IOP PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **20** (2008) 064209 (9pp) doi:10.1088/0953-8984/20/6/064209

Recent progress in linear-scaling density functional calculations with plane waves and pseudopotentials: the ONETEP code

Chris-Kriton Skylaris[1,4], Peter D Haynes[2], Arash A Mostofi[2] and Mike C Payne[3]

[1] School of Chemistry, University of Southampton, Highfield, Southampton SO17 1BJ, UK
[2] Departments of Materials and Physics, Imperial College London, Exhibition Road, London SW7 2AZ, UK
[3] Theory of Condensed Matter group, Cavendish Laboratory, J J Thomson Avenue, Cambridge CB3 0HE, UK

E-mail: cks@soton.ac.uk

Received 12 November 2007
Published 24 January 2008
Online at stacks.iop.org/JPhysCM/20/064209

Abstract

The ONETEP program employs the single-particle density matrix reformulation of Kohn–Sham density functional theory to achieve computational cost and memory requirements which increase only linearly with the number of atoms. As the code employs a plane wave basis set (in the form of periodic sinc functions) and pseudopotentials it is able to achieve levels of accuracy and systematic improvability comparable to those of conventional cubic-scaling plane wave approaches. The code has been developed with the aim of running efficiently on a variety of parallel architectures ranging from commodity clusters with tens of processors to large national facilities with thousands of processors. Recent and ongoing studies which we are performing with ONETEP involve problems ranging from materials to biomolecules to nanostructures.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

Simulations from first principles have been particularly successful in predicting properties and processes of matter. The approach of density functional theory (DFT) [1] in particular, in the formalism developed by Kohn and Sham [2] has been most widely used for simulations as it allows excellent approximations for the exchange and correlation energy with a computational cost which is significantly lower and scales more favourably than methods that use correlated wavefunctions. The computational machinery for DFT calculations has matured over the years to a number of robust and reliable approaches that can be used routinely even by non-experts. Major milestones include the plane wave pseudopotential approach [3] and Car–Parrinello approach [4] which combines nuclear dynamics with the relaxation of the electronic degrees of freedom. DFT has now become established as a practical tool for materials design and has been used in cutting edge research in academia [5]. As a consequence of these achievements, DFT calculations have been adopted by industrial researchers as a useful tool, even though not all needs of applied research have been met by the progress in theory and algorithms.

Despite the excellent progress, the computational scaling of DFT calculations is a severe obstacle that prevents them from achieving their full potential. The computational cost of conventional DFT typically increases with the third power of the number of atoms N. While this scaling is still more favourable than the steeper scaling of most correlated wavefunction approaches, it still limits the applicability of DFT methods to no more than a few hundred atoms, even when supercomputers are employed for the calculations. There are therefore difficulties in the applicability of DFT in cases where the description of the interactions between thousands of atoms is necessary as in the realistic modelling of nanostructures [6] or in problems in molecular biology. Therefore efforts to produce DFT approaches with linear-scaling cost began more

[4] http://www.soton.ac.uk/chemistry/research/skylaris/skylaris.html

than a decade ago; the main theoretical principles that underlie the various methods are summarized in [7]. The development of practical linear-scaling methods requires the solution of many technical problems and was therefore not as rapid as originally predicted. Significant progress has been made however and currently a number of codes [8–14] with linear-scaling capabilities are available.

The purpose of this paper is to provide an outline of current applications with our linear-scaling DFT code ONETEP [8] which aims to achieve the high level of accuracy and systematicity of plane wave approaches. In section 2 we provide an overview of the theory on which ONETEP is based and of its implementation for parallel computers. Then in section 3 we describe recent and ongoing applications with the code and we conclude with a summary and future outlook in section 4.

2. Overview of the ONETEP method

2.1. Density matrix formulation

The set of Kohn–Sham orbitals $\{\psi_n(\mathbf{r})\}$ provides a complete description of the fictitious system of non-interacting particles in DFT. An equivalent description may also be given by the single-particle density matrix

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_n f_n \psi_n^*(\mathbf{r}) \psi_n(\mathbf{r}') \tag{1}$$

which possesses the property of idempotency,

$$\rho^2(\mathbf{r}, \mathbf{r}') = \int \mathrm{d}^3 r'' \rho(\mathbf{r}, \mathbf{r}'') \rho(\mathbf{r}'', \mathbf{r}') = \rho(\mathbf{r}, \mathbf{r}') \tag{2}$$

which implies the orthonormality of the orbitals $\psi_n(\mathbf{r})$ and the requirement that the occupancies f_n are equal either to one for all states up to the chemical potential or zero for all other states, according to the antisymmetry requirement of the electronic wavefunction. The density matrix in Kohn–Sham theory is thus the position representation of the projection operator onto the space of occupied states $\hat{\rho}$. The density $n(\mathbf{r})$ may be obtained from the diagonal elements of the density matrix,

$$n(\mathbf{r}) = 2\rho(\mathbf{r}, \mathbf{r}) \tag{3}$$

where the factor of 2 accounts for spin degeneracy (assuming no spin polarization). The total energy of the interacting system is given by

$$E[n] = -\int \mathrm{d}^3 r \, [\nabla^2_{\mathbf{r}'} \rho(\mathbf{r}, \mathbf{r}')]_{\mathbf{r}'=\mathbf{r}} + \int \mathrm{d}^3 r \, v_{\mathrm{ext}}(\mathbf{r}) n(\mathbf{r}) + E_{\mathrm{H}}[n] + E_{\mathrm{xc}}[n] \tag{4}$$

where the terms on the right-hand side of the above equation are the kinetic energy, the energy due to the external potential(s), the Hartree energy and the exchange and correlation energy, in accordance with standard notation [15] in atomic units and the definition of density in equation (3). Using a variational methodology [16], the total energy can be obtained by minimizing the expression of equation (4) with respect to the density matrix, subject to the constraints of idempotency (2) and normalization,

$$N_{\mathrm{e}} = 2 \int \mathrm{d}^3 r \, \rho(\mathbf{r}, \mathbf{r}) \tag{5}$$

i.e. the density matrix corresponds to a system of N_{e} electrons.

