STM experiment and atomistic modelling hand in hand: individual molecules on semiconductor surfaces

G.A.D. Briggs\textsuperscript{a,*}, A.J. Fisher\textsuperscript{b,1}

\textsuperscript{a}Department of Materials, University of Oxford, Parks Road, Oxford OX1 3PH, UK
\textsuperscript{b}Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK
Contents

1. Introduction – atoms in materials 3
2. Scanning tunnelling microscopy 6
   2.1. Instrumentation – what do you need? 6
   2.2. Images – what do you see? 11
3. Modelling electronic structure of surfaces 17
   3.1. Quantum chemical methods 17
   3.2. Density functional theory 20
   3.3. Approximations to density functional theory 27
   3.4. Other methods 28
   3.5. The surface geometry – clusters, slabs and embedding 30
4. Modelling electron tunnelling 31
   4.1. Perturbation theory 32
   4.2. Scattering theory 34
   4.3. Educated intuition 38
   4.4. Tip–sample interactions 38
5. The Si(001) Surface 39
   5.1. Clean Si(001) surfaces 39
   5.2. C₂H₄ and C₂H₂ on Si(001) 45
   5.3. Hydrogen on Si(001) 57
   5.4. Disilane on Si(001) 61
6. The GaAs(001) surface 68
   6.1. The clean surface 68
   6.2. Adsorption of hydrocarbons 71
   6.3. Adsorption of metallorganic compounds 72
7. Conclusions 74

Acknowledgements 76
References 76
STM experiment and atomistic modelling hand in hand: individual molecules on semiconductor surfaces

G.A.D. Briggs\textsuperscript{a,*}, A.J. Fisher\textsuperscript{b,1}

\textsuperscript{a}Department of Materials, University of Oxford, Parks Road, Oxford OX1 3PH, UK

\textsuperscript{b}Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK

Manuscript received in final form 7 July 1998

Abstract

When the scanning tunnelling microscope was invented, the world was amazed at the atomic resolution images of surfaces which could be obtained. It soon became apparent that it was one thing to obtain an image, and quite another to understand the structure that was seen. Happily the developments in real space experimental techniques for studying surfaces have been accompanied by developments in real space theoretical techniques for modelling electronic structure and bonding at surfaces. The aim of this review is to describe how and why STM experiments and atomistic modelling should be combined and what they can then be expected to tell us. A summary is given of the experimental methods for theorists and vice versa, and their relationship is illustrated using a number of case studies where they have been used together. To give the review a coherent focus the examples are confined to studies of adsorbed molecules on semiconductor surfaces, in particular Si(001) and GaAs(001). Questions thus addressed include: How are experimental images and structural modelling linked by tunnelling theory? What can they tell us together that we could not learn from experiment or theory alone? What can we learn about atomic positions and bonding at semiconductor surfaces with and without adsorbed molecules? How many different ways are there to relate images to calculations? © 1999 Published by Elsevier Science B.V.

1. Introduction – atoms in materials

The world first became convinced of the reality of molecules, not because they could be seen, but because of the consistency with which Avogadro’s number could be determined. In 1905, the year in
which he published his seminal paper on special relativity (a paper with 11 sections, one
acknowledgement and no references), Albert Einstein developed no less than five different methods
for determining Avogadro’s number [1]. The overdetermination of Avogadro’s number, from
phenomena as unrelated as radioactivity, Brownian motion, and the blue colour of the sky, clinched
the argument for believing in molecules long before they could be seen with a microscope.

There are now at least three different techniques available by which not only molecules but even
individual atoms can be directly imaged. For two of them, electron microscopy and scanning tunnelling
microscopy, a Nobel prize has been awarded. The third, field ion microscopy, has been less widely
taken up, but provides information not available from the other two. All three techniques offer the
capability to image solids in real space. Historically the arrangement of atoms in materials has been
studied by diffraction techniques. Classically this has been achieved using X-rays; an early Nobel prize
was awarded for the discovery that by diffracting X-rays from planes of atoms in a solid, its crystalline
structure and parameters could be measured. Subsequently it was found that particles such as neutrons
and electrons could also be diffracted by solids. Imaging of defects such as dislocations in metals arose
out of microdiffraction experiments; indeed, diffraction plays a key role in the contrast in electron
microscopy. As instruments improved, it eventually became possible to image individual columns of
atoms in materials, and hence to image defects (i.e. deviations from perfect crystalline regularity) with
atomic resolution.

Electron microscopy uses lenses. Field ion microscopy was developed as a lenseless technique for
imaging single atoms, subject neither to diffraction in the material being imaged nor to diffraction or
aberration limitations of lenses [2]. A very sharp tip is made out of the material to be studied, and this is
placed in an evacuated chamber into which a low pressure of helium or neon gas is introduced. By
raising the potential of the tip, tunnelling from its sharp point is induced. This ionizes rare gas atoms
attracted to the tip by polarization in the strong field gradient, which are then strongly accelerated away
from the tip by the field itself. The tip is made of atoms, and if the tip is sharp enough, the process is
significantly enhanced by individually protruding surface atoms. The rare-gas ions are used to produce
secondary electrons, which then strike a fluorescent screen placed some distance away from the sample;
a magnified image of the individual atomic protrusions can then be seen. If now the helium is removed
and a large enough electrical or thermal impulse is imparted to the sample (for example, by means of a
nanosecond laser pulse), the atoms of the material can themselves be removed. The trajectory followed
in an inhomogeneous electrostatic field is independent of the mass to charge ratio of the particles, but
their velocity is not. Therefore, by measuring the time of flight of the atoms, they can be individually
weighed and their chemical identity established. Thus the field ion microscope can be used as an atom
probe to perform chemical analysis on one atom at a time.

Scanning tunnelling microscopy (STM) is also lenseless. Like field ion microscopy it uses a sharp tip
and relies on quantum mechanical tunnelling of electrons, but now the tip is essentially part of the
instrument, and the sample can be a flat surface of any conducting or semiconducting material. The
quantum mechanical tunnelling takes place between the tip and the sample. The voltages employed are
generally smaller than the work functions of either material, and the tunnelling probability changes by
typically an order of magnitude for every Ångström change in the spacing between the tip and the
sample. Individual atomic protrusions on the tip are usually random, and with luck one atom may
protrude sufficiently to dominate the tunnelling geometry. In this way, perhaps to some extent
fortuitously, it is possible to obtain atomic resolution by scanning the tip over the surface and using it as
a near field probe of the surface atomic arrangement. The contrast, which will be a key issue in this
review, is due to the combined effects of topography and electronic structure: indeed, on this scale the distinction between them begins to become semantic. Although both electron microscopy and field ion microscopy can be used to study the position and motion of atoms on surfaces, scanning tunnelling microscopy provides a unique method for observing bonding on surfaces, both of individual atoms and of molecules.

Just as these experimental techniques brought the atomic picture to life in materials science, so modern developments in computational physics have given a new significance to our understanding of the underlying quantum mechanics of the particles in a material. We have believed for nearly seventy years that the Schrödinger equation and its relativistic generalizations give us, in principle, all that we need to know about structure and bonding; it was in 1929 that Dirac made his famous claim that ‘the underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble’ [3]. By the 1950s our understanding had progressed to the point where quantitative calculations of the electron states in bulk materials could be made, and compared with the results of experiments such as those using the de Haas–van Alphen effect. Thanks to further advances in computing power and in theoretical methods, it was possible by the early 1980s to make successful predictions of the relative stability of different bulk phases of silicon [4,5] that proceeded from first principles in the sense that only the laws of quantum mechanics, some general approximations, and the atomic numbers of the atoms concerned were needed as input to the calculations.

By this time, a separate series of theoretical developments was under way that focussed on the local electronic properties of materials which determine chemical bonding, rather than on the extended states necessary to describe the very low-energy excitations of perfect crystals. Beginning in the 1960s with electronic structure calculations on point defects, this approach gradually spread to encompass the theory of electronic states in alloys, interfaces, surfaces, grain boundaries, and dislocations. This process received a boost in 1985 with the publication of the seminal paper of Car and Parrinello [6], showing how state-of-the-art electronic structure methods could be combined with molecular dynamics and used to study even highly disordered systems such as liquids.

More recently, it has been realized that many aspects of electronic structure are intrinsically local, and that the electronic properties of a particular atom can be computed with an effort that is essentially independent of the size of the system being considered. The resulting computational effort for the whole system is $O(N)$, where $N$ is the number of atoms; this should be contrasted with the $O(N^3)$ scaling that is obtained when the calculation is performed by a traditional diagonalization of the Hamiltonian for the entire system.

We have arrived at a point where we have two very powerful tools at our disposal. One is experimental – the STM – and is uniquely sensitive to the local atomic and electronic structure of the surface. The other – electronic structure calculations – allows us to make predictions of these same quantities with increasing precision, and in systems of ever growing complexity. It is the purpose of this review to illustrate some of the benefits that can accrue when these techniques are combined. One of the fields in which this is proving immensely fruitful is in the study of adsorbates on surfaces of semiconductors, and to give a limited objective to the review we shall use examples involving small molecules on (001) surfaces of silicon and gallium arsenide. We hope we shall be excused the bias towards our own work and interests which has resulted from this approach. A complementary review focussing on the imaging of adsorbates at metal surfaces has recently appeared [7]. We start by
reviewing separately the status of STM experiments and of electronic structure calculations. We go on to describe how the two may be combined to calculate STM tunnelling currents, and then illustrate this synthesis with examples from the physics and chemistry of two technologically important semiconductor surfaces: the (001) surfaces of silicon and gallium arsenide.

2. Scanning tunnelling microscopy

2.1. Instrumentation – what do you need?

The principle of scanning tunnelling microscopy (STM) is now well established. A sharp tip is held a nanometre or so above the surface to be studied (see Fig. 1). A bias voltage is applied between them, and the resulting quantum mechanical tunnelling current is measured. In the usual mode of operation, the current is required to be equal to some chosen value (generally measured in nanoamps). A feedback loop is employed to achieve this. The tip (or equivalently the sample) is supported on a combination of piezoelectric actuators which permit movement in three axes. The one that controls the movement normal to the sample surface (traditionally designated the z direction) is connected to the output of the feedback amplifier; any departure of the tunnelling current from the set value causes a movement that changes the gap in such a way as to restore the set current. Movement of the tip in the other two directions, x and y, is controlled so as to scan an area of the surface in a raster (from the German Raster meaning scanning pattern). For standard topographic imaging, the voltage applied to the z actuator is used to modulate the brightness of the image, with bright meaning the actuator is retracted and dark.

Fig. 1. A schematic diagram of an STM operating in a typical ultra-high vacuum system, in which a number of other surface science experimental tools are also available (courtesy of M. Iwatsuki, JEOL).
Fig. 2. The tunnelling gap between a gold sample and a gold tip, in a micromachined STM specially designed to operate while being observed in a transmission electron microscope. The resolution was adequate to resolve (200) planes in the sample. The tip geometry was neither ideal nor perfectly stable, and abrupt changes were seen sometimes as often as every 5 s [11].

meaning the actuator is extended. In analysing the images it should be remembered that this is what is actually displayed; everything else is interpretation.

2.1.1. The tunnelling gap

One of the early surprises in the development of STM was that atomic resolution could be achieved. Before any experimental results had been obtained, estimates had been made of the resolution that could be expected using a spherical tip of radius 100 nm, which seemed a reasonable specification. The answer was about 4 nm. Imagine, therefore, the delight when it was found that on a surface of gold resolution better than 1 nm was obtained [8], and subsequently that individual atoms approximately 0.9 nm apart could be resolved on a Si(111)-(7 x 7) reconstructed surface [9]. The imperfections of nature work in favour of the experiments. No tips ever approximate to smooth mesoscopic spheres. Although tips have been demonstrated which consist of a single atom placed at the end of a tungsten single crystal [10], almost all tips used in practice are atomically rough. A picture of the tunnelling gap between a gold sample and a tip (in this case also made of gold) is shown in Fig. 2 [11]. This picture was taken in a transmission electron microscope (TEM) using a specially designed micromachined tunnelling microscope. The gap width was found to change abruptly every few seconds, perhaps because of molecules adsorbed from the TEM column which were not visible in the TEM contrast. The question of the stability of a tip is an important one in STM, especially when considering spectroscopy (see Section 2.2.2 below), but the key point here is that the tip is certainly not a smooth sphere. This is what makes atomic resolution possible. The tunnelling current density varies approximately exponentially with separation, typically increasing by about a factor of 10 for every 0.1 nm reduction in the separation. Thus it requires the last atom on the tip to be only slightly closer to the sample than its fellows for the current through that atom to dominate. This is the reason why atomic resolution can be obtained, if not easily, then at least routinely.
The successful development of STM has been an astonishing series of technological achievements. The tip and sample must be brought together over a distance of a millimetre or so, and then held rigid relative to one another with picometre stability. This has been achieved in a variety of ingenious ways. For the present purpose, however, discussion of experimental details of the STM is confined first to achieving and maintaining the appropriate sample surface, and second to the different ways in which data can be obtained and displayed.

2.1.2. Ultra-high vacuum

All the experimental studies described in this review were conducted in ultra-high vacuum (UHV). If you are going to look at a surface with atomic resolution, then you had better make sure that the atoms which you are seeing are the ones that you are interested in. Having prepared the surface, you do not want it to become cluttered with unspecified molecules landing on it from the atmosphere. Therefore the experiments should be carried out in the best vacuum available.

Theoretical physics has shown us that even a perfect vacuum is seething with activity, in the form of a zoo of quasiparticles. Experimental work with vacuum does not demand such virtual concepts; a real vacuum is seething with real particles. This may be best considered on a logarithmic scale of pressure. The standard atmospheric pressure for international air navigation is 1.0132 × 10^5 Pa. Using the simple kinetic theory expression for the number of collisions on the surface of a container per unit area per unit time, \( n = \frac{1}{4} N \bar{c} \), at atmospheric pressure each atomic site experiences an impact from a gas molecule approximately every nanosecond. If the pressure is reduced by nine orders of magnitude, to 10^{-4} Pa, then each atomic site experiences an impact about once every second. There is some flexibility over the time taken to scan an STM image, but most high quality images in the constant current mode require some tens of seconds to scan. It would be intolerable to find that the surface had been covered with molecules when the scan had scarcely begun.

Detailed studies of atomically clean surfaces require the use of UHV, in which pressures of order 10^{-8} Pa or lower are used. These pressures make it possible to work for several hours before the majority of sites experience arriving molecules. Even then they may remain clean. Many species, such as \( \text{N}_2 \) and \( \text{H}_2 \) have low sticking coefficients on most of the surfaces included in this review. The atmosphere in UHV tends to be fairly reducing. Reactive surfaces such as clean silicon are most vulnerable to species such as oxygen and water, and the partial pressure of these will generally be somewhat lower than the total pressure. Less reactive surfaces, such as many oxides, can be even more tolerant.

Much of the expertise of working with UHV is learnt on the job. However, it may be useful for readers of this review who come from different backgrounds (for example theoretical!) to have some appreciation of what is involved in achieving and maintaining such low pressures. Alternative units, common in the trade, are mbar and Torr: 1 mbar = 1 hPa = 100 Pa. 1 Torr is the pressure measured by 1 mm Hg under standard gravity. Thus 1 Torr = 1.33322 mbar = 133.322 Pa. An exposure to a gas at a pressure of 10^{-6} Torr for 1 s is defined as 1 Langmuir; this would give a coverage of approximately one monolayer (1 ML) if the sticking coefficient were unity. The inside of the UHV apparatus must be made of materials which have a vapour pressure lower than the pressure to be achieved. Stainless steel fulfills this condition, and the sight of welded stainless steel chambers covered with circular ports has become familiar. High-vapour-pressure metals and alloys such as zinc and solder must be avoided. Almost all polymers are excluded, although viton can be used for seals which must be
closed after the vacuum has been established, and kepton can be used for flexible electrical insulation. The seals which must be made as the apparatus is bolted together are made of soft copper between knife edges machined on the stainless steel. Organic products on human skin have a quite unacceptable vapour pressure, and so all interior parts once cleaned must be handled with gloves. Even with all these precautions, there will still be adsorbed layers of atmospheric gases on the inner surfaces of the equipment, and especially water molecules. At room temperature this would desorb at a rate fast enough to limit the vacuum severely (typically $10^{-6}$ Pa), but slow enough that the vacuum would not be ready for the experiment to start within the lifetime of a Ph.D. To address this, the equipment is baked for a day or two at a temperature of 150°C or higher to desorb and pump away gas molecules from the inner surfaces. It can be imagined that work with UHV is time consuming. Even a simple internal repair or adjustment can easily take a week by the time the system has been opened and closed and the vacuum restored.

The study of individual molecules by STM requires surfaces which are atomically clean and flat. Metal surfaces must be prepared by a patient series of annealing and bombarding with inert gas ions (usually Ar$^+$). For silicon samples the preparation is slightly different for each surface orientation, but basically it consists of an acid etch which inhibits oxide formation by creating a hydrogen-terminated surface, followed by outgasing in vacuum at a temperature of a few hundred degrees Centigrade for several hours, and then heating to a temperature of about 1150°C or so to cause sublimation of all oxide layers. The temperature is then lowered as fast as possible through the range in which silicon is most reactive to about 800°C, from which it is cooled slowly to avoid quenching in defects. The structure of silicon surfaces can be drastically affected by minute quantities of metals such as nickel, and so all contact with any materials which might contain these metals must be avoided. Compound semiconductor surfaces are more difficult to prepare. For example, if gallium arsenide is heated, the arsenic sublimes leaving the gallium to form into droplets. After growing a surface of interest on a Group III-V compound substrate, it is possible to apply a protective cap of arsenic which can later be removed by heating to give a clean surface corresponding to the growth surface. However, the preferred method is to have a molecular beam epitaxy (MBE) facility within the same UHV system as the STM, so that the growth surface can be cooled under controlled arsenic pressure, while monitoring with reflection high energy electron diffraction (RHEED) to ensure that the surface maintains the reconstruction and composition corresponding to growth conditions.

The tip also must be clean. The vast majority of STM work is carried out with tungsten tips. These invariably have a layer of oxide on them when they are introduced to the UHV system. Techniques for treating the tip vary from doing almost nothing through modest outgasing and annealing treatments all the way to very thorough cleaning by bombarding it with a beam of electrons from a hot filament, using the tip itself as the anode in the circuit. Although the tip can in principle be characterized using a specially installed field ion microscope (FIM) [12], this is seldom done in practice. One of the key skills necessary for successful operation of an STM is to know whether the tip is giving results as good as can be reasonably expected from a new surface, and what to do about it if it is not. One commercial microscope is equipped with a facility for applying a short sharp voltage pulse to the tip while it is scanning a surface, to improve the tip quality. This is rather like kicking a television set when it is malfunctioning, but somewhat more successful. The uncertainty in the exact geometry, and even chemical composition, of the tip remains a major challenge in relating image simulations to experimental results.
2.1.3. Introducing molecules

In this review several studies of small molecules on surfaces are summarized. The molecules must be introduced to the UHV chamber in a controlled way. Some gases, such as ethene, can be admitted through a needle valve, or leak valve, which is designed to introduce gases at a rate that allows the partial pressure to be readily controlled at levels comparable with the total UHV pressure. Thus, for example, an exposure of 1 L could be achieved by exposing the sample to a pressure of $10^{-4}$ Pa for a little over two minutes. The slight vagueness is deliberate. There are several reasons why it is difficult to be confident of absolute exposure. First, the pressure measured by the ion gauge in the UHV chamber is difficult to calibrate, and anyway it may be somewhat different from the pressure at the surface of the sample. Second, some gases, such as trimethyl gallium and disilane, take a long time to pump away, so that the integrated pressure over time may be difficult to estimate. Third, some gases need to be cracked in the chamber in order to react with the sample. For example molecular hydrogen is almost inert on silicon; it can be made to react as atomic hydrogen by passing over a filament at about 1800 K with line of sight to the sample. This cracking process is notoriously inefficient, and since is not possible to measure the partial pressure of atomic hydrogen, the atomic hydrogen exposure can only be guessed at. Calibration can be done by titration on a surface which has effectively unity sticking coefficient [13], but this is somewhat indirect. Although absolute exposure is difficult to estimate, local absolute coverage can usually be readily determined from direct examination of STM images. Some gases require special handling. Organometallic gases may be toxic, or pyrophoric, or both. Some of them condense at room temperature, and the pipework must be warmed. Arsene has a lethal toxicity even in trace amounts, and the handling of it requires very special precautions. Germane is both toxic and explosive. Disilane is among the most explosive gases known, and the fact that it is not toxic is of little comfort. Clearly these gases must be handled with the utmost caution. Fortunately, for studies of their adsorption and behaviour in STM, only a few molecules (i.e. a small fraction of Avagadro’s number) are required.

2.1.4. Variable temperature

For studies of molecules on surfaces there is great advantage in being able to operate at variable temperature, especially elevated temperatures. The early STM studies were mostly performed at room temperature. Even then the question of thermal stability is a challenging one. There is strong motivation to make the scanning stage of an STM as small and rigid as possible simply to make the mechanical resonant frequencies as high as possible and thus minimize the effect of vibrations. But even with a compact design the total mechanical circuit between the tip and sample is many orders of magnitude greater than the tip-sample separation. Thus even modest thermal expansion can lead to noticeable drift. Lowering the temperature leads to a reduction in the coefficients of thermal expansion, and cryogenic microscopes have been used for fascinating studies of physisorbed inert gas molecules [14]. The development of UHV STM’s which can operate at elevated temperatures up to 1500 K represents a very remarkable achievement [15]. It was made possible by careful attention to two aspects. First, the materials used must be either capable of operating at such high temperatures without degradation, or else thermally protected so that, for example, the piezoelectric ceramic elements do not become depoled. Second, the sample stage is supported solely on planes passing through the area to be imaged. Thus $z$ location is achieved by a surface on the sample holder coplanar with the sample surface resting against supports, and $x$ and $y$ location are achieved by planes containing diameters through the centre of the sample. In this way thermal drift is compensated to first order. No microscope has yet been made.
with zero drift, and studying samples while simultaneously varying the temperature remains a challenge; nevertheless it is now possible routinely to study adsorbed molecules at any temperature up to the point at which they desorb.

The benefits of being able to operate at variable temperature are very great. The time window available in the STM is limited. Since a typical high resolution scan requires some tens of seconds, it is generally difficult to follow events that occur very much faster than that. It is sometimes possible to deduce rates of adsorbate hopping which are of the order of the time taken for a line scan, i.e. some tens of milliseconds, and special techniques have been developed for tracking adsorbate motion on a surface [16]. But it is difficult to extend the range of possible observation windows over more than a few orders of magnitude in time at most. But rates of events that are subject to thermal activation can be altered over many orders of magnitude by changes in temperature. The activation energies for many of the kinds of process which may be of interest are of the order of 1 eV, which gives a change in rate of nearly nine orders of magnitude in doubling the temperature from 300 to 600 K; lower activation energies give even more dramatic increases. Thus for observation of atomic and molecular movement and reactions on surfaces, and for real-time observation of growth processes, there is enormous gain in having temperature as a variable under independent control.

