First-principles simulation: ideas, illustrations and the CASTEP code

2002 J. Phys.: Condens. Matter 14 2717

(http://iopscience.iop.org/0953-8984/14/11/301)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **14** (2002) 2717–2744 PII: S0953-8984(02)32831-5

First-principles simulation: ideas, illustrations and the CASTEP code

M D Segall[1,2], Philip J D Lindan[3,7], M J Probert[4], C J Pickard[1], P J Hasnip[5], S J Clark[6] and M C Payne[1]

[1]TCM, Cavendish Laboratory, University of Cambridge, UK
[2]Camitro (UK) Ltd, Cambridge, UK
[3]Centre for Materials Research, University of Kent at Canterbury, Canterbury CT2 7NR, UK
[4]Physics Department, University of York, UK
[5]Department of Materials, University of Cambridge, UK
[6]Department of Physics, University of Durham, UK

Received 17 January 2002
Published 8 March 2002
Online at stacks.iop.org/JPhysCM/14/2717

Abstract

First-principles simulation, meaning density-functional theory calculations with plane waves and pseudopotentials, has become a prized technique in condensed-matter theory. Here I look at the basics of the subject, give a brief review of the theory, examining the strengths and weaknesses of its implementation, and illustrating some of the ways simulators approach problems through a small case study. I also discuss why and how modern software design methods have been used in writing a completely new modular version of the CASTEP code.

1. Overview

The simulator builds a model of a real system and explores its behaviour. The model is a mathematical one and the exploration is done on a computer, and in many ways simulation studies share the same mentality as experimental ones. However, in a simulation there is absolute control and access to detail, the ability to compute almost any observable, and given enough computer muscle, exact answers for the model. These strengths have been exploited for the last fifty years and have led to many advances in the theory of condensed matter. However, it is only in the last fifteen years or so that we have been able to compute the properties of condensed matter from first principles. The first-principles approach is vastly ambitious because its goal is to model real systems using no approximations whatsoever. That one can even hope to do this is down to the accuracy of quantum mechanics in describing the chemical bond. Dirac's apocryphal quip that after the discovery of quantum mechanics 'the rest is chemistry' sums it up: if one can solve the Schrödinger equation for something—an atom, a

[7] Author who was invited to submit this article (the full author list is the CASTEP Development Group.)

molecule, assemblies of atoms in solids or liquids—one can predict every physical property. Perhaps Dirac didn't realise quite how difficult doing 'the rest' is, and it has taken great effort and ingenuity to take us to the point of calculating *some* of the properties of materials with *reasonable* accuracy from *small* model systems! However, to focus limitations is to miss the point: the impact of simulations on our thinking about condensed matter problems is immense.

Here I shall concentrate on just a few elements of what is a very large subject. First I shall discuss the first-principles rationale and what makes the task so difficult. I shall focus on one of the most successful approaches, the application of density-functional theory using plane waves and pseudopotentials, and consider why this method turned out to be so important. I shall also spend some time discussing the simulation approach in general, and the types of information that come out of a calculation. To illustrate the application of these methods I shall present highlights of an extensive study of the adsorption of water on an oxide surface, a problem that has turned out to be surprisingly difficult.

In reviewing the methodology I shall in the main look at practicalities rather than the full details of the formalism. This is because I hope to cover a few topics that seldom appear in the literature, and because the formal theory and its implementation are already treated in many excellent reviews [1–9]. Throughout I have slanted things towards the 'user' and the non-specialist.

Finally, I shall describe briefly the latest developments around the CASTEP computer code. Traditional scientific programming has many shortcomings which are serious for large and complex codes such as those required for modern first-principles calculations. With this in mind, CASTEP has been redesigned and rewritten completely. The primary aim of this is to make new developments as rapid and reliable as possible. This is achieved through a modular design which embodies the ideas of data abstraction and encapsulation. Consider that our codes are more than just the 'equipment' we use to probe condensed matter, they are also the language and notation we use to implement new theoretical approaches. In new CASTEP we strive to make this language flexible, powerful and ideal for the task.

2. Theory

2.1. The first-principles rationale

Quantum mechanics provides a reliable way to calculate what electrons and atomic nuclei do in any situation. The behaviour of electrons in particular governs most of the properties of materials[8]. This is true for a single atom or for assemblies of atoms in condensed matter, because quantum mechanics describes and explains chemical bonds. Therefore we can understand the properties of any material from *first-principles*, that is, based on fundamental physical laws and without using free parameters, by solving the Schrödinger equation for the electrons in that material[9]. This, however, is a tall order. We rapidly run into difficulty because electrons interact strongly with each other. The alarming consequence is that *exact* pencil-and-paper solutions exist only for a single electron in simple potentials: solving the Schrödinger equation for the hydrogen atom is a classic undergraduate task, but solving it for helium requires a computational approach. The problem of interacting electrons in condensed-matter physics, one manifestation of the many-body problem, is the defining challenge of the subject.

[8] Setting aside nuclear processes.

[9] Because electrons are so much lighter than atomic nuclei it is possible to freeze the positions of the latter while dealing with the former. The idea is that the electrons move so rapidly compared to the nuclei that they are always in the ground state. This is the Born–Oppenheimer approximation.

Despite this difficulty, one can read dozens of papers each week describing the application of first-principles calculations to systems containing hundreds or thousands of atoms and electrons, yielding accurate, quantitative information. This is a great triumph of condensed-matter science, and it has changed for good the way we approach the subject. How is it then that we can do these calculations?

2.2. Coping with interacting electrons

For practical calculations on condensed matter, most first-principles approaches recast the problem from one where electron interactions are explicit to one where the interactions are represented by an effective potential acting on apparently independent electrons. The interactions are 'hidden' in the effective potential, and one deals with one electron at a time. The result is a set of one-electron Schrödinger-like equations:

$$H\psi_n = \left(-\frac{\hbar^2}{2m}\nabla^2 + V_{\text{ext}} + V_{\text{eff}}\right)\psi_n = \epsilon_i\psi_n \ . \tag{1}$$

Here, ψ_n are the n one-electron wavefunctions, V_{ext} is the external potential of the nuclei[10], and V_{eff} the effective potential. The methods used to achieve this trick may or may not rely on neglecting part of the *e–e* interaction, but almost always involve writing part of it in a mean-field manner. However, it is not necessary at this stage to make any approximations at all, since what is left out of a mean-field term may be added back elsewhere.

An early approach was developed by Hartree. He set V_{eff} to the average of the Coulomb potential between an electron and all others in the system, giving what is now called the Hartree potential. An electron experiencing this potential is said to move in the mean field of the other electrons. Of course this is an approximation, and for two reasons. In the real case the interaction depends explicitly on the position of the other electrons. Something is missed when the interaction is averaged to form V_{eff}. Also, electrons are fermions, and they obey the Pauli exclusion principle and Fermi statistics. This gives rise to an effective interaction, called the exchange interaction, which is not accounted for. The Hartree approach neglects *exchange* and *correlation*, and as one may guess it gives rather poor results.

Adding Fermi statistics to Hartree's method yields the Hartree–Fock approach. The effective potential is now non-local, and arises from the demand that the total wavefunction be antisymmetric upon exchange of any two electrons [10]. The exchange interaction is treated exactly, but the method remains inherently approximate because it neglects correlation. Nonetheless it has enabled advances in quantitative theory and structural studies of molecules and solids, and remains the platform on which highly accurate quantum-chemical theories are built.

2.3. Density-functional theory

Density-functional theory (DFT) takes a radically different approach than the foregoing wavefunction methods. It is both a profound, exact theory for interacting electrons [11], and a practical prescription calculating in terms of single-electron equations [12]. Its contribution in both these respects received the highest recognition with the award of the Nobel prize for chemistry in 1998 to Walter Kohn and John Pople [13]. It has become a runaway success, enabling great advances in practical first-principles calculations. DFT is predicated on two deceptively simple principles [11, 14]:

[10] It may of course arise from other influences on the electrons, such as an applied electric field.

- The total energy of a system of electrons and nuclei is a unique functional of the electron density
- The variational minimum of the energy is exactly equivalent to the true ground-state energy

An alternative form of the first principle is that the density uniquely determines the potential acting on the electrons, and vice-versa. Noting that the energy is a functional of this potential brings us to the statement above[11]. The second principle is a variational statement for the energy in terms of the density.

The beauty of DFT is that one makes no attempt to compute the many-body wavefunction. Instead the energy is written in terms of the electron density. This seems to be a quite remarkable step, throwing out the fearsome complexity of a multidimensional wavefunction and instead working with a simple scalar field. The degree of simplification is immense, yet the theory remains completely general. The energy is written as

$$E = E[\rho(\boldsymbol{r})] = \int \mathrm{d}\boldsymbol{r} V_{ext}(\boldsymbol{r})\rho(\boldsymbol{r}) + F[\rho(\boldsymbol{r})] \ . \tag{2}$$

Both right-hand terms in this equation are *functionals*: in the same way that a function gives a number from a variable, a functional gives a number from a function [15].

Working with a density functional of the energy does not mean that the *e–e* interaction is approximated. DFT asserts that the state energy is given exactly for any density[12], no matter what arrangement the electrons that generate that density are in. It connects this to the true ground-state energy for a given external potential via a variational principle. It also asserts the existence of an exact functional that by construction handles exchange and correlation in any situation. What it does not do is to tell us what that functional is, or how to find it.

The practical tools for applying DFT are the Kohn–Sham (KS) equations [12]. These are n Schrodinger-like equations for n non-interacting electrons moving in an effective potential, just as in equation (1). They arise when the second (variational) principle of DFT is applied to the energy functional, with the density written in terms of the wavefunctions of the non-interacting electrons[13]:

$$\rho(\boldsymbol{r}) = \sum_{n=1}^{N} \psi_n^*(\boldsymbol{r})\psi_n(\boldsymbol{r}) \ . \tag{3}$$

Then F is written in as:

$$F[\rho(\boldsymbol{r})] = E_K[\rho(\boldsymbol{r})] + E_H[\rho(\boldsymbol{r})] + E_{xc}[\rho(\boldsymbol{r})] \ . \tag{4}$$

Here, E_H is the familiar Hartree Coulomb term, which as mentioned, does not include exchange and correlation effects. The electron kinetic energy is given by E_K, and it is defned as the kinetic energy of a system of non-interacting electrons that gives rise to the density $\rho(\boldsymbol{r})$. The exchange-correlation functional E_{xc} is defined to be whatever is needed, and therefore not present in $E_K + E_H$, to make $F[\rho(\boldsymbol{r})]$ exact. In the KS equations the effective potential is the Kohn–Sham potential, defined as the functional derivative

$$V_{KS}(\boldsymbol{r}) = \frac{\delta}{\delta\rho(\boldsymbol{r})}\left(E_H[\rho(\boldsymbol{r})] + E_{xc}[\rho(\boldsymbol{r})]\right) = V_H(\boldsymbol{r}) + \frac{\delta E_{xc}[\rho(\boldsymbol{r})]}{\delta\rho(\boldsymbol{r})} \ . \tag{5}$$

[11] In fact, not only the energy but all ground-state properties are determined by the density: the potential, the wavefunction, the Hamiltonian and so on. See page 51ff of reference [1].

