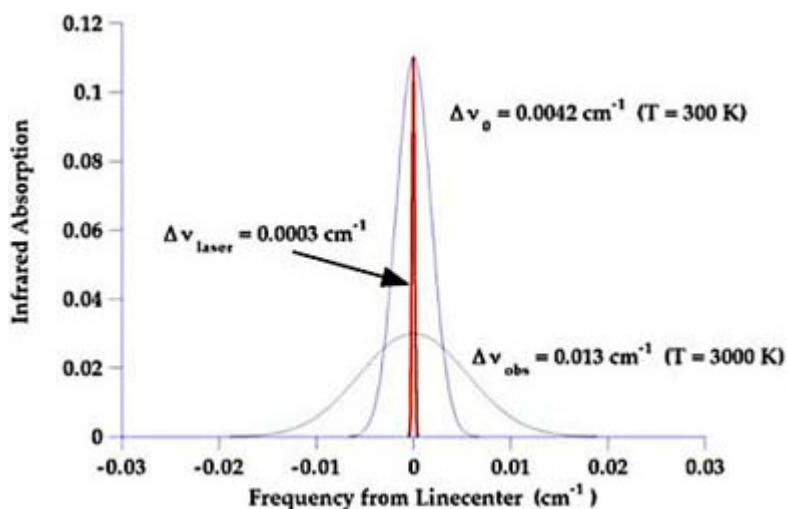


Figure C3.3.3. A schematic drawing of the Doppler-lineshape profile for a typical infrared transition in a small molecule at $T = 300$ K and $T = 3000$ K. The width of the absorption profile scales as $T^{1/2}$ reflecting the thermal spread and isotropic distribution of molecular velocities. Such profiles are easily measured using a high-resolution infrared diode laser whose typical line profile, having a width of 0.0003 cm^{-1} , is also shown in the figure.

C3.3.2.3 EXPERIMENTAL APPARATUS

Figure C3.3.4 shows a schematic diagram of an apparatus that can be used to study collisions of the type described above [5, 9, 12, 16]. Donor molecules in a 3 m long collision cell (a cylindrical tube) are excited along the axis of the cell by a short-pulse excimer laser (typically 25 ns pulse width operating at 248 nm), and bath molecules are probed along this same axis by an infrared diode laser (wavelength in the mid-infrared with continuous light-output power of $100\text{ }\mu\text{W}$ in a bandwidth of 0.0003 cm^{-1}). The pump and probe beams are joined in front of the collision cell with a dichroic beamsplitter coated to reflect the ultraviolet laser light and pass the infrared laser light. The beams propagate collinearly along the cell axis. At the end of the cell the ultraviolet beam is discarded and the infrared beam passes through a monochromator to select the appropriate diode laser mode. About 10% of the infrared laser light is split from the main beam and sent through a reference cell, a scanning etalon, and a monochromator. This fraction of the light beam is detected with an InSb infrared detector and fed to a lock-in amplifier. The output of the lock-in amplifier can be used as an error signal and fed back to the diode-laser-control electronics to lock the diode either to a specific spectral reference line or to an etalon fringe. This reference loop is used to identify the exact frequency of the diode laser and to sweep the laser over small ranges (for Doppler profile measurements) by scanning the frequency of the etalon to which the laser can be locked. The reference cell also has electrodes for exciting a discharge in the reference gas in cases where the frequencies of interest originate in vibrational levels not populated at the ambient temperature of the cell.

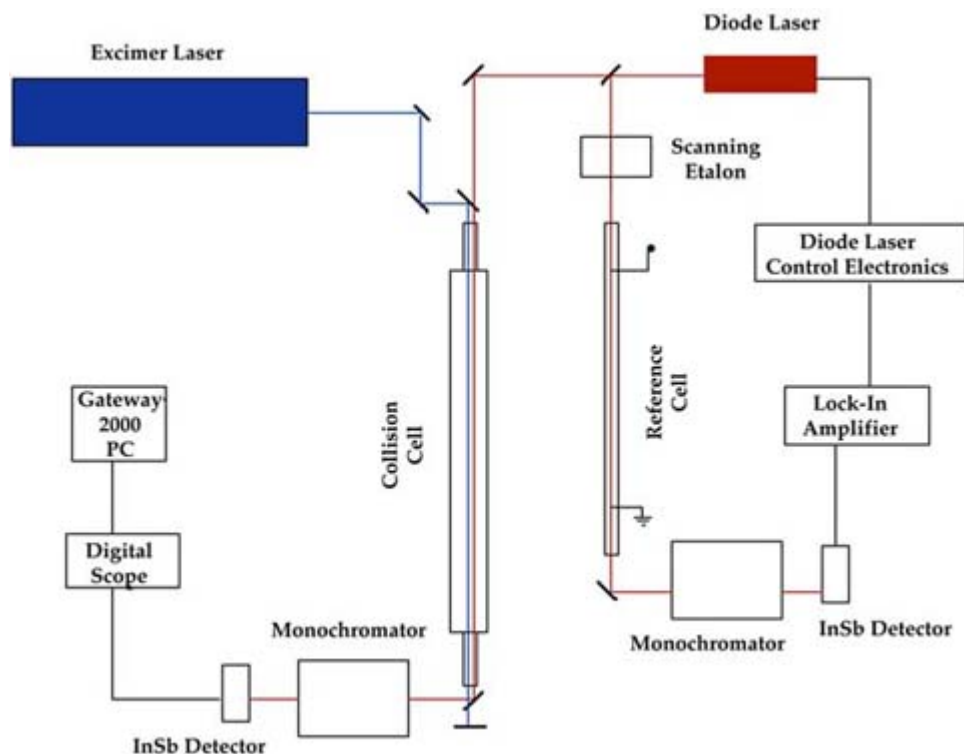


Figure C3.3.4. A schematic diagram of an apparatus described in the text for studying vibrational energy transfer to small bath molecules from donor molecules having chemically significant amounts of internal vibrational energy.

The infrared light passing through the collision cell impinges on a second InSb-solid-state detector (cooled to 77 K) producing both DC and AC signals at the detector output. Since most of the IR light is not absorbed by the sample, the DC signal that measures the continuous output level of laser light is much larger than the AC signal. It is in fact fluctuations in the DC light level that constitute one of the main noise sources in the experiment. Both AC and DC signals are fed to a high-speed transient recorder with at least two channels where the time-resolved ratio of the AC and DC currents is recorded and stored in memory. Single-collision data are obtained from this time-dependent absorption data. Signals from a series of ultraviolet laser pulses can be added in memory with subsequent signal averaging and noise reduction. The ratio of the AC and DC infrared light levels, $\Delta I/I$, is related to the pressure of absorbing molecules, P , the molecular absorption coefficient, α and the cell path length, L :

$$\ln[(\Delta I/I) + 1] = \alpha PL.$$

The path length is set by the experimental configuration while α is known for each transition (such as $00^0_0: J \rightarrow 00^0_1, J \pm 1$ or $00^0_1; J \rightarrow 00^0_2, J \pm 1$). Thus, a measurement of $\Delta I/I$ provides the partial pressure P of molecules produced in probed states such as $00^0_0; J$ or $00^0_1; J$. (Strictly, optical probing measures the *difference* in the partial pressures between the upper and lower states of the probed transition; however, in practice, the lower state population is always much larger than the upper state population so that the probe senses only the lower state population in the experiment.)

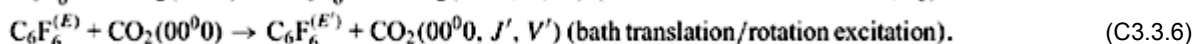
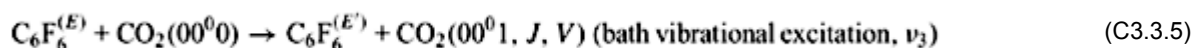
C3.3.3 DATA ANALYSIS

C3.3.3.1 KINETIC PROCESSES

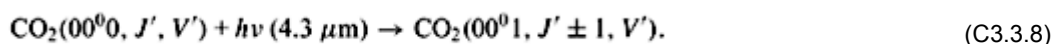
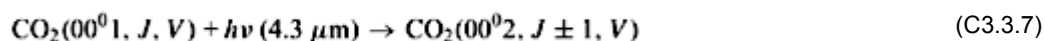
As a first step in understanding the analysis of energy transfer experiments, it is worthwhile to summarize the steps in a typical experiment where C_6F_6 is the hot donor and carbon dioxide is the bath receptor molecule. First, excited C_6F_6 molecules ($C_6F_6^{(E)}$) are produced at energy $E = 41\,822\text{ cm}^{-1}$ by an excimer laser pulse (25 ns),



The C_6F_6 electronic state excited in this energy region rapidly interconverts (collision-free), on a time scale of less than 30 ns, to high vibrational energy levels of the electronic ground state. Collisions of this hot donor with CO_2 ‘inert’ bath gas then cause translational, rotational and vibrational excitation of the ν_3 stretching (00^01 ; 2349 cm^{-1}) vibrational state, as well as rotational and translational excitation in the ground vibrationless (00^00) level,



A tunable diode laser operating at $4.3\text{ }\mu\text{m}$ is used to probe the P and/or R branch bands of the following transitions,



Velocity recoils are measured at short times after the initial ultraviolet excitation pulse by probing the ‘nascent’ Doppler profiles for the different spectral lines probed in these last steps.

C3.3.3.2 INITIAL RATES TECHNIQUE

Using the above equations, the rate equation for production of bath molecules in a given quantum state due to collisions with a hot donor molecule can be written (e.g. for equation (c3.3.5))

$$d[CO_2(00^01, J, V)]/dt = k_r[C_6F_6^{(E)}][CO_2(00^00)]$$

where k_r is the rate constant for production of $CO_2(00^01, J, V)$ from collisions between $C_6F_6^{(E)}$ molecules excited to an energy E and initially unexcited $CO_2(00^00)$ bath molecules. For times t after the excimer laser excitation pulse that are short compared to the mean collision time in the gas, this equation can be solved to a good degree of approximation by using the initial rate technique with

$$[CO_2(00^01, J, V)] = k_r[C_6F_6^{(E)}]_0[CO_2(00^00)]t.$$

Here $[C_6F_6^{(E)}]_0$ is the initial concentration of excited donor molecules produced at time $t = 0$ by the excimer laser pulse, and $[CO_2(00^00)]$ is the concentration of bath molecules that, for all practical purposes, can be assumed constant. The infrared diode laser probe provides an experimental value for $[CO_2(00^01, J, V)]$, while the number of absorbed excimer laser photons provides a measure of $[C_6F_6^{(E)}]_0$. $[CO_2(00^00)]$ is known from the partial pressure of the bath gas in the cell and the time t is easily determined from the transient recorder time base, leaving only k_r to be determined from these experimentally measured parameters. Experiments of this kind provide three important pieces of data: (1) the distribution of populations in the final bath vibration/rotation state ($00^01, J, V$); (2) the distribution of recoil velocities, V , from a measure of the Doppler lineshape for the carbon dioxide spectral transition ($00^01, J, V) + h\nu(4.3\text{ }\mu\text{m}) \rightarrow (00^02, J \pm 1, V)$ and (3) the rate constant, k_r , or probability that a collision produces a CO_2 molecule with a given final J and V in the 00^01 vibrational state.

The initially excited C_6F_6 molecules can produce de-excited species, such as $C_6F_6(E')$, that are also able to excite CO_2 via collisions. This fact emphasizes the importance of choosing t short enough that a given bath molecule has time for only a single collision with an excited donor and does not collide a second time with either another hot donor or a bath molecule. Collisions with other bath molecules tend to relax the ‘nascent’ population and velocity distributions formed initially by collisions with the hot donor, while second collisions with excited donors produce ‘doubly’ excited bath molecules. In contrast, rotational and vibrational-state-population distributions, velocity distributions and k_T values measured at a time t after the excimer laser pulse corresponding to 0.1–0.25 of a gas kinetic collision provide data specific to a collision of a hot donor of energy E with a cold bath molecule. Thus, in experiments of this type, it is possible to correlate the donor energy, its vibrational density of states and other molecular properties with excitation probabilities, population distributions and the kinetic energy distributions of the scattered bath species.

C3.3.4 DEDUCING ENERGY TRANSFER MECHANISMS FROM POPULATION AND VELOCITY DISTRIBUTIONS OF THE SCATTERED BATH MOLECULES’ ROTATIONAL STATE POPULATION DISTRIBUTIONS FOR VIBRATIONAL EXCITATION OF THE BATH

Figure C3.3.5 shows typical data obtained from experimental studies of the type described above, where the hot donor is the nitrogen heterocycle pyrazine, $C_4H_4N_2$, initially excited by an excimer laser to an energy of $40\,640\text{ cm}^{-1}$. Here the process probed is excitation of a vibrationally-excited bath state where all three degrees of freedom of the bath—vibration, rotation, and translation—can become excited. In this particular case the vibrational state excited by collision is the first asymmetric stretch level of CO_2 , 00^0_1 that has 2349 cm^{-1} of vibrational energy, roughly ten times the mean thermal energy in these experiments ($kT = 208\text{ cm}^{-1}$, where k is Boltzmann’s constant). Shown in the upper half of the figure is a ‘Boltzmann plot’ of the natural log of the measured rotational state populations for just the 00^0_1 level, divided by their degeneracy, $2J + 1$, versus $J(J + 1)$ that is proportional to the molecular rotational energy. The slope of such plots ($-1/kT_R$) gives the temperature (T_R) describing the rotational state distribution for a system at equilibrium at temperature T_R [6, 9, 10 and 11, 13, 18]. There are two remarkable things about this figure. First, the rotational state population distribution does give a straight-line ‘Boltzmann plot’ suggesting that the CO_2 molecules, scattered into this excited vibrational level by collisions with vibrationally hot pyrazine molecules, have a ‘pseudo-equilibrium’ distribution. Second, and far more amazing, is that the temperature, T_R , deduced from the slope of this plot is only $383 \pm 40\text{ K}$, just slightly warmer than the initial, ambient cell temperature of 298 K !

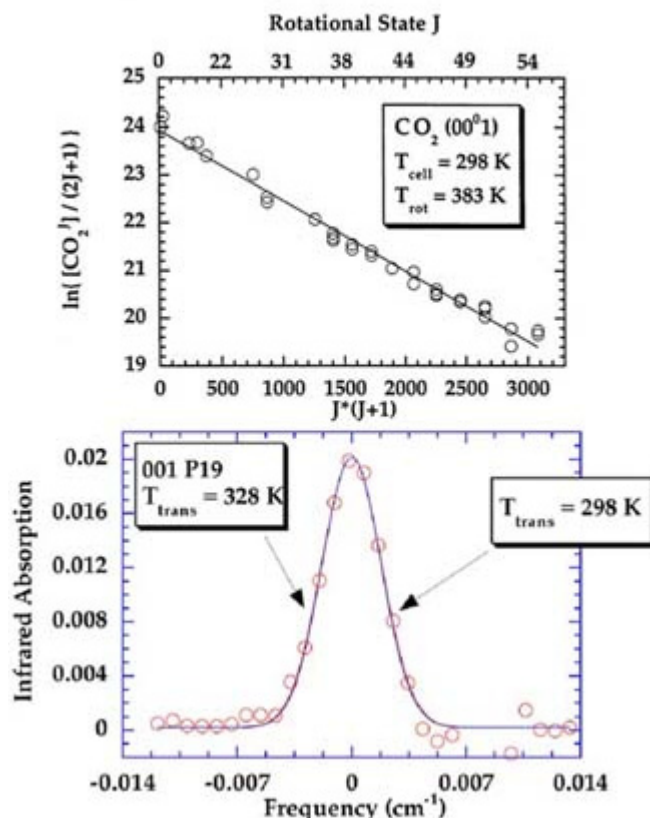
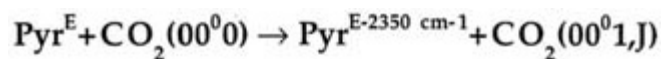


Figure C3.3.5. The upper half of the figure is a ‘Boltzmann’ plot of the natural log of the population scattered into the $\text{CO}_2(00^0_1; J)$ vibration–rotation level divided by $2J + 1$, for collisions with an excited pyrazine molecule as depicted by the equation at the top of the figure. The slope of such a plot is $1/kT_R$ where k is Boltzmann’s constant and T_R is the temperature that characterizes the rotational-state distribution in the $\text{CO}_2(00^0_1)$ vibrational level. Shown in the lower half of the figure is a Doppler-lineshape profile for the $\text{CO}_2(00^0_1; J = 19) \rightarrow \text{CO}_2(00^0_2; J = 18)$ transition, where the molecules in $\text{CO}_2(00^0_1; J = 19)$ have been excited by the collision process depicted at the top of the figure. The best fit of the data (circles) to a Gaussian profile gives a translational temperature of 328 K, indistinguishable in this case from the ambient 298 K Doppler profile.

C3.3.4.1 VELOCITY PROFILES FOR VIBRATIONAL EXCITATION OF THE BATH

In the lower half of figure C3.3.5 is a plot of the spectral lineshape for the transition $00^0_1; J = 19 \rightarrow 00^0_2; J = 18$ that provides a measure of the recoil velocity distribution for molecules scattered into the $00^0_1; J = 19$ state. The width of this distribution is characterized by a temperature of $328 \pm 30 \text{ K}$ that is again only slightly larger than the 298 K gas temperature of the pre-excited molecules. These collisions have managed to insert into the bath acceptor molecules an energy equivalent to more than ten times the mean ambient energy of the initial molecular ensemble (before excimer laser excitation) without significantly exciting either the translational or rotational degrees of freedom of the molecule [6, 9, 10 and 11, 13, 18]!

In fact, the two observations represented by the upper and lower halves of [figure C3.3.5](#) paint a remarkably consistent picture of the physical process that leads to vibrational excitation of the bath molecules in collisions with molecules having chemically significant amounts of vibrational energy. Such collisions must be ‘soft’, taking place at some distance short of the repulsive potential wall for interaction of the two molecules. Note that an impulsive collision sampling the steep repulsive wall of the intermolecular potential of a highly vibrationally excited molecule with its rapidly oscillating ‘atomic pistons’ would perforce be kicked rather hard, thereby developing substantial translational recoil. Furthermore, since CO_2 is cigar shaped, it is nearly impossible to strike the

molecule anywhere along its molecular axis (except perpendicular to the axis at the C atom and parallel to the axis at the O atom end) without inducing significant rotational motion [20]. Long-range energy transfer of this type has been characterized for small molecules at low excitation energies [21, 22, 23, 24, 25, 26, 27, 28 and 29], but came as a complete surprise for encounters of the highly energetic nature described here. In such cases excitation of the vibrations of the acceptor is believed to arise from resonant vibrational-energy transfer in which the donor and the acceptor lose and gain, respectively, equal amounts of internal energy. Such an energy exchange is known to occur via long-range forces of which the most important is the transition-dipole–transition-dipole interaction moment. (The contribution to the dipole moments arises from vibrations of the molecules and is proportional to the derivative of the dipole moment with respect to the molecular, internal, nuclear coordinates) [21, 22, 23, 24, 25, 26, 27, 28 and 29]. It is worth emphasizing that such energy transfer processes do not occur with high probability in a typical collision. (The rate constant for the process depicted in [figure C3.3.5](#) has been measured and indicates that the probability for exciting all of the 00^0_1 molecules without regard to their rotational state designation, is about 1% per gas kinetic encounter [6, 8, 11, 13].)

C3.3.4.2 VELOCITY PROFILES FOR TRANSLATIONAL–ROTATIONAL EXCITATION OF THE BATH

In stark contrast to the results shown in [figure C3.3.5](#) are data obtained for collision processes that lead to no vibrational excitation of the bath molecule, leaving CO_2 in its ground vibrationless state, 00^0_0 , but in highly rotational excited levels [5, 9, 11, 12, 14, 16, 17]. The mean J for CO_2 at $T = 298$ K is 24, and [figure C3.3.6](#) shows linewidth data obtained for the transition $00^0_0; J = 72 \rightarrow 00^0_1; J = 71$ of CO_2 produced by collisions with methylpyrazine molecules excited to energies of 37 000 and 41 000 cm^{-1} , respectively. Such high rotational levels of CO_2 are essentially unpopulated at the cell temperature of 298 K. The linewidths measured for this transition correspond to temperatures of 1340 ± 250 and 1160 ± 220 K, respectively, a little over four times the initial ambient value, indicating significant recoil for this rapidly rotating molecule. Again, the picture presented by these data is very consistent if the behaviour of both the rotational and translational motions of the recoiling CO_2 bath molecule are considered. A hard, impulsive, collision that samples the repulsive wall of the methylpyrazine donor, with its rapidly oscillating atomic pistons, provides a significant kick to the bath acceptor molecule causing large translational recoil. Again, because CO_2 is cigar shaped, such a violent hit on the molecule almost always leads to significant rotational excitation accompanying the translational recoil [20]! While rate constant measurements for such processes indicate that the probability for exciting a given J level of 00^0_0 is of the order of 0.1–1% per gas-kinetic collision [5, 9, 11, 12, 14, 16, 17] the total excitation probability for this process, summed over all J levels in 00^0_0 , is probably greater than 90%. The key point is that a collision between a medium-sized donor, with very high levels of internal vibrational excitation, and a small bath acceptor molecule is most likely to put donor vibrational energy into rotation and translation of the bath with little going into overall bath rotation.

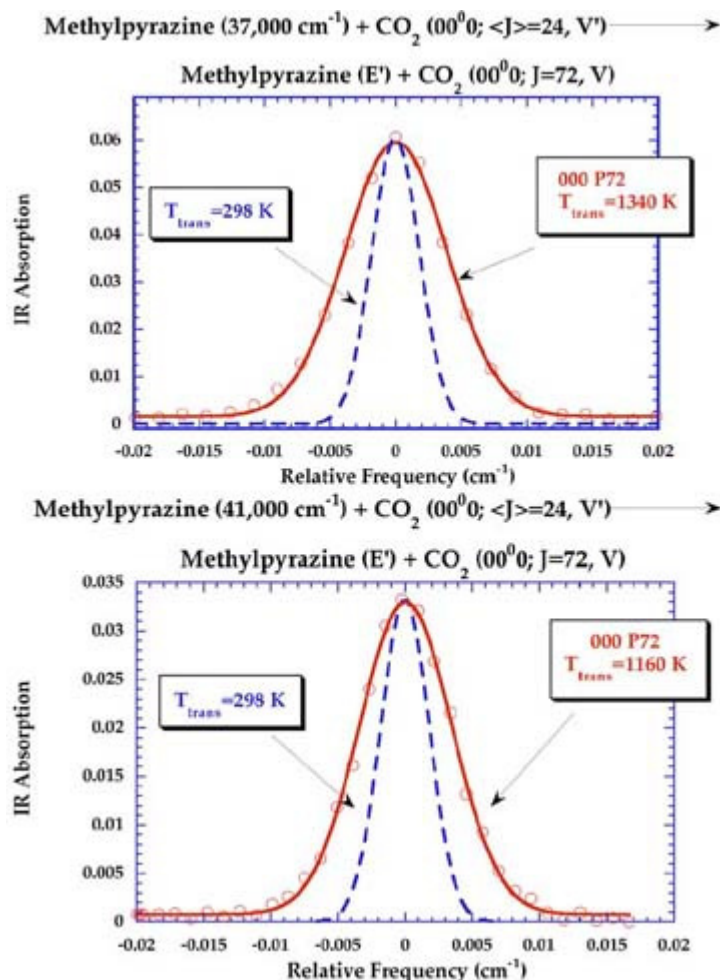


Figure C3.3.6. Doppler-line profiles for molecules scattered into the $\text{CO}_2(00^0_0; J=72)$ state by collisions with hot methylpyrazine molecules as depicted by the equations above each half of the figure. The energy of methylpyrazine in the upper half of the figure is $37\,000\text{ cm}^{-1}$ (excitation at 266 nm) while the energy of methylpyrazine in the lower half of the figure is $41\,000\text{ cm}^{-1}$ (excitation at 248 nm).

C3.3.4.3 QUALITATIVE CORRELATION OF ENERGY TRANSFER DATA TO THE INTERMOLECULAR POTENTIAL

Figure C3.3.7 shows a typical intermolecular potential and the types of recoil linewidth that are observed for collisions that sample the long- and short-range forces acting during a collision. For all cases studied so far, vibrational excitation of the bath acceptor species is accompanied by almost no excitation of translational or rotational motion. The clear signature of the kind of collision that samples the long-range, attractive part of the intermolecular potential is narrow recoil linewidths as shown in the upper half of Figure C3.3.7. Figure C3.3.8 shows what might be a typical trajectory for this kind of interaction—a distant fly-by in which energy exchange is brought about by long-range electrical forces acting at a distance. On the other hand, collisions that sample the steep repulsive part of the potential shown in the lower half of Figure C3.3.7 have a signature characterized by wide recoil linewidths as shown in the upper half of Figure C3.3.7. A typical ‘direct hit’ trajectory that samples this part of the potential is shown in figure C3.3.9.

Close encounters of the bath molecule with the rapidly vibrating atoms of the donor lead to strong recoil (and corresponding rotational motion) with accompanying loss of energy from the donor vibrations. The distant fly-by is characterized by vibration–vibration (resonant) energy transfer, while the impulsive, violent collision is characterized by vibration–translation/rotation energy transfer. It is reasonable to ask why these mechanisms are so clearly separated in nature, for example, why the vibrationally excited bath molecules do not have a recoil linewidth intermediate between their original, ambient value and the large recoil linewidths exhibited in vibration–

translation/rotation energy transfer. While a definitive answer to this question is at present lacking, some clues can be found in the nature of the bath vibrations that have been studied so far. In all cases investigated to date with this technique, the bath modes excited by collisional energy transfer have had vibrational states whose separation is large compared to the mean thermal energy ($h\nu \gg kT$). These are sometimes referred to as ‘stiff’ vibrations. There are good theoretical reasons to expect ‘soft’ acceptor modes with energy separations comparable to kT to be excited by a combination of both short- and long-range forces [30, 31, 32, 33, 34, 35, 36, 37, 38 and 39].

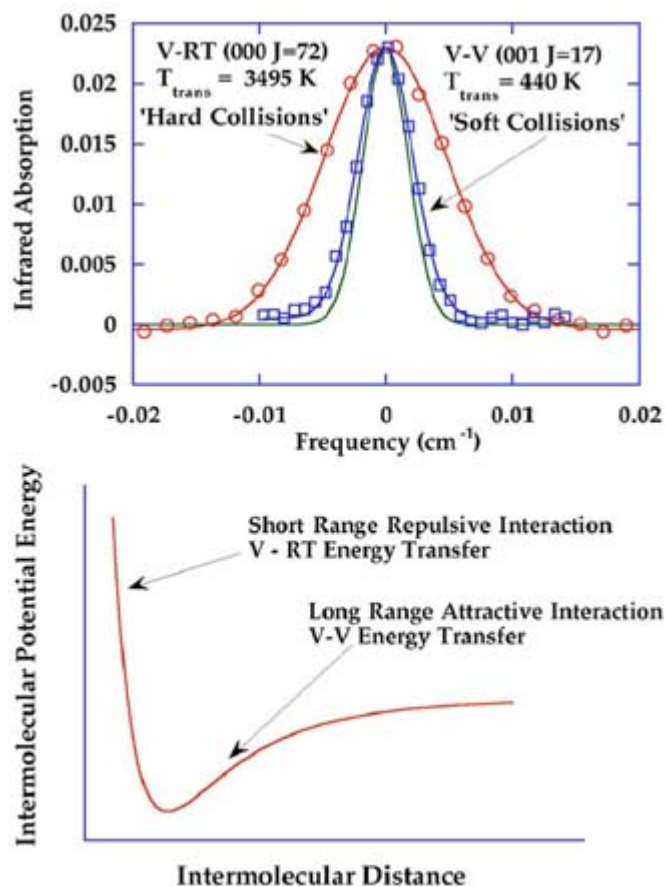


Figure C3.3.7. In the upper half of the figure are shown typical measured Doppler profiles for molecules scattered into the (00⁰0; $J = 72$) or (00⁰1; $J = 17$) states of CO₂ by collisions with hot pyrazine having an energy of 40 640 cm⁻¹. In the lower half of the figure is shown a typical intermolecular potential identifying the ‘hard’ and ‘soft’ collision regimes and the kind of energy transfer they effect.

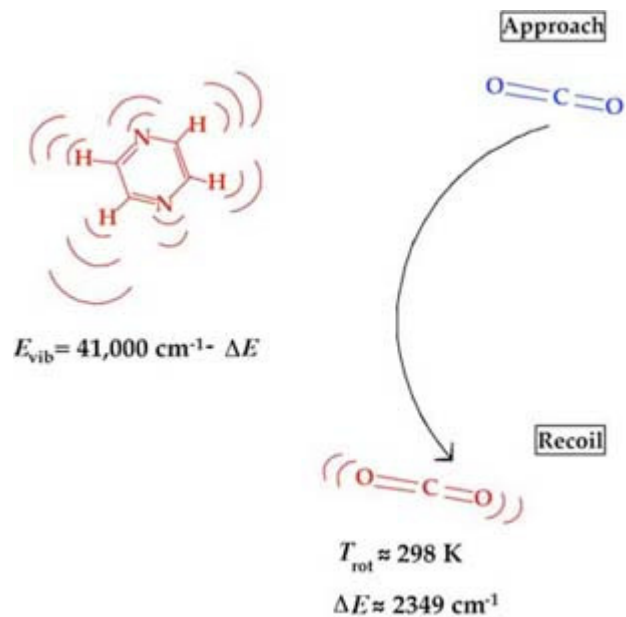


Figure C3.3.8. A typical trajectory for a ‘soft’ collision between a hot pyrazine molecule and a CO_2 bath molecule in which the CO_2 becomes vibrationally excited.

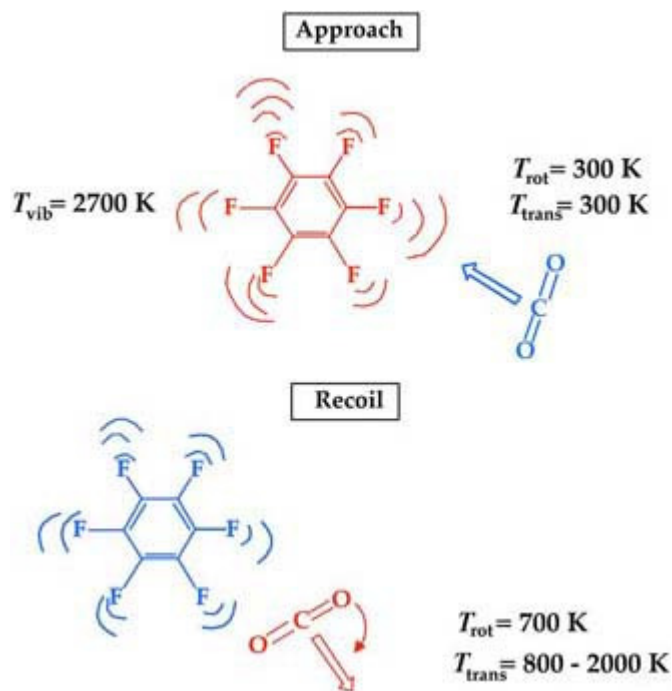


Figure C3.3.9. A typical trajectory for a ‘hard’ collision between a hot donor molecule and a CO_2 bath molecule in which the CO_2 becomes translationally and rotationally excited.

C3.3.5 QUANTITATIVE DATA ANALYSIS

C3.3.5.1 MASTER EQUATION ANALYSIS OF UNIMOLECULAR REACTION DYNAMICS

The analysis of the data presented so far, while qualitative, has provided a clear picture of the general mechanisms for loss of energy from a molecule with chemically significant amounts of energy. Nevertheless, a quantitative representation of this information would be highly desirable to use as a bench mark for comparison with theory and in practical applications where rates of unimolecular reactions are modelled with master equation techniques [40, 41, 42, 43 and 44]. Figure C3.3.10 shows a schematic, energy-level diagram for a molecule undergoing unimolecular decomposition. Above the reaction barrier the molecule has sufficient energy to undergo decomposition at a rate k_E represented by the arrows going to the right. On the other hand, quenching collisions of the type we have been discussing carry molecules down to lower energy as represented by the downward arrows in the figure. When a quenching collision carries a molecule from an energy above the reaction barrier to an energy below this barrier, it snuffs out the reaction process. In the original Lindemann unimolecular reaction scheme [2], both k_E and the rate constant for quenching collisions were assumed to be the same, independent of energy. This assumption is too simple. Both k_E and k_q , the quenching rate constant, are energy dependent. Quenching collisions can not only bring molecules below the reaction barrier depicted in figure C3.3.10, they can also reshuffle molecules within the different energy states above the barrier, changing the rate of unimolecular decomposition because of the dependence of k_E on E . An expression that takes into account these different energy dependences is the master equation giving the rate of loss of substrate $S^*(E)$ at an energy E [40, 41, 42, 43 and 44]:

$$d[S^*(E)]/dt = k_{LJ}[B] \int_0^\infty \{P(E, E')[S^*(E')] - P(E', E)[S^*(E)]\} dE' - k_E[S^*(E)].$$

The last term in this expression is simply the unimolecular rate of loss of substrate $S^*(E)$, k_{LJ} is the Lennard-Jones rate constant, [B] the concentration of bath molecules and $P(E, E')$ the energy transfer probability distribution function. $P(E, E')$ gives the probability that a substrate molecule $S^*(E)$ at energy E will be carried to an energy E' in a collision with a bath quenching molecule. Note that, while k_E is a property of the substrate molecule alone, $P(E, E')$ depends on the identity of both substrate and bath molecules (as well as on the energies E and E'). From this expression, a full description of a unimolecular reaction process clearly requires a knowledge of the distribution function $P(E, E')$ for energy loss from the substrate $S^*(E)$.