Since the number of occupied states is directly proportional to N and each state extends over the whole system, the amount of information in the density matrix defined by equation (1) scales as N^2. Any calculation involving manipulation of this density matrix will therefore scale quadratically with system size at best. In order to obtain a linear-scaling method, it is necessary to exploit the nearsightedness property [17, 18] of many-body quantum mechanics.

Both analytical [19, 20] and numerical [21] studies have demonstrated that, for an insulating system, both the Wannier functions and density matrix decay exponentially, so that

$$\rho(\mathbf{r}, \mathbf{r}') \sim \exp(-\gamma |\mathbf{r} - \mathbf{r}'|) \to 0 \qquad \text{as } |\mathbf{r} - \mathbf{r}'| \to \infty. \tag{6}$$

The decay rate γ depends only on the energy gap between the highest occupied and lowest unoccupied states, and not on the system size, therefore the total amount of significant information it contains scales linearly with N. In practice this is exploited by writing the density matrix in the following form:

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{\alpha\beta} \phi_\alpha(\mathbf{r}) K^{\alpha\beta} \phi_\beta^*(\mathbf{r}') \tag{7}$$

where the $\{\phi_\alpha\}$ are a set of spatially localized non-orthogonal functions which span a superspace of the Hilbert space of the set of occupied Kohn–Sham orbitals and in ONETEP they are called non-orthogonal generalized Wannier functions (NGWFs) [22]. The matrix $K^{\alpha\beta}$, known as the density kernel [23] is the representation of the density matrix in the set of duals of the NGWFs. The advantage of this form is that it allows the nearsightedness to be exploited with the use of spatial cut-offs. First, the NGWFs which are exponentially localized are truncated, by allowing them to be non-vanishing only in spherical regions of fixed radii r_α and centred at positions $\mathbf{R}_\alpha$. In ONETEP, a number of NGWFs are associated with each atom in the system, so that the regions are centred on atoms and their radii depend only on the atomic species. Second, the density kernel is required to be a sparse matrix, and this is achieved by discarding elements $K^{\alpha\beta}$ corresponding to NGWFs centred further apart than some cut-off $r_{\mathbf{K}}$. Note that since the density kernel is constructed from the duals of the NGWFs, $r_{\mathbf{K}}$ cannot simply be defined as the sum of the NGWF radii. In ONETEP the total energy (4) is optimized by direct minimization with respect to both the density kernel and the NGWFs [24]. This *in situ* optimization of the NGWFs leads to high accuracy and eliminates difficulties such as the basis set superposition error (BSSE) [25] which are present in methods with fixed localized functions and can severely compromise the reliability of the results obtained. The lack of BSSE is a consequence of the flexibility of the NGWFs to adopt any shape within their localization region and is not affected by the fact that the r_α are finite. Imposing spatial cut-offs on the NGWFs and the density kernel results in a density matrix whose information content scales linearly with system size, an approximation which is controlled by adjusting the r_α and $r_{\mathbf{K}}$. In practice, these cut-offs are increased until the desired physical properties of the system converge [26].

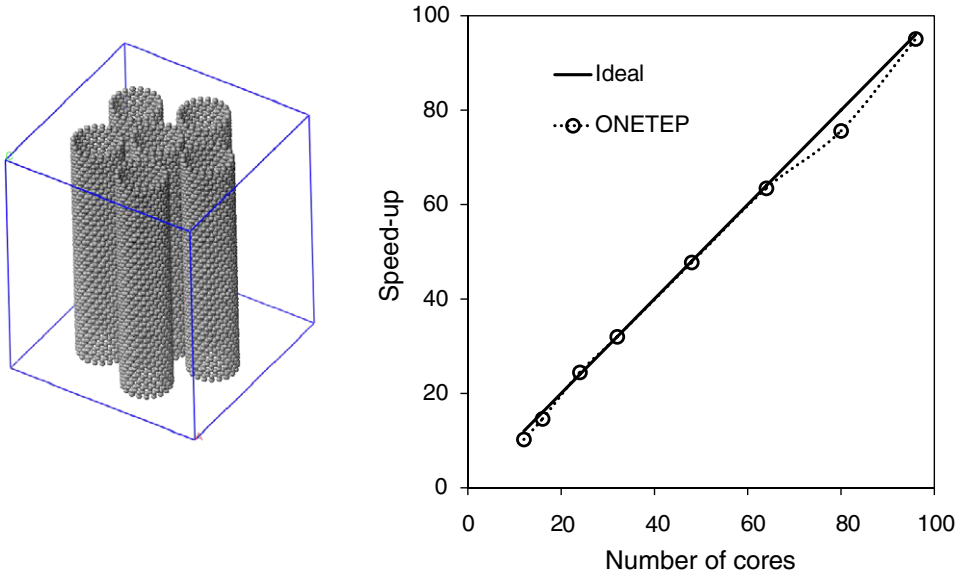


Figure 1. Speed-up of ONETEP calculations on a 9600-atom carbon nanorope segment (shown on the left) as a function of the number of processors. The calculations were performed on a commodity cluster consisting of 24 dual-socket dual-core nodes.