[12] More exactly, for all *N-representable* densities, which are those arising from an antsymmetrized wavefunction. This means any 'reasonable' density, requiring only that (a) the density is non-negative, (b) it is normalized, so that its integral gives N electrons, and (c) that it is continuous in the sense that $\int \mathrm{d}\boldsymbol{r}\|\nabla\rho(\boldsymbol{r})^{1/2}\|^2 \leqslant \infty$. See [1, 16, 17].

[13] A density of this form is said to be *non-interacting v-representable*. This simply means it is associated with the solutions of a set of Hamiltonians for N independent electrons. This is a stronger restriction than N-representability, and in fact not all densities are v-representable. This rarely matters in practise because the exceptions are those which require a full many-body treatment, and for which a ground-state, single-determinant theory is inapplicable. Once again, reference [1] has a good discussion.

But why do we need to swap a single equation in the density (equation (2)) for a system of n equations, and in the process re-introduce wavefunctions? The answer is that it is an excellent means of treating the kinetic energy properly. Efforts to find a direct route to $E_K[\rho]$, starting from the early Thomas–Fermi theory [18, 19], have not yielded accurate prescriptions. Calculating E_K indirectly[14], that is, from the one-electron wavefunctions rather than the density, gives the major part of the kinetic energy exactly, with the remainder accounted for in the exchange-correlation functional. The cost is a considerable loss of simplicity, but this is worth paying. It is important to note that the wavefunctions ψ_n are really auxiliary quantities that are there simply to make the maths work: however, there is plently of evidence that the KS wavefunctions do have physical meaning, but the interpretation needs care.

One must bear in mind that although one deals with apparently independent electrons, the e–e interactions are still present, hidden in $V_{\rm eff}$ (equation (1)), and they have important consequences for the solution of the KS equations, or indeed any equations like equation (1). Because the effective potential depends on the density and hence the ψ_n, the solutions for the latter must be self-consistent. This means that the KS equations and equation (5) must be satisfied together by the same ψ_n. In practise, solutions for ψ_n are found for a fixed V_{eff}, then the latter is updated. The procedure is repeated until self-consistency is achieved. I'll say more on how this is done in section 3.3.

2.4. Making DFT tractable: approximate functionals

At first sight, DFT has done nothing more than repackage an impossible problem. All the complexity of interacting electrons is still there, and the task of finding a functional E_{xc} that embodies the required information seems just as hopeless as that of calculating the exact many-body wavefunction for hundreds of electrons. What saves us is that very simple-minded approximate functionals work, and they work incredibly well. The most widely-used approximation is the local-density approximation (LDA) which was introduced by Kohn and Sham along with the KS equations. The LDA states that E_{xc} can be given by assuming, for each infinitesimal element of density $\rho(\boldsymbol{r})d\boldsymbol{r}$, the exchange-correlation energy is that of a *uniform electron gas* of density $\rho = \rho(\boldsymbol{r})$. Then,

$$E_{xc} = \int d\boldsymbol{r}\rho(\boldsymbol{r})\epsilon_{xc}(\rho(\boldsymbol{r}))\ , \tag{6}$$

where $\epsilon_{\rm xc}(\rho)$ is the exchange-correlation energy per electron in a uniform gas of density ρ. The LDA is clearly wrong, because the charge density is highly non-uniform around atoms. However, the uniform electron gas remains the only system for which E_{xc} can be calculated, and hence from which $\epsilon_{xc}[\rho]$ can be constructed [20]. The LDA seems patently wrong, but it works: its justification is unashamedly *post-hoc*, taking the form of thousands of successful applications which prove it to be remarkably useful and capable of yielding accurate calculated properties for many systems. In some cases though the LDA description is poor. Recognising that LDA failures must in part be due to ignoring spatial variations in the density, functionals have been developed that include dependence on the gradient of the density. This scheme goes under the name of the generalized-gradient approximation (GGA). The GGA improves predicted binding and dissociation energies, especially for hydrogen-containing systems [21]. Almost all workers use the LDA or GGA in one of the several parametrizations available.

Despite the success of the LDA and GGA they are far from ideal, and finding an accurate and universally-applicable E_{xc} remains the great challenge in DFT [1]. In fact, we can be sure of only two things with functionals: that the universal density functional exists, and

[14] As the sum of expectation values of the kinetic energy operator, $< \psi_n | -\frac{\hbar^2}{2m}\nabla^2 | \phi_n >$.

that we shall never find it. Practical E_{xc} functionals are the major approximation made to DFT: they are *not* derived from first principles, but are postulated from physically reasonable assumptions, and their use is justified *a posteriori* by their success. Some may argue that the first-principles ideal is lost by making these approximations. The reality is that the theory must be made practical, and moreover practical for calculations on systems large enough to model condensed phases realistically. The spirit of the approach remains intact though: it is always to make as few approximations as necessary, to drive them ever lower, and to accept the predictions the resulting scheme gives.

3. Machinery

One of the most successful styles of DFT calculation has become synonymous with Roberto Car and Michele Parrinello. Their seminal paper [22] was a turning point in the first-principles story, not because it introduced new theories or radical methods, but because it showed which key elements could be combined to make very efficient calculations possible. Broadly speaking, five features characterize the Car–Parrinello approach:

(a) A plane-wave basis to represent the wavefunctions
(b) Pseudopotentials replacing the ionic cores
(c) The use of fast-Fourier transforms (FFT's)
(d) Minimization of the total energy to find the ground state (originally via simulated annealing)
(e) 'Fictitious dynamics' for the electrons in a unified Lagrangian formalism for molecular dynamics

Not all of these need be used together, and in fact several groups pioneered the approach, each with its own emphasis. For example, the drive towards large-scale calculations and the exploitation of parallel computers was headed by Mike Payne with his conjugate-gradients approach [2, 23] in the CASTEP code. These methods are united by the use of plane waves, pseudopotentials, FFT's and some form of minimization. I shall use the terms plane-wave pseudopotential (PWP) and first-principles to mean an approach of this kind.

3.1. Supercells, plane waves and pseudopotentials

Plane waves and pseudopotentials are a hallmark of the method, and they form a very natural alliance. They are so fundamental that their strengths and weaknesses deserve special attention.

In the PWP method the model system is constructed as a 3D periodic supercell which allows Bloch's theorem to be applied to the electron wavefunctions:

$$\psi_{n,\boldsymbol{k}}(\boldsymbol{r}) = u_{n,\boldsymbol{k}}(\boldsymbol{r}) \exp(i\boldsymbol{k} \cdot r) \ . \tag{7}$$

The function $u(\boldsymbol{r})$ has the periodicity of the supercell. It can be of any suitable mathematical form, and usually one chooses a series expansion in terms of a set of basis functions. In PWP, plane waves are used for this expansion, so that each single-electron wavefunction $\psi_{n,\boldsymbol{k}}$ is written as

$$\psi_{n,\boldsymbol{k}}(\boldsymbol{r}) = \sum_{\boldsymbol{G}} u_{n,\boldsymbol{k}}(\boldsymbol{G}) \exp(i(k+G) \cdot \boldsymbol{r}) \ . \tag{8}$$

The $u_{n,\boldsymbol{k}}$ are the expansion coefficients. The wavevectors $\boldsymbol{G}$ are such that the plane waves are commensurate with the supercell. Both the number of $\boldsymbol{G}$-vectors in the sum and the number of $\boldsymbol{k}$'s considered should in principle be infinite. The exponential term is a plane wave of wavevector $\boldsymbol{k}$ which must be commensurate with the entire system (i.e. not just the

periodically-replicated cell). For an infinite system there is an infinite number of $\boldsymbol{k}$ vectors, at each of which solutions for $\psi_{n,\boldsymbol{k}}$ exist. This simply reflects the fact that the number of electrons is infinite. However, a great simplification comes about when one realises that the change in $\psi_{n,\boldsymbol{k}}$ with $\boldsymbol{k}$ becomes negligible for $\boldsymbol{k}$-points that are close together. This means that one may calculate at a finite number of $\boldsymbol{k}$-points. We speak of this idea as $\boldsymbol{k}$-point sampling. The set of vectors $\{\boldsymbol{G}\}$, on the other hand, should in principle be infinite to obtain an exact representation of the wavefunction. This is never necessary because summing over a finite number of $\boldsymbol{G}$'s will yield sufficient accuracy. Keeping this number small enough for practical purposes is still a technical challenge, as discussed later. Note that the $\{\boldsymbol{G}\}$ is usually chosen as the set of points on a regular grid covering reciprocal space. The subject of grids is discussed further in section 3.2.

A plane-wave basis set has many advantages:

- It is unbiased, so all space is treated the same
- It is complete[15]
- There is a single convergence criterion
- Plane waves are mathematically simple, and their derivatives are products in k-space
- Plane waves do not depend on atomic positions

and one or two important disadvantages:

- The number of plane waves needed is determined by the greatest curvature of the wavefunction
- Empty space has the same quality of representation—and cost—as regions of interest

The advantages speak for themselves. The first three mean that one can always ensure that the basis set is adequate for a calculation by increasing the number of plane waves until the quantity of interest stops changing. In other words, the quality of the basis set depends on a single parameter, usually expressed as the energy a of free electron whose wavefunction has the same wavevector as the largest wavevector in the plane-wave basis,

$$E_c = \frac{\hbar^2(\boldsymbol{G}+\boldsymbol{k})^2}{2m} \,. \tag{9}$$

All plane waves of 'energy' less than the cutoff energy E_c are used in the expansion. This guarantee is incredibly valuable: ask anyone who works with a local basis set. The mathematical simplicity of plane waves means the method is easier to implement, crucially so for the calculation of ionic forces which adds little complexity or cost to the calculation. Equally important in this context is the originless nature of plane waves. Their independence from atomic positions means that the forces do not depend on the basis set—there are no 'Pulay' or 'wavefunction' forces[16]. Even more important, new developments are easiest in plane-wave codes. An idea to calculate a property is most rapidly realised in a plane-wave basis, and even if other methods catch up in time, the plane-wave approach remains as the reference.

From a computational viewpoint the first of the disadvantages appears to be very serious. Remember that in studying condensed matter one is interested mainly in the valence electrons

[15] Completeness, in a mathematical sense, means that the members of the basis 'span all space'. A crude way of thinking this means that given enough members of the basis in the expansion, any function can be represented to arbitrary accuracy.

[16] This is true even if the basis is incomplete, as it always is in practise. Changes in cell shape and size *do* produce Pulay terms in the stress, since the basis is altered by cell changes. One should also note that there is always some error in the forces due to residual non-self-consistency, since a numerical calculation can never be fully self-consistent. This rather subtle matter is discussed in reference [8]

and what they do, and as is well known from theory, a free-electron (plane-wave) picture is not terrible for valence states. However, the valence wavefunction is far from free-electron like near atomic cores. It varies very rapidly, both because of the strong Coulomb potential there, and the requirement for the wavefunction to be orthogonal to core electron states. For the localized, tightly-bound core states matters are even worse. Millions of plane waves would be required to represent the electronic wavefunctions accurately. A combination of pseudopotentials and FFT's is used to overcome this problem. A pseudopotential replaces both the atomic nucleus and the core electrons by a fixed effective potential that in a special sense reproduces the effect of the nuclear potential and the orthogonality requirement. Crucially, we can arrange things so that the pseudopotential is weak compared to what it replaces, and therefore the curvature of the valence wavefunctions in the core regions is much lower. In addition there are fewer electrons in the calculation, because the core electron have been removed. FFT's give an efficient means of transforming the wavefunctions and charge density between real and reciprocal space. The advantage to this is that parts of the calculation scale differently in the two spaces, and therefore they are done in the 'cheaper' space.