-16-

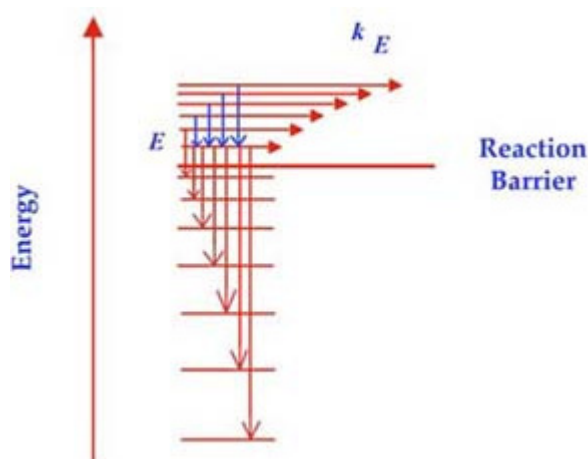


Figure C3.3.10. A schematic energy-level diagram for a molecule capable of undergoing unimolecular reaction above the energy depicted as the reaction barrier. Arrows to the right indicate reaction (collision-free) at a rate k_E that depends on the energy E . Down arrows represent collisional redistribution of the hot molecules both above and below the reaction barrier.

C3.3.5.2 EXTRACTING THE ENERGY TRANSFER PROBABILITY DISTRIBUTION FUNCTION $P(E, E')$

The data obtained in the infrared-diode-laser-probe studies described above provides quenching information at a given substrate donor energy E . By varying the laser excitation wavelength for production of vibrationally hot

species, equivalent data at other excitation energies can be obtained. Experiments of this type have already begun [45, 46 and 47]. Even to extract $P(E, E')$ at a single donor energy E by inverting the experimental data from these studies is a formidable task [15]. First, the information obtained provides a measure of the probability, $P(00^0_0; J, V)$, for producing a given final quantum state of the bath acceptor molecule, such as $00^0_0, J, V$, with a specific vibrational (00^0_0), rotational (J) and velocity (V) signature. In order to turn such a distribution into a $P(E, E')$ distribution, some method of identifying the initial state of the bath molecule must be found so that the energy change $\Delta E = E' - E$ occurring during the collision can be determined. (The conventional way to define ΔE gives it a negative sign for energy loss ($E' < E$) from the donor.) The initial state of the bath molecule consists of a Boltzmann distribution of velocities and rotational state populations described by the cell temperature T . Thus, each final state of the bath molecule can arise from a number of different initial states leading to a distribution of ΔE values. Fortunately, for large ΔE , this spread is not too significant because the initial distribution for cell temperatures near $T = 300$ K is not large. In addition, by studying the final $P(00^0_0; J, V)$ distributions as a function of cell temperature, the initial states of the bath that contribute significant population to a given scattered $00^0_0, J, V$ state can be narrowed still further [15, 16]. Second, in the case of translational motion, the quantity of interest is the energy transferred in the centre-of-mass frame that takes into account the recoil of both the bath acceptor and the donor. Thus, the data obtained in the experiments that measures the laboratory frame recoil velocities of the bath molecules, as described here, must be transformed into the centre-of-mass frame. The procedure for doing this is lengthy and has been described elsewhere [9, 12]. Third, the results for collision-induced scattering into a large number of different final states of the bath molecule must be summed in order to obtain the complete distribution function $P(E, E')$. Finally, there is no way at present to take into account (no experimental measure of) the change in rotational energy of the donor molecule during the collision. For heavy donors, this is not expected to cause much error in the determination of the distribution functions because angular-momentum constraints limit the maximum change in angular momentum during the collision [9, 20].

-17-

For heavy molecules with very small rotational state spacing, this limit on ΔJ puts severe upper limits on the amount of energy that can be taken up in the rotations of a heavy molecule during a collision. Despite these limitations, $P(E, E')$ distributions have been obtained by inverting data of the type described here for values of ΔE in the range $-1500 \text{ cm}^{-1} > \Delta E > -8000 \text{ cm}^{-1}$ for the two donor molecules pyrazine and hexafluorobenzene with carbon dioxide as a bath acceptor molecule [15, 16]. Figure C3.3.11 shows these experimentally derived probability distributions for events that leave the CO_2 bath molecule in its ground vibrational level 00^0_0 ('pure' vibration-rotation/translation energy transfer). Even though limited to large values of ΔE , these probability distribution functions are very revealing. First and foremost, we see that the probability for very large energy transfers (e.g. $-\Delta E = 6000\text{--}8000 \text{ cm}^{-1}$) is small but measurable. These so-called 'super collisions' were a great surprise when first discovered a number of years ago [48, 49, 50, 51, 52, 53, 54, 55 and 56]. The distributions in figure C3.3.11 can be thought of as the 'supercollision tail' of the $P(E, E')$ distribution function. A second interesting feature of the data in figure C3.3.11 is the difference between the two molecules. Evidently, $P(E, E')$ is significantly larger for hexafluorobenzene than for pyrazine at small ΔE . Since the *average* energy transferred in collisions of this type is always weighted heavily by low ΔE values, we would expect that the mean energy transferred from hexafluorobenzene to bath acceptors would be larger than the mean energy transferred from pyrazine to the same bath acceptors. Experimental measurements of these average energy transfer values have not yet been made for the same CO_2 quencher bath molecule, but the trends from self-quenching data indicate that C_6F_6 will have a substantially larger mean energy transfer value than pyrazine [57, 58]. Such a trend is consistent with the large number of low-frequency modes in C_6F_6 compared to pyrazine, a factor that usually drives up the mean-energy-transfer values. Finally, the probabilities for transferring large amounts of energy become larger in pyrazine than in C_6F_6 for ΔE values more negative than about -3000 cm^{-1} . While no data are generally available on the variation in 'supercollision' probability with molecular parameters, the trend in figure C3.3.11 can be rationalized by recognizing that pyrazine has more high-frequency vibrations than C_6F_6 . In transferring, e.g. 6000 cm^{-1} from pyrazine to CO_2 , only two C-H-stretch quanta are required, while four C-F-stretch quanta would be needed in the case of C_6F_6 [16]. As a general rule, energy-transfer probabilities increase if the number of vibrational quanta surrendered in the exchange process are minimized [30, 31, 32, 33, 34, 35, 36, 37, 38 and 39].

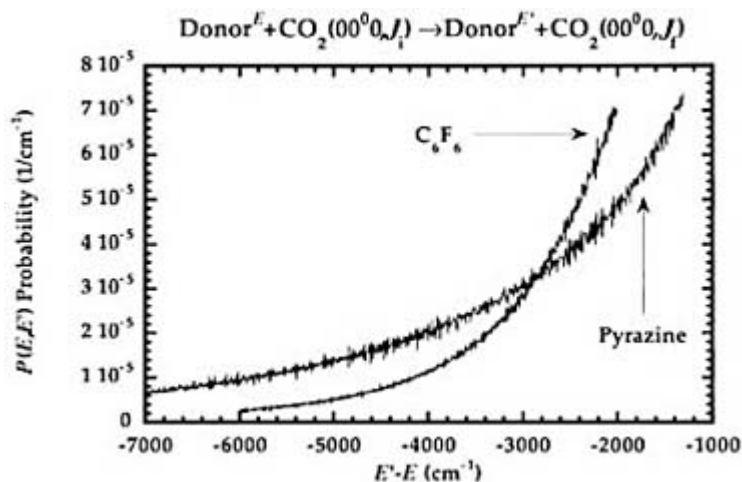


Figure C3.3.11. The energy transfer probability distribution function $P(E, E')$ (see figure C3.3.2) for two molecules, pyrazine and hexafluorobenzene, excited at 248 nm, arising from collisions with carbon dioxide molecules. Only those collisions that leave the carbon dioxide bath molecule in its ground vibrationless state $00^0 0$ have been included in computing this probability.

The probability distribution functions shown in figure C3.3.11 are limited to events that leave the bath molecule vibrationally unexcited. Nevertheless, we know that the vibrations of the bath molecule are excited, albeit with low probability in collisions of the type being considered here. Figure C3.3.12 shows how these $P(E, E')$ distribution functions of Figure C3.3.11 are changed if the probability for exciting the $\text{CO}_2(00^0 1)$ level is also included. Because such bath vibrational excitation is accompanied by essentially no translational or rotational energy gain, the probability increase from this channel is confined to a narrow region at 2349 cm^{-1} , the energy of the $\text{CO}_2 00^0 0 \rightarrow 00^0 1$ vibrational transition. Thus, the quantum nature of the bath vibrations, coupled with a mechanism in which bath vibrations are resonantly excited by vibration–vibration energy exchange from a hot donor, leads to the appearance of resonances or spikes in the $P(E, E')$ distribution function [15, 16]!

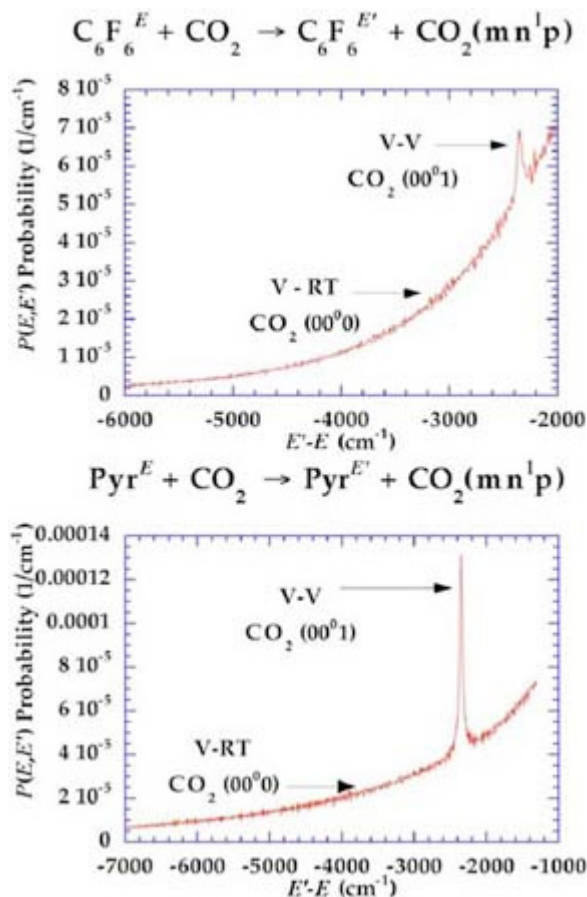


Figure C3.3.12. The energy-transfer-probability-distribution function $P(E, E')$ (see [figure C3.3.2](#) and [figure C3.3.11](#)) for two molecules, pyrazine and hexafluorobenzene, excited at 248 nm, arising from collisions with carbon dioxide molecules. Both collisions that leave the carbon dioxide bath molecule in its ground vibrationless state, 00^00 , and those that excite the 00^01 vibrational state (2349 cm^{-1}), have been included in computing this probability. The spikes in the distribution arise from excitation of the carbon dioxide bath 00^01 vibrational mode.

C3.3.6 SUMMARY

A variety of interconnected, bimolecular, collision studies have been described that employ laser devices to investigate the detailed energy disposal in product molecular species and to provide a direct experimental measure of the energy-transfer distribution function $P(E, E')$. A knowledge of this function is important both in testing detailed theoretical energy-transfer calculations and in modelling unimolecular chemical reactions using master-equation methods. Although there have been a number of extremely informative studies of the energy relaxation of highly vibrationally excited molecules in the past, with rare exception these studies are not able to follow all the degrees of freedom of the quencher molecules. The experimental approach described here, designed as it is to investigate the detailed dynamics of these collisions by *separately* probing the vibrational, rotational and translational degrees of freedom, can significantly increase our understanding of the mechanisms for these fundamental processes that are of such importance in studies of photochemistry, unimolecular and bimolecular reactions. All of these experiments provide data of fundamental chemical interest since the information obtained is sensitive to molecular-potential-energy surfaces and can serve as a test for necessarily approximate dynamical theories. In addition, many of the experimental data obtained will be of practical interest in the study and control of unimolecular chemical reactions and photochemical processes in the development of optically pumped molecular

lasers and in the development of an improved understanding of atmospheric chemical reactions.

ACKNOWLEDGMENTS

I am indebted to my students, post-doctoral fellows and collaborators, particularly Chris Michaels, Eric Sevy, Amy Mullin, Zhen Lin, Charles Tapalian, Professor Mark Muyskens and Dr Ralph Weston who have contributed to the insights and experimental efforts described here.

REFERENCES

- [1] Kauzmann W 1966 *Kinetic Theory of Gases* (New York: Benjamin) pp 165–84
 - [2] Lindemann F 1922 A discussion on the 'radiation theory of chemical action' *Trans. Faraday Soc.* **17** 598–606
 - [3] Oref I and Tardy D C 1990 Energy transfer in highly excited large polyatomic molecules *Chem. Rev.* **90** 1407–45
 - [4] Tardy D C and Rabinovitch B S 1977 Intermolecular vibrational energy transfer in thermal unimolecular systems *Chem. Rev.* **77** 369–408
 - [5] Chou J Z and Flynn G W 1990 Energy dependence of the relaxation of highly excited NO₂ donors under single collision conditions: vibrational and rotational state dependence and translational recoil of CO₂ quencher molecules *J. Chem. Phys.* **93** 6099–101
 - [6] Zheng L, Chou J and Flynn G W 1991 Relaxation of molecules with chemically significant amounts of energy: vibrational, rotational and translational energy recoil of an N₂ O bath due to collisions with NO₂ ($E = 63.5$ kcal/mole) *J. Phys. Chem.* **95** 6759–62
-

-20-

- [7] Sedlacek A J, Weston R E Jr and Flynn G W 1991 Interrogating the vibrational relaxation of highly excited polyatomics with time-resolved diode laser spectroscopy: C₆H₆, C₆D₆, and C₆F₆ + CO₂ *J. Chem. Phys.* **94** 6483–90
- [8] Weston R E Jr and Flynn G W 1992 Relaxation of molecules with chemically significant amounts of vibrational energy: the dawn of the quantum state resolved era *Annu. Rev. Phys. Chem.* **43** 559–89
- [9] Mullin A S, Park J, Chou J Z, Flynn G W and Weston R E Jr 1993 Some rotations like it hot: selective energy partitioning in the state resolved dynamics of collisions between CO₂ and highly vibrationally excited pyrazine *Chem. Phys.* **175** 53–70
- [10] Flynn G W and Weston R E Jr 1993 Diode laser studies of collisional energy transfer *J. Phys. Chem.* **97** 8116–27
- [11] Flynn G W and Weston R E Jr 1995 Glimpses of a mechanism for quenching unimolecular reactions: a quantum state resolved picture *Advances in Chemical Kinetics and Dynamics* vol 2B, ed J Barker (Greenwich, CT: JAI Press) pp 359–91
- [12] Mullin A S, Michaels C A and Flynn G W 1995 Molecular supercollisions: evidence for large energy transfer in the collisional relaxation of highly vibrationally excited pyrazine by CO₂ *J. Chem. Phys.* **102** 6032–45
- [13] Michaels C A, Mullin A S and Flynn G W 1995 Long and short range interactions in the temperature dependent collisional excitation of the antisymmetric stretching CO₂(00⁰1) level by highly vibrationally excited pyrazine *J. Chem. Phys.* **102** 6682–95
- [14] Michaels C A, Tapalian C, Lin Z, Sevy E and Flynn G W 1995 Supercollisions, photofragmentation and energy transfer in mixtures of pyrazine and carbon dioxide *Faraday Discuss.* **102** 405–22
- [15] Michaels C A and Flynn G W 1997 Connecting scattering data directly to chemical kinetics: energy transfer

distribution functions for the collisional relaxation of highly vibrationally excited molecules from state resolved probes of the bath *J. Chem. Phys.* **106** 3558–66

- [16] Michaels C A, Lin Z, Mullin A S, Tapalian H C and Flynn G W 1997 Translational and rotational excitation of the CO₂(00⁰) vibrationless state in the collisional quenching of highly vibrationally excited perfluorobenzene: evidence for impulsive collisions accompanied by large energy transfers *J. Chem. Phys.* **106** 7055–71
- [17] Flynn G W, Michaels C A, Tapalian C, Lin Z, Sevy E and Muyskens M A 1997 Infrared laser snapshots: vibrational, rotational and translational energy probes of high energy collision dynamics *Highly Excited Molecules: Relaxation, Reaction, and Structure* ed A Mullin and G Schatz (Washington, DC: ACS)
- [18] Michaels C A, Mullin A S, Park J, Chou J Z and Flynn G W 1998 The collisional deactivation of highly vibrationally excited pyrazine by a bath of carbon dioxide: excitation of the infrared inactive (10⁰), (02⁰), and (02²) bath vibrational modes *J. Chem. Phys.* **108** 2744–55
- [19] Smith M A H, Rinsland C P and Fridovich B 1985 Intensities and collision broadening parameters from infrared spectra *Molecular Spectroscopy: Modern Research Volume III* ed K N Rao (New York: Academic) pp 118–19
- [20] Kreutz T G and Flynn G W 1990 Analysis of translational, rotational, and vibrational energy transfer in collisions between CO₂ and hot hydrogen atoms: the three dimensional 'breathing ellipse' model *J. Chem. Phys.* **93** 452–65
- [21] Stephenson J C and Moore C B 1972 Temperature dependence of nearly resonant vibration–vibration energy transfer in CO₂ mixtures *J. Chem. Phys.* **56** 1295–308
- [22] Rosser W A Jr, Sharma R D and Gerry E T 1971 Deactivation of vibrationally excited carbon dioxide (001) by collisions with carbon monoxide *J. Chem. Phys.* **54** 1196–205
- [23] Gueguen H, Arditi I, Margottin-Maclou M, Doyennette L and Henry L 1971 Vibrational energy transfer of nitrogen oxide and carbon dioxide excited on ν_3 to carbon dioxide or nitrous oxide, molecular nitrogen, carbon monoxide, hydrogen chloride, hydrogen bromide and hydrogen iodide *C. R. Acad. Sci. Paris* **272** 1139–42
- [24] Rosser W A Jr and Gerry E T 1971 De-excitation of vibrationally excited CO₂(001) by collisions with CO₂, H₂, and C₁₂ *J. Chem. Phys.* **54** 4131–2

- [25] Margottin-Maclou M, Doyennette L and Henry L 1971 Relaxation of vibrational energy in carbon monoxide, hydrogen chloride, carbon dioxide and nitrous oxide *Appl. Opt.* **10** 1768–80
- [26] Stephenson J C, Finzi J and Moore C B 1972 Vibration–vibration energy transfer in CO₂–hydrogen halide mixtures *J. Chem. Phys.* **56** 5214–21
- [27] Osgood R M Jr, Sackett P B and Javan A 1974 Measurement of vibrational–vibrational exchange rates for excited vibrational levels ($2 \leq \nu \leq 4$) in hydrogen fluoride *J. Chem. Phys.* **60** 1464–80
- [28] Sharma R D and Brau C A 1969 Energy transfer in near-resonant molecular collisions due to long-range forces with application to transfer of vibrational energy from the ν_3 mode of CO₂ to N₂ *J. Chem. Phys.* **50** 924–30
- [29] Sharma R D and Brau C A 1967 Near-resonant vibrational energy transfer in nitrogen carbon dioxide mixtures *Phys. Rev. Lett.* **19** 1273–5
- [30] Schwartz R N, Slawsky Z I and Herzfeld K F 1952 Calculation of vibrational relaxation times in gases *J. Chem. Phys.* **20** 1591–9
- [31] Levine R D and Bernstein R B 1974 *Molecular Reaction Dynamics* (New York: Oxford)
- [32] Cottrell T L and McCoubrey J C 1961 *Molecular Energy Transfer in Gases* (London: Butterworths)
- [33] Yardley J T and Moore C B 1967 Intramolecular vibration-to-vibration energy transfer in carbon dioxide *J. Chem. Phys.* **46** 4491–5
- [34] Moore C B 1969 Laser studies of vibrational energy transfer *Accounts Chem. Res.* **2** 103–9
- [35] Grabiner F R, Flynn G W and Ronn A M 1973 Vibration–vibration equilibration in laser excited CH₃F and CH₃F–X mixtures *J. Chem. Phys.* **59** 2330–4

- [36] Weitz E and Flynn G W 1974 Laser studies of vibrational and rotational relaxation in small molecules *Annu. Rev. Phys. Chem.* **25** 275–315
- [37] Weitz E and Flynn G W 1981 Vibrational energy flow in the ground electronic states of polyatomic molecules *Advances in Chemical Physics Vol. XLVII, Photoselective Chemistry* part 2, ed J Jortner, R D Levine and S A Rice, pp 185–235
- [38] Flynn G W 1981 Collision induced energy flow between vibrational modes of small polyatomic molecules *Accounts Chem. Res.* **14** 334–41
- [39] Yardley J T 1980 *Introduction to Molecular Energy Transfer* (New York: Academic)
- [40] Troe J 1977 Theory of thermal unimolecular reactions at low pressures. I. Solutions of the master equation *J. Chem. Phys.* **66** 4745–57
- [41] Troe J 1977 Theory of thermal unimolecular reactions at low pressures. II. Strong collision rate constants. Applications *J. Chem. Phys.* **66** 4758–75
- [42] Gilbert R G and Smith S C 1990 *Theory of Unimolecular and Recombination Reactions* (Oxford: Blackwell)
- [43] Baer T and Hase W L 1996 *Unimolecular Reaction Dynamics: Theory and Experiments* (Oxford)
- [44] Barker J R, Brenner J D and Toselli B M 1995 *Advances in Chemical Kinetics and Dynamics* vol 2B, ed J Barker (Greenwich, CT: JAI Press) pp 393–425
- [45] Wall M C and Mullin A S 1998 Supercollision energy dependence: state resolved energy transfer in collisions between highly vibrationally excited pyrazine ($E_{\text{vib}} = 37,900 \text{ cm}^{-1}$ and $40,900 \text{ cm}^{-1}$) and CO_2 *J. Chem. Phys.* **108** 9658–67
- [46] Wall M C, Lemoff A and Mullin A S 1998 An independent determination of supercollision energy loss magnitudes and rates in highly vibrationally excited pyrazine with $E_{\text{vib}} = 36,000$ to $41,000 \text{ cm}^{-1}$ *J. Phys. Chem.* at press
- [47] Wall M C, Stewart B A and Mullin A S 1998 State resolved collisional relaxation of highly vibrationally excited pyridine ($E_{\text{vib}} = 38,000 \text{ cm}^{-1}$) and CO_2 : influence of a permanent dipole moment *J. Chem. Phys.* **108** 6185–96
-

-22-

- [48] Loesch H J and Herschbach D R 1972 Ballistic mechanism for vibrational and rotational energy transfer in Ar + Csl collisions *J. Chem. Phys.* **57** 2038–50
- [49] Crim F F, Chou M S and Fisk G A 1973 Inelastic scattering of vibrationally excited KBr by small nonpolar and essentially nonpolar partners *Chem. Phys.* **2** 283–92
- [50] Crim F F, Bente H B and Fisk G A 1974 Inelastic scattering of vibrationally excited potassium bromide by polyatomic partners *J. Phys. Chem.* **78** 2438–42
- [51] Fisk G A and Crim F F 1977 Single collision studies of vibrational energy transfer mechanisms *Accounts Chem. Res.* **10** 73–9
- [52] Brown N J and Miller J A 1984 Collisional energy transfer in the low-pressure-limit unimolecular dissociation of HO_2 *J. Chem. Phys.* **80** 5568–80
- [53] Hassoon S, Oref I and Steel C 1988 Collisional activation of quadricyclane by azulene: an example of very strong collisions *J. Chem. Phys.* **89** 1743–4
- [54] Morgulis J M, Sapers S S, Steel C and Oref I 1989 Collisional activation of cyclobutene by hexafluorobenzene: a chemical probe for highly energetic collisions in reactive systems *J. Chem. Phys.* **90** 923–9
- [55] Sharma R D and Sindoni J M 1992 Relaxation of highly vibrationally excited KBr by Ar *Phys. Rev. A* **45** 531–4
- [56] Oref I 1995 Supercollisions *Advances in Chemical Kinetics and Dynamics* vol 2B, ed J Barker (Greenwich, CT: JAI Press) pp 285–98
- [57] Miller L A and Barker J R 1996 Collisional deactivation of highly vibrationally excited pyrazine *J. Chem. Phys.* **105** 1383–91
- [58] Lenzer T, Luther K, Troe J, Gilbert R G and Lim K F 1995 Trajectory simulations of collisional energy transfer in highly excited benzene and hexafluorobenzene *J. Chem. Phys.* **103** 626–41
-

C3.4 Electronic energy transfer in condensed phases

Andrey A Demidov and David L Andrews

C3.4.1 INTRODUCTION

In condensed phase matter, the process of electronic excitation through the absorption of light commonly results in a system with an initially high degree of energetic instability. One of the most important and rapid means by which the system begins to accommodate its energy is a redistribution of excitation between the component optical centres (molecules or chromophores). The fundamental mechanism for this redistribution is the phenomenon of *electronic energy transfer*. At the molecular level this is a pairwise process in which the energy of electronic excitation held by one molecule transfers to another. As a result the former molecule, called the *donor*, imparts its energy of electronic excitation to the latter *acceptor*.

In modelling energy transfer in a bulk medium containing more than two chromophores or molecules, one has to consider all the individual pairwise processes and conduct the appropriate averaging. Such an approach is valid when excitation can be localized on individual molecules. This is called *localized excitation*; the localized energy is termed a *localized exciton*. The process of energy migration is then known as *incoherent energy transfer*, and the jump of an exciton from one molecule to another is a good way to describe it. The physical condition for incoherent energy transfer is weak coupling between the donor and acceptor species.

In the opposite case, i.e. strong coupling, the model of energy transfer changes. Then the donor and acceptor form a dimer-like structure as their excited states mix together and form a joint excited state split by twice the coupling energy. With such excitonic states we can no longer specify the molecular location of the electronic excitation, as it is 'spread out' between both molecules. Now we have to use a different language, and the issue is energy transfer between excitonic states.

In systems comprising a large number of molecules one can find various types of energy transfer. If the molecules are strongly coupled we observe multi-excitonic behaviour, where not just two but a number of excitonic excited states are formed. Energy transfer in this case is called *coherent energy transfer*. Generally it may happen that in a bulk medium some molecules have strong coupling, whereas others have weak coupling, resulting in a mixture of coherent and incoherent energy migration. Let us now consider in more detail the different mechanisms that lead to these types of electronic energy transfer.

C3.4.2 INCOHERENT ENERGY TRANSFER

In our everyday life we move in a world that we largely experience through processes of electronic energy transfer. For example, you can read the printed words on this page because the print absorbs radiation emitted by your source of light. At the atomic level that is a process of electronic energy transfer from, let us say, the excited tungsten atoms of your reading lamp to the atoms of carbon on the page. The same principle applies across cosmic scales as we peer up at

distant stars in the night sky, the very process of vision requiring molecules of rhodopsin in the retina to capture

photons of starlight and so enter an electronic excited state. In a very real sense this radiative process is a direct transfer of electronic energy from the star to the retina. Obviously the coupling between the donor and acceptor in such a case is close to zero.

The coupling remains relatively weak even at the other extreme of distance, where the energy transfer occurs within a single piece of matter and the donor and acceptor are separated by, let us say, only a few nanometres. Given the weak coupling, we still have incoherent energy transfer, but now the photon that conveys the energy has to be conceptualized as a virtual quantity—its involvement can only be inferred, and the energy transfer is for all practical purposes radiationless. This kind of energy transfer was first investigated by Förster in 1948 [1], and addressed with a perturbation theory based on dipole–dipole interaction between the excited donor and unexcited acceptor. Later the theory was rectified in a number of works [2, 3] and became the proven workhorse for much of the modern research on energy transfer in condensed matter [4, 5].

More recently Andrews and Juzeliunas [6, 7] developed a unified theory that embraces both radiationless (Förster) and long-range radiative energy transfer. In other words this theory is valid over the whole span of distances ranging from those which characterize molecular structure (nanometres) up to cosmic distances. It also addresses the intermediate range where neither the radiative nor the Förster mechanism is fully valid. Below is their expression for the rate of pairwise energy transfer w from donor to acceptor, applicable to transfer in systems where the donor and acceptor are embedded in a transparent medium of refractive index n :

$$\begin{aligned}
 w &= w_F + w_I + w_{\text{rad}} \\
 w_F &= \frac{9\kappa_3^2 c^4}{8\pi \tau_D n^4 R^6} \int F_D(\omega) \sigma_A(\omega) \frac{d\omega}{\omega^4} \\
 w_I &= \frac{9c^2}{8\pi \tau_D n^2 R^4} (\kappa_3^2 - 2\kappa_1 \kappa_3) \int F_D(\omega) \sigma_A(\omega) \frac{d\omega}{\omega^2} \\
 w_{\text{rad}} &= \frac{9\kappa_1^2}{8\pi \tau_D R^2} \int F_D(\omega) \sigma_A(\omega) d\omega.
 \end{aligned} \tag{C3.4.1}$$

In the above expression the three terms respectively represent a ‘Förster’ rate contribution, w_F ; a radiative (fluorescence) term, w_{rad} ; and an ‘intermediate’ term, w_I . The various parameters featuring in (C3.4.1) are defined as follows: $F_D(\omega)$ is the normalized spectrum of donor fluorescence; $\sigma_A(\omega)$ is the absorption cross section of the acceptor; ω is the optical frequency ($2\pi\nu$); c is the speed of light; τ_D is the radiative lifetime of the donor, n is the refractive index; and R is the distance between the donor and acceptor. Lastly, $\kappa_{1,3}$ are orientation factors (their detailed form to be given below). It may be noted that if the medium across which energy transfer occurs is not transparent, one has to take into account the frequency dependence of the refractive index, and in particular its imaginary part, leading to a more complex result (see [7]). One can find elsewhere the following alternative form of the Förster formula:

$$w_F = \frac{2\kappa^2}{3} \frac{1}{\tau_D} \left(\frac{R_0}{R} \right)^6 \tag{C3.4.2}$$

-3-

where R_0 is called the ‘critical’ or Förster radius. Obviously (C3.4.2) and the Förster component of (C3.4.1) are related, and the Förster radius is calculable from the overlap integral of the appropriate fluorescence and absorption spectra. Typical values of the Förster radius range over tens of ångströms, see [4, 8].

In (C3.4.1) we find three major forms of dependence on distance: R^{-6} , R^{-4} and R^{-2} which correspond to Förster, intermediate and radiative types of energy transfer. The intermediate part usually makes a significant contribution at distances of about a hundred or a few hundred nanometres, where all three components w_F , w_I and w_{rad} are comparable in magnitude. At shorter distances w_F dominates, whereas at larger distances w_{rad} dominates. It is

important to clarify the meaning of τ_D . It represents the rate of radiative transition of the donor from its excited state to the ground state, related to the measured fluorescence lifetime τ_{fluor} through the fluorescence quantum yield η , with $\tau_{\text{fluor}} = \eta \tau_D$. For example, chlorophyll *a* in various solutions exhibits a fluorescence lifetime of about 5–7 ns; with $\eta \sim 0.3$ –0.35, that makes $\tau_D \sim 15$ –20 ns [8, 9 and 10].

The κ factors in (C3.4.1) represent another very important facet of the energy transfer [4, 11]. These factors depend on the orientations of the donor and acceptor. For certain orientations they can reduce the rate of energy transfer to zero—for others they effect an ‘enhancement’ of the energy transfer to its maximum possible rate. Figure C3.4.1 exhibits the angles which define the mutual orientation of a donor and acceptor pair; in terms of those angles the orientation factors κ_1 and κ_3 are given by [6, 7]

$$\kappa_j = (\hat{\mu}_D \cdot \hat{\mu}_A) - j(\hat{R} \cdot \hat{\mu}_D)(\hat{R} \cdot \hat{\mu}_A) = \cos \alpha - j \cos \beta \cos \gamma \quad j = 1, 3 \quad (\text{C3.4.3})$$

where the case $j = 3$ is the conventional orientation factor, usually written simply as κ .

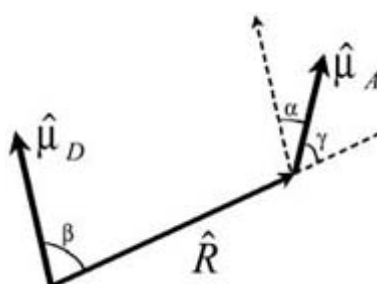


Figure C3.4.1. Example of the spatial orientation of three major vectors: $\hat{\mu}_D$, $\hat{\mu}_A$ and \hat{R} , unit vectors of the donor and acceptor transition dipole moments, and the donor–acceptor displacement-vector.

The dipole–dipole interaction which leads to (C3.4.2) and (C3.4.3) for energy transfer is in certain cases not applicable, as for example, if either the donor or acceptor transition is dipole (E1)-forbidden or exceptionally weak. Then, the coupling necessarily involves the electric quadrupole moment (E2), higher electric multipoles (E_n) or even magnetic multipoles (M_n), in each case leading to an orientation and distance of a different form. In the most common case of predominantly electric coupling, then if (E_n) and (E_m) are the leading non-zero moments of the donor and acceptor, the distance dependence takes the form $R^{-2(n+m+1)}$. Details of the functional form of the distance and angle dependence is discussed for example in [12, 13 and 14].

Knowledge of the pairwise energy transfer rates forms a basis for finding the average rate of energy transfer in an ensemble of N molecules. To this end, a system of ‘master equations’ is commonly employed [15, 16 and 17]. Then, the probability, p_i , to find excitation on molecule i can be calculated as:

$$\begin{aligned}
\frac{dp_1}{dt} &= -\frac{p_1}{\tau_1} - \sum_{j=2}^N w_{1j} p_1 + \sum_{j=2}^N w_{1j} p_j + F_1 \\
&\quad \vdots \\
\frac{dp_i}{dt} &= -\frac{p_i}{\tau_i} - \sum_{\substack{j=1 \\ j \neq i}}^N w_{ij} p_i + \sum_{\substack{j=1 \\ j \neq i}}^N w_{ij} p_j + F_i \\
&\quad \vdots \\
\frac{dp_N}{dt} &= -\frac{p_N}{\tau_N} - \sum_{j=1}^{N-1} w_{Nj} p_N + \sum_{j=1}^{N-1} w_{Nj} p_j + F_N.
\end{aligned} \tag{C3.4.4}$$

Here τ_i is the intrinsic lifetime of the excitation residing on molecule i (i.e. the fluorescence lifetime one would observe for the isolated molecule), w_{ij} is the pairwise energy transfer rate and F_i is the rate of excitation of the molecule i by the external source (the photon flux multiplied by the absorption cross section). The master equation system (C3.4.4) allows one to calculate the complete dynamics of energy migration between all molecules in an ensemble, but the computation can become quite complicated if the number of molecules is large. Moreover, it is commonly the case that the ensemble contains molecules of two, three or more spectral types, and experimentally it is practically impossible to distinguish the contributions of individual molecules from each spectral pool.