Recently we have implemented atomic forces and the capability to perform geometry optimization and Born–Oppenheimer molecular dynamics simulations. The theoretical methodology behind these developments and examples demonstrating their capabilities will be presented elsewhere.

2.2. Parallel implementation

The ONETEP code has been developed from the beginning for parallel computers. The implementation is highly portable and is based on a distributed data model where communication between processors is achieved through use of the message passing interface (MPI) [27] library. A detailed description of the parallel implementation has been presented [28]. The code and its parallel algorithms are under constant development and we have made several important improvements in the parallel algorithms since the publication of this paper. Most notably these include improvements in our sparse matrix algebra module which now distributes the matrices to processors by columns which are of course related to atoms for which we have a very efficient data distribution scheme [28].

To assess the efficiency of the code as it currently stands we have performed a series of tests which are summarized in the plot of figure 1. The plot shows the speed-up in the time of a self-consistent iteration in ONETEP (density kernel and NGWF geometry optimization steps) as a function of the number of processors. The calculations were performed on a 'nanorope' made of six (20,0) nanotube segments, with 9600 atoms in total. An orthorhombic unit cell of 70.0 Å × 70.0 Å × 85.2 Å and a plane wave kinetic energy cut-off of 470 eV [29] for the psinc basis were used. The density matrix cut-offs were set to $r_\alpha = 3.7$ Å and $r_{\mathbf{K}} = 7.9$ Å. The calculations were performed on a cluster consisting of 24 IBM 326m nodes connected with a 24-port Infiniband switch. Each node contained 8 GB of RAM and two Dual Core AMD Opteron 285 processors operating at a clock speed of 2.6 GHz. The plot of figure 1 was obtained by setting the speed-up at 32 processors exactly equal to 32, the smallest number of processors where memory limitations allowed all four cores per node to be used (at 24, 16 and 12 processors 2, 1 and 1 cores per node were used respectively). The wall clock time taken for one SCF iteration on 96 processors is 2.5 h.

The plot shows that a speed-up of 95.1 is achieved when 96 processing cores are used, which is the maximum available on the cluster on which these tests were conducted. Similar clusters are nowadays widely available amongst research groups as they are rather inexpensive and relatively easy to maintain. Our results indicate that the code is likely to show the same extremely good parallel scalability on such commodity clusters. For far larger calculations, national supercomputing facilities are now available which offer opportunities to run on thousands of cores. The performance of the code on such platforms has not been explored yet but we are currently seeking opportunities to do this in the near future.

3. Recent applications with ONETEP

ONETEP is a newly developed code but already it has found use in a multitude of problems involving materials that range from bulk crystalline solids, to biomolecules, to nanostructures. In this section we take a tour though some of the applications which we we have recently completed or are currently working on with the code.

3.1. Crystalline silicon

Periodic crystalline solids are particularly challenging for linear-scaling approaches as their structure involves a high atom-density that consequently reduces the sparsity of matrices such as the local-orbital representation of the Hamiltonian. This leads to an increase in the number of atoms where a linear-scaling approach begins to outperform conventional cubic scaling approaches, known as the 'cross-over' point.

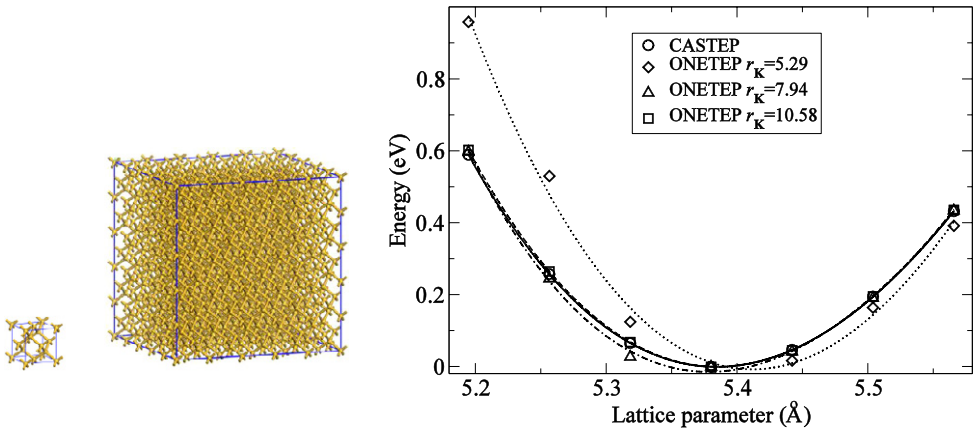


Figure 2. Energy as function of lattice for crystalline silicon as obtained from ONETEP calculations with different values of the kernel cut-off $r_{\mathbf{K}}$, in Å. The curve obtained with the cubic-scaling plane wave code CASTEP is also shown as well as the simulation cells used for the calculations (1000-atom cell for ONETEP and 8-atom cell for CASTEP).

If the solid happens to be a semiconductor the difficulties are exacerbated as the small band gap decreases the rate of decay of the density matrix. Crystalline silicon is therefore a particularly difficult example for a density matrix based linear-scaling code.