2.2. Images – what do you see?

2.2.1. Topographical imaging

The Si(001) surface provides a paradigm for many of the principles expounded in this review for relating STM observations to theoretical modelling. The (001) surface of silicon is the most widely used surface in the whole of semiconductor technology, yet it has been less studied by STM than the Si(111) surface. The Si(111) surface has a range of reconstructions, which make for photogenic STM pictures. In particular, the Si(111)-(7 x 7) reconstruction has a rather open structure, which means that it is possible to get achieve atomic resolution at less demanding resolution than almost any other regular crystalline surface. It was an inspired choice on the part of the inventors of the STM to select the Si(111)-(7 x 7) structure as a demonstrator of their new instrument, and the subsequent pictures of this surface have confirmed beyond doubt the generally accepted model of this surface. Its intrinsic elegance and the frequency with which it has been reproduced have made it the Mona Lisa of the STM world.

The Si(001) surface has one basic reconstruction, the (2 x 1), in which surface atoms pair up into dimers. An STM image of a clean Si(001) surface is shown in Fig. 3. The dimers are formed into straight rows, though there are plenty of defects. There is a step towards the left of the picture. The step is a quarter of a unit cell high, 0.136 nm; the total height range represented by the gray scale in the picture is less than 0.5 nm. The direction of the dimers changes above and below the step; this particular step runs parallel to the dimer rows on the upper terrace, which tends to lead to a fairly straight step edge. Some of the dimers appear symmetrical in the picture; these can just be resolved as individual dimers. In other places the dimer rows have a zigzag appearance, as though the dimers have alternating tilt along a row (cf. Fig. 12 in Section 5.1.1). The question of dimer tilt will be one of the first to be addressed in modelling the clean surface.

The appearance of dimers with their tilt pinned provides an excellent demonstration of the conceptual impossibility of separating topography from electronic structure in the STM. The ‘up’ end of the dimer has a greater density of filled states below the Fermi level than the ‘down’ end, while the
Fig. 3. A bare silicon surface with (0 0 1) orientation. The surface silicon atoms pair up into dimers which form the rows apparent in this picture. At a step the height changes by a quarter of a unit cell, causing a change in the direction of the dimer rows. Although the dimers prefer to be tilted, at room temperature most of them flip from one tilt to the other faster than the microscope can scan, so that on the average they appear symmetrical. But some dimers are pinned with a tilt in one sense, and because adjacent dimers prefer to adopt alternate tilts the rows then have a zigzag appearance.

‘down’ end has a higher density of empty states above the Fermi level than the ‘up’ end. If the sample is biased negative in the STM, so that electrons tunnel out of the filled sample states into the tip to provide a filled-states image, then the electronic effects enhance the geometrical effects, exaggerating the appearance of the tilt in the picture. But if the sample is biased positive, so that electrons tunnel from the tip to the empty states of the sample to give an empty-states image, then the electronic effects diminish and partially compensate for the geometric tilt, so that the appearance of the tilt is considerably flattened. An example of filled and empty-states imaging of molecules on Si(0 0 1) will be given in Section 5.2.1.

2.2.2. Scanning tunnelling spectroscopy

The invention of the STM arose from experiments to measure local electronic properties [8]. If the separation between the tip and a metallic sample is modulated at a constant low bias voltage, then the derivative of the logarithm of the tunnelling current with separation can be directly related to the square root of an effective mean work function via a simple relationship. If one assumes that a simple one-dimensional tunnelling model applies, and that the modulation of the piezoelectric drive is applied directly to the tunnel barrier (we shall see in Section 4 that both assumptions are in fact very questionable), one finds \( d \ln I / ds \propto \sqrt{\phi} \) (where \( s \) is the displacement and \( \phi \) is the work function). This relationship was used in early studies by STM as a check on the condition of the surface being studied.

As the theory of STM became better developed [17,18], attention was shifted to measurements of density of states. If the tip is held over a given point on the surface of the sample, and the voltage is swept to make a series of measurements of current as a function of voltage, then an \( I-V \) curve can be
Fig. 4. Tunnelling spectra acquired over different sites on a clean Si(0 0 1)-(2 x 1) surface on which some dimers appeared symmetrical and others were pinned in an asymmetrical zigzag configuration: (a) — over a symmetrical dimer; (b) — between two symmetrical dimers along a row; (c) — over the ‘up’ end of a pinned dimer; (d) — over the ‘down’ end of a pinned dimer. By definition all the curves have the same value for the current at -2 V. The left-hand plot shows \( I \) versus \( V \), the right-hand plot shows \( \frac{dI}{dV} \) versus \( V \) [19].

measured corresponding to that location. Assuming a relatively smooth distribution of states in the tip, then the derivative of this curve, specifically \( (V/I)\frac{dI}{dV} \), or equivalently \( d \ln I/d \ln V \), gives a direct indication of the density of states of the sample at that location. The underlying theory will be expounded in Section 4. A number of specific experimental techniques have been developed for measuring spectra. They have in common that the feedback loop is switched off for a time short enough that negligible thermal drift occurs but long enough to measure the current at a series of values of voltage.

A series of \( I-V \) curves and their derivatives for a clean Si(0 0 1) surface are shown in Fig. 4 [19]. By definition \( I-V \) curves from different places will show the same current at the bias voltage used to determine the topography, regardless of the local electronic properties, since this is the starting point for each curve. As a general rule tunnelling is more sensitive to variations in the empty states into which the electrons are tunnelling than in the filled states from which they come [20], and so it is preferable where possible to work from a filled states topography. The curves in Fig. 4 coincide at \( V_s = -2 \) V, and are insensitive to small differences in the density of filled states near the bottom of the band gap. Each \( I-V \) curve is the average of several measurements made over equivalent sites. Curves (a) and (b) were taken over a dimer row where the dimers appear symmetrical, over a dimer and between two dimers respectively. They indicate the presence of both filled and empty states within the bulk bandgap, associated with the bonding and antibonding dimer \( \pi^* \)-orbitals. Curves (c) and (d) were measured over dimers whose tilt was pinned. Significant differences are seen, however, in the empty states at the top of the band gap. The ‘down’ end (c) of a tilted dimer has a higher density of empty states than the ‘up’ end (d), because it finds itself in a more nearly planar configuration, corresponding to \( sp^2 \)-like hybridization, leaving the remaining p-orbital with a relatively high energy. The up end is more \( sp^3 \)-like, so that the
two electrons in the dimer π-bond are to some extent localized on the down atom, leading to the spectra seen in Figs. 4 (c) and (d). This corresponds to the compensation of the tilt contrast in empty-states imaging described in Section 2.2.1.

The effect of molecular adsorption can be seen in Fig. 5, which presents the average from several spectra over ethene molecules adsorbed on Si(001), together with a clean dimer spectrum for comparison [19]. There are no bandgap states in the ethene spectrum, indeed there is an extended voltage range over which the current is too small to measure. This has consequences for imaging. It means that in order to achieve the required tunnelling current the feedback loop may move the tip down when it is over an adsorbed ethene molecule, even though topographically the molecule is higher than nearby clean dimers. In spectra of semiconductor surfaces acquired at a fixed tip-sample separation, the current often becomes too small to measure at small values of the magnitude of the bias voltage. The derivative d ln I/d ln V then becomes too noisy to be useful; this is why it has not been plotted in Fig. 5. To overcome this problem a technique has been developed in which the tip height is changed as the spectrum is acquired, so as to give a larger effective dynamic range [21]. The data are subsequently processed to compensate for the different values of separation at different bias voltages.

Tunnelling spectra of semiconductors can exhibit an exaggerated bandgap. This is because of band bending induced by the field between the tip and the sample. The phenomenon is similar to the band bending which occurs at a biased Schottky barrier [22]. The effect is illustrated in Fig. 6, which shows spectra obtained from GaAs(110) surfaces on materials with three different levels of doping [23]. The most heavily doped sample exhibits a bandgap similar to the bulk value, but for lower doping levels the apparent bandgap becomes extended at positive sample bias. Perfect GaAs(110) surfaces have no
Fig. 6. Tunnelling current vs sample bias for GaAs(110) surfaces with different doping: (—) $n^+$; (—) $n^0$: ··· $n^-$. For each sample the gap was set to give $I = 0.2\,\text{nA}$ at a sample bias $V = -2.7\,\text{V}$ [23].

surface states, and so in the absence of an external field the bands are flat. Fig. 7 illustrates how band bending is induced by the tip. At negative sample bias not much band bending can occur with n-doped material (Fig. 7(b)), because as soon as the conduction band meets the Fermi level it begins to fill with electrons and so becomes electrostatically buoyant, but at positive sample bias the bands can bend upwards (Figs. 7(c) and (d)). These effects have been modelled both by a simple depletion approximation and using a full solution of the Poisson equation, and although the level of agreement is not perfect it does suggest that this account is along the right lines [23]. Band bending is likely to be important wherever the free carrier concentration in the sample is low, and it has also been observed in spectroscopy of oxides [24]. It must be carefully distinguished from the much more local effects of polarization which are also caused by the tip–sample field, which will be presented in Sections 4 and 5.

The task of achieving reproducible and reliable scanning tunnelling spectra is bedevilled by uncertainties in the geometrical and chemical nature of the tip. The $I$–$V$ curves depend on the electronic properties of both tip and sample, and confident interpretation demands knowledge of the tip which is in general simply not available. A study of metal atoms on metal surfaces has shown that while for a given tip the systematic differences in spectra could be found according to the species of metal atom under the tip, the absolute form of the spectrum varied dramatically from one tip to another [25]. A study of Si(111)-(7 × 7) surface using current imaging tunnelling spectroscopy (to be described below) gave unrecognizably different results following a tip change part way through a scan [26]. Repeated measurements on an oxide surface can give a sufficient range of band gaps and offsets to fit almost any postulated model.² There is, alas, no easy solution to this kind of difficulty. Even when field ion microscopy is used to characterize the atomic arrangement on the tip [12], the chances of the atomic arrangement remaining unchanged for a long time are negligible. Moreover, it is sometimes found that the best tips for imaging are not the best tips for spectroscopy [21]. Scanning tunnelling spectroscopy is

² M.R. Castell (private communication).
Fig. 7. Schematic diagram of the band edges as a function of position for the (1 1 0) surface of n-type GaAs in proximity to a metal tip. In (a) the only band-bending arises from the depletion of carriers near the surface as a result of the work function difference between the tip and the sample. For negative sample voltage (b) there is only small band bending before the conduction band starts to fill; for positive sample voltage (c) a large band-bending is possible before inversion occurs, and the band-bending can become still greater if the surface and bulk are not in effective equilibrium (d) [23].

much too useful to dismiss because of these difficulties; nevertheless, until the tip uncertainties can be overcome quantitative use of the results must proceed with extreme caution.

Despite these caveats, scanning tunnelling spectroscopy can reveal remarkable information about local electronic structure. A rather comprehensive technique for acquiring and displaying spectroscopic information with atomic resolution is current imaging tunnelling spectroscopy (CITS) [27]. At each pixel of image acquisition, the feedback loop is frozen, and a complete set of $I-V$ data is measured at chosen values of bias voltage. The feedback loop is then restored, and the process is repeated at the next pixel. This process can result in a large data set, so that CITS is usually performed over only a limited area or with limited resolution. Instead of treating the results as a series of spectra obtained at different places, the data can be thought of as a series of image planes. The first plane is a topography image at a given bias and set current. The remaining planes are images of the current at different voltages for those
tip heights, hence the term current imaging. The definitive demonstration for CITS was the Si(111)-(7 × 7) structure, in which it proved possible to relate states previously identified from photoemission experiments with specific bonding sites on the surface [27].

For many purposes a simpler approach is adequate to reveal how the structure of the sample depends on the bias voltage. Topographic images obtained at different bias voltages often reveal the key spatial distributions associated with different states. Many microscopes enable different parameters to be set for forward and reverse line scan directions, so that a pair of images at different bias voltages can be obtained. Careful registration between images is necessary, because of piezo hysteresis, but this is usually possible, and it can also be applied to images obtained in different frame scans provided the thermal drift is not too great. For reasons that may be in part related to human pattern recognition skills, and the natural preference for selecting good images for analysis, it often seems to be easier to use voltage-dependent imaging techniques with confidence than it is to be sure of spectroscopy curves alone. An example of the use of voltage-dependent imaging to identify the correct model for a missing dimer defect structure in a Si(001) surface will be given in Section 5.1.2. In that case empty states are imaged at different bias voltages, and the enhancement of states near the defect at relatively low bias voltage is exploited. It is also often useful to compare filled and empty states images of molecules on semiconductor surfaces. For example, ethene molecules on Si(001) look somewhat similar at both positive and negative sample bias, but the contrast from adjacent clean dimers is utterly different, so that having both images makes it easy to see where molecules have been adsorbed. STM images of these molecules will be presented in Section 5.2.1. The image simulation in that case depends crucially on taking into account the tip-sample electrostatic field, and this will be explained in Section 5.2.3. In our experience, providing a satisfactory account of STM images generally proves to be a more reliable criterion than prediction of tunnelling spectra for evaluating candidate models of structures studied by STM. In Section 6 we shall see how other STM information, such as observations about the extent to which missing dimer trenches run in straight lines on GaAs(001) surfaces, enables the plausibility of different models of a surface to be evaluated.

3. Modelling electronic structure of surfaces

Much of the power of the STM comes from its sensitivity to the geometrical and electronic structure of the surface. However there is no inversion theorem from which we can deduce this geometrical and electronic structure, given an STM image. A full interpretation of the information from the STM therefore requires us to be able to generate a plausible model for the surface, solve its electronic structure, and compute the resulting STM image to compare with experiment. This section reviews some of the approaches to these calculations; it is a sign of the importance of surfaces to present-day physics, chemistry, materials science and engineering that the techniques that have been applied to this problem include virtually all those that are available for studying the electronic structure of condensed phases.

3.1. Quantum chemical methods

We first discuss a set of techniques which are often grouped together as ‘quantum chemical’, because they are extensions of the methods used by theoretical chemists to study the electronic structure of molecules.
3.1.1. The Hartree–Fock approximation

The basis of the quantum chemical techniques is the Hartree–Fock approximation [28]. It is assumed that the many-electron wavefunction $\Psi$ can be written as a determinant of one-electron wavefunctions $\{\psi_i\}$:

$$\Psi(r_1, \sigma_1, \ldots, r_N, \sigma_N) = \left| \begin{array}{ccc} \psi_1(r_1, \sigma_1) & \cdots & \psi_1(r_N, \sigma_N) \\ \vdots & \ddots & \vdots \\ \psi_N(r_1, \sigma_1) & \cdots & \psi_N(r_N, \sigma_N) \end{array} \right| \sqrt{N!}$$

(1)

where $N$ is the number of electrons, the $r_i$ are the electron positions, and the $\sigma_i$ are the electron spins (up or down along any chosen direction). This form automatically ensures that $\Psi$ is antisymmetric with respect to the exchange of pairs of particles, as required by the Pauli exclusion principle. The one-electron wavefunctions $\{\psi_i\}$ are found by minimizing the energy of the trial wavefunction $\Psi$; they turn out to obey a set of one-electron-like Schrödinger equations:

$$-\frac{\hbar^2}{2m_e} \nabla^2 \psi_i(r) + V_{\text{ext}}(r)\psi_i(r) + V_H(r)\psi_i(r) + \int dr'V_X(r, r')\psi_i(r') = \epsilon_i\psi_i(r),$$

(2)

where $V_{\text{ext}}$ is the external potential arising from the presence of the atomic nuclei and any external fields, $V_H$ is the Hartree potential (the potential arising if the electron density is regarded as a classical charge density)

$$V_H(r) = \frac{e^2}{4\pi\epsilon_0} \sum_{j} \frac{|\psi_j(r)|^2}{|r - r'|},$$

(3)

and the nonlocal ‘exchange potential’ $V_X$ is given by

$$V_X(r, r') = \frac{e^2}{4\pi\epsilon_0} \sum_{j} \frac{\psi_j^*(r)\psi_j(r')}{|r - r'|}.$$

(4)

Note that the sum in Eq. (3) goes over all the single-particle states, whereas that in Eq. (4) goes only over states with the same spin.

3.1.2. Implementation of the Hartree–Fock approximation

Since the Hartree–Fock equations are a set of coupled integro-differential equations, they cannot be solved analytically except in extremely simple cases and some numerical representation of the one-electron functions $\psi_i$ is required. In the vast majority of implementations, one writes

$$\psi_i(r) = \sum_{\nu} c_{\nu} \phi_\nu(r),$$

(5)

where the functions $\phi_\nu$ are localized basis functions, typically centred on the nuclear sites. The analogue of the Hartree–Fock equations is now the Hartree–Fock–Roothaan equations, which take the form of a matrix eigenvalue equation for the coefficients $c_{\nu}$:

$$\sum_{\nu} F_{\mu\nu} c_{\nu} = \epsilon_i c_{\mu},$$

(6)

where $F$ is the Fock matrix and is the matrix representation of the Fock operator (kinetic energy plus external, Hartree and exchange potentials) in the chosen basis set.
The angular parts of the basis functions \( \phi_n \) are fixed by the requirement that they represent states of a certain definite angular momentum; the radial parts are usually taken as Gaussians, because this makes the evaluation of the matrix elements of the Fock operator relatively straightforward. A number of very sophisticated Hartree–Fock codes have been developed along these lines, such as the GAUSSIAN packages for molecular systems [29] and the CRYSTAL program for periodic systems [30]. A further discussion of the geometry which may be used in calculations on a surface may be found in Section 3.5 below.

3.1.3. Strengths and weaknesses of Hartree–Fock

Particularly when compared to the density functional methods described below, the great strength of the Hartree–Fock method is that it gives an explicit expression for the many-electron wavefunction. Being the result of a well-defined variational procedure, it therefore lends itself to a series of natural improvements by making the class of variational wavefunction more general than a single Slater determinant (see Section 3.4.2 below).

Therein, however, also lies one of its main disadvantages: because of the simple form of the wavefunction, all correlation effects are neglected – the electrons are assigned a definite set of single-particle states, despite the scattering events that can in fact occur because of their mutual Coulomb interaction. It is true that the ‘correlation energy’ (the change in the total energy associated with these correlation effects) is generally small in comparison with the total energy in an atom or molecule; however, changes in the correlation energy can have a significant effect on the relatively small energy changes that govern changes in bonding (such as surface reconstructions and chemical reactions). One reason for this is that electrostatic interactions in many solids are very effectively screened; this behaviour is not captured in the effective potentials (3) and (4). Indeed, in metals, the unscreened long-range nature of the Coulomb interaction has such a dramatic effect within the Hartree–Fock approximation that the Fermi surface disappears altogether [31], in direct contradiction to experiment. Any effects which rely on the properties of the Fermi surface are therefore very poorly reproduced by the Hartree–Fock approximation.

Another disadvantage of the Hartree–Fock approximation is that the exchange potential (4) is non-local; this makes it in general more expensive to implement than methods based on a local potential (Section 3.2). Moreover, while it is possible to calculate not just the ground-state energy of the system but also the atomic forces (gradients of the total energy) and the force constants (second derivatives of the total energy), these calculations become progressively more time-consuming. It is partly for this reason that semi-empirical approximations to Hartree–Fock theory were developed; these have had considerable use in applications to surface systems, and are described in Section 3.1.4.

3.1.4. Semi-empirical approximations to Hartree–Fock

A full Hartree–Fock calculation in principle involves the calculation of \( N^4 \) matrix elements of the Coulomb interaction in order to determine the Fock matrix. (In practice, for sufficiently large systems where the basis functions are localized, the locality of the Hartree potential and the short range of the nonlocality in the exchange potential can be exploited to reduce this scaling to \( N^3 \) – the same scaling as the diagonalization of the Fock matrix.) Particularly in the early days of computer-based calculations, the expense of the construction of the full Fock matrix led to simplified methods which use a minimal set of basis orbitals (typically only one orbital of each angular momentum relevant in the valence shell)
and rely on the fitting of some experimental data to compensate for the errors which are thereby introduced.

There are many different semi-empirical methods and we give here only a list of the most common, with references where further information may be found.

- **CNDO (complete neglect of differential overlap) [32]**: This is the original technique and although it is now seldom used, it has provided the basis for much subsequent work. All three-centre and four-centre terms (i.e., integrals of the form $\langle \mu \nu | \lambda \sigma \rangle$ where the orbitals are on three or four different atoms) are neglected, as are all one-centre and two-centre exchange terms. In order to make the theory invariant under transformations of the basis states, the remaining two-atom two-electron integrals (representing the Coulomb interaction of electrons localized on different atoms) must depend only on the distance between the atoms, and must be the independent of the angular character of the electronic states.

- **INDO (intermediate neglect of differential overlap) [32]** is similar to CNDO, but includes two-orbital, one-centre exchange terms.

- **MINDO (modified intermediate neglect of differential overlap) [33]**: This method is similar to INDO, except that different orbital exponents are allowed for basis functions with different values of the orbital angular momentum $l$, and an empirical modification is made to the repulsive potential between the atomic cores.

- **MNDO (modified neglect of diatomic overlap) [34]**: This is based on an older approximation (NDDO – neglect of diatomic differential overlap [32]), which is now seldom used.

- **AM1 (Austin Model 1, whose inventors work at the University of Texas at Austin) [35]**: Despite the break in naming conventions, this method is closely modelled on MNDO but with a modified treatment of the core repulsion.

Recently, methods have been developed which attempt to include electron correlations within a similar semi-empirical framework. They involve constructing a configuration interaction calculation (see Section 3.4.2) in which the Hamiltonian parameters are optimized to reproduce the excitation spectra of certain sets of trial molecules. Note that these parameters will in general be different from those in a semi-empirical Hartree–Fock calculation, which are optimized to reproduce bond lengths and reaction energies.

The semi-empirical methods have produced good results for molecules within a given class; for example, a method whose parameters are fitted to a database of heterocyclic organic compounds will usually give a good account of other heterocyclic compounds. However, caution must be exercised when using them outside this domain; in particular, surface problems can involve unusual coordination or bonds between atoms which normally do not react. This is not to say that semi-empirical methods should not be used for surface problems, but it does underlie the need for careful checking with ab initio calculations (where possible) and with experiment if one is to have confidence in the results.