The measurement of fluorescence intensity from a compound containing chromophores of two spectral types is an example of a system for which it is reasonable to operate with the average rates of energy transfer between spectral pools of molecules. Let us consider the simple case of two spectral pools of donor and acceptor molecules, as illustrated in [figure C3.4.2](#) [18]. The average rate of energy transfer can be calculated as

$$k_{DA} = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M w_{ij} g_i \tag{C3.4.5}$$

where N and M are the number of donor and acceptor molecules respectively and g_i is the probability to find excitation on the donor i , commonly we have $g_i \sim 1$. In this case the master equation system can be simplified as:

$$\begin{aligned}
\frac{dn_D}{dt} &= -\frac{1}{\tau^D} n_D - k_{DA} n_D + k_{AD} n_A + F_D \\
\frac{dn_A}{dt} &= -\frac{1}{\tau^A} n_A + k_{DA} n_D - k_{AD} n_A + F_A.
\end{aligned} \tag{C3.4.6}$$

In these equations n_D and n_A are the excited state populations of the donor and acceptor molecules¹ and τ^D and τ^A are the lifetimes of the donor and acceptor molecules in the excited state; the notation τ^D is used to distinguish it from the radiative constant τ^D (in other words $\tau^D = \tau_{\text{fluor}}^D$ for the donor); k_{DA} is given by (C3.4.5) and k_{AD} , the corresponding rate constant for the backward energy transfer from acceptors to donors can be found by the same means. Finally, F_D and F_A represent external sources of excitation, for example the absorption of laser light by the donor and acceptor molecules. Commonly, for example in the case of δ -pulse excitation (in practice an ultrashort laser pulse), (C3.4.6) yields exponential decay kinetics for $n_D(t)$ and $n_A(t)$. The opposite case of steady excitation (CW light), yields the equilibrium ratio

$$\frac{n_D}{n_A} = \frac{(F_D/F_A)\tau_A^{-1} + k_{AD}(F_D + 1)}{\tau_D^{-1} + k_{DA}(F_D/F_A + 1)} \quad (\text{C3.4.7})$$

Master equation methods are not the only option for calculating the kinetics of energy transfer and analytic approaches in general have certain drawbacks in not reflecting, for example, certain statistical aspects of coupled systems. Alternative approaches to the calculation of energy migration dynamics in molecular ensembles are Monte Carlo calculations [18, 19 and 20] and probability matrix iteration [21, 22], amongst others.

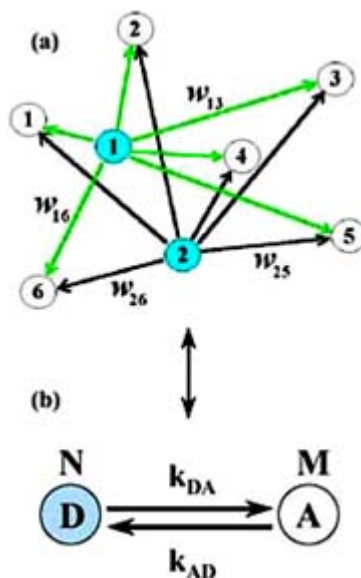


Figure C3.4.2. Schematic presentation of energy transfer between: (a) two donor molecules and six acceptor molecules; and (b) a general case of energy transfer involving a pool of N donor molecules and a pool of M acceptor molecules.

C3.4.3 POLARIZATION ANISOTROPY

Let us consider the case of a donor–acceptor pair where the acceptor, after capturing excitation from the donor, can emit a photon of fluorescence. If the excitation light is linearly polarized, the acceptor emission generally has a different polarization. Common quantitative expressions of this effect are the anisotropy of fluorescence, r , or the degree of polarization, P ;

$$r = \frac{I_{\parallel} - I_{\perp}}{I_{\parallel} + 2I_{\perp}} \quad P = \frac{I_{\parallel} - I_{\perp}}{I_{\parallel} + I_{\perp}} \Rightarrow r = \frac{2P}{3 - P} \quad (\text{C3.4.8})$$

where I_{\parallel} and I_{\perp} are measured components of the acceptor fluorescence parallel and perpendicular to the polarization of the incident excitation beam. The anisotropy of fluorescence is a valuable source of information about the structure of molecular complexes. The key factor determining the change in polarization is the angle α between the directions of the transition dipole moments of the donor, $\hat{\mu}_D$, and the acceptor, $\hat{\mu}_A$ (see [figure C3.4.1](#)). In the 1920s Levshin [23, 24] and Perrin [25] derived the following well known formula for an ensemble of randomly oriented donor–acceptor pairs

$$P = \frac{3 \cos^2 \alpha - 1}{3 + \cos^2 \alpha}. \quad (\text{C3.4.9})$$

This formula allows one to directly calculate the angle α , a microscopic parameter, by measuring the macroscopic value P (or equally the anisotropy r).

When applying polarization techniques one has to bear in mind that the above result is derived for an ensemble of independent donor–acceptor pairs with a ‘rigid’ structure—i.e., a system for which α is a constant. If we deal with an ensemble of randomly oriented donor and acceptor molecules the result is dramatically different, reflecting a very rapid loss of polarization ‘memory’. Then, one single act of energy transfer yields $r_0 = 1/25$ [2] and two or more energy transfer jumps, to all intents and purposes, totally destroys any polarization in the emitted fluorescence. This feature may be used as a powerful test for energy transfer in a sample of freely rotating or disordered molecules. If energy transfer occurs in a non-ordered system, the acceptor fluorescence is depolarized; if not, the fluorescence remains polarized. The one exception is where the angle $\alpha = 54.7^\circ$, the so-called ‘magic angle’, when the polarization will be zero anyway. In the case of incoherent energy transfer, the maximum magnitude of the anisotropy ($r = 0.4$) happens when the donor and acceptor have linear transition dipole moments oriented in parallel.

Naturally occurring molecular ensembles such as proteins from photosynthetic systems (plants, algae, photosynthetic bacteria, etc) are usually relatively rigid systems that contain various chromophores and hold them at fixed positions and orientations relative to each other. That is why, despite the numerous energy jumps between the chromophores, the resulting emitted fluorescence is polarized. The extent of this polarization thus affords invaluable information about the internal structure of molecular complexes.

-7-

C3.4.4 NONLINEAR PHENOMENA

C3.4.4.1 BLEACHING OF THE GROUND STATE

Equations (C3.4.5) and (C3.4.6) cover the common case when all molecules are initially in their ground electronic state and able to accept excitation. The system is also assumed to be impinged upon by sources F . The latter are usually expressible as the product $\sigma\Phi\hat{n}_0$, where σ is an absorption cross section, Φ is the photon flux and \hat{n}_0 is the population in the ground state. The common assumption is that $\hat{n}_0 \cong n_0$, i.e. practically all molecules are in the ground state because $n \gg n_0$. This is the assumption of linear excitation, where the system exhibits a linear response to the excitation intensity. This assumption does not hold when the extent of excitation is significant, i.e. when the rate of excitation inflow ($\sigma\Phi$) is larger than excitation dissipation, τ^{-1} . In this case we have a significant depletion of the ground state: $\hat{n}_0 = (n_0 - n) < n_0$, resulting in a nonlinear response from the ensemble.

C3.4.4.2 SINGLET–SINGLET (S–S) ANNIHILATION

Let us now consider the case where there is more than one exciton in the given molecular ensemble. The presence of two or more excess excitons not only creates two or more ‘holes’ in the ground state (see case (a) above) but it also opens up the possibility of two excitons being found on neighbouring molecules. Then the following two-stage process can take place [26]:



In the first stage, where at first we have two excitons S_1 , excitation jumps from one of the excited molecules to

another excited molecule. As a result, the latter molecule is promoted to a higher excited state S_n , while the former loses its energy and finds itself in the ground state S_0 . This is followed by the second step, a fast internal relaxation of the highly excited molecule from S_n to S_1 . Thus, where we started with two S_1 excitations we end up with only one, i.e. one excitation has effectively been annihilated (the energy lost through intramolecular relaxation ultimately manifests as heat). A simplified mathematical expression accounting for S–S annihilation in a homogeneous ensemble of molecules is as follows:

$$\frac{dn}{dt} = -\frac{n}{\tau} - \gamma n^2 + F. \quad (\text{C3.4.11})$$

The solution to this equation reflects a nonlinear response—the kinetics $n(t)$ are strongly dependent on the magnitude of F and/or the initial conditions $n(t=0)$.

S–S annihilation phenomena can be considered as a powerful tool for investigating the exciton dynamics in molecular complexes [26]. However, in systems where that is not the objective it can be a complication one would prefer to avoid. To this end, a measure of suitably conservative excitation conditions is to have the parameter $\sigma\phi\tau < 0.01$. Here τ is the effective rate of intrinsic energy dissipation in the ensemble if the excitation is by CW light, and $\tau = \tau_{\text{las}}$ is the

-8-

pulse duration if the source is a laser pulse faster than all dissipation processes: $1/\tau_{\text{las}} \gg \langle 1/\tau_i \rangle$.

One other common source of nonlinear response, singlet–triplet annihilation, is often the reason for a discrepancy between fluorometric and absorption kinetic measurements [27, 28 and 29].

C3.4.5 COHERENT ENERGY TRANSFER

C3.4.5.1 DIMER

Let us first consider the simplest case of two identical chromophores (or molecules) that are in interaction with each other such that their ‘dimer’ Hamiltonian can be written [30, 31] as

$$\hat{H} = \hat{H}_1 + \hat{H}_2 + \hat{V}_{12} \quad (\text{C3.4.12})$$

where \hat{H}_1 and \hat{H}_2 are the Hamiltonians of the individual molecules and \hat{V}_{12} is the operator for their dipole–dipole interaction

$$\hat{V}_{12} = (4\pi\epsilon_0)^{-1} [(\mu_1 \cdot \mu_2) R_{12}^{-3} - 3(\mu_1 \cdot \mathbf{R}_{12})(\mu_2 \cdot \mathbf{R}_{12}) R_{12}^{-5}] \quad (\text{C3.4.13})$$

each μ being an operator on molecular wavefunctions (see also figure C3.4.1 and equation (C3.4.3) for notation and definitions). The solution of the Schrödinger equation with the above Hamiltonian yields two wavefunctions:

$$\left. \begin{aligned} \Psi^+ &= \frac{1}{\sqrt{2}}(\Psi_{1a} + \Psi_{2a}) \\ \Psi^- &= \frac{1}{\sqrt{2}}(\Psi_{1a} - \Psi_{2a}) \end{aligned} \right\} \quad \Psi_{1a} = \varphi_{10}\varphi_{2a} \text{ and } \Psi_{2a} = \varphi_{1a}\varphi_{20}. \quad (\text{C3.4.14})$$

Here ϕ_{10} , ϕ_{20} and ϕ_{1a} , ϕ_{2a} are the wavefunctions of the non-excited and excited molecules if there is no interaction between them. In the case we consider the molecules do interact and as a result the dimer exhibits properties different from those of the monomers it comprises. In particular, the energy level of the excited state is different from the monomer—it is *split* into two states:

$$\left. \begin{aligned} h\nu^+ &= h\nu_{0a} + V_{12} \\ h\nu^- &= h\nu_{0a} - V_{12} \end{aligned} \right\} \quad (\text{C3.4.15})$$

termed symmetric and antisymmetric respectively. In (C3.4.15), V_{12} is the energy of dipole–dipole interaction between the monomers (the expectation value in the monomer product basis set of the operator given by (C3.4.13)). The dimer has a common ground state and excitation may terminate in either the ν^+ or ν^- excited state (see the solid arrows in [figure C3.4.3](#)). The transition dipole moments of these transitions are defined as:

-9-

$$\left. \begin{aligned} \mu^+ &= \mu_1 + \mu_2 \\ \mu^- &= \mu_1 - \mu_2 \end{aligned} \right\} \quad (\text{C3.4.16})$$

(see [figure C3.4.4](#)). The dipole strength of these transitions are calculated as

$$D^\pm = D_{0a}(1 \pm \cos \alpha) \quad (\text{C3.4.17})$$

where $\cos \alpha = \hat{\mu}_1 \cdot \hat{\mu}_2$ and α expresses the angle between transition dipole moments of unperturbed molecules, D_{0a} is the dipole strength of the individual monomer. One can see from (C3.4.17) that the total dipole strength is conserved, i.e. $D^+ + D^- = 2D_{0a}$. The singly excited dimer, in either its symmetric or antisymmetric state, can capture another photon and undergo transition to the doubly excited state ($\phi_{1a} \phi_{2a}$) of energy $2h\nu_{0a}$. Depending in what excited state the dimer was originally, ν^+ or ν^- , the resonance energy for the double excitation would be ν^- or ν^+ respectively, see [figure C3.4.3](#).

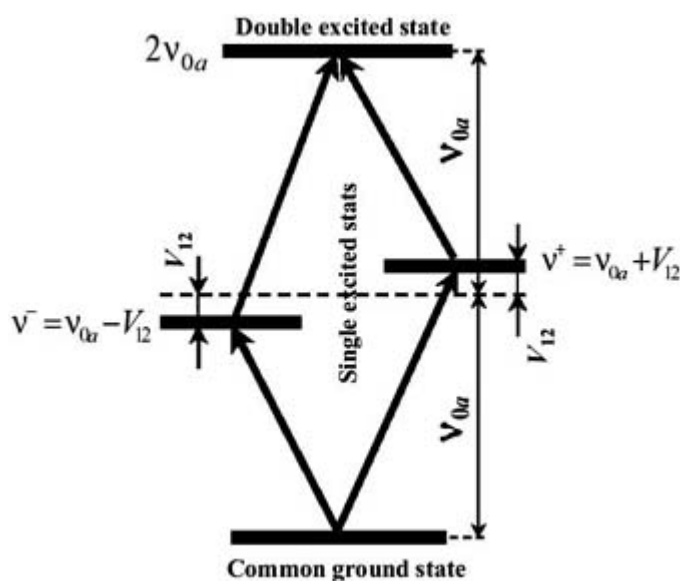


Figure C3.4.3. Energy levels of a dimer (complete description in text).

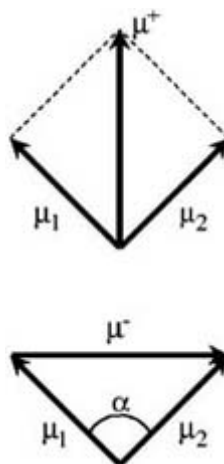


Figure C3.4.4. Definition of the dimer transition dipole moments μ^+ and μ^- on the basis of the monomer transition dipole moments μ_1 and μ_2 .

Now let us consider the implications of these results for energy transfer. First we recognize that there is no directed energy transfer of the form considered in the incoherent case. Molecules in the dimer cannot be recognized as well defined separate entities that can capture and translate excitation from one to another. The captured excitation belongs to the dimer, in other words, it is shared by both molecules. The only counterpart to energy migration relates to ‘internal’ transitions between the dimer states ν^+ and ν^- . In other words, if in the case of incoherent transfer, energy migrates between two or more distinct *sites* and an exciton is a localized entity that can at any time be localized on either of the molecules, then in the case of a dimer we have to change our language and speak about an internal transfer between energy *states*, as the excitation is delocalized between the contributing molecules [32].

Before moving further to the multimer case we should outline one further significant feature of a dimer. Dimer excited states have a well defined rotational strength [33, 34];

$$R_{\pm} = \pm \frac{\pi}{2} \nu_0 \mu^2 R(\hat{R} \cdot \hat{\mu}_2 \times \hat{\mu}_1) \quad (\text{C3.4.18})$$

i.e. the dimer has a definite circular dichroism (CD) in each of its excited bands, whereas the monomers it comprises may not. This is a useful test to verify if one has dimers or monomers in a sample. To confirm the presence of a dimer, the spectrum should exhibit two bands (‘+’ and ‘-’) manifesting CD of equal magnitude and opposite sign. Moreover, the dimer transition dipole moments μ^+ and μ^- are perpendicular to each other (see figure C3.4.4), yielding the linear anisotropy $r = (3\cos^2 90^\circ - 1)/5 = -0.2$ when the excitation is in the ‘+’ band and the response is recorded in the ‘-’ band. Also, the bleaching of the ground state measured in the ‘+’ and ‘-’ bands must happen synchronously because their ground state origin is common, whereas the ‘antibleach’ (absorption from the excited state into a higher state) and also stimulated emission would reflect the kinetics of an internal energy transfer between the dimer states.

Recent theoretical [35, 36 and 37] and experimental [38] research has revealed anomalous behaviour of the dimer anisotropy under certain excitation conditions. If the dimer is excited by broadband light that covers both excitonic transitions, or by a relatively narrow band properly positioned between the maxima of the excitonic transitions, the

anisotropy would have the maximum value $r = 0.7$. This magnitude far exceeds the theoretical limit $r = 0.4$ for uncoupled molecules.

C3.4.5.2 MULTIMER

When more than two chromophores exhibit significant coupling, the ensemble can be described as multimer. This means that the energy levels of individual molecules are replaced by a number of excitonic levels following the same rule as for a dimer and the wavefunctions of the multimer are linear combinations of the wavefunctions for the unperturbed molecules [30, 39, 40]. The number of excitonic levels is equal to the number of chromophores if there is no degeneracy. The light-harvesting antennae of photosynthetic bacteria and some other photosynthetic preparations are believed to be an example of a multimer [39, 40 and 41]. Energy migration in the multimer proceeds in a form of energy equilibration between excitonic levels in the same manner as described above.

C3.4.6 EXCHANGE MECHANISM OF ENERGY TRANSFER IN FORBIDDEN TRANSITIONS

In the previous section we analysed cases of energy transfer where transitions between excited and non-excited states are allowed, i.e. the common case of dipole–dipole interaction. In 1953 Dexter [42] offered a theory that revealed the possibility of energy transfer in the case of dipole-forbidden transitions. The major idea of this theory is that when the electron distributions of the donor and acceptor are close enough to strongly overlap, the energy of electronic excitation can pass directly to the acceptor, essentially being channelled by the overlapped electron clouds. According to this theory the probability of energy transfer is described as

$$P = \frac{2\pi}{\hbar} Z^2 \int F_d(E) F_a(E) dE \quad (\text{C3.4.19})$$

where

$$Z^2 \approx Y \frac{e^4}{k^2 R_0^2} \exp\left(-\frac{2R}{L}\right). \quad (\text{C3.4.20})$$

Here the parameter k is the dielectric constant of the medium; e is the charge of the electron; R_0 is the critical radius; R is the distance of donor–acceptor separation; and L is a parameter introduced as ‘an effective Bohr radius for the excited donor and unexcited acceptor’ (see the original paper, [42]). The constant of proportionality Y is a dimensionless scaling entity $\ll 1$. The coupling integral between the donor and acceptor (written in the energy domain) is analogous to the integrals introduced in (C3.4.1). The problem with this integral is that $F_a(E)$, which represents the absorption spectrum of the acceptor, cannot always be measured directly from experiment (if the transition is forbidden) but must rather be calculated—and the latter can be quite a complicated procedure [42, 43].

One can see that the Dexter exchange mechanism is exponentially dependent on the distance between the donor and acceptor, and as such it begins to play a visible role only at very short distances when the electron clouds begin to

overlap strongly. The exponential distance dependence manifest in (C3.4.20) essentially reflects a typical asymptote for the wavefunctions of the molecular orbitals. It should be emphasized that the Dexter mechanism will operate over short distances both for dipole–dipole forbidden and allowed transitions. With increasing distance between the donor and acceptor, the Förster mechanism will come into play for dipole–dipole allowed transitions, whereas in the forbidden case one should account for transitions via higher multipoles. A review on further developments in this area, including superexchange mechanisms, can be found in [44], and for an account of how the Dexter and Förster mechanisms seamlessly merge for the dipole-allowed case, the reader is referred to [45].

C3.4.7 NUCLEAR MOTIONS AND ENERGY TRANSFER

With the development of femtosecond laser technology it has become possible to observe in resonance energy transfer some apparent manifestations of the coupling between nuclear and electronic motions. For example in photosynthetic preparations such as light-harvesting antennae and reaction centres [32, 46, 47 and 49] such observations are believed to result either from oscillations between the coupled excitonic levels of dimers (generally multimers), or the nuclear motions of the chromophores. This is a subject that is still very much open to debate, and for extensive discussion we refer the reader for example to [46, 47, 50, 51 and 53]. A simplified view of the subject can nonetheless be obtained from the following semiclassical picture.

In light of the theory presented above one can understand that the rate of energy delivery to an acceptor site will be modified through the influence of nuclear motions on the mutual orientations and distances between donors and acceptors. One aspect is the fact that ultrafast excitation of the donor pool can lead to collective motion in the excited donor wavepacket on the potential surface of the excited electronic state. Another type of collective nuclear motion, which can also contribute to such observations, relates to the low-frequency vibrations of the matrix structure in which the chromophores are embedded, as for example a protein backbone. In the latter case the matrix vibration effectively causes a collective motion of the chromophores together, without direct involvement on the wavepacket motions of individual chromophores. For all such reasons, nuclear motions cannot in general be neglected. In this connection it is notable that observations in protein complexes of low-frequency modes in the range 40–150 cm⁻¹ reflect vibrational periods of about 200–800 fs, comparable to typical rates of donor–acceptor energy transfer.

C3.4.8 SPECTROSCOPIC METHODS AND TECHNIQUES

C3.4.8.1 PUMP-PROBE ABSORPTION

The spectroscopic methods called ‘pump-probe absorption’, or ‘pump-probe transient absorption’ involve the use of at least two laser pulses. One pulse excites the sample and the second probes changes in optical properties caused by the first pulse—thus, the first pulse is called the ‘pump’ and the second one the ‘probe’. Obviously there needs to be a time interval between these pulses as the pump pulse precedes the probe. A typical configuration for a one-colour pump-probe installation (probing at the same wavelength as the excitation) is presented in [figure C3.4.5](#). Experiments based on probing at a different wavelength from the excitation require the use of at least two source laser beams, though the two-colour experimental set-up has the same principal elements. In either case the time resolution is defined by the difference in the optical paths to the sample of the pump and probe pulses. The book [54] can be recommended

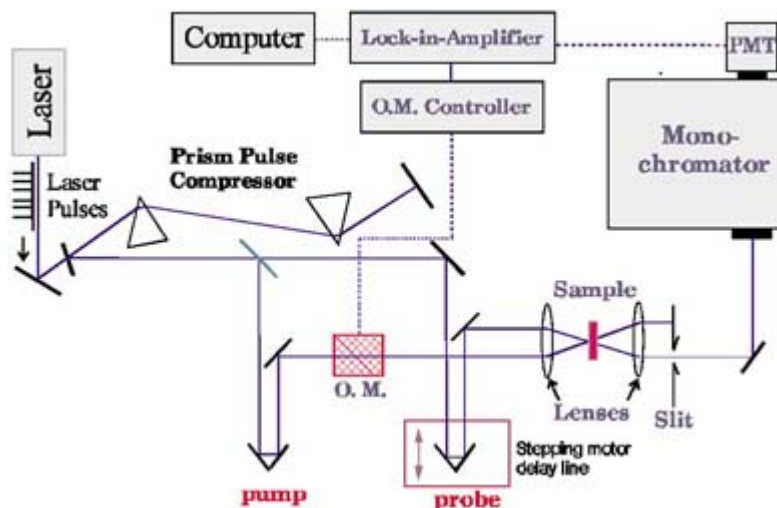


Figure C3.4.5. Typical scheme of a single-colour pump-probe experiment utilizing lock-in-amplifier detection.

Optical detectors can routinely measure only intensities (proportional to the square of the electric field), whether of optical pulses, CW beams or quasi-CW beams; the latter signifying conditions where the pulse train has an interval between pulses which is much shorter than the response time of the detector. It is clear that experiments must be designed in such a way that pump-induced changes in the sample cause changes in the intensity of the probe pulse or beam. It may happen, for example, that the absorption coefficient of the sample is affected by the pump pulse. In other words, due to the pump pulse the transparency of the sample becomes larger or smaller compared with the *unperturbed sample*. Let us stress that even when the optical density (OD) of the sample is large, let us say $OD \sim 1$, and the pump-induced change is relatively weak, say 10^{-4} , it is the latter that carries positive information.

Thus we are challenged by the problem of measuring a small signal against the background of one much stronger. The problem is usually solved by one of two means: (a) lock-in-amplifier detection; and (b) a boxcar type of detection (to some extent we can include double-input optical multichannel detection in this category).

- (a) *Lock-in-amplifier detection.* This method involves ‘chopping’ or rapidly switching the pump beam by a mechanical chopper or by an electro-optic or acousto-optic modulator. Figure C3.4.5 shows the set-up where this chopping is achieved by an optical modulator (OM) of the above kind. The chopping introduces alternating periods of time when the pump beam is affecting the sample and when it is not, causing the detected signal to fluctuate at the chopping frequency. In that case the detector, for example a photodiode or photomultiplier tube (PMT), would see a change in the intensity of the probe beam corresponding to the value of 10^{-4} OD. The next stage is to filter out this signal. This is achieved by using the device called a lock-in-amplifier, which is actually an amplifier with a very narrow spectral bandwidth. Tuned into resonance with the chopping frequency, it will register only the part of the total detected signal that is modulated by the chopping frequency, dumping all other components.
- (b) *Boxcar detection.* This also uses signals associated with ‘pump on’ versus ‘pump off’ conditions, but in a different manner. Whereas the lock-in detector measures *differences* in the intensity of the probe beam, the boxcar detector can measure signal *ratio*. The latter scheme involves the use of two channels with two

time/space gates. Through one gate the boxcar records the unperturbed probe beam (with the pump off), whereas through the other it records the probe beam with the pump on. The ratio of the perturbed over the unperturbed signals from the probe beam is the desired output.

The gates referred to above can be created in various ways. For example, suppose that the probe beam goes through the sample, but only half of its physical width (in the sample) is crossed with the pump beam. Now, if we have two photodiodes, one can measure the intensity of the perturbed part of the probe beam, whilst the second measures the unperturbed part; as a result of creating spatial gates, the two recorded output signals can be used to measure the

requisite ratio. At the same time, the signals from the detectors can be gated in time to improve the signal/noise ratio, i.e. the boxcar records signals only in the short time gate that opens in the presence of the light pulse signal. The rest of the time the gate is closed, and the boxcar receives no input. Averaging over a large number of input signals is another option that is used in boxcar instrumentation to improve signal/noise ratio.

C3.4.8.2 UP/DOWN CONVERSION

The nonlinear optical techniques of up- and down-conversion are based on mixing optical beams in a suitable crystal (BBO, LiNbO₃, KDP, etc) with the generation of new optical frequencies: the physical principle is as follows. If two beams having optical frequencies ω_1 , ω_2 and wavevectors k_1 , k_2 are mixed in a nonlinear optical crystal at the appropriate angle, a new optical frequency ω_3 can be coherently generated with the following conditions satisfied:

$$\omega_3 = \omega_1 \pm \omega_2 \quad \text{and} \quad k_3 = k_1 \pm k_2.$$

The sum-frequency case of $\omega_3 = \omega_1 + \omega_2$ is called up-conversion, the difference-frequency $\omega_3 = \omega_1 - \omega_2$ down-conversion, reflecting the increase or decrease of the generated optical frequency ω_3 from the input frequencies ω_1 and ω_2 .

Now, if we consider the ω_1 input as sample fluorescence caused by the excitation pulse (the pump) and ω_2 as the probe pulse, the process of up/down conversion affords a means of analysing the kinetics of sample fluorescence through observing the intensity of the ω_3 output. In this case the time selection happens in the nonlinear crystal, not in the sample. It is as if the probe beam creates a very narrow time-gate that results in generation of the ω_3 signal only when the gate is open. The kinetics is measured by delaying the time-gate versus the time of sample excitation by the pump pulse.

C3.4.8.3 STREAK CAMERA

Monitoring the kinetics of energy transfer in many systems calls for ultrafast (sub-picosecond) time resolution. Streak camera detection relies on fast electronics to achieve real-time detection of an ultrafast signal registering the transfer—usually a fluorescence signal. Modern streak cameras allow such measurements to be performed with picosecond resolution. The principle of detection is simple: the pump pulse excites the sample and triggers a fast electronic camera similar to the tube in a common oscilloscope. Fluorescence from the sample is collected by the input slit of the camera; the photons either hit the camera photocathode directly or are preamplified by use of an electro-optic amplifier. The beam of electrons so produced, created by the initial pulse of fluorescence and featuring its temporal profile, propagates towards the display phosphor screen. *En route* the beam is deflected by an electric field created by a fast generator triggered by the pump pulse. The result is a track on the phosphor screen with an intensity proportional to

the intensity of the fluorescence. Usually the streak camera is coupled with a CCD camera or a diode array to record this track.

C3.4.8.4 SAMPLE CELL

When performing measurements of energy transfer one has to pay attention to a number of factors that could invalidate the data. In particular, the ‘reset time’ of the sample (the time for conversion of excited molecules back to the ground equilibrated state) could be larger than the period between laser pulses. This creates a build-up of the excited states and/or intermediate products—the latter significantly complicating analysis of the collected data. Yet another complication relates to the local heating of the sample, which modifies the local optical properties from

those which apply under equilibrium ambient conditions (thermal lensing, for example). The common way to eliminate these problems is to refresh the sample in the laser-illuminated area at a rate equal to or faster than the laser pulse repetition. This can be achieved using a flow cell, shaking cell, spinning cell or flow jet. Another kind of problem occurs if the laser pulses have significant noise (exhibiting spatial/temporal spikes for example) and no longer can be considered as propagating with a smooth Gaussian profile. This is a problem which is difficult to overcome without improving the quality of the pulses.

In a flow cell the sample flows through a cuvette by use of a pump, the most popular kind being a peristaltic pump. A *shaking cell* is usually a cell that resides on a mount that can move laterally in a plane perpendicular to the laser beam(s). A *spinning cell* is a disc-like cell mounted on a motor shaft that rotates with a speed up to thousands of revolutions per minute. Finally the *flow jet* is an assembly in which liquid sample is ejected through the special jet nozzle to create a uniform fast stream of sample. The laser beam is focused in this stream. The spinning cell is generally the best choice to achieve the fastest refreshing rate in anaerobic conditions without damaging the sample molecules.

¹ The population of excited states (n) and the probability to find excitation on an individual molecule (p) are related by $n = pN$, where N is the total number of molecules.