Despite these difficulties, we have recently performed a detailed study with ONETEP on crystalline silicon [26] through which we have shown that it is possible, while using linear-scaling conditions and algorithms, to achieve results of plane wave accuracy. To perform this study we have used a simulation cell of 1000 atoms which is sufficiently large to allow testing of various values of the density kernel cut-off $r_{\mathbf{K}}$. The CASTEP [30] code which is an implementation of the conventional cubic-scaling plane wave pseudopotential approach, was used as an accuracy benchmark by performing calculations in a unit cell of 8 atoms with a mesh of $5 \times 5 \times 5$ k-points set up so that it imitates the 1000-atom cell of ONETEP. Some of the calculations are summarized in figure 2 which shows the behaviour of the energy as a function of the lattice parameter for several values of $r_{\mathbf{K}}$ with r_α fixed to 3.7 Å. In going from $r_{\mathbf{K}} = 5.29$ to 10.58 Å, the value of the lattice constant calculated by fitting the graph to the Birch–Murnaghan equation of state [31] goes from 5.412 to 5.385 Å. Also the bulk modulus, a property which is very sensitive to calculation parameters, varies from 122.7 to 97.8 GPa. The 'correct' values from CASTEP are 5.384 Å and 96.3 GPa respectively. Accuracy comparable to this of conventional cubic-scaling plane wave calculations can therefore be achieved even for this rather demanding case.

3.2. $BaTiO_3$

Another example of a crystalline solid that is being studied using ONETEP is the perovskite barium titanate (BTO), $BaTiO_3$. It is viewed as the model ferroelectric, undergoing a displacive phase transition from cubic to tetragonal structure at 120 °C. Ferroelectric materials possess a permanent electric dipole moment which may be reversed or reoriented by the application of an electric field. They are thus of great technological interest, particularly within the microelectronics industry where one promising application is computer memory. Alloyed with strontium titanate, $Ba_{1-x}Sr_xTiO_3$ is a leading candidate for the next generation of dynamic random access memory. The relatively small sizes of traditional DFT calculations restrict the variation of Sr concentration x attainable by straightforward atomic substitution, a constraint that may be relaxed using ONETEP without resorting to the virtual crystal approximation. Larger simulations also permit the realistic simulation of defects such as oxygen vacancies, which play a vital role in these materials, particularly their interaction with grain boundaries. The study of such materials also provides an opportunity to assess the physical significance of the NGWFs generated by ONETEP. The electronic polarization of an insulating crystalline solid may be related to the centres of charge of the valence band Wannier functions [32, 33], allowing the calculation of the Born effective charges for BTO [34]. This formal connection applies equally to the maximally-localized Wannier functions (MLWFs) [35] generated from the Bloch functions of traditional DFT calculations. The extent to which the NGWFs reflect this relationship has yet to be established. Figure 3 shows the qualitative change that occurs when pseudoatomic orbitals are optimized *in situ* to generate NGWFs. The familiar s, p and d orbital shapes of the former, resulting from the spherical symmetry of the closed-shell ionic cores, is evident, whereas the symmetry of the NGWFs reflects that of the crystal. In addition, radial nodes have been introduced as would be required in the MLWFs to maintain orthogonality, even though this constraint is not imposed on the NGWFs. Band structure calculations obtained by taking linear combinations of NGWFs to generate Bloch functions also indicate that in addition to giving an accurate measure of average properties (such as the total energy), they also describe the details of the electronic structure. It is therefore likely that the NGWFs themselves may be used to calculate observable properties directly, providing a further advantage to linear-scaling methods based on *in situ* optimized local orbitals.

3.3. Protein–ligand binding

Studies of biomolecules are usually carried out with classical force field approaches [36]. The major advantage of force

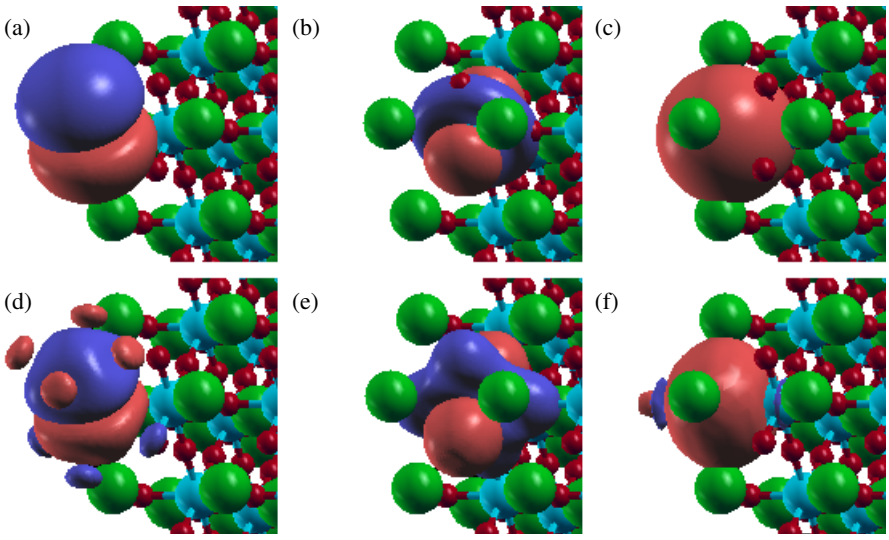
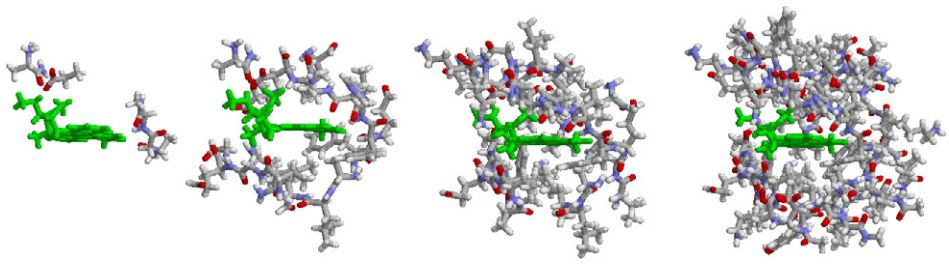