### 3.2. Density functional theory

In contrast to the Hartree–Fock method, density functional theory is based, not on the all-electron wavefunction $\Psi$, but on the electron density $\rho(r)$. This approach was initiated by Hohenberg and Kohn [36], who demonstrated the remarkable fact that the ground-state energy of an electron gas can, in principle, be deduced from the electron density alone. Their proof consists of the following steps:
(i) Since the external potential $V_{\text{ext}}(\mathbf{r})$ is local and is the only thing which can differ between two systems containing the same number of electrons, it is shown to be impossible for two different external potentials to give rise to ground states with the same charge density. This establishes that the potential is a unique functional of the ground state density.

(ii) For any given density, one can then in principle construct the external potential and solve the (many-body) Schrödinger equation to determine the ground-state energy.

This is a remarkable result; the trouble is, it tells us nothing about the form of the functional linking the density and the energy. The next step, which began to provide a useful framework for practical calculations, was taken the following year by Kohn and Sham [39]. They constructed a fictitious, noninteracting electron system out of a set of single-particle orbitals $\psi_i(\mathbf{r})$ (now known as Kohn–Sham orbitals); the noninteracting and interacting electron systems are supposed to share a common charge density

$$\rho(\mathbf{r}) = \sum_i f_i |\psi_i(\mathbf{r})|^2, \quad (7)$$

(where $f_i$ is the occupation number: for a non-spin-polarized system, $f_i = 2$ for filled states and 0 for empty states at $T = 0$). The energy of the interacting system is then written as a functional of the electron density in the following form:

$$E[\rho] = T_0[\rho] + E_{\text{ext}}[\rho] + E_{\text{H}}[\rho] + E_{\text{XC}}[\rho], \quad (8)$$

where

$$T_0[\rho] = \sum_i f_i \int d\mathbf{r} \psi_i^*(\mathbf{r}) \left[ -\frac{\hbar^2}{2m_e} \nabla^2 \psi_i(\mathbf{r}) \right] \quad (9)$$

is the kinetic energy of the noninteracting electron system, $E_{\text{ext}}$ is the energy of interaction with the external potential

$$E_{\text{ext}} \equiv \int d\mathbf{r} \rho(\mathbf{r}) V_{\text{ext}}(\mathbf{r}), \quad (10)$$

$E_{\text{H}}$ is the energy of interaction with the Hartree potential (3), and the so-called ‘exchange–correlation energy’ is simply the correction term to the other three. If one now varies the Kohn–Sham orbitals $\psi_i$ (and hence the charge density $\rho$) so as to minimize the energy, one finds that they obey exactly a set of effective one-electron Schrödinger-like equations

$$-\frac{\hbar^2}{2m_e} \nabla^2 \psi_i(\mathbf{r}) + V_{\text{ext}}(\mathbf{r}) \psi_i(\mathbf{r}) + V_{\text{H}}(\mathbf{r}) \psi_i(\mathbf{r}) + V_{\text{XC}} \psi_i(\mathbf{r}) = \epsilon_i \psi_i(\mathbf{r}), \quad (11)$$

where $V_{\text{ext}}$ and $V_{\text{H}}$ are as before the external and Hartree potentials, respectively, while the exchange–correlation potential $V_{\text{XC}}$ is the functional derivative of the exchange–correlation energy with respect to

---

3 Provided that the density concerned corresponds to the ground-state energy for some potential. This class does not, in fact, include all ‘physically reasonable’ charge densities, and a number of efforts have been made to stop this loophole [37,38].

4 This step also gives rise to subtle problems [38].
the charge density at position \( r \):

\[
V_{XC}(r) \equiv \frac{\delta E_{XC}}{\delta \rho(r)}.
\]  

(12)

Note that, unlike the Hartree–Fock equations, the Kohn–Sham equations are in principle exact; the exchange–correlation potential exactly represents both the effects of exchange and of the correlation between the electrons. Note also that the effective potential appearing in them is local, unlike the exchange potential of Hartree–Fock theory.

### 3.2.1. The local density approximation (LDA)

So far we have simply defined the exchange–correlation energy as the difference between those parts of the energy functional we know and the correct answer. This is not very helpful if we actually want to perform calculations. Luckily, Kohn and Sham [39] also came up with a simple approximation for the exchange–correlation energy that is surprisingly useful. Let \( \epsilon_{XC}(\rho) \) be the exchange–correlation energy density per unit charge density (i.e. the exchange–correlation energy per electron) of a uniform electron gas of density \( \rho \). Then the ‘local density approximation’ (LDA) consists in assuming that the exchange–correlation energy density at each point in our system is the same as it would be in a uniform electron gas of the same charge density, i.e. in writing

\[
E_{XC}^{\text{LDA}} = \int d\rho(r) \epsilon_{XC}(\rho(r)).
\]

(13)

Using eq. (12) immediately gives an expression for the corresponding exchange–correlation potential:

\[
V_{XC}^{\text{LDA}}(r) = \epsilon_{XC}(\rho(r)) + \frac{d\epsilon_{XC}}{d\rho}.
\]

(14)

Modern implementations of the local density approximation generally use values for \( \epsilon_{XC} \) obtained from quantum Monte Carlo calculations on the uniform electron gas [40]; it is frequently convenient to adopt a parametrization of these results due to Perdew and Zunger [41].

Despite its simplicity, the LDA has proved to be surprisingly accurate for the structural properties of strongly bonded systems. Although there is typically a substantial error in the binding energy relative to isolated atoms, the total energy differences between different condensed-phase structures are given correctly (see, for example, the early work on silicon [5] and more recent calculations on 3d transition metals [42]). Where the LDA fails is in calculations involving more open or weakly bonded structures; the total energies for isolated atoms are rather poor (hence the errors in the formation energies of solids), Van der Waals bonded dimers are dramatically overbound [43], and the LDA hydrogen bond is much too strong [44].

### 3.2.2. The generalized gradient approximation (GGA)

Unfortunately, there is no systematic way to improve on the LDA. Initial attempts to extend it by treating it as the first term in an expansion of the exchange–correlation energy in powers of the gradient of the charge density proved unsuccessful; this is now understood to be because such a series is anyway divergent, and because terms after the first violate a sum-rule which states that the total deficit

\[\text{Notice that } \epsilon_{xc} \text{ is a function, not a functional, of } \rho; \text{ it depends only on the charge density at a single point.} \]
in the number of electrons produced by exchange and correlations around an electron as it moves through the system (the so-called ‘exchange–correlation hole’) must total exactly one electron [45].

The difficulties surrounding these gradient expansions have been partially overcome by more recent ‘generalized gradient approximations’ (GGAs) [46–49]. These are formulae which are chosen to satisfy the sum-rule on the exchange–correlation hole, and also to give the correct asymptotic behaviour of the exchange–correlation energy in the tail of the charge density of an atom. (They generally fail, however, to give also the correct behaviour of the potential.) While they can be made plausible, they lack any rigorous theoretical justification. It is therefore best to judge them by their results, which are on the whole encouraging; use of the GGA greatly improves the hydrogen bond [44], other weakly bonded systems [50], and transition states [51,52], as well as certain phenomena involving comparing energies of closely packed and open structures in strongly bonded systems, such as the phase behaviour of SiO2 under high pressures [53]. There have been some suggestions, however, that the GGA spoils the generally excellent agreement obtained in LDA with elastic moduli and phonon frequencies of solids.

One combination of techniques which initially gave quite reasonable results was to calculate the GGA total energy, but using a charge density derived from a previous LDA calculation; this is often referred to as a post hoc GGA correction. It has the advantage that one does not have to carry out a fully-self-consistent calculation using GGA – a procedure in which some care is necessary if spurious results are not to be obtained [54,55].

There is therefore sufficient empirical evidence to suggest that gradient corrections are more likely to improve the accuracy of LDA calculations than to worsen them, particularly if comparisons of structures involving weak bonds or broken bonds is involved. In the absence of a more systematic theoretical foundation, however, each case should be treated on its merits and the different approximations judged by their results.

3.2.3. Localized and augmented basis sets

Having decided on the exchange–correlation functional, there are then several options available for the representation of the wavefunctions in a density functional calculation. The most similar in spirit to traditional ‘quantum chemical’ implementations of the Hartree–Fock method (see Section 3.1.2) is to expand the Kohn–Sham wavefunctions in terms of localized basis functions. These might be Gaussians or atomic-like wavefunctions. One of the most ingenious methods for solids of this type is the LMTO (linear muffin tin orbitals) technique, in which a small atom-centred basis set is produced and adapted during the calculation in such a way that the expansion of the wavefunctions converges extremely efficiently. However, the scheme is most efficient only if the ‘atomic sphere approximation’ is made, in which only the spherical component of the potential around each atom is considered; this approximation is often poor for atoms at surfaces. Further examples of such localized basis sets are discussed in Section 3.3.

A quite different approach is motivated by Bloch’s theorem; for a crystalline solid, the wavefunction can always be written in the form \( \psi(r) = \exp(i \mathbf{k} \cdot \mathbf{r}) u(r) \), where \( u \) is a function with the periodicity of the crystal lattice. It is therefore natural to expand \( u \) in a Fourier series (i.e. in terms of the plane waves \( \exp(i \mathbf{G} \cdot \mathbf{r}) \), where \( \mathbf{G} \) is a reciprocal lattice vector). Such an expansion has the advantage that it is unbiased, straightforward, and its convergence can be controlled by a single parameter (the magnitude of the maximum or ‘cutoff’ value of \( \mathbf{G} \)). However it is very wasteful in the interior of the atoms, where the potential and the wavefunction change very rapidly, requiring a correspondingly enormous number of Fourier components. In the augmented plane wave (APW) method and the (now much more
widespread) LAPW (linearized augmented plane wave) method [56], the plane-wave expansion is used only for the interstitial region between the atoms, while a completely separate solution of the Schrödinger equation is performed for each angular momentum around each atomic site. The results are then matched, by value and normal derivative, onto the rest of the wavefunction at the boundary of the atomic and interstitial regions. (Note that we have implicitly assumed that we are dealing with a periodic structure – this may seem strange in a review devoted to the properties of surfaces, but we shall return to this point in Section 3.5 below.)

3.2.4. Pseudopotential techniques

Another method of solving the Kohn–Sham equations relies on the use of pseudopotentials. Since the atomic cores contribute a very large term to the total energy of the system which is essentially independent of how the atoms are bonded and therefore does not contribute to atomic forces or to energy differences between structures, the idea is to replace the true core potential and core electrons with an effective potential, the pseudopotential, which has the same scattering properties as the true potential as far as the valence electrons are concerned. This idea has a long history [57,58] but the modern era of quantitative pseudopotential calculations started with the discovery of how to produce norm-conserving pseudopotentials [59] which keep the overall normalization of the valence electrons correct. This enabled pseudopotentials to be produced directly from all-electron atomic calculations on a very wide range of elements [60], and used in self-consistent electronic structures without introducing major errors in the electrostatics.

Because the pseudopotential is much weaker and more slowly varying than the full atomic potential, its Fourier representation converges relatively quickly. This makes it feasible to use the same very simple basis set of plane waves as in the APW techniques, expanding the one-electron wavefunctions in the form

$$\psi(\mathbf{r}) = \sum_{\mathbf{G}} \exp[i(\mathbf{k} + \mathbf{G}) \cdot \mathbf{r}],$$

where $\mathbf{k}$ is the Bloch wave vector of the state and the sum goes over reciprocal lattice vectors of the structure, but this time throughout the whole of the system. Without a pseudopotential, there would be serious problems involved in representing the true (rapidly varying) wavefunction close to the nucleus using Eq. (15). The first methodology for using this combination of plane waves and pseudopotentials to calculate the total energy of solids was developed by Yin and Cohen [4].

However, the standard pseudopotentials can still cause problems for atoms where the valence electrons do not contain a node; these include the first-row elements (2p electrons), the first-row transition elements (3d electrons) and the rare earths (4f electrons). In such cases the pseudopotential is still relatively strong, and the corresponding pseudo-wavefunctions still vary rapidly in space and are poorly Fourier-converged. In recent years there have been a number of developments of the pseudopotential method to overcome this problem, ranging from ‘optimizing’ the pseudopotential generation procedure [61–63] to the more radical approach of abandoning the norm-conservation constraint on the wavefunction and compensating elsewhere in the calculation for the errors thereby introduced [64].

The procedure of pseudopotential generation is relatively complex and is still something of a ‘black art’. The plane-wave convergence properties of a given pseudopotential are difficult to predict and must be carefully checked. One must also check the transferability of a pseudopotential: it must reproduce
the scattering properties of the true potential in the condensed-phase or surface environments where it is to be used, as well as in the isolated atom. This check is usually done by calculating a quantity such as the lattice constant of a material where the result can be checked with experiment or with a carefully converged all-electron calculation. If a comparison with experiment is used, it can be difficult to separate errors arising from the pseudopotential or the plane-wave cutoff from those that are intrinsic to the exchange–correlation approximation (LDA or GGA) being used.

3.2.5. The projector augmented wave (PAW) technique

Many of the most desirable properties of pseudopotential techniques (such as the use of a plane-wave basis set, easy calculation of atomic forces) and of all-electron methods such as LAPW (availability of the full wavefunction throughout all space, ability to treat transition metals and noble metals) are shared the projector augmented wave (PAW) method [65]. In this approach, the full wavefunction \( \psi(\mathbf{r}) \) can be recovered from the smooth pseudo-wavefunction \( \widetilde{\psi}(\mathbf{r}) \) via the transformation

\[
\psi(\mathbf{r}) = \widetilde{\psi}(\mathbf{r}) + \psi_1(\mathbf{r}) - \widetilde{\psi}_1(\mathbf{r}),
\]

where \( \psi_1(\mathbf{r}) \) and \( \widetilde{\psi}_1(\mathbf{r}) \) are nonzero only within the atomic region (the superscript ‘1’ denotes one-centre, i.e. they refer to a single atomic centre). Outside the atomic region, \( \psi = \widetilde{\psi} \) (which is represented in plane waves, just as for a pseudopotential calculation).

So, there is an augmentation to the wavefunction inside the atoms, just as in the LAPW method. However, it is calculated in an entirely different way: the one-centre functions \( \psi_1(\mathbf{r}) \) and \( \widetilde{\psi}_1(\mathbf{r}) \) are expanded as

\[
\psi_1(\mathbf{r}) = \sum_L c_L \phi_L(\mathbf{r}) \quad \text{and} \quad \widetilde{\psi}_1(\mathbf{r}) = \sum_L \tilde{c}_L \tilde{\phi}_L(\mathbf{r}),
\]

where \( \phi_L(\mathbf{r}) \) and \( \tilde{\phi}_L(\mathbf{r}) \) are a set of basis functions chosen to describe the true atomic wavefunctions and nodeless atomic pseudo-wavefunctions, respectively. The coefficients \( c_L \) are determined from the overlap of the smooth wavefunction \( \widetilde{\psi}(\mathbf{r}) \) with ‘projector functions’ which, like \( \phi_L \) and \( \tilde{\phi}_L \), are derived from an atomic calculation. This contrasts with the LAPW approach, where the modified form for the wavefunction inside the atomic region is determined by matching value and derivative of the wavefunction at the sphere boundary. Since the projector functions are well-behaved at the edge of the atomic region, it is much more straightforward to apply the Hellmann–Feynman theorem to calculate the forces on the atoms than it is in the LAPW method; however, the advantage of having access to the full wavefunction near the nucleus, and being able to include corresponding corrections to the total energy, is retained.

3.2.6. The Car–Parrinello revolution

The combination of pseudopotentials and plane waves gives a very efficient means of calculating total energies for periodic systems of atoms. The scope of the problems that could be tackled using it was dramatically enhanced with the work of Car and Parrinello in 1985 [6]. The Car–Parrinello technique takes the radical step of treating the atomic positions and the plane-wave coefficients of the electron wavefunctions as dynamical variables on the same footing. A classical dynamical system is

\[\text{In fact, it can be shown that both the pseudopotential and LAPW methods can be derived as special cases of the PAW technique [65].}\]
constructed which includes, as well as the LDA energy and the kinetic energy of the atoms, a wholly fictitious kinetic energy associated with the electron wavefunctions [66]. This approach leads to several important benefits:

(i) The minimization of the electronic energy can be carried out as a direct minimization problem, rather than by solving a succession of eigenvalue problems. This can be done extremely efficiently, as the gradient of the LDA energy with respect to the wavefunctions can be computed using fast Fourier transforms (FFTs).

(ii) The Hellmann–Feynman theorem can be used to find the forces on the atoms in terms only of the derivatives of those parts of the electronic Hamiltonian which depend explicitly on the atomic positions (i.e. the atomic pseudopotentials). This process is also extremely efficient – indeed the expense is negligible once the total energy has been computed.

(iii) The minimization of the energy with respect to electronic wavefunctions and atomic positions can be performed simultaneously.

(iv) Molecular dynamics can be performed, in which the atoms follow Newton’s equations of motion with forces derived from the LDA Born–Oppenheimer potential surface.

As well as ‘pure’ Car–Parrinello approaches in which the wavefunctions are treated as dynamical variables, similar ideas have been extended to find the ground-state of the electronic wavefunctions by approaches such as conjugate-gradients minimization [67,68] or iterative diagonalization [69]. These methods have been developed in response to slightly different problems – for example, the need to perform dynamics, or to find ground states efficiently, or to treat metallic systems. They all share the application of fast Fourier transforms to avoid explicitly computing and storing the entire electronic Hamiltonian, and all have in common the ability to calculate atomic forces very easily.

3.2.7. Prospects for O(N) methods in density functional theory

Despite the efficiency of Car–Parrinello methods and their derivatives, the CPU time for the calculations still scales as $N^3$ in the large-$N$ limit, where $N$ is the number of atoms in the system. This scaling manifests itself in different places according to the algorithm used: it appears in the diagonalization of the Hamiltonian matrix, or in the orthogonalization of the wavefunctions if one of the iterative minimization schemes are used. However there are good reasons for believing that the total energy can, in principle, be evaluated in a way that depends only on the local environment of each point in the system, and hence that should scale only linearly with the size [70,71].

An improvement of the Car–Parrinello technique to O(N) scaling is obtained if each Kohn–Sham wavefunction is localized in one part of the system, as it is then not necessary to orthogonalize each function to all the others and the full fast Fourier transform grid is not necessary [72]. There has recently been a great deal of activity in this field, with a number of approaches being tried. These have included the use of a direct representation of the wavefunctions on a real-space grid rather than in terms of plane waves [73,74]; this requires care in the representation of the kinetic energy, but does away with the requirement for periodic boundary conditions and makes the localization of the functions more straightforward. Another ingenious suggestion was made by Vanderbilt et al. [75]; the Kohn–Sham total energy can be written in terms of the one-particle density matrix

$$\rho(r, r') = \sum_n f_n \psi(r) \psi^*(r'),$$  \hspace{1cm} (18)
where the sum goes over the eigenfunctions of the system and $f_n$ is the occupation number of states $n$. The density matrix elements may themselves be used as variational parameters; they go to zero as $|\mathbf{r} - \mathbf{r}'| \to \infty$ (exponentially rapidly for insulators, as a power law for metals) and so may be truncated after a certain point, leaving a number of parameters scaling linearly with the size of the system. A trick is used to generate a Kohn–Sham density matrix that is approximately ‘idempotent’ (i.e. has eigenvalues all 0 or 1, as the true Kohn–Sham density matrix must at low temperatures) from a nonidempotent trial density matrix which is itself expanded in localized support functions [76–78]. A third possibility is to revert to a more ‘quantum chemical’ approach and to expand all quantities in terms of strictly localized ‘atomic-like’ orbitals [79–81].

At the time of writing impressive demonstration calculations have been performed demonstrating linear scaling in static calculations up to ordered systems of several thousand atoms; it remains to be seen how effectively the convergence problems which tend to arise when disordered systems are considered can be resolved. For the moment, then, these first-principles methods are promising newcomers rather than the workhorses of surface physics. In the area of tight-binding calculations, where the extra complications of electronic self-consistency do not need to be addressed, the situation is quite different and O($N$) methods are already making a significant impact, as we shall see in Sections 3.3.2, 5 and 6.

3.3. Approximations to density functional theory

Like the Hartree–Fock method, implementations of density functional theory require a self-consistency cycle to ensure that the effective potential used in solving the Kohn–Sham equations does indeed correspond to the charge density generated by the Kohn–Sham orbitals. An approximate theory which does not require self-consistency was derived by Harris [82] (and independently by Foulkes [83]); it relies on the fact that the true charge density minimizes the total energy. The method starts with a trial charge density (for example, overlapping atomic charge densities); an expression for the total energy is obtained which involves constructing the effective potential and solving the Kohn–Sham equations just once. The error is then known to be of second order in the deviation of the trial charge density from the true (self-consistent) charge density. Although the straightforward superposition of atomic charge densities is now known to be insufficiently accurate for surface calculations [84,85], the method is very useful as a starting point for further approximations.

3.3.1. Tight binding

Tight-binding methods, based on a linear combination of a very limited basis set of orbitals localized at atomic sites, have a long history in the electronic structure of solids [86–91]. It is now recognized that they can be derived from the Harris–Foulkes approximation [92] – although the derivation does not give an unambiguous prescription for obtaining the tight-binding parameters. These must therefore be obtained by fitting to a database of bulk or molecular properties [86,89,93], in a similar fashion to the parameters for semi-empirical Hartree–Fock theories; the same words of caution apply with regard to extending the methods to applications far away from this fitted database. This said, the ability of tight-binding methods to cope with large systems that are currently beyond the scope of ab initio treatments

---

7 The explicit solution of the Kohn–Shan equations is still a process scaling as $N^3$; in practice, however, the method is seldom implemented in this way.
has been crucial in identifying a number of important surface phases observed in the STM (see Section 6).

Tight-binding methods are generally non-self-consistent, and do not include any explicit electron–electron interaction terms. Their predictions must therefore be treated with particular caution when large transfers of charge are involved. In metallic or semiconducting systems it is often reasonable to exclude such charge transfer altogether, and introduce a simplified self-consistent procedure in which the ‘on-site’ Hamiltonian matrix elements are adjusted to make every atom charge-neutral [92].

3.3.2. Linear scaling methods based on tight binding

The O(N) procedures, based on localizing the wavefunctions or density matrix of the system (see Section 3.2.7) can also be applied to tight-binding methods – indeed, because of the simplicity of the Hamiltonian and the freedom from complications related to charge self-consistency, the progress of O(N) methods has been most rapid in this area [75,94,95].

Because they are among the simplest methods to capture the essence of quantum-mechanical bond formation, tight-binding approaches also form a natural foundation on which to build truly quantum-mechanically based interatomic potentials. The quest for a simple and universal interatomic force law has a very long history [96]. The simplest assumption is that the total potential energy can be written as a sum of interactions between pairs of atoms; however this proves to be a very poor approximation for metals and covalent semiconductors, because the formation of covalent or metallic bonds does not in reality occur independently for each pair. (An elegant ab initio demonstration of this for a sequence of metallic structures may be found in [97].) A number of ‘many-body’ potentials have been proposed which capture some of the essential cooperative nature of the bonding in these systems [98–101].