REFERENCES

- [1] Förster Th 1948 Zwischenmolekulare Energiewanderung und Fluoreszenz *Ann. Phys.* **2** 55–75
 - [2] Agranovich V M and Galanin M D 1982 *Electronic Excitation Energy Transfer in Condensed Matter* (Amsterdam: Elsevier/North-Holland)
 - [3] Förster Th 1965 Delocalized excitation and excitation transfer *Modern Quantum Chemistry* ed O Sinanoglu (New York: Academic) pp 93–137
 - [4] van der Meer B W, Coker J III and Chen S-Y S 1994 *Resonance Energy Transfer: Theory and Data* (New York: VCH)
 - [5] Clegg R M 1998 Fluorescence resonance energy transfer *Fluorescence Imaging Spectroscopy and Microscopy* ed X F Wang and B Herman, pp 179–251
 - [6] Andrews D L and Juzeliunas G 1992 Intermolecular energy transfer: retardation effects *J. Chem. Phys.* **96** 6606–12
 - [7] Juzeliunas G and Andrews D L 1999 Unified theory of radiative and radiationless energy transfer *Resonance Energy Transfer* ed D L Andrews and A A Demidov (New York: Wiley) pp 65–107
 - [8] Birks J B 1970 *Photophysics of Aromatic Molecules* (London: Wiley)
-
- [9] Demidov A A and Ivanov I G 1989 Laser fluorometric method for measuring the concentrations of chlorophyll *a* and pheophytin *a* in acetone solutions *Biol. Bull. Acad. Sci. USSR* **16** 228–34
 - [10] 1975 *Bioenergetics of Photosynthesis* ed Govindjee (New York: Academic)
 - [12] Dale R E, Eisinger J and Blumberg W E 1979 The orientation freedom of molecular probes. The orientation factor in intramolecular energy transfer *Biophys. J.* **26** 161–94
 - [13] Basiev T T, Orlovskii Y V and Privis Y S 1996 High order multipole interaction in nanosecond Nd–Nd energy transfer *J. Luminescence* **69** 187–202
 - [14] Scholes G D and Andrews D L 1997 Damping and higher multipole effects in the quantum electrodynamic model for electronic energy transfer in the condensed phase *J. Chem. Phys.* **107** 5374–84
 - [15] Scholes G D, Clayton A H A and Ghiggino K P 1992 On the rate of radiationless intermolecular energy transfer *J. Chem. Phys.* **97** 7405–13
 - [15] Kulak L and Bojarski C 1995 Forward and reverse electronic-energy transport and trapping in solution: 1. Theory *Chem. Phys.* **191** 43–66
 - [16] Pailotin G, Geacintov N E and Breton J 1983 A master equation theory of fluorescence induction, photochemical yield,

and singlet-triplet exciton quenching in photosynthetic systems *Biophys. J.* **44** 65–77

- [17] Pullerits T, Visscher K J, Hess S, Sundstrom V, Freiberg A, Timpmann K and Van Grondelle R 1994 Energy-transfer in the inhomogeneously broadened core antenna of purple bacteria—a simultaneous fit of low-intensity picosecond absorption and fluorescence kinetics *Biophys. J.* **66** 236–48
- [18] Demidov A A 1999 Use of Monte-Carlo method in the problem of energy migration in molecular complexes *Resonance Energy Transfer* ed D L Andrews and A A Demidov (New York: Wiley) pp 435–65
- [19] Agranovich V M, Efremov N A and Kirsanov V V 1980 Computer simulation of kinetics of excitation bimolecular quenching by Monte-Carlo method *Fiz. Tverd. Tela* **22** 2118–27
- [20] Barzykin A V, Barzykina N S and Fox M A 1992 Electronic excitation transport and trapping in micellar systems—Monte-Carlo simulations and density expansion approximation *Chem. Phys.* **163** 1–12
- [21] Fetisova Z G, Borisov A Y and Fok M V 1985 Analysis of structure-function correlations in light-harvesting photosynthetic antenna—structure optimization parameters *J. Theoret. Biol.* **112** 41–75
- [22] Fetisova Z G, Fok M V, Shibaeva L V and Borisov A Y 1984 Ways of optimizing conversion of light energy in the initial-stages of photosynthesis: 2. Optimizing the structure of the grid of a homogeneous photosynthetic unit *Molecular Biol.* **18** 1359–65
- [23] Levshin V L 1925 Poliarizovannaia fluorescenciia i phosphorescenciia rastvorov krasok *Zhurnal Russkogo Physiko-Chimicheskogo Obschestva, Fizika (Russian)* **57** 283–300
- [24] Lewshin W L 1925 Polarisierete Fluoreszenz und Phosphoreszenz der Farbstofflosungen: IV *Z. Phys.* **32** 307–26
- [25] Perrin F 1929 La fluorescence des solutions *Ann. Phys.* **12** 169–275
- [26] Valkunas L, Trinkunas G and Liuolia V 1998 Exciton annihilation in molecular aggregates *Resonance Energy Transfer* ed D L Andrews and A A Demidov (New York: Wiley) pp 244–307
- [27] Valkunas L, Liuolia V and Freiberg A 1991 Picosecond processes in chromatophores at various excitation intensities *Photosynthesis Res.* **27** 83–95
- [28] Kolubayev T, Geacintov N E, Paillotin G and Breton J 1985 Domain sizes in chloroplasts and chlorophyll-protein complexes probed by fluorescence yield quenching induced by singlet-triplet exciton annihilation *Biochimica Biophys. Acta* **808** 66–76
- [29] van Mourik F, van der Oord C J R, Visscher K J, Parkes-Loach P S, Loach P A, Visschers R W and van Grondelle R 1991 Exciton interactions in the light-harvesting antenna of photosynthetic bacteria studied with triplet-singlet spectroscopy and singlet-triplet annihilation on the b820 subunit form of *Rhodospirillum rubrum* *Biochimica Biophys. Acta* **1059** 111–9
- [30] Cantor C R and Schimmel P R 1980 *Techniques for the Study of Biological Structure and Function* (San Francisco, CA: Freeman)

- [31] Tinoco I 1963 The exciton contribution to the optical rotation of polymers *Radiat. Res.* **20** 133–9
- [32] Pullerits T, Chachisvilis M, Fedchenia I and Sundstrom V 1994 Coherent versus incoherent energy transfer in the light-harvesting complexes of photosynthetic bacteria *Lietuvos Fizikos Zurnalas* **34** 329–38
- [33] Pearlstein R M 1991 Theoretical interpretation of antenna spectra *Chlorophylls* ed H Scheer (Boca Raton, FL: CRC) pp 1047–78
- [34] Craig D P and Thirunamachandran T 1984 *Molecular Quantum Electrodynamics* (New York: Academic)
- [35] Wynne K and Hochstrasser R M 1995 Anisotropy as an ultrafast probe of electronic coherence in degenerate systems exhibiting Raman-scattering, fluorescence, transient absorption and chemical-reactions *J. Raman Spectrosc.* **26** 561–9
- [36] Wynne K and Hochstrasser R M 1993 Coherence effects in the anisotropy of optical experiments *Chem. Phys.* **171** 179–88
- [37] Knox R S and Gulen D 1993 Theory of polarized fluorescence from molecular pairs—Förster transfer at large electronic coupling *Photochem. Photobiol.* **57** 40–3
- [38] Galli C, Wynne K, Lecours S M, Therien M J and Hochstrasser R M 1993 Direct measurement of electronic dephasing using anisotropy *Chem. Phys. Lett.* **206** 493–9
- [39] Gnanakaran S, Haran R, Kumble R and Hochstrasser R M 1999 Energy transfer and localization: application to photosynthetic systems *Resonance Energy Transfer* ed D L Andrews and A A Demidov (New York: Wiley) pp 308–65
- [40] Savikhin S, Buck D R and Struve W S 1999 The Fenna–Mathews–Olson protein: a strongly coupled photosynthetic

antenna *Resonance Energy Transfer* ed D L Andrews and A A Demidov (New York: Wiley) pp 399–434

- [41] Durrant J R, Klug D R, Kwa S L S, Van Grondelle R, Porter G and Dekker J P 1995 A multimer model for P680, the primary electron donor of photosystem II *Proc. Natl Acad. Sci. (USA)* **92** 4798–802
- [42] Dexter D L 1953 A theory of sensitized luminescence in solids *J. Chem. Phys.* **21** 836–50
- [43] Lax M 1952 The Franck–Condon principle and its application to crystals *J. Chem. Phys.* **20** 1752–60
- [44] Scholes G D 1999 Theory of coupling in multichromophoric systems *Resonance Energy Transfer* ed D L Andrews and A A Demidov (New York: Wiley) pp 212–43
- [45] Scholes G D and Ghiggino K P 1994 Electronic interactions and interchromophore excitation transfer *J. Phys. Chem.* **98** 4580–90
- [46] Kumble R, Palese S, Visschers R W, Dutton P L and Hochstrasser R M 1996 Ultrafast dynamics within the B820 subunit from the core (LH-1) antenna complex of *Rs. rubrum* *Chem. Phys. Lett.* **261** 396–404
- [47] Vos M H, Jones M R, Breton J, Lambry J C and Martin J L 1996 Vibrational dephasing of long-lived and short-lived primary donor excited-states in mutant reaction centers of *Rhodobacter sphaeroides* *Biochemistry* **35** 2687–92
- [48] Streltsov A M, Aartsma T J, Hoff A J and Shuvalov V A 1997 Oscillations within the B_L absorption band of *Rhodobacter sphaeroides* reaction centers upon 30 femtosecond excitation at 865 nm *Chem. Phys. Lett.* **266** 347–52
- [49] Chachisvilis M and Sundstrom V 1996 Femtosecond vibrational dynamics and relaxation in the core light-harvesting complex of photosynthetic purple bacteria *Chem. Phys. Lett.* **261** 165–74
- [50] Vos M H, Jones M R and Martin J L 1998 Vibrational coherence in bacterial reaction centers: spectroscopic characterisation of motions active during primary electron transfer *Chem. Phys.* **233** 179–90
- [51] Kim Y R, Share P, Pereira M, Sarisky M and Hochstrasser R M 1989 Direct measurements of energy-transfer between identical chromophores in solution *J. Chem. Phys.* **91** 7557–62
- [52] Souaille M and Marchi M 1997 Nuclear dynamics and electronic transition in a photosynthetic reaction center *J. Am. Chem. Soc.* **119** 3948–58.
- [53] Jonas D M, Lang M J, Nagasawa Y, Joo T and Fleming G R 1996 Pump-probe polarization anisotropy study of femtosecond energy transfer within the photosynthetic reaction-center of *Rhodobacter sphaeroides* R26 *J. Phys. Chem.* **100** 12660–73
- [54] Fleming G R 1986 *Chemical Applications of Ultrafast Spectroscopy* (Oxford: Oxford University Press)

FURTHER READING

We would like to recommend the following books for further reading. Recently (1999) Wiley published a book entitled *Resonance Energy Transfer* (ed D L Andrews and A A Demidov). This book contains a detailed overview of the major subjects briefly discussed in the present section. Previously, van der Meer *et al* [4] published a good work concerning the Förster mechanism of energy transfer, and in particular the impact of the orientation factor on the efficiency of energy transfer. Those who are interested in theoretical aspects of energy transfer in condensed matter are referred to the classic book ‘*Electronic Excitation Energy Transfer in Condensed Matter*’ by Agranovich and Galanin [2]. Copious examples of energy transfer in photosynthetic systems are presented in the books entitled *Bioenergetics of Photosynthesis* [10] and *Excitation Energy and Electron Transfer in Photosynthesis* (1987, edited by Govindjee, Kluwer Academic Publishers). Finally, a great many details and tips on practical applications of laser spectroscopy for the investigation of energy transfer, and its instrumentation, can be found in the book *Chemical Applications of Ultrafast Spectroscopy* by Fleming [54].

C3.5 Vibrational energy transfer in condensed phases

Lawrence K Iwaki and Dana D Dlott

C3.5.1 INTRODUCTION

Vibrational energy relaxation (VER) of molecules in condensed phases is a fundamental dynamical process [1, 2, 3, 4, 5 and 6]. Isolated molecules can undergo only *intramolecular* vibrational energy redistribution (IVR), and cannot lose vibrational energy except by slow radiative processes. VER is usually associated with condensed phases or high pressure gases [7]. VER refers to loss of energy from a specific vibrational mode (the ‘system’) to some or all of the other mechanical degrees of freedom (the ‘bath’). VER in condensed phases ordinarily occurs on the 10^{-13} – 10^{-9} s time scale, although slower VER has been observed in diatomics.

VER occurs as a result of fluctuating forces exerted by the bath on the system at the system’s oscillation frequency Ω [5]. Fluctuating dynamical forces are characterized by a force–force correlation function. The Fourier transform of this force correlation function at Ω , denoted $\eta(\Omega)$, characterizes the quantum mechanical frequency-dependent friction exerted on the system by the bath [5, 8].

The multiple roles of VER in essentially all condensed phase chemical processes have been extensively discussed [8]. In chemical reactions, the ‘system’ is the specific mode of the reactant (or a coupled set of reactant and solvent modes [8]) associated with the reaction coordinate. Chemical reactions are ‘catalysed’ by vibrational energy. The system becomes activated by vibrational energy from the bath. Then the barrier is crossed. Then that vibrational energy plus the enthalpy of reaction flows back into the bath (figure C3.5.1). Much has been written about dynamical effects of VER on barrier crossings [8, 9 and 10]. Chemical reactions are slower when VER is too slow, due to multiple barrier recrossings. Reactions are faster when VER is too fast, due to VER during the barrier crossing. The ‘Kramers turnover’ between these two regimes is located at the maximum possible rate, which is also that given by transition-state theory [8, 9] (figure C3.5.1). The VER rate is varied in practice by pressure-tuning solvent density, as in classic studies of photoisomerization of stilbene [11, 12] and boat–chair isomerization of cyclohexane [13, 14].

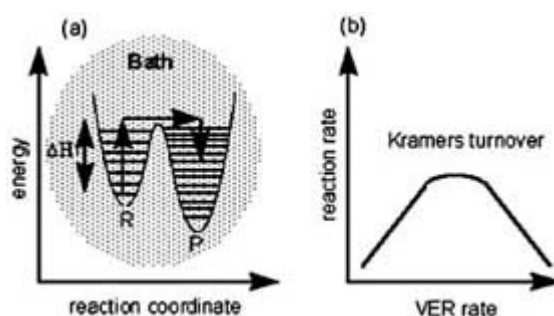


Figure C3.5.1. (a) Vibrational energy catalyses chemical reactions. The reactant R is activated by taking up the enthalpy of activation ΔH^\ddagger from the bath. That energy plus the heat of reaction is returned to the bath after barrier crossing. (b) VER influences chemical reaction rates by modulating the system during barrier crossing. For a particular VER rate, the reaction rate has a maximum at the Kramers turnover.

C3.5.2 BRIEF HISTORY OF VER

Condensed phase vibrational or vibronic lineshapes (vibronic transitions create vibrational excitations of electronic excited states) rarely provide information about VER (see example C3.5.6.4). Experimental measurements of VER need much more than just the vibrational spectrum. The earliest VER measurements in condensed phases were ultrasonic attenuation studies of liquids [15], which provided an overall relaxation time for slowly (>10 ns) relaxing small molecule liquids.

Lasers revolutionized VER measurements, providing the needed time resolution and specific state resolution. Even early nanosecond solid-state lasers (not tunable) could produce large vibrational populations in liquids or solids via stimulated Raman scattering (SRS) [16]. Early infrared (IR) lasers had limited tuning ranges, but could pump certain molecules (e.g. CO laser pumping solid CO [17]). An excellent history of early VER studies of small molecules is given by Oxtoby [5]. An amazing result from this era, due to Brueck and Osgood [18], is the incredible 56 s VER lifetime of liquid N₂, showing the average N₂ molecule undergoes $\sim 4 \times 10^{15}$ oscillations before VER.

Other early work, which continues to this day, involved vibronic relaxation [6] of large colored molecules such as chrysene [19], pyrene [20] and perylene [21], due to the relative ease of using visible or near-UV light to pump and probe these systems (see example C3.5.6.5 below).

Major breakthroughs in early ultrafast VER measurements were made in 1972 by Laubereau *et al* [22], who used picosecond lasers in an SRS pump–incoherent anti-Stokes Raman probe configuration, to study VER of C–H groups in ethanol and methanol ($\sim 3000 \text{ cm}^{-1}$), and by Alfano and Shapiro [23], who monitored both the decay of the initially excited C–H stretch excitation and the appearance and subsequent decay of a daughter vibration, a C–H bending vibration ($\sim 1460 \text{ cm}^{-1}$). Several reviews described these early studies of liquids [1, 6, 16].

Another important breakthrough occurred with the 1974 development by Laubereau *et al* [24] of tunable ultrafast IR pulse generation. IR excitation is more selective and reliable than SRS, and IR can be used in pump–probe experiments or combined with anti–Stokes Raman probing (IR–Raman method) [16]. Ultrashort IR pulses have been used to study simple liquids and solids, complex liquids, glasses, polymers and even biological systems.

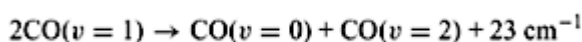
C3.5.3 OVERVIEW OF VER PHENOMENA

C.3.5.3.1 SIMPLE (DIATOMIC) SYSTEMS

The simplest condensed phase VER system is a dilute solution of a diatomic in an atomic (e.g. Ar or Xe) liquid or crystal. Other simple systems include neat diatomic liquids or crystals, or a diatomic molecule bound to a surface. A major step up in complexity occurs with polyatomics, with several vibrations on the same molecule. This feature guarantees enormous qualitative differences between diatomic and polyatomic VER, and casts doubt on the likelihood of understanding polyatomics by studying diatomics alone.

-3-

Diatomic molecules have only one vibrational mode, but VER mechanisms are paradoxically quite complex (see examples C3.5.6.1 and C3.5.6.2). Consequently there is an enormous variability in VER lifetimes, which may range from 56 s (liquid N₂ [18]) to 1 ps (e.g. XeF in Ar [25]), and a high level of sensitivity to environment. A remarkable feature of simpler systems is spontaneous concentration and localization of vibrational energy due to anharmonicity. Collisional up-pumping processes such as



proceed spontaneously with a decrease in enthalpy and a decrease in entropy. For very long-lived vibrations equilibrium may be reached with quite highly excited vibrational states. Anex and Ewing [26] have observed collisional up-pumping of CO up to $v = 20$ for CO in liquid N₂, and high overtone emission was observed in a

monolayer of CO on NaCl by Chang and Ewing [27].

VER of diatomic molecules bound to surfaces [28] was first studied by Heilweil and co-workers [29], who used ultrafast IR to find a surprisingly long VER lifetime of 150 ps for OH stretching vibrations ($\sim 3650 \text{ cm}^{-1}$) on the surface of silica [30]. Guyot-Sionnest *et al* [31] found nanosecond lifetimes for H chemisorbed to Si(111) surfaces ($\sim 4000 \text{ cm}^{-1}$). Surface-bound diatomics have much shorter lifetimes on metals than on dielectrics or semiconductors, due to interactions with conduction electrons [32]. For example, when CO is bound to NaCl (100), its VER lifetime is in the millisecond range [27], but when CO is bound to Pt [33], its lifetime is a few picoseconds.

C.3.5.3.2 COMPLEX (POLYATOMIC) SYSTEMS

In polyatomics, a completely different VER process may occur, termed ‘ladder relaxation’, where energy is transferred from one excited vibrational mode to another mode on the same molecule (the ‘rungs’ of the ‘vibrational ladder’), while the bath takes up the remaining energy [34, 35]. Vibrational energy running down this ladder is termed a ‘vibrational cascade’. Ladder processes are so efficient that VER lifetimes of polyatomics in condensed phases hardly ever exceed one nanosecond. Due to short VER lifetimes in polyatomics, intermolecular energy transfer is thought to be noncompetitive with intramolecular ladder processes. However, there are not many data yet, and intermolecular transfer has been directly observed in a few studies. Apkarian and co-workers observed energy transfer among CH_3F molecules in a rare-gas matrix, resulting in vibrational up-pumping to $\nu = 2$ [36]. Ambroseo and Hochstrasser [37] observed energy transfer from pyrrole to benzene. Hong *et al* [38] observed energy transfer from OH stretch of alcohols to C–H stretch of nitromethane.

The most powerful technique for studying VER in polyatomic molecules is the IR–Raman method. Initial IR–Raman studies of a few systems appeared more than 20 years ago [16], but recently the technique has taken on new life with newer ultrafast lasers such as Ti:sapphire [39]. With more sensitive IR–Raman systems based on these lasers, it has become possible to monitor VER by probing virtually every vibration of a polyatomic molecule, as illustrated by recent studies of chloroform [40], acetonitrile [41, 42] (see example C3.5.6.6 below) and nitromethane [39, 43].

There does not yet exist a simple theory, analogous say to Marcus theory for electron transfer, which predicts VER rates of polyatomic molecular vibrations, or how those rates depend on vibrational frequency. A simple rule was proposed by Nitzan and Jortner [35]. Those authors identified three frequency regimes for VER, a lower-energy regime I where short-lived vibrations decay directly into the bath, an intermediate regime II where longer-lived vibrations undergo ladder relaxation and a higher-energy regime III where vibrations undergo ultrafast intramolecular

-4-

vibrational relaxation. This hardly rigorous rule has been observed (always with numerous exceptions) in several systems (see examples C3.5.6.5 and C3.5.6.6).

One simplifying motif for VER studies of polyatomic molecules is to probe a diatomic ligand bound to a more complex system. CO is a ubiquitous ligand and many studies have been made of VER of CO ligands, e.g. $\text{W}(\text{CO})_6$ in CCl_4 , where the VER lifetime is 280 ps but quite dependent on solvent [44, 44] or CO bound to metalloporphyrins [46, 47, 48 and 49] where the VER lifetime is ~ 20 ps and dependent on haem structure. CO has also been used to study VER of biomolecules, specifically haem proteins. For myoglobin and haemoglobin at 300 K, the CO VER lifetime was ~ 20 ps [46, 47, 50]. A study of CO bound to native and genetically engineered haem proteins showed that vibrational energy flows from CO to haem via π -electron coupling, and that VER rate could be influenced by the protein [50]. Recently, VER measurements have been made of proteins themselves. The $\sim 1600 \text{ cm}^{-1}$ amide I stretching vibrations of several proteins have been found to have an ~ 1 ps lifetime [51, 52 and 53].

C3.5.4 THEORY OF VIBRATIONAL ENERGY RELAXATION

Consider an excited condensed-phase quantum oscillator Ω , with reduced mass μ and normal coordinate q_Ω . The bath exerts fluctuating forces on the oscillator. These fluctuating forces induce VER. The quantum mechanical Hamiltonian is [54, 55]

$$\hat{H} = \hat{H}_\Omega(q_\Omega) + \hat{H}_B(Q) + \hat{V}(q_\Omega, Q) \quad (\text{C3.5.1})$$

where $\hat{H}_\Omega(q_\Omega)$ is the Hamiltonian for the oscillator Ω , $\hat{H}_B(Q)$ is the Hamiltonian for the bath, where Q represents collective bath coordinates, and \hat{V} represents the Hamiltonian for interaction between Ω and the bath. In solids these collective bath states are phonons, which extend from zero frequency to a cut-off frequency, the Debye frequency, ω_D . In liquids, the collective states have been termed instantaneous normal modes [56, 57 and 58]. Since these play the same role as phonons in VER [34, 56, 59, 60], we will call them liquid phonons or simply phonons.

The fluctuating forces $F(t)$ on the rigid oscillator Ω are characterized by a time-dependent force-force correlation function [54, 55],

$$\langle \hat{F}(t)\hat{F}(0) \rangle_Q = \frac{\text{Tr}[e^{-\hat{H}_B/kT} \hat{F}(t)\hat{F}(0)]}{\text{Tr}[e^{-\hat{H}_B/kT}]} \quad (\text{C3.5.2})$$

where

$$\hat{F}(t) = e^{i\hat{H}_B t} \hat{F} e^{-i\hat{H}_B t}$$

is the Heisenberg operator for the fluctuating forces and Tr denotes trace.

-5-

Equation (C3.5.2) is a function of bath coordinates only. The VER rate constant is proportional to the Fourier transform, at the oscillator frequency Ω , of the bath force-correlation function. This Fourier transform is proportional as well to the frequency-dependent friction $\eta(\Omega)$ mentioned previously. For example, the rate constant for VER of the fundamental ($\nu = 1$) to the ground ($\nu = 0$) state of an oscillator with frequency Ω is [54]

$$k_{1 \rightarrow 0} = \frac{1}{2\mu\hbar\Omega} \int_{-\infty}^{\infty} dt e^{i\Omega t} \langle \hat{F}(t)\hat{F}(0) \rangle_Q. \quad (\text{C3.5.3})$$

Equation (C3.5.3) shows the VER lifetime can be determined if the quantum mechanical force-correlation function is computed. However, it is at present impossible to compute this function accurately for complex systems. It is straightforward to compute the classical force-correlation function using classical molecular dynamics (MD) simulations. With the classical force-correlation function, a ‘quantum correction’ factor Q is needed [5],

$$k_{1 \rightarrow 0} = \frac{Q}{2\mu\hbar\Omega} \int_{-\infty}^{\infty} dt e^{i\Omega t} C(t) \quad (\text{C3.5.4})$$

where $C(t)$ is the classical force-force correlation function $C(t) = \langle F(t)F(0) \rangle$. For a harmonic bath, Bader and Berne [61] give the exact quantum correction,

$$Q = \frac{\hbar\Omega/kT}{1 - e^{-\hbar\Omega/kT}}. \quad (\text{C3.5.5})$$

Most realistic problems involve an anharmonic bath. How to determine \mathcal{Q} for an arbitrary anharmonic bath is not yet known. Other correction methods have been discussed, including the method proposed by Egelstaff [62].

In diatomic VER, the frequency Ω is often much greater than ω_D , so VER requires a high-order multiphonon process (see example C3.5.6.1). Because polyatomic molecules have several vibrations ranging from higher to lower frequencies, only lower-order phonon processes are ordinarily needed [34]. The usual practice is to expand the interaction Hamiltonian $\hat{V}(q_\Omega, \mathcal{Q})$ in equation (C3.5.2) in powers of normal coordinates [34, 63],

$$\begin{aligned}\hat{V} = & \sum_{\alpha} \frac{\partial \hat{V}}{\partial Q_{\alpha}} \Big|_{\{Q=0\}} Q_{\alpha} + \frac{1}{2!} \sum_{\alpha, \gamma} \frac{\partial^2 \hat{V}}{\partial Q_{\alpha} \partial Q_{\gamma}} \Big|_{\{Q=0\}} Q_{\alpha} Q_{\gamma} \\ & + \frac{1}{3!} \sum_{\alpha, \gamma, \delta} \frac{\partial^3 \hat{V}}{\partial Q_{\alpha} \partial Q_{\gamma} \partial Q_{\delta}} \Big|_{\{Q=0\}} Q_{\alpha} Q_{\gamma} Q_{\delta} \\ & + \frac{1}{4!} \sum_{\alpha, \gamma, \delta, \epsilon} \frac{\partial^4 \hat{V}}{\partial Q_{\alpha} \partial Q_{\gamma} \partial Q_{\delta} \partial Q_{\epsilon}} \Big|_{\{Q=0\}} Q_{\alpha} Q_{\gamma} Q_{\delta} Q_{\epsilon} + \dots\end{aligned}\quad (\text{C3.5.6})$$

For polyatomics, ordinarily only the last two terms of equation (C3.5.6), the cubic and quartic anharmonic terms, need be considered [34]. In a cubic anharmonic process, excited vibration Ω relaxes by interacting with two other states, say another vibration ω and one phonon (or alternatively two phonons). In the quartic process, Ω relaxes by interacting with three other states, say two vibrations and one phonon. The total rate constant for energy loss from Ω for cubic

-6-

coupling was given by Fayer and co-workers as [34]

$$K_{\Omega} = \sum_{\omega} [n_{\omega}(1 + n_{\Omega+\omega})\rho_{\Omega+\omega}C_{\Omega+\omega} + (1 + n_{\omega})(\alpha + n_{|\Omega+\omega|})\rho_{\Omega-\omega}C_{|\Omega-\omega|}] \quad (\text{C3.5.7})$$

where n_{ω} is the thermal occupation number,

$$n_{\omega} = (e^{\hbar\omega/kT} - 1)^{-1} \quad (\text{C3.5.8})$$

ρ_{ω} is the density of phonon states at ω , C_{ω} is a product of coupling constants which contains factors such as $\hbar/2\mu\omega$ and the derivatives of V in equation (C3.5.6) and $\alpha = 1$ if $\Omega > \omega$ or $\alpha = 0$ if $\Omega < \omega$. When $T \rightarrow 0$, all the thermal occupation factors in equation (C3.5.7) vanish, but the VER rate does not vanish. VER is then said to occur via spontaneous emission of phonons. As T is increased, two new thermally activated processes turn on. One involves stimulated phonon emission and the other phonon absorption. Spontaneous and stimulated emission processes convert Ω to a lower-energy vibration ω (down-conversion). Phonon absorption converts Ω to a higher-energy vibration ω (up-conversion). Figure C3.5.2 from Kenkre *et al* [34] shows all the possible VER processes which can occur via cubic or quartic anharmonic coupling.

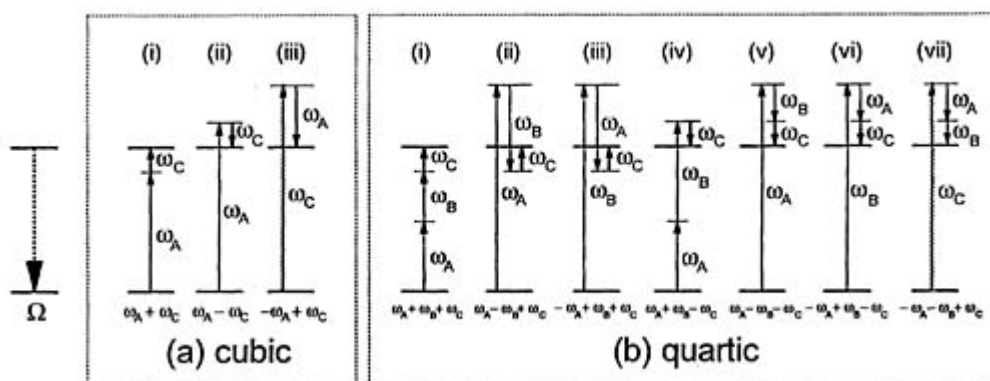


Figure C3.5.2. VER transitions involved in the decay of vibration Ω by cubic and quartic anharmonic coupling (from [34]). Transitions involving discrete vibrations are represented by arrows. Transitions involving phonons (continuous energy states) are represented by wiggly arrows. In (a), the transition denoted (i) is the ladder down-conversion process, where Ω is annihilated and a lower-energy vibration ω_A and a phonon ω_C are created.

-7-

C3.5.5 EXPERIMENTAL TECHNIQUES

Experimental techniques have been reviewed extensively, e.g. [6]. Only a brief discussion will be presented here.

C3.5.5.1 CREATING VIBRATIONAL EXCITATIONS

The easiest method for creating many vibrational excitations is to use convenient pulsed visible or near-UV lasers to pump electronic transitions of molecules which undergo fast nonradiative processes such as internal conversion (e.g. porphyrin [64, 65] or near-IR dyes [66, 67, 68 and 69]), photoisomerization (e.g. stilbene [12] or photodissociation (e.g. HgI_2 [8]). Creating a specific vibrational excitation Ω in a controlled way requires more finesse. The easiest method is to use visible or near-UV pulses to resonantly pump a vibronic transition (e.g. $S_0^0 \rightarrow S_1^\Omega$, where S_n denotes an electronic singlet state) of a coloured molecule [6]. Vibronic relaxation may be complicated by the presence of multiple electronic states (see example C3.5.6.1). Nonresonant pumping using visible or near-IR pulses can pump an S_0^Ω vibration by stimulated Raman scattering (SRS) [70]. SRS excites a specific vibration (with the largest Raman cross-section) but one cannot select which vibration that will be and the high intensities needed for SRS often produce unwanted parasitic optical effects. Resonant mid-IR pumping of specific vibrational transitions ($S_0^0 \rightarrow S_1^\Omega$) seems to produce fewer parasitic effects and has become increasingly popular, given recent improvements in tunable mid-IR pulse generation.

C3.5.5.2 PROBING VIBRATIONAL EXCITATIONS

Vibronic excitations are relatively easy to probe using resonant visible or near-UV processes such as vibronic absorption or vibronic fluorescence. Here we will concentrate mainly on probing vibrational excitations produced by mid-IR pumping. Two powerful but technically difficult techniques which are becoming increasingly important are probing by IR absorption (two-colour IR pump-probe method) or probing by incoherent anti-Stokes Raman scattering with a nonresonant pulse (IR-Raman method).

C3.5.5.3 EXPERIMENTAL TECHNIQUES

Schematic diagrams of modern experimental apparatus used for IR pump-probe by Fayer and co-workers [50] and for IR-Raman experiments by Dlott and co-workers [39] are shown in figure C3.5.3. Ultrafast mid-IR pulse generation by optical parametric amplification (OPA) [71] will not be discussed here. Single-colour IR pump-probe or vibrational echo experiments have been performed with OPAs or free-electron lasers. Free-electron lasers use

relativistic electron bunches travelling through periodic magnetic fields to generate tunable light pulses [72]. Two-colour IR pump–probe experiments use a pair of OPAs or one OPA and a continuous probe laser and a high-speed optical detection scheme [73]. In pump–probe experiments, the modulation of a weaker probe pulse by the pump pulse is monitored with an IR detector. Vibrational echo experiments (see example C3.5.6.4) use essentially the same apparatus, but instead two intense pulses are directed to the sample and a third pulse (the echo), emitted by the sample in a unique direction, is detected.

-8-

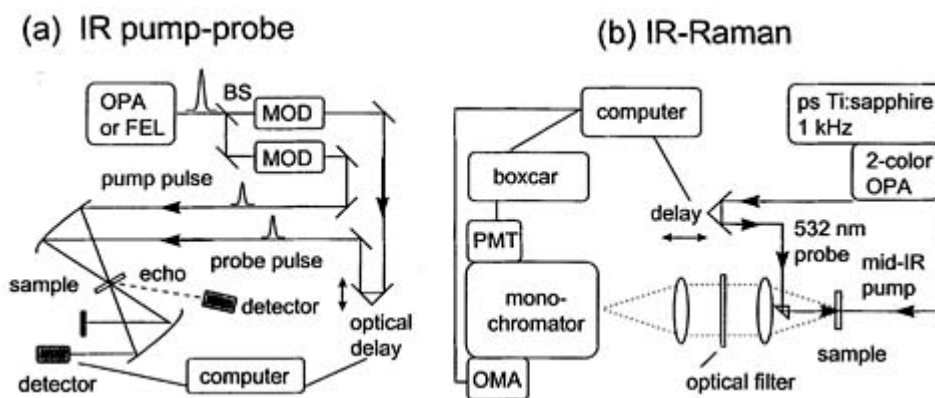


Figure C3.5.3. Schematic diagram of apparatus used for (a) IR pump–probe or vibrational echo spectroscopy by Fayer and co-workers [50] and (b) IR–Raman spectroscopy by Dlott and co-workers [39]. Key: OPA = optical parametric amplifier; FEL = free-electron laser; MOD = high speed optical modulator; PMT = photomultiplier; OMA = optical multichannel analyser.

For IR–Raman experiments, a mid-IR pump pulse from an OPA and a visible Raman probe pulse are used. The Raman probe is generated either by frequency doubling a solid-state laser which pumps the OPA [16], or by a two-colour OPA [39]. Transient anti-Stokes emission is detected with a monochromator and photomultiplier [39], or a spectrograph and optical multichannel analyser [40].

C3.5.5.4 INTERPRETING EXPERIMENTAL MEASUREMENTS

Most molecular vibrations are well described as harmonic oscillators with small anharmonic perturbations [5]. For an harmonic oscillator, all single-quantum transitions have the same frequency, and the intensity of single-quantum transitions increases linearly with quantum number ν . For the usual anharmonic oscillator, the single-quantum transition frequency decreases as ν increases. Ultrashort pulses have a non-negligible frequency bandwidth. For a 1 ps duration pulse, the bandwidth is at minimum $\sim 15 \text{ cm}^{-1}$; for a 100 fs pulse $\sim 150 \text{ cm}^{-1}$. We need to consider two cases, narrow-band pulses (bandwidth less than the anharmonicity) or broad-band pulses (bandwidth greater than the anharmonicity).

Consider narrow-band pulses pumping and probing the fundamental transition of an anharmonic oscillator Ω . The pulse will not be resonant with overtone transitions, so the oscillator is viewed as a two-level system. A two-level system can be saturated, so a pump–probe experiment can measure the VER rate by pumping the system and measuring the absorption recovery rate [45]. There is, however, an interpretation problem. It is difficult to distinguish between decay of Ω with direct repopulation of the ground state, and decay of Ω with population of a different vibration ω by the ladder process $\Omega \rightarrow \omega$ [74]. What the probe sees depends on the frequency shift of the Ω fundamental transition when ω is excited, caused by $Q_{\Omega}Q_{\omega}^2$ and other terms in equation (C3.5.6). If this shift is too small for the probe to resolve, then energy transfer from Ω to ω causes the Ω absorption to recover with decay rate constant K_{Ω} from equation (C3.5.7). If the shift is larger, absorption does not recover with rate constant K_{Ω} , but instead absorption recovery involves both decay constants K_{Ω} and K_{ω} . A two-colour pump–probe experiment greatly alleviates this problem, since a broadly tunable probe pulse could monitor fundamental and overtone transitions of both Ω [73] and ω transitions. Anti-Stokes Raman probing is perhaps even better, since the Raman pulse may simultaneously probe all transitions [40], including transitions of ω and Ω .

With broad-band pulses, pumping and probing processes become more complicated. With a broad-bandwidth pulse it is easy to drive fundamental and overtone transitions simultaneously, generating a complicated population distribution which depends on details of pulse structure [75]. Broad-band probe pulses may be unable to distinguish between fundamental and overtone transitions. For example in IR–Raman experiments with broad-band probe pulses, excitation of the first overtone of a transition appears as a fundamental excitation with twice the intensity, and excitation of a combination band $\Omega + \omega$ appears as excitation of the two fundamentals [76].