There are two reasons why we can play with the valence wavefunctions near to ionic cores. Firstly, the details of the valence wavefunctions near atomic nuclei are unimportant to the bonds they form, and it is these bonds that are crucial in determining most material properties. Secondly, it is often true to an excellent approximation that core orbitals are unaffected by a change in the environment the atom finds itself in. If this is not true, changes in core orbitals will influence the valence electrons and hence change the properties of interest, and preclude the use of a fixed pseudopotential. This idea of *transferability* is a key one in the construction of pseudopotentials.

Modern pseudopotentials are constructed from first principles. The basic idea is to replace the real potential, arising from the nuclear charge and the core electrons, with an effective potential, within a core region of radius r_c. Certain demands are then placed on this effective potential. It must be such that the valence orbital eigenvalues are the same as those in an all-electron calculation on the atom[17]. It must also preserve the continuity of the wavefunctions and their first derivatives across the core boundary. Finally, integrating the charge in the core region should give the same answer for the pseudo-atom and the all-electron one, that is, the pseudopotential must be *norm-conserving*. A pseudopotential that satisfies these demands also has the same scattering properties, at energies corresponding to valence eigenvalues, as the ionic core it replaces. In fact, norm conservation ensures the scattering properties remain correct away from the eigenvalues to linear order in the energy [14]. Norm conservation also ensures that the electrostatics of the pseudoatom are approximately correct outside the core region. Note that for an atom spherical symmetry leads to angular momentum quantization, and therefore for each value l the construction may be done differently. If it is, the potential is said to be 'non-local'. A non-local potential acts differently on wavefunctions of different angular momenta.

There are two problems with all this. Firstly, what valence electronic configuration does one work with? Ideally, if the pseuodpotential were completely transferrable, it would not matter. In practise, for some elements, and especially for norm-conserving pseudopotentials, one obtains a markedly different pseudopotential from a neutral atom and an ionized one. Clearly this does matter. Take NaCl as an example. To a good approximation the the crystal contains Na^+ and Cl^- ions, not neutral atoms. As a more extreme example, oxygen in MgO, UO_2, TiO_2 and many other oxides can be viewed as an O^{2-}ion. Like it or not, because

[17] Of course, the two calculations are performed within the same theory. Here that means DFT with an appropriate choice of exchange-correlation functional.

pseudopotentials are different for different valence electronic structures[18], they have to be constructed according their target environment[19]. A related question is, which electrons are core electrons and can be pseudized away? In some cases this is a clear-cut matter, but in others, notably the transition metals, the overlap of valence (bonding) and core states blurs the distinction. Of course one can always treat these 'semi-core' states as valence states but the pseudopotential will be correspondingly stronger and E_c higher[20]. The second problem is that one can never guarantee that valence eigenvalues for pseudopotentials are identical to the all-electron eigenvalues across the entire range of valence-band energies. This means that even in the most favourable cases there is still some approximation in using a pseudopotential constructed from an isolated atom in condensed-matter calculations.

Though successful in many cases. norm-conserving pseudopotentials for first-row and transition-metal elements require very large basis sets and high values of E_c, despite the best attempts to optimize their performance [25, 26]. Vanderbilt [27, 28] realised that by relaxing the norm-conservation requirement for the valence wavefunctions, *ultrasoft* pseudopotentials could be generated, requiring far fewer plane waves for a given accuracy. In his scheme the charge within the pseudopotential core comes from a 'hard' augmentation function, and the 'soft' valence states. Relaxation of norm conservation loses the guarantee of correct scattering properties to linear order in the valence eigenvalues. To compensate, ultrasoft pseudopotentials use several reference energies, usually two or three, that span the valence band, with a set of projectors for each reference energy. Experience seems to show that ultrasoft pseudopotentials are much more transferable than their norm-conserving counterparts. In fact, it is simply not necessary to use anything other than the neutral atom in the generation of ultrasoft potentials. Typically E_c is half that for a norm-conserving pseudopotential, which means less than one-third as many plane waves are required. Even with multiple projectors, ultrasoft pseudopotentials are significantly more efficient.

The forgoing discussion should have revealed that the construction of pseudopotentials is a non-trivial business. After E_{xc}, the use of pseudopotentials it is the next most serious source of approximation in the method—and if the pseudopotentials are badly-constructed it becomes the major source of approximation. However, careful construction and testing will produce accurate potentials for any element, provided that one pays any concomitant computational cost.

3.2. Grids and FFT's

Real- and reciprocal-space grids are another key feature of the PWP method, and a brief digression is needed to discuss them. Expressing the wavefunction as an expansion in a finite set of plane waves leads naturally to the idea of a reciprocal-space grid. However, it is advantageous to have a real-space representation too, on the related real-space grid. FFT's are used to transform the data between the two spaces in a highly efficient manner.

We denote the direct lattice vectors (the sides) of the real-space supercell $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ and $\boldsymbol{a}_3$. The reciprocal lattice vectors $\boldsymbol{b}_i$ are defined by the relation $\boldsymbol{a}_i \cdot \boldsymbol{b}_j = 2\pi\delta_{ij}$. In practise we construct the $\boldsymbol{b}_i$ using

$$\boldsymbol{b}_1 = \boldsymbol{a}_2 \times \boldsymbol{a}_3/(\boldsymbol{a}_1 \cdot \boldsymbol{a}_2 \times \boldsymbol{a}_3)\ , \tag{10}$$

[18] The cart is before the horse here. The physical reason is that the core electron configuration changes, albeit slightly, depending on the environment.

[19] Confusingly, this does not mean the pseudoatoms must be fully ionized. Rather, it is often enough to consider a partly-ionized pseudoatom. Neither is it necessary to use the same electronic configuration for all angular momentum channels.

[20] Alternatively, so-called non-linear core corrections may be applied to E_{xc} [24]. This is much cheaper, but less accurate.

$$\boldsymbol{b}_2 = \boldsymbol{a}_3 \times \boldsymbol{a}_1/(\boldsymbol{a}_1 \cdot \boldsymbol{a}_2 \times \boldsymbol{a}_3) \,, \tag{11}$$

$$\boldsymbol{b}_3 = \boldsymbol{a}_1 \times \boldsymbol{a}_2/(\boldsymbol{a}_1 \cdot \boldsymbol{a}_2 \times \boldsymbol{a}_3) \,. \tag{12}$$

A reciprocal lattice vector $\boldsymbol{G}$ is given by

$$\boldsymbol{G} = n_1\boldsymbol{b}_1 + n_2\boldsymbol{b}_2 + n_3\boldsymbol{b}_3 \,, \tag{13}$$

where n_i are integers. A plane wave $\exp(i\boldsymbol{G} \cdot \boldsymbol{r})$ is commensurate with the supercell, and the set plane waves whose wavevectors are defined by equation (13) is an orthogonal set. The real-space grid is formed by dividing the lattice vectors $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ and $\boldsymbol{a}_3$ into N_1,N_2 and N_3 points. A point in the supercell is then denoted

$$\boldsymbol{r}(l_1, l_2, l_3) = \frac{l_1}{N_1}\boldsymbol{a}_1 + \frac{l_2}{N_2}\boldsymbol{a}_2 + \frac{l_3}{N_3}\boldsymbol{a}_3 \,, \tag{14}$$

where the l_i are integers in the range $0 \leqslant l_i \leqslant (N_i - 1)$. We can view the real-space grid as the lattice of points for the lattice vectors $\alpha_i = \boldsymbol{a}_i/N_i$. The corresponding reciprocal lattice vectors are given by $\beta_i = N_i\boldsymbol{b}_i$ because of the relation $\alpha_i \cdot \beta_j = 2\pi\delta_{ij}$. The vectors β_i are the reciprocal-space supercell vectors. The reciprocal-space grid is the lattice of points for the vectors $\boldsymbol{b}_i$. Within the reciprocal-space supercell a point is given by equation (13) with $0 \leqslant n_i \leqslant (N_i - 1)$. In each supercell there are $N_1N_2N_3 = N$ points. These relationships are depicted in figure 1. One could say that discrete Fourier transforms, or at least plane waves, impose these relationships between the grids. The products $\boldsymbol{G} \cdot \boldsymbol{r}$ are independent of the supercell dimensions.

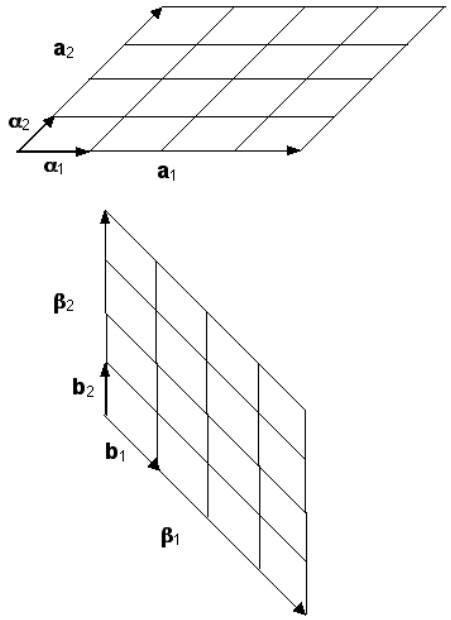

Figure 1. Real- and reciprocal-space grids, vectors and supercells. For clarity the figures show two-dimensional spaces.

Equation 8 can now be seen to be a discrete inverse Fourier transform of the wavefunction from reciprocal space to real space on grids of N points. The forward transform gives the the wavefunction on the reciprocal-space grid:

$$u_{n,k}(\boldsymbol{G}) = \frac{1}{N}\sum_{\boldsymbol{r}} \psi_{n,k}(\boldsymbol{r}) \exp(-i\,(k + G) \cdot \boldsymbol{r}) \,. \tag{15}$$

The summation goes over the N points on the regular real-space grid.

Note that quantities expressed as functions of $\boldsymbol{r}$ or $\boldsymbol{G}$ on the grids, that is, at points given by equations (13) or (14), are related *exactly* by the transforms of the kind in equations (8) and (15). If their values are exact at the points in one space, their Fourier transforms will be exact at the points in the other space. In the PWP method this turns out to be a great advantage. This is because the calculation can be split into terms that are easier to calculate in reciprocal space or in real space, and the wavefunctions and the density are transformed between the spaces to take advantage of this. For example, the evaluation of the electron kinetic energy is a simple (diagonal) sum over the wavefunction coefficients in reciprocal space. However, this would be of no use if the computational cost of the transforms outweighed the advantage of working in a particular space. This is where FFT's come in. Fast-Fourier transform algorithms scale as $N \log_2(n)$ instead of as N^2 for a direct application of equations (8) and (15). A typical grid size in a PWP calculation is $N \sim 100^3$, so the difference in scaling has a huge effect on the computational workload. In addition, FFT algorithms, because they are widely used, are usually available in highly optimized form within numerical libraries. In summary, although pseudopotentials reduce the number of plane waves required that number is still large: FFT's play a role of equal importance because they allow the calculation to scale well with system size.