Figure 3. Examples of the *in situ* optimization of NGWFs in barium titanate. (a)–(c) The pseudoatomic orbitals used as initial guesses for the NGWFs are generated from spherically symmetric ionic cores: (a) p orbital on Ba, (b) d orbital on Ti and (c) s orbital on O. (d)–(f) The optimized NGWFs on the same atoms now reflect the crystal symmetry.



Binding energy (eV):				
ONETEP	1.404	1.686	1.760	1.755
CASTEP	1.388	---	---	---

Figure 4. Binding energy of a staurosporine ligand into the ATP pocket of the CDK2 protein as a function of the CDK2 fragment size.

fields is that they can provide the energy as a function of atomic coordinates at a computational cost which is several orders of magnitude smaller that that of a first principles quantum mechanical calculation. As a consequence, force fields are used routinely to perform dynamical simulations as they allow to access timescales (usually a few ns, depending on total number of atoms) that are often long enough to allow sufficient sampling of phase space for the simulation of dynamical processes and also the calculation of various thermodynamic quantities through the laws of statistical mechanics. Nevertheless, classical force fields are by construction empirical and the quality of their parameterization is a constant concern. Most force fields also have inherent limitations such as the inability to form or break chemical bonds and the lack of electronic effects such as polarization and charge transfer.

These effects need to be taken into account for a qualitatively correct description of the energetic contributions in various situations such as for example when studying the binding of a small ligand molecule into the cavity of a protein. The procedure that is usually followed to tackle such problems involves using a mixed quantum mechanical/molecular mechanical (QM/MM) description where the quantum mechanical description includes the ligand and a small part of its binding cavity. This approach is far from ideal as it gives rise to other concerns such as convergence of the important contributions with the size of the quantum region and how and where to define the interface of the quantum region with the classical region.

Having the capability to perform large-scale first principles calculations with ONETEP we have decided to use a fully quantum description to investigate the binding of small ligands to the CDK2 protein [37], thus avoiding the ambiguities that exist in QM/MM approaches. The first part of our study consisted of a calibration stage where the binding energy was calculated for increasing sizes of the protein binding pocket, as shown in figure 4. Our results show that convergence to less than 0.01 eV ($\simeq$0.23 kcal mol^{-1}) is achieved with protein fragments with about 1000 atoms, which is better than the 'chemical accuracy' threshold of 1 kcal mol^{-1}.

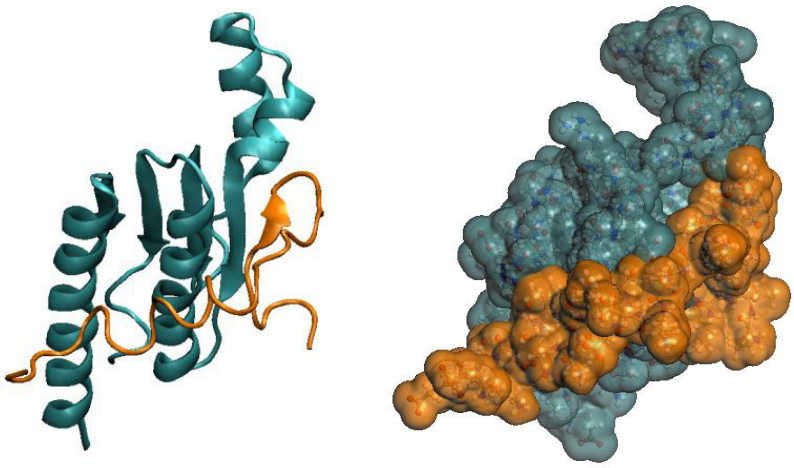


Figure 5. Left: a ribbon diagram showing the tertiary structure of one RAD51-BRC complex. Right: density isosurfaces of the constituent proteins of the complex from ONETEP calculations.

We have subsequently carried out long classical molecular dynamics simulations of the proteins in explicit water at room temperature. This approach is necessary in order to capture the dynamical behaviour of these systems as certain ligand-protein interactions are periodically established and disrupted during the simulation. DFT calculations on ensembles of structures (including water molecules stabilized in the protein) were then used to compute the relative average binding energies of the ligands. The relative binding affinities of the ligands obtained from these average energies are in remarkable agreement with experimentally determined values [37]. Therefore such a combination of linear-scaling DFT with classical dynamics simulations results in an approach that appears to have the sensitivity to further screen inhibitors that have been selected by cruder drug optimization approaches.

3.4. Protein–protein interactions

Most biological processes essential for cellular regulation involve the regulated formation and disassembly of macromolecular complexes between proteins. The rational design of small molecules that inhibit protein–protein interactions is an important goal for drug discovery but it also offers the potential to deepen our understanding of the chemical interactions that underpin biological function. However, the design of such small molecule inhibitors (SMIs) remains very difficult, despite the growing evidence that such an approach is feasible, at least in principle, for many protein–protein interactions of key biological interest. One major difficulty in evaluating the 'druggability' of specific protein–protein interactions, and subsequently in rational SMI design, is the high complexity of the atomic interactions that contribute to the binding energy of the complex. A major challenge, therefore, is to develop approaches to dissect the energetic contributions of particular amino acid residues to the overall binding energy of a given protein–protein interaction. This problem is more challenging than that of a drug molecule docking into the binding pocket of one protein as the area of interaction usually involves the entire proteins.