C3.5.6 VIBRATIONAL RELAXATION EXAMPLES

C3.5.6.1 DIATOMIC MOLECULES IN RARE GAS CRYSTALS

In rare gas crystals [77] and liquids [78], diatomic molecule vibrational and vibronic relaxation have been studied. In crystals, VER occurs by multiphonon emission. Everything else held constant, the VER rate should decrease exponentially with the number of emitted phonons (exponential gap law) [79, 80]. The number of emitted phonons scales as, and should be close to, the ratio Ω/ω_D , where ω_D is the Debye frequency. A possible complication is the perturbation of the local phonon density of states by the diatomic molecule guest [77].

Apkarian and co-workers used ultrafast spectroscopy to investigate vibrational relaxation of I_2 ($\Omega = 212 \text{ cm}^{-1}$) in solid Kr ($\omega_D = 50 \text{ cm}^{-1}$) and vibronic relaxation of XeF ($\Omega = 424 \text{ cm}^{-1}$) in solid Ar ($\omega_D = 65 \text{ cm}^{-1}$). I_2 , in a Kr lattice at 15 K, is photodissociated by a subpicosecond visible pulse and probed by transient absorption [81]. The impulsive photodissociation hurls the two I atoms against the walls of a Kr cage. The atoms recoil and then recombine to form nascent vibrationally excited I_2 . Figure C3.5.4 [81] shows the rate of ensemble-averaged vibrational energy loss from I_2 . In the first 5 ps, energy loss is fast and stepwise due to individual binary collisions. After that time, energy loss from any individual molecule most likely involves a stepwise relaxation, but ensemble averaging wipes out the details, so in the longer time region the ensemble averaged energy loss is exponential in time with a time constant of 12 ps. This VER process is much faster than VER of liquid O_2 discussed below, probably because the ratio $\Omega/\omega_D \sim 4$ is relatively small.

Vibronic relaxation of XeF in solid Ar at 25 K was studied by pumping vibronic transitions with a subpicosecond UV pulse, and detecting frequency-resolved emission with a fast optical gate [25]. XeF has two sites in Ar, one which emits only from the $B(^2\Sigma_{1/2})$ state and one which emits only from the $C(^2\Pi_{3/2})$ state. Very fast VER was observed in the C-emitting site. Excitation near $v = 20$ results in a return to $v = 0$ in 13 ps, about the same as I_2 in Kr. In the B-emitting states, a slower stepwise relaxation was observed. Figure C3.5.5 shows the possible modes of relaxation for B-emitting XeF and some experimentally determined time constants. Although a diatomic in an atomic lattice seems to be a simple system, these vibronic relaxation experiments are rather complicated to interpret, because of multiple electronic states which are involved due to energy transfer between B and C sites.

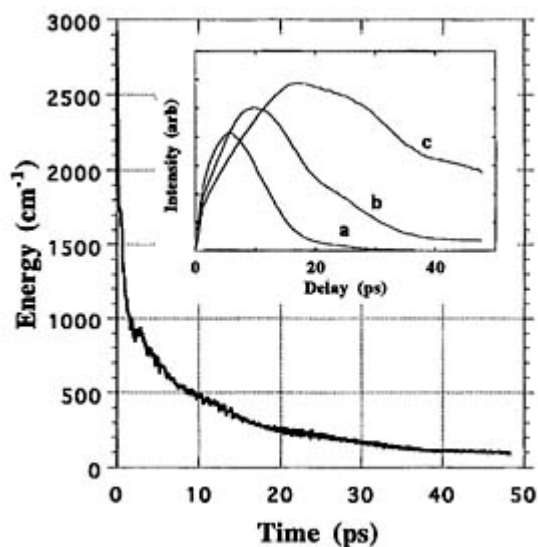


Figure C3.5.4. Ensemble-averaged loss of energy from vibrationally excited I_2 created by photodissociation and subsequent recombination in solid Kr, from [81]. The inset shows calculated transient absorption (pump-probe) signals for inner turning points at 3.5, 3.4 or 3.3 Å.

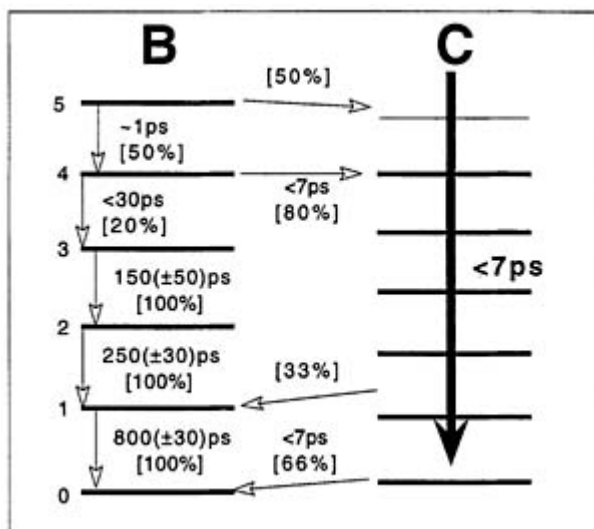


Figure C3.5.5. Vibronic relaxation time constants for B- and C-state emitting sites of XeF in solid Ar for different vibrational quantum numbers ν , from [25]. Vibronic energy relaxation is complicated by electronic crossings caused by energy transfer between sites.

C3.5.6.2 LIQUID OXYGEN

The VER lifetime of liquid O_2 is 280 ms at 70 K [82]. Predicting such a slow rate from theory is a formidable challenge taken up by Skinner and co-workers [54] in a recent paper. The fundamental frequency of O_2 is 1552 cm^{-1} , whereas the maximum characteristic frequency for motion in the liquid (analogous to ω_D) is $\sim 50 \text{ cm}^{-1}$, so an extremely high-order multiphonon process ($\Omega/\omega_D \sim 30$) is needed for VER [54].

VER in liquid O_2 is far too slow to be studied directly by nonequilibrium simulations. The force-correlation function, equation (C3.5.2), was computed from an equilibrium simulation of rigid O_2 . The VER rate constant given in equation (C3.5.3) is proportional to the Fourier transform of the force-correlation function at the O_2 frequency. However, there are two significant practical difficulties. First, the Fourier transform, denoted $\hat{C}(\Omega)$ in

[54] ($\hat{C}(\Omega)$ is not an operator but rather a classical mechanical function), is needed at a frequency $\Omega \gg \omega_D$ where its value is very small. It is difficult to compute $\hat{C}(\Omega)$ accurately at large Ω given the statistical noise in the simulation. Second, the simulation uses classical mechanics, and a quantum correction factor defined in equation (C3.5.4) is needed. The first problem is alleviated using the Wiener–Khintchine theorem [54]. Instead of Fourier transforming the correlation function over the time range of $-\infty$ to $+\infty$, it is more accurate to Fourier transform the fluctuating force itself over a finite time range of $-\tau$ to $+\tau$.

The computed Fourier transform is shown in figure C3.5.6. The O_2 frequency is $\Omega = 2.925 \times 10^{14} \text{ s}^{-1}$. Above $\Omega = 0.8 \times 10^{14} \text{ s}^{-1}$, is very noisy. The result from the Wiener–Khintchine theorem in figure C3.5.6 is believed to be accurate up to $\Omega = 1.5 \times 10^{14} \text{ s}^{-1}$. In order to extend the result to the needed high frequency, an *ansatz* was made that the correlation function $C(t)$ must have the form

$$C(t) = C_0 \frac{\cos(bt)}{\cosh(at)} \quad (\text{C3.5.9})$$

where the coefficients a and b are determined so that equation (C3.5.9) has the same short-time expansion, through order t^4 , as the exact correlation function [54]. By Fourier transforming equation (C3.5.9) analytically, a result for the VER rate can be determined,

$$k_{1 \rightarrow 0} = \frac{\pi Q C_0 \cosh(\pi \Omega / 2a) \cosh(\pi b / 2a)}{\mu \hbar \Omega a [\cosh(\pi \Omega / a) + \cosh(\pi b / a)]} \quad (\text{C3.5.10})$$

Figure C3.5.6 compares the result of this *ansatz* to the numerical result from the Wiener–Khintchine theorem. They agree well and the *ansatz* exhibits the expected exponential energy-gap law (VER rate decreases exponentially with Ω). The *ansatz* was used to determine the VER rate with no quantum correction ($Q = 1$), with the Bader–Berne harmonic correction [61] and with a correction based [83, 84] on Egelstaff’s method [62]. The Egelstaff corrected results were within a factor of five of experiment, whereas other corrections were off by orders of magnitude. This calculation represents the present state of the art in computing VER rates in such difficult systems, inasmuch as the authors used only a model potential and no adjustable parameters. However the *ansatz* procedure is clearly not extendible to polyatomic molecules or to diatomic molecules in polyatomic solvents.

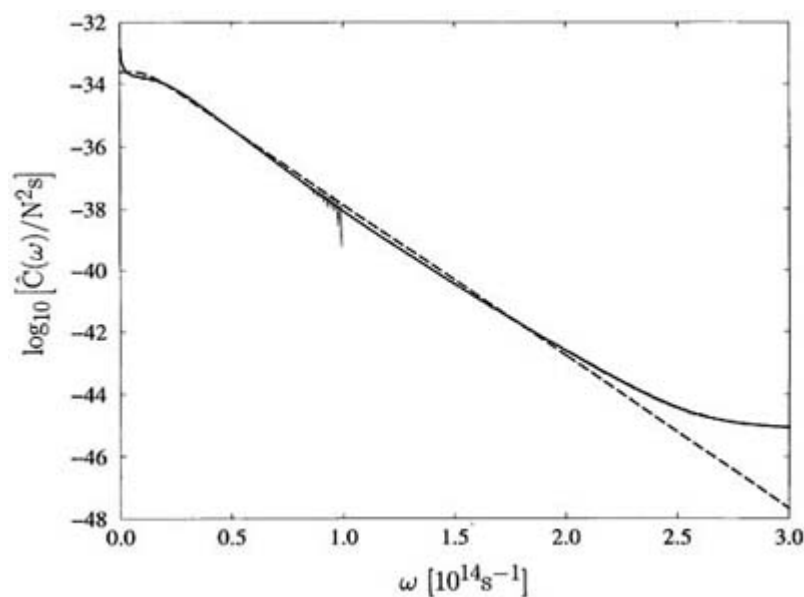


Figure C3.5.6. The computed Fourier transform at frequency ω , $\hat{C}(\omega)$, of the classical mechanical force–force correlation function for liquid O₂ at 70 K from [54]. The VER rate is proportional to the value of $\hat{C}(\omega)$ at the O₂ vibrational frequency $2.925 \times 10^{14} \text{ s}^{-1}$, multiplied by the quantum correction Q . The solid line is obtained from direct Fourier transformation of the simulation. The thick solid line is obtained from the Wiener–Kjhintchine theorem, and the dashed line is an *ansatz* proposed by the authors.

C3.5.6.3 SMALL-MOLECULE REACTION DYNAMICS IN SOLUTION

Chemical reaction dynamics is an attempt to understand chemical reactions at the level of individual quantum states. Much work has been done on isolated molecules in molecular beams, but it is unlikely that this information can be used to understand condensed phase chemistry at the same level [8]. In a bath, the reacting solute’s potential energy surface is altered by both dynamic and static effects. The static effect is characterized by a potential of mean force. The dynamical effects are characterized by the force-correlation function or the frequency-dependent friction [8].

Photodissociation of a linear triatomic such as I₃⁻ [85, 86] or HgI₂ [8] to produce a vibrationally excited diatomic, or cage recombination of a photodissociated diatomic such as I₂ [78, 81] are classic model simple systems for reaction dynamics. Here we discuss the HgI₂ → HgI + I reaction studied by Hochstrasser and co-workers [87, 88 and 89]. The important issues are how energy is partitioned, the degree of coherence in the formation of the product and whether the reaction is adiabatic (solvent easily follows reactant motions) or nonadiabatic [8]. A nonadiabatic theory would be much more complicated.

Pumping HgI₂ in ethanol with a femtosecond UV pulse causes impulsive photodissociation, producing HgI with average vibrational quantum number $\nu = 15$ [88]. There are several possibilities for nascent HgI, as depicted in [figure C3.5.7](#) [8]. The smooth Gaussian function represents the ensemble-average of HgI vibrational displacements (vibrational wavepacket). In possibility A, there is no VER and no vibrational dephasing. All HgI fragments simply oscillate coherently. In B, VER is faster than a vibrational period, so HgI loses energy before vibrating even once.

An

-13-

ensemble of coherently vibrating ground state HgI results. In C, the VER rate is comparable to a vibrational period, so there will be some coherence in the ground state. In D, dephasing (caused by quartic anharmonic coupling to a dynamic bath) is faster than VER, so the wavepacket spreads (dephases) much faster than it loses energy.

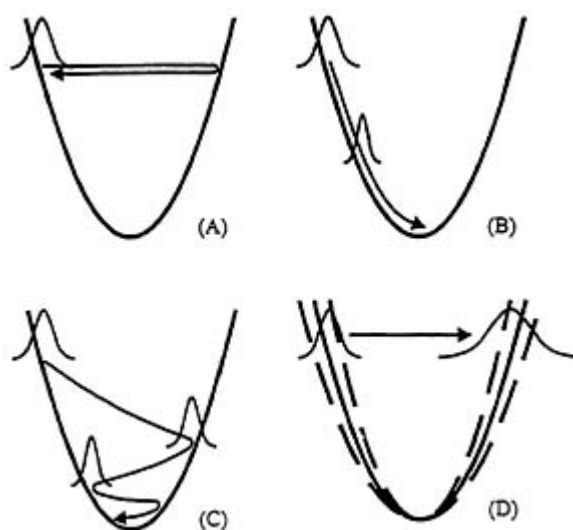


Figure C3.5.7. Possible modes of vibrational wavepacket (smooth Gaussian curve) motion for a highly vibrationally excited diatomic molecule produced by photodissociation of a linear triatomic such as HgI₂, from [8].

(A) no VER and no dephasing. (B) VER is faster than a vibrational period. Once the vibrational ground state is reached, the wavepacket begins to oscillate coherently. (C) The VER rate is comparable to a vibrational period so some coherence is seen in both excited and ground states. (D) Dephasing is faster than VER.

In HgI, possibility C is the best description [8]. The dephasing time constant is ~ 150 fs and the overall time for vibrational cooling is ~ 200 fs. Thus coherence is seen in the vibrational excited states, and in the ground state as well. A molecular dynamics simulation of rigid HgI in ethanol was used to understand the VER mechanism [90]. The computed frequency-dependent friction is shown in [figure C3.5.8](#) [90]. Notice this function is much more complicated than in liquid O_2 ([figure C3.5.6](#)), and an exponential gap law is not observed. The simulation results were used to conclude VER at the 130 cm^{-1} frequency of HgI was dominated by Lennard-Jones interactions between solute and solvent. The solvent nuclear response is a few times faster than the vibrational period, so the HgI–ethanol system is close to the adiabatic limit.

-14-

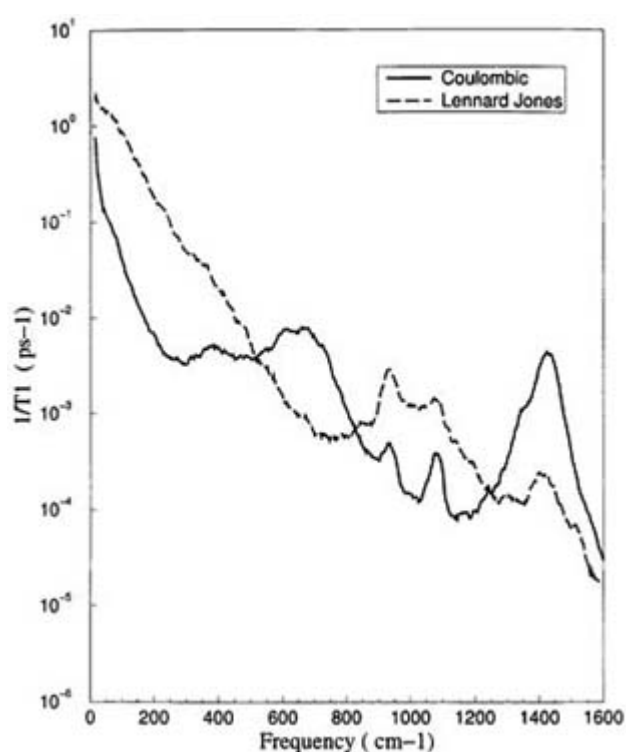


Figure C3.5.8. Computed frequency-dependent friction (inversely proportional to the VER lifetime T_1) from a classical molecular dynamics simulation of rigid HgI molecules in ethanol solution, from [90]. The HgI vibrational frequency is 125 cm^{-1} . The Lennard-Jones contribution to the friction dominates the Coulombic contribution at that frequency.

C3.5.6.4 TEMPERATURE DEPENDENCE OF VIBRATIONAL LINESHAPES

VER rates cannot be predicted from vibrational lineshapes alone [5] except for a few exceptional cases [2, 91]. The most detailed deconstruction of vibrational lineshapes to date is the work of Fayer and co-workers [92], who studied $W(CO)_6$ in glass-forming liquids including 2-methyl pentane (2MP; $T_g = 80$ K). The triply degenerate asymmetric $C\equiv O$ stretch of $W(CO)_6$ ($\Omega \sim 1980\text{ cm}^{-1}$) has a VER lifetime which is solvent and temperature dependent. In 2-MP at 300 K, the lifetime is ~ 150 ps.

Infrared absorption was used to measure the total lineshape. Vibrational echo spectroscopy (the vibrational analogue of spin echoes in magnetic resonance [93]) was used to remove inhomogeneous broadening, to reveal the underlying homogeneous lineshape. One-color pump–probe with magic-angle polarization was used to measure the VER lifetime. Pump–probe polarization anisotropy was used to measure the orientational relaxation rate.

The vibrational echo experiments yielded exponential decays at all temperatures. The Fourier-transform of the echo decay gives the homogeneous lineshape, in this case Lorentzian. The echo decay time constant is $4T_2$, where T_2 is the vibrational dephasing time constant, and the corresponding homogeneous linewidth $\Gamma_{\text{hom}} = 1\pi T_2$. At low temperature (~ 10 K) in 2-MP glass, the total linewidth is $\Gamma_{\text{tot}} = 300$ GHz (30 GHz = 1 cm $^{-1}$), whereas $\Gamma_{\text{hom}} \sim 1.5$ GHz.

-15-

Therefore, the absorption line is massively inhomogeneously broadened at low temperature. An inhomogeneous lineshape can be used to determine the static or quasistatic frequency spread of oscillators due to a distribution of environments, but it provides no dynamical information whatsoever [94, 95]. As T is increased to 300 K, the absorption linewidth Γ_{tot} decreases and Γ_{hom} increases. At 300 K, the lineshape is nearly homogeneously broadened and dominated by vibrational dephasing, because fast dephasing wipes out effects of inhomogeneous environments, a well known phenomenon termed ‘motional narrowing’ [95].

The homogeneous linewidth Γ_{hom} can be subdivided into distinctly different contributions,

$$\Gamma_{\text{hom}} = \frac{1}{\pi T_2} = \frac{1}{\pi T_2^*} + \frac{1}{2\pi T_1} + \Gamma_{\text{or}}$$

where T_2^* is the time constant for ‘pure dephasing’ processes which modulate only the oscillator phase, and Γ_{or} is the contribution from orientational relaxation [92]. Orientational relaxation refers either to molecular rotation (impossible below T_g) or interconversion among nearly degenerate $C\equiv$ stretching transitions of different symmetry [96]. The parameters T_2 , T_1 and Γ_{or} are measured from echo and pump–probe experiments, and T_2^* is computed by substituting those parameters into this equation.

Figure C3.5.9 shows all contributions to the homogeneous linewidth of $W(\text{CO})_6$ in 2-MP. Γ_{hom} increases from about 1.5 to 100 GHz from 10 to 300 K. Pure dephasing and orientational contributions to Γ_{hom} vanish as $T \rightarrow 0$, leaving only the VER contribution. As T is increased, the VER lifetime decreases slightly with increasing temperature. This counterintuitive temperature dependence is not seen explicitly in equation (C3.5.7). However, a thermal slowdown of VER can happen if the C_m or ρ_m factors decrease with increasing T and the occupation numbers $n_m(T)$ involve only high frequencies $\hbar\omega \gg kT$ [34]. The orientational contribution increases with T , but it never contributes much to the total lineshape. At ~ 50 K the homogeneous line undergoes a transition from VER dominance to pure dephasing dominance. Thus the VER contribution to the lineshape is hidden at lower temperatures by inhomogeneous broadening and, at higher temperatures, by pure dephasing.

-16-

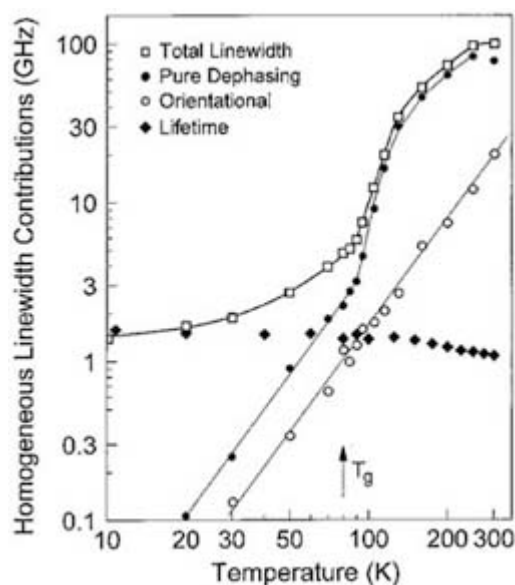


Figure C3.5.9. Contributions to the homogeneous linewidth of a C≡O stretching transition of $W(CO)_6$ ($\Omega \sim 2000 \text{ cm}^{-1}$) in a glass-forming liquid, 2-methyl pentane, from [92]. The total homogeneous linewidth Γ_{hom} is measured with vibrational echo spectroscopy, which removes inhomogeneous broadening. The VER lifetime τ_1^{hom} and orientational relaxation contribution Γ_{or} are measured with pump–probe experiments. The pure dephasing time constant τ_2^{h} is computed from the other results. VER dominates the homogeneous lineshape at low temperature. Pure dephasing dominates at high temperature.

C3.5.6.5 POLYATOMIC MOLECULES IN LOW-TEMPERATURE CRYSTALS—FREQUENCY DEPENDENCE

Much of our knowledge of the frequency dependence of VER rates in polyatomic molecules stems from low-temperature studies of molecular crystals [2] such as pentacene (PTC; $C_{22}H_{14}$) guest molecules in a crystalline naphthalene (N; $C_{10}H_8$) host. In naphthalene, the phonon cut-off frequency is $\sim 180 \text{ cm}^{-1}$ [97]. At low temperature, PTC has well resolved vibronic transitions ($S_0^0 \rightarrow S_1^{\Omega}$) in a convenient wavelength range for picosecond dye lasers (560–605 nm).

Vibronic relaxation of PTC/N was studied by Hesselink and Wiersma [98, 99], who used ultrafast vibronic echoes to measure dephasing rates of 16 PTC vibronic transitions at low temperature (1.5 K), where vibronic dephasing is dominated by VER. Their results are shown in figure C3.5.10. The lower-frequency vibrations ($< 350 \text{ cm}^{-1}$) have shorter lifetimes in the 2 ps range. The mid-frequency vibrations (350 to 1000 cm^{-1}) have generally longer lifetimes up to 40 ps. The higher-frequency vibrations ($> 1000 \text{ cm}^{-1}$) have shorter lifetimes. These results generally support the three-regime model proposed by Nitzan and Jortner [35]. Below regime I ($\leq 360 \text{ cm}^{-1}$), vibrations relax efficiently by two-phonon emission. Above 360 cm^{-1} vibrational lifetimes become longer (regime II) until IVR becomes dominant in regime III, because the vibrational density of states becomes large enough ($> 10^2 \text{ states cm}^{-1}$) for efficient IVR.

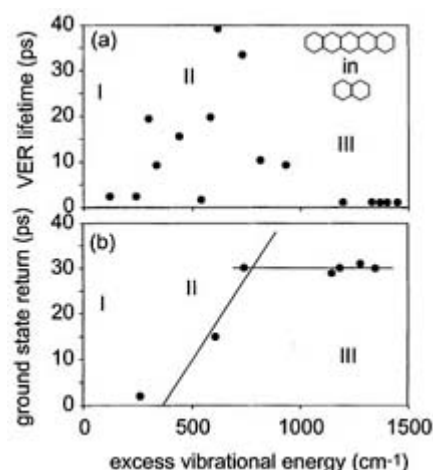


Figure C3.5.10. Frequency-dependent vibronic relaxation data for pentacene (PTC) in naphthalene (N) crystals at 1.5 K. (a) Vibrational echoes are used to measure VER lifetimes (from [99]). The lifetimes are shorter in regime I, longer in regime II, and become shorter again in regime III. (b) Two-colour pump–probe experiments are used to measure vibrational cooling (return to the ground state) from [102].

VER measurements of individual levels do not necessarily indicate the overall rate of vibrational energy loss from PTC molecules. Vibrational energy loss is a multistep process termed ‘vibrational cooling’ (VC) if ladder processes (vibrational cascades) are important [100]. Chang and Dlott measured VC rates in PTC/N with two-color pump probe experiments [101, 102]. The first pulse pumps an $S_0^0 \rightarrow S_1^0$ vibronic transition and the second pulse measures the return to S_1^0 by probing the $S_1^0 \rightarrow S_0^0$ transition. The return kinetics are nonexponential in time. By fitting the data to a VC model, time constants shown in figure C3.5.10 were extracted. The decay of regime I vibrations directly repopulates the ground state, so VER and VC rates are identical. The decay of regime II vibrations occur by a ladder process, so the VC rate increases higher up the ladder. The decay of regime III vibrations occurs by fast IVR. Some of that redistributed energy populates longer-lived regime II states, so in regime III the VC rate levels out at about the maximum VC rate in regime II.

C3.5.6.6 VER OF A POLYATOMIC LIQUID

The decay of C–H stretching (and OH and NH) vibrations in liquids has been studied by IR–Raman spectroscopy [6]. Early work on ethanol by Alfano and Shapiro [23] indicated that C–H stretch excitations ($\sim 3000\text{ cm}^{-1}$) decayed by populating C–H bending excitations ($\sim 1500\text{ cm}^{-1}$). However, until recently it was not known where the rest of the energy went, or the subsequent fate of the daughter C–H bending excitations. Dlott and co-workers used IR–Raman techniques to monitor the flow of vibrational energy through several polyatomic molecule systems [39, 41, 42]. Data for one example, acetonitrile ($\text{CH}_3\text{C}\equiv\text{N}$), are shown in figure C3.5.11 [6]. Acetonitrile is a model for nonassociated polar polyatomic liquids [103, 104 and 105].

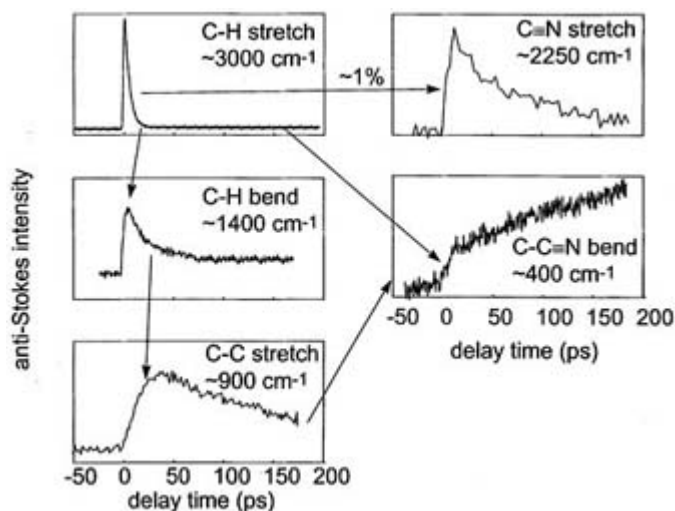


Figure C3.5.11. IR–Raman measurements of vibrational energy flow through acetonitrile in a neat liquid at 300 K, adapted from [41]. An ultrashort mid-IR pulse pumps the C–H stretch, which decays in 3 ps. Only 1% of the energy is transferred to the C≡N stretch, which has an 80 ps lifetime. Most of the energy is transferred to the C–H bend plus about four quanta of C–C≡N bend. The daughter C–H bend vibration relaxes by exciting the C–C stretch. The build-up of energy in the C–C≡N bend mirrors the build-up of energy in the bath, which continues for about 250 ps after C–H stretch pumping.

An ultrashort mid-IR pulse excited a C–H stretching vibration ($\sim 3000\text{ cm}^{-1}$) of neat acetonitrile at 300 K. The loss of C–H stretching energy occurred in 3 ps. Only 1% of that energy was transferred to the C≡N stretch (2250 cm^{-1}), where it remained for ~ 80 ps. Most of the energy was lost from the C–H stretch by the process,



which accounts for the ~ 3 ps rise of excitation of the C–C≡N bending vibration ($\sim 400\text{ cm}^{-1}$). The C–H bending excitation decays in 15 ps, exciting lower-energy C–C stretch, C–C bend and C≡N bending modes. The C–C stretch and C–C bend decay with quite long lifetimes, in the 50 ps range. All vibrational energy is dissipated to the bath in ~ 250 ps. There is a net temperature increase of $\sim 10^\circ\text{C}$ in the bath, which accounts for the long-time-scale build-up of the C–C≡N bending population. Whereas C–H stretch lifetime measurements indicate a VER lifetime of 3 ps, vibrational cooling actually takes ~ 250 ps. Frequency-dependent VER lifetimes in acetonitrile generally support the model of Nitzan and Jortner [35], with an exception being the long-lived high-frequency C≡N stretch [41].

C3.5.7 CONCLUDING REMARKS

Understanding VER in condensed phases has proven difficult. The experiments are hard. The structurally simple systems (diatomic molecules) involve complicated relaxation mechanisms. The structures of polyatomic molecules are obviously more complex, but polyatomic systems are tractable because the VER mechanisms are somewhat simpler.

There are encouraging signs that condensed-phase VER is an important fundamental problem which is due for a major breakthrough. Theoreticians are finally having success in predicting VER rates of simple systems, and have begun to understand the multiple roles of VER in chemical reaction dynamics. Experimental technology has improved a great deal. As a result, we are beginning to accumulate data on several different molecular liquids and solids, vibrational energy flowing through a polyatomic molecule can be monitored in real time—even protein VER can be measured—and the first direct observations of chemical reaction dynamics in solution have been

reported. Nevertheless, much remains to be done in the areas of predicting and understanding VER rates, developing simple but robust conceptual frameworks and incorporating our emerging understanding of VER into theories of chemical reaction dynamics, electron transfer, protein dynamics and other condensed-phase dynamical processes.