3.3. Finding the ground state

The original Car–Parrinello paper emphasized the use of an 'extended Lagrangian' in which the wavefunction coefficients $u_{n,\boldsymbol{k}}$ are treated as dynamical variables[21][22]. This is a very smart idea, leading to a unified scheme for treating the electronic structure and the motion of ions in a single molecular dynamics (MD) framework. The formalism could be used for MD of ions and electrons simultaneously, or with some form of damping was applied to the electronic degrees of freedom, for minimization of the total energy. What turned out to be the key feature was the use of minimization to reach the electronic ground state, rather than direct diagonalization (of the KS equations) used in traditional self-consistent field (SCF) methods. Many workers use variants of conjugate gradients minimization instead of the simulated annealing of the electronic variables proposed by Car and Parrinello [2, 6, 7]. The arguments about which is better are still going on [6, 8], but the truth is that either can possess the advantage depending on the system under study and the kind of calculation, and both are vastly more efficient than diagonalization of a huge matrix of plane-wave coefficients. Note that DFT relies on the variational principle, and the KS equations are derived from it. When a system is at its variational minimum the solutions of the KS equations (the wavefunction coefficients) are self-consistent ones.

As for FFT's, it is better scaling with system size that gives minimization methods their advantage. The computational cost and the storage requirement of matrix diagonalization scale as N_{pw}^3 and N_{pw}^2 respectively, where N_{pw} is the number of plane-wave coefficients. For conjugate gradients minimization the scaling becomes N_{pw}^2 for the computation and N_{pw} for the storage[23]. Typically, $N_{pw} = 100 \times N_I$, the number of ions. All of this means that where

[21] Associating fictitious masses with the electronic coefficients allows one to write equations of motion for them. Moreover, if the masses are small enough to lead to separation of the ionic and electronic frequency spectra, adiabatic dynamics may be achieved, keeping the system near to the Born–Oppenheimer surface automatically. These ideas are subtle: for further discussion see references [5] and [8].

[22] It's a slight irony that nowadays first-principles calculations are associated mainly with plane waves, pseudopotentials, FFT's and minimization, *not* fictitious dynamics.

[23] The same scaling holds for fictitious dynamics, though for a given system the cost of a 'step'—a minimization

conventional matrix diagonalization was limited to systems of a few atoms, minimization approaches could handle hundreds.

Some elements of traditional SCF methods have recently been 'rediscovered' and applied in the PWP approach. In particular it has been found that density-mixing schemes greatly increase efficiency, especially for metals [6]. In these schemes the update of V_{eff} does not use the new density given by the KS wavefunctions in its entirety. Rather, the new density is mixed with the old, according to one of the various prescriptions that exist for this mixing [6]. For a fixed V_{eff} CG minimization may still be used, of course. The disadvantage of density mixing schemes is that they are non-variational. The most annoying consequence of this is that the ionic forces may be very hard to converge.

3.4. Computers

Aside from all the ingenuity spent on theories and algorithms, there is another factor which drives forward first-principles applications: computers just keep on getting faster. A cheap personal computer of today is close on six orders of magnitude quicker than the best research machine available in the 50's. Roughly speaking, processors have doubled in speed every 2–3 years, and this trend shows no sign of abating. PWP is memory-hungry, and that too is cheap and plentiful since manufacturing efficiency and competition crashed prices in the 90's. However, the advent of parallel machines was perhaps the most significant hardware advance for first-principles work. Linking N processors together[24] to perform a single task is an obvious attraction, since in principle the available memory increases N times and the computational time decreases by $1/N$. The PWP approach works well in parallel, as first demonstrated with the CETEP code [23] running on early parallel computers. Useful scaling up to hundreds of processors is possible, and the biggest, most difficult calculations of the day have always been tackled on parallel computers. Groundbreaking work in the 90's [29, 30], the decade which saw the coming of age of the first-principles approach [31], relied to a large extent on the development and use of parallel machines.

3.5. Why the plane-wave pseudopotential method was and is important

The PWP approach is a way of applying DFT that is ideally balanced for the study of condensed matter. For large, practical calculations the PWP approach is accurate, general, robust and efficient in the right measures. The PWP method marks a watershed in first-principles simulation. To see this, take the now-remarkable fact that at the time Car and Parrinello wrote their paper, in their words 'the theoretical prediction of equilibrium geometries ... still remains an unsolved problem in most cases'. PWP approaches broke the log-jam that was preventing large-scale DFT applications, so that today these calculations are often the primary source of structural and other data. More than this though, our ability to perform reliable first-principles calculations on realistic systems is changing the way we think about condensed matter.

along a CG search direction or an MD step—is different, as is the number of steps required.

[24] In the early days of parallel computing it was hoped that N could be 10's of thousands. In practise most machines have fewer than a thousand processors, because inter-processor communications time and inherently serial tasks eventually dominate, limiting useful scaling. Non-uniform memory architecture (NUMA) is a promising scheme to allow the effective use of more processors.

4. Running the program

Having examined the anatomy of the PWP approach I want to turn to the question of how to apply it, what kind of results one obtains, and how to use them. This is a hopelessly big subject, and a difficult one to discuss. In a sense it is like trying to describe good experimental technique: the description is rather elusive but the thing itself is easily recognized. Instead of getting bogged down in technical details I shall try to explain *what simulation is* and *what it is used for*, in the hope that this will reveal the guiding principles of the field, for example, why setting up and testing is so important in establishing the validity of results. I shall also look at a few very simple elements of building a model system and setting up a simulation run.

It is worth a reminder that we use simulation because of the many-body problem: the interactions between atoms, between charged particles and so on, make it impossible to solve the equations describing these interactions. Once we have the means to solve these equations there are two main uses of simulation:

- To test approximate analytic theories
- To measure the properties of 'real' materials

In principle these aims are compatible, but in practise they are differentiated by the 'working substance' that goes into the simulation box. By this I mean the description of the interatomic interactions that is used. An example will help to illustrate this. One of the first applications of computer simulation was to liquids, and in particular to testing the predictions of analytic theory. This does not necessarily require a good description of a real liquid: rather, the interactions between molecules must be simple enough for the pencil-and-paper theory to be tractable. In fact a lot of work of this kind used 'hard spheres' as the working substance. In a hard-sphere material the only interactions are at a single sphere separation where the interaction energy changes stepwise from zero to infinity. The simulation yields essentially exact results for this model substance, and in a sense it takes the place of experiment. Towards the other end of the spectrum one can use simulation to measure the properties of real materials. By this I mean the simulated material behaves in the same way as the real material, and we use the simulation to measure properties of interest. Here the premium is placed on getting the interatomic interactions right. One route to this end is to construct better, more sophisticated potential models of the interatomic interactions. This can undoubtedly lead to a *realistic* description of a material, and simulations based on this technology provide a huge amount of insight into real materials. However, there is an important difference with the first-principles approach because we take the best description of electronic behaviour we have, quantum mechanics, and make enough controlled approximations for it to be possible to find solutions for assemblies of atoms. In this sense we are trying to make the working substance *real* rather than merely realistic. Of course, first-principles simulations retain the 'computer experiment' quality discussed earlier, in that we must use the simulation as an instrument to probe the properties of the substance in the simulation box. Many of the technical details of a simulation have a direct analogue in experimental technique. However, we are also furnishing the solutions of theory: the difference compared with traditional theory is that they are arrived at using numerical methods. The fact that simulation is both theory and experiment rolled into one has led to rather slow acceptance and even scepticism in both camps [32].

4.1. What comes out

In first-principles simulations there are, roughly speaking, two main classes of calculation: *static*, or total energy, and molecular dynamics. This reflects the fact that across the entire range of atomistic simulation methods the basic output is the energy of an assembly of atoms in a given

configuration, and perhaps the forces on the ionic cores, and what one does with these is largely the same regardless of the particular approach being used. Of course a first-principles approach also yields the electronic structure, which is usually of interest in its own right. Therefore, from a static calculation one obtains, at a literal level, charge densities, Kohn–Sham orbitals, relaxed ionic positions or forces and so on. However, the single most important number printed out is the total energy. This is because it enables one to discriminate between structures and to predict stability. This means that, for example, one can compute crystal parameters and surface structures, or calculate energy differences between different arrangements of atoms. The latter is exploited extensively to calculate, for example, adsorption energies, point and extended defect energies, and relative phase stabilities. One can also consider the change in total energy with respect to various quantities, leading to elastic constants, phonon frequencies, reaction barriers and much more[25].

Molecular dynamics is a simple but powerful idea. The ionic positions and forces for a given configuration at time t are used in the integration of Newton's equations of motion. The integration must be piecewise: the forces are assumed constant over a small time interval δt. For the new configuration, at time $t + \delta t$, new forces are computed. This whole process is iterated as many times as is required (or can be afforded!) to generate ionic trajectories. From these we use statistical mechanics to link the microscopic motions to macroscopic observables. These range from simple quantities like temperature (the average kinetic energy), through structural measurements like the radial distribution function or the structure factor, to transport properties including diffusion coefficients, thermal conductivities and viscosities, and fundamental thermodynamic entities including the chemical potential and free energies.

First-principles molecular dynamics (FPMD) is defined by the use of a first-principles method to compute the ionic forces. However, there is a further distinction based on the way the system is evolved in time. The original Car–Parrinello scheme introduced the very appealing idea of treating the electrons as dynamic variables. This is done by associating a mass with them, and allows the writing of an extended Lagrangian and hence equations of motion. The second part of the trick is to make the fictitious electron masses small enough to keep the vibrational spectra of electrons and nuclei separate. Finally, the adiabatic principle is employed by arranging the initial conditions so that the electrons are in their ground state, and the electronic degrees of freedom are 'cold' compared to the ionic ones. Then, when the ionic cores move along their trajectories, the electrons follow adiabatically, staying very close to the ground state. A different approach is followed in Born–Oppenheimer dynamics, where the electronic energy is minimized for every successive configuration. There are pros and cons to both methods: fictitious dynamics has the advantages of stability and error cancellation [5], a smaller computational task at each iteration, and consistent forces [8], while Born–Oppenheimer methods can employ much longer time steps to offset the greater cost of minimizing at each configuration, and can use a variety of wavefunction and density projection methods to improve performance. The fictitious dynamics scheme runs into serious difficulties with metals since it is then impossible to stop rapid energy transfer from the ionic cores to the electrons without recourse to thermostats or similar remedies [5]. However, both methods can be made to work very well, and the advantages or weaknesses are often exaggerated.

[25] Calculations that involve a response to a perturbation are in general best done within a linear-response framework [33–36] A simpler approach is often used, relying on computing energy differences for finite displacements, as for example in frozen-phonon calculations.