Recently a new form of potential, the bond order potential (BOP), has been proposed [102–104]; it relies on a computation of the strength of the local bond order (effectively the difference between the occupation of the local bonding and antibonding states), based on information about neighbouring atoms at successively greater and greater distances. The calculation of the bond strength is then based on a tight-binding model; to a first approximation the method reduces to the Tersoff potential [100] for tetrahedral semiconductors, and to the Finnis–Sinclair potential [99] for metals. It is in principle possible to obtain a systematic analytic expansion beyond this point, but in practice it becomes algebraically very cumbersome and it is more convenient to make a direct numerical evaluation of the bond order. In this form the method is essentially another O(N) approach to electronic structure within the tight-binding approximation. Both the bond-order and density-matrix [75] approaches have been implemented in the OXON package at the Department of Materials in Oxford. In systems where direct comparisons have been made, it has been found that the density-matrix approach is more suitable for semiconducting systems (where the density matrix elements decay rapidly in space, but information on a very large number of surrounding atoms is necessary for BOP to reconstruct correctly the gap in the density of states), while BOP is more suitable for metals (where the density of states is smooth, but the density matrix elements decay only slowly) [105].

3.4. Other methods

The vast majority of calculations on the electronic structure of surfaces are based on density-functional theory or its approximations, or (less frequently) on the Hartree–Fock approach. However, it will become possible over the next few years to apply techniques which go beyond this level of theory
in their treatment of electron correlations, and it seems clear that this will become a major area of activity.

3.4.1. Quantum Monte Carlo

The idea of Quantum Monte Carlo methods is to evaluate the ground-state energy of a system of electrons at a certain fixed set of atomic positions by using a sequence of random numbers. The version of the technique which is simplest conceptually, and the most advanced in applications, is variational Monte Carlo. Here a fully antisymmetrized, many-electron wavefunction \( \Psi \) is selected, which depends on a set of variational parameters. Typically, the wavefunction is a Slater determinant multiplied by a Jastrow factor, which has the effect of preventing the electrons from coming close together; the variational parameters determine how strongly the electrons avoid each other. The total energy of the \( N \)-electron system is then evaluated, in the form of an integral over the \( 3N \)-dimensional configuration space of the electrons. The number of points needed to perform this integration by any grid-based technique would grow exponentially with \( N \), so the random numbers are used to generate a selection of points ‘sampling’ the region of integration. The ground-state energy is then minimized as a function of the variational parameters appearing in the wavefunction.

Such Monte Carlo methods have now been applied to find very accurate total energies for atoms [106] and simple semiconductors [107]. Applications to more strongly correlated systems such as NiO, to defects and to surfaces are just beginning. A much more complete survey of this field, along with a description of other algorithms such as diffusion Monte Carlo and Green’s-function Monte Carlo, can be found in [108].

3.4.2. Wavefunction methods beyond Hartree–Fock

Theoretical chemistry is used to the idea that the description of the wavefunction obtained with the Hartree–Fock approximation may not be adequate; this applies particularly to the chemistry of open-shell species and radicals.

There are two approaches to improving the quality of the wavefunction, and hence of the total energy. One is to make a perturbation expansion (of the so-called Möller–Plesset type) in the electron–electron interaction; these techniques are usually referred to as MP\( n \) methods, where \( n \) is the order to which the perturbation expansion is taken. The total energy in an MP\( n \) calculation takes the form

\[
E = E_{\text{HF}} + \sum_{k=1}^{n} E_k, \tag{19}
\]

where \( E_{\text{HF}} \) is the Hartree–Fock energy and \( E_k \) is a term of order \( k \) in the electron–electron interaction operator (see [28] for further details). The second approach is to write the wavefunction as a sum of different Slater determinants:

\[
\Psi = \sum_i \alpha_i \Psi_i, \tag{20}
\]

where each \( \Psi_i \) is a Slater determinant (or ‘configuration’), as in Eq. (1). In principle, if one has a complete basis of single-particle states, and includes in the sum of Eq. (20) all the possible ways of constructing Slater determinants, the description of the many-electron wavefunction will be complete and “all” one has to do is to find the coefficients \( \alpha_i \) (for example, by setting up a matrix representation of the Hamiltonian operator in the basis of the Slater determinants \( \{ \Psi_i \} \) and diagonalizing it). However,
in practice this leads to a problem size that grows exponentially large, even for very simple systems. A more usual approach, therefore, is first to perform a Hartree–Fock calculation to obtain a single ‘best’ set of one-electron orbitals. One then chooses only the set of determinants \( \Psi_i \) which can be formed by taking a small number of electrons (typically up to two) from the occupied orbitals, and replacing them in unoccupied ones. This type of calculation would be referred to as ‘configuration interaction with single and double excitation’. Another approach is so-called ‘multi-configuration SCF’, where a relatively small number of terms are taken in Eq. (20) and then both the coefficients \( \alpha_i \) and the Slater determinants \( \Psi_i \) are chosen variationally to minimize the total energy.

Both these types of calculation can be performed on molecules, and by extension also on clusters of atoms. However, the perturbation approach is much more suitable for extended systems, because (unlike configuration interaction) it is extensive – the total energy is guaranteed to be proportional to system size for a homogeneous system. We will now study more carefully the question of how the geometry of a surface should be represented in calculations.

3.5. The surface geometry – clusters, slabs and embedding

The question of what geometry to use for a surface calculation is not a straightforward one. Historically, the two most popular choices have been a cluster geometry or a slab geometry (see Fig. 8). For a cluster calculation (Fig. 8(a)), a small portion of the solid near the surface is excised and treated in isolation using methods originally developed for molecules; the potential pitfall is then in the treatment of the cluster boundary. For a slab calculation (Fig. 8(b)), the system is extended with two-dimensional periodicity parallel to the surface, and a second surface is created a few atomic layers below the first. Depending on the details of the calculation it may also be necessary to create an artificially periodic system in the third (perpendicular) dimension (Fig. 8(c)), so that the system can be treated exactly like a three-dimensional crystal (by the use of plane waves as a basis set, for example).

Both cluster and slab geometries introduce artificial features: in a cluster calculation an ‘artificial surface’ is created at the cluster boundary, whereas in a slab calculation a semi-infinite crystal is represented by a slab typically only a few atoms thick, so one must ensure that the surfaces do not interfere with each other and also that possible long-range fields (electrical monopoles or dipoles, or elastic effects) do not propagate unintentionally from one cell to another, either parallel to the surface (in cases (b) and (c)) or perpendicular to the surface (in case (c) only). Broadly speaking, the cluster approximation is more acceptable in an ionic insulator where there is no breaking of bonds at the cluster boundary, whereas the slab approach is most suitable for metals, in which the geometrical and electronic influence of the surface decays very rapidly into the bulk. In a covalent semiconductor there are both strong bonds to be broken at a cluster boundary, and relatively long-range relaxations and electronic effects arising from the surface, so there are potential pitfalls with either approach.

The ideal solution is to perform a calculation which takes full account of the fact that the surface is the boundary of a semi-infinite crystal (Fig. 8(d)). This is relatively straightforward as regards long-range polarization and atomic distortion [109] by themselves; however, the question of the correct boundary conditions for the electronic wavefunctions is much more difficult, particularly in cases where there is a long-range component to the perturbation in crystal potential produced by the presence of the surface [110–112]. The applications of this type of method are therefore most advanced in metals, where the issue is at the same time most easy to resolve (because the screening length is short and there are no long-range fields) and most important (because the electrons are most delocalized); the methods
used to perform the matching include the embedding potential approach [110,113–115] and Green's function LMTO method [116], although in neither case has atomic relaxation near the surface been included.

This issue is particularly important when a calculation of electronic structure of a surface system is to be compared with an STM measurement. In the STM, it is precisely the fact that both tip and sample are connected to an effectively infinite reservoir of electrons in the external circuit that leads to a non-equilibrium situation with a nonzero current flowing, rather than simply the re-establishment of a local equilibrium under the applied bias.

4. Modelling electron tunnelling

We have finished our survey of the available theoretical models for the electronic structure of a surface. As we have seen, there are a very large number of techniques, some of them very mature, and
there is a long history of applications to calculate many different quantities. When we turn to the more specific field of calculating STM tunnel currents the field is altogether less mature, with far fewer well established techniques available. On the other hand the youth of this part of the subject, and the rate of experimental and theoretical developments, make it a very exciting field.

4.1. Perturbation theory

As a first step in the treatment of the STM, we start from our hope that the tip and the sample behave as two weakly coupled systems, and that it is therefore reasonable to perform perturbation theory in this coupling. Formally, we might hope that we can set up two Hamiltonians $\hat{H}_1$ and $\hat{H}_2$ which describe the uncoupled tip and sample, respectively, and write the total Hamiltonian when the tip and the sample are brought into proximity as

$$\hat{H} = \hat{H}_1 + \hat{H}_2 + V = \hat{H}_0 + V,$$

where $V$ is the ‘tunnelling perturbation’ and $\hat{H}_0$ is the Hamiltonian of the separated tip and sample, $\hat{H}_0 = \hat{H}_1 + \hat{H}_2$. We would then wish to perform perturbation theory in $V$, which would be small for a tip weakly coupled to the sample.

If we are working within a tight-binding formulation (see Section 3.3) this is not a problem — $\hat{H}_1$ and $\hat{H}_2$ can reasonably be taken as the separate tight-binding Hamiltonians of the tip and sample, and $V$ as the hopping matrix elements which couple an atom in the tip with an atom in the sample. The eigenfunctions of $\hat{H}_0$ can therefore be taken to be entirely localized in either the tip or the sample. However, if we wish to make a more complete solution of the Schrödinger equation, we run into trouble even at this stage. The obvious choices for $\hat{H}_1$ and $\hat{H}_2$ now appear to be the Hamiltonians for isolated tip and sample, i.e. for the respective surfaces with a semi-infinite vacuum region outside. However, this choice suffers from the serious problem that the ‘unperturbed Hamiltonian’ $\hat{H}_0 = \hat{H}_1 + \hat{H}_2$ does not correspond to any physical system; this is because, even if it is correct to superpose the potentials of the tip and sample systems, the kinetic energy operator appears twice and is therefore double-counted in $\hat{H}_0$.

4.1.1. The transfer Hamiltonian

The difficulty of constructing a single Hamiltonian which can be used as a starting point for perturbation theory pervades the whole subject. It was addressed many years ago by Bardeen [117], who was considering the many-body problem of superconducting tunnelling (indeed, the title of his paper was *tunnelling from a many-particle point of view*); nevertheless his discussion can easily be recast in language appropriate to single-particle tunnelling in the STM. His idea was to work with two sets of states, which in the STM problem are the eigenfunctions of an isolated tip and an isolated sample, recognizing that each set of states solves the Schrödinger equation correctly only in part of space. (He assumed, however, that there was some overlap between the two domains of validity, and therefore that there existed a region between the two surfaces within which the potential was vacuum-like.) Bardeen then asserted that the quantity

$$M_{ij} = \int dr \, \psi_i^*(\hat{H} - E_0) \psi_j$$

(22)
could be used in time-dependent perturbation theory to calculate the rate of scattering from state \( i \) to state \( j \) in the same way that the matrix element of a perturbation is used when discussing transitions between a single set of states all belonging to the same unperturbed Hamiltonian (see, for example, [118]) and therefore that the rate of tunnelling is given by Fermi’s ‘golden rule’ result

\[
\Gamma_{ij} = \frac{2\pi}{\hbar} |M_{ij}|^2 \delta(E_i - E_j).
\]  

Finally, Bardeen performed an ingenious integration by parts to show that \( M_{ij} \) could be written in the form

\[
M_{ij} = \frac{\hbar^2}{2m_e} \int_S d^2r [\psi_i^*(r) \nabla \psi_j(r) - \psi_i^*(r) \nabla \psi_j(r)],
\]  

where \( S \) is a surface running through the vacuum between the two systems. The quantity in square brackets can be recognized as being proportional to the matrix element of the quantum mechanical current operator taken between states \( \psi_i \) and \( \psi_j \), and \( M_{ij} \) in this form is generally known as the ‘transfer Hamiltonian’ between states \( \psi_i \) and \( \psi_j \).

Although Eqs. (22) and (23) appear very reasonable, they are not, in fact, at all easy to justify. This is because the states \( \psi_i \) and \( \psi_j \) do not belong to the same Hamiltonian and are therefore not orthogonal. This introduces extra terms into the time-dependent perturbation theory, some of which must be neglected to arrive at Bardeen’s result. Extensive work by Feuchtwang [119] and others [120,121] has shown that the transfer Hamiltonian is, indeed, a correct starting point for an expansion of the tunnelling in powers of the tip–sample interaction; we leave this until our discussion of non-perturbative approaches to tunnelling in Section 4.2.

### 4.1.2. The Tersoff–Hamann Approximation

The seminal work of Tersoff and Hamann [17,18] transformed the Bardeen formulae (23) and (24) into a very useful and widely used relationship between the tunnel current and the properties of the sample. They supposed that the following conditions are satisfied:

(i) perturbation theory in the tip–sample interaction is indeed appropriate;

(ii) there exists a region between the tip and the sample within which the potential is equal to its vacuum value, so the applied bias between the two systems must be zero (this is necessary in order that Bardeen’s construction for the transfer Hamiltonian can be used);

(iii) the predominant tip state involved in the tunnelling is an s-state.

In that case, they showed that the known asymptotic forms of the surface wavefunction and of the presumed s-state on the tip (taken here to be the state \( \psi_i \)) enable one to perform the integration in equation (24) explicitly. The result is

\[
M_{ij} = \frac{\hbar^2}{m_e \kappa} \Omega_{\text{tip}}^{-1/2} \kappa R e^{\kappa R} \psi_j(r_0),
\]  

where \( \Omega_{\text{tip}} \) is the normalization volume used for the tip wavefunction, \( R \) is the radius of the spherical potential well assumed for the tip, and \( r_0 \) is the position of the tip centre. The current flowing at a bias
which is small relative to the energy scale of features in the tip or sample density of states is then

$$I = A \sum_{\text{sample states } j} |\psi_j(r_0)|^2,$$

where we have absorbed all the tip-dependent and bias-dependent quantities into the constant $A$. The small-bias tunnelling rate is therefore directly proportional to the sample density of states at the Fermi energy, evaluated at the centre of the tip.

4.1.3. Beyond Tersoff–Hamann

The Tersoff–Hamann approximation is very convenient, and gives an appealing and simple picture of the STM’s operation. There are, however, situations in which one needs to include more details of the electronic structure of the tip to understand the experiments, while still remaining within the basic domain of perturbation theory. Usually these situations arise because the tip electronic structure has significant spatial modulation (rather than being a simple s-state as assumed by Tersoff and Hamann) or because the tip density of states shows significant dependence on energy within the tunnelling energy range.

An elegant example of this type of idea is provided by the work of Tsukada et al. [122,123], which is based on separate electronic structure calculations of the tip and the sample using the LDA and a linear combination of atomic orbitals in a cluster geometry. The Bardeen matrix element equation (24) was then evaluated between tip states and sample states as a function of the applied bias. This model was able to describe phenomena such as negative differential resistance at a boron impurity on the Si(111) surface [124], which arises because peaks in the tip and sample densities of states move out of coincidence as the bias increases. The calculations of Bass and Matthai [125] described in Section 6 are similar in spirit, except for the use of a periodic slab geometry and a plane-wave basis set.

Lawumni and Payne [126] and Tersoff [127] pointed out that the spatial dependence of the tip orbital can be very important at points on the surface where a selection rule would prevent any tunnelling to or from an s-state. For example, on the bare graphite surface there is no tunnelling at the Fermi energy when a state of s symmetry is held above one of the so-called ‘B-type’ atoms; a constant-current image computed at low bias within the Tersoff–Hamann approximation would therefore have an infinitely large ‘corrugation’ as the tip moves closer and closer to the B-type atoms in a hopeless quest for nonzero current. Indeed, very large corrugations are observed on graphite, but these are now believed to be due to mechanical interactions between the tip and the sample (see Section 4.4); Lawumni and Payne showed that only a very small admixture of a p-state on the tip is sufficient to reduce dramatically the corrugation arising from purely electronic effects.

Chen [128–130] has discussed the images that are produced if a d-state on the tip dominates the tunnelling; this case is very relevant in practice, since many of the materials commonly used for tips are transition metals with valence d-electrons. The images obtained are proportional to spatial derivatives of the images obtained with a purely s-wave tip.

4.2. Scattering theory

A quite different type of approach is necessary when perturbation theory itself can no longer be applied. This can happen in two different cases: the first is if the tunnelling becomes intrinsically strong, so the probability of a certain set of incident electrons making the jump from tip to sample is of
Fig. 9. Let transport occur by a tunnelling event between terminals P and L. For the use of perturbation theory to be valid, the tunnelling part of the transport process must be the least rapid; this is not necessarily the case if another slow transport process is involved.

In either case, a single treatment of the transport of electrons through the whole system (tip and sample) is necessary. It is well understood how to do this kind of calculation in principle; however, the small number of applications that have appeared testifies to the difficulties in practice. The earliest, and still in many ways the most complete, treatment of the problem predates the STM by some years and is due to Feuchtwang [119,134–136]. In these papers, the current is written in terms of the density matrix of the (nonequilibrium) system of tip and sample with different chemical potentials on the two sides; this density-matrix is then evaluated using the approach of Keldysh [137,138], which allows the nonequilibrium Green’s functions of the system to be calculated using a perturbation approach similar to those which apply in thermal equilibrium. The ‘zeroth-order’ system from which the perturbation expansion proceeds involves the separate tip and sample systems, with Green’s functions chosen to obey a homogeneous boundary condition involving their value and normal derivative on a surface separating the two. The complete perturbation series can be formally summed using a Dyson equation; the first term in the series, however, turns out to involve the Bardeen tunnelling matrix element.

order unity; this happens if, for example, a point contact is formed [131]. The other case arises if, even though the tunnelling is itself weak, it is nevertheless more rapid than some other part of the electron transport process; this could happen if, for example, there is a second barrier to the passage of the electrons, such as the presence of an insulating layer on one side of the tunnel junction, as in experiments involving self-assembled alkyl thiol monolayers on conducting substrates [132,133]. An analogy which may be familiar to European readers of this review is illustrated in Fig. 9; the Eurostar train makes slower progress wending its way along the picturesque track into London than it does while tunnelling under the Channel, so the overall rate of transit cannot be calculated from the tunnelling step alone!
A number of authors have pointed out that results equivalent to Feuchtwang’s can be obtained starting from different zeroth-order systems. Pendry et al. [121] took as their starting states the tip and the sample separated by a thin impermeable barrier which had the effect of forcing the wavefunctions of both to decay to zero; Todorov et al. [139] took a tight-binding model in which the hopping matrix elements connecting the tip and the sample are set to zero. In either case, ordinary perturbation theory can be performed to give Green’s functions of the full system and hence the current in terms of the $t$-matrix of the system:

$$I = \frac{e^2}{h} \text{Tr}[\rho t^1 \rho].$$ \hfill (27)

(An extra factor of 2 is sometimes included in this result; it is not needed if the ‘up’ and ‘down’ electrons are counted separately in the trace.) This is effectively the celebrated Landauer formula [140], which arises from consideration of the resistance introduced by the scattering of a stream of electrons [141]. The most complete formulation is due to Meir and Wingreen [142]; they showed that the result can be generalized to the case where arbitrarily strong electron–electron interactions are present in one region of the system, provided that the electrons are effectively noninteracting in the ‘leads’ which carry the current towards or away from the system. Their result reduces to Eq. (27) when the electrons are noninteracting.

Eq. (27) can also be obtained, with the aid of some lengthy mathematics [143], from the Kubo formula [144,145] which approaches the transport problem from a more traditional direction, by asking how much current is produced in the system (to first order) by a given applied potential. Interestingly, in view of the wide use of the result in other areas of physics, direct evaluation of the Kubo formula has been very little used in the theory of the STM. For systems of independent (or effectively independent) electrons such as those we consider here, the Kubo and Landauer approaches are equivalent. For strongly interacting systems, however, in which the quasiparticle excitations carrying the current do not closely resemble free electrons, there can be differences: the response of the charge within the system to an applied field is not then related so simply to the transmission of electrons arriving from outside [146].

While (resummed) perturbation theory is used as a tool for deriving these expressions, it is not necessary in actual calculations. Thus in the method developed by Joachim and Sautet, which was originally applied to elastic electron transport through organic molecules [147,148], the whole system (including the tip) is treated together on the same footing from the outset. A simple tight-binding type Hamiltonian is used, with the matrix elements representing the tunnelling through the vacuum fitted to ensure a reasonable exponential dependence of the tunnelling on the separation. A set of transmission and reflection coefficients is then calculated for the successive layers of the system (including the vacuum), and from these the $t$-matrix for the entire tunnel junction is calculated and the current is found using Eq. (27). There have been many applications of this technique, particularly to problems involving molecules on metal surfaces [149–151]. It has the great advantage that the contributions of the different molecular states to the tunnelling can readily be separated [152,153]; however the basis set that is usually used is somewhat restricted, and the method is not at present combined with a reliable total-energy technique which enables the positions of the atoms on the surface to be predicted. There is no reason in principle, however, why this method should not be combined with one of the tight-binding total energy schemes; it would then be interesting to see whether the same tight-binding parameters were able to describe both the chemical binding and the tunnelling.
The nearest approach to this scheme in the context of an ab initio electronic structure calculation is provided by the work of Hirose and Tsukada, who solved LDA problem of a silicon tip and a silicon surface self-consistently with open boundary conditions for the electrons [154,155]. This involved a calculation of transmission and reflection coefficients on a real-space grid self-consistently with the charge density.\(^8\) The system studied involved atom transfer between the tip and the surface; the total energy and electron transport were therefore computed as a function of the position of one atom across the tunnel junction. However no forces were calculated, so no relaxation of the other atoms was possible; in another approximation, the asymptotic states in the tip and the sample were taken to propagate in jellium, rather than in silicon. Despite these limitations, this calculation represents the state of the art in incorporating a genuine current flow into a first-principles calculation.