REFERENCES

- [1] Chesnoy J and Gale G M 1984 Vibrational energy relaxation in liquids *Ann. Phys., Paris* **9** 893–949
 - [2] Dlott D D 1988 Dynamics of molecular crystal vibrations *Laser Spectroscopy of Solids II* ed W Yen (Berlin: Springer) pp 167–200
 - [3] Owrutsky J C, Raftery D and Hochstrasser R M 1994 Vibrational relaxation dynamics in solutions *Annu. Rev. Phys. Chem.* **45** 519–55
 - [4] Oxtoby D W 1981 Vibrational relaxation in liquids *Annu. Rev. Phys. Chem.* **32** 77–101
 - [5] Oxtoby D W 1981 Vibrational population relaxation in liquids *Photoselective Chemistry Part 2 (Advances in Chemical Physics 47)* ed J Jortner, R D Levine and S A Rice (New York: Wiley) pp 487–519
 - [6] Seilmeier A and Kaiser W 1988 Ultrashort intramolecular and intermolecular vibrational energy transfer of polyatomic molecules in liquids *Ultrashort Laser Pulses and Applications (Topics in Applied Physics 60)* ed W Kaiser (Berlin: Springer) pp 279–315
 - [7] Flynn G W, Parmenter C S and Wodtke A M 1996 Vibrational energy transfer *J. Phys. Chem.* **100** 12817–38
 - [8] Voth G A and Hochstrasser R M 1996 Transition state dynamics and relaxation processes in solutions: a frontier of physical chemistry *J. Phys. Chem.* **100** 13034–49
 - [9] Kramers H A 1940 Brownian motion in a field of force and the diffusion model of chemical reactions *Physica* **7** 284–304
 - [10] Grote R F and Hynes J T 1981 Reactive modes in condensed phase reactions *J. Chem. Phys.* **74** 4465–75
 - [11] Lee M, Holtom G R and Hochstrasser R M 1985 Observation of the Kramers turnover region in the isomerism of *trans*-stilbene in fluid ethane *Chem. Phys. Lett.* **118** 359–63
 - [12] Fleming G and Hänggi P 1993 *Activated Barrier Crossing* (River Edge, NJ: World Scientific)
 - [13] Hasha D L, Eguchi T and Jonas J 1981 Dynamical effects on conformational isomerization of cyclohexane *J. Chem. Phys.* **75** 1571–3
 - [14] Hasha D L, Eguchi T and Jonas J 1982 High-pressure NMR study of dynamical effects on conformational isomerization of cyclohexane *J. Am. Chem. Soc.* **104** 2290–6
 - [15] Herzfeld K F 1952 The origin of the ultrasonic absorption in liquids II *J. Chem. Phys.* **20** 288–9
 - [16] Laubereau A and Kaiser W 1978 Vibrational dynamics of liquids and solids investigated by picosecond light pulses *Rev. Mod. Phys.* **50** 607–65
 - [17] Legay-Sommaire N and Legay F 1980 Observation of a strong vibrational population inversion by CO laser excitation of pure solid carbon monoxide *IEEE J. Quantum Electron.* **16** 308–14
 - [18] Brueck S R J and Osgood R M Jr 1976 Vibrational energy relaxation in liquid N₂-CO mixtures *Chem. Phys. Lett.* **39** 568–72
-
- [19] Miyasaka H, Hagihara M, Okada T and Mataga N 1992 Femtosecond laser photolysis studies on the cooling process of chrysene in the vibrationally hot S₁ state in solution *Chem. Phys. Lett.* **188** 259–64
 - [20] Foggi P, Pettini L, Sànta I, Righini R and Califano S 1995 Transient absorption and vibrational relaxation dynamics of the lowest excited singlet state of pyrene in solution *J. Phys. Chem.* **99** 7439–45
 - [21] Jiang Y J and B G 1994 Vibrational population relaxation of perylene in its ground and excited electronic states *J. Phys. Chem.* **98** 9417–21
 - [22] Laubereau A, von der Linde D and Kaiser W 1972 Direct measurement of the vibrational lifetimes of molecules in liquids *Phys. Rev. Lett.* **28** 1162–5

- [23] Alfano R R and Shapiro S L 1972 Establishment of a molecular-vibration decay route in a liquid *Phys. Rev. Lett.* **29** 1655–8
- [24] Laubereau A, Greiter L and Kaiser W 1974 Intense tunable picosecond pulses in the infrared *Appl. Phys. Lett.* **25** 87–9
- [25] Hoffman G J, Imre D G, Zadayan R, Schwentner N and Apkarian V A 1993 Relaxation dynamics in the B(1/2) and C(3/2) charge transfer states of XeF in solid Ar *J. Chem. Phys.* **98** 9233–40
- [26] Anex D S and Ewing G E 1986 Transfer and storage of vibrational energy in liquids: collisional up-pumping of carbon monoxide in liquid argon *J. Phys. Chem.* **90** 1604–10
- [27] Chang H-C and Ewing G E 1990 Infrared fluorescence from a monolayer of CO on NaCl(100) *Phys. Rev. Lett.* **65** 2125–8
- [28] Eienthal K B 1992 Equilibrium and dynamical processes at interfaces by second harmonic and sum frequency generation *Annu. Rev. Phys. Chem.* **43** 627–61
- [29] Heilweil E J, Casassa M P, Cavanagh R R and Stephenson J C 1989 Picosecond vibrational energy transfer studies of surface adsorbates *Annu. Rev. Phys. Chem.* **40** 143–71
- [30] Heilweil E J, Casassa M P, Cavanagh R R and Stephenson J C 1984 Picosecond vibrational energy relaxation of surface hydroxyl groups on colloidal silica *J. Chem. Phys.* **81** 2856–8
- [31] Guyot-Sionnest P, Lin P-H and Hiller E M 1995 Vibrational dynamics of the Si–H stretching modes for the Si(100)/H:2 × 1 surface *J. Chem. Phys.* **102** 4269
- [32] Gomez M and Tully J C 1993 Electronic and phonon mechanisms of vibrational relaxation: CO on Cu(100) *J. Vac. Sci. Technol. A* **11** 1914–20
- [33] Beckerle J D, Casassa M P, Cavanagh R R, Heilweil E J and Stephenson J C 1990 Ultrafast infrared response of adsorbates on metal surfaces: vibrational lifetime of CO/Pt(111) *Phys. Rev. Lett.* **64** 2090–3
- [34] Kenkre V M, Tokmakoff A and Fayer M D 1994 Theory of vibrational relaxation of polyatomic molecules in liquids *J. Chem. Phys.* **101** 10618–29
- [35] Nitzan A and Jortner J 1973 Vibrational relaxation of a molecule in a dense medium *Mol. Phys.* **25** 713–34
- [36] Apkarian V A, Wiedman L, Janiesch W and Weitz E 1986 Vibrational energy transfer and migration processes in matrix isolated CH₃F *J. Chem. Phys.* **85** 5593–610
- [37] Ambroseo J R and Hochstrasser R M 1988 Pathways of relaxation of the N–H stretching vibration of pyrrole in liquids *J. Chem. Phys.* **89** 5956–7
- [38] Hong X, Chen S and Dlott D D 1995 Ultrafast mode-specific intermolecular vibrational energy transfer to liquid nitromethane *J. Phys. Chem.* **99** 9102–9
- [39] Deák J C, Iwaki L K and Dlott D D 1997 High power picosecond mid-infrared optical parametric amplifier for infrared–Raman spectroscopy *Opt. Lett.* **22** 1796–8
- [40] Graener H, Zürl R and Hofmann M 1997 Vibrational relaxation of liquid chloroform *J. Phys. Chem.* **101** 1745–9

- [41] Deák J C, Iwaki L K and Dlott D D 1998 Vibrational energy relaxation of polyatomic molecules in liquids: acetonitrile *J. Phys. Chem.* **102** 8193–201
- [42] Deák J C, Iwaki L K and Dlott D D 1998 When vibrations interact: ultrafast energy relaxation of vibrational pairs in polyatomic liquids *Chem. Phys. Lett.* **293** 405–11
- [43] Chen S, Hong X, Hill J R and Dlott D D 1995 Ultrafast energy transfer in high explosives: vibrational cooling *J. Phys. Chem.* **99** 4525–30
- [44] Tokmakoff A and Fayer M D 1995 Infrared photon echo experiments: exploring vibrational dynamics in liquids and glasses *Accounts Chem. Res.* **28** 437–45
- [45] Heilweil E J, Cavanagh R R and Stephenson J C 1988 CO ($\nu = 1$) population lifetimes of metal-carbonyl cluster compounds in dilute CHCl₃ solution *J. Chem. Phys.* **89** 230–9
- [46] Owrutsky J C, Li M, Locke B and Hochstrasser R M 1995 Vibrational relaxation of the CO stretch vibration in hemoglobin–CO, myoglobin–CO, and protoheme–CO *J. Phys. Chem.* **99** 4842–6
- [47] Hill J R *et al* 1994 Vibrational dynamics of carbon monoxide at the active site of myoglobin: picosecond infrared free-electron laser pump-probe experiments *J. Phys. Chem.* **98** 11213–19

- [48] Hill J R *et al* 1995 Vibrational relaxation of carbon monoxide in model heme compounds: 6-coordinate metalloporphyrins (M = Fe, Ru, Os) *Chem. Phys. Lett.* **224** 218–23
- [49] Hill J R, Ziegler C J, Suslick K S, Dlott D D, Rella C W and Fayer M D 1996 Tuning the vibrational relaxation of CO bound to heme and metalloporphyrin complexes *J. Phys. Chem.* **100** 18023–32
- [50] Hill J R *et al* 1996 Ultrafast infrared spectroscopy in biomolecules: active site dynamics of heme proteins *Biospectroscopy* **2** 277–99
- [51] Hamm P, Lim M and Hochstrasser R M 1998 Ultrafast dynamics of amide-I vibrations *Biophys. J.* **74** A332–
- [52] Hamm P, Limm M and Hochstrasser R M 1998 Structure of the amide I band of peptides measured by femtosecond nonlinear-infrared spectroscopy *J. Phys. Chem. B* **102** 6123–38
- [53] Peterson K A, Engholm J R, Rella C W and Schwettman H A 1997 Picosecond infrared studies of protein vibrational modes *Accelerator-Based Infrared Sources and Applications* eds G P Williams and P Dumas (Bellingham, WA: SPIE) pp 149–58; *Proc. SPIE* vol 3153
- [54] Everitt K F, Egorov S A and Skinner J L 1998 Vibrational energy relaxation in liquid oxygen *Chem. Phys.* **235** 115–22
- [55] Velsko S and Oxtoby D W 1980 Vibrational energy relaxation in liquids *J. Chem. Phys.* **72** 2260–3
- [56] Moore P, Tokmakoff A, Keyes T and Fayer M D 1995 The low frequency density of states and vibrational population dynamics of polyatomic molecules in liquids *J. Chem. Phys.* **103** 3325–34
- [57] Seeley G and Keyes T 1989 Normal-mode analysis of liquid-state dynamics *J. Chem. Phys.* **91** 5581–6
- [58] Xu B-C and Stratt R M 1990 Liquid theory for band structure in a liquid. II. p orbitals and phonons *J. Chem. Phys.* **92** 1923–35
- [59] Goodyear G and Stratt R M 1996 The short-time intramolecular dynamics of solutes in liquids. I. An instantaneous-normal-mode theory for friction *J. Chem. Phys.* **105** 10050–71
- [60] Goodyear G, Larsen R E and Stratt R M 1996 Molecular origin of friction in liquids *Phys. Rev. Lett.* **76** 243–6
- [61] Bader J S and Berne B J 1994 Quantum and classical rates for classical simulations *J. Chem. Phys.* **100** 8359–66
- [62] Egelstaff P A 1962 Neutron scattering studies of liquid diffusion *Adv. Phys.* **11** 203–32
- [63] Califano S, Schettino V and Neto N 1981 *Lattice Dynamics of Molecular Crystals* (Berlin: Springer)
- [64] Lingle R J, Xu X, Zhu H, Yu S-C and Hopkins J B 1991 Picosecond Raman study of energy flow in a photoexcited heme protein *J. Phys. Chem.* **95** 9320–31

- [65] Lingle R J, Xu X B, Zhu H P, Yu S-C and Hopkins J B 1991 Direct observation of hot vibrations in photoexcited deoxyhemoglobin using picosecond Raman spectroscopy *J. Am. Chem. Soc.* **113** 3992–4
- [66] Chen S, Lee I-Y S, Tolbert W, Wen X and Dlott D D 1992 Applications of ultrafast temperature jump spectroscopy to condensed phase molecular dynamics *J. Phys. Chem.* **96** 7178–86
- [67] Chen S, Tolbert W A and Dlott D D 1994 Direct measurement of ultrafast multiphonon up pumping in high explosives *J. Phys. Chem.* **98** 7759–66
- [68] Wen X, Tolbert W A and Dlott D D 1992 Multiphonon up pumping and molecular hot spots in superheated polymers studied by ultrafast optical calorimetry *Chem. Phys. Lett.* **192** 315–20
- [69] Wen X, Tolbert W A and Dlott D D 1993 Ultrafast temperature jump in polymers: phonons and vibrations heat up at different rates *J. Chem. Phys.* **99** 4140–51
- [70] Giordmaine J A and Kaiser W 1966 Light scattering by coherently drive lattice vibrations *Phys. Rev.* **144** 676–88
- [71] Dmitriev V G, Gurzadyan G G and Nikogosyan D N 1997 *Handbook of Nonlinear Optical Crystals* 2nd edn, ed H K V Lotsch (Berlin: Springer)
- [72] Brau C A 1988 Free-electron lasers *Science* **239** 1115–21
- [73] Owrutsky J C, Li M, Culver J P, Sarisky M J, Yodh A G and Hochstrasser R M 1993 Vibrational dynamics of condensed phase molecules studied by ultrafast infrared spectroscopy *Time Resolved Vibrational Spectroscopy VI (Springer Proc. in Physics 74)* ed A Lau (New York: Springer) pp 63–7
- [74] Heilweil E J, Casassa M P, Cavanagh R R and Stephenson J C 1985 Vibrational deactivation of surface OH chemisorbed on SiO₂: solvent effects *J. Chem. Phys.* **82** 5216–31
- [75] Tokmakoff A, Kwok A S, Urdahl R S, Francis R S and Fayer M D 1995 Multilevel vibrational dephasing and vibrational

anharmonicity from infrared photon echo beats *Chem. Phys. Lett.* **234** 289–95

- [76] Hofmann M and Graener H 1995 Time resolved incoherent anti-Stokes Raman spectroscopy of dichloromethane *Chem. Phys.* **206** 129–37
- [77] Dubost H 1984 Spectroscopy of vibrational and rotational levels of diatomic molecules in rare-gas crystals *Inert Gases. Potentials, Dynamics, and Energy Transfer in Doped Crystals (Springer Ser. Chem. Phys. 34)* ed M L Klein (Berlin: Springer) pp 145–256
- [78] Harris C B, Smith D E and Russell D J 1990 Vibrational relaxation of diatomic molecules in liquids *Chem. Rev.* **90** 481–8
- [79] Egorov S A and Skinner J L 1997 Vibrational energy relaxation of diatomic molecules in rare gas crystals *J. Chem. Phys.* **106** 1034–40
- [80] Nitzan A, Mukamel S and Jortner J 1975 Energy gap law for vibrational relaxation of a molecule in a dense medium *J. Chem. Phys.* **63** 200–7
- [81] Zadoyan R, Li Z, Martens C C and Apkarian V A 1994 The breaking and remaking of a bond: caging of I₂ in solid Kr *J. Chem. Phys.* **101** 6648–57
- [82] Faltermeier B, Protz R and Maier M 1981 Concentration and temperature dependence of electronic and vibrational energy relaxation of O₂ in liquid mixtures *Chem. Phys.* **62** 377–85
- [83] Berne B J, Jortner J and Gordon R 1967 Vibrational relaxation of diatomic molecules in gases and liquids *J. Chem. Phys.* **47** 1600–8
- [84] Borysow J, Moraldi M and Frommhold L 1985 The collision induced spectroscopies. Concerning the desymmetrization of classical line shape *Mol. Phys.* **56** 913–22
- [85] Banin U, Waldman A and Ruhman S J 1992 Ultrafast photodissociation of I₃⁻ in solution: direct observation of coherent product vibrations *J. Chem. Phys.* **96** 2416–19

-23-

- [86] Banin U and Ruhman S 1993 Ultrafast photodissociation of I₃⁻. Coherent photochemistry in solution *J. Chem. Phys.* **98** 4391–403
- [87] Pugliano N, Palit D K, Szarka A Z and Hochstrasser R M 1993 Wave packet dynamics of the HgI₂ photodissociation reaction in solution *J. Chem. Phys.* **99** 7273–6
- [88] Pugliano N, Szarka A Z, Gnanakaran S, Triechel M and Hochstrasser R M 1995 Vibrational population dynamics of the HgI photofragment in ethanol solution *J. Chem. Phys.* **103** 6498–511
- [89] Pugliano N, Szarka A Z and Hochstrasser R M 1996 Relaxation of the product state coherence generated through the photolysis of HgI₂ in solution *J. Chem. Phys.* **104** 5062–79
- [90] Gnanakaran S and Hochstrasser R M 1996 Vibrational relaxation of HgI in ethanol: equilibrium molecular dynamics simulations *J. Chem. Phys.* **105** 3486–96
- [91] Decola P L, Hochstrasser R M and Trommsdorff H P 1980 Vibrational relaxation in molecular crystals by four-wave mixing: naphthalene *Chem. Phys. Lett.* **72** 1–4
- [92] Tokmakoff A and Fayer M D 1995 Homogeneous vibrational dynamics and inhomogeneous broadening in glass-forming liquids—infrared photon echo experiments from room temperature to 10 K *J. Chem. Phys.* **103** 2810–26
- [93] Tokmakoff A, Zimdars D, Sauter B, Francis R S, Kwok R S and Fayer M D 1994 Vibrational photon echoes in a liquid and glass—room temperature to 10 K *J. Chem. Phys.* **101** 1741–4
- [94] Kubo R and Tomita K 1954 A general theory of magnetic resonance absorption *J. Phys. Soc. Japan* **9** 888–919
- [95] Anderson P W 1953 A mathematical model for the narrowing of spectral lines by exchange or motion *J. Phys. Soc. Japan* **9** 316–39
- [96] Tokmakoff A, Sauter B, Kwok A S and Fayer M D 1994 Phonon-induced scattering between vibrations and multiphoton vibrational up-pumping in liquid solution *Chem. Phys. Lett.* **221** 412–18
- [97] Schosser C L and Dlott D D 1984 Temperature dependent libron relaxation in naphthalene *J. Chem. Phys.* **80** 1369–70
- [98] Hesselink W H and Wiersma D A 1980 Optical dephasing and vibronic relaxation in molecular mixed crystals: a picosecond photon echo and optical study of pentacene in naphthalene and *p*-terphenyl *J. Chem. Phys.* **73** 648–63

- [99] Hesselink W H and Wiersma D A 1983 Theory and experimental aspects of photon echoes in molecular solids *Spectroscopy and Excitation Dynamics of Condensed Molecular Systems (Modern Problems in Condensed Matter Sciences 4)* eds V M Agranovich and R M Hochstrasser (Amsterdam: North-Holland) pp 249–300
- [100] Hill J R, Chronister E L, Chang T-C, Kim H, Postlewaite J C and Dlott D D 1988 Vibrational relaxation and vibrational cooling in low temperature molecular crystals *J. Chem. Phys.* **88** 949–67
- [101] Chang T-C and Dlott D D 1988 Picosecond vibrational cooling in mixed molecular crystals studied with a new coherent Raman scattering technique *Chem. Phys. Lett.* **147** 18–24
- [102] Chang T-C and Dlott D D 1989 Vibrational cooling in large molecular systems: pentacene in naphthalene *J. Chem. Phys.* **90** 3590–602
- [103] Berg M and Vanden Bout D A 1997 Ultrafast Raman echo measurements of vibrational dephasing and the nature of solvent–solute interactions *Accounts Chem. Res.* **30** 65–71
- [104] Benjamin I, Barbara P F, Gertner B J and Hynes J T 1995 Nonequilibrium free energy functions, recombination dynamics, and vibrational relaxation of I_2^- in acetonitrile: molecular dynamics of charge flow in the electronically adiabatic limit *J. Phys. Chem.* **99** 7557–67
- [105] Stratt R M and Maroncelli M 1996 Nonreactive dynamics in solution: the emerging molecular view of solvation dynamics and vibrational relaxation *J. Phys. Chem.* **100** 12981–96
-

FURTHER READING

Seilmeier A and Kaiser W 1988 Ultrashort intramolecular and intermolecular vibrational energy transfer of polyatomic molecules in liquids *Ultrashort Laser Pulses and Applications (Topics in Applied Physics 60)* ed W Kaiser (Berlin: Springer) pp 279–315

Excellent discussion of experimental methods and summary of experimental results.

Dlott D D 1988 Dynamics of molecular crystal vibrations *Laser Spectroscopy of Solids II* ed W Yen (Berlin: Springer) pp 167–200

A review of VER in molecular crystals.

Harris C B, Smith D E and Russell D J 1990 Vibrational relaxation of diatomic molecules in liquids *Chem. Rev.* **90** 481–8

An excellent summary of theoretical approaches to VER, their significance and comparison to experiment.

Owrutsky J C, Raftery D and Hochstrasser R M 1994 Vibrational relaxation dynamics in solutions *Annu. Rev. Phys. Chem.* **45** 519–55

A review of vibrational energy relaxation of small molecules in solution.

Stratt R M and Maroncelli M 1996 Nonreactive dynamics in solution: the emerging molecular view of solvation dynamics and vibrational relaxation *J. Phys. Chem.* **100** 12981–96

A review of the multiple roles of vibrational energy transfer in liquids.

Voth G A and Hochstrasser R M 1996 Transition state dynamics and relaxation processes in solutions: a frontier of physical chemistry *J. Phys. Chem.* **100** 13034–49

An excellent discussion of developments in chemical reaction dynamics in liquids.

C3.6 Chaos and complexity in chemical systems

Raymond Kapral and Simon J Fraser

C3.6.1 INTRODUCTION

Complex chemical mechanisms are written as sequences of elementary steps satisfying detailed balance where the forward and reverse reaction rates are equal at equilibrium. The laws of mass action kinetics are applied to each reaction step to write the overall rate law for the reaction. The form of chemical kinetic rate laws constructed in this manner ensures that the system will relax to a unique equilibrium state which can be characterized using the laws of thermodynamics.

Most chemically reacting systems that we encounter are not thermodynamically controlled since reactions are often carried out under non-equilibrium conditions where flows of matter or energy prevent the system from relaxing to equilibrium. Almost all biochemical reactions in living systems are of this type as are industrial processes carried out in open chemical reactors. In addition, the transient dynamics of closed systems may occur on long time scales and resemble the sustained behaviour of systems in non-equilibrium conditions. A reacting system may behave in unusual ways: there may be more than one stable steady state, the system may oscillate, sometimes with a complicated pattern of oscillations, or even show chaotic variations of chemical concentrations.

Analogous considerations apply to spatially distributed reacting media where diffusion is the only mechanism for mixing chemical species. Under equilibrium conditions any inhomogeneity in the system will be removed by diffusion and the system will relax to a state where chemical concentrations are uniform throughout the medium. However, under non-equilibrium conditions chemical patterns can form. These patterns may be regular, stationary variations of high and low chemical concentrations in space or may take the form of time-dependent structures where chemical concentrations vary in both space and time with complex or chaotic forms.

In this chapter we shall examine how such temporal and spatial structures arise in far-from-equilibrium chemical systems. We first examine spatially uniform systems and develop the theoretical tools needed to analyse the behaviour of systems driven far from chemical equilibrium. We focus especially on the nature of chemical chaos, its characterization and the mechanisms for its onset. We then turn to spatially distributed systems and describe how regular and chaotic chemical patterns can form as a result of the interplay between reaction and diffusion.

This account is not exhaustive but provides a guide to the main theoretical ideas and experimental methods that have emerged in this subject. Fuller accounts and broad background are given in recent books devoted to this topic [[1](#), [2](#), [3](#), [4](#) and [5](#)].

C3.6.2 CHEMICAL REACTIONS AS DYNAMICAL SYSTEMS

Consider a spatially homogeneous reacting mixture where concentration gradients are removed by stirring or rapid

diffusion of the chemical species. In this circumstance the instantaneous state of the system is described by a vector set of chemical concentrations for the n chemical species, $\mathbf{c}(t) = (c_1(t), c_2(t), \dots, c_n(t))$, whose evolution is specified by the ordinary differential equations (ODEs) of mass action kinetics

$$\frac{dc(t)}{dt} = R_M(c(t); k). \quad (C3.6.1)$$

Here $R_M(c(t); k)$ is a vector of reaction velocities which are usually nonlinear functions of the chemical concentrations whose form is determined by the reaction mechanism. The reaction velocities also depend on the chemical rate constants, collectively described by the vector $k=(k_1, k_2, \dots)$, for the steps in the reaction mechanism.

For a closed chemical system with a mass action rate law satisfying detailed balance these kinetic equations have a unique stable (thermodynamic) equilibrium, $\lim_{t \rightarrow \infty} c(t) = c_{eq}$. In general, however, we shall be concerned with chemical reactions that are maintained far from chemical equilibrium by flows of reagents into and out of a continuously stirred tank reactor (CSTR). In this case the chemical kinetic equation (C3.6.1) must be supplemented with flow terms

$$\frac{dc(t)}{dt} = R_M(c(t); k) - k_f(c(t) - c_r) \equiv R(c(t); \mu) \quad (C3.6.2)$$

where k_f is the flow rate constant and c_r the vector of the feed concentrations. We have denoted the reaction rate in this general non-equilibrium case by $R(c(t); \mu)$ with μ as symbol for the collection of all parameters that characterize R : rate constants, feed concentrations and flow rates. Suppose the flow terms increase from zero; then this open system's stable state moves from the thermodynamic equilibrium, c_{eq} , to a nearby, non-equilibrium steady state, c_s , on the so-called thermodynamic branch. This non-equilibrium stable state of the system is the solution, $c=c_s$, of $R(c; \mu)=0$. However, if the flow terms become sufficiently large, this steady state becomes unstable and is replaced by new, non-equilibrium states characteristic of this well stirred system. (Transients mimicking the behaviour of the non-equilibrium states can be observed in closed reactors starting from initial conditions that are far from equilibrium.)

It is convenient to analyse these rate equations from a dynamical systems point of view similar to that used in classical mechanics where one follows the trajectories of particles in phase space. For the chemical rate law (C3.6.2) the 'phase space', conventionally denoted by Γ , is n -dimensional and the chemical concentrations, c_1, c_2, \dots, c_n , are taken as orthogonal coordinates of Γ , rather than the particle positions and velocities used as the coordinates in mechanics. In analogy to classical mechanical systems, as the concentrations evolve in time they will trace out a trajectory in Γ . Since the velocity functions in the system of ODEs (C3.6.2) do not depend explicitly on time, a given initial condition in Γ will always produce the same trajectory. The vector R of velocity functions in (C3.6.2) defines a phase-space (or trajectory) flow and in it is often convenient to think of these ODEs as describing the motion of a fluid in Γ with velocity field $R(c; \mu)$.

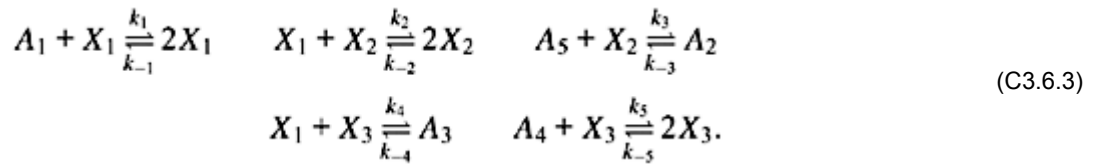
C3.6.2.1 CHEMICAL ATTRACTORS

Because of the underlying dissipative nature of the chemical systems that the ODEs (C3.6.2) represent, they have another important property: any volume in Γ will shrink as it evolves. For a given set of initial chemical concentrations the time evolution under the chemical rate law will approach arbitrarily close to some final set of points in

Γ after transients have decayed. This final set of phase-space points is the *attractor*, and the set of all initial conditions that eventually reaches the attractor is called its basin of attraction.

Attractors can be simple time-independent states (points in Γ), limit cycles (simple closed loops in Γ) corresponding to oscillatory variations of the chemical concentrations with a single amplitude, or chaotic states (complicated trajectories in Γ) corresponding to aperiodic variations of the chemical concentrations. To illustrate

the representation of chemical dynamics in concentration phase space and the existence of chemical attractors, we consider the Willamowski–Rössler (WR) model chemical system based on the following reaction mechanism [6]:



The species A_1, A_2, \dots, A_5 are pool chemicals whose concentrations are assumed to be fixed by flows of reagents into and out of the reactor while X_1, X_2 and X_3 are the species whose concentrations vary with time. For mechanism (C3.6.3) the mass action rate law is the system of ODEs

$$\begin{aligned}
 \frac{dc_1}{dt} &= k_1 c_{A_1} c_1 - k_{-1} c_1^2 - k_2 c_1 c_2 + k_{-2} c_2^2 - k_4 c_1 c_3 + k_{-4} \\
 \frac{dc_2}{dt} &= k_2 c_1 c_2 - k_{-2} c_2^2 - k_3 c_{A_5} c_2 + k_{-3} c_{A_2} \\
 \frac{dc_3}{dt} &= -k_4 c_1 c_3 + k_{-4} c_{A_3} + k_5 c_{A_4} c_3 - k_{-5} c_3^2.
 \end{aligned}
 \tag{C3.6.4}$$

For this model the parameter set μ consists of the rate constants and the constant pool chemical concentrations $\{c_{A_i}\}$ (Most chemical rate laws are constructed phenomenologically and often have cubic or other nonlinearities and irreversible steps. Such rate laws are reductions of the full underlying reaction mechanism.)

For certain parameter values this chemical system can exhibit fixed point, periodic or chaotic attractors in the three-dimensional concentration phase space. We consider the parameter set

$\mu = \{k_1 c_{A_1} = 31.2, k_{-1} = 0.2, k_2 = 1.45, k_3 c_{A_5} = 10.8, k_{-3} c_{A_2} = 0.12, k_4 = 1.02, k_{-4} = 0.01, k_5 c_{A_4}\}$. The rate constant k_{-2} will be taken as the control or *bifurcation* parameter which is varied to examine how the system attractor changes. As an example, the single-banded chaotic attractor at $k_{-2}=0.072$ for the WR model is shown in [figure C3.6.1\(a\)](#).

-4-

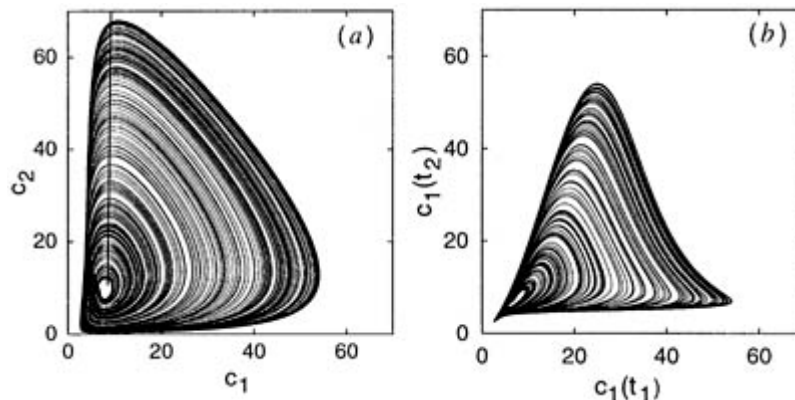


Figure C3.6.1 (a) WR single-banded chaotic attractor for $k_{-2} = 0.072$. This attractor is projected onto the (c_1, c_2) plane. The maximum value reached by $c_1(t)$ is $c_1^{\max} \approx 54.1$ and the minimum reached by $c_1^{\min} \approx 2.5$. The vertical line, at $c_1 = 8.5$ for $\dot{c}_1 < 1$, shows the position of the Poincaré section of the attractor used later. (b) A projection, onto the $(c_1(t_1), c_1(t_2))$ plane, of the chaotic attractor reconstructed from the set of delayed coordinates $\{c_1(t), c_1(t_1), c_1(t_2)\}$, where $t_1 = t + \tau_1$ and $t_2 = t + \tau_2$, for $0 \leq t < \infty$, and fixed delays $\tau_1 = 137$ and $\tau_2 = 200$. Note that both $c_1(t_1)$ and $c_1(t_2)$ reach a maximum of c_1^{\max} and a minimum of c_1^{\min} so that the three-dimensional reconstructed attractor is

confined to a cube with sides of length $c_1^{\max} - c_1^{\min}$. The central hole of the attractor is bounded along the diagonal in a similar way.

C3.6.2.2 PHASE-SPACE RECONSTRUCTION

The description of chemical reactions as trajectories in phase space requires that the concentrations of all chemical species be measured as a function of time, something that is rarely done in reaction kinetics studies. In addition, the underlying set of reaction intermediates is often unknown and the number of these may be very large. Usually, experimental data on the time variation of the concentration of a single chemical species or a small number of species are collected. (Some experiments focus on the simultaneous measurement of the concentrations of many chemical species and correlations in such data can be used to deduce the chemical mechanism [7].)

The trajectory description problem of chemical reactions is resolved by using phase-space reconstruction from a single time series [8]; this method uses delayed data at times: $t, t+\tau_1, t+\tau_2, \dots, t+\tau_{n-1}$ for an n -dimensional attractor, where usually $n \leq 3$. One may show that in place of the set of all chemical concentration one may use, say, $c_1(t), c_1(t+\tau_1), c_1(t+\tau_2), \dots$ to represent trajectories in the concentration phase space. Such phase-space reconstruction methods all rely on Whitney's embedding theorem [9] which allows a multi-dimensional attractor to be reconstructed from a single time series. Since phase-space volumes contract for dissipative chemical systems, as noted above, the final attractor may have a dimension much smaller than the original n -dimensional phase space. The effective behaviour of the system may often be captured in a phase space of only few dimensions even though many chemical intermediates are involved. To illustrate this reconstruction method, the set of delayed c_1 coordinates for chaotic attractor shown in figure C3.6.1(a): namely $(c_1(t), c_1(t+\tau), c_1(t+\tau_2))$, for t going from zero to some large value, was used to reconstruct the topologically equivalent attractor shown in figure C3.6.1(b).

C3.6.3 CHEMICAL CHAOS

We shall now analyse the structure of a chemical strange attractor and describe why the dynamics may be classified as chaotic.

C3.6.3.1 STRANGE ATTRACTORS

We begin by describing the features of a strange attractor. Figure C3.6.1(a) was constructed from a single chaotic trajectory which corresponds closely to a chaotic strange attractor, meaning that any such chaotic trajectory would look similar. Any point in the basin of the attractor approaches it asymptotically. A point moving on the strange attractor at some time comes arbitrarily close to any other point lying in the attractor, so motion on the attractor is *ergodic*; this is necessary but not sufficient for chaotic behaviour. To be chaotic the motion on the attractor must be *sensitive to initial conditions*, so that as time increases points on the attractor, however close together they may be initially, separate to distances comparable to the size of the attractor over a sufficiently long time. The rate of separation is measured by the Lyapunov number. In order that such separation be compatible with bounded motion, i.e. with the observation that the strange attractor lies in a finite volume of phase space, the chaotic (phase-space) flow must stretch and fold back onto itself. If we imagine a parcel of (compressible) fluid in phase space we see that this folding implies creation of infinitely many layers like the repeated folding of mille-feuille pastry. Correspondingly, the chaotic attractor is the result of an infinity of similar foldings. This dynamical recursion produces a self-similarity or fractal structure in the chaotic attractor. In summary, the chaotic attractor displays (exponential) separation of points or orbit segments and self-similar structure in the way these orbit segments are arranged in space.

C3.6.3.2 POINCARÉ SECTIONS AND NEXT-AMPLITUDE MAPS

For the strongly contracting phase volumes associated with chemical reactions, the three-dimensional continuous-

time flow can be reduced to a one-dimensional discrete-time map as follows. We first construct the Poincaré section of the attractor flow. For the projection shown in [figure C3.6.1\(a\)](#) the flow is counter-clockwise implying \dot{c} above the central hole in the single-banded attractor. Therefore this flow will circulate round this hole and repeatedly intersect the Poincaré surface of section ($c_1 \equiv c_1^{\text{sect}} = 8.5$), indicated by the heavy vertical line in [\(C3.6.1\(a\)\)](#), from right to left. Suppose that at time t_0 the trajectory intersects this Poincaré surface at a point $(c_2(t_0), c_3(t_0))$; at time t_1 it makes its next or so-called *first* return to the surface at point $(c_2(t_1), c_3(t_1))$. This process continues for times t_2, t_3, \dots , the difference $t_{n+1} - t_n$ being the period of the n th first-return trajectory segment. The sequence of points generated by these intersections is the Poincaré section and is plotted in [figure C3.6.2\(b\)](#). The thin line-like form arises from the strong contraction of the flow onto the attractor; thus the attractor resembles a two-dimensional surface formed from extremely tightly compressed and folded layers; sufficiently close to the attractor the trajectories tend to separate from one another across the attractor band. The function that takes $(c_2(t_n), c_3(t_n))$ into $(c_2(t_{n+1}), c_3(t_{n+1}))$ is the Poincaré map. The line-like form of the Poincaré section and its single-valuedness as a function of either coordinate permits a one-dimensional representation of the two-dimensional Poincaré map. To do this second part of the next-amplitude map construction we plot the c_2 component of the Poincaré map corresponding to the n th intersection of the chaotic trajectory with the Poincaré surface, $c_2(n) \equiv c_2(t_n)$, versus its value at the $(n + 1)$ th intersection, $c_2(n + 1) \equiv c_2(t_{n+1})$. This next-amplitude map is displayed in [figure C3.6.2 \(b\)](#). The map has a quadratic extremum, a roughly parabolic shape, and is densely filled by intersection points. We may then represent trajectories of the flow by iterates of this

-6-

map. To represent these iterates graphically, we first draw the bisectrix of the map, i.e. the diagonal line B in [figure C3.6.2\(b\)](#). By construction, any point on the map whose abscissa is $c_2(n)$ has ordinate $c_2(n + 1)$. This ordinate is given by moving horizontally to the bisectrix. Moving vertically from the bisectrix to the map makes $c_2(n + 1)$ the new abscissa and $c_2(n + 2)$ the corresponding ordinate. These two steps correspond to an iteration of the next-amplitude map and the procedure can be repeated to obtain a discrete trajectory indicating how the chaotic attractor structure is built up. A portion of such a discrete chaotic trajectory is shown in [figure C3.6.2\(b\)](#).