4.2. What goes in

I have spent a lot of time describing the electronic structure theory that underpins a first-principles simulation. One could take away the impression that the performance of the functional was the limiting feature of a calculation, but this is very rarely true: it is in much more mundane matters that the limits lie. In rough order of their effect on precision, these are

- The simulation supercell and its contents
- How entropy is taken into account
- Pseudopotentials
- The basis set and other computational tolerances

These make up the framework of the calculation: no matter how accurate our electronic theory is, it is close to useless if this framework cannot support it. To take a pertinent example, many surface studies are seriously weakened by a model system, usually a slab, that is too thin, or worse still by failure to relax the structure to mechanical equilibrium. Of course we would like to make all model systems so large that any effects due to their finite size were negligible, but this is not possible. On the other hand we cannot sail too close to the wind, and we must resist the temptation to draw conclusions that we doubt would survive if we did the calculations on a better model system with tighter tolerances.

What I am definitely *not* saying is that our model systems must capture every feature of the real systems they correspond to. This is just not possible. What we always have to accept is that we work with an idealized system which should capture the essence of the scientific problem. Take surface studies as an example. Surface scientists experiment on carefully-prepared crystals under ultra-high vacuum, but even under these special conditions the surface is far from the theorist's ideal. It is stepped, full of defects, probably covered in hydrogen and other impurities, and at 100–200 K, whereas our models are flat, defect-free and almost always at zero K. The value in studying the model system is thus to make *definitive statements about an idealized system*. To be able to make such definitive statements we must follow some guiding principles in building and exploring a simulation model. First and foremost, we must do exactly what has just been said: we must explore the predictions of a model, without attempting to fit its properties to experiment or to our own intuition. Unlike some other approaches, first-principles simulation is not a fitting and extrapolation exercise. Neither is the main point to 'agree with experiment'! Going hand in hand with this is the need to understand how all the various choices we have made in setting up the calculation affect the results. We already have made a major choice by using DFT with an approximate functional, and this carries implications for the accuracy of the calculation. Before considering these though, the precision of our results, in other words how uncertain we are that 'the total energy is x eV', must be determined. Our results will vary with system size, number of $\boldsymbol{k}$-points, choices in constructing the pseudopotential, the tolerances applied to minimization of the electronic energy or in structural relaxation, and so on. These model parameters must be varied to improve precision in the quantities of interest, so ideally one tests the variation in the latter with respect to the former. Of course it is not always possible to do the whole calculation many times with different parameters, and instead suitable quantities (the unit cell parameter, bond lengths, surface energies, a single vibrational frequency . . .) are used to test the setup. All of this is just good scientific practise in the context of computer simulation. Having quantified the uncertainty in our results, we are finally in a position to say what the model predicts. This is both a statement about the model system, and about DFT applied to that model.

4.3. Trump cards

A great strength of all simulation approaches is that no thermodynamic conditions, no matter how extreme, are out of bounds. Simulations can be performed at any pressure or temperature. This is always useful, but in certain situations it is crucially important. For example, earth science requires knowledge of mineral phases and compounds under pressures as great as 350 GPa found in the Earth's mantle and core, but the best diamond-anvil cells can only reach around 150–200 GPa. High-pressure phase diagrams determined from first principles are therefore elevated to the status of primary or even unique information. The same applies to temperature, where examples might include nuclear reactors (2000–3000 K) and the Earth's core (∼6000 K). These examples also show how simulations may be applied to systems that are physically inaccessible.

A related advantage is that every conceivable atomic-scale detail is available to the simulator. As a simple example, the precise positions of the ionic cores is trivially known from a calculation, while experiments must always use indirect methods (x-ray or electron diffraction, neutron scattering) or probes (STM and AFM, ion scattering). These comments apply equally well to electronic properties. In general it is of course possible to relate this microscopic data to observables, and the onus is on the simulator to exploit this link. Access to atomic details need not be passive, and any number of thought experiments may be played out on the computer. Good examples include application of constraints and searching for transition states, and 'numerical alchemy', where atomic identity is transmuted in the course of free energy calculations [37]. Thus, even if we are interested in the physical properties of real materials, we need not follow a physical path in their determination. The trickery employed by the simulator in this context is a major part of the art of the approach.

4.4. What we learn

Despite the remarkable advances of recent years, I think it would be wrong to hope for a time when we could plug any system into our codes and perform a perfect simulation, thereby usurping experiment forever. Such a pursuit would not be science in any case: do we really want to be reduced to saying 'because the computer says so' when asked why? What we really seek is more powerful ways of thinking that lead to new ideas and concepts. Simulation is increasingly able to help us find such insights.

First-principles simulation studies, or better, the conclusions from them, may be divided into three categories. There are studies that provide reference or benchmark data for ideal systems, ones which reveal new ideas or mechanisms, and ones which are genuinely predictive. It goes without saying that in all cases we need precise answers, but only in the first case are we also seeking absolute accuracy. Some examples will help to make these points.

Great strides have been taken in determining surface structures from first principles. A spectacular success and a real milestone in the story was the determination of the 7×7 reconstruction of the (111) silicon surface in 1992 [38, 39]. Another good example is the explanation of anomalous thermal contraction of the Al (110) surface [40]. Structural studies of oxide surfaces are particularly useful because oxide surfaces are the least well characterized by experiment. Oxides are inherently complex crystallographically, and the majority are good insulators. In addition, transition-metal oxides readily depart from stoichiometry. These factors make experiments difficult. The (110) surface of TiO_2 must be the most studied of all oxide surfaces, yet the details of its structure remain controversial. The debate centres on the position of the so-called bridging-oxygen atom relative to the substrate (see figure 2). A series of first-principles studies [41–43] had asserted that the distance between the

bridging-oxygen atom and the nearest titanium atom, was around 0.3 Å longer than found in experiment. All these calculations employed quality methods, large slabs to model the surface and full structural relaxation. In the details of the other surface atom positions no such discrepancy existed. Muscat *et al* undertook rigorous benchmark-style calculations to resolve this question [44]. By careful application of DFT, quantifying all uncertainties for their calculations, they arrived at a surface structure which could honestly be labelled as the 'DFT answer'. This answer still disagreed with experiment, but of course the difference was that a fixed point had been established, providing a platform for further investigations. These revealed a surprising answer to the riddle. Soft anharmonic modes involving the bridging oxygen atom are the likely cause of mis-interpretation of the experimental data [44].

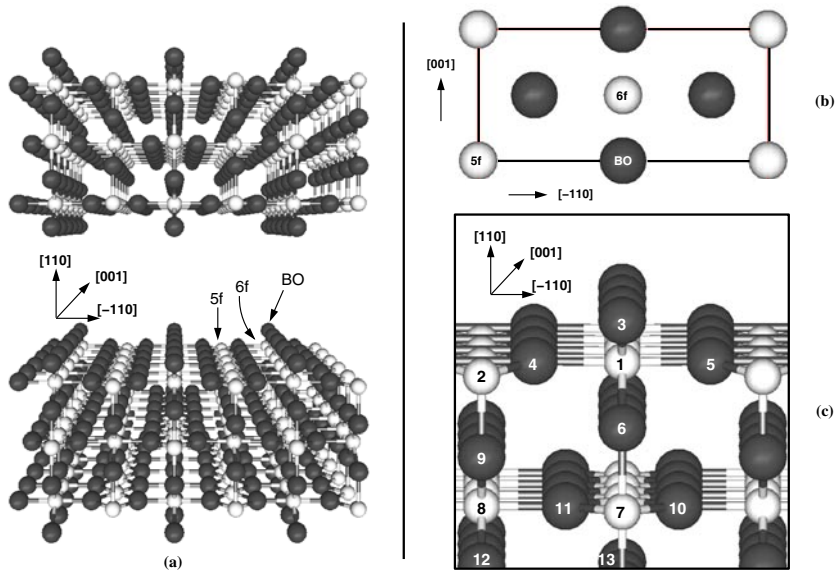

Figure 2. The (110) surface of rutile. Light and dark spheres indicate titanium and oxygen ions respectively. (a) perspective view showing slab geometry. (b) The (110) surface unit cell. The fivefold- and sixfold-coordinated titanium sites and the bridging oxygen site are labelled 5f, 6f and BO respectively. (c) Side view. The relaxation of the labelled ions is discussed in reference [**?**]

A good recent example of completely predictive work comes from Mike Gillan's group and their work on the properties of the Earth's core. Of course, such studies are bound to be predictive because they treat systems that experiment cannot access directly. As mentioned, simulations can be performed under any conditions, and here that strength is put to good use. One of the key questions addressed was 'what is the viscosity of the Earth's core?'. It is an important question because seismic measurements rely on such basic data on the core and mantle, and theories of the Earth's magnetic field depend directly on the properties of the core. A remarkable range of answers was to hand, spanning no fewer that twelve orders of magnitude. Using first-principles MD, Gillan and co-workers calculated the viscosity of liquid iron under Earth's core conditions, and found that their model gave an answer near the lower end of that range [45]. To be fair, there are a huge number of factors which could affect this answer: a small model system, neglect of the effect of impurities, technical details of

the calculations such as the treatment of d electrons and the length of the simulations, and so on. Nevertheless, even allowing for large uncertainty in their results, they have succeeded in closing by many orders the range of uncertainty in the real value. This in turn will allow progress through greater confidence in a key piece of geophysical data.

The case study that follows should illustrate how mechanisms nay be revealed by first-principles simulation, so I'll give little further discussion here.

5. A case study: water chemistry at an oxide surface

Interfaces between oxides and water or aqueous solutions are ubiquitous, occurring in the natural environment, for example in soils and aquatic systems, and in technological applications such as catalysis and gas sensing. The interfacial chemistry is pivotal to a huge diversity of phenomena and processes, ranging from the weathering of rocks to electrochemistry. For a variety of reasons it is also very challenging to study. The upshot is that our fundamental knowledge is poor for these extremely important systems. Here I hope to show that first-principles studies can be used to enrich our understanding of oxide surface chemistry. I will highlight some new ideas that have emerged, and demonstrate the distinct but complementary role that simulations can play alongside theory and experiment. This case study illustrates many features of the simulation approach: there is re-interpretation of experiment; new ideas are generated and some prediction attempted; a range of techniques and styles of calculation are employed; there is the need to understand how approximations affect results; simulation results are connected to experimental ones; and it shows what conclusions may safely be drawn even at the limit of accuracy of the method.

This study is focused on titanium dioxide surfaces. It addresses two main questions: how does water adsorb on the (110) surface, and can apparently conflicting experiments be reconciled?

The water–TiO_2 system has attracted more attention than any other comparable system, starting from the observation of the photoelectrolysis of water on rutile TiO_2 surfaces [46]. However, TiO_2 has now become a model transition-metal oxide system for both theory and experiment. A vast effort has gone into studying the (110) surface of the rutile-structured form of TiO_2 (see figure 2). All the calculations here deal with this surface. Titanium dioxide is more than just a scientific curiosity though. It is widely used in powder form as a white pigment and opacifier, and is found in paints, plastics, paper, cosmetics, foodstuffs and sunblock. It is also used as a supporting material for metal catalysts. Recently, many interesting new applications of the material have been found, for example in biological waste treatment, solar panels and self-cleaning glass coatings, all of which rely on the photocatalytic properties of TiO_2 nanoparticles.