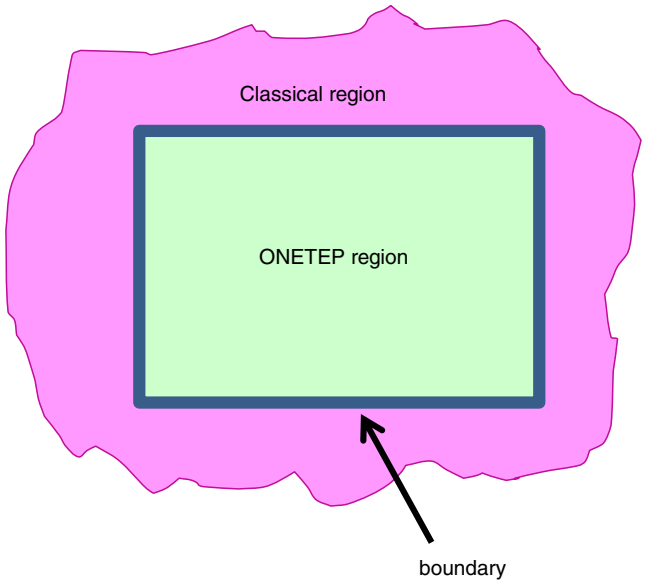


Figure 6. ONETEP with Dirichlet boundary conditions: the Hartree potential is obtained by solving the Poisson equation in the ONETEP region.

An important biomolecule whose function involves protein–protein interactions is the breast cancer susceptibility protein BRCA2 [38, 39]. This is a massive protein of approximately 3500 amino acids. The active region of this protein consists of 8 similar subunits, of 30–40 amino acids each, which are called the BRC repeats. Each BRC repeat binds with different strengths to the RAD51 DNA recombinase protein. This is necessary in order to position RAD51 onto double strand breaks in DNA and initiate its repair though the process of homologous recombination. Initial molecular dynamics simulations we have carried out [40] in this system show that the beta-hairpin structure of the BRC repeat is stabilized by numerous intramolecular hydrogen bonding backbone–backbone, backbone–sidechain and sidechain–sidechain interactions and as a result, even when it is not bound to RAD51, it retains a conformation that closely