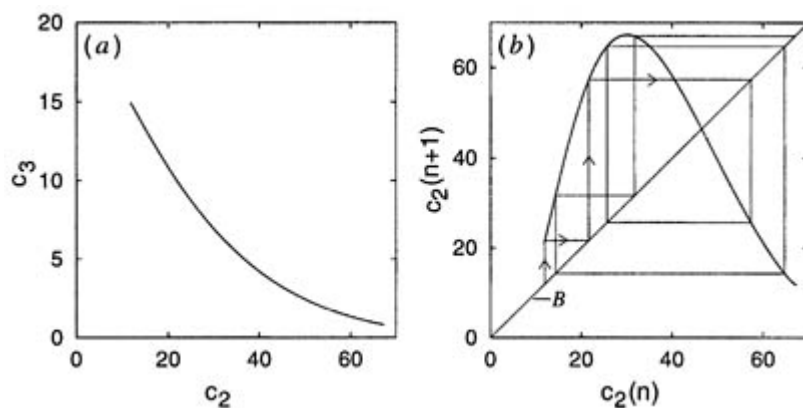


Figure C3.6.2 (a) The (c_2, c_3) Poincaré surface of a section of the phase flow, taken at $c_1^{\text{sect}} = 8.5$ with $c_1 < 0$, for the WR chaotic attractor at $k_{-2} = 0.072$. (b) The next-amplitude map constructed from pairs of intersection coordinates $\{\dots, (c_2(n+1), c_2(n+2), c_2(n+1)), \dots\}$. The sequence of horizontal and vertical line segments, each touching the diagonal B and the map, comprise a discrete trajectory. The direction on the first four segments is indicated.

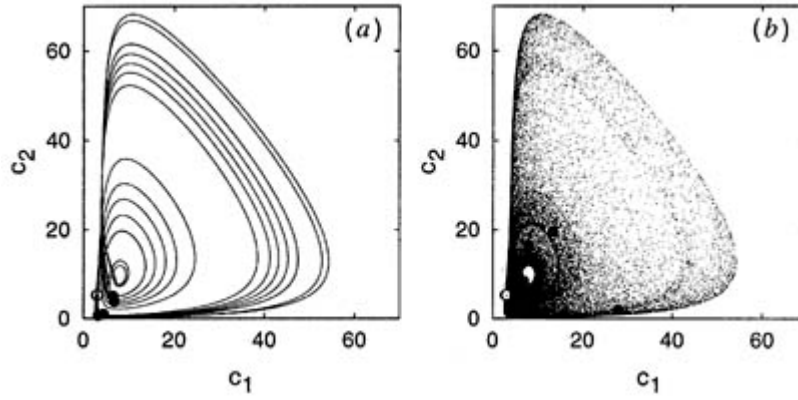


Figure C3.6.3 The spreading of an ensemble of four points on the WR chaotic attractor. (a) The initial tight, four-point ensemble of open circles (o) at $c_2 = 5.287 \dots$, $c_3 = 24.065 \dots$ and variable $c_1 = 2.884 \dots, 2.984 \dots, 3.084 \dots$, and $3.184 \dots$ spreading to the set of four filled circles (•) at time $t = 4.0$. The filled circles overlap in two pairs. (b) The spread from the same initial ensemble at time $t = 800.0$. The dispersion of the initial ensemble has the size of a full stop in the centre of the four overlapping open circles. The attractor is shown as a dust of stroboscopically plotted points so that the final ensemble of four filled circles (•) can be seen. One point lies on the inner edge of the central hole in the attractor. The density of the dust is an indicator of the *coarse-grained* density on the attractor.

-7-

C3.6.3.3 LYAPUNOV NUMBER AND FRACTAL DIMENSION

Chaotic attractors are complicated objects with intrinsically unpredictable dynamics. It is therefore useful to have some dynamical measure of the strength of the chaos associated with motion on the attractor and some geometrical measure of the structural complexity of the attractor. These two measures, the Lyapunov exponent or number [1] for the dynamics, and the fractal dimension [10] for the geometry, are related. To simplify the discussion we consider three-dimensional flows in phase space, but the ideas can be generalized to higher dimension.

As already mentioned, the motion of a chaotic flow is sensitive to initial conditions [11]; points which initially lie close together on the attractor follow paths that separate exponentially fast. This behaviour is shown in [figure C3.6.3](#) for the WR chaotic attractor at $k_{-2}=0.072$. The instantaneous rate of separation depends on the position on the attractor. However, a chaotic orbit visits any region of the attractor in a recurrent way so that an infinite time average of this exponential separation taken along any trajectory in the attractor is an invariant quantity that characterizes the attractor. If $\gamma(t)$ is a trajectory for the rate law ([c3.6.2](#)) then we can linearize the motion in the neighbourhood of γ to get

$$\frac{d\delta\mathbf{c}}{dt} = \left. \frac{\partial \mathbf{R}}{\partial \mathbf{c}} \right|_{\gamma} \delta\mathbf{c}. \quad (\text{C3.6.5})$$

The formal (or numerical) integration of this equation can be written as

$$\delta\mathbf{c}(t) = \mathbf{L}(t)\delta\mathbf{c}(0) \quad (\text{C3.6.6})$$

where $\mathbf{L}(t)$ is the displacement evolution matrix along γ and $\delta\mathbf{c}(t)$ is the solution of equation [C3.6.5] for initial displacement $\delta\mathbf{c}(0)$. Then the Lyapunov number is defined by

$$\bar{\lambda} = \lim_{t \rightarrow \infty} \frac{1}{2t} \ln[\text{Tr} \mathbf{L}^\dagger(t)\mathbf{L}(t)] \quad (\text{C3.6.7})$$

where L^\dagger is the adjoint of L . If the Lyapunov number $\bar{\lambda}$ is positive this indicates chaotic behaviour since $\bar{\lambda}$ is a measure of (the exponent for) the average rate at which trajectories separate on the attractor.

A chaotic attractor comprises line-like trajectory segments and so is topologically a one-dimensional object. However, the trajectories may lie arbitrarily close together in some regions of space, at least in the infinite time limit. In such regions a chaotic attractor has almost surface-like ‘filling’ properties. This unusual structure motivates the definition of a geometrical measure of chaos: the fractal, or more often, the box counting dimension of an attractor

$$D = \lim_{\varepsilon \rightarrow 0} \frac{\ln N(\varepsilon)}{\ln(1/\varepsilon)} = - \lim_{\varepsilon \rightarrow 0} \frac{\ln N(\varepsilon)}{\ln \varepsilon} \quad (\text{C3.6.8})$$

where, in two-dimensional or three-dimensional space $N(\varepsilon)$ is the minimum number of squares or cubes, respectively, of side ε that covers the attractor. This dimension can be calculated for the Poincaré section of the phase flow by covering it with successively smaller squares or for the entire attractor by covering it with successively smaller cubes

-8-

and measuring D as the slope of $N(\varepsilon)$ versus $\ln(1/\varepsilon)$ implied in (C3_6_8). This dimension is typically non-integer and is less than the phase-space dimension.

C3.6.3.4 EXPERIMENTAL OBSERVATIONS OF CHEMICAL CHAOS

The existence of chaotic oscillations has been documented in a variety of chemical systems. Some of the earliest observations of chemical chaos have been on biochemical systems like the peroxidase–oxidase reaction [12] and on the well known Belousov–Zhabotinskii (BZ) [13] reaction. The BZ reaction is the Ce-ion-catalyzed oxidation of citric or malonic acid by bromate ion. Early investigations of the BZ reaction used the techniques of dynamical systems theory outlined above to document the existence of chaos in this reaction. Apparent chaos in the BZ reaction was found by Hudson *et al* [14] and the data were analysed by Tomita and Tsuda [15] using a return-map method. Chaos was confirmed in the BZ reaction carried out in a CSTR by Roux *et al* [16, 17] and by Hudson and Mankin [18] who also used reconstruction from the electrode potentials of Pt and Br^- , and $d[\text{Pt}]/dt$ as independent variables. These demonstrations of true chemical chaos were achieved by a number of then new methods: power-spectral analysis [16, 19], trajectory reconstruction in phase space [16], and next-amplitude-map analysis [15, 20, 21]. The existence of true chemical chaos was signalled by a positive Lyapunov exponent calculated from the experimental return map. Since these early investigations, chaos has been documented in a variety of chemical systems. One aspect of these CSTR experiments was the observation that the stirring rate moved the bifurcation point(s) even if this rate was very large [22]. This effect depends on turbulent mixing and can be controlled but not eliminated by keeping the stirring rate constant. We now give examples of two related dynamical systems techniques used by experimentalists: phase-space reconstruction of chaotic attractors and the analysis of the associated next-amplitude maps. First, we discuss a study where an attractor was reconstructed from experimental data and then used to obtain a next-amplitude map [17].

Figure C3.6.4(a) shows an experimental chaotic attractor reconstructed from the Br^- electrode potential, i.e. the logarithm of the Br^- ion concentration, in the BZ reaction [17]. Such reconstruction is defined, in principle, for continuous time t . However, in practice, data are recorded as a discrete time series of measurements $\{X(t_i); i = 1, 2, \dots, i_{\max}\}$, consisting of thousands (i_{\max}) of data points. In our example $X(t_i)$ is proportional to $\ln[\text{Br}^-](t_i)$. The experimental attractor was reconstructed [17] in the space of the three variables, $X(t_i)$, $X(t_i + \tau)$ and $X(t_i + 2\tau)$, and C3.6.4(a) shows the projection of this attractor onto the $(X(t_i), X(t_i + \tau))$ plane. This attractor resembles that of the chaotic attractor shown in figure C3.6.1(a) and it can be demonstrated that the reconstructed attractor possesses the signatures of chaos: regions where trajectories locally spread or diverge and regions of re-injection and folding of the phase-space flow. Furthermore, we see that because the chaotic attractor is surface-like it has a fractal dimension close to two in spite of the fact that there are likely to be 30–40 chemical species involved in the

reaction so that the Euclidean dimension of the full concentration phase space is large. This points to the usefulness of phase-space reconstruction methods for low-dimensional chaotic attractors, especially for systems with a high but unknown phase-space dimensionality.

-9-

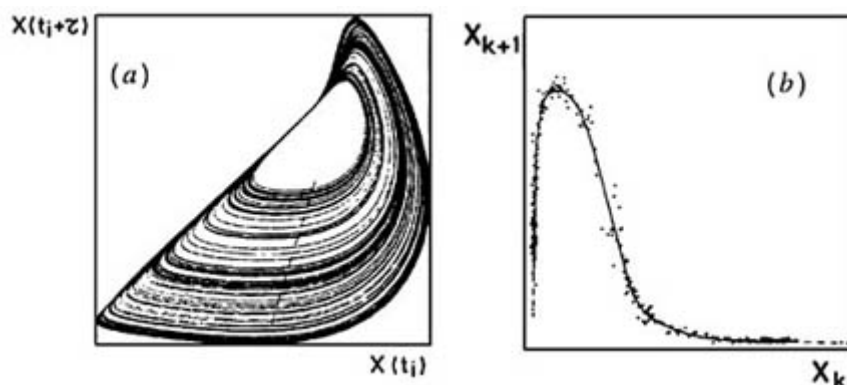


Figure C3.6.4 Single-banded chaotic attractor and next-amplitude map reconstructed from experimental data for the BZ reaction. (a) The reconstructed attractor projected onto the $(X(t_i), X(t_i + \tau))$ plane (see the text for a discussion of the notation). (b) The next-amplitude map in the (X_k, X_{k+1}) plane drawn from the surface of section taken at the dashed curve on the lower part of the attractor in (a). See the text for an explanation of the map construction. Reproduced by permission from Roux and Swinney [17].

We now examine how a next-amplitude-map was obtained from the attractor shown in figure C3.6.4(a) [17]. Consider the plane in this space whose projection is the dashed curve; i.e. a plane orthogonal to the $(X(t_i), X(t_i + \tau))$ plane. Then, for the k th intersection of the (continuous) trajectory with this plane, there will be a data point $(X(t_{i_k}), X(t_{i_k} + \tau), X(t_{i_k} + 2\tau))$ on the attractor that lies closest to the intersection of the continuous trajectory. A second discretization produces the set $\{X_k \equiv X(t_{i_k} + 2\tau) : k = 1, 2, \dots, k_{i_{\max}}\}$. This set is used in the construction of the next-amplitude map shown in figure C3.6.4(b) from the pairs of points $\{(X_k, X_{k+1}) : k = 1, 2, \dots, k_{i_{\max}} - 1\}$. This map has a single quadratic extremum, similar to that of the WR model described in detail earlier. Such maps (together with the technical constraint of negative Schwarzian derivative) [23] possess *universal* properties. In particular, the universal (U) sequence in which the periodic orbits appear [24] was observed in the BZ reaction in accord with this picture of the chemical dynamics.

C3.6.4 ROUTES TO CHAOS

The next problem to consider is how chaotic attractors evolve from the steady state or oscillatory behaviour of chemical systems. There are, effectively, an infinite number of routes to chaos [25]. However, only some of these have been examined carefully. In the simplest models they depend on a single control or bifurcation parameter. In more complicated models or in experimental systems, variations along a suitable curve in the control parameter space allow at least a partial observation of these well known routes. For chemical systems we describe period doubling, mixed-mode oscillations, intermittency, and the quasi-periodic route to chaos.

C3.6.4.1 THE PERIOD-DOUBLING ROUTE TO CHAOS

We first examine how chaos arises in the WR model using the rate constant k_{-2} as the bifurcation parameter. However, another parameter or set of parameters could be used to explore the behaviour. (Independent variation of p parameters

-10-

produces a p -dimensional bifurcation diagram.) In the context of experiments carried out in CSTRs the bifurcation parameter is usually taken to be the flow rate or a reservoir concentration. If we start with a value of $k_{-2} > 0.1715$ with all other rate constants fixed at the values given in [section C3.6.2](#), the WR reaction has a stable steady state or fixed point. We examine the sequence of transformations that takes place as k_{-2} decreases. At a certain value of $k_{-2} = k_{-2}^H \approx 0.1715$ the fixed point, $(c_1^H, c_2^H, c_3^H) \approx (8.801, 11.494, 15.048)$, loses its stability and the concentrations begin to oscillate with period T_0 . This is the Hopf bifurcation point and for $k_{-2} < k_{-2}^H$ the chemical attractor is a limit cycle. As k_{-2} decreases further the amplitude of the limit cycle grows (initially as $|k_{-2} - k_{-2}^H|^{1/2}$) until the system undergoes a further bifurcation at $k_{-2} \approx 0.1 \dots$ where the orbit undergoes a subharmonic bifurcation and its period doubles. To understand this bifurcation, imagine that the limit cycle lies on the surface of a Möbius band which is effectively a strip with a single twist in it. Motion on this band represents the slow relaxation of the system. At bifurcation the limit cycle becomes unstable but a stable orbit is born adjacent to it in the strip; this newborn orbit is geometrically equivalent to the edge of the Möbius strip as the width of the strip becomes arbitrarily small; because of the twist, the strip has only one edge of twice the length of the unstable limit cycle it contains. Therefore, the new stable orbit has twice the period of its parent limit cycle. An infinite sequence of these local twists or braids occur in the phase flow, generating an infinite subharmonic sequence of period-doubled orbits. The first two orbits of the main WR sequence are shown in [figure C3.6.5](#) there is a period-4 attractor at $k_{-2} = 0.095$. At the n th period doubling the period of the oscillation is $T_n \approx 2^n T_0$. In the limit $n \rightarrow \infty$ we arrive at the strange attractor where the time variation of the concentrations is no longer periodic. This is the *period-doubling route to chaos*.

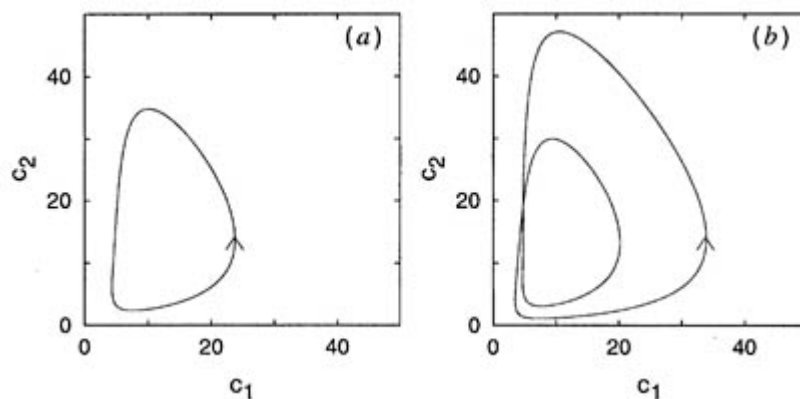


Figure C3.6.5 The first two periodic orbits in the main subharmonic sequence are shown projected onto the (c_1, c_2) plane. This sequence arises from a Hopf bifurcation of the stable fixed point for the parameters given in the text. The arrows indicate the direction of motion. (a) The limit cycle or period-1 orbit at $k_{-2} = 0.11$. (b) The first subharmonic or period-2 orbit at $k_{-2} = 0.095$.

It is instructive to view this sequence of transformations in terms of a bifurcation diagram. We use the procedure described earlier to examine the chaotic orbit: the intersections of the periodic trajectories with the Poincaré surface are recorded for each value of the rate constant k_{-2} . In [figure C3.6.6](#) we plot the concentration c_2 on the Poincaré plane versus k_{-2} . One can clearly see the sequence of period-doubling bifurcations leading eventually to the chaotic attractor. One can understand the origin of this sequence of bifurcations by considering the next-amplitude map discussed earlier. We remarked in [section C3.6.3.2](#) that this map has the nearly parabolic functional form shown in [figure C3.6.2\(b\)](#) so that, after suitable scaling, we can write the next-amplitude map in the standard quadratic form $c_2(n+1) = \lambda c_2(n)(1 - c_2(n))$, thereby preserving the local and global features of the bifurcation [diagram C3.6.6](#). We know what happens

as the standard map parameter λ is changed. This reduction of the problem to the study of a quadratic map allows one to make a detailed examination of the universal properties of this route to chaos since the only requirement is that the map function be quadratic in the vicinity of its maximum. Such an analysis was carried out by Feigenbaum [26] where the following scaling relation was derived: let λ_n be the value of λ at the n th period doubling and λ_∞ be

its value in the $n \rightarrow \infty$ limit. Then for sufficiently large n , $\lambda_n - \lambda_c = \delta(\lambda_{n+1} - \lambda_c)$ with $\delta = 4.6692 \dots$ a universal number for such period-doubling cascades for quadratic maps.

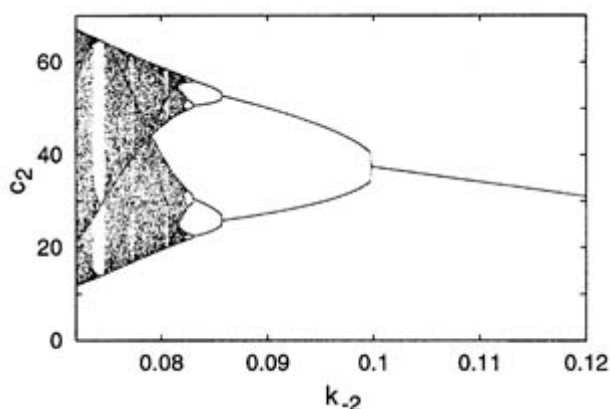


Figure C3.6.6 The figure shows the c_2 coordinate, for $c_1 < 0$, of the family of trajectories intersecting the (c_2, c_3) Poincaré surface at $c_1 = 8.5$ as a function of bifurcation parameter k_{-2} . As the ordinate k_{-2} decreases, the first subharmonic cascade is visible between $k_{-2} \approx 0.1$, the value of the first subharmonic bifurcation to $k_{-2} \approx 0.083$, the subharmonic limit of the first cascade. Periodic orbits that arise by the tangent bifurcation mechanism associated with type-I intermittency (see the text for references) can also be seen for values of k_{-2} smaller than this subharmonic limit. The left side of the figure ends at $k_{-2} = 0.072$, the value corresponding to the chaotic attractor shown in figure C3.6.1(a). Other regions of chaos can also be seen.

C3.6.4.2 OTHER ROUTES TO CHAOS

In addition to the period-doubling route to chaos there are other routes that are chemically important: mixed-mode oscillations (MMOs), intermittency and quasi-periodicity. Their signature is easily recognized in chemical experiments, so that they were seen early in the history of chemical chaos.

MMOs have been observed in many experiments. Typically, a MMO consists of one or more large amplitude oscillations followed by several small amplitude oscillations. The size of the small oscillations may grow slowly. Suppose L large oscillations are followed by s small oscillations, where L and s are integers, then this MMO can be encoded by L^s . For example, one large oscillation followed by one small oscillation is written 1^1 , and so on. Since large and small have a specific meaning in a series of chemical experiments we may find only small oscillations, encoded 0^1 or only large oscillations encoded 1^0 in the series. Chaotic MMOs consist of L^s oscillations interspersed randomly by $L^{s'}$, L'^s , or $L'^{s'}$ oscillations. Experimental observations and theoretical descriptions for the origins of such oscillations have been given [27].

Intermittency, in the context of chaotic dynamical systems, is characterized by long periods of nearly periodic or 'laminar' motion interspersed by chaotic bursts of random duration [28]. Within this broad phenomenological

description, three kinds of intermittency have been distinguished theoretically and some detected experimentally [1, 29]. The onset of the laminar phase is statistical but its subsequent evolution is deterministic until the start of the next burst, whereas the behaviour of the chaotic phase is largely probabilistic. For this kind of onset of chaotic motion the bifurcation parameter, ϵ say, is close to its critical bifurcation value ϵ_c for periodic motion. As ϵ passes from a 'chaotic' value through ϵ_c the motion goes from intermittent to marginally stable, to strictly stable periodic motion.

The quasiperiodic route to chaos is historically important. It arises from a succession of Hopf bifurcations. As already noted, a single Hopf bifurcation results in a limit cycle. The next Hopf bifurcation produces a phase flow that can be represented on the surface of a torus (doughnut). This flow is associated with two frequencies: if the ratio of these frequencies is irrational then the torus surface is densely covered by the phase trajectory, whereas if

the ratio is rational the orbit winds periodically on the torus surface with both frequencies determining the overall period. However, a further Hopf bifurcation leads to an unstable torus flow which deforms into a chaotic flow. The nature of this instability was first discussed independently by Kupka and Smale; an equivalent theory of this breakdown of quasi-periodic flow to chaotic flow was proposed by Ruelle and Takens and was developed by them and Newhouse [30]. The quasiperiodic route to chaos was important because it was the first example of a transition to chaos that involved *few* modes, in contrast to the classical model of Landau of a gradual wandering into chaos as successive modes become unstable.

C3.6.5 CHEMICAL PATTERNS AND SPATIO-TEMPORAL CHAOS

Thus far we have considered systems where stirring ensured homogeneity within the medium. If molecular diffusion is the only mechanism for mixing the chemical species then one must adopt a local description where time-dependent concentrations, $c(\mathbf{r}, t)$, are defined at each point \mathbf{r} in space and the evolution of these local concentrations is given by a reaction-diffusion equation

$$\frac{\partial c(\mathbf{r}, t)}{\partial t} = R(c(\mathbf{r}, t); \mu) + D\nabla^2 c(\mathbf{r}, t) \quad (\text{C3.6.9})$$

where D is a matrix of diffusion coefficients. In addition to the temporal behaviour described above, one now has the possibility of chemical pattern formation which may lead to spatio-temporal chaos. In order to investigate chemical pattern formation under controlled non-equilibrium conditions, experiments are now carried out in continuously fed unstirred reactors (CFURs) [31]. In such reactors, well stirred reagent baths are in contact with a gel or porous medium within which the chemicals mix and react in the absence of stirring effects other than diffusion. Since the reagent baths are CSTRs they continuously supply and remove reactants and products from the reaction-diffusion medium, and chemical pattern formation can be controlled and maintained indefinitely. This has allowed experimentalists to make detailed studies of chemical pattern formation.

We shall describe some of the common types of chemical patterns observed in such experiments and comment on the mechanisms for their appearance. In keeping with the theme of this chapter we focus on states of spatio-temporal chaos or on regular chemical patterns that lead to such turbulent states. We shall touch only upon the main aspects of this topic since there is a large variety of chemical patterns and many mechanisms for their onset [2, 3, 5, 32].

C3.6.5.1 EXCITABLE MEDIA

Excitable media are some of the most commonly observed reaction-diffusion systems in nature. An excitable system possesses a stable fixed point which responds to perturbations in a characteristic way: small perturbations return quickly to the fixed point, while larger perturbations that exceed a certain threshold value make a long excursion in concentration phase space before the system returns to the stable state. In many physical systems this behaviour is captured by the dynamics of two concentration fields, a fast activator variable u with cubic nullcline and a slow inhibitor variable v with linear nullcline [33]. The FitzHugh-Nagumo equation [34], derived as a simple model for nerve impulse propagation but which can also apply to a chemical reaction scheme [35], is one of the best known equations with such activator-inhibitor kinetics:

$$\begin{aligned} \frac{du}{dt} &= -u^3 + u - v = R_u(u, v) \\ \frac{dv}{dt} &= \varepsilon(v - au + b) = R_v(u, v). \end{aligned} \quad (\text{C3.6.10})$$

Figure C3.6.7(a) shows the $\dot{u}=0$ and $\dot{v}=0$ nullclines of this system along with trajectories corresponding to sub- and super-threshold excitations. The trajectory arising from the sub-threshold perturbation quickly relaxes back to the stable fixed point. Three stages can be identified in the trajectory resulting from the super-threshold perturbation: an excited stage where the phase point quickly evolves far from the fixed point, a refractory stage where the system relaxes back to the stable state and is not susceptible to additional perturbation and the resting state where the system again resides at the stable fixed point.

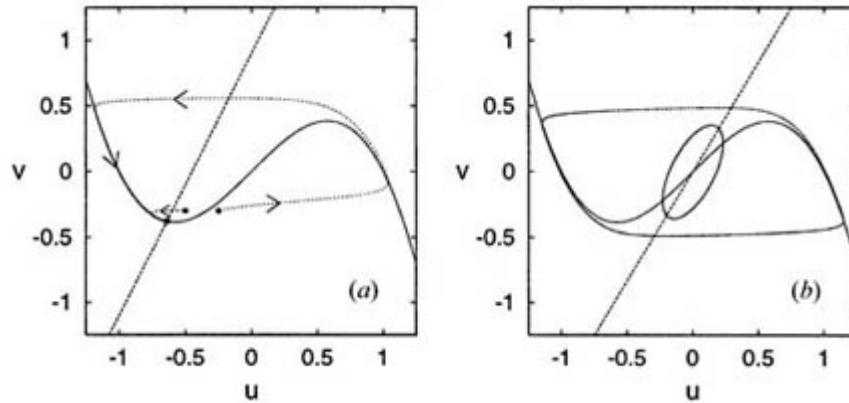


Figure C3.6.7 Cubic ($\dot{u}=0$) and linear ($\dot{v}=0$) nullclines for the FitzHugh-Nagumo equation. (a) The excitable domain showing trajectories resulting from sub- and super-threshold excitations. (b) The oscillatory domain showing limit cycle orbits: small inner limit cycle close to Hopf point; large outer limit cycle far from Hopf point.

An excitable medium is a diffusively coupled array of such local excitable elements described by the reaction–diffusion equation (C3.6.9) with \mathbf{R} given by (C3.6.10) and $\mathbf{c} = (u, v)$. Imagine a local super-threshold perturbation applied to the system in the homogeneous resting state. Due to diffusive coupling, the perturbation will excite neighbouring regions of the medium. The originally perturbed region will then relax to the refractory stage where it is no longer susceptible

to perturbation, and finally back to the stable steady state. Consequently, a circular wave of excitation with a refractory tail will propagate outward through the medium (see figure C3.6.8(a)). If the excitable system is periodically stimulated in a local region of the medium (a pacemaker region) a target pattern comprising a set of concentric rings of excitation will be observed.

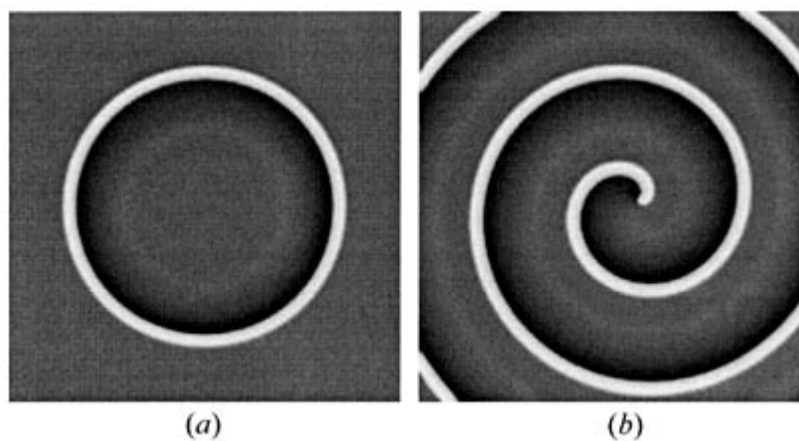


Figure C3.6.8 (a) A growing ring of excitation in an excitable FitzHugh–Nagumo medium. (b) A spiral wave in the same system.

If an excitable wave is broken, for instance by an obstacle or inhomogeneity in the medium, since the front velocity is smaller at the tip than the rest of the wave front, free ends of wave fronts will curl leading to the formation of spiral waves in the system. An example of a spiral wave is shown in figure C3.6.8(b). Excitable waves are seen in many chemical and biological systems. The often studied Belousov-Zhabotinskii (BZ) reaction was one of the first systems in which such waves were observed [13, 36]. Chemical waves of this type have been studied extensively in catalytic oxidation of CO on Pt[37]. In biological contexts, waves of this type occur in the aggregation stage of the slime mould *Dictyostelium discoideum* where the chemical signalling is through periodic waves of cAMP; also the Ca^{2+} waves in systems like *Xenopus laevis* oocytes and pancreatic β cells fall into this category 38. Electrochemical waves in cardiac and nerve tissue have this origin and the appearance and/or breakup of spiral wave patterns in excitable media are believed to be responsible for various types of arrhythmias in the heart [39, 40]. Figure C3.6.9 shows an excitable spiral wave in dog epicardial muscle [41].

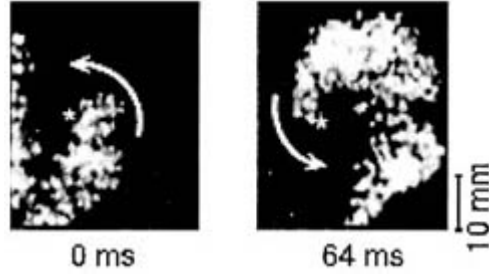


Figure C3.6.9 Spiral electrochemical wave in dog epicardial muscle visualized using a voltage-sensitive dye. Reproduced by permission from Pertsov and Jalife [41].

-15-

The cores of the spiral waves need not be stationary and can move in periodic, quasi-periodic or even chaotic ‘flower’ trajectories [42, 43]. In addition, spatio-temporal chaos can arise if such spiral waves break up and the spiral wave fragments spawn pairs of new spirals [42, 44].

C3.6.5.2 OSCILLATORY AND CHAOTIC MEDIA

We described earlier how a stable steady state may give rise to a periodic oscillation through a Hopf bifurcation. The steady state of the FitzHugh–Nagumo model can undergo such a Hopf bifurcation. Consider the situation shown in figure C3.6.7(b) for $b = 0$ where there is a single fixed point at the origin $(u^*, v^*) = (0, 0)$. This fixed point is stable if $\varepsilon > a^{-1}$, $a < 1$, and becomes unstable at $\varepsilon = \varepsilon_H = a^{-1}$ through a Hopf bifurcation spawning a limit cycle encircling the origin.

Consider the analogue of such a bifurcation in a spatially distributed system and imagine tuning a bifurcation parameter μ (in the parameter set μ) in (C3.6.9) through such a bifurcation point, μ_H , and let $\lambda = |\mu - \mu_H|^{1/2}$ gauge the distance from the bifurcation point. One may then expand the local concentration about the steady state c^* as $c(\mathbf{r}, t) = c^* + A(\mathbf{r}, t)\hat{e} + \text{c.c.}$, where $A(\mathbf{r}, t)$ is a complex amplitude and \hat{e} is an eigenvector of the linearized reaction–diffusion problem. Then, in the vicinity of the Hopf bifurcation point, it is possible to transform the reaction–diffusion equation into a universal equation for the complex amplitude $A(\mathbf{r}, t)$ [45]:

$$\frac{\partial A(\mathbf{r}, t)}{\partial t} = A - (1 + i\beta)|A|^2 A + (1 + i\alpha)\nabla^2 A. \quad (\text{C3.6.11})$$

This complex Ginzburg–Landau equation describes the space and time variations of the amplitude A on long distance and time scales determined by the parameter distance from the Hopf bifurcation point. The parameters α and β can be determined from a knowledge of the parameter set μ and the diffusion coefficients of the reaction–diffusion equation. For example, for the FitzHugh–Nagumo equation we have $\alpha = (\mathbf{D}_v - \mathbf{D}_u)/[\omega_0(\mathbf{D}_v + \mathbf{D}_u)]$ and $\beta = -1/\omega_0$. The Ginzburg–Landau equation parameters may also be extracted from the experimental data and this has been done for the BZ reaction [46]. Through such an analysis one can study general features of oscillatory media,

independent of specific features of the reaction kinetics.