Titanium dioxide has a 3.1 eV bandgap [47], making pure crystals transparent to visible light. This is what makes TiO_2 powders such good white pigments, the tiny crystals scattering across the visible spectrum by internal reflection. It is a d_0 transition-metal oxide: the valence band is formed from O2p states, and the conduction band is predominantly Ti3d at its lower edge. Under ambient conditions, TiO_2 is commonly found in either the rutile or anatase structures, though its full phase diagram is quite rich [48]. The structural motif is the octahedron formed by the six nearest oxygen neighbours surrounding the titanium atoms. The rutile structure, shown in the figure, is orthorhombic. The rutile (110) surface is shown in figure 2. Some details of its geometry will turn out to be very important. The bridging-oxygen (bo) atoms sit proud of the surface by about 2 Å making the surface corrugated. They conceal sixfold-coordinated (5f) titanium atoms, which lie in-plane with fivefold-coordinated (5f) titaniums and some in-plane oxygens. On chemical grounds it seems reasonable that the

under-coordinated 5f site should be basic and therefore a good candidate for water adsorption, while the bridging oxygens may be expected to be acid sites.

From experiment, it has variously been proposed that on the (110) surface, water adsorbs molecularly and only dissociates at defect sites [49], dissociatively at low coverages and thereafter molecularly [50], either molecularly or dissociatively depending on temperature [51], or that the surface is inert [52]. One point of consensus is that if dissociation does occur, it is only at low coverages. In contrast, theorists were unanimous in predicting dissociation at all coverages [53–57]. They all looked at the adsorption of a single water molecule at the titanium 5f site, employing a range of first-principles techniques.

Over the last couple of years my co-workers and I have pursued a programme of work aimed at unravelling this question. The starting point was to check that previous theoretical work had not simply missed a possible adsorption site. To do this we used FPMD to explore the surface and its interaction with water. This is a simple idea: one chooses a variety of initial configurations, in this case placing a water molecule above various positions on the surface and in a variety of orientations. Then the system is 'let go' and the molecule finds its own way down to the surface, or sometimes is repelled into the vacuum. This approach can greatly increase our confidence in having found all possible sites for the molecule, but for it to work a number of conditions must be met:

- It helps if barriers on the energy surface are small compared to the temperature
- There must be relatively few degrees of freedom to explore (mainly molecular conformation)
- A few initial configurations must be enough to cover most possibilities

In the present case we looked at the adsorption of a single molecule in a 2×1 cell of the (110) surface using three starting points. The choice of these starting configurations was guided by chemical reasoning: the first hopes to see the molecule react at the 5f site, the second tries to access the 6f site, and the third looks for hydrogen bonds with the bridging oxygen atoms. Each of these was run for 0.5–1.0 ps, and in fact only the first yielded a bound molecule. We took the end configuration from this run and minimized the structural energy by relaxing all the ions to zero-force positions. The resulting geometry is shown in figure 3.

This certainly looks like a dissociated state, and therefore ties in with previous theory. There are two hydroxyl groups, one formed from a bridging oxygen and donated proton (bridging hydroxyl or BH), and one from the water minus a proton (terminal or TH). So far nothing is different, and all the discrepancies remain. However, we can go further in the analysis. Experiment does not produce pictures like figure 3, but infers structures from indirect measurements. One such technique is high-resolution electron energy loss spectroscopy (HREELS), which measures the vibrational spectra of species adsorbed on surfaces. Mike Henderson had performed HREELS measurements on the water-rutile (110) system, and to compare we used FPMD to compute the power spectrum for the vibrations of the OH groups[26]. The results is shown in figure 4, and reveals the first surprise. OH groups are strongly bound, though the vibrational frequency of their stretch mode would be affected slightly by their environment. Therefore, the most obvious expectation would have been to see two sharp, high-frequency features in this spectrum. Instead, only one of the OH's gives this sharp feature, the other contributing a much broader and fuzzier signal.

This was real progress, since the high-frequency region of Henderson's HREELS spectra exhibited exactly this broadening. Moreover, inspection of the MD trajectories *via* computer

[26] The vibrational frequencies of the protons are shifted in these simulations because an artificially-high mass was used for them. This is a technical trick employed to allow a longer time step to be used. It does not affect equilibrium statistical-mechanical quantities, but the shift in frequencies should be borne in mind.

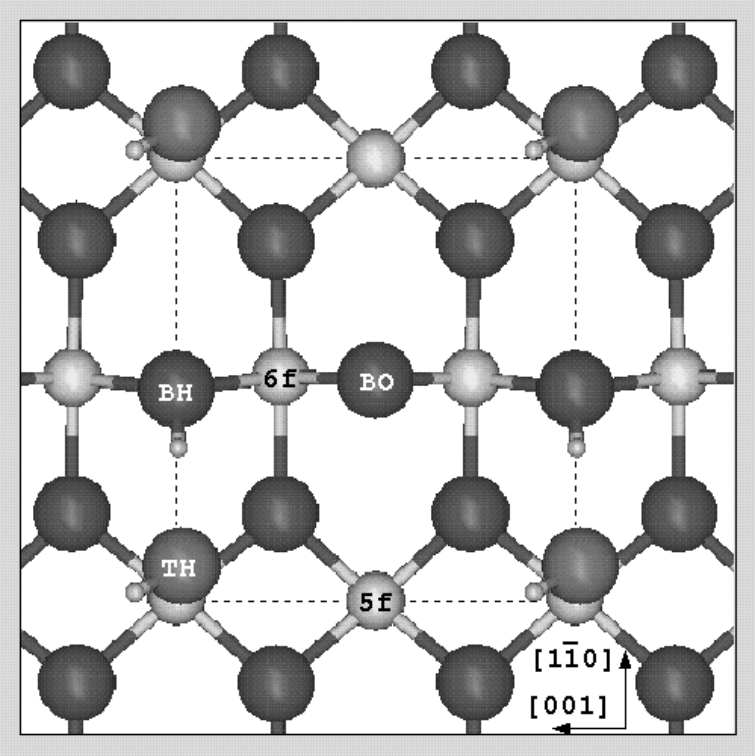

Figure 3. Geometry of a single H_2O at half-monolayer coverage on the (110) surface. Indications as in figure 2 and in addition small grey spheres represent hydrogen.

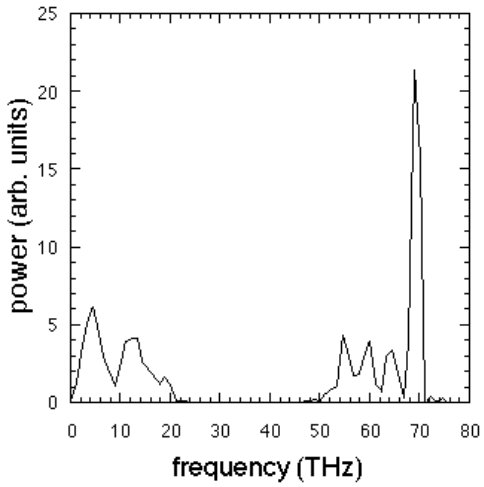

Figure 4. The calculated vibrational spectrum for hydrogen atoms in the system shown in figure 3.

graphics and 'movies' readily showed what was causing it. The H on the bridging hydroxyl was able to make considerable excursions towards the TH group, and in doing so was clearly

experiencing much softer restoring forces. Substrate vibrations were strongly involved, as they altered the distance between BH and TH groups considerably. This picture of dissociation is much more complicated than we are accustomed to. Moreover, the energy separation between dissociated and molecular forms was small: static calculations showed that the structure in figure 3 has a corresponding adsorption energy of 0.91 eV per H_2O, while H_2O in molecular form is metastable on the surface, and its adsorption energy is 0.87 eV. The experimentalists were very confident in stating that molecular water is present at almost all coverages: their most persuasive evidence came from the HREELS data, which showed a clear signal from the medium-frequency 'water wave' vibrational mode in which the two 'arms' of the water molecule move in opposite directions. This is entirely absent in figure 4, where it should appear around the centre of the plot. We have to ask at this point whether theory and experiment are in agreement. Although there is a *metastable* molecular state, it remains that the most stable state is dissociated, and its key signature, its vibrational spectrum, does not agree with experimental measurements.

What are the possible resolutions of this situation? Perhaps the barrier between the dissociated and molecular forms is large, and most H_2O is trapped in the metastable molecular state. Perhaps the adsorption state is in some way coverage dependent. Perhaps intermolecular interactions affect the adsorption state. Whatever the case more work was required.

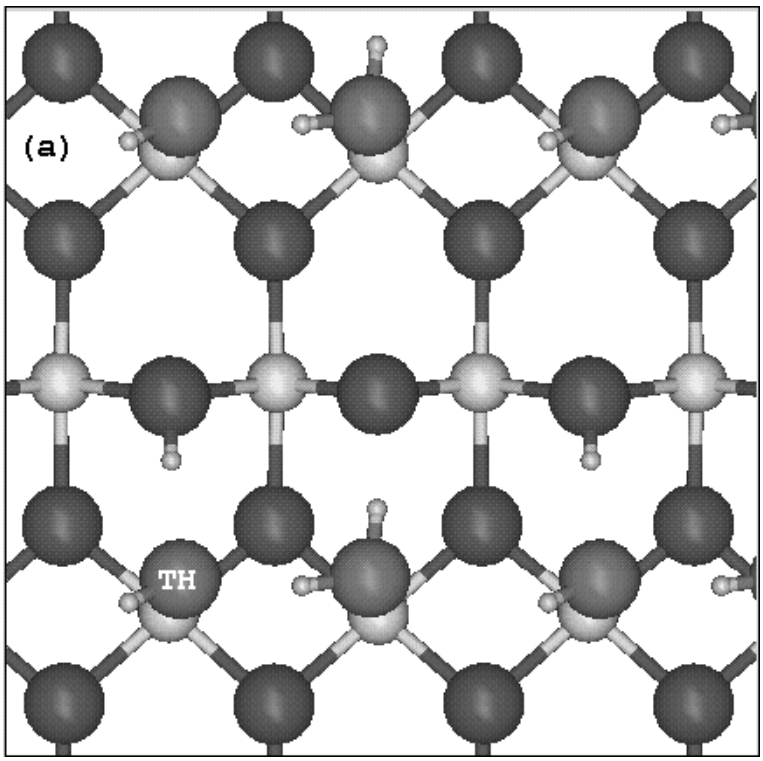

Figure 5. Geometry of two H_2O adsorbates at monolayer coverage on the (110) surface. Indications as in figure 3.