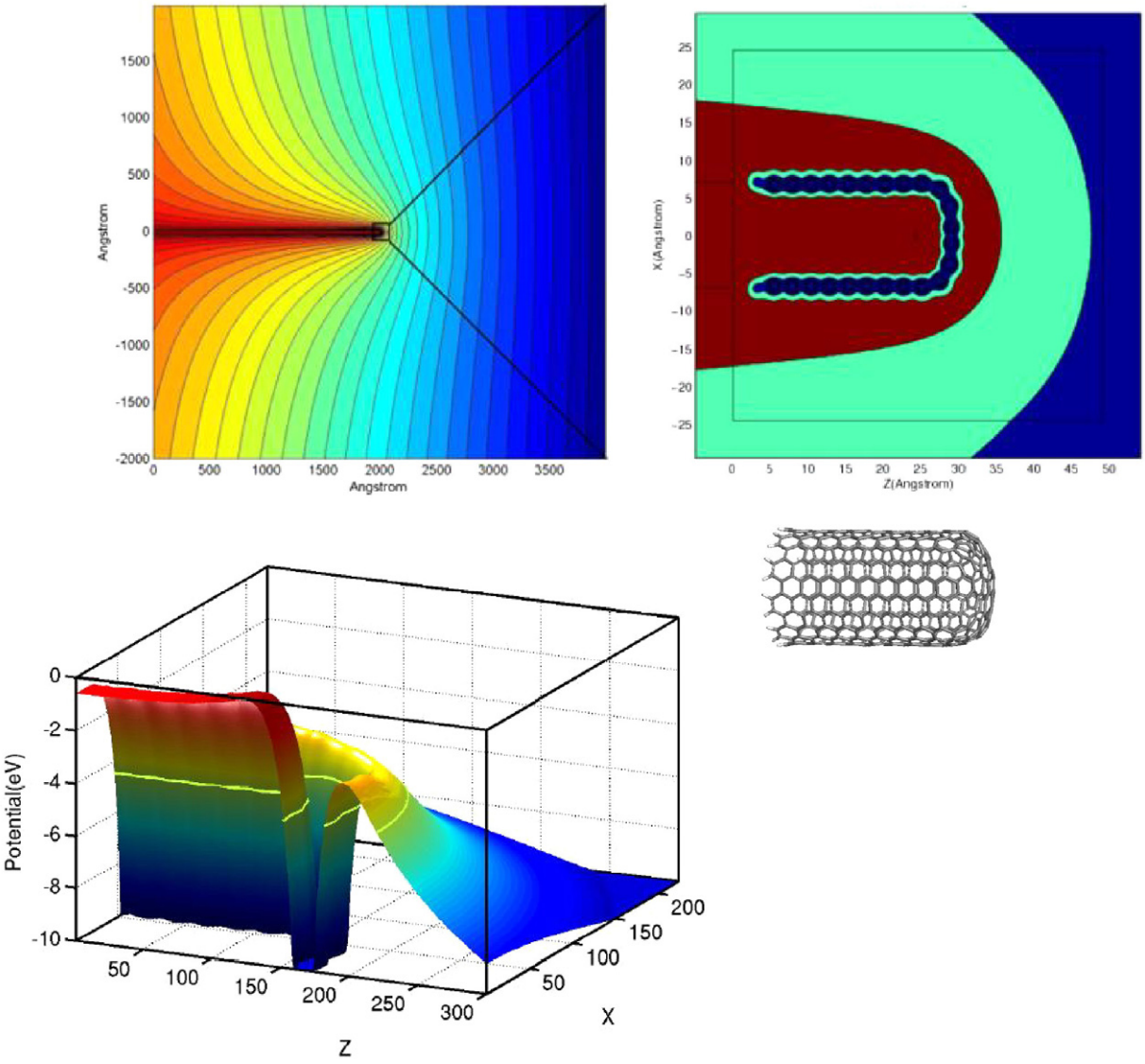


Figure 7. The top left panel shows a classical electrostatic calculation where the nanotube is represented as a 200 nm-long classical conductor in a cell where the plates that create the electrostatic field are 400 nm apart. The top right panel shows the cell of the ONETEP calculation which has a side of 5 nm and is embedded in the classical calculation. The bottom panel shows a slice of the electrostatic potential at the tip, where atomic details and the barrier that the electrons have to overcome are clearly visible. The Fermi level is also depicted.

resembles the bound form. We are currently using ONETEP calculations to evaluate the relative binding strengths of the different BRC repeats. Figure 5 shows the protein structure in one of our calculations. For each repeat the binding energy is obtained from ONETEP calculations on an ensemble of structures from the classical dynamics simulation. About 2500 atoms are involved in each such single point energy calculation. The role of the solvent (i.e. water and counter ions) in driving the association of the protein complexes (e.g. the hydrophobic force), is currently not taken into account in our calculations. We are currently implementing in ONETEP implicit solvation models with which we will subsequently investigate these effects.

3.5. Carbon nanotubes in electric fields

The ONETEP code has been constructed to use periodic boundary conditions. This however does not prevent calculations on isolated molecules. Isolated molecules can be described accurately and efficiently through the supercell approximation as the FFT box technique [41] on which the code is based allows for calculations in very large simulation cells so that molecules can be effectively isolated from their periodic images. There are nevertheless cases where it is useful to have an explicitly-defined non-periodic boundary, and we have recently implemented [42] such a scheme in ONETEP.

In standard ONETEP calculations Fourier transform techniques are used to solve the Poisson equation for the Hartree potential in momentum space where, by definition, a periodic boundary is imposed. Our new scheme involves solving for the Hartree potential in real space with its values predefined in the simulation cell boundary (Dirichlet boundary conditions). The values of the potential at the boundary obviously need to be provided. These are usually obtained from a classical elec-

trostatic model as demonstrated in figure 6, which can be of much larger scale than the ONETEP calculation. The rest of the calculation is performed using the standard ONETEP methodology [22, 43, 44]. We have used this scheme to calculate the electrostatic potential around the tip of capped metallic single-walled carbon nanotubes inside a uniform external electric field. The strength of the external field was set below but near the values where electronic field emission is experimentally observed [45]. The actual set up, including the electrostatic potential obtained in one of our calculations is shown in figure 7. It is obvious that this system cannot be described correctly with periodic boundary conditions. A much larger classical electrostatic calculation is performed where the nanotube is simply represented as a 'conductor' and has length comparable to nanotubes used in experiments (hundreds of nm). This allows the correct charge built up at the tip of the tube and the potential distribution around it to be obtained. Then the ONETEP calculation is 'embedded' in the classical calculation to obtain in atomic detail the potential around the tip as well as other relevant quantities such as the electronic charge density and the energy levels. As a result, the exact shape of the barrier that the electrons have to overcome is obtained and its dependence on the type of nanotube and the shape of its cap can be investigated and compared with experiment [46].

Obviously this electrostatic embedding approach is very flexible and can be applied to other kinds of problems which require explicit non-periodic boundary conditions.

4. Conclusions and outlook

We have reviewed recent methodological developments and applications with the ONETEP linear-scaling DFT code. The code is capable of running efficiently on parallel computers with an arbitrary number of processors and we have already been using it in numerous applications. Results show that with ONETEP we are able to achieve plane wave accuracy in large-scale calculations on crystalline solids such as in the case of a 1000-atom unit cell of crystalline silicon. Calculations on other crystalline materials of industrial importance such as barium titanate ferroelectric perovskite are in progress. Applications of the code in protein–ligand and protein–protein interactions have proved that it is a very powerful tool for biomolecular studies as it allows for the accurate description of all energetic contributions on entire biomolecules. The code is also being used for calculations on nanostructures such as capped carbon nanotubes where is allows us to obtain a correct atomistic description of the electronic density and electrostatic potential under external electric fields.

It is obvious that linear-scaling DFT is now a reality and it allows problems which require large-scale DFT calculations to be tackled. ONETEP is a state-of-the-art method for such calculations. It is under continuous development and we therefore expect its range of capabilities to increase in the near future. Its user base in academia and industry is steadily increasing so increases are also expected in the number of studies that are performed with it.

Acknowledgments

C-KS and PDH would like to thank the Royal Society for University Research Fellowships. We would also like to thank the Southampton University Information Systems services for support with supercomputing facilities.