The complex Ginzburg–Landau equation also supports spiral wave solutions [47]. The core of a spiral wave is a point topological defect where the complex amplitude A vanishes [48]. In certain parameter regions, one finds a type of spatio-temporal chaos termed defect-mediated turbulence where the average number of topological defects is stationary but their instantaneous number fluctuates: defects of opposite topological charge may collide and annihilate or defects may nucleate in pairs as a result of ‘pinching’ of wave fronts [49]. [Figure C3.6.10](#) shows the system in the defect-mediated turbulence regime and illustrates distribution of spiral defects in the turbulent dynamics described above. Such defect-mediated turbulence has been observed in experiments on the BZ reaction [50]. [Figure C3.6.11\(a\)](#) shows the chemical pattern near the onset of the instability giving rise to spatio-temporal turbulence. Note that small well defined spirals can still be seen embedded in a sea of turbulent dynamics while in [figure C3.6.11\(b\)](#) well beyond the instability, one sees fully developed turbulence.

-16-

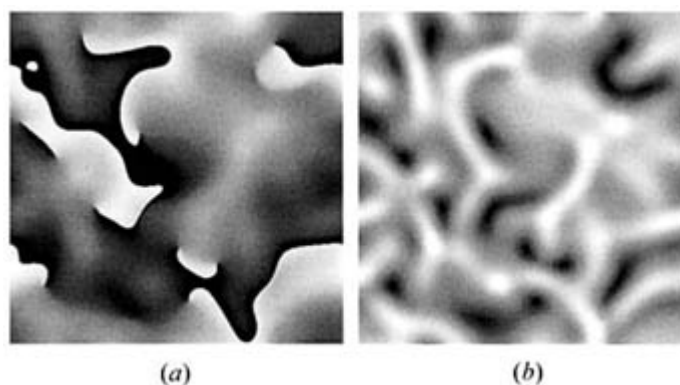


Figure C3.6.10 Defect-mediated turbulence in the complex Ginzburg–Landau equation. (a) The phase, $\arg(A)$, as grey shades. (b) The amplitude $|A|$, with a similar color coding. In the left panel topological defects can be identified as points around which one finds all shades of grey. Note the apparently random spatial pattern of amplitudes.

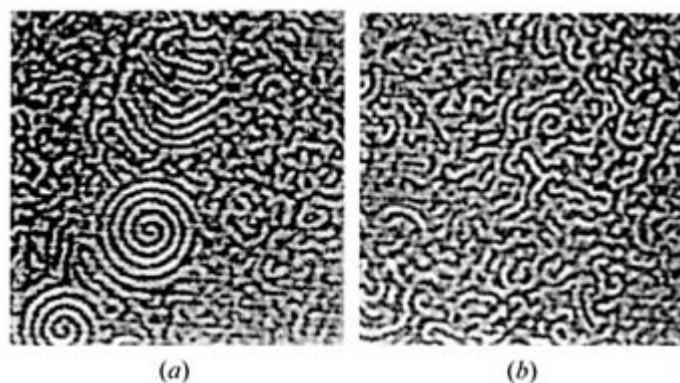


Figure C3.6.11 Defect-mediated turbulence in the BZ reaction. (a) Spatial structure close to the instability. (b) Fully developed spatio-temporal turbulence. The control parameter is the concentration of H_2SO_4 in the feed reactor. Reproduced by permission from Ouyang and Flesselles [50].

The local dynamics of the systems considered thus far has been either steady or oscillatory. However, we may consider reaction–diffusion media where the local reaction rates give rise to chaotic temporal behaviour of the sort discussed earlier. Diffusional coupling of such local chaotic elements can lead to new types of spatio-temporal periodic and chaotic states. It is possible to find phase-synchronized states in such systems where the amplitude varies chaotically from site to site in the medium whilst a suitably defined phase is synchronized throughout the medium [51]. Such phase synchronization may play a role in layered neural networks and perceptive processes in mammals. Somewhat surprisingly, even when the local dynamics is chaotic, the system may support spiral waves

[52, 53 and 54]. The origin of such spiral waves in chaotic media can again be traced to the phenomenon of phase synchronization. The notion of a defect at the core of the spiral remains valid even for these chaotic media, so the phase-coherent dynamics necessary for the existence of a spiral wave survives the amplitude turbulence. New phenomena can arise: in addition to point topological defects one can find synchronization line defects whose dynamics may be chaotic. Such synchronization line-defect dynamics has been observed in the BZ medium reaction [54, 55 and 56].

-17-

C3.6.5.3 TURING PATTERNS

If the diffusion coefficients of the chemical species are sufficiently different, new types of chemical instability arise which can lead to the formation of chemical patterns and ultimately to spatio-temporal chaotic behaviour.

One of the best known such instabilities is the Turing bifurcation proposed in 1952 as a possible mechanism for morphogenesis [57]. While the relevance of this type of pattern-forming instability for biological systems is still a matter of debate, such Turing patterns have been observed in laboratory chemical experiments [58, 59 and 60]. A Turing bifurcation involves the destabilization of a homogeneous steady state to form an inhomogeneous state or chemical pattern whose wavelength depends on the kinetic parameters and diffusion coefficients of the system. Turing bifurcations are often discussed in terms of activator–inhibitor kinetics like that of the FitzHugh–Nagumo equation above [61]. Consider two chemical species, X_1 and X_2 , with concentration vector, $\mathbf{c} = (c_1, c_2)$, that satisfies a two-variable reaction–diffusion equation where, as usual, $\mathbf{R}(\mathbf{c})$ describes the kinetics and \mathbf{D} is a diagonal diffusion-coefficient matrix with elements D_1 and D_2 . We suppose the system possesses a homogeneous stable steady state, \mathbf{c}^* , obtained from the solution of $\mathbf{R}(\mathbf{c}^*) = 0$. To determine the conditions for a Turing bifurcation to occur we consider the perturbation of this homogeneous steady state to *inhomogeneous* perturbations, $\mathbf{c}(\mathbf{r}, t) = \mathbf{c}^* + \delta\mathbf{c}(\mathbf{r}, t)$. We may linearize the reaction–diffusion to obtain

$$\frac{\partial \delta\mathbf{c}(\mathbf{r}, t)}{\partial t} = \mathbf{A}\delta\mathbf{c}(\mathbf{r}, t) + \mathbf{D}\nabla^2\delta\mathbf{c}(\mathbf{r}, t) \quad (\text{C3.6.12})$$

where $\mathbf{A} = (\partial\mathbf{R}/\partial\mathbf{c})_{\mathbf{c}=\mathbf{c}^*}$ is the matrix that specifies the chemical rate evolution about the steady state \mathbf{c}^* . To determine the stability of the steady state it is useful to examine the behaviour of the Fourier components of the concentration field, $\hat{\mathbf{c}}_k$, which satisfies the Fourier transform of equation C3.6.12:

$$\frac{\partial \delta\hat{\mathbf{c}}_k(t)}{\partial t} = (\mathbf{A} - k^2\mathbf{D})\delta\hat{\mathbf{c}}_k(t) \equiv \mathbf{B}\hat{\mathbf{c}}_k(t). \quad (\text{C3.6.13})$$

Now we may state the well known conditions for a Turing bifurcation. If $A_{11} < 0$ and $A_{22} < 0$ we say species X_1 is the activator and species X_2 is the inhibitor. Then, for a Turing bifurcation to occur we must have $\det \mathbf{B} = 0$, $\text{Tr } \mathbf{B} > 0$ and $A_{11}D_2 + A_{22}D_1 > 0$. The (unique) wavenumber at the bifurcation is

$$k_c = \left(\frac{\det \mathbf{A}}{D_1 D_2} \right)^{1/4}. \quad (\text{C3.6.14})$$

Furthermore, since the bifurcation must occur from a stable homogeneous steady state we must have $D_1/D_2 < 1$; i.e. the diffusion coefficient of the inhibitor is greater than that of the activator. The critical diffusion ratio at the bifurcation is

-18-

$$\frac{D_1}{D_2} = A_{22}^{-1} (\det \mathbf{A} - A_{12}A_{21} + 2(A_{12}A_{21} \det \mathbf{A})^{1/2}). \quad (\text{C3.6.15})$$

Consequently, when D_1/D_2 exceeds the critical value, close to the bifurcation one expects to see the appearance of chemical patterns with characteristic length $\ell = 2\pi / k_c$. Beyond the bifurcation point a band of wave numbers is unstable and the nature of the pattern selected (spots, stripes, etc.) depends on the nonlinearity and requires a more detailed analysis. Chemical Turing patterns were observed in the chlorite–iodide–malonic acid (CIMA) system in a gel reactor [58, 59 and 60]. Figure C3.6.12(a) shows an experimental CIMA Turing spot pattern [59].

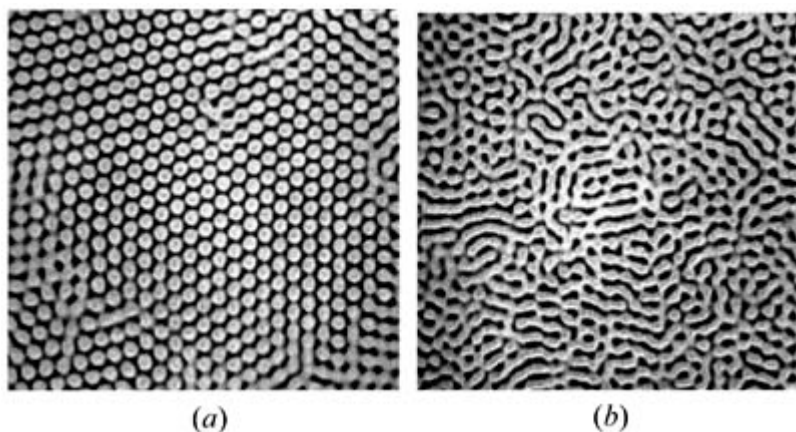


Figure C3.6.12 (a) Turing spot pattern in the CIMA reaction. (b) Tio-temporal turbulence near the Turing bifurcation. Reproduced by permission from Ouyang and Swinney [59].

The Turing mechanism requires that the diffusion coefficients of the activator and inhibitor be sufficiently different; but the diffusion coefficients of small molecules in solution differ very little. The chemical Turing patterns seen in the CIMA reaction used starch as an indicator for iodine. The starch indicator complexes with iodide which is the activator species in the reaction. As a result, the complexing reaction with the immobilized starch molecules must be accounted for in the mechanism and leads to the possibility of Turing pattern formation even if the diffusion coefficients of the activator and inhibitor species are the same [62].

One may also observe a transition to a type of defect-mediated turbulence in this Turing system (see figure C3.6.12 (b)). Here the defects divide the system into domains of spots and stripes. The defects move erratically and lead to a turbulent state characterized by exponential decay of correlations [59]. Turing bifurcations can interact with the Hopf bifurcations discussed above to give rise to very complicated spatio-temporal patterns [63, 64].

C3.6.5.4 CHEMICAL FRONT INSTABILITIES

Another class of instabilities that are driven by differences in the diffusion coefficients of the chemical species determines the shapes of propagating chemical wave and flame fronts [65, 66].

As an example of chemical front instability consider a simple cubic autocatalytic reaction, $A + 2B \rightarrow 3B$, occurring in

a two-dimensional geometry where the ‘fuel’ A occupies the right-hand region and the autocatalyst occupies the left-hand region [67]. We suppose the reaction occurs under isothermal conditions which can be achieved for condensed phase reactions. The species B will consume the fuel A and the chemical front that separates the A and B species will move to the right. (For flame fronts one must generally couple the reaction kinetics to the variations in the temperature of the system.)

If the diffusion coefficient of species A is less than that of B ($D_A < D_B$) the propagating front will be planar. However, if D_A is sufficiently greater than D_B , the planar front will become unstable to transverse perturbations and chaotic front motion will ensue. To understand the origin of the mechanism of the planar front destabilization consider the following: suppose the interface is slightly non-planar. We would like to know if the dynamics will tend to eliminate this non-planarity or accentuate it. Let $D_B \gg D_A$. The situation is depicted schematically in figure C3.6.13 where large diffusion fluxes are indicated by \rightarrow and smaller diffusion fluxes by \dashrightarrow . For the part of the B front that protrudes into the A region, fast diffusion of B leads to dispersal of B and suppresses the autocatalytic reaction that requires two molecules of B . The front will have difficulty advancing here. In the region where A protrudes into B , A will react leading to advancement of the front. The net effect is to remove any initial non-planarity and give rise to a planar front.

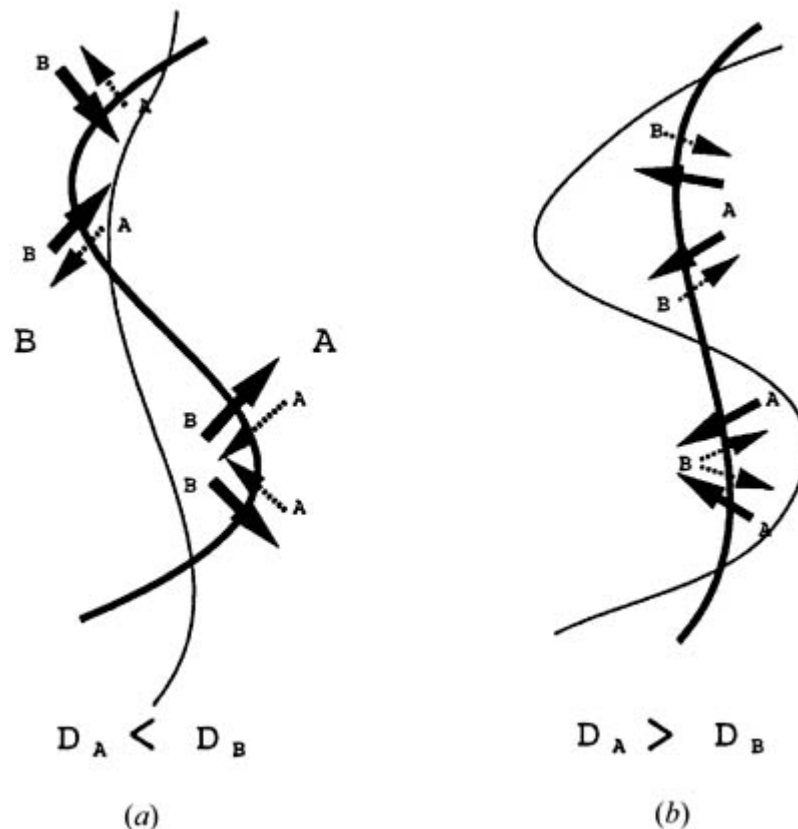


Figure C3.6.13 Schematic illustration of how the front instability arises for the case (a) $D_B \gg D_A$ and (b) $D_B \ll D_A$.

If $D_B \ll D_A$, in regions where B protrudes into A , rapid A diffusion will lead to conversion of A to B leading to front

advance. In regions where A protrudes into B , small diffusion of B into the A region does not favour the autocatalytic conversion so the front will not advance rapidly here. Consequently, any small non-planarity will grow to make the front even more non-planar. Therefore, we expect that for some ratio of diffusion coefficients, $d = D_A/D_B > 1$, the planar front will lose its stability [68]. An example of the front dynamics for $d = 5$ is shown in figure C3.6.14 where the minima in the front profile are plotted versus time. The resulting space-time plot shows the chaotic nature of the front dynamics. The (black) minima act like ‘particles’ in the system: they move and when they collide they coalesce to form a single minimum. If the distance between two minima is too large, a new minimum is formed. Thus, the average density of ‘particles’ per unit length of the interface remains constant but the instantaneous number of ‘particles’ fluctuates due to the creation and annihilation events for the minima.

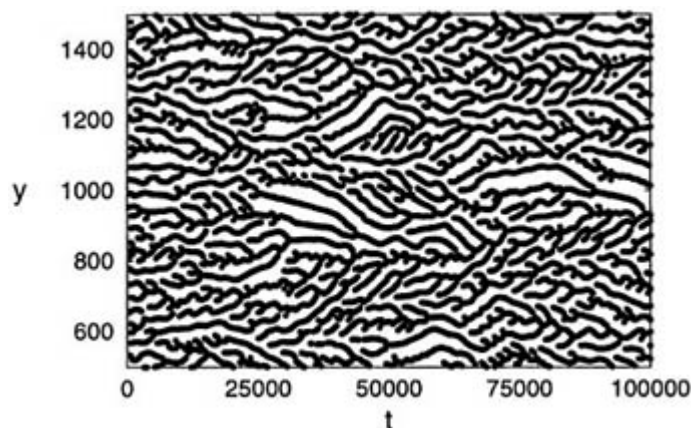


Figure C3.6.14 Space–time (y,t) plot of the minima (black) in the cubic autocatalysis front $\phi(y,t)$ in equation C3.6.16 showing the nature of the spatio-temporal chaos.

In order to investigate such front instabilities quantitatively one may derive an equation for the profile $\phi(y,t)$ of the front directly from the reaction–diffusion equation. This Kuramoto–Sivashinsky equation [69]

$$\frac{\partial \phi(y, t)}{\partial t} = v \frac{\partial^2 \phi}{\partial y^2} - \frac{v}{2} \left(\frac{\partial \phi}{\partial y} \right)^2 - \kappa \frac{\partial^4 \phi}{\partial y^4} \quad (\text{C3.6.16})$$

describes a number of general features of such front dynamics. The parameters v , ν and κ may be related to the parameters in the original reaction–diffusion equation. The nonlinear term accounts for the fact that the velocity of the front depends on its curvature, while the gradient terms arise from diffusive effects. The coefficient of the fourth-order gradient is positive while the sign of v depends on the diffusion coefficient ratio d : for $d > d_c$, where d_c is a critical value of d , v changes from being positive to negative. This negative value of the ‘diffusion coefficient’ leads to an instability whose growth is controlled by the stabilizing fourth-order term. Instead of studying the full reaction–diffusion equation we may now explore the front dynamics directly through equation (C3.6.16). This equation yields front dynamics like that described above.

In addition to flame fronts, which have been extensively studied experimentally, front instabilities have been investigated for the isothermal cubic autocatalytic iodate arsenous acid system [70] as well as for polymerization

reactions where thermal and hydrodynamic effects lead to complicated front patterns [71]. Front instabilities also play a role in determining the labyrinth patterns seen in recent chemical experiments [72].

C3.6.6 CONCLUSION

Our understanding of the development of oscillations, multi-stability and chaos in well stirred chemical systems and pattern formation in spatially distributed systems has increased significantly since the early observations of these phenomena. Most of this development has taken place relatively recently, largely driven by development of experimental probes of the dynamics of such systems. In spite of this progress our knowledge of these systems is still rather limited, especially for spatially distributed systems.

Several important topics have been omitted in this survey. We have described only a few of the routes by which chaos can arise in chemical systems and have made no attempt to describe in detail the features of the different kinds of chemical strange attractor seen in experiments. A wide variety of chemical patterns have been observed and while the many aspects of the mechanisms for their appearance are understood, some features like nonlinear

pattern selection still present challenges and new patterns continue to be discovered. An ubiquitous class of chemical patterns that was not discussed here are those that arise from diffusion-limited aggregation (DLA) [73]. Such DLA clusters are seen in many contexts, including electrochemical deposition processes, and are often analysed using the concepts of fractal geometry [10] and wavelets [74, 75]. Also, methods for controlling chemical chaos [76] have not been discussed in this chapter although they have potential applications for both industrial processes and biological systems.

In spite of these limitations it is hoped that this chapter will provide an introduction to the unusual phenomena that chemically reacting systems exhibit when driven far from equilibrium and an indication of how these phenomena may be analysed. Although such systems were often regarded as curiosities in the past, it is now clear that they are the rule rather than the exception in nature and deserve our full attention.

REFERENCES

- [1] Bergé P, Pomeau Y and Vidal C 1984 *Order within Chaos: Towards a Deterministic Approach to Turbulence* (New York: Wiley)
- [2] Nicolis G 1995 *Introduction to Nonlinear Science* (Cambridge: Cambridge University Press)
- [3] Kapral R and Showalter K (eds) 1994 *Chemical Waves and Patterns* (Dordrecht: Kluwer)
- [4] Scott S K 1991 *Chemical Chaos* (New York: Oxford University Press)
- [5] Walgraef D 1997 *Spatio-Temporal Pattern Formation* (New York: Springer)
- [6] Willamowski K D and Rössler 1980 *Z. Naturfor.* **35** 317
- [7] Arkin A, Shen P and Ross J 1977 *Science* **277** 1275

- [8] Roux J C, Simoyi R H and Swinney H L 1983 *Physica D* **8** 257
Packard N H, Crutchfield J P, Farmer J D and Shaw R S 1980 *Phys. Rev. Lett.* **45** 712
- [9] Whitney H 1936 *Ann. Math.* **37** 645
- [10] Mandelbrot B B 1982 *The Fractal Geometry of Nature* (San Francisco: Freeman)
Falconer K J 1990 *Fractal Geometry: Mathematical Foundations and Applications* (New York: Wiley)
- [11] Ruelle D and Takens F 1971 *Commun. Math. Phys.* **20** 167
Ruelle D and Takens F 1971 *Commun. Math. Phys.* **23** 343
Guckenheimer J 1979 *Commun. Math. Phys.* **70** 133
- [12] Degn H 1968 *Nature* **217** 1047
Hauser M J B, Olsen L F, Bronnikova T V and Schaffer W M 1997 *J. Phys.: Condens. Matter* **101** 5075
Larter R, Olsen L F, Steinmetz C G and Geest T 1993 *Chaos in Chemistry and Biochemistry* ed R J Field and L Györgyi (Singapore: World Scientific) p 175
Aguda B D, Frisch L L H and Olsen L F 1990 *J. Am. Chem. Soc.* **112** 6652
Nakamura S, Yakota K and Yamazaki I 1969 *Nature* **222** 794
Olsen L F and Degn H 1978 *Biochem. Biophys. Acta* **523** 321
Fed'kina V R, Bronnikova T V and Atuallakhanov F I 1981 *Studia Biophys.* **24** 165
Fed'kina V R, Atuallakhanov F I and Bronnikova T V 1988 *Theor. Exp. Chem.* **24** 165
Hauck T and Schneider F W 1993 *J. Chem. Phys.* **97** 391
Hauck T and Schneider F W 1994 *J. Phys.: Condens. Matter* **97** 2072
Geest T, Steinmetz C G, Larter R and Olsen L F 1992 *J. Phys.: Condens. Matter* **96** 5678
Hauser M B J, Anderson S and Olsen L F 1996 *Plant Peroxidases: Biochemistry and Physiology* ed C Obinger *et al* (Geneva: University of Geneva Press) p 88

- [13] Belousov B 1958 *Sb. Ref. Radiats. Med.* (Moscow) p 145
 Zhabotinskii A M 1964 *Biofizika* **9** 306
 Zaikin A N and Zhabotinsky A M 1970 *Nature* **225** 535
- [14] Hudson J L, Hart M and Marinko D 1979 *J. Chem. Phys.* **71** 1601
- [15] Kazuhisa Tomita and Ichuro Tsuda 1980 *Prog. Theor. Phys.* **64** 1138
- [16] Roux J C, Rossi A, Bachelart S and Vidal C 1980 *Phys. Rev. Lett.* A **77** 391
- [17] Roux J C and Swinney H L 1981 *Nonlinear Phenomena in Chemical Dynamics* ed C Vidal and A Pacault (New York: Springer) p 38
- [18] Hudson J L and Mankin J C 1981 *J. Chem. Phys.* **74** 6171
- [19] Vidal C, Roux J C, Bachelart S and Rossi A 1980 *Nonlinear Dynamics* ed R H G Helleman (New York: Academy of Sciences) p 377
- [20] Lorenz E N 1963 *J. Atmos. Sci.* **20** 130
- [21] Olsen L F and Degn H 1977 *Nature* **267** 177

-23-

- [22] Roux J C, Boissonade J and De Kepper P 1983 *Phys. Lett.* A **168**
 Menzinger M, Boissonade J, Boukalouch M, De Kepper P, Roux J C and Saadaoui H 1986 *J. Phys. C.: Solid State Phys.* **90** 313
- [23] Guckenheimer J and Holmes P 1983 *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields* (New York: Springer) pp 270, 306
- [24] Metropolis N, Stein M L and Stein P R 1973 *J. Combin. Theor.* A **15** 25
- [25] Eckmann J P 1981 *Rev. Mod. Phys.* **53** 643
- [26] Feigenbaum M J 1978 *J. Stat. Phys.* **19** 25
 Feigenbaum M J 1979 *J. Stat. Phys.* **21** 669
- [27] Schmitz R A, Graziani K R and Hudson J L 1977 *J. Chem. Phys.* **67** 3040
 Hudson J L, Hart M and Marinko J 1979 *J. Chem. Phys.* **71** 1601
 Maselko J and Swinney H L 1986 *J. Chem. Phys.* **85** 6430
 Maselko J and Swinney H L 1987 *Phys. Lett.* A **119** 403
 Richetti P, Roux J C, Argoul F and Arneodo A 1987 *J. Chem. Phys.* **86** 3339
 Ringland J, Issa N and Schell M 1990 *Phys. Rev.* A **41** 4223
 Koper M T M and Gaspard P 1991 *J. Phys.: Condens. Matter* **95** 4945
 Koper M T M and Gaspard P 1992 *J. Chem. Phys.* **96** 7797
 Koper M T M, Gaspard P and Sluyters J H 1992 *J. Chem. Phys.* **97** 8250
 Petrov V, Scott S K and Showalter K 1992 *J. Chem. Phys.* **97** 6191
 Koper M T M 1995 *Physica D* **80** 72
 Goryachev A, Strizhak P and Kapral R 1997 *J. Chem. Phys.* **107** 2881
- [28] Pomeau Y, Roux J C, Rossi A, Bachelart S and Vidal S 1981 *J. Physique Lett.* **42** L-271–L-273
- [29] Pomeau Y and Manneville P 1980 *Commun. Math. Phys.* **74** 189
- [30] Kupka I 1963 *Contributions to Differential Equations* **2** 457
 Kupka I 1964 *Contributions to Differential Equations* **3** 411
 Smale S 1963 *Ann. Scuola Norm. Sup. Pisa* (3) **17** 97
 Newhouse S, Ruelle D and Takens T 1978 *Commun. Math. Phys.* **64** 35

- [31] Tam W Y, Horsthemke W, Noszticzius Z and Swinney H L 1988 *J. Chem. Phys.* **88** 3395
- [32] Manneville P 1990 *Dissipative Structures and Weak Turbulence* (New York: Academic)
- [33] Fife P 1984 *Non-Equilibrium Dynamics in Chemical Systems* ed C Vidal and A Pacault (Berlin: Springer) p 76
Mikhailov A S 1994 *Foundations of Synergetics I. Distributed Active Systems* (Berlin: Springer)
Zykov V S 1988 *Simulation of Wave Processes in Excitable Media* (Manchester: Manchester University Press)
- [34] FitzHugh R 1961 *Biophys. J.* **1** 445
Nagumo J, Arimoto S and Yoshikawa 1962 *Proc. IRE* **50** 2061
- [35] Malevanets A and Kapral R 1997 *Phys. Rev. E* **55** 5657
- [36] Winfree A T 1972 *Science* **175** 634

-24-

- [37] Ertl G 1990 *Adv. Catal.* **40** 231
- [38] Goldbeter A 1996 *Biochemical Oscillations and Cellular Rhythms* (Cambridge: Cambridge University Press)
- [39] Winfree A T 1987 *When Time Breaks Down: The Three Dimensional Dynamics of Electrochemical Waves and Cardiac Arrhythmias* (Princeton, NJ: Princeton University Press)
- [40] 1998 *Chaos Focus Issue: Fibrillation in Normal Ventricular Myocardium* **8** (1)
- [41] Pertsov A M and Jalife J 1995 *Cardiac Electrophysiology: From Cell to Bedside* ed D P Zipes and J Jalife (Philadelphia: Saunders) p 403
- [42] Winfree A T 1991 *Chaos* **1** 303
- [43] Winfree A T 1994 *Chemical Waves and Patterns* ed R Kapral and K Showalter (Dordrecht: Kluwer) p 3
Plesser T, Müller S C and Hess B 1990 *J. Chem. Phys.* **94** 7501
Skinner G S and Swinney H L 1991 *Physica D* **48** 1
Mikhailov A S and Zykov V S 1991 *Physica D* **52** 379
Barkley D 1994 *Phys. Rev. Lett.* **72** 164
- [44] Bär M and Eiswirth M 1993 *Phys. Rev. E* **48** R1635
- [45] Newell A C 1989 *Complex Systems* ed D Stein (New York: Addison-Wesley)
Newell A C, Passot T and Lega J 1993 *Ann. Rev. Fluid Mech.* **25** 399
Kuramoto Y 1984 *Chemical Oscillations, Waves and Turbulence* (Berlin: Springer)
- [46] Sorensen P G and Hynne F 1989 *J. Phys.: Condens. Matter* **93** 5467
- [47] Hagen P S 1982 *SIAM J. Appl. Math.* **42** 672
- [48] Mermin N D 1979 *Rev. Mod. Phys.* **51** 591
- [49] Bohr T, Pedersen A W, Jensen M H and Rand D A 1989 *New Trends in Nonlinear Dynamics and Pattern Forming Processes* ed P Coulet and P Heurre (New York: Plenum)
Coulet P, Gil L and Lega J 1989 *Phys. Rev. Lett.* **62** 161
Huber G, Alstrom P and Bohr T 1992 *Phys. Rev. Lett.* **69** 2380
- [50] Ouyang Q and Flesselles J M 1996 *Nature* **379** 143
- [51] Pikovsky A, Zaks M, Rosenblum M, Osipov G and Kurths J 1997 *Chaos* **7** 680

- [52] Klevecz R, Pilliod J and Bolen J 1991 *Chronobiol. Int.* **8** 6
- [53] Brunnet L, Chaté H and Manneville P 1994 *Physica D* **78** 141
- [54] Goryachev A and Kapral R 1996 *Phys. Rev. Lett.* **76** 1619
Goryachev A and Kapral R 1996 *Phys. Rev. E* **54** 5469
- [55] Goryachev A, Chaté H and Kapral R 1998 *Phys. Rev. Lett.* **80** 873
- [56] Yoneyama M, Fujii A and Maeda S 1995 *J. Am. Chem. Soc.* **117** 8188

-25-

- [57] Turing A 1952 *Phil. Trans. R. Soc. B* **237** 37
- [58] Castets V, Dulos E, Boissonade J and De Kepper P 1990 *Phys. Rev. Lett.* **64** 2953
De Kepper P, Castets V, Dulos E and Boissonade J 1991 *Physica D* **49** 161
- [59] Ouyang Q and Swinney H L 1991 *Nature* **352** 610
Ouyang Q and Swinney H L 1991 *Chaos* **1** 411
- [60] Lengyel I, Kadar S and Epstein I 1993 *Phys. Rev. Lett.* **69** 6315
- [61] Murray J D 1989 *Mathematical Biology* (New York: Springer)
- [62] Lengyel I and Epstein I 1991 *Science* **251** 650
Lengyel I and Epstein I 1992 *Proc. Natl Acad. Sci. USA* **89** 3977
Hunding A and Sorensen P G 1988 *J. Math. Biol.* **26** 27
Pearson J E 1992 *Physica A* **188** 178
Pearson J E and Bruno W J 1992 *Chaos* **2** 513
- [63] Borckmans P, De Wit A and Dewel G 1992 *Physica A* **188** 137
- [64] Boissonade J and De Kepper P J 1980 *J. Phys. Chem.* **84** 501
- [65] 1964 *Non-Steady Flame Propagation* ed G H Markstein (New York: Macmillan)
- [66] Sivashinsky G I 1983 *Ann. Rev. Fluid Mech.* **15** 179 and references therein
- [67] Horváth D and Showalter K 1995 *J. Chem. Phys.* **102** 2471
- [68] Zhang Z and Falle S A E G 1994 *Phys. R. Soc. S A* **446** 1
Milton R A and Scott S K 1995 *J. Chem. Phys.* **102** 5271
Horváth D, Petrov V, Scott S K and Showalter K 1993 *J. Chem. Phys.* **98** 6332
Malevanets A, Careta A and Kapral R 1995 *Phys. Rev. E* **52** 4724
- [69] Kuramoto Y and Tsuzuki T 1976 *Prog. Theor. Phys.* **55** 356
Sivashinsky G I 1977 *Combust. Sci. Tech.* **15** 137
- [70] Saul A and Showalter K 1985 *Oscillations and Travelling Waves in Chemical Systems* ed R J Field and M Burger (New York: Wiley)
- [71] Pojman J A, Gunn G, Patterson C, Owens J and Simmons C 1988 *J. Phys. Chem. B* **102** 3927
- [72] Lee K J, McCormack W D, Ouyang Q and Swinney H L 1993 *Science* **261** 192
Lee K J and Swinney H L 1995 *Phys. Rev. E* **51** 1899
Lee K J, McCormack W D, Pearson J and Swinney H L 1994 *Nature* **214** 215
Hagberg A and Meron E 1994 *Phys. Rev. Lett.* **72** 2494
Elphik C, Hagberg A and Meron E 1995 *Phys. Rev. E* **51** 3052
Petrich D M and Goldstein R E 1994 *Phys. Rev. Lett.* **72** 1120

- [73] Whitten T A and Sander L M 1981 *Phys. Rev. Lett.* **47** 1400
Barabási A L and Stanley H E 1995 *Fractal Concepts in Surface Growth* (Cambridge: Cambridge University Press)
Avnir D (ed) 1989 *The Fractal Approach to Heterogeneous Chemistry: Surfaces, Colloids, Polymers* (New York: Wiley)
- [74] Meyer Y 1990 *Ondelette et Opérateur* vols I and II (Paris: Hermann)
Meyer Y and Coifmann R R 1991 *Ondelette et Opérateur* vol III (Paris: Hermann)
Daubechies I 1992 *Wavelets (CBMS-NFS Series in Appl. Math.)* (Philadelphia: SIAM)
Chui C K 1992 *An Introduction to Wavelets* (New York: Academic)
- [75] Arneodo A, Argoul F, Muzy J F and Tabard M 1992 *Phys. Lett. A* **171** 31
- [76] Ott E, Grebogi C and Yorke J A 1990 *Phys. Rev. Lett.* **64** 1196
Peng B, Petrov V and Showalter K 1991 *J. Phys. Chem.* **95** 4957
Petrov V, Scott S K and Showalter S 1992 *J. Chem. Phys.* **96** 7506
-