We chose to work on the hypothesis that inter-molecular interactions are important [58–60]. The next step therefore was to increase the coverage to one monolayer by adding a second molecule to the simulation. This was done as before, by choosing a number of

starting positions and tracking their evolution with MD. The outcome was the adsorption geometry shown in figure 5. This was quite unexpected, the second molecule adsorbing intact because of its strong hydrogen bond to the terminal OH group. This configuration is stable over pico-seconds of MD at around 150 K, and subsequent static calculations confirmed it to be the lowest energy configuration for the two molecules. However, this so by only a small margin, a point I shall return to shortly. The vibrational spectra for the protons in this 'mixed' configuration are shown in figure 6, and the sought-after water-wave feature is evident at 29 THz. Thus, theory was at last giving a picture that was consistent with experiment. Starting with monolayer coverage, what are the main conclusions from the calculations? First, we have that inter-molecular interactions can affect the adsorption state of H_2O. Also, it seems possible that water can adsorb in both molecular and dissociated forms at the same time due to these interactions. Remember that there are no defects (e.g. oxygen vacancies) on this surface, so dissociation need not be linked with defect sites. However, the results do *not* show that a monolayer consists of a 50/50 mix of OH groups and water molecules. More configurations would need to be explored to shed light on this question. We might also tentatively suggest that intermolecular interactions would be important at other coverages, but again this needs more work. However, at this point we can certainly help re-interpret the experiments and to resolve the various discrepancies in the light of these findings. First of all, we can see why previous theoretical work failed. It simply did not take into account the hydrogen bonds explicitly, but instead imposed the restrictions of a 1×1 periodicity. Next, at 'low' coverage the calculations still predict dissociation, but for *single molecules*. However, it seems plausible that at low coverages the H_2O molecules will cluster because of their interactions, provided the H_2O is able to diffuse on the surface. Then the calculations suggest that some H_2O will remain in molecular form, which if true might explain the data showing that molecular water is present at all coverages. Remember too that for a single H_2O there is a metastable molecular state: an understanding of the adsorption dynamics would help to decide whether adsorbing molecules get trapped in this state. Our results do indicate that dissociation does occur on non-defective surfaces. Finally, we have seen that small energy differences separate dissociated and molecular forms, and that the protons are mobile at moderate temperatures. These suggest that the observations may indeed be temperature dependent, in other words, the adsorption state (and its vibrational and other signatures) may depend on temperature. Calculations of the free energy barrier between dissociated and molecular states would help to to solidify these speculations.

I'll return now to the small energy differences already mentioned. The mixed state of figure 5 has an adsorption energy of 1.01 eV per H_2O. The two other obvious candidate states are metastable: with two molecules the energy is 0.99 eV, and with all OH's it is 0.91 eV. The latter is sufficient to state reliably that an all-OH configuration may be ruled out, but the difference between all molecular and mixed is within the uncertainty of the method. This is despite our attempts to refine the calculations in their weakest aspect by increasing the slab thickness. The fact remains that a change in pseudopotential, functional or $\boldsymbol{k}$-point sampling could tip the balance by an equal amount in the opposite direction. One has to ask if this weakens the arguments I have just outlined. The answer is no, because the key points remain intact: these are the two ideas about inter-molecular interactions and a 'interacting' dissociated H_2O, and their consequences for the vibrational spectra. More work will strengthen this story though, and recent work on the coverage dependence of adsorption state reinforces these key conclusions as well as revealing more unexpected behaviour [61].

It is very important to realise that without the experimental data this full and appealing interpretation would not be possible. By combining evidence from theory and experiment we generate far more than the sum of the parts, and can make stronger conclusions than we could

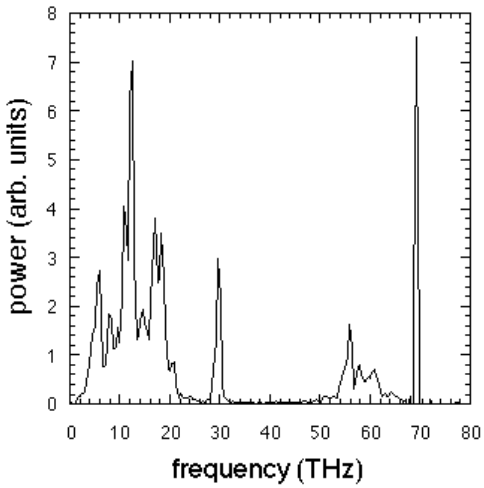

Figure 6. The calculated vibrational spectrum for hydrogen atoms in the system shown in figure 5.

from separate analyses. Exploiting this virtuous circle is a must.

6. CASTEP

Mike Payne wrote the first version of CASTEP in 1986 [2]. He took about two months to produce a fully-functioning program: cubic cell, local pseudopotentials, a simple minimizer, the total energy and the ionic forces. He employed the practises of that time, using FORTRAN77 and 'functional programming', that is, basing the program structure on the operations that must be performed on the data. By 1992 he and more than ten co-authors had added many features to the code: arbitrary cell shape, stresses, partial occupancies for metals, non-local pseudopotentials, real-space pseudopotential evaluation, molecular dynamics, parallel execution and much more. The following years saw many more authors and more developments, including gradient-corrected functionals, spin-polarization and ultrasoft pseudopotentials. By the late '90s the code stood at around 120 000 lines, all cast around the original design, all in FORTRAN77.

This is how almost all scientific codes are developed. A new idea or method is implemented in a fairly simple form as quickly as possible. Early on the need is to get working quickly so as to capitalize in the new field. Improved techniques are developed and added, and sometimes radically different new methods must be incorporated, but very rarely is a code completely rewritten. The reasons for this are as much to do with scientific careers and funding as they are with software practises and training, as we examine more fully elsewhere [62]. In a nutshell, no-one can afford to spend the 20–30 man-years required to redesign and implement a mature code, let alone do it with high-quality software engineering. However, if the code works and works well, why worry?

The problem is that development grinds to a halt. Changes start to take longer and longer, partly because more lines of code need to be altered (the changes are not localized), partly because the structure grows more and more complicated. One must know how almost everything in the code works in order to add something new, and this becomes daunting, especially for those new to the code. Debugging becomes very difficult. New ideas are

simply not explored. These are very bad things indeed because science is all about change and exploration.

So far it might seem that the argument is as follows: once methods and algorithms have become tried and tested it is worthwhile to rewrite a code in the 'right' way for them. This is certainly so, but it is only half the story. If we were to do this using the same software approach we would soon run into the same old problems once new developments were tried. The problem is this: it is very easy to write a simple program quickly in a language such as FORTRAN77, indeed that is one of their advantages. However, this does not lay a good foundation for a large code, as no design thought has been given to the full range of purpose and function the code must have. What works for a small code is terrible for a large one. Large codes must be specified and designed.

A brief digression is required here to introduce some ideas. Over the last decade 'object-oriented' design (OOD) and coding have supplanted other approaches in almost every field where large codes are required. Object-oriented designs and programs are more robust, easier to develop and debug, and far more extensible than equivalent functional or structured programs. In the context of scientific programming there are two concepts that are absolutely crucial to achieving these advantages, and they are data hiding and data encapsulation. Data hiding, and to some extent data encapsulation, are familiar ideas in mathematics. Writing $f(x)$ to denote a function of a single independent variable hides the details of what that function is, what its value is at any particular point x. In just the same way we can write a software function, a piece of code that returns $f(x)$ if given a value of x, without the user needing to see the inner workings of the function. This is data hiding. Data encapsulation goes much further than this. Continuing the analogy we know that there are many operations allowed on $f(x)$, such as addition, subtraction, integration, differentiation, and so on. If we imagine writing a piece of software that encapsulates these operations as well as the data, we have an 'object'. The user may then not only interrogate the object for the value of $f(x)$, but also $f'(x)$, $f(x) + 102.6$ or any other operation that is on offer. The programmer defines which operations may be performed on the data, and the only way to interact with the data is through these operations [63]. This is not full object orientation but it is already of immense value. Here are some reasons why:

- The data are protected from side effects
- There is no need to know the data structures at every point in a large program
- Changes are mostly local to the object
- Almost forces the designer to consider the best generalizations of operations on the data.

The argument for encapsulation becomes clear when changes to the data structures or operations upon them are considered. Since the interface with the object is its set of operations, changes inside the object *do not* require changes elsewhere. In other words, the changes are localized. In contrast, changes to the data structures in a non-OO code must be echoed throughout, wherever that structure is accessed. We note that this produces a natural reticence to make major changes to a large program, even when those changes are highly desirable.

6.1. New CASTEP

Over the last two and a half years we [64] have specified, designed and written a completely new version of CASTEP. The code is written in FORTRAN90 to a modular design. A great deal of our efforts went on getting this design right, and in fact coding did not start until almost a year after the project began. The key achievement has been to abstract, in the manner just described, the data and associated operations that are required in a first-principles calculation.

The design is complex and its description will be given elsewhere [62], but here I'll give an illustration of the style.

In the new code operations involving the ionic pseudopotentials, the wavefunctions, the grids and so on reside in separate modules. For example, there are 33 operations involving the pseudopotentials. The modules are arranged in a hierarchy, such that at the highest level virtually all of the detail of the calculation is hidden. The high-level modules are to do with the overall functions of the code and the major elements of implementation: electronic minimization, first derivatives (forces and stresses), second derivatives, band structure, and so on. Beneath these 'Functional' modules lie the 'Fundamental' modules that provide all of the PWP machinery: density, wavefunctions, ions and grids. Beneath all of this lie the 'Utility' modules: FFTs, parallel communications and others.

In principle, coding at the high level involves the use of the operations that the modules immediately below provide and nothing more: in practise, new developments may necessitate the addition of new operations to (any of) the modules. An example will help to illustrate things. In the calculation of band structures a steepest-descent vector is required, and here is a pseudocode version of the calculation:

```
! Calculate the kinetic energy and store del^2|band> in H_band
call wavinetic_energy(band,nk,ke,H_band)

! Apply the local potential
call pot_apply(pot_loc,band,nk,V_band,ener_loc)

! Add on pot_loc|band> to H|band>
call wave_add(V_band,H_band)

! Apply the non-local potential
call nlpot_apply_Vnl_ES(band,nk,ns,nl_d,ener_loc+ke,ener_nl,V_band)

! Add on V_nl|band> to total H|band>
call wave_add(V_band,H_band)

! For USP's search direction is Sinv * (H-ES)
call nlpot_apply_Sinv(H_band,nk)

! Initial search direction is negative H|band>
call wave_scale(H_band,(-1.0_dp,0.0_dp))
```

The detailed meaning of the variables and calls is not required to make the main points. The calculation takes only seven lines of code. Anyone familiar with the old way of coding may wonder where all of the code has gone to: this calculation would have taken perhaps 200 lines of code. Because of the data abstraction there is no need to be concerned with array limits or data structures for the wavefunction and potentials, nor to worry about which pieces of data need to be fetched from other nodes when running in parallel, or to remember to use library calls to do things efficiently. All of that is taken care of in the underlying modules, and the operations they offer up are completely general. Writing the code at the high level is a facsimile of writing the algebra.

This is all very elegant but what are the advantages? There must be a very substantial payback for the considerable extra effort involved, and there is: new developments are very, very much more rapid. Not only this, but developments of much greater complexity remain

completely manageable because the modules are largely self-contained and their interfaces with each other rather simple. Debugging becomes much easier: the scope of a bug tends to be within a module, and there is far less replicated code to work with. On a related theme, optimizations once done in a module are always available. All of these characteristics are absolutely critical because change is always necessary, the code is never finished, and new ideas will always come along.