It is also possible to perform an exact scattering calculation by starting from a soluble model of the combined tip–sample system. A very influential early calculation of this type was that of Lang [156], who studied the crossover from tunnelling to point contact formation in the STM observed by Gimzewski and Müller [131]. Lang started from the solution of the jellium–vacuum–jellium problem [157] and solved the Lipmann–Schwinger equation for scattering from the self-consistent LDA potential produced by an adsorbed atom on one electrode. Subsequently, Lucas et al. [158] studied a similar junction with the inclusion of the image potential for the tunnelling electron, and calculated the scattering from a hemispherical protruberance on one electrode. Such an approach has been carried furthest towards real systems by Doyen and co-workers [159,160], who began from a pair of planar metal surfaces whose electronic structure could be calculated using the KKR bandstructure approach and the layer-doubling method originally developed for low-energy electron diffraction problems [161]. They then included the scattering from a localized potential, representing a tip protrusion, added to one of the surfaces.

Finally, we come to a class of approaches that concentrate on the local properties of the STM tunnel junction. The most rigorous mathematically is that of Noguera [120,162], who transformed the scattering problem to one contained wholly within the tunnel junction; the rest of the system is represented by effective potentials at the tip and sample ends. The relationship between this method and the techniques based on the scattering of electrons from a perturbation is reminiscent of that between scattering theory and ‘embedding’ methods in the theory of defects or surfaces [110,113]; both are exact representations of the same physics. In Noguera’s discussion the vacuum is treated separately, and the embedding is performed at either end of the vacuum gap; there is no reason in principle, however, why her approach cannot be extended to treat any desired part of the tip–sample system.

The difficulty in practical applications is still the coupling of the electronic structure of the region near the tunnel junction onto the bulk. The movement of charge across this boundary is apt to cause problems in a self-consistent calculation. This has led to suggestions that the embedding process could be simplified, with the introduction of an artificial ‘source’ introducing electrons at one chemical potential on one side of the junction, and/or a ‘sink’ removing them at a lower chemical potential on the other side. Molotkov et al. proposed adding an imaginary part to the Hamiltonian to represent removal

\(^8\)In principle, the current density should also be included in such a nonequilibrium situation; however, the corrections that arise from this are believed to be small.
of electrons from the ‘drain’, and evaluating the current from the imaginary part of the eigenvalues associated with states on the other side of the junction [163,164]. This presupposes, of course, that all states can be definitely associated with one side or other of the junction – this will not be the case if the tunnelling is strong. An alternative approach [165] is to calculate explicitly the rate at which electrons propagate from the source to the sink, without using any perturbation theory. This can be done by a calculation of Green’s function of the full tip-sample system [165,166], and is effective for arbitrarily strong coupling of the tip and the sample. This approach has the advantage that it can easily be combined with efficient ab initio calculations of the electronic structure of the system around the tunnel junction, from which quantities such as forces can be efficiently calculated; however, there is no unique recipe for choosing the initial and final states, and one’s choice will not in general reproduce correctly the injection of electrons into the system from the leads attached to the external circuit. One must therefore test carefully for the dependence of the results on the initial and final states, and must be wary in interpreting the absolute conductance predicted by the method, since it will not include terms such as the spreading resistance correctly.

4.3. Educated intuition

Despite the work that has gone into quantitative theories of electronic structure and of tunnel current, much work on STM image interpretation still relies on what one might call ‘educated intuition’. The Tersoff–Hamann model relates the image to a density of states; if one has a picture of the local surface physics and chemistry which tells one the shape of the density of states, one can therfore make deductions about the images. One example is in the dark line running down the centre of the dimer rows on the Si(001) surface in empty-states images, and its relationship to the node in the antibonding wavefunction for the dimer (Section 5).

In using such intuition, it is important to remember just how quickly a wavefunction decays with distance at a typical surface – a factor of 10 decrease in probability density for each Ångström of distance is an often-quoted rule of thumb. It therefore requires quite strong electronic effects to counteract even relatively small geometrical ones, such as the fact that one atom lies above another on the surface. This is illustrated by the case of C₂H₄ on Si(001) considered in Section 5: the hydrocarbon molecules appear darker than the neighbouring Si dimers in the filled-states experiments, and one might expect that this could be explained by the lesser weight of the density of states at the top of the valence band on the adsorbates relative to the substrate. But the detailed calculation shows this is not the case; in fact, the inclusion of a much more subtle extra electronic effect (the distortion of the surface electronic structure by the tip-induced electric field) is necessary to overturn the geometric considerations and obtain agreement with experiment.

4.4. Tip–sample interactions

While it is very easy to become absorbed in the subtleties of quantum-mechanical transport theory and its application to the tunnelling problem, one should not forget that there will also be strong mechanical interactions between the tip and the sample, and these can be at least as important as electronic effects in determining the actual image. An early example of this was seen in imaging on graphite [167], where very large ‘atomic’ corrugations were seen (as much as several hundred Angstroms in magnitude) which could not possibly arise from the true topography of the surface. While
electronic effects indeed appear to give a large corrugation when treated within the Tersoff–Hamann approximation with an s-wave tip, the electronic corrugation is significantly reduced when the non-s-wave component of the tip wavefunction is accounted for (see Section 4.1.3). The experiments can be understood by appreciating that graphite is very compressible perpendicular to the atomic layers; as the tip ‘tries’ to move towards the surface to keep the tunnel current constant, the surface recedes because of the additional force exerted on it.

Quantitative investigations of the tip–sample interactions of this kind became possible with the development of the atomic force microscope [168], and particularly of combined STM/AFM instruments which enable the force to be measured during scanning. The tip–sample force has a profound effect on the magnitude of the corrugation as a function of height, and also on experimental attempts to measure the tunnel barrier height by watching the decay of the current as a function of tip–sample distance [169]. These points were also previously emphasized from a theoretical point of view by Baratoff and collaborators [170].

A particularly extreme form of tip–sample interaction occurs when a physical contact is formed between the tip and the sample. Even before this occurs, the barrier for electron transfer between the two materials has disappeared; a unified non-perturbative treatment of electron transport through the system is then essential [139]. As the contact is broken through the withdrawal of the tip, there is a complex interplay between the mechanical properties of the resulting ‘quantum wire’ as it contracts and the electronic properties that determine the electron transport [171–173], resulting in stepwise reductions of conductance as the wire contracts.

5. The Si(001) Surface

The Si(001) surface is fairly reactive. At room temperature hydrogen in molecular form does not react, but atomic hydrogen does react. At modest exposures one hydrogen attaches to each surface silicon atom, to form a monohydride phase. The hydrogen atoms attach to the dangling bonds, leaving the dimers still bonded and the (2 × 1) reconstruction still remaining. At higher exposures a dihydride phase forms, in which two hydrogen atoms attach to each surface silicon atom, destroying the dimer bonds and causing the surface to revert to a (1 × 1) periodicity. Other structures can be seen at different exposures of atomic hydrogen, for example the (3 × 1) periodicity. Adsorption and desorption of hydrogen atoms on monohydride surfaces occurs by a pairing mechanism, in which atoms seem to be able to move relatively freely on the surface until they find a mate. At all but the lowest submonolayer coverages dimers generally have two adsorbed hydrogen atoms or none. The study of adsorbed hydrogen provides an excellent foundation for investigating growth by gas source molecular beam epitaxy (GSMBE) from hydride sources. But for our first example of the relationship between STM experiment and theory we have chosen molecules which chemisorb on Si(001) without dissociating at room temperature.

5.1. Clean Si(001) surfaces

In order to model adsorbates on a surface, it is essential first to understand the structure of the clean surface, and so we begin by describing the results of some ab initio atomic and electronic structure calculations on Si(001). The calculations were performed using the Car–Parrinello algorithm and the
PAW approach to the description of the wavefunctions, in a $(4 \times 2)$ unit cell, with the experimental cubic lattice constant of 5.43 Å, containing eight surface atoms which in the undistorted diamond lattice are equivalent. Four layers of silicon atoms were used; the height of the unit cells in this direction is 10.86 Å. This allows a vacuum gap of approximately 6.9 Å between periodic images of the silicon slab, which is ample to allow the inclusion of an adsorbed hydrocarbon without introducing any significant interaction between the slabs in the total energy calculations. A plane-wave cutoff of 25.0 Rydbergs was used for the electron wavefunctions and only the $I$-point of the reduced Brillouin zone was sampled. The energy cutoff is considerably more than required for calculations on silicon, but is necessary to give a good description of the adsorbates.

5.1.1. Perfect dimerised surfaces

The modelling of the bare surface serve both as a check on the method of calculation and as a starting point for assessing the effect of the adsorbates. As expected, the calculations give a surface reconstruction in which neighbouring surface Si atoms move towards each other to form dimers. The nearest-neighbour distance between atoms in the surface layer is reduced from 3.84 Å in the unrelaxed diamond structure to approximately 2.3 Å, and the dimers tilt through an angle of about 15°. The tilting of individual dimers agrees with other LDA results [174], which were carefully converged with respect to both plane-wave cutoff and slab thickness. Those calculations gave rather deep minima of the potential surface around the equilibrium dimer tilt angles (about 0.1 eV per dimer), suggesting that it is a reasonable approximation to treat the dimer as a two-state system that can be described by a variable $s$ which takes the value +1 for left tilt and −1 for right tilt. When the coupling between neighbouring dimers along a row deduced above is included, the effective Hamiltonian for the dimer chain becomes that of a one-dimensional antiferromagnetic Ising model:

$$ H = J \sum_i s_i s_{i+1}. $$

The spin–spin correlation function of such a system takes the form [175]

$$ \langle s_i s_{i+n} \rangle = (-1)^n \exp(-nK), $$

with

$$ K = \log[\coth(J/k_B T)]. $$

(An earlier expression for $K$ [19] corresponds to a type of mean-field approach to the Ising system which underestimates the correlation length at low temperatures.) For example, for a coupling of $J = 30$ meV we have a correlation length $K^{-1} = 5$ dimers at 300 K.

Generally, no tilt is observed in the dimers in STM experiments away from surface defects [19,176,177]. It is conceivable that this might be a result of structural changes induced by the STM current [178] or by mechanical interaction with the tip [179], but an alternative explanation involves a rapid flipping of the dimers back and forth between the lowest-energy configurations which averages out the tilt on the timescale of an STM measurement. However, as in Fig. 3 zigzag structures of oppositely tilted dimers are often observed along dimer rows [19,180]; it is believed that these are related to the presence of pinning centres [181] which force a dimer preferentially into one tilt direction. We would expect the decay of the tilt asymmetry along a dimer row away from such a pinning centre to follow the exponential law of Eq. (29). Observations at low
temperature [180] show zigzag dimer rows with much longer decay lengths, which would tend to confirm this account.

Another clean Si(0 0 1) surface which was subsequently used to study the adsorption of ethene will be shown in Fig. 12(a) in Section 5.2 below [182]. Like Fig. 3 it shows two terraces with a step a quarter of a unit cell high between them. There are several places in which the dimer tilt appears to have been pinned by a defect; the rate of decay of the zigzag appearance can be compared with the expression given in Eq. (29) in Section 5 above. The dimer tilts appear with considerably stronger contrast in the filled-states image than in the empty-states image. As explained in Section 2.2.1, in a tilted dimer there is a concentration of the filled $\pi$-state on the *up* atom and a concentration of the empty $\pi^*$ state on the *down* atom, so that in the filled-states image the geometric and electronic effects reinforce one another, enhancing the topographic contrast of the tilted dimers, whereas in the empty-states image, the electronic effect partially compensates the topographic effects, leading to weak contrast which is scarcely visible even where we know from the filled-states image that the dimers are pinned. There is another important difference between the filled- and empty-states images. On the upper terrace in the filled-states image a number of defects can be seen which plainly correspond to one or more missing dimers on a given row, and they are aligned with the bright stripe that corresponds to the filled $\pi$ dimer states along the row. In the empty-states image the same defects are aligned with a dark stripe. This is the node in the empty $\pi^*$ state, which occurs in the middle of the dimer bond. Thus in a filled-states image the dimer row is characterized by a broad bright stripe, whereas in an empty-states image it is characterized by a narrow dark stripe. As explained in Section 5.2.1, this is helpful when identifying adsorbed molecules in a pair of filled- and empty-states images.

### 5.1.2. Missing-dimer defects

Defects in surfaces can have a crucial effect on the behaviour of molecules. For example, on surfaces such as TiO$_2$(110) many common molecules such as water react only at defects, and not at all on a perfect region of a surface. A great benefit of studying the adsorption of molecules on surfaces by STM lies in the ability to observe the behaviour in relation to particular sites on the surface, and thus to identify site-specific effects. In Fig. 10(a) considerable number of missing dimer defects are present. The nature of single dimer defects in Si(00 1) surfaces has been identified by careful comparison of STM images and modelling [183,184]. The key steps were to observe the voltage-dependent contrast from single dimer defects by STM, and then to observe that one, and only one, of the possible candidate defect structures would be expected to give contrast consistent with what was observed.

Images of a Si(00 1) surface studied in this way are presented in Fig. 10 [184]. Both pictures show filled states in the sample, but they are at different bias. Fig. 10(a) was taken at a sample bias of $-1.0$ V. It contains plenty of defects with one or more dimers missing, known as dimer vacancies (DV). The contrast of all these defects is similar at this bias voltage: in each case the adjacent dimers have normal contrast right up to the defect, and the defect then appears as a pit in the surface. It has been found from scanning tunnelling spectroscopy [185] that 1-DV’s (single-dimer vacancies), and also 2-DV’s, exhibit a reduced bandgap relative to the bandgap of a perfect surface. Fig. 10(b) was taken at a sample bias of $-0.6$ V. Away from defects the dimer rows are similar in appearance to Fig. 10(a). But in the vicinity of the defects there are differences. Where a defect consists of several missing dimers (for example, in the lower left circle in each picture), the adjacent dimers appear darker than normal. But the dimers next to a single-missing-dimer defect (for example, in the other circle in each picture) look bright.
Candidate structures for the single-missing-dimer defect were modelled using a two-stage approach. The atomic positions were first found by a linear scaling tight-binding program (see Section 3.3), using established parameters. A \((4 \times 6)\) unit cell was used, with eight layers of silicon, making a total of 190 atoms, with periodic boundary conditions and the lower surface terminated by a \((4 \times 2)\) reconstructed surface. An even number of dimers along each dimer row is necessary in order to permit alternating tilt. Having found the relaxed atomic positions, the minimum-energy electronic configuration was then found by a local density method with pseudopotentials and a plane-wave basis set (see Section 3.2.4). This calculation was performed for a \((4 \times 4)\) unit cell six layers thick, making 94 atoms altogether, with a vacuum gap of a further six atomic layer spacings separating the periodic slabs, and with a single \(k\)-point \((0,1/4,0)\) and a planewave cutoff of 115 eV. The final energies and charge densities were expressed in terms of the Kohn–Sham eigenvalues and eigenstates. The perfect surface had two more atoms in each case.

The perfect surface and three candidate structures considered for the I-DV are illustrated in the left column of Fig. 11 [184]. The first structure is the perfect dimer reconstructed surface. In the second structure the four second-layer atoms nearest to the missing dimer are bonded to form two bonds parallel to the dimer row (i.e. perpendicular to the dimer bonds). This eliminates all the dangling bonds, at the expense of some additional strain energy. Only one of these bonds is present in the third possibility, resulting in an asymmetrical tilted and twisted structure in which the positions of the atoms on either side are not equivalent. In the final structure both of the bonds between the second layer atoms next to the missing dimer are broken, releasing some strain energy but leaving four dangling bonds. The first and third structures had been postulated in an earlier theoretical study [186], in which the bonded structure was calculated to be 0.42 eV lower in energy than the broken structure, but it was not possible to predict the bandgap of either structure. Calculations performed using the two-
stage approach outlined in the previous paragraph also found that the bonded structure was the lowest in energy, being 0.45 eV lower than the broken structure and 0.48 eV lower than the twisted structure (comparison with the perfect surface is not straightforward because of the different number of atoms involved). In all the modelled structures the dimers adopted one of two possible tilted configurations.

To make detailed use of the information in the voltage-dependent images, it is necessary to be able to simulate the STM images that would be expected from each of the structures. Methods for calculating images with various degrees of sophistication were introduced in Section 4, and their use will be illustrated for the adsorption of C₂H₄ on Si(0 0 1) in Section 5.2. For the present comparison, a simpler educated-intuitive approach was taken (motivated by the Tersoff–Hamann approximation, even although the conditions for its applicability are not strictly met): each eigenstate in the sample was assumed to make a contribution to the tunnelling current proportional to the availability of empty states in the tip at the corresponding energy, the bias being represented by an equivalent offset between the Fermi levels of tip and sample. The distribution of states in the tip is represented by a uniform density, with occupation described by a Fermi function with \( kT = 0.025 \) eV. Thus, with some smearing due to
temperature, states in the sample will contribute to the tunnelling current from the Fermi level down to an energy below it equal to the bias voltage. The difference in tunnelling probability at different electron energies is neglected; for incident electron energy differences of a few tenths of an eV the tunnelling probability through a barrier several eV high and about 1 nm wide may be 10% or more, and more exact calculations would allow for this. Nevertheless the present method allows a simulated STM image to be directly visualized without significant further calculation, and is useful for evaluating the plausibility of the different candidate structures. Pinned zigzag dimer structures are not generally seen in the vicinity of 1-DV's. Since in the modelling the dimers adopt a particular tilt, alternating from one dimer to the next, it is helpful to take the average of two complementary tilt configurations, corresponding to changes of tilt within a given experimental observation. Simulations are then presented as the height of an ideal point tip required to maintain a given constant tunnelling current; the height is displayed in an image as brightness, just as in experimental STM images.

Such simulations for sample bias voltages of $-1.0 \text{ eV}$ and $-0.6 \text{ eV}$ are presented in columns (ii) and (iii) of Fig. 11, for direct comparison with Figs. 10(a) and (b). The simulations at $-1.0 \text{ V}$ in column (ii) show little difference between the three candidate structures, though the broken structure does show slight darkening of the dimers next to the missing dimer. All of them are compatible with Fig. 10(a). At this bias the tunnelling current seems to be dominated by the normal dimer surface states, and to be insensitive to particular details of the defect. The situation at $-0.6 \text{ V}$ sample bias is quite different. At this voltage the contribution from regular surface states is greatly reduced (Figs. 11(a)(ii) and (iii) are separately normalized), and so the contrast shows enhanced sensitivity to states within the perfect surface band gap. The broken structure shows reduced contrast next to the defect, which is the opposite of what is seen experimentally. The twisted structure shows enhanced brightness on one side of the missing dimer and reduced brightness on the other side (there is asymmetry in the structure which is not represented in Fig. 11(c)(i)). The single bond in the twisted structure might also lead to dimer tilt pinning, which was consciously removed in the preparation of the simulated images, and which would appear as zigzag dimer structures as in Fig. 12(a). If this is correct, then the fact that zigzag structures are not present in Fig. 10 is a further argument against the twisted structure. The bonded structure alone gives enhanced brightness on either side of the 1-DV. This is consistent with the lower energy found for this structure, which tends to confirm that it is indeed the right one. Thus it is apparent that using the experimental information from STM images at different voltages, together with rather simple image simulations from good atomistic calculations, makes interpretation of the detailed bonding arrangement in a surface structure possible with more confidence that would be possible from topographic images and geometrical calculations alone.

Is it a general principle that greater strain in a semiconductor leads to greater contrast near the Fermi level? The broken structure has more dangling bonds, but those bonds that are formed are able to relax with less constraints. On a simple bonding picture this would lead to a greater splitting in energy between the bonding and the antibonding states. In an n-doped semiconductor the bonding states are filled, but if they are further from the Fermi energy they will not contribute at low enough negative bias voltages. In the bonded structure the geometry suffers an additional constraint since the atoms with the additional bonds are pulled towards one another. Some of this strain must be transmitted to the dimer $\pi$-bonds responsible for the STM contrast, reducing the bond-antibond energy splitting and thus enhancing the contrast at low bias voltage. If this reasoning is valid, then the structure of defects like the one in the lower left circles in Fig. 11 should be less constrained. Similar arguments also apply to the structure of steps on Si(001) [183].
5.2. C$_2$H$_4$ and C$_2$H$_2$ on Si(001)

Extensive studies have been made of the adsorption of ethene (or ethylene, C$_2$H$_4$) and ethyne (or acetylene, C$_2$H$_2$) on Si(001)-(2 × 1) surfaces, and of subsequent desorption [187,188]. Such measurements indicated a saturation coverage of 0.80 ± 0.06 C$_2$H$_4$ molecules per Si$_2$ dimer site on the surface: if the surface had 20% defective dimer sites then this would correspond to one adsorbed molecule per perfect dimer. Experiments with hydrogen capping of the silicon sites gave a fall-off in adsorption which behaved as $(1 - \theta)^2$, where $\theta$ is the fractional hydrogen coverage [13], which indicates that two adjacent Si sites are needed for the adsorption of a molecule (and also, incidentally, that individual hydrogen atoms pair up upon adsorption at all but the lowest coverages). Similar post-hydrogenation experiments with C$_2$D$_4$ and atomic H confirmed all these results, and showed no isotopic exchange between hydrogen and deuterium. The conclusion from these and other observations is that the adsorption of ethene and ethyne occurs by di-sigma attachment to silicon dimers, i.e. each carbon atom has a sigma bond to a silicon atom.

Although the adsorption properties of ethene and ethyne are very similar, the desorption of the two molecules is quite different. Temperature-programmed desorption (TPD) studies have been performed on Si(001) surfaces exposed to each of the gases [187,188]. In the case of ethene, desorption begins at 550 K, and is almost complete by 600 K, depending on the initial coverage. There is scarcely any dissociation of the ethene on the surface (no more than 2%), and almost all the adsorbed molecules come off as C$_2$H$_4$; the absence of residual carbon on the surface has been confirmed by Auger analysis. From the TPD measurements the energy of adsorption, extrapolated to zero coverage, is deduced to be 1.65 ± 0.07 eV per molecule (38.0 ± 1.5 kcal mol$^{-1}$). In the case of ethyne the temperature for desorption is higher, but less than 5% of a monolayer is desorbed as C$_2$H$_2$. The remainder comes off as H$_2$, leaving carbon on the surface. The adsorption energy of ethyne cannot therefore be deduced from temperature-programmed desorption, but it is certain to be somewhat higher than that of ethene.