References

[1] Hohenberg P and Kohn W 1964 *Phys. Rev.* **136** B864
[2] Kohn W and Sham L J 1965 *Phys. Rev.* **140** A1133
[3] Payne M C, Teter M P, Allan D C, Arias T A and Joannopoulos J D 1992 *Rev. Mod. Phys.* **64** 1045
[4] Car R and Parrinello M 1985 *Phys. Rev. Lett.* **55** 2471
[5] Hafner J, Wolverton C and Ceder G 2006 *MRS Bull.* **31** 659
[6] Marzari N 2006 *MRS Bull.* **31** 681
[7] Goedecker S 1999 *Rev. Mod. Phys.* **71** 1085
[8] Skylaris C-K, Haynes P D, Mostofi A A and Payne M C 2005 *J. Chem. Phys.* **122** 084119
[9] Bowler D R, Choudhury R, Gillan M J and Miyazaki T 2006 *Phys. Status Solidi* b **243** 989
[10] Soler J M, Artacho E, Gale J D, García A, Junquera J, Ordejón P and Sánchez-Portal D 2002 *J. Phys.: Condens. Matter* **14** 2745
[11] Fattebert J L and Bernholc J 2000 *Phys. Rev.* B **62** 1713
[12] Challacombe M 1999 *J. Chem. Phys.* **110** 2332
[13] VandeVondele J, Krack M, Fawzi M, Parrinello M, Chassaing T and Hutter J 2005 *Comput. Phys. Commun.* **167** 103
[14] Fattebert J-L and Gygi F 2006 *Phys. Rev.* B **73** 115124
[15] Martin R M 2004 *Electronic Structure. Basic Theory and Practical Methods* (Cambridge: Cambridge University Press)
[16] Skylaris C-K, Diéguez O, Haynes P D and Payne M C 2002 *Phys. Rev.* B **66** 073103
[17] Kohn W 1996 *Phys. Rev. Lett.* **76** 3168
[18] Prodan E and Kohn W 2005 *Proc. Natl Acad. Sci.* **102** 11635
[19] Kohn W 1959 *Phys. Rev.* **115** 809
[20] Des Cloizeaux J 1964 *Phys. Rev.* **135** A685
[21] Ismail-Beigi S and Arias T A 1999 *Phys. Rev. Lett.* **82** 2127
[22] Skylaris C-K, Mostofi A A, Haynes P D, Diéguez O and Payne M C 2002 *Phys. Rev.* B **66** 035119
[23] McWeeny R 1960 *Rev. Mod. Phys.* **32** 335
[24] Haynes P D, Skylaris C-K, Mostofi A and Payne M C 2006 *Phys. Status Solidi* b **243** 2489
[25] Haynes P D, Skylaris C-K, Mostofi A A and Payne M C 2006 *Chem. Phys. Lett.* **422** 345
[26] Skylaris C-K and Haynes P D 2007 *J. Chem. Phys.* **127** 164712
[27] *Message Passing Interface Forum* http://www.mpi-forum.org/
[28] Skylaris C-K, Haynes P D, Mostofi A A and Payne M C 2006 *Phys. Status Solidi* b **243** 973
[29] Skylaris C-K, Haynes P D, Mostofi A A and Payne M C 2005 *J. Phys.: Condens. Matter* **17** 5757
[30] Clark S J, Segall M D, Pickard C J, Hasnip P J, Probert M I J, Refson K and Payne M C 2005 *Z. Kristallogr.* **220** 567
[31] Murnaghan F D 1944 *Proc. Natl Acad. Sci.* **30** 244
[32] King-Smith R D and Vanderbilt D 1993 *Phys. Rev.* B **47** 1651
[33] Vanderbilt D and King-Smith R D 1993 *Phys. Rev.* B **48** 4442
[34] Ghosez P, Gonze X, Lambin P and Michenaud J-P 1995 *Phys. Rev.* B **51** 6765
[35] Marzari N and Vanderbilt D 1997 *Phys. Rev.* B **56** 12847
[36] Case D, Cheatham T III, Darden T, Gohlke H, Luo R, Merz K M Jr, Onufriev A, Simmerling C, Wang B and Woods R J 2005 *J. Comput. Chem.* **26** 1668
[37] Heady L, Fenandez-Serra M, Joyce S, Venkitaraman A R, Artacho E, Skylaris C-K, Ciacchi C and Payne M C 2006 *J. Med. Chem.* **49** 5141
[38] Venkitaraman A R 2002 *Cell* **108** 171

[39] Pellegrini L, Yu D S, Lo T, Anand S, Lee M, Blundell T L and Venkitaraman A 2002 *Nature* **420** 287
[40] Buis N, Skylaris C-K, Grant G, Payne M C, Venkitaraman A R and Rajendra E 2007 *J. Chem. Theor. Comput.* submitted
[41] Skylaris C-K, Mostofi A A, Haynes P D, Pickard C J and Payne M C 2001 *Comput. Phys. Commun.* **140** 315
[42] Skylaris C-K 2006 unpublished results
[43] Mostofi A A, Skylaris C-K, Haynes P D and Payne M C 2002 *Comput. Phys. Commun.* **147** 788
[44] Mostofi A A, Haynes P D, Skylaris C-K and Payne M C 2003 *J. Chem. Phys.* **119** 8842
[45] Amaratunga G 2003 Spectrum *IEEE* **40** 28
[46] Skylaris C-K, Peng J, Gsanyi G, Payne M, Nevidomskyy A and Edgcombe C 2007 in preparation