I have already mentioned the long design process, but it should also be understood that this must be supported by fairly rigorous documentation. The specification of new CASTEP now runs to several hundred pages, and it was substantially complete before coding began: indeed, the coding was done *to the specification*, and this was so successful that authors in separate institutions wrote modules that were completely compatible because they adhered to the specification. Of course iteration was required, with the specification being updated constantly in the light of implementation. Another major piece of documentation is the implementation guide, which describes how the functionality of a module was achieved: the mathematics, the algorithms and the tests. A computer scientist reading this story would no doubt point out that we have merely adopted the classic software cycle using object-oriented design ideas: this is true but in the scientific programming context this is a radical step.

6.2. Is there a future for the plane wave pseudopotential method?

The argument to rewrite CASTEP is predicated on the long-term utility of the plane wave pseudopotential method. It is fair to question this view: after all, it is well known that at best the method scales as N^2 and ultimately as N^3, and that to get anywhere pseudopotentials, will all their potential problems, must be used. Many groups are striving to produce linear-scaling DFT codes with plane-wave accuracy, and they will succeed in the end. What then is the argument for the PWP method?

First of all, the vast majority of applications are well away from the bad scaling regime, and will remain so for some time. For example when doing MD calculations it is often not feasible to use a very large model system, and the electronic part of the calculation scales as N^2. A more practical restriction is that of computer size, meaning large systems are out of reach. However, we will reach the point where N^3 scaling starts to be a limiting factor. The most promising way forward is to embed the PWP system within a cheaper method, either semi-empirical or classical. This approach works well because it is very often the case that only a small region of the system requires a full quantum-mechanical treatment. Of course, for both linear-scaling codes and embedding approaches, reference calculations are done with the PWP method. Another point is that pseudopotentials can be made to work extremely well these days. Ultrasoft pseudopotentials are highly transferable, and the closely-related projector-augmented wave method [65] is another alternative. Pseudopotential core reconstruction [66] allows the computation of properties that depend on the core electrons, such as NMR. In short, pseudopotentials are accurate and non-restrictive.

By far the most persuasive argument is that new ideas in first-principles electronic structure calculations will always be tried first with the PWP method. This is because of the mathematical simplicity of a plane-wave basis. A good example is the computation of response properties. Calculations of the phonon dynamical matrix [33,34], ferroelectric phare transitions [67], solid-state NMR chemical shifts [66] and the electron G-tensor [68] have all appeared recently, and all used the PWP method. Developments that would take years in other frameworks take months in the PWP scheme.

We can of course hope for further improvements the PWP method and in the algorithms used in its application. The Car–Parrinello paper itself marked a huge step forward in methods

rather than theory, and more recently Vanderbilt's ultrasoft pseudopotentials brought a very substantial improvement to the approach. As a final point, PWP codes are probably the most 'automatic', and therefore the most accessible to non-specialists. Classical simulations on systems that are difficult to model with potentials, such as water and silicon, may now employ the PWP method as an energy and force 'engine'. At the other end of the spectrum, biosciences is perhaps the most rapidly growing area of application for PWP calculations.

It is hard to think that DFT will be replaced by a theory of equivalent accuracy that can be implemented at a fraction of the cost. Moreover, new functionals [69] are promising to improve the accuracy of our calculations very considerably. As far as it is sensible to look into the future the PWP method has an important role to play.

References

[1] Parr R and Yang W 1989 *Density-functional theory of atoms and molecules* Oxford
[2] Payne M C, Teter M P, Allan D C, Arias T A and Joannopoulos J D 1992 *Rev. Mod. Phys.* **64** 1045
[3] Jones R and Gunnarsson O 1989 *Rev. Mod. Phys.* **61** 689
[4] Srivastava G P and Weaire D 1987 *Adv. Phys.* **36** 463
[5] Remler D K and Madden P A 1990 *Mol. Phys.* **70** 921
[6] Kresse G and Furthmüller J 1996 *Phys. Rev.* B **54** 11169
[7] Gillan M J 1989 *J. Phys.: Condens. Matter* **1** 689
[8] Marx D and Hutter J 2000 Ab initio molecular dynamics: theory and implementation *Modern methods and algorithms of quantum chemistry* ed J Grotendorst, NIC Series
[9] Koch W and Holthausen M C 2001 *A chemist's guide to density-functional theory* 2nd edn (New York: Wiley)
[10] Szabo A and Ostlund N S 1989 *Modern quantum chemistry—introduction to advanced electronic structure theory* (New York: McGraw-Hill)
[11] Hohenberg P and Kohn W 1964 *Phys. Rev.* **136** B864
[12] Kohn W and Sham L J 1965 *Phys. Rev.* **140** A1133
[13] The 1998 Nobel Prize in chemistry: http://www.nobel.se/chemistry/laureates/1998/
[14] Gillan M J 1991 *Computer Simulation in Materials Science* ed M Meyer and V Pontikis (Dordrecht: Kluwer) p 257
[15] The mathematics of functionals is discussed in appendix A of reference [1].
[16] Gilbert T L *Phys. Rev.* B **12** 2111
[17] Levy M 1982 *Phys. Rev.* A **26** 1200
[18] Thomas L H 1926 *Proc. Cantab. Philos. Soc.* **23** 542
[19] Fermi E 1928 *Z. Phys.* **48** 73
[20] Ceperley D M and Alder B J 1980 *Phys. Rev. Lett.* **45** 566
[21] Perdew J P, Chevary J A, Vosko S H, Jackson K A, Singh D J and Fiolhais C 1992 *Phys. Rev.* B **46** 6671
[22] Car R and Parrinello M 1985 *Phys. Rev. Lett.* **55** 22
[23] Clarke L J, Štich I and Payne M C 1992 *Comp. Phys. Comm.* **72** 14
[24] Louie S G, Froyen S and Cohen M L 1982 *Phys. Rev.* B **26** 1738
[25] Lin J S, Qteish A, Payne M C and Heine V 1993 *Phys. Rev.* B **47** 4174
[26] Rappe A M and Joannopoulos J D 1991 *Computer Simulation in Materials Science* ed M Meyer and V Pontikis (Dordrecht: Kluwer) p 409
[27] Vanderbilt D 1990 *Phys. Rev.* B **41** 7892
[28] Lee C Y, Vanderbilt D, Laasonen K, Car R and Parrinello M 1993 *Phys. Rev.* B **47** 4863
[29] Perez R, Payne M C and Simpson A D 1995 *Phys. Rev. Lett.* **75** 4748
[30] Štich I, Gale J D, Terakura K and Payne M C 1998 *Chem. Phys. Lett.* **283** 402
[31] Parrinello M 1997 *Solid State Commun.* **102** 107
[32] Frenkel D and Smit B 1996 *Understanding molecular simulation* New York: Academic
[33] Gonze X 1997 *Phys. Rev.* B **55** 10337
[34] Gonze X and Lee C 1997 *Phys. Rev.* B **55** 10354
[35] Baroni S, Gianozzi P and Testa A 1987 *Phys. Rev. Lett.* **58** 1861
[36] de Gironcoli S 1991 *Phys. Rev.* B **51** 6773
[37] Alfé D, Gillan M J and Price G D 2000 *Nature* **405** 172
[38] Štich I, Payne M C, King-Smith R D, Lin J-S and Clarke L J 1992 *Phys. Rev. Lett.* **68** 1351
[39] Brommer K, Needels M, Larson B and Joannopoulos J D 1992 *Phys. Rev. Lett.* **68** 1355

[40] Marzari N, Vanderbilt D, De Vita A and Payne M C 1999 *Phys. Rev. Lett.* **82** 3296
[41] Ramamoorthy M, Vanderbilt D and King-Smith R D 1994 *Phys. Rev.* B **49** 16721
[42] Lindan P J D, Harrison N M, Holender J M and Gillan M J 1996 *Chem. Phys. Lett.* **261** 246
[43] Bates S P, Kresse G and Gillan M J 1997 *Surf. Sci.* **385** 386
[44] Harrison N M, Wang X G, Muscat J and Scheffler M 1999 *Farad. Discus.* **114** 305
[45] de Wijs G A, Kresse G, Vocadlo L, Dobson D, Alfé D, Gillan M J and Price G D 1998 *Nature* **392** 805
[46] Fujishima A and Honda K 1972 *Nature* **238** 37
[47] Henrich V E and Kurtz R L 1981 *Phys. Rev.* B **23** 6280
[48] Dubrovinsky L S, Dubrovinskaia N A, Swamy V, Muscat J, Harrison N M, Ahuja R, Holm B and Johansson B 2001 *Nature* **410** 653
[49] Henderson M A 1996 *Surf. Sci.* **355** 151
[50] Hugenschmidt M B, Gamble L and Campbell C T 1994 *Surf. Sci.* **302** 329
[51] Kurtz R L, Stockbauer R, Madey T E, Román E and de Segovia J L 1989 *Surf. Sci.* **218** 178
[52] Henrich V E and Cox P A 1994 *The Surface Science of Metal Oxides* (Cambridge: University Press)
[53] Goniakowski J, Bouette-Russo S and Noguera C 1993 *Surf. Sci.* **284** 315
[54] Goniakowski J and Noguera C 1995 *Surf. Sci.* **330** 337
[55] Bredow T and Jug K 1995 *Surf. Sci.* **327** 398
[56] Fahmi A and Minot C 1994 *Surf. Sci.* **304** 343
[57] Goniakowski J and Gillan M J 1996 *Surf. Sci.* **350** 145
[58] Lindan P J D, Harrison N M and Gillan M J 1998 *Phys. Rev. Lett.* **80** 762
[59] Gillan M J, Lindan J D, Kantorovich L N and Bates S P 1998 *Mineral. Mag.* **62** 669
[60] Lindan P J D, Muscat J, Bates S, Harrison N M and Gillan M 1998 *J. Chem. Soc. Farad. Disc.* **106** 135
[61] Lindan P J D 2002 *Chem. Phys. Lett.* submitted
[62] Segall M D, Lindan P J D, Probert M J, Pickard C J, Hasnip P J, Clark S J and Payne M C 2002 *Comp. Phys. Comm.* in preparation
[63] In object-oriented languages the term 'class' is often used for an object, and the operations it allows on the data are called 'member functions' or 'methods'
[64] In this context 'we' means the CASTEP Developers' Group: M D Segall (Cambridge), P J D Lindan (Kent), M J Probert (York), C J Pickard (Cambridge), P J Hasnip (Cambridge), S J Clark (Durham) and M C Payne (Cambridge)
[65] Kresse G and Joubert D D 1999 *Phys. Rev.* B **59** 1758
[66] Pickard C J and Mauri F 2001 *Phys. Rev.* B **63** 245101
[67] Waghmare U V and Rabe K M 1997 *Phys. Rev.* B **55** 6161
[68] Pickard C J and Mauri F 2002 *Phys. Rev. Lett.* submitted *preprint* cond–mat/0110092
[69] Rushton P P, Tozer D J and Clark S J 2002 *Phys. Rev. Lett.* submitted