5.2.1. Observations and energetics

Surfaces exposed to ethene and ethyne can be studied by scanning tunnelling microscopy [19,182]. It is possible to image a clean silicon surface, withdraw the tip by 1 μm to reduce shadowing effects, give a controlled exposure of ethene, and examine the same area again. Images can be obtained with the sample biased both negative and positive relative to the tip (±2 V), thus imaging both the filled and the empty states. A series of pairs of such images is presented in Fig. 12 [182]. The images were obtained in an Omicron UHV STM system. The filled- and empty-states images were obtained during the forward and reverse line scans of a given frame. Between each pair of images the surface was exposed to a nominal 2 L of ethene (an exposure of 1 L has a pressure-time product of 10$^{-6}$ Torr s, which gives about a monolayer coverage if all the incident gas molecules stick to the surface); the absolute exposure is difficult to calibrate and since ethene has a sticking coefficient close to unity on Si(001) it must have been somewhat less than that. The total height range in the images of Fig. 12 is about 0.15 nm, much of which can be attributed to the step. Fig. 12(a) shows the clean Si(001) surface before it was exposed to ethene. Because of the difference between the $\pi$- and the $\pi'$-orbitals, the dimer rows appear quite different at negative and positive biases. In the filled states image there are dark lines between the dimer rows. In the empty states images a dark line appears along the centre of the dimer row, corresponding to the node in the $\pi'$-antibonding orbital. The empty states of the dangling bonds of adjacent rows almost seem to merge into one another, so that in the empty-states image it is scarcely possible to see any gap.
between rows. These differences can be confirmed by comparing the relative positions of missing dimer defects in the filled- and empty-states images: in the former they are aligned with bright strips whereas in the latter they are aligned with dark strips.

A surface after a nominal 2L exposure of ethene is shown in Fig. 12(b). In the filled-states image there are many more dark sites than there were on the clean surface in Fig. 12(a). How can you tell which are defects and which are adsorbates? One way to tell is to compare the exposed and the clean images directly. Because the same areas are imaged, it is possible to see which of the dark sites on the exposed surface was imaged as a normal dimer on the clean surface. A statistical method has been developed for describing the effect of neighbouring occupied sites on adsorption [189]. A second way is to compare the filled- and empty-states images of the exposed surface. It happens that in the empty-states image the appearance of the molecules is rather similar to the filled-states image, but because of the difference in the contrast of the dimers they now appear as a spot along an otherwise dark strip, and so they can be unambiguously identified. One such site is marked; there are plenty of others. From all the previous evidence from other surface science techniques the images of the adsorbates are consistent with their identification as complete ethene molecules di-c bonded above silicon dimers, but there is a counter-intuitive aspect. If the molecules are adsorbed over the dimer, should we not expect the tip to be higher over them, so that they should appear brighter than the clean dimers in STM images? Of course an STM images a combination of electronic and geometrical properties, and not geometrical structure alone, but the C–C bond will be some 0.2 nm higher than the clean dimer bonds, and it is not obvious that it should have a density of states 2V below the Fermi level two orders of magnitude less than a Si–Si bond. A full account of this phenomenon will be presented in Section 5.2.3; it cannot be explained without including electrostatic effects, and it turns out to be because the clean dimers are much more polarizable than the molecules sitting on them.

Further exposure led to approximately 50% coverage, as seen in Fig. 12(c). At this coverage a remarkable phenomenon is seen, which had not been detected by any diffraction technique. The rows of adsorbed molecules appear like pearls on a string, with a spacing equal to the spacing between dimer rows, indicating that the molecules are on alternate dimers along a row. From the pictures at this coverage alone it would be hard to know which are the molecules and which are the clean dimers, but from Fig. 12(b) we can deduce that the pearls are clean dimers and the knots between the pearls are ethene molecules. In some places molecules are lined up on adjacent rows, forming 2 × 2 domains; in other places the occupied sites alternate between rows, forming c(2 × 4) domains. There are roughly equal areas of each of these, and the domains are quite small, which suggests that the interaction between rows is not nearly as strong as the effect which leads to alternate site occupation along a row. By a curious coincidence, in the empty states image the contrast from adsorbed molecules is about the same as the contrast from a pair of dangling bonds from clean dimers in adjacent rows, which leads to the fine checkerboard patterns seen in the 2 × 2 domains in Fig. 12(c).

Fig. 12. As series of filled-states, \( V = -2 \text{ V} \), and empty-states, \( V = +2 \text{ V} \), images of an area of Si(001) after successive exposures to ethene; \( I = 1 \text{ nA} \). The notional exposures (uncalibrated) were: (a) 0L; (b) 2L; (c) 4L; (d) 6L. The same step profile can be seen in each picture, and the antinode in the \( \pi^*\)-orbital means that in the empty-states images the centre of the dimer rows appears dark, whereas in the filled-states images a dark line is seen between dimer rows; the adsorbed molecules appear similar in both filled- and empty-states images. Some pinned dimers are present in the filled-states image of the clean surface, with a zigzag appearance which decays exponentially away from the pinning sites. At coverages of up to half a monolayer of ethene the molecules prefer to sit on alternate sites along a dimer row, and they appear darker or smaller than the clean dimers. At higher coverages they are forced to occupy adjacent sites, and they then appear brighter [182].
The result of higher exposure still is shown in Fig. 12(d). Hitherto empty sites between occupied dimers begin to fill in, leading to coverage greater than 50%. Where two or more adjacent sites are occupied, there is a contrast reversal and the molecules now appear bright relative to the clean dimers; several cases of this are present in Fig. 12(d). If a surface which has been exposed to ethene is heated until some, but not all, of the ethene is desorbed (this can be monitored by checking for particles with mass/charge = 28m_p/e in a mass spectrometer), then pictures similar to Fig. 12(b) are seen. This is entirely consistent with previous observations that ethene adsorbs and desorbs intact on Si(001)-(2 x 1) [187].

Turning now to theory, there have been several studies of the adsorption of small hydrocarbons on Si(001), mostly using semi-empirical quantum chemical techniques (see Section 3.1.4). MNDO cluster calculations for C_2H_2 [190] and C_2H_4 [191] suggested that a ‘bridge site’, in which the molecules bridge two surface atoms in neighbouring dimers, is the most stable, with heats of adsorption of 3.9 and 5.12 eV respectively, while the AM1 technique gave an adsorption energy of 3.99 eV for C_2H_4. In these calculations no relaxation of the substrate atoms from their bulk positions was permitted, either in the presence of an adsorbate or in the case of a clean surface. Cluster calculations using a non-self-consistent molecular orbital technique have also been performed, in which the total energy is expressed as the sum of a one-electron Hückel-theory part and a repulsive interaction [192]. Once again no relaxation of the substrate atoms was permitted, although this time their positions were fixed according to the symmetric-dimer model of the surface reconstruction (the best ab initio calculations [174,193] suggest that for a clean surface this structure is not the most stable one). In the case of C_2H_2 [194] it was found that adsorption occurs atop a silicon dimer, with the carbon–carbon bond stretching by 0.18 Å. The adsorption energy in this case was found to be 1.2 eV per molecule. For C_2H_4, the same technique gave a similar adsorption site [195] with an increase in the carbon–carbon bond length by 0.25 Å and an adsorption energy of 1.0 eV.

We modelled the structure of adsorbed ethene molecules on Si(001)-(2 x 1) using the PAW-based method described for the clean surface in Section 5.1 [196]. The same (4 x 2) surface unit cell was used, giving four dimers and allowing both alternate site adsorption and adsorption on every site along a dimer row to be modelled, with 2× or 4× periodicity in the perpendicular direction. The most stable position for adsorption of a C_2H_4 molecule proves to be directly over and approximately parallel to a dimer. The geometry is illustrated in Fig. 13(a) [196]. The carbon–carbon bond lengths from 1.33 Å in the free molecule to 1.53 Å, which is very close to the bond length of the saturated hydrocarbon C_2H_6. Thus the carbon–carbon double bond is effectively reduced to a single bond, and the electrons occupying the bonding π state of the free molecule now form bonds (of length 1.95 Å) to the surface silicon atoms. This in turn weakens the dimer bond, although the lengthening of the silicon–silicon distance is rather small (from 2.30 to 2.39 Å); this point is discussed in more detail in Section 5.2.2 below. The tilt of the dimer is reduced from around 15° on the clean surface to 7°. The carbon–hydrogen bonds of the C_2H_4 molecule are tipped upwards away from the surface: the H–C–H bond angle is 106.7° and the H–C–C bond angle is 113°. The reduction of both angles from 120° reflects the transformation of the carbon atoms from sp^2 to sp^3 hybridization.

The adsorption energy for a single C_2H_4 molecule in the 4 x 2 unit cell is 1.57 eV. Within the periodic boundary conditions, this corresponds to a surface coverage of 0.25 molecules per dimer, with alternating occupied and empty sites both along and perpendicular to the dimer rows. The adsorption energy is in good agreement with the result of 1.65 ± 0.07 eV from thermal desorption measurements. The adsorption energy for a second molecule at an adjacent site along a dimer row is lower, only
1.07 eV. This indicates a repulsion between two nearest-neighbour C$_2$H$_4$ molecules along a dimer row of 0.25 eV, since each molecule has two neighbours in the calculation. This cannot be explained simply by a direct steric repulsion between the C$_2$H$_4$ molecules: the direct interaction energy of the molecules in the same positions and with the same unit cell and plane-wave cutoff gives a repulsion of only 0.04 eV (0.02 eV per nearest-neighbour pair).

The repulsion between the adsorbates must therefore be surface-mediated. There are several mechanisms that might contribute to such a repulsion. One important term is the elastic interaction that occurs through the distortion of the substrate by the molecules [197]. Removing the adsorbed molecules from the dimer sites, while retaining the substrate geometry induced by the adsorbate, gives an estimate of the total elastic distortion energy accompanying the adsorption; this yields an energy difference of 1.03 eV for the entire cell, containing both adsorption sites. On allowing the atoms to relax, but constraining the dimer atoms at one of the adsorption sites to remain in the same relative positions that they had in the presence of the adsorbate, the strain energy is reduced to 0.42 eV. This gives the elastic interaction energy as 1.03 – 2 × 0.42 = 0.19 eV per unit cell, or 0.10 eV per nearest-neighbour pair, and suggests that while elastic interaction may account for a substantial part of the coupling between the adsorption sites, it is not the only effect. The origin of this relatively large elastic interaction might be associated with the removal of the tilt in the dimer above which adsorption takes place; this seems reasonable in view of the energy difference of about 0.1 eV per dimer [174] found between the tilted and untitled dimer structures.

Another possible interaction mechanism involves electrostatic effects. The Si–C bond is believed to be polar, so one would expect some charge transfer on adsorption. Taking the dipole moment $\mu = 0.735$ Debyes of the methylsilane molecule (CH$_3$SiH$_3$) as an indication of the polarization of the bond yields a repulsion of 0.07 eV per dimer pair at a separation of 3.8 Å. Once again, this seems too small to explain all of the observed effect; however, in combination with the elastic energy contribution and the direct molecular interaction discussed above, it gives a total interaction of the correct order of magnitude. The remainder of the interaction may be due to local changes in the chemical bonding which do not have simple classical interpretations.

The adsorption of ethyne on Si(001) has been modelled by the same method [196]. The most favoured adsorption site for C$_2$H$_2$ is similar to that for C$_2$H$_4$: the molecule sits atop a surface dimer. The geometry is illustrated in Fig. 13(b). Once again the carbon–carbon bond of the molecule lengthens, from 1.21 Å in the isolated molecule to 1.37 Å. This is close to the equilibrium bond length of C$_2$H$_4$ and reflects a weakening of the molecular bond as a result of adsorption, this
time from a triple bond to a double bond. The hydrogen atoms are once again tipped up away from the surface, reducing the H–C–C bond angle from 180° in the free molecule to 123°. The H–C–Si bond angle is 132°; this accords with the view that the carbon atoms are approximately sp² hybridized.

The adsorption energy in this case for a single molecule in the (4 × 2) unit cell of the calculation (i.e. for adsorption on every other site along a dimer row) is calculated to be 2.38 eV. This cannot be compared with a TPD measurement because dissociation of the molecules occurs on the surface before they can desorb, although it is larger than the corresponding energy for C₂H₄, as we would expect. However, it might be possible to measure the adsorption energy directly using surface calorimetry techniques which are capable of resolving heats of reaction from very small samples of adsorbed gases [198,199]. There is a repulsive interaction between molecules adsorbed along a dimer row, similar to that found for C₂H₄. In this case, the adsorption energy of the second adsorbate is 2.00 eV; this implies a surface-mediated intermolecular interaction of 0.19 eV per nearest-neighbour pair. This may again originate from a mixture of elastic and electronic effects.

5.2.2. Charge density and the fate of the dimer bond

The above calculations predict that when either C₂H₂ or C₂H₄ is adsorbed onto a dimer, there is only a slight increase in the separation of the silicon atoms. There has been some debate about how the Si bond changes on adsorption – whether it remains intact [200–202] or is broken [187,188,203]. The small change observed in the geometrical structure suggests that the dimer bond remains unbroken and implies that adsorption occurs onto the silicon dangling bonds, which is geometrically somewhat awkward. On the other hand, if the dimer bond has been broken to provide electrons for the Si–C bonds, then one would expect the dangling bonds to remain present and available for further bonding.

To investigate this experimentally [204], a Si(0 0 1)-(2 × 1) surface saturated with deuterated ethene was exposed to atomic hydrogen, obtained by passing H₂ over a hot filament. High resolution electron energy loss spectroscopy (HREELS), temperature programmed desorption (TPD) and isotopic mixing experiments were then performed. The HREELS measurements gave a spectrum almost identical to that following initial ethene chemisorption, with the addition of a strong Si–H vibration at 2100 cm⁻¹. The TPD experiments exhibited almost the same ethene desorption peak as is observed without post-hydrogenation, while isotopic mixing experiments showed no exchange between ethene and hydrogen to within the isotopic purity of the chemicals used. These results all point to the conclusion that upon post-hydrogenation, hydrogen attaches to the Si dangling bonds; since there are then also a Si–C bond and two Si–Si bonds to the layer below for each silicon atom in the bonded dimer, the dimer bond can no longer exist.

The calculated pseudoelectron densities with and without an adsorbed ethene molecule are shown in Fig. 14 [196]. Fig. 14(a) shows a (1 1 0) section through a plane including a dimer with an adsorbed ethene; there is a high density associated with the carbon atoms, while the hydrogen atoms (which do not lie in this plane) are not visible. Fig. 14(b) shows a (1 1 0) section through the structure, through a plane bisecting the dimer bond. There is a high density between the silicon atoms, and the near-circular symmetry in (b) is suggestive of a σ-bond. Figs. 14(c) and (d) show the same sections calculated in the absence of the ethene molecule, but with the substrate atoms fixed in the slightly distorted positions that they occupy with the adsorbate present. Within the dynamic resolution of this presentation of the data, the density around the dimer bond appears to be the same. However, on subtracting the two densities for a (1 1 0) section and amplifying the dynamic range of the greyscale, Fig. 14(e) is obtained; this shows
that the difference density possesses a node in the $x$--$y$ plane, as would be expected for a $\pi_z$-orbital. This orbital is more delocalized than the $\sigma$-bond, so its charge density is lower.

The calculated length of the carbon--carbon bond in the adsorbed ethyne molecule (1.37 Å) is slightly greater than the length of the double bond in a free ethene molecule (1.331 Å). If this difference is significant, then it may be related to the frequency of the C--C stretch vibration frequency of 1450 cm$^{-1}$ observed by HREELS for adsorbed ethyne, which is a little lower than the 1623 cm$^{-1}$ of a free ethene molecule [204]. What is the reason for this extension? One possibility would be that the dimer $\sigma$-bond breaks, with the dimer $\pi_z$-bond remaining intact. Hybridization of the dimer $\pi_z$-bond with the $\pi_z$-bond of the adsorbed ethyne would then contribute to the bond lengthening. However, there is a problem with this explanation because the $\pi_z$-bond of the adsorbate must already have been broken, being the only $\pi$-bond available with the correct symmetry to interact with the available orbitals of the silicon dimer.°

The calculations suggest an alternative explanation: electrons are not removed from the dimer $\sigma$-bond on adsorption, but rather from the weak $\pi_z$-bond formed by the overlap of the Si dangling bonds. The resulting bonding configuration is necessarily somewhat strained, because the bond length of the hydrocarbons is less than would ideally be required for bonding to the silicon dangling bonds. This strain is more severe for the shorter bond in ethyne than for the longer bond in ethene; this may help to explain why the carbon--carbon bond length in adsorbed ethyne is increased somewhat beyond the normal length of a carbon--carbon double bond, whereas the bond length in adsorbed ethene is almost exactly that of a carbon--carbon single bond.

How are these results to be reconciled with the results of the post-hydrogenation experiments, which indicate that hydrogen attaches to the dangling bonds? There are at least two possible explanations. The first is that STM images of C$_2$H$_4$ adsorbed on Si(001)-(2 × 1) show unoccupied dimers at all exposures. Absolute measurements of C$_2$H$_2$ on Si(001)-(2 × 1) by TPD also indicate a coverage less than unity (normalized to monohydride coverage).° The post-hydrogenation results summarized earlier apply only to low hydrogen exposure; higher exposures convert the ethene to adsorbed ethyl groups, with a (1 × 1) LEED pattern [205]. It is not easy to quantify the flux of atomic hydrogen from a hot filament, but it was estimated that the adsorbed ethene was unaffected up to an exposure of about 5.6 × 10$^{18}$H m$^{-2}$. Thus it might be that the attachment of atomic hydrogen to the dangling bonds on clean silicon dimers is responsible for the post-hydrogenation results discussed above. A piece of evidence against this, however, is that hydrogen post-adsorption produces a very small downshift in the carbon--carbon stretch frequency of the adsorbed ethene, which might suggest that the hydrogen is, after all, adsorbed on the same dimers as the ethene.

Calculations have been therefore performed on a surface with co-adsorbed C$_2$H$_2$ and H; these suggest a second explanation [196]. Fig. 15(a) shows the pseudo-charge density in a (1 1 0) plane before the hydrogen was added; the geometry and the charge distribution around the silicon is very similar to that shown for the case of C$_2$H$_4$ adsorption in Fig. 14(a), and in particular the Si--Si $\sigma$-bond is still present. On introducing two hydrogen atoms close to the dimer silicon atoms and relaxing the system, however, the situation changes to that shown in Fig. 15(b). Two changes are evident: the dimer atoms have moved apart (to a separation of 3.64 Å), and the electron density between them has almost vanished. The hydrogen atoms are attached to the dangling bonds on the silicon atoms, just as was suggested by the experiments. These results indicate that the dimer bond is indeed broken in the

---

9We are grateful to Professor W.H. Weinberg and Dr. C.D. Bain for these points

10W. Widdra (private communication).
Fig. 14. Charge densities (in units of electrons per Bohr radius cubed) for the Si(001) surface in the presence and absence of adsorbed C\textsubscript{2}H\textsubscript{4}: (a) charge density in a (1\,1\,0) plane through a surface dimer and an adsorbed C\textsubscript{2}H\textsubscript{4} molecule; (b) charge density in a (1\,1\,0) plane through the centre of the dimer bond for the same system as in (a), showing the carbon–carbon \sigma bond; (c) charge density in a (\overline{1}\,1\,0) plane with the Si atoms in the same relaxed positions as (a), but with the molecule removed; (d) charge density in a (\overline{1}\,1\,0) plane for the same system as in (c); (e) difference between charge densities (b) and (d), with amplified greyscale showing only the regions where charge density has decreased on adsorption, showing that charge density corresponding to \pi-bonding of the Si dimer has disappeared [196].
presence of coadsorbed hydrogen and ethyne, but that the bond cleavage occurs as a result of the post-adsorption of the hydrogen rather than of the initial adsorption of the hydrocarbon.

Total energy calculations have been also performed for the case in which only a single hydrogen atom is coadsorbed on the same dimer as the C₂H₂ molecule. (In order to avoid a system containing an odd number of electrons, which would strictly necessitate a more time-consuming spin-polarized calculation, a second hydrogen atom was adsorbed on a neighbouring dimer.) This case gives an
intermediate result: the silicon dimer bond is stretched to a separation of 2.84 Å and some σ-bond charge density remains. This suggests that weakening of the dimer bond at low coverages of atomic hydrogen might provide a pathway leading to complete bond cleavage at higher coverages. The calculations have addressed only the location of stable minima on the potential energy surface of the system and no kinetics associated with transitions between them have been considered; in particular, it has not been demonstrated that there is an accessible low-energy path between the configuration in which the hydrogen is desorbed and the dimer bond is intact, and the situation in which the hydrogen is adsorbed and the dimer bond broken.

Hydrogen on Si(001) has been well studied both experimentally (for example [206]) and theoretically [207,208]. The monohydride phase (i.e. two hydrogen atoms and therefore two covalent bonds per dimer) does not exhibit cleavage of the dimer σ-bond. Higher coverages of hydrogen are required to break up the dimer structure and produce a 3 × 1 or 1 × 1 surface periodicity with up to two covalent bonds per surface atom. This is very similar to our finding that hydrocarbon adsorption (producing a maximum of two covalent bonds per surface dimer) is not sufficient to break the dimer bonds; an increase in the number of bonds per dimer to four by hydrogen coadsorption is necessary to make this happen. A comparison can also be made with the adsorption of the same hydrocarbons on the As-rich GaAs(001) surface (see Section 6.2).

5.2.3. STM images – the effect of electric fields

In all the experimental filled-states images of ethene on silicon at low coverages, the adsorbed molecules appear smaller or lower than the clean dimers. This observation emphasizes the well-established point that STM images must not be naïvely interpreted as simple topography images if the electronic properties are not uniform. The fact that the tip must come closer to the surface in order to achieve the required tunnelling current relates nicely to the changes that we attribute to the π-bond upon adsorption. It is this bond that is responsible for the tunnelling in a clean dimer: if it is changed
Fig. 15.  (a) Charge density in a (1 1 0) plane through a surface dimer and an adsorbed C₂H₂ molecule; (b) the same system but with two H atoms added. Note how the dimer bond has been broken in (b) and that the dimer Si atoms have relaxed away from one another [196].

and if the bonding and antibonding orbitals in the carbon–carbon bond are more widely separated, then the tip comes closer to maintain the tunnelling current.

Despite the good account of the structural properties given by the calculations described in Section 5.2.2, simulations of the STM image derived from the calculated structural and electronic properties using either the Tersoff–Hamann approximation (Section 4.1.2) or a non-perturbative Green’s function technique (Section 4.2) give qualitatively incorrect answers. Fig. 16 shows the results of the nonperturbative simulation; it shows a filled states picture in which the molecules appear
Fig. 16. Theoretical filled-states STM picture for C\textsubscript{2}H\textsubscript{4} on Si(00 1), calculated using a Green's function technique without the inclusion of the tip electric field [209]. Note that the molecules are more prominent than the bare dimers, in disagreement with experiment.

considerably \textit{brighter} than the empty dimers [209]. This result is opposite to what is found experimentally, and is independent of voltage (over a range from from -2 to 0 V in the filled states), and of tip–sample separation (over a range from 2 to 4 Å above the clean parts of the surface). There might seem to be many conceivable reasons for this discrepancy. The structure being modelled might be incorrect; however, the range of successful predictions of vibrational and energetic properties encourages us to think is this unlikely. The calculations might be inadequately converged with respect to plane-wave cutoff, \( k \)-point sampling, or slab thickness; there might be a large error arising from the use of the LDA, or from the chosen method of modelling the STM current. One aspect of the experiments, however, is conspicuously omitted from the calculations: a finite bias (of the order of 1 V) has to be applied to the tunnel junction in order to image a semiconductor. The tip–sample separation is probably 5–10 Å, so in the experiment there is a tip-induced electric field of order \( 10^9 \) Vm\(^{-1} \) near the surface.

Such an electric field will have both mesoscopic and microscopic effects. On a mesoscopic scale, as mentioned in Section 2.2.2, the field causes band-bending near the surface; this affects the absolute position in energy of the surface energy levels and the apparent size and location of the bandgap, but has no significant effect on the distribution of electrons in a given state near the surface. These mesoscopic effects can be straightforwardly (albeit approximately) included at the image simulation stage, when the correspondence is made between tip–sample bias and energy levels. However, since the field is significant even by atomic standards, there will also be microscopic distortions of the individual wavefunctions and atomic positions. These microscopic effects must be included at the point where the structure itself is calculated.

What effect might the microscopic field-induced effects have on the STM image? When the sample bias is negative, the field ‘pulls’ electrons out of the surface towards the tip. This ‘pull’ aids the tunnelling process, and is more effective when the tip is over a more polarizable part of the surface,
Fig. 17. Theoretical filled-states STM picture for \( \text{C}_2\text{H}_4 \) on \( \text{Si}(001) \), calculated using a Green’s function technique with the tip electric field included [209]. Note that the bare dimers are now more prominent than the molecules, in agreement with experiment.

thereby enhancing the visibility of more polarizable components. In this example, silicon is a relatively polarizable material (with a bulk relative permittivity \( \epsilon_r \approx 14 \)), whereas hydrocarbons are only slightly polarizable \( (\epsilon_r \approx 1 \text{ or } 2) \); the field will therefore make the silicon surface brighter relative to the adsorbed molecules.

This expectation is born out by detailed calculations. When the electric field is included in the self-consistent calculations, the polarization is sufficient to reverse the zero-field contrast [209,210] and the STM image shown in Fig. 17 is predicted. Analytical calculations on a simple model of a semiconducting surface with an impurity which has a larger band-gap than the rest of the surface (and hence is less polarizable) confirm that the relative enhancement of tunnelling over the bulk surface in the presence of a field is a general effect [211].

5.3. Hydrogen on \( \text{Si}(001) \)

The previous examples have been concerned with the structure of the surface and how molecules are adsorbed on it, and with comparison between what is seen in the STM and simulated images of the modelled structures. We shall return to that kind of question when we consider growth from hydride sources in Section 5.4. But first we want to consider a somewhat different kind of comparison between experiment and theory, namely the motion of adsorbates on the surface. In this case the comparison will be between observed rates of movement as a function of temperature and theoretically calculated activation energies. This will be illustrated from the diffusional hopping of individual hydrogen atoms along the dimer rows on a \( \text{Si}(001) \) surface.

A crucial role in the growth of silicon alloys from hydride sources is played by hydrogen on the surface. In the growth of Si–Ge alloys from solid sources it is difficult to maintain the exact composition required, especially at interfaces, because the germanium atoms tend to segregate to the
surface, leading to gradations in the germanium content where it should be constant. In solid source growth Group V elements can be used as surfactants. In growth from hydride sources hydrogen on the surface behaves as a rather effective surfactant, reducing the energy difference between surface Ge and Si atoms, and it would be helpful to understand how this works. Hydrogen coverage also affects the sticking probability of arriving hydride molecules, thus giving rise to site specific growth phenomena, and again it would be useful to understand the details of such processes. There is evidence that unusual first-order thermal desorption kinetics of molecular hydrogen can be accounted for in terms of pairing of adsorbed hydrogen atoms on silicon dimers prior to desorption [212]. A similar mechanism appears to operate during adsorption of atomic hydrogen on Si(001). Even at temperatures as low as 150 K, hydrogen atoms are found to adsorb on the surface in pairs for coverages of 0.2 or more [13]. This can be accounted for by an adsorption mechanism in which the incident hydrogen becomes dynamically trapped in a vibrational state in which it is mobile above the barrier for diffusion, and is thus able to sample different sites with a strong preference for adsorption on a dimer which already contains one hydrogen atom (ca. 0.2 eV [213]). About 20 such sites can be sampled in this way before the kinetic energy is dissipated and the atom becomes subject to the conventional diffusion barrier.

At lower coverages, it is nevertheless possible to have single hydrogen atoms on a Si(001) surface [206]. An example is given in Fig. 18 [214]. The structure is not controversial. The hydrogen atom bonds with one of the Si atoms of a dimer, breaking the dimer π-bond but leaving the σ-bond intact. Breaking the π-bond costs 0.2–0.3 eV in energy [213]; this is why a second atom prefers to bond on a dimer whose π-bond is already broken if it can. The Si–H bond lies about 2 eV below the Fermi level, and therefore looks dark in an STM image. As a result of the breaking of the π-bond, the clean end of the dimer now has a half-filled dangling bond rather close to the Fermi level. This therefore looks bright in both filled- and empty-states STM pictures. The bright spots indicated by circles in Fig. 18 indicate single hydrogen atoms on dimers, specifically they mark the clean ends of dimers which have a hydrogen atom on the other end. The images in Fig. 18 were obtained with the substrate at a temperature of 500 K and with a sample bias of −1.3 V. Fig. 18(b) was taken 15 min. after Fig. 18(a), and none of the bright spots indicated in the seven circles has moved. It can therefore be concluded that at this temperature the probability of a hydrogen atom making a thermally activated move along a dimer row is small. With an attempt frequency of $10^{14}$ s$^{-1}$, this indicates an activation energy in excess of about 1.5 eV.

By observing the motion of the hydrogen atoms at a series of higher temperatures, it is possible to obtain a series of points for an Arrhenius plot. Fig. 19 shows a pair of images, recorded 10 s apart at a temperature of 600 K [214]. There is considerable motion of the hydrogen atoms in this time interval. Indeed, the slightly fuzzy appearance of the dots is due to the motion of the atoms during the time taken to scan them. Of the two atoms circled, one has moved only about two dimer positions, while the other has moved considerably further. In every case the movement is along dimer rows.

At higher temperatures still, there is too much motion to be able to compare the positions of individual hydrogen atoms in consecutive scans. Instead, it is possible to observe the effect of thermal motion in a single picture. Fig. 20 was taken at a sample temperature of 640 K [214]. STM pictures are scanned images, not snapshots, and at this temperature the hydrogen atoms move several times during a scan. There is an interval of 40 ms (18 secs/512 lines) between successive scans, so if a hydrogen atom moves in that time the image of a bright spot gets broken up into a sequence of bright and dark streaks, according to whether or not an atom was present during each line scan. At this temperature the hydrogen atom can move many times during the time required to scan an entire image. This produces
Fig. 18. Hydrogen on Si(0 0 1) at 500 K; width 14 nm, $V = -1.3$ V, $I = 0.08$ nA. The white dots ringed in (a) indicate the presence of individual hydrogen atoms (they are actually the clean Si atom of a dimer which has a hydrogen atom on the other end). Pictures (a) and (b) were separated by 15 minutes, during which time there was no change in the position of the adsorbed hydrogen atoms [214].

so-called smudges on the pictures. Two such smudges are indicated in white ovals on the picture. Even at this temperature the movement of the hydrogen atoms is still along the dimer rows.

From analysis of observations such as these at different temperatures it is possible to construct an Arrhenius plot of the hopping probability. This is presented in Fig. 21 [214]. It happens that the relevant temperatures are quite difficult to measure accurately. They are too low to measure with a disappearing filament pyrometer, and the sample is both too small and too transparent for an infra-red pyrometer at these temperatures. Higher measured sample temperatures can be related to the power through a law of the form $aT + bT^4$. For the experiments here the sample temperature was interpolated from such a relationship, but there is an uncertainty of ±30 K and this is indicated by horizontal error bars. The hopping rates deduced at the lower temperatures represent an upper limit, while those at the higher temperatures represent a lower limit, and these are indicated by unidirectional error arrows in each case.

It is possible to devise sophisticated statistical models for analysing the hopping rates from the smudges at higher temperatures, but since the hopping probability appears only in a logarithm it would be necessary to improve considerably the accuracy in the temperature measurement in order to take advantage of more accurate hopping data. The size of the error bars allows for considerable variation in slope of a straight line to be drawn through the experimental data, and it is therefore useful to have an addition point at $1/T = 0$, corresponding to an attempt frequency. The median value for the pre-exponential factor from the literature is $10^{13}$ s$^{-1}$, and the solid line passes through that point, with a slope corresponding to an activation energy of 1.68 eV. To indicate the effect of varying this assumption, the dashed line corresponds to a pre-exponential factor of $10^{14}$ s$^{-1}$, with an activation energy of 1.80 eV. It is possible that the influence of the STM tip locally reduces the barrier to diffusion, but if such an effect exists it is probably small at the voltages and currents used.
Fig. 19. Hydrogen on Si(0 0 1) at 600 K; width 20 nm, \( V = -1 \) V, \( I = 0.08 \) nA. At this temperature the white dots are a little fuzzy, due to the motion of the hydrogen atoms. The pictures were separated by 10 s, during which time some of the atoms have moved along the dimer rows. Two hydrogen atoms are circled in both (a) and (b); one appears to have moved a considerable distance along the dimer row, the other has moved about two dimers along [214].

The diffusion of individual hydrogen atoms on Si(0 0 1) was modelled by a combination of tight binding and LDA, one after the other. Atomic positions were first calculated using the density matrix method which enables tight binding calculations to be performed with computational effort which scales only linearly with the number of atoms involved. A slab was used which was one dimer wide, six dimers long, and ten layers deep. The lower five layers were given bulk coordinates, and the bottom surface was terminated with hydrogen (145 atoms altogether). Using a nearest-neighbour parametrization for the hydrogen–silicon interactions deduced from silane molecule data, minimum energy configurations were found for the start and end points with the hydrogen atom on the down atom of a buckled silicon dimer. For the midpoint configuration the coordinate parallel to the dimer row of the hydrogen atom was fixed, and situations were considered in which the hydrogen atom was between two up atoms, or between two down atoms, or between one of each. Two up atoms gave the lowest energy. This calculation gave a barrier of 1.75 eV for the hydrogen atom to diffuse from one dimer to the next. To increase the accuracy of the results, the atomic coordinates of the initial and midpoint configurations were then used in a density functional computation [68], using a slab two dimers wide, two dimers long, and five layers deep, with the bottom layer given bulk coordinates and terminated with hydrogen (58 atoms altogether). This LDA calculation gave a barrier of 1.2 eV. LDA has a tendency to make bonds stronger than they really are, giving a fractional error which is greatest at the mid-point between two dimers where the electron density is relatively low, and thus a barrier energy which is too small. To alleviate this underestimate, generalized gradient approximation (GGA) corrections were applied using the self-consistent LDA charge density, and this increased the calculated barrier energy to 1.55 eV. Further inclusion of spin-polarization in the calculation further increases the value to 1.60 eV; this probably represents the most reliable value available.
Fig. 20. Hydrogen on Si(001) at 640 K; width 30 nm, \( V = -1 \, \text{V}, \, I = 0.08 \, \text{nA} \). The fuzzy white dots have broken up into elongated smudges, two of which are indicated with ellipses. These smudges correspond to hydrogen atoms moving rapidly relative to the scanning time. Each line takes about 40 ms to scan, and smudges begin to appear when the hopping probability in that time approaches one third [214].

Both the tight-binding and the GGA values for the energy barrier are in agreement with the experimental observations, viz. that individual hydrogen atoms on Si(001) are immobile at room temperature, they begin to move at 570 K, and by 640 K they move faster than the line scan rate of the STM, with an energy barrier of \( 1.68 \pm 0.15 \, \text{eV} \). These results add weight to the conclusion from the earlier adsorption studies [13], that a mobile delocalized state must exist prior to chemisorption, since it would be impossible for thermal surface diffusion to account for the observed pairing at low temperatures.

5.4. Disilane on Si(001)

Designer silicon based semiconductors can be produced by growing materials with spatially varying germanium content. The combined effects of alloying, strain, and quantum confinement give endless scope for engineering the structural, electronic, and optical properties. At relatively low germanium content, for example, graded layers can be made as virtual substrates, while when pure germanium is grown on silicon, quantum dots spontaneously form. Silicon–germanium growth processes are affected by surfactants, which can be used to control roughness and segregation. Hydrogen can be an effective surfactant, so that Ge–Si alloys grown from hydride sources give better controlled changes in composition than materials grown by solid source MBE. Growth processes during gas-source molecular beam epitaxy (GSMBE) can be studied with atomic resolution in UHV STM. The study of GSMBE of silicon provides a marvellous foundation for eventual studies of germanium on silicon and eventually Ge–Si alloys. Although silane (SiH\(_4\)) is widely used as a source for reasons of cost, for laboratory
Fig. 21. Arrhenius plot for H diffusion on Si(001). The solid line is fitted to a pre-exponential rate of $10^{13} \text{s}^{-1}$, giving a hopping barrier of 1.68 eV; the dashed line is fitted to a pre-exponential rate of $10^{14} \text{s}^{-1}$, giving a hopping barrier of 1.80 eV. Data points with arrows pointing up denote experimental lower limits on the hopping rate; conversely, points with arrows pointing down denote experimental upper limits [214].

studies disilane ($\text{Si}_2\text{H}_6$) is preferred because of the lower temperatures at which it decomposes in contact with the substrate and a higher sticking coefficient. Rather detailed studies have been performed from the very earliest stages of deposition right up to the growth of whole monolayers. For this review we have selected two examples where comparison between STM pictures and atomistic modelling are available. In both cases the exposure to disilane was carried out in situ in the STM.

5.4.1. $\text{SiH}_2$ fragments on Si(001)

A common pathway for the decomposition of disilane on silicon is first for it to break into two $\text{SiH}_3$ fragments, and then for each of these to lose a hydrogen atom to form adsorbed $\text{SiH}_2$ [215]. The second of these processes has been captured in the remarkable sequence of STM images presented in Fig. 22 [181]. These pictures were obtained at room temperature, so that a process which would occur far too rapidly to be captured at actual growth temperatures is slowed down sufficiently that it occurred during a time interval of a few minutes in this experiment. Fig. 22(a) is a filled-states image ($V = -2.2 \text{V}$) of the clean Si(001) surface before exposure. The tip was then withdrawn 2 $\mu\text{m}$, and the sample was exposed to 0.1 L disilane. After bringing the tip back into tunnelling range, Fig. 22(b) was obtained. The area shown in these images has been selected from considerably larger scans, which enabled the same area to be found unambiguously, and also served as a check that diffusion and reaction of adsorbates from elsewhere on the surface were not responsible. In Fig. 22(b), two fragments are visible. The upper fragment is identified as an $\text{SiH}_2$ fragment. A combination of voltage dependent imaging and statistical analysis of large areas indicates that at low exposures the predominant fragments are $\text{SiH}_2$ and H, and that these occur in equal numbers to within 10%. The lower fragment in Fig. 22(b) is
Fig. 22. Sequential filled-states images of a Si(001) surface showing adsorption of SiH$_2$ and SiH$_3$, and the dissociation reaction SiH$_3$(ads)$\rightarrow$SiH$_2$(ads)+H(ads): (a) the clean Si(001) surface before exposure to Si$_2$H$_6$; (b) the same area after exposure to 0.1 L Si$_2$H$_6$ at 300 K, showing an adsorbed SiH$_2$ group (upper) and an SiH$_3$ group (lower); (c) the same area 8 min. later, showing the shift in bonding location as the SiH$_3$ group dissociates to SiH$_2$, and adsorbed H (dark); picture width 4.7 nm, $V = -2.2$ V, $I = 0.5$ nA [181].

identified as an SiH$_3$ fragment. Over the course of the next 8 min. the SiH$_3$ fragment decomposed into SiH$_2$ + H. The new SiH$_2$ fragment in Fig. 22(c) is similar in appearance to the one previously identified in Fig. 22(b).

What is the nature of the bonding of this SiH$_2$ fragment to the underlying silicon surface? In order to investigate this, density functional calculations were performed for four candidate structures [216]. The candidate structures are shown schematically in Fig. 23; each has full sp$^3$ hybridization of the adsorbate, with the silicon atom in the silane fragment bonded to two substrate silicon atoms. The calculated energies are given in the caption to Fig. 23. Structures A and B have the silane fragment bonded to a single dimer: in A the sigma bond of the dimer is intact, whereas in B it is broken. Breaking the dimer bond leads to an increase in the Si–Si separation from 0.243 to 0.329 nm, at a cost of 0.183 eV in energy. It is interesting to compare this result with the adsorption of ethene and ethyne on a silicon dimer. Although structure B is metastable, the high energy associated with the broken sigma bond rules it out. There are two possibilities for bonding to silicon atoms from different dimers, either in the same row (C) or in adjacent rows (D). The distance between adjacent dimer rows is so great that bridging is possible only if both dimer bonds break. This, together with the necessary bond distortions, has an enormous energy cost, 0.313 eV greater than C, which effectively rules it out. Therefore the two remaining candidates are A and C. Stable configurations for these structures and also the clean surface are shown in Fig. 24, together with plots of electron density ($e$A$^{-3}$) through a chosen section [216]. It is remarkable that the sterically unfavourable angles in A lead to such a low energy structure. Structure C has a four membered ring, with bonds to the adsorbed silicon as strong as the dimer back bonds; we shall see another example of a silicon four membered ring as a candidate nucleation site for silicon growth. The difference in energy between structures A and C (0.004 eV) is less than the accuracy of these calculations, and in any case would not be significant at room temperature or above. Direct evaluation of these two must therefore come from comparison with what is seen in the STM. In Fig. 22 the SiH$_2$ fragments are placed asymmetrically on the dimer rows. This strongly suggests that they are more likely to be in configuration C than A.

The identification of C rather than A is highly relevant for what happens next in the growth of silicon from disilane. There is general agreement that two SiH$_2$ fragments must come together to form a single dimer of monohydride, losing a molecule of H$_2$ in the process. For case A this requires adsorption on two adjacent dimers along a row, and the calculated energy is 0.20 eV per fragment greater than for two
Fig. 23. A schematic plan view of the possible bonding sites for an SiH₂ fragment, with their energies relative to a clean surface plus a half of a separate disilane molecule: (A) on top of a dimer (on-dimer, 0.313 eV); (B) inserting into a dimer bond (in-dimer, 0.187 eV); (C) between two dimers in the same row (intra-row, 0.004 eV); (D) between two dimer rows (inter-row, 0.000 eV) [216]. The large ovals represent surface dimers, the pair of large empty circles represents a cleaved dimer, the small empty circle represents the Si atom, and the filled circles represent hydrogen atoms.

Independently adsorbed fragments. This positive energy would inhibit the required adjacent adsorption, rather in the way that is observed for ethene (see Section 5.2 above). For case C there is little interaction between the rows, and the adsorption energy for two fragments facing each other between dimer rows is within 0.01 eV of the independent adsorption energies. The question of how a new dimer formed from two fragments coming together in this way can rotate to the correct orientation for the growth of the next layer and also lose their hydrogen atoms at relatively low growth temperatures is a matter of active consideration.

5.4.2. Four-membered silicon rings

In the study of gas-source molecular beam epitaxy growth, it is possible to perform in situ experiments in a UHV STM that is capable of operating with the sample at elevated temperature, such as the JEOL JSTM-4500XT. By performing experiments at a range of temperatures, features can be observed corresponding to successive points along the reaction pathway as the temperature of the substrate increases. At appropriate disilane flux, the first development at 400 K is the formation of single epitaxial dimers and the four-membered ring structures to be described in this section [214]. At 450 K dimer strings are formed, generally an odd number of dimers in length and terminated by rebonded ends except where this is prevented by surface defects. At temperatures up to 600 K continued growth is blocked by adsorbed hydrogen, but at higher temperatures layer by layer growth can proceed. When one layer is almost complete, the next layer begins, and at modest disilane fluxes four membered rings are again among the first features of the new layer to form.

A pair of filled- and empty-states images from a study of GSMBE silicon growth is shown in Fig. 25 [217]. These pictures were obtained at 550 K after dosing with disilane at that temperature. In the filled
Fig. 24. Charge densities for (a) a clean silicon dimer, and for the two SiH2 fragment configurations with the lowest energies, (b) on a dimer (on-dimer) and (c) between two dimers in the same row (intra-row). For the clean dimer and the on-dimer the charge densities are taken in the plane of the dimer. For the intra-row structure the atomic positions are shown rotated 90° with respect to the other two structures, and the charge density is taken in a plane containing the adsorbed silicon atom and the two dimer atoms to which it has bonded. The charge density contours are in units of electrons per cubic Angstrom. The silicon atoms appear as charge minima, and the bonds as charge maxima, because the core electrons are not included in the charge density maps [216].
Fig. 25. Si(001) surface exposed to disilane: (a) $V = +2.0 \, V$, $I = 0.08 \, nA$, width 20 nm. Three of the so-called square features are shown circled; they are found after dosing with disilane at substrate temperatures above 400 K, and appear to be stable up to 600 K. (b) $V = -1.0 \, V$, width 8 nm, showing the rotation of the nodes in two of the square features which accompanies the change in bias voltage [217].

states image, the ring structure has nodes at 45° to the dimer rows; in the empty states image, there are weaker nodes parallel and perpendicular to the rows. Remembering that in empty-states images of Si(001) the dark stripes correspond to the node in the $\pi^*$-antibonding orbitals, the rings can be seen to lie between dimer rows. Comparison of the filled- and empty-states images, and also of filled-states images at different voltages, enables these structures to be unambiguously distinguished from short dimer strings which can appear similar under certain imaging conditions such as Fig. 25(b), in which a pair of dimers is visible near the bottom. The position of the nodes would be consistent with a structure consisting of four silicon atoms arranged in a square, each with one back bond to one silicon atom of a dimer. Filled-states images would reveal the $\pi$-bonds between adjacent silicon atoms of the square, and empty-states images the $\pi^*$-antibonding states, which may be considered as dangling bonds associated with each of the four atoms.

To model this structure, tight binding was used for an initial evaluation. A density matrix method was used [183,218], with a unit cell six dimers long, two dimers wide, and ten layers deep, the bottom five layers being constrained to lie in bulk positions and terminated with hydrogen. The starting conjecture of four silicon atoms arranged in a square between two pairs of dimers proved to be unstable, and spontaneously distorted to the rhombohedral configuration sketched in Fig. 26 [217]. The tight-binding calculations gave an energy 1.1 eV lower than two isolated dimers on the surface, although
Fig. 26. The proposed structure for the square on Si(001). The filled circles are the four atoms which make up the distorted square; the open circles are the substrate dimer atoms [217].

since tight-binding underestimates the lowest energy of dimers this may be an overestimate of the difference. The rhombohedral configuration is not planar, and in Fig. 26 the upper atoms are drawn larger. There is a symmetrical arrangement in which the upper and lower atoms swap, with a corresponding mirror symmetry in the plan view, and the calculated barrier for this to occur is 0.15 eV. At 550 K a barrier of this height permits millions of flips per second, so that in an STM image the symmetrical configurations would appear averaged, just as happens with dimer tilts if they are not pinned. For each configuration, two of the ring bonds (opposite one another) exhibit strong \( \pi \)-bonding, and the other two have weaker, more bulk-like, \( \pi \)-bonds. The two possible arrangements of the stronger \( \pi \)-bonding are almost degenerate, and so the STM image is expected to be an average of the four possibilities, with fourfold symmetry and with the node orientations seen in Fig. 25. More accurate calculations of the energy of the relaxed rhombohedral structure were performed using LDA, with a unit cell two dimers long, two dimers wide, and five dimers deep, with the bottom layer constrained to bulk positions and again terminated with hydrogen. The LDA calculations gave an energy 0.18 eV lower than two isolated dimers, and also enabled the spatial distribution of charge density to be calculated.

Steps on Si(001) surfaces are designated ‘A’ type or ‘B’ type depending on whether they run parallel or perpendicular to the dimer rows on the upper terrace. The kink energy for ‘A’ type steps is relatively high, so that they generally appear straight, whereas ‘B’ type steps have a low kink energy and are often ragged. The difference is related to the relative ease of making a dimer row shorter or longer, compared to starting a new dimer row alongside a previous one. It is found in solid-source MBE that ‘B’ type steps are favourable sites for growth. At low bias voltages ‘B’ type steps appear bright in the STM, and by a similar argument to that for the rebonded missing dimer defect (Section 5.1.2) this suggests a rebonded configuration for ‘B’ type steps. The four membered ring can be thought of as two ‘B’ steps back to back, the steps being parallel to an incipient dimer row at the level of the ring. It therefore provides two good sites for a single ad-dimer to attach and hence form the shortest string with an odd number of dimers. In this way, the four-membered ring may provide an important nucleation site for the growth of a new silicon layer.
6. The GaAs(001) surface

6.1. The clean surface

The clean (001) surface of GaAs has been the subject of more controversy concerning its structure than the corresponding surface of silicon. The {001} planes consist alternately of Ga and As atoms; the composition of the stable surface varies depending on the chemical potentials of As and Ga atoms in the environment. We concentrate here on the As-rich surface, which predominates in normal growth conditions; in a bulk-terminated surface, the As atoms are arranged in the same pattern as the Si atoms in bulk-terminated Si(001), and so one’s first guess might be that they would undergo the same stages of reconstruction, involving the formation and tilting of As dimers.

The existence of dimers on the surface is indeed firmly established [219], but the situation differs from that in Si because there is one additional valence electron per As atom, and one fewer per Ga atom. So, each bulk Ga–As bond contains 5/4 electrons provided by the As, and 3/4 by the Ga. The two unbonded orbitals per As atom in the bulk-terminated As-rich surface therefore contain 5/2 electrons between them; forming an As–As dimer uses up one of these in the dimer $\sigma$-bond, leaving 3/2 electrons in an unbonded orbital on each atom. This observation is the basis of the ‘electron counting rule’ [220] – the unbonded orbitals on the As atoms must either become bonded in some way, or additional electrons must be supplied from some other part of the structure to produce unreactive lone pairs.

The different reconstructions that are actually observed on the GaAs(001) surface are, in effect, different solutions to the problem of supplying the extra electrons. The most commonly observed reconstruction is based on a $(2 \times 4)$ periodicity, which in turn forms in three different structures labelled $\alpha$, $\beta$ and $\gamma$ in order of decreasing substrate temperature during growth [221]. The surface coverage of As is found to be less than 100% [222], and STM observations of the surface in all the structures show a series of dark lines which have been identified as missing As$_2$ dimers [223]. It is difficult to assess exactly how many missing dimers there are: the initial STM evidence suggested one [223], but the resolution of the pictures is insufficient to rule out the possibility that the bright parts are caused by just two dimers per cell, and therefore that two are missing [224]. The situation is complicated by the extreme surface-specificity of the STM, which makes it difficult to address the structure of the ‘gap’ where the dimer is missing.

Four models of the missing-dimer structure have been proposed: the ‘three-dimer’ and ‘trench-dimer’ [225,226], the ‘Ga-rebonded’ and ‘extra-dimer’ models [221] (Fig. 27 [227]). As its name implies, the three-dimer model has a single missing dimer per cell, while the trench-dimer and Ga-rebonded models have two. The extra-dimer model actually has a total of eight As atoms in the surface layer, but they are not arranged in four conventional dimers. Because the structures differ in the number of atoms, their relative stability will depend on the chemical potential of different atomic species. Even then, calculating the relative energies of the different structures is a delicate problem. Ab initio pseudopotential calculations performed by Northrup and Froyen [228,229] suggested that the extra-dimer model is never stable; these authors identified the $\alpha$-phase as the rebonded-Ga structure, which they found to be stable at higher Ga chemical potential, and the $\beta$-phase as the three-dimer

---

11 By convention, the axes on the surface are chosen in such a way that the $(2 \times 4)$ reconstruction terminates with As dimers, whereas the $(4 \times 2)$ terminates with Ga dimers.
structure. (The trench-dimer model was not included in these calculations.) On the other hand, an analysis of RHEED data [230] suggested that the \( \alpha \)-phase corresponds to the Ga-rebonded structure, and the \( \beta \)-phase to the trench-dimer structure.

How could the local information from the STM be brought to bear on the problem of making these assignments? An attempt to attempt a direct comparison of experimental and calculated STM pictures was inconclusive; such calculations have been performed [125,231] based on the three-dimer model. An ab initio pseudopotential treatment of the surface was combined with a calculation of the current from the Bardeen tunnelling matrix element (Section 4.1.1), using a small Al cluster for the tip. Pictures were computed in the constant-height mode, and it was found that at negative sample bias (tunnelling out of the surface) the bright spots in the simulated image were associated with As atoms in the reconstruction, but that the details of this association depended on the exact bias. Although the results were not inconsistent with the experimental pictures, they do not avoid the basic difficulty of getting good enough resolution in the STM to differentiate between the models, so the other structures could not be ruled out. In a subsequent paper [232], the same group proposed that a mixture of trench-dimer (80\%) and Ga-rebonded (20\%) provides the best fit to the infra-red reflectance anisotropy data from the surface.

This Gordian knot was cut by asking what other information is hidden in the STM pictures that might be used to differentiate between the structures. In the images of the different phases in Fig. 28, it is striking that the \( \beta \)-phase of the surface exhibits long, straight missing-dimer trenches with a minimum of kinks, whereas the \( \alpha \)- and \( \gamma \)-phases have a high kink density. This absence of kinks in the \( \beta \)-phase provides the potential for a new means of discriminating between models of the surface – one which is not beset by the difficulties of resolution in counting the number of dimers in a unit cell, provided that the tendency to kink of the different candidate structures can be calculated [227]. This is a very demanding problem, because of the large size of surface unit cell necessary to compare the energies of kinked and uninkinked structures, but it can be tackled using an \( O(N) \) tight-binding method of the type discussed in Section 3.2.7.
Fig. 28. STM pictures of (a) the \( \alpha \) phase of GaAs(001)-(2\times4) (the dark lines running across the image are the missing-dimer rows) and (b) the \( \gamma \) phases of GaAs(001)-(2\times4) (the dark lines running from bottom left to top right are the missing-dimer rows) [227].

The system required to evaluate the kinking energies contained eight (2\times4) surface cells. The extradimer reconstruction was excluded because of the high energy found previously [228,229], leaving the trench-dimer, three-dimer and Ga-rebonded structures. Of these, it was found that only the trench-dimer structure has a kink energy greater than 1 eV (to be compared to 0.09 and 0.12 eV for the three-dimer and Ga-rebonded structures). The large kinking energy in the trench-dimer system occurs because the kink leaves a high-energy twofold coordinated As atom in the trench between the dimers (Fig. 29(a)). Only in the trench-dimer structure, therefore, would a low density of kinks be expected at growth temperatures, giving the long straight rows that are observed in the \( \beta \)-phase reconstruction. This is definitive evidence that the trench-dimer structure should be assigned to the experimentally observed \( \beta \)-phase.

The use of the presence or absence of kinks as a diagnostic of the different structures can be extended to the relationship between the reconstructions. It is observed [233] in the STM that the kinked \( \gamma \)-phase can be easily obtained from the \( \beta \)-phase by the addition of As, or when the substrate is doped \( n \)-type with Si [234] — the surface being made more electron-rich in either case. One can then ask theoretically how additional As atoms can be accommodated in the \( \beta \)-phase [227]; it is found that two can be fitted into the trench between the dimers. Furthermore, if the kink energy is re-computed with the additional As atoms in place, the energy of the kink is found to be negative, so that a kinked structure is predicted to form spontaneously (Fig. 29(b)). The ease with which the strongly kinked \( \gamma \)-phase is produced from the \( \beta \)-phase can therefore be understood. This example illustrates how all the aspects of a seemingly intractable problem fall into place once the crucial point of comparison between the experiments and the theory — in this case the tendency of the different structures to kink — has been identified.
6.2. Adsorption of hydrocarbons

Since the GaAs(001) surface is in some ways similar to Si(001) (being built around a basic structural motif of dimers), it is natural to ask whether the same hydrocarbon molecules that we considered in Section 5 will also chemisorb on GaAs and, if so, what the response of the dimers will be.

Initial calculations to investigate this question were carried out with the local density approximation [95,183], using a plane-wave pseudopotential technique with periodic boundary conditions. It was found that both \( \text{C}_2\text{H}_4 \) and \( \text{C}_2\text{H}_2 \) adsorb stably on As dimers, and that the dimer bond breaks rather than remaining intact. The adsorption energies found were: for \( \text{C}_2\text{H}_4 \) 0.25 eV (intact dimer) and 0.76 eV (broken dimer), and for \( \text{C}_2\text{H}_2 \) 0.84 eV (intact dimer) and 1.74 eV (broken dimer). It was found that the broken-dimer structure was stabilized by the formation of shorter (and presumably, therefore, stronger) C–As bonds.

It is interesting to make the comparison between these results and those obtained for Si(001) in Section 5, where no dimer bond cleavage was found until atomic H was coadsorbed with the hydrocarbons. The difference, once again, lies in the greater number of valence electrons present in the As dimer; it is as if the dangling bonds on the As atoms have already been passivated (by the extra electron), and so do not require additional H to fulfil this role in order to break up the dimer.

Unfortunately, no STM pictures are available to illustrate this structure; this is because experimentally the chemisorbed state of these molecules on GaAs is found not to exist! Energy loss spectroscopy, of the same type that shows up the C–Si bond stretch that is a signature of the adsorption on Si, finds no such signal on GaAs. Similarly, experiments searching for signals corresponding to the molecular adsorption have been unsuccessful. It has been concluded that the molecules are only physisorbed. What, then, is the cause of this serious discrepancy with the calculations?

The local density approximation is known to make errors in the relative energies of bonded and unbonded structures (see Section 3.2.1). In order to confirm this, the calculations were repeated at the same geometries, but with the electronic structure re-computed using the generalized gradient approximation. The result was to reduce markedly the adsorption energies in the broken-dimer structure.
Fig. 30. STM image of GaAs(001)-(2×4) with adsorbed TMGa. The dark lines running up and down the image are missing-dimer rows, and the bright features result from exposure to TMGa [218,241].

(from 0.76 to 0.05 eV for C₂H₄ and from 1.74 to 0.99 eV for C₂H₂), and to remove altogether the binding in the intact-dimer structure. It is not known by which mechanism chemisorption might occur, but it seems very likely that it would involve physisorption and then chemisorption atop an intact dimer. The instability of the intact-dimer state relative to the bare surface plus an isolated molecule places a barrier in the path of the reaction which makes it very unlikely that molecules can reach the stable broken-dimer structure, despite the prediction that it is stable.

6.3. Adsorption of metallorganic compounds

While the adsorption of C₂H₄ and C₂H₂ is an interesting model system, the deposition of metallorganic compounds such as Ga(CH₃)₃ (trimethyl gallium or TMGa) on GaAs is of more direct technological importance, since it can be used in the growth of GaAs crystals and heterostructures by MOVPE. A great deal of experimental work, including electron energy loss spectroscopy [235], infra-red spectroscopy [236] and mass spectrometry [237,238] has been performed on this system; the conclusions are (i) that TMGa adsorbs intact at −80°C; (ii) that the saturation coverage of Ga is about 0.1 ML (corresponding to slightly less than one atom per (2 × 4) cell) for temperatures up to 300°C; (iii) for low coverage and at room temperature, one methyl group becomes attached to an As atom; (iv) at higher temperatures (500°C) carbon desorbs from the surface, and the saturation Ga coverage rises to 0.2 ML.

The nature and structure of the species produced on the surface were unknown at the time the first STM images of GaAs exposed to TMGa at room temperature were obtained [239]. These show (Fig. 30) a series of bright features, covering at saturation about 75% of the β(2 × 4) unit cells; the apparent
Fig. 31. Scheme of energy differences between different structures for TMGa on GaAs [242]. When methyl groups are not shown as part of the structure, they are taken to be individually adsorbed on distant As dimers (denoted by CH$_3$(s)). Energy differences are in eV/molecule, and structures which are locally stable (in the sense that they are more stable than any of their neighbours in the diagram to which they are connected by arrows) are in bold boxes.

The tendency to cluster along a given row parallel to the missing-dimer trenches can be confirmed by a statistical analysis [218,240,241].

To go further in the analysis of these results, calculations are essential. Density functional calculations were performed [95,183] in a 4 x 4 surface unit cell – the large cell is essential because of the bulky nature of the adsorbate. All structures were considered that could be produced from a single TMGa molecule with the methyl groups remaining intact, that involved only a single As dimer, and that contained only threefold- and fourfold-coordinated As, Ga and C atoms. This resulted in the eight structures considered in Fig. 31 [242]. Each involves either TMGa itself or the dimethyl gallium (DMGa) or monomethyl gallium (MMGa) which can be produced from it by removal of methyl groups:
(1) TMGa adsorbed at one end of an As dimer;
(2) as (1) but with a methyl group removed;
(3) as (2) but with the methyl group added to the other As atom of the dimer;
(4) DMGa inserted into a dimer;
(5) as (4) but with a methyl group bound to one As atom of the dimer;
(6) MMGa inserted into a dimer;
(7) as (6) but with a methyl group bound to one As atom of the dimer;
(8) as (6) but with methyl groups bound to both As atoms of the dimer.

In comparing the energy of structures differing in the number of methyl groups, the methyl chemical potential was fixed at the energy of an isolated methyl group bound to one end of an As dimer; this corresponds to assuming that any methyl group not involved in the local bonding is free to diffuse away, but must remain bound to the surface.

The theoretical stability of the resulting structures is shown in Fig. 31, in which structures which differ by a single chemical step (such as addition or removal of a methyl group, or insertion into the dimer bond) are connected by arrows. The arrow runs in each case from the higher-energy to the lower-energy structure, and the associated number represents the energy change in eV. While all the structures shown are locally stable in the sense that small displacements from equilibrium produce a restoring force on the atoms, two of them (numbers (1) and (5)) have the additional property that they are stable with respect to all neighbouring structures in the diagram. Of these, structure (5) is the global minimum (provided that the methyl group chemical potential is fixed as described above). It should be noted that all these energies were obtained using the GGA but at the LDA relaxed geometry; in all cases, the LDA bound the structure more strongly than the GGA (by between 0.11 and 0.98 eV). Given the experience with hydrocarbons on GaAs, and with other, less complex, systems in which it is possible to compare energetics from density functional theory with quantum chemistry calculations, it is to be presumed that the GGA adsorption energies are more accurate.

It seems, therefore, that below room temperature, the TMGa adsorbs intact into structure (1), and that at room temperature and above it can insert into the dimer bond to produce structure (5). The size of the adsorbates in either case provides a steric barrier to the adsorption of a molecule on the other dimer within the $(2 \times 4)$ cell. The desorption of methyl groups above $450^\circ C$ [238] probably produces MMGa in structure (6), whose smaller size could allow an additional adsorption event on a neighbouring dimer and result in an increase of Ga coverage to around 0.2 ML, as observed. The precise origin of the bright spots in the STM image (Fig. 30) has not been definitively explained, but it seems reasonable to assign them to the large DMGa inserted into the As dimers. The observed clustering of adsorbates could be a kinetic effect, resulting from collisions of mobile precursors with previously adsorbed molecules.

7. Conclusions

The examples presented in this review illustrate powerfully the benefits of interdependent theoretical and experimental work in this field. Without the guidance of real-space experiments, theoretical modelling of surfaces would either be restricted to structures with sufficient long-range order to be visible to diffraction measurements, or would have to operate entirely without the guidance of experiment. The enormous freedom available to the atoms at a surface, and the complexity of the
structures that result, would make this a formidable undertaking. It is very hard to imagine that a
structure as complex as the Si(111)-(7 × 7) reconstruction, or the missing-dimer defects described in
Section 5.1.2, would have been found from theoretical considerations alone.

On the other hand, without theoretical input, our experience has taught us that it can be very difficult,
or indeed impossible, to extract the crucial information about the identity of structures observed in the
STM. This is partly because of the extreme surface-sensitivity of STM as a probe – information on
anything below the very surface atoms is at best indirect – and partly because of the subtle interplay of
electronic and geometrical effects which determine an STM image. Seen from this perspective, the
interdependence of geometry and electronic structure appears both as a weakness and a strength: a
weakness because it prevents one from making instant deductions about either aspect separately, but a
strength because it forces one to keep in mind the physical (and chemical) fact that the atomic structure
is determined by the electronic bonding.

This synthesis demands commitment from both sides. The examples of work from our own
laboratories discussed in this review are the outcome of five years of collaboration which has been by
turns frustrating and rewarding. In some cases the theory and experiment has been performed by
separate groups working in parallel; in others, experimentalists and theoreticians worked side by side in
the same laboratory. Each side must be prepared to calculate, or measure, not just the quantities which
are most accessible to their own techniques, but those which are most useful to the other.

How many different ways are there to relate STM observations and theoretical predictions? Both
start with simple geometrical structure. Semiconductor substrates are generally imaged at bias voltages
within the bulk bands, to reveal the position of surface atoms. By comparing filled- and empty-states
images it is often possible to distinguish bonding from antibonding orbitals; this made the bonding sites
of ethene on Si(001) very clear. By further varying the bias voltage within the bulk bandgap even more
information can be obtained, and sensitivity to details of bonding at defects is often available, as in the
case of the missing dimer defect in Si(001). Such information can sometimes be made quantitative at a
particular site using tunnelling spectroscopy, though in practice voltage dependent imaging may
sometimes be more robust against tip uncertainties. In looking at geometrical structure it is important to
note what is missing as well as what is present; the absence of kinks provided the key to determining
the structure of the GaAs(001)-2×4 structure. Each of these points of experimental observation has a
counterpart in modelling: voltage dependent imaging corresponds to energy dependent density of
states; absence of kinks corresponds to a high calculated kink energy, and so on. For studies of
molecules on surfaces such points can be supplemented with observations of bonding sites and their
exact relationship to the underlying substrate. But neither the experiments nor the modelling are
confined to static structures; it is also possible to study the behaviour of adsorbed atoms and molecules.
By observing the dependence of movement on temperature, activation energies can be deduced, and
these can be related to calculations of barrier heights. It is possible to go further and observe chemical
reactions and growth processes on surfaces, and these in turn can be related to modelling and
simulations, including calculations of energy barriers and of postulated mechanisms. At each of these
points of contact the applicability of the modelling and the interpretation of the experiments can be
mutually evaluated, and each can suggest to the other what to do next.

Of course, modelling and experiment are not the same, and the contact points at which they can be
compared serve to give confidence in the results which each can give that are not available from the
other, and in the subsequent interpretation. For example, the time-scales which are accessible to each
are different. Although there are some tricks for observing faster changes, most STM images are taken
at least a few seconds apart, and thus give the beginning and end states of changes that take place within that time frame. Molecular dynamics, on the other hand, tends to deal in steps separated by femtoseconds, and can therefore be used for modelling much faster processes. Moreover, good modelling is able to provide a picture of what is going on with details that will never be accessible to experimental STM. Molecular mechanics modelling has given invaluable insight into the deformation processes which occur when a large molecule is moved over a surface by pushing it with an STM tip [243]. Cu-TBP-porphyrin has four legs which because of steric repulsion orient themselves out of the plane of the porphyrin ring. By positioning the STM tip between two legs it is possible to push the molecule across a surface. The modelling shows that a crucial role in the motion is played by rotation of legs around the \( \sigma \)-bond attaching them to the central ring. Each leg undergoes stick-slip motion, but the motion of different legs appears to be uncorrelated. A rigid molecule would have a significantly higher barrier to motion.

Results such as this serve as a healthy reminder of the limited scope of the examples given in this review. Neither experimental methods for manipulation of molecules nor theoretical methods of molecular mechanics have been described in the detail which they deserve. New theoretical techniques are becoming available for modelling materials with strongly correlated electrons such as oxides, and these have been applied with remarkable success to STM images of NiO to explain the contrast and enable confident interpretation to be made of the atomic sites being imaged [244,245]. In preparing this review we are conscious that new results are being generated faster than we can write, so that if we were to delay until every piece of work worthy of inclusion was given its proper due, Achilles would indeed never catch the tortoise. But this simply serves to emphasize our central point, which is the enormous benefit from an active combination of theory and experiment.

Finally, we should not forget that the end goal of work like this is not a beautiful STM image, or a perfectly converged set of atomic positions – or even a successful synthesis of the two. It is to achieve some advance in our understanding of processes involving surfaces and the molecules that react with them, so that we can predict what will happen in analogous situations, and contribute to the development of technologies based upon them. We hope that this review may serve to illustrate how a joint experimental–theoretical approach can contribute to such understanding.

Acknowledgements

We wish to thank our colleagues for the many occasions when they have inspired, educated and corrected us. The references to publications in which figures have appeared serve as acknowledgement in each case. We are grateful to Dr. D.R. Bowler for his comments on drafts of the text, and to Dr. C.M. Goringe for his assistance with the figures. We also wish to express our thanks to Professor B.A. Joyce, whose inspiration lies behind so much of the work on semiconductor surfaces presented here, and to Professor A.M. Stoneham, who first introduced us to each other in the hope that we might indeed be able to relate STM experiments and atomistic modelling